Stratégies de génotypage pour la sélection génomique chez la poule pondeuse

Florian Herry

UNIVERSITE ECOLOGIE
BRETAGNE GEOSCIENCES
LOIRE AGRONOMIE ALIMENTATION

AGRO CAMPUS OUEST

THESE DE DOCTORAT DE

AGROCAMPUS OUEST
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 600
Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation
Spécialité : Génétique, Génomique et Bio-informatique

Par
Florian HERRY

Stratégies de génotypage pour la sélection génomique chez la poule pondeuse

Thèse présentée et soutenue à Rennes, le 19 décembre 2019
Unité de recherche : Physiologie, Environnement, et Génétique pour l'Animal et les Systèmes d'Élevages – INRA – Agrocampus Ouest – Équipe Génétique et Génomique
Thèse N° : B-330_2019-25

Rapporteurs avant soutenance :

Tom DRUET
Maître de Recherches, F.R.S.-FNRS
Unité de Génomique Animale, GIGA-R

Juliette RIQUET
Directrice de Recherches, UMR GenPhySE
Génétique, Physiologie et Systèmes d'Élevage

Composition du Jury :

Maria MANZANARES-DAULEUX, Présidente
Professeure, Agrocampus Ouest, UMR IGEPP

Tom DRUET, Rapporteur
Maître de Recherches, F.R.S.-FNRS, GIGA-R

Juliette RIQUET, Rapportrice
Directrice de Recherches, UMR GenPhySE

Sophie BRARD-FUDULEA, Examinatrice
Chef de Projet Génomique, SYSAAF, UMR BOA

Thierry BURLOT, Co-Directeur de thèse
Directeur R&D, Novogen

Pascale LE ROY, Directrice de thèse
Directrice de Recherches, UMR PEGASE

Sophie ALLAIS, Co-encadrante de thèse
Maître de conférences, Agrocampus Ouest, UMR PEGASE

À ma famille

Remerciements

En regardant quasiment 4 ans en arrière et le déroulement de ces années passées, il serait presque possible de se dire que, finalement, la thèse « est un long fleuve tranquille ». Il est important et nécessaire de souligner la qualité de l'encadrement et des personnes ayant permis, de près comme de loin, de faire en sorte que ma thèse se passe de cette façon. Je souhaite donc remercier toutes ces personnes qui m'ont aidé lors de ce travail.

En tout premier lieu, je tiens à remercier ma directrice de thèse Sophie Allais, pour avoir d'abord accepté de me prendre en stage de fin d'études. Et ceci alors que durant mon cursus d'ingénieur agronome je me sois, à chaque fois, à chaque examen, planté dans mon calcul de déséquilibre de liaison… Mais cela ne nous a pas empêché de poursuivre l'aventure avec une thèse. Cela fut un réel plaisir de travailler avec toi au cours de ces années. Dynamique, bienveillante, rassurante lorsqu'il le fallait et toujours de bon conseil. Tu as également su trouver le parfait équilibre entre autonomie et encadrement, ce qui m'a permis d'avancer à mon rythme et de réaliser cette thèse. Merci d'avoir corrigé tout ce que j'ai soumis à ta relecture (et il y en a eu des pages de relues !) et d'avoir su raccourcir lorsqu'il le fallait, mes articles « un chouïa » longs. J'ai énormément appris au cours de ces années et je suis ravi d'avoir été ton premier thésard. Merci pour tout.

Je tiens aussi à remercier ma directrice de thèse Pascale Le Roy, pour avoir également accepté de me prendre en stage de fin d'études et pour poursuivre avec une thèse. Merci pour tes conseils et remarques qui ont permis de faire progresser les différents travaux lors de nos réunions hebdomadaires, ainsi que pour nos discussions. Merci également d'avoir relu un grand nombre de pages que j'ai soumis à ta relecture !

Une thèse réussie passe souvent par un ou des bons directeurs de thèse. Je le dis sans aucun doute, cette thèse fut excellente sur ce point.

Je souhaite remercier Thierry Burlot et l'entreprise Novogen d'avoir financé ma thèse et d'avoir accepté de se lancer dans l'aventure des thèses CIFRE. Je tiens aussi à remercier Amandine Varenne. Merci à vous deux pour avoir répondu à toutes mes questions et pour m'avoir fait découvrir le monde de la sélection avicole. Merci également de m'avoir fait confiance dans les différents travaux et de m'avoir permis de les présenter à de très nombreux congrès nationaux comme internationaux dans des pays comme la République Tchèque, la Turquie ou encore la Nouvelle-Zélande.

Un grand merci également à Frédéric Hérault pour tes conseils, tes remarques et ton aide, aussi bien pendant qu'en dehors de nos réunions hebdomadaires. Merci aussi d'avoir dépassé la barrière du nombre de pages et d'avoir relu tout ce que j'ai soumis à ta relecture.

Je tiens également à remercier tous les membres de mon comité de thèse, à savoir Nabeel Alnahhas, Alban Bouquet, Stéphane Nicolas, Christèle Robert-Granié et Romain Saintilan. Merci également à Olivier Demeure. Je vous remercie tous pour votre présence aux différents comités de thèse, pour nos échanges toujours très constructifs et vos conseils qui ont contribué et permis de faire avancer les différents travaux.

Merci aussi à Yannick Le Cozler d'avoir accepté d'être mon tuteur au cours des 3 dernières années. Tu as parfaitement rempli ton rôle en ayant su me conseiller et me rassurer à certains moments. Merci également d'avoir assisté aux différents comité de thèse malgré un sujet quelque peu éloigné de ton domaine de prédilection.

Je remercie évidement Tom Druet et Juliette Riquet d'avoir accepté les rôles de rapporteur et rapportrice de la thèse et pour les retours bienveillants. Merci également à Maria Manzanares-Dauleux d'avoir accepté le rôle de Présidente du jury et à Sophie Brard pour le rôle d'examinatrice.

Merci également à mon collègue thésard David Picard Druet pour nos échanges et tes travaux. Et pour finir le tour des participants à nos réunions hebdomadaires, merci à Nicolas Bédère pour tes remarques « tatillonnes » et conseils pour la préparation de la soutenance et pour les bons moments passés à Prague !

Merci également à Frédéric Lecerf pour toutes nos nombreuses discussions et tes conseils plus appliqués « informatique » ou sur mes différentes présentations. Tu es également un grand contributeur à la bonne ambiance des pauses café ou du repas du midi ! En ce sens, je remercie également Frédéric Jehl ainsi que sa directrice de thèse Sandrine Lagarrigue, Laetitia Lagoutte (merci aussi pour ton temps passé sur le fameux projet RAD-Seq !), Morgane Boutin, Colette Désert, Kévin Muret, ainsi que Pauline Philippe et ton énergie faisant oublier une arrivée seulement très récente ! Et merci également à toute l'équipe de Productions Animales.

Vient maintenant le tour des amis extérieurs à la thèse mais qui, par leur présence, ont contribué à ce que cette thèse soit ce qu'elle est. Je remercie donc Alban, mon ami le plus « ancien » ainsi qu'Alan pour les bons moments passés lors de nos soirées repas. Merci également à Gwenaël et Charlaine pour les bons moments passés au Bagad ainsi qu'en dehors. Et encore merci pour tous vos conseils de voyage en Nouvelle-Zélande ! Merci aussi à Gwendal pour nos soirées aux Fauvettes (ça fait un moment !), les bons moments passés au Bagad et nos discussions. Enfin je tiens à remercier particulièrement trois amies de l'Agro et de la promo

164, Adeline, Caroline et Floriane. Même si l'on ne peut pas tout le temps se retrouver, merci pour tous ces bons moments passés ensemble et qui ont rendu les 3 années d'école particulièrement agréables. Enfin, parce que je ne pourrai pas tous les citer (et il ne faut pas m'en vouloir), merci à mes amis de la promo 164 ainsi que de la prépa de Chatô (vive la BioSup 2 !).

Je tiens également à remercier toute ma famille pour m'avoir supporté et encouragé dans les moments du quotidien (parfois bien compliqués avant le stage et la thèse), dans la réussite et dans les échecs, et de m'avoir permis d'être là où je suis aujourd'hui. Merci à vous tous et merci pour tout. Merci également à toute ma belle-famille.

Enfin, je tiens à te remercier, pour avoir supporté depuis plus de 3 ans au quotidien mon caractère un peu têtu (de temps en temps), mes moments de « grommelage », mes moments passés à râler sur Orion, mais aussi mes moments de délires ou encore tous mes jeux de mots pourris. Si la thèse a pu être ce qu'elle est, c'est également en grande partie grâce à toi. Merci Lucile.

Je voulais être assez bref lors de ces remerciements, mais j'ai finalement encore écrit quelques pages. Désolé, c'était plus fort que moi !

Et je voulais finir avec une petite image. Gwenaël, je te l'avais promis, ta poule est donc dans ce manuscrit de thèse !

Table des matières

Liste des abréviations

ACP : Analyse en Composantes Principales

ADN : Acide Désoxyribonucléique

BD : Basse Densité

BLUP : Best Linear Unbiased Prediction

CI : Cage Individuelle

CM : Cage Multiple

CRoPS : Complexity Reduction of Polymorphic Sequences

DAG : Graphe orienté acyclique

ddRAD-Seq : Double-digest Restriction site-Associated DNA Sequencing

DGV : Direct Genomic Value

DL : Déséquilibre de Liaison

DYD : Daughter Yield Deviation

EQ : Équidistant

FF : Force de Fracture

GBLUP : Genomic Best Linear Unbiased Prediction

GBS : Genotyping-By-Sequencing

GEBV : Genomic Estimated Breeding Value

GGRS : Genotyping by Genome Reducing and Sequencing

HA : Hauteur d'Albumen

HD : Haute Densité

IBD : Identité Par Descendance

IBS : Identité Par État

INRA : Institut National de la Recherche Agronomique

ITAVI : Institut Technique de l'Aviculture

L : Leghorn

Lab : Couleur de la coquille

MAF : Fréquence Allélique Mineure

MD : Moyenne Densité

MSG : Multiplexed Shotgun Sequencing

NGS : Next-Generation Sequencing

PCR : Polymerase Chain Reaction

PO : Poids d'Œuf

QTL : Quantitative Trait Loci

RAD-Seq : Restriction site-Associated DNA Sequencing

RI : Rhode Island

RRL : Reduced Representation Libraries

SNP : Single Nucleotide Polymorphism

ssGBLUP : Single-step Genomic Best Linear Unbiased Prediction

SYSAAF : Syndicat des Sélectionneurs Avicoles et Aquacoles Français

TBV : True Breeding Value

YD : Yield Deviation

Liste des figures

Liste des tableaux

Introduction Générale

Face à la forte croissance démographique mondiale, l'œuf est une ressource importante pour assurer en partie les besoins en protéine de l'humanité. En effet, l'œuf est consommé partout dans le monde et est une source de protéine reconnue et peu couteuse comparativement à d'autres sources de protéines animales. La Chine est, de loin, le premier pays producteur d'œuf avec 24 millions de tonnes produites en 2016. Suivent ensuite l'Union Européenne et les États-Unis avec une production respective de 7.1 et 5.4 millions de tonnes d'œufs. En Europe, l'évolution de la production est quasi-stable (+0.5%) sur la période 2007-2017. La France est la premier pays producteur d'œuf Européen avec une production de 925 000 tonnes d'œufs suivi par l'Espagne (838 000 tonnes) et l'Allemagne (820 000 tonnes) (ITAVI, 2018). La filière ponte est donc une filière importante caractérisée par une forte production et une haute qualité de production. Ceci a été rendu possible par la mise en place, entre autres, d'une sélection génétique des individus. Cette sélection génétique, en choisissant les meilleurs reproducteurs sur différents critères à chaque génération, a permis de créer du progrès génétique au cours du temps.

La formalisation des bases de la sélection génétique a été rendue possible par la redécouverte au XXème siècle des lois de Mendel et le développement du modèle polygénique. Ce modèle caractérise le fait qu'un caractère quantitatif (par exemple le poids d'un animal ou d'un œuf) est gouverné par un très grand nombre de gènes, leurs effets s'additionnant et chaque gène ayant un effet infinitésimal. Ce caractère dépend aussi de l'environnement dans lequel il s'exprime. Ce modèle a évolué au fil du temps avec notamment le développement du BLUP (Best Linear Unbiased Prediction) permettant de calculer les valeurs génétiques de tous les individus ayant ou non des performances et des descendants, en les corrigeant simultanément des effets fixes du milieu, tout en tenant compte des individus apparentés.

En 2001, Meuwissen et al. ont proposé une méthode appelée « sélection génomique » utilisant plusieurs milliers de marqueurs moléculaires pour prédire la valeur génomique des individus. Les marqueurs moléculaires sont définis comme des fragments d'ADN, polymorphes, qui peuvent servir de balises pour suivre la transmission d'un segment chromosomique d'une génération à une autre. Avec une population de référence génotypée et phénotypée, il est alors possible de prédire la valeur génomique des candidats à la sélection dont on ne connait uniquement que le génotype.

En 2004, le génome de la poule a été le premier génome d'une espèce d'élevage rendu public (International Chicken Genome Sequencing Consortium, 2004 ; International Chicken Polymorphism Map Consortium, 2004b) permettant la découverte de plusieurs millions de SNP (Single Nucleotide Polymorphism). Ces SNP sont des marqueurs moléculaires et correspondent à des changements d'une seule base nucléotidique (A, T, G ou C), à un locus donné, très fréquents et apparaissant de façon régulière le long de l'ADN. Le développement des biotechnologies permettant le génotypage de ces marqueurs a abouti en 2013 à la création, en filière poule, d'une puce commerciale haute densité de 600 000 marqueurs (Kranis et al., 2013). Cette puce a permis d'envisager la sélection génomique dans cette filière. Le projet ANR UtOpIGe (2011-2015) a alors été lancé pour étudier les particularités des schémas de sélection pyramidaux avicoles et porcins quant à la mise en place de la sélection génomique. En parallèle, le développement des nouvelles techniques de séquençage (Next Generation Sequencing) permet dès à présent d'envisager des solutions autres que les puces à SNP pour réaliser cette sélection génomique.

Un point clé de ces différentes méthodes est l'obtention du génotype des individus pour pouvoir estimer de façon précise la valeur génétique des individus. Même si les coûts de génotypage ou de séquençage ont diminué au cours des années, le génotypage en routine pour la sélection génomique coûte encore cher, surtout pour une espèce comme la poule où la valeur marchande du reproducteur est très faible.

Actuellement, les grands enjeux de la sélection génomique consistent en une optimisation des coûts à l'échelle du schéma et en une optimisation de la précision des évaluations génomiques. Un des leviers d'optimisation des coûts de la sélection génomique consiste à développer des outils de génotypage à bas coût des candidats à la sélection tels que les puces à SNP basse densité (BD) mais aussi le séquençage basse profondeur. Il s'agit ensuite de déduire les génotypages des puces HD grâce à la méthode de l'imputation. Cette méthode consiste à prédire les génotypes HD des candidats à la sélection à partir de leur génotypes BD et des génotypes HD de la population parentale. Cette méthode s'appuie sur les règles de la transmission mendélienne et sur le déséquilibre de liaison (DL).

Un des leviers d'optimisation de l'efficacité de la sélection génomique concerne les individus reproducteurs dont l'obtention précise des génotypes (actuellement avec des puces HD) est importante pour réaliser des évaluations génomiques précises des candidats à la sélection. Le séquençage haute profondeur peut permettre d'identifier 5 à 10 fois plus de polymorphismes qu'avec une puce à SNP HD et pourrait donc, en théorie, permettre d'améliorer la précision des évaluations génomiques.

Les entreprises avicoles comme Novogen se tournent donc aujourd'hui vers la révolution que constitue la sélection génomique. Dans un monde très compétitif et concurrentiel, les entreprises de sélection doivent mettre en place les meilleures stratégies de génotypages pour leurs schémas de sélection dans le but de produire les meilleurs animaux et produits.

L'objectif global de cette thèse est d'identifier les meilleures stratégies de génotypages des candidats à la sélection, de façon à optimiser la précision des évaluations génomiques, tout en minimisant les coûts du schéma de sélection. Après un premier chapitre consistant en une revue bibliographique des différents points clés importants pour la compréhension et la réalisation de cette thèse, le deuxième chapitre correspond à une étude de l'impact de différents facteurs concernant le développement des puces à SNP BD ou la constitution de la population de référence sur l'efficacité de l'imputation. Ces travaux ont donné lieu à une publication « Design of low density SNP chips for genotype imputation in layer chicken » dans BMC Genetics. Dans le troisième chapitre, l'intérêt de l'utilisation de l'imputation pour les évaluations génomiques a été étudié. Une deuxième publication « Interest of using imputation for genomic evaluation in layer chicken » a été soumise dans Poultry Science. Dans un quatrième chapitre, une puce à SNP BD multi-lignée poule a été développée pour les différentes lignées de pondeuses Novogen et son impact sur la précision des évaluations génomiques a été étudié. L'objectif était d'obtenir une puce permettant de bons résultats d'imputation des candidats à la sélection ainsi qu'une bonne précision des évaluations génomiques. Dans un cinquième chapitre, les puces BD ont été remplacées par différentes méthodes simulées de génotypage par séquençage (RAD-Seq). L'objectif était d'identifier si les méthodes RAD-seq simulées pouvaient constituer une bonne alternative aux puces BD pour les candidats à la sélection. Une troisième publication « Interest of using Restriction site-associated DNA Sequencing technologies as an alternative to low density SNP chips for genomic selection in layer chicken: in silico results » a été soumise dans Genetic Selection Evolution. Enfin, les questions et les perspectives soulevées tout au long de la thèse et des études sont discutées dans le sixième et dernier chapitre.

Chapitre I. Revue Bibliographique

I. De la sélection génétique à la sélection génomique

A. Début de la sélection génétique à la découverte des marqueurs moléculaires

1. Des débuts de la sélection génétique aux premiers modèles d'évaluation génétiques

Depuis des dizaines de milliers d'années, l'homme a domestiqué les animaux pour les élever. Bien loin d'identifier que la valeur génétique d'un animal correspondait, en moyenne, à la moitié de la valeur génétique du père et de la mère, il a néanmoins constaté que certaines performances des animaux semblaient dépendre des performances des parents. Il a dès lors cherché à sélectionner ces animaux en se concentrant sur les performances propres de chaque animal. À partir du XVIII-XIXème siècle, ces pratiques se sont organisées, principalement en Europe, et plus particulièrement au Royaume-Uni, pour permettre la sélection d'animaux avec des caractéristiques homogènes. Ceci a conduit à la standardisation des races et a eu beaucoup d'importance dans l'homogénéisation de l'aspect extérieur des populations d'élevage. Chaque race peut ainsi être identifiée par une apparence et des caractéristiques propres. La sélection génétique est devenue plus appliquée au XXème siècle avec la « redécouverte » des lois de Mendel (Laloë, 2011) et avec le développement du modèle polygénique de Fisher en 1918. Ce modèle formalise le fait qu'un caractère quantitatif (par exemple le poids d'un animal ou d'un œuf) est gouverné par un très grand nombre de gènes, chaque gène ayant un effet infinitésimal, et que ce caractère dépend aussi de l'environnement dans lequel il s'exprime.

Enfin, en 1943, Hazel a développé une méthode des index de sélection permettant de combiner les performances propres de chaque individu avec celles de ses apparentés (père, mère, frères, sœurs, ou descendants). Mais il y avait toutefois deux points de complexité. Le premier point était la nécessité de distinguer les effets de milieux qui sont des effets fixes (effets du bâtiment, de la saison, etc.) des effets génétiques qui sont des effets aléatoires. Ces deux types d'effets contribuent tous deux au phénotype et il était impossible de comparer les valeurs génétiques d'animaux élevés dans des conditions différentes. Le deuxième point était la nécessité de calculer un très grand nombre de covariances entre apparentés. La combinaison des différents apparentements rendait alors les calculs complexes à formaliser (Laloë, 2011).

2. Développement du BLUP et du modèle mixte

Le premier point de complexité est résolu avec le développement du BLUP (Best Linear Unbiased Prediction) grâce à Henderson en 1975. La modélisation des performances peut alors s'écrire selon l'équation suivante :

$$y = X\beta + Zu + e$$

Avec :

- y le vecteur des performances
- β le vecteur des effets fixes du modèle
- u le vecteur des valeurs génétiques, avec $Var(u) = A\sigma_u^2$,
- e le vecteur des résiduelles du modèle, avec $Var(e) = I\sigma_e^2$
- X et Z les matrices d'incidence reliant les performances respectivement aux effets fixes et aux effets aléatoires.
- A la matrice des coefficients de parenté
- I la matrice identité
- σ_u^2 la variance des valeurs génétiques
- σ_e^2 la variance de la résiduelle

Les estimations des effets fixes et des valeurs génétiques sont alors les solutions du système d'équation du modèle mixte suivant :

$$\begin{pmatrix} X'X & X'Z \\ Z'X & \frac{\sigma_e^2}{\sigma_u^2}A^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ u \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

Avec A^{-1} l'inverse de la matrice des coefficients de parenté.

Cette méthode permet donc de calculer les valeurs génétiques de tous les individus ayant ou non des phénotypes et des descendants, en les corrigeant simultanément des effets fixes du milieu, tout en tenant compte des individus apparentés.

Enfin, le deuxième point de complexité concernant le calcul des covariances entre individus est résolu par le même Henderson en 1976. Il simplifie le calcul de l'inverse de la matrice de parenté en s'appuyant sur les corrélations partielles entres individus.

Le BLUP est ainsi devenu la référence internationale pour réaliser des évaluations génétiques.

3. La découverte des marqueurs moléculaires et l'évolution des modèles d'évaluations

Dans les années 80, les généticiens ont développé de nombreuses techniques permettant d'étudier l'ADN et ont découvert plusieurs types de marqueurs moléculaires polymorphes dans le génome (Boichard et al., 1998). Les marqueurs moléculaires sont définis comme des

fragments d'ADN, polymorphes, qui peuvent servir de balises pour suivre la transmission d'un segment chromosomique d'une génération à une autre. Parmi les différents types de marqueurs détectés, les microsatellites puis les SNP (Single Nucleotide Polymorphism) se sont révélés les plus intéressants pour la sélection animale.

Les microsatellites ont été identifiés en 1989 et correspondent à la répétition de séquence d'un motif constituée d'un à quatre nucléotides (Figure 1). Ce type de marqueur présente l'avantage d'être très polymorphe et bien réparti sur l'ensemble du génome, mais est finalement assez peu fréquent. À titre d'exemple, plus de 800 microsatellites ont été identifiés et positionnés sur le génome de la poule (Groenen et al., 2000), et plus de 3900 sur le génome bovin (Ihara et al., 2004). En parallèle en 1990, Lande et Thomson ont proposé d'utiliser ces marqueurs et le déséquilibre de liaison pour suivre la transmission des portions de chromosomes ayant un effet sur un caractère quantitatif, appelés QTL. Le déséquilibre de liaison caractérise l'association non-aléatoire sur un même chromosome entre allèles de deux loci. Cette notion sera abordée plus en détail dans la section II.B. En cas de DL avec un allèle du QTL, l'allèle du marqueur moléculaire peut permettre de suivre la transmission du QTL. Ils calculaient ensuite un « score moléculaire » et l'incluaient dans le calcul des valeurs génétiques des individus. Puis en 1998, Haley et Visscher ont imaginé l'utilisation de plusieurs milliers de marqueurs moléculaires répartis sur l'ensemble du génome pour calculer la valeur génétique des individus. L'idée était qu'en utilisant l'ensemble des marqueurs du génome, il serait possible, grâce au DL, de suivre d'une génération à une autre la transmission de l'ensemble des QTL ayant un effet sur différents caractères d'intérêt. Cette idée a été rendue possible en 2001 avec le premier séquençage du génome humain qui a permis de découvrir les SNP. Ces SNP correspondent à des changements d'une seule base (A, T, G, C) à un locus donné, très fréquents et apparaissant de façon assez uniforme le long de l'ADN (Figure 1). Ce sont donc des marqueurs bialléliques, moins informatifs que les microsatellites. Toutefois, ces SNP constituent la majorité des variations de séquence des génomes. La moindre informativité de ce type de marqueur est ainsi compensée par son abondance.

Avec ces différents marqueurs moléculaires, il est possible d'estimer les effets des portions de chromosomes sur différents caractères d'intérêt. Le concept des évaluations génomiques et leur mise en place ont alors commencé à être étudiés plus en profondeur.

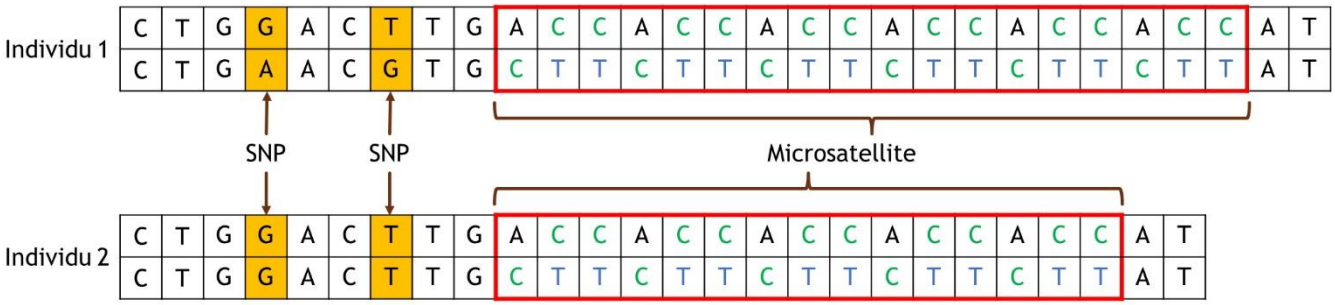


Figure 1. Exemple de SNP avec un polymorphisme (A/G) et (T/G) et de microsatellite avec répétitions du motif (ACC/CTT).

4. Le séquençage du génome de la poule et ses particularités

En 2004, la poule a été la première espèce d'élevage dont le génome a été rendu public (International Chicken Genome Sequencing Consortium, 2004 ; International Chicken Polymorphism Map Consortium, 2004). En 2016, une cinquième version du génome a été publiée (Warren et al., 2017). Cette version a été utilisée tout au long de la thèse. Le caryotype de la poule (Figure 2) présente la particularité d'être sous-divisé en 38 paires de chromosomes autosomaux de tailles très variables. Il est ainsi possible de distinguer des macro-chromosomes (1 à 5) mesurant de 198Mb à 91Mb, des chromosomes intermédiaires (6 à 10) mesurant de 36Mb à 21Mb et des micro-chromosomes (11 à 38) mesurant de 20Mb à 3Mb. Il est à noter que le chromosome 1 représente 20% du génome de la poule. Par ailleurs, il y a également une paire de chromosomes sexuels Z et W dont la taille se rapproche respectivement des macro-chromosomes et des chromosomes intermédiaires. Les mâles sont homogamétiques ZZ et les femelles hétérogamétiques ZW (Vignal, 2000). Les micro-chromosomes sont caractéristiques des oiseaux et d'une partie des reptiles et des poissons. Ils présentent une densité de gènes plus élevée que les macro-chromosomes (International Chicken Genome Sequencing Consortium, 2004.) Enfin, Qanbari et al. (2010) ont mis en évidence une persistance du DL différente entre type de chromosomes. Les macro-chromosomes présentent une forte persistance du DL et donc un DL à longue distance, alors que les micro-chromosomes présentent une persistance plus faible du DL, et donc un DL à plus faible distance. Des exemples sont présentés dans la section II.B.3 et dans le chapitre IV.

Le génome de la poule mesure environ 1Gb et plus de 24 millions de SNP sont référencés dans la base de données Ensembl (2019) de SNP. Il est donc possible d'estimer les effets des portions de chromosomes sur différents caractères d'intérêt à partir de ces marqueurs.

Enfin, à ce jour, seuls les chromosome 1 à 28, 33, un groupe de liaison LGE64 et les deux chromosomes sexuels Z et W sont référencés, les micro-chromosomes n'étant pas encore bien référencés.

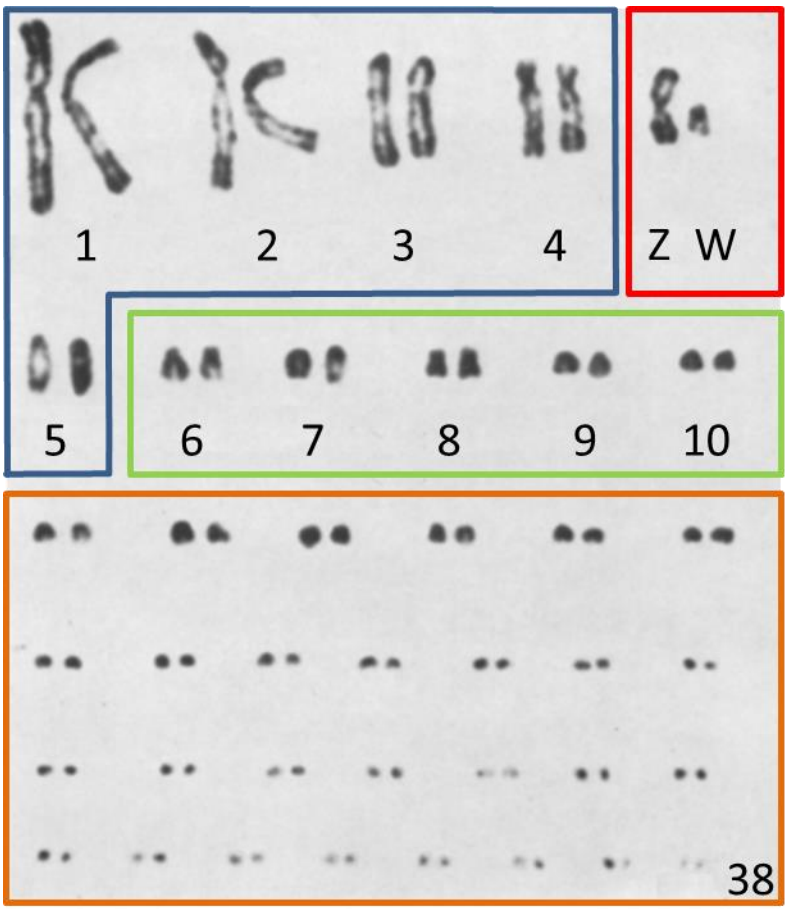


Figure 2. Caryotype de la poule. Les macro-chromosomes sont ici représentés par les chromosomes 1 à 5 en bleu, les chromosomes intermédiaires par les chromosomes 6 à 10 en vert, et les micro-chromosomes par les chromosomes 11 à 38 en orange. Les chromosomes sexuels Z et W sont en rouge. Adapté de Owen (1965).

5. L'émergence des outils de génotypages haut-débit

La découverte des SNP ainsi que le séquençage du génome des espèces d'élevages ont débouché sur la possibilité de création de puces à SNP. Bien que la poule ait été la première espèce d'élevage séquencée en 2004, c'est chez les bovins, séquencés en 2006, que les premières puces à SNP ont été développées dans le cadre d'un consortium international de laboratoires de recherches. Une première puce de 50K SNP (Figure 3a) a ainsi été développée par l'entreprise Illumina (Matukumalli et al., 2009 ; Illumina Inc, 2011b). Les recherches se sont poursuivies pour aboutir en 2012 à la création d'une puce commerciale haute-densité de 777K SNP (Illumina Inc, 2012a). À cette densité et avec un génome bovin mesurant 2.7Gb, la distance médiane entre deux marqueurs consécutifs est inférieure à 3kb. Les QTL sont forcément proches des marqueurs SNP, et peuvent donc être suivis de génération en génération.

Chez la poule, les puces à SNP ont été plus longues à se mettre en place. La première puce développée par Illumina contenait 3K SNP (Muir et al., 2008), ce qui n'était pas suffisant compte tenu de l'informativité faible d'un SNP par rapport à un microsatellite. En 2008, une première puce de 60K a été conçue grâce à des fonds du Ministère de l'Agriculture des États-Unis, de l'université de Wageningen et des entreprises Cobb et Hendrix, et a été utilisée en interne par ces deux entreprises (Groenen et al., 2011). Puis en 2009, une deuxième puce de 42K a été créée par le groupe Wesjohann pour un usage également en interne. Ce n'est finalement qu'en 2013 qu'une puce commerciale haute densité (HD) de 600K SNP a été développée (Kranis et al., 2013) (Figure 3b) et a permis à tous les groupes et chercheurs de pouvoir se lancer dans la sélection génomique. En effet, la densité de marqueurs est devenue suffisamment importante pour estimer les effets des portions de chromosomes sur différents caractères d'intérêt.

Enfin, de nouvelles puces à moyenne (MD) voire basse densité (BD) ont également été mises au point. Ces puces présentent donc l'avantage de coûter moins cher que les puces HD, mais permettent d'obtenir des génotypages tout aussi intéressants que les génotypages HD. Ceci repose sur la méthode de l'imputation, permettant de déduire les génotypages manquants sur les puces HD mais présents sur les puces BD ou MD. Cette méthode sera détaillée dans la section II. Ainsi, Boichard et al. (2012a) ont élaboré en bovin une puce commerciale BD de 7K SNP et Liu et al. (2019) ont conçu pour la poule une puce commerciale MD de 55K. Enfin, les entreprises de sélection avicoles ont continué de développer elles-mêmes des panels de SNP à plus basse densité que la puce HD. À titre indicatif, l'utilisation de la puce HD en poule coûte environ 150€ par individu quand l'utilisation d'une puce d'environ 10K SNP coûte 35€.

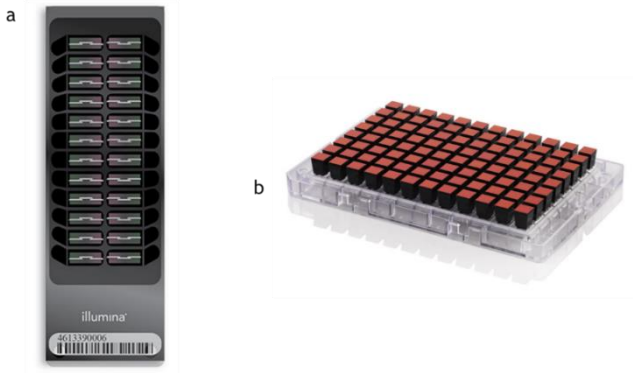


Figure 3. Illustration de la puce Bovine SNP50 Genotyping BeadChip (a) et de la puce Affymetrix Axiom Chicken Array (b).

B. Mise en place de la sélection génomique

1. Principes de la sélection génomique

En 2001, Meuwissen et al. ont proposé une méthode appelée « Sélection Génomique » utilisant de nombreux marqueurs moléculaires tels que les SNP pour prédire la valeur génomique des individus, sans connaissance préalable de leur phénotype (Figure 4). La mise en place de la sélection génomique nécessite une population de référence constituée d'un grand nombre d'individus pour lesquels on dispose de génotypages et de performances. Cette population est sous-divisée en deux populations : une population d'apprentissage correspondant environ au ¾ de la population de référence, et une population de validation correspondant environ au ¼ de la population de référence.

La sélection génomique se déroule en deux étapes. La première est d'estimer, sur la population d'apprentissage, les effets de chaque région du génome. Ces effets sont estimés grâce à un système d'équations associant les génotypes des différents marqueurs aux phénotypes des individus. Le système d'équation est ensuite validé sur la population de validation en comparant les valeurs génomiques prédites et performances mesurées. Une fois le système d'équation validé, il est appliqué à l'ensemble des candidats à la sélection pour lesquels on ne dispose que des génotypages. Il est alors possible de déterminer dès la naissance d'un individu sa valeur génétique, avant même de disposer de ses performances, ou de celles de ses descendants.

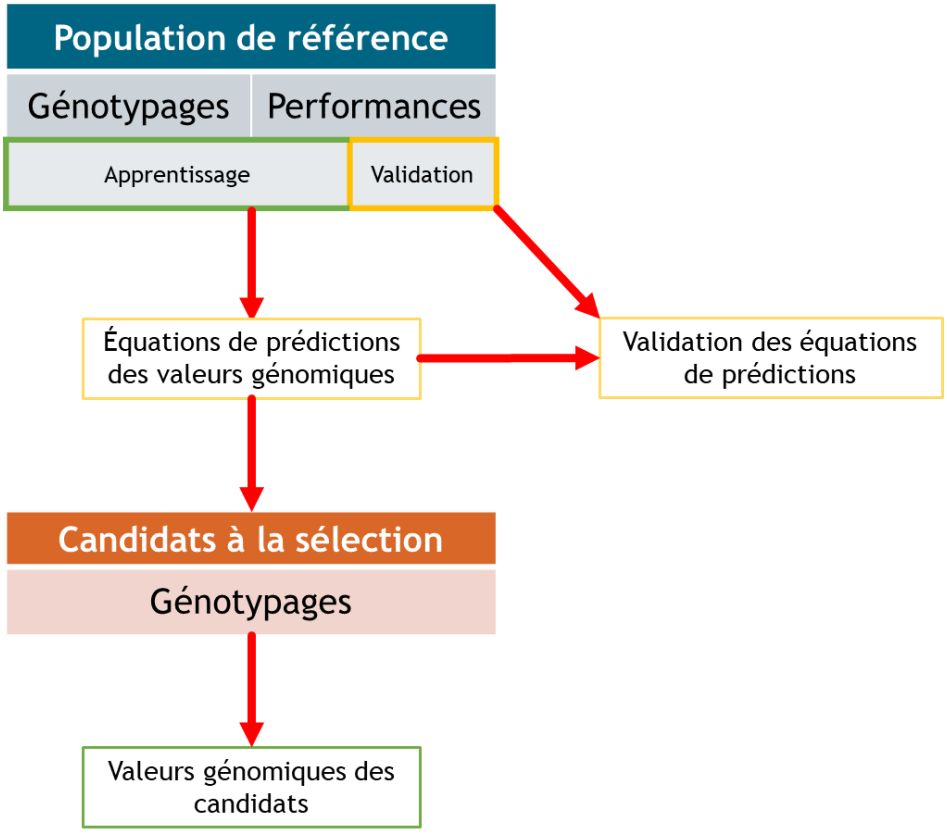


Figure 4. Principe de la sélection génomique

2. Les différentes propriétés de la sélection génomique

La sélection génomique peut donc être appliquée à un ensemble de candidats à la sélection pour différents caractères, à condition que ceux-ci aient été mesurés sur la population de référence. En effet, les équations de prédictions des valeurs génomiques associent les génotypes des différents marqueurs aux phénotypes des individus de la population de référence. Sans phénotype sur cette population, il devient alors impossible d'estimer les valeurs génomiques des candidats pour les différents caractères d'intérêts.

L'efficacité de la sélection génomique est également fortement dépendante de la taille de la population de référence et du nombre de marqueurs. Goddard et Hayes (2009) ont montré que la précision des évaluations génomiques était d'autant plus grande que le nombre d'individus dans la population de référence était élevé. Cette précision est également fonction de l'héritabilité du caractère, qui est définie pour une population comme la part de variance phénotypique d'origine génétique (Figure 5). Daetwyler et al. (2010) ont ensuite formalisé le calcul de la précision des évaluations :

$$r = \sqrt{\frac{Nh^2}{Nh^2 + 4N_eL}}$$

Avec :

- N la taille de la population de référence
- h^2 l'héritabilité du caractère étudié
- N_e la taille efficace de la population
- L la longueur du génome (en Morgan)

La taille efficace de la population correspond à la taille d'une population idéale (non sélectionnée avec une reproduction aléatoire) de même diversité génétique que la population étudiée. Enfin Solberg et al. (2008) ont montré qu'une augmentation de la densité de marqueurs pouvait permettre une augmentation de la précision des évaluations génomique.

Pour que la sélection génomique soit efficace, il est également important de renouveler la population de référence en fonction de la population candidate. Legarra et al. (2008) ont montré que la sélection génomique est d'autant plus efficace que les individus dans les populations de référence et candidate sont proches. Réaliser une sélection génomique sur des candidats à la sélection trop éloignés en termes de générations de la population de référence est moins efficace à cause de la chute du DL entre marqueurs au fil des générations. Ce point sera détaillé dans la section III.C. Ceci impacte les équations de prédiction des valeurs génomiques qui ne sont alors plus adaptées et entraîne une chute de la précision des évaluations. Ces observations sont également retrouvées par Wolc et al. (2011) en poule pondeuse qui ont montré une diminution

de la précision des évaluations génomiques au fur et à mesure que les générations candidates s'éloignaient des générations de référence (Figure 6). De la même façon, lorsque des observations supplémentaires sont apportées pour la population de référence, les équations de prédictions sont modifiées. Finalement, dès que la population de référence est modifiée, soit par changement de génération, cumul d'individus ou disponibilité de nouvelles observations phénotypiques, il est nécessaire de ré-estimer les effets des différentes régions chromosomiques.

Les évaluations sont par ailleurs efficaces quel que soit le sexe des individus et dès leur plus jeune âge. En effet, il n'y a plus besoin d'obtenir le phénotype des candidats à la sélection pour les sélectionner, seule l'information des marqueurs étant nécessaire pour réaliser une sélection génomique. Ceci est toutefois conditionné par la disponibilité des phénotypes pour la population de référence.

Il devient également possible de sélectionner de nouveaux caractères qui étaient jusque-là bien trop complexes et/ou couteux à mesurer en routine sur un grand nombre d'individus. Par exemple, il est possible de réaliser une sélection sur la résistance des individus à diverses maladies, sur des caractères de qualité de produits qui nécessitaient l'abattage des animaux, ou encore sur des caractères assez lourds à mesurer en plaçant l'individu dans un environnement particulier. Là encore, ces possibilités dépendent de la disponibilité des phénotypes pour la population de référence.

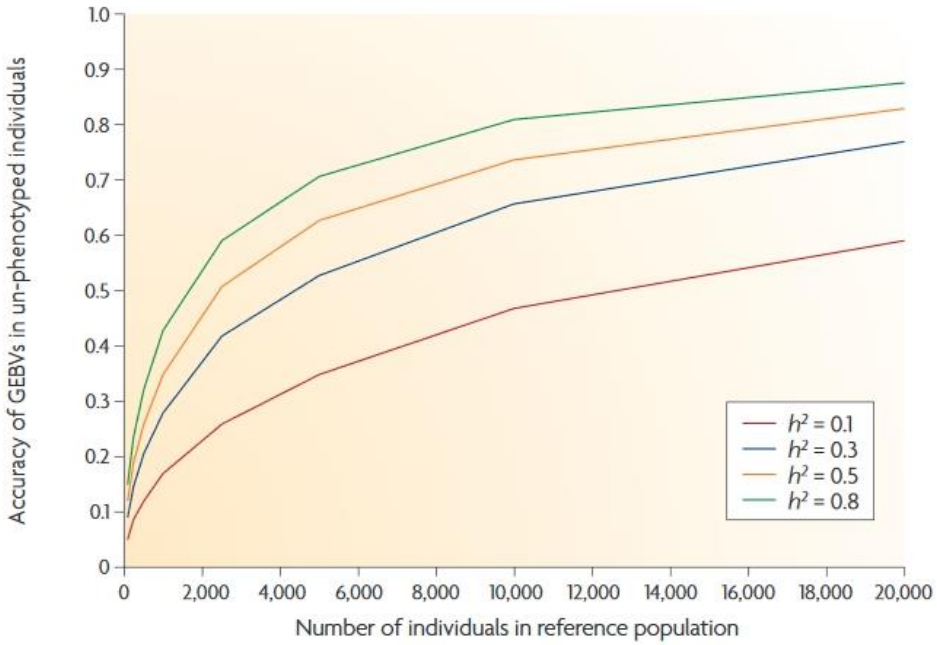


Figure 5. Précision des évaluations génomiques en fonction du nombre d'individus dans la population de référence et de l'héritabilité du caractère étudié. D'après Goddard et Hayes (2009).

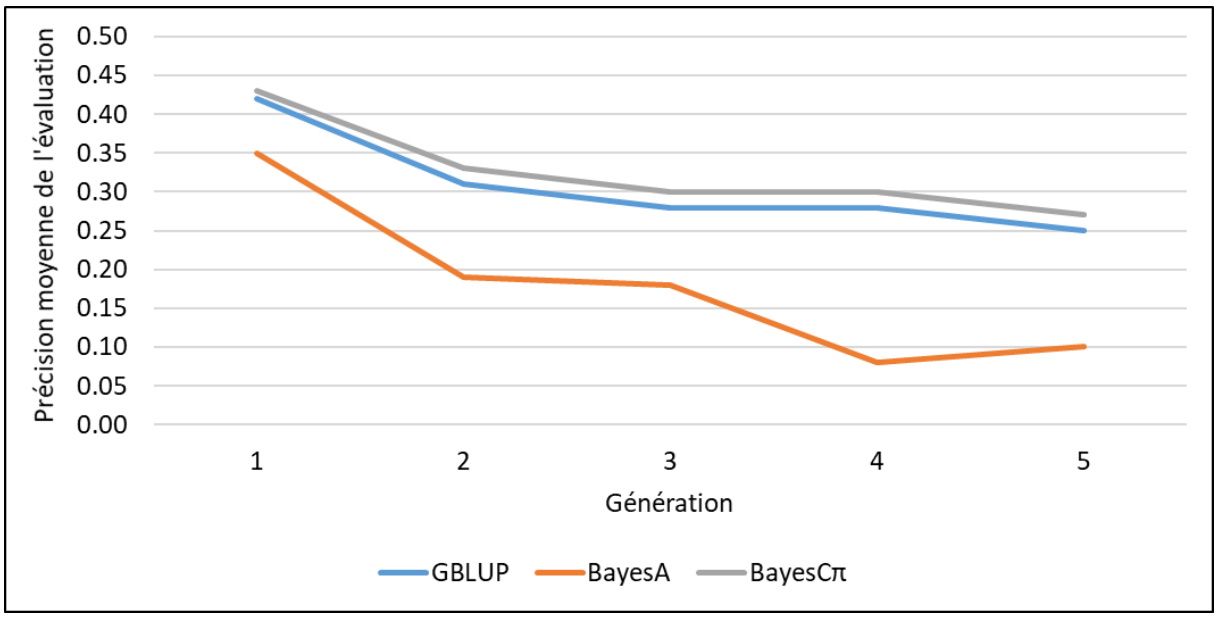


Figure 6. Précision moyenne de l'évaluation des générations candidates à partir d'une population de référence constituée des deux générations précédant la génération 1. La précision est évaluée pour les méthodes GBLUP, Bayes A et Bayes Cπ. Elle correspond à la moyenne des précisions de 16 caractères, la précision de chaque caractère étant calculée comme la corrélation entre les valeurs génomiques estimées et les phénotypes divisée par la racine carrée de l'héritabilité du caractère. D'après Wolc et al. (2011).

3. Les différents modèles d'évaluations génomiques

Avec la méthode développée par Meuwissen et al. (2001), il se pose très vite la question de savoir comment estimer avec suffisamment de précision l'effet de p SNP à partir de n observations, sachant que le nombre de marqueurs est largement supérieur au nombre d'observations. Plusieurs types de méthodes permettent de répondre à ce problème de grande dimension « $p \gg n$ » (Le Roy et al., 2014). Nous ne développerons ici brièvement que les trois principaux types de méthodes.

a) Méthode GBLUP

Le premier type de méthode repose sur le GBLUP (Genomic Best Linear Unbiased Prediction), incluant tous les SNP dans l'évaluation en supposant que les effets des marqueurs suivent une loi normale de même variance. Cette méthode suppose donc que tous les marqueurs contribuent à l'expression du caractère, chaque marqueur ayant un effet faible sur le caractère. Le GBLUP est finalement une extension du BLUP polygénique, avec remplacement de la matrice de parenté pedigree A par une matrice de parenté génomique G. Les éléments de cette matrice mesurent la proportion moyenne d'allèles partagée par deux individus (modèle animal). Van

Raden (2008) a montré qu'il était également possible de sommer les effets des SNP considérés comme aléatoires (modèle SNP). Ces deux façons de faire permettent d'aboutir aux mêmes estimations des valeurs génomiques. En fonction du nombre de marqueurs et d'individus disponibles, on préférera soit estimer les parentés génomiques lorsque le nombre de marqueurs est supérieur au nombre d'individu, soit estimer les effets des marqueurs lorsque le nombre d'individu est supérieur aux nombre de marqueurs (Croué, 2017). Ces méthodes sont efficaces et robustes pour des caractères contrôlés par de nombreux gènes ayant chacun un effet faible sur les caractères (Robert-Granié et al., 2011).

b) Méthodes Bayésiennes

D'après Hayes et Goddard (2001), la variabilité génétique des caractères dépend d'un grand nombre de loci à effets faibles et d'un petit nombre de loci à effets forts sur le caractère. C'est le cas par exemple du gène *DGAT1* qui a un effet fort sur la teneur en matière grasse du lait chez les bovins (Coppieters et al., 1998), du gène de la caséine α_{s1} qui a un effet fort sur le teneur en protéine du lait de chèvre (Grosclaude et al., 1987) ou encore du gène du rendement Napole (RN) chez les porcins qui a un effet fort sur le pH ultime de la viande (Le Roy et al., 1990). Les méthodes bayésiennes permettent de considérer des effets différents sur les caractères pour les différents marqueurs. Il est ainsi possible de considérer une proportion de SNP à effet nul et une autre proportion de SNP à effet variable plus ou moins fort. La distribution de ces effets est considérée comme un *a priori* et plusieurs méthodes se distinguent en fonction des distributions *a priori* considérées. On peut ainsi citer le Bayes A, Bayes B, Bayes C, Bayes Cπ, etc. En comparaison du GBLUP, ces méthodes sont souvent intéressantes lorsqu'il y a des QTL à effets forts sur les caractères d'intérêt.

c) Méthode Single-Step GBLUP

Il existe deux principaux inconvénients aux méthodes développées précédemment. Le premier inconvénient est que l'évaluation génomique est réalisée à partir de données phénotypiques corrigées, ou pseudo-phénotypes, obtenues à partir d'évaluations génétiques des individus apparentés aux candidats à la sélection. Par exemple pour un coq, il semble de prime abord compliquer de vouloir mesurer sa ponte… On peut en revanche lui attribuer un pseudo-phénotype qui correspond à la moyenne des performances de ses filles, corrigée par l'ensemble des effets du modèle. Mais ces pseudo-phénotypes peuvent parfois manquer de précision à cause d'un nombre trop faible d'observations ou à cause d'observations non précises (Legarra et al., 2014). Le deuxième inconvénient est que l'évaluation génomique n'est réalisée que pour

les individus génotypés. Il peut apparaître intéressant d'utiliser l'information des apparentés non génotypés qui ne sont pas des descendants directs des individus génotypés. Utiliser uniquement les individus génotypés et les apparentés directs peut entrainer une perte d'information, un risque de perte de précision et un risque de biais, les animaux génotypés n'étant généralement pas choisis au hasard (Patry et Ducrocq, 2011 ; Croué, 2017).

Pour résoudre ces problèmes, il est alors devenu intéressant d'analyser simultanément les données phénotypiques et les données génomiques. C'est ainsi que le single-step GBLUP (ou GBLUP en une étape) a été développé. Cette méthode conserve les mêmes principes que le BLUP polygénique mais remplace la matrice de parenté pedigree A par une matrice de parenté mixte H combinant les informations de parenté pedigree et de parenté génomique (Aguilar et al., 2010). D'après la formule développée dans la section I.A.2, le ssGBLUP nécessite le calcul de l'inverse de la matrice H. Cette matrice inverse a justement une structure très simple :

$$H^{-1} = A^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{pmatrix}$$

Avec :

- A^{-1} l'inverse de la matrice de parenté pedigree
- G^{-1} l'inverse de la matrice de parenté génomique
- A_{22}^{-1} l'inverse de la matrice de parenté pedigree des animaux génotypés

Par cette méthode il est alors possible de faire bénéficier à l'ensemble des animaux non génotypés de l'information génomique des individus apparentés. Cela permet également de faire bénéficier aux individus génotypés des informations de performances des individus non génotypés. Cette méthode permet donc d'obtenir de bons résultats, sans biais (Legarra et al., 2014). C'est d'ailleurs cette méthode qui a été utilisée tout au long de la thèse pour réaliser les différentes évaluations génomiques.

Il reste toutefois quelques inconvénients à cette méthode. En effet, le calcul de la matrice H^{-1} peut se révéler assez long car il nécessite le calcul préalable de la matrice G^{-1} qui est une matrice dense et de grande taille, ainsi que le calcul de la matrice A_{22}^{-1}. Enfin, l'hypothèse de départ considérant que les effets des marqueurs suivent une loi normale de même variance peut entraîner des résultats moins bons que certaines méthodes bayésiennes dans le cas de caractères sous contrôle de QTL à effets forts. Des modèles existent pour essayer de mieux prendre en compte des QTL à effets forts sur certains caractères d'intérêts (Wang et al., 2012 ; Zhang et al., 2016).

4. Impact de la sélection génomique sur la sélection des espèces d'élevage

a) Définition du progrès génétique

L'efficacité de la sélection génétique se mesure avec le progrès génétique théorique. Ce progrès s'exprime de la façon suivante :

$$\Delta G = \frac{i * \sqrt{CD} * \sigma_g}{T}$$

Avec :

- ΔG le progrès génétique
- i l'intensité de sélection
- $\sqrt{CD}$ la précision de l'évaluation génétique
- σ_g l'écart-type génétique du caractère étudié
- T l'intervalle de génération

La mise en place de la sélection génomique chez les bovins puis progressivement étendue aux autres espèces d'élevages a permis une augmentation du progrès génétique chez ces espèces. Cette augmentation peut passer par une augmentation de l'intensité de sélection, une amélioration de la précision des évaluations génomiques ainsi qu'une diminution de l'intervalle de génération. En revanche, la sélection génomique n'a aucun impact sur la variabilité des caractères (Boichard et al., 2016).

b) Augmentation de l'intensité de sélection

L'intensité de sélection peut être augmentée si l'on peut diminuer la proportion d'individus sélectionnés qui permettront de constituer une nouvelle génération d'individus. En supposant un nombre de candidats à la sélection fixe, en diminuant le nombre d'individu sélectionnés, on augmente l'intensité de sélection. Toutefois, avec la sélection génomique, il est possible et préférable d'élargir la base de sélection si le génotypage est plus facile et plus rapide à réaliser que le phénotypage. Avec les coûts décroissants de génotypage, en utilisant des puces à moyennes ou basse densité, puis en remontant à l'information d'une puce haute-densité grâce à la technique de l'imputation, il devient alors possible de génotyper un grand nombre de candidats à la sélection. En supposant la sélection de 50 individus parmi 5000 au lieu de 50 parmi 500, la proportion d'individus sélectionnés est divisée par 10, ce qui permet d'augmenter l'intensité de sélection. Meuwissen et al. (2016) ont ainsi rapporté pour les bovins laitiers que la combinaison des puces basse ou moyenne densité avec l'imputation pour remonter à des hautes densités a permis le génotypage de plus de 2 millions d'individus au niveau mondial. Une augmentation de l'intensité de sélection a ainsi pu être observée, ces génotypages

concernant des individus qui étaient des candidats à la sélection au moment de leur génotypage. En porcin, l'intensité de sélection pratiquée dans le cadre d'une sélection génétique classique est déjà élevée (Tribout et al., 2011). La mise en place d'une sélection génomique en porcin n'entraînerait donc pas de forte augmentation de l'intensité de sélection et donc de gain de progrès génétique en jouant sur ce facteur. Enfin, en poule pondeuse, Sitzenstock et al. (2013) ont réalisé des simulations sur des lignées commerciales de pondeuses dont les individus sont sélectionnés selon un index multi-caractères. Ils montrent qu'un progrès génétique est réalisable sur la plupart des caractères étudiés comme l'intensité de ponte en passant de 800 à 4800 candidats à la sélection.

c) Diminution des intervalles de génération

Il est également possible de diminuer l'intervalle de génération pour augmenter le progrès génétique. C'est d'ailleurs principalement sur ce facteur qu'est permise l'augmentation de progrès génétique pour plusieurs espèces d'élevages. Schaeffer (2006) montre ainsi que la sélection génomique permet d'utiliser un taureau dès sa maturité sexuelle. Le passage de l'intervalle de génération d'un taureau de 5 à 2 ans permet de doubler le progrès génétique. En poule pondeuse, Wolc et al. (2015) ont étudié deux groupes d'animaux : un groupe soumis à une sélection génétique classique et un autre groupe soumis à une sélection génomique. Au bout de 3 ans d'élevage, deux générations ont été obtenues pour le premier groupe et quatre générations pour le deuxième groupe. Les résultats des évaluations génomiques sont meilleurs pour 12 des 16 caractères analysés pour le groupe soumis à une sélection génomique. En divisant par deux l'intervalle de génération, un progrès génétique a donc été possible. En revanche, chez les porcins, l'intervalle de génération est déjà très court (Tribout et al., 2011). Les choix de sélection ou de réforme des candidats mâles et femelles sont pris au moment de la fin de leur contrôle de performance en ferme, avant 6 mois d'âge. Les individus retenus sont ensuite mis en reproduction à l'âge de 8 mois. Ceci limite le gain potentiel de progrès génétique par la voie de l'intervalle de génération.

d) Augmentation de la précision des évaluations

Enfin, il est également possible d'améliorer les précisions des évaluations grâce à la sélection génomique. Comme expliqué dans la section I.B.2, la précision des évaluations dépend de la taille de la population de référence, du nombre de marqueurs, de l'héritabilité du caractère étudié ainsi que de l'architecture génétique du caractère. En bovin laitier, Meuwissen et al. (2016) rapportent que des précisions élevées sont atteignables, dépassant même 0.8 pour des

caractères de production et 0.7 pour des caractères de fertilité ou encore pour le taux de cellules somatiques. À titre de comparaison, la précision des évaluations génétiques pour ces mêmes caractères est comprise entre 0.2 et 0.4. Les précisons d'évaluations génomiques élevées s'expliquent par la disponibilité de très grandes populations de référence et par la qualité des phénotypages des individus de la population de référence, un grand nombre étant récolés sur des individus testés sur descendance. En revanche, en bovin allaitant, Van Eenennaam et al. (2014) rapportent des précisions comprises entre 0.3 et 0.7 pour des caractères de poids à différents âges, de conformation ou encore de production de viande. Ces précisions plus faibles sont toutefois supérieures à celles obtenues avec une évaluation génétique (Lourenco et al., 2015). En effet, pour les caractères de poids à la naissance, poids de sevrage, gain de poids post-sevrage et facilité de vêlage, les précisions des évaluations génétiques sont respectivement de 0.29, 0.34, 0.23 et 0.12. Les précisions des évaluations génomiques pour ces mêmes caractères sont respectivement de 0.39, 0.38, 0.29 et 0.13. Ces précisions plus faibles s'expliquent par des populations de référence plus petites en fonction des races considérées, populations qui ne sont pas constituées d'individus testés sur descendance (Meuwissen et al., 2016). Enfin, les populations de référence et candidate peuvent être moins liées entre elles, ce qui diminue le niveau de DL entre QTL et marqueurs et donc la précision des effets de SNP estimés (Le Roy et al., 2014). En porcin, du fait d'une durée de carrière courte des reproducteurs, le nombre de mesures phénotypiques est limité. Cela a donc pour conséquences une précision des valeurs génétiques assez faibles comprise entre 0.1 et 0.4 en fonction des caractères étudiés (Tribout et al., 2011). Lillehammer et al. (2011) indiquent que pour des candidats mâles, une précision des valeurs génomiques de 0.5 peut être obtenue pour des caractères maternels. Enfin, en poules pondeuses, Picard Druet et al. (2019) montrent dans le cas d'une évaluation génomique sur ascendance des candidats mâles à la sélection, pour une période de production en cage individuelle, que la précision des évaluations génomiques de différents caractères de qualité d'œuf est inférieure à 0.5 mais au moins toujours supérieure à la précision obtenue avec des évaluations génétiques.

C. Utilisation des puces à SNP et contrôle de la qualité des génotypages

1. Fonctionnement des puces et détection du génotypage des SNP

En fonction de la technologie utilisée, le génotypage peut se faire de plusieurs façons différentes. Les explications suivantes résument le protocole mis en place pour les dernières technologies Illumina et Thermo Fisher (Figure 7).

La première étape consiste à extraire l'ADN des individus à analyser. Cet ADN peut être extrait à partir de n'importe quel tissu. Le sang est principalement utilisé comme source d'ADN mais il est également possible de réaliser des prélèvements d'autres tissus comme un bout de crête ou encore une plume en volaille. Cet ADN est extrait par lyse de la membrane et du noyau des cellules du tissu puis il est purifié et dénaturé. L'ADN est ensuite amplifié puis fragmenté par des enzymes de restrictions. Les fragments d'ADN sont alors précipités puis mis en suspension afin de les récupérer pour les disposer ensuite sur la puce à SNP.

Les puces à SNP correspondent à des lames de verres ou de silicium sur lesquelles sont déposés de façon ordonnée des oligonucléotides de moins de 100 pb dont la séquence est divisée en deux parties : une première partie qui va se fixer sur la puce à SNP et une deuxième partie qui va s'hybrider avec le fragment d'ADN extrait et amplifié précédemment et qui s'arrête une base avant le SNP d'intérêt. Ces fragments sont déposés dans des micropuits et constituent les sondes de la puce à SNP. Pour une sonde donnée, seul le brin complémentaire peut s'hybrider avec le monobrin de la sonde. Les puces sont ensuite lavées pour éliminer les fragments d'ADN qui ne se seraient pas fixés à la puce. Puis des nucléotides marqués et détectables par fluorescence sont ajoutés et se lient à la sonde de la puce en fonction du nucléotide complémentaire du fragment d'ADN et donc du SNP à analyser. La fluorescence du nucléotide ajouté et complémentaire de l'allèle du SNP à analyser permet ensuite de juger de la présence de tel ou tel allèle pour le SNP d'intérêt. Un SNP sera caractérisé par deux sondes, permettant d'identifier les 2 allèles du SNP. Le niveau de l'intensité de fluorescence de chaque sonde est ensuite mesuré avec un scanner et représente la force du signal associée à chaque allèle du SNP. L'analyse de l'intensité de fluorescence du SNP est réalisée sur l'ensemble des individus, puis un algorithme permet la formation de clusters distinguant les individus présentant un génotypage AA, AB ou BB pour le SNP étudié. Il arrive toutefois des cas où les clusters sont difficiles à identifier. C'est pourquoi il est nécessaire de réaliser un contrôle de la qualité des génotypages.

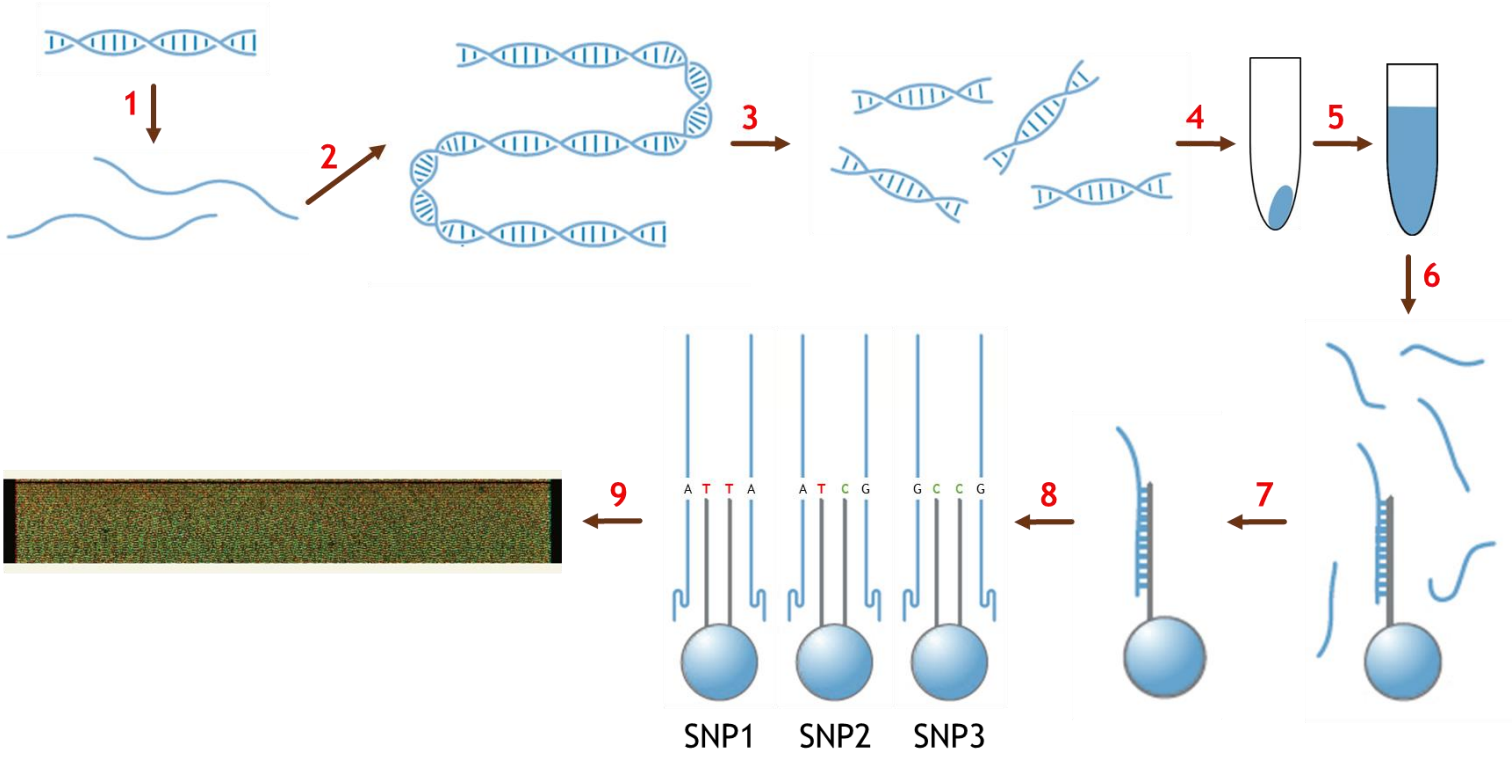


Figure 7. Principe du génotypage. D'après les technologies Illumina et Thermo Fisher. 1 : Dénaturation et purification de l'ADN ; 2 : Amplification de l'ADN ; 3 : Fragmentation enzymatique de l'ADN ; 4 : Précipitation des fragments d'ADN ; 5 : Mise en suspension des fragments d'ADN ; 6 : Hybridation des fragments d'ADN aux sondes de la puce ; 7 : Lavage de la puce ; 8 : Fixation des nucléotides fluorescents ; 9 : Détection de la fluorescence (avec ici TT qui renvoie un signal rouge, CC un signal vert, et TC un signal jaune).

2. Mise en place du contrôle de la qualité des génotypages et de la compatibilité entre pedigree et génotypages

Les erreurs de génotypages ou des données manquantes peuvent avoir des conséquences néfastes sur les estimations des valeurs génomiques. Le contrôle de la qualité des SNP pouvait se réaliser facilement avec le regretté package R GenABEL (Aulchenko et al., 2007). Ce package a été utilisé au début de la thèse mais n'est aujourd'hui plus disponible du fait d'un arrêt du principal soutien financier du projet et donc d'un arrêt de la maintenance du package. Il existe toutefois d'autres solution avec le logiciel Plink V1.9 (Chang et al., 2015). Plusieurs critères doivent ainsi être contrôlés pour vérifier la qualité des génotypages. Ces contrôles sont effectués soit au niveau des individus, soit au niveau des SNP, et s'organisent en 6 étapes successives.

a) Contrôle du call rate individu

Le premier contrôle mis en place concerne la qualité de génotypage par individu avec le calcul du pourcentage de marqueurs génotypés par rapport au nombre de marqueurs total de la puce utilisée pour obtenir les génotypages. On parle de contrôle du call rate individu. Ces problèmes de call rate individu peuvent souvent s'expliquer par des problèmes au niveau de la qualité de l'ADN prélevé pour l'individu qui n'aura alors pas un génotypage considéré comme fiable. Le seuil est généralement fixé à 95% de marqueurs génotypés mais peut varier en fonction des études et des espèces.

b) Contrôle du call rate SNP

Le deuxième point de contrôle consiste à étudier la qualité de génotypage par SNP en calculant le pourcentage d'individus génotypés pour un SNP par rapport au nombre total d'individus. Les problèmes de call rate SNP peuvent s'expliquer par des problèmes techniques de génotypage du SNP. Le seuil est généralement fixé à 95% mais peut là encore varier en fonction des études et des espèces. Guo et al. (2014) rapportent qu'un call rate moyen de 98% est généralement attendu pour des études avec un grand nombre d'individus.

Un problème de call rate SNP peut être visualisé lors de l'assignation des génotypages (basée sur la fluorescence). Il est alors difficile pour certains individus de savoir si le génotypage du SNP étudié est AA, AB ou BB (Figure 8a). En supprimant les individus problématiques pour le SNP étudié, il devient plus aisé de distinguer clairement les trois génotypes, et donc de déterminer le génotypage du SNP pour les individus restants (Figure 8b).

Marees et al. (2017) recommandent également de réaliser ce contrôle en deux temps en le liant au contrôle du call rate individu :

- Un premier contrôle en fixant les seuils des call rate individu et SNP à 80% pour enlever les individus et SNP avec de gros problèmes de génotypage.
- Un deuxième contrôle en fixant un seuil de call rate plus élevé à 95%.

Ce fonctionnement en deux temps permet de ne pas supprimer des individus ou des SNP trop vite. On peut par exemple imaginer un individu avec un call rate à 88%. Le premier contrôle de call rate avec un seuil à 80% permettrait de supprimer les SNP avec de gros problèmes de génotypage. En supposant que la valeur faible de son call rate était due aux SNP supprimés, le premier contrôle pourrait permettre de faire passer le call rate de l'individu au-dessus de 95%, et donc de le retenir à l'issu du deuxième contrôle. En choisissant de fixer directement le seuil à 95%, l'individu n'aurait pas été inclus dans l'étude.

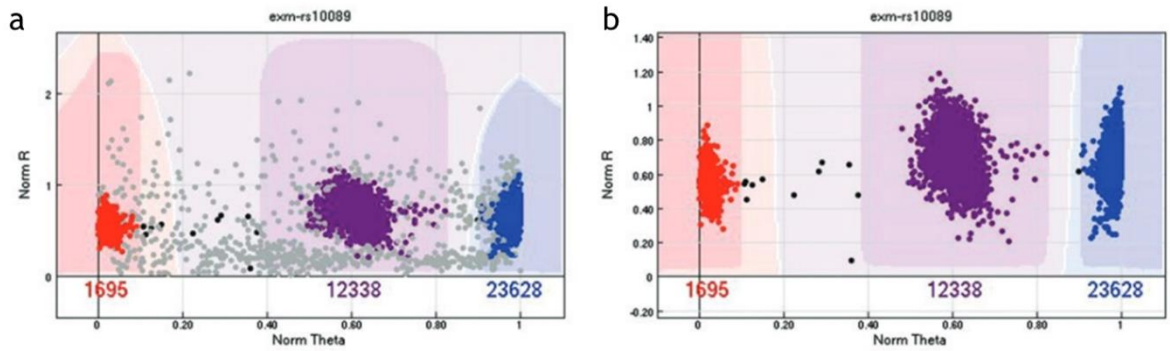


Figure 8. Exemple de SNP avant (a) et après (b) traitement du problème de call rate. Après suppression des individus avec un mauvais call rate pour le SNP, les clusters deviennent plus facilement identifiables. D'après Zhao et al., 2018.

c) Contrôle des fréquences alléliques

Le troisième point de contrôle concerne la fréquence allélique de l'allèle mineur (MAF). Lors du développement des différentes puces à SNP, plusieurs lignées peuvent être utilisées pour mettre au point et fixer les SNP à inclure sur la puce. C'est par exemple le cas de la puce commerciale à SNP haute densité en poule (Kranis et al., 2013) qui a été développée à partir de 24 lignées différentes, avec deux tiers des lignées correspondant à des lignées de poulets de chair et un tiers des lignées correspondant à des lignées de poules pondeuses. Cette puce permet donc de génotyper un grand nombre de lignées différentes mais le nombre de SNP informatifs (MAF > 0) est compris entre 100K et 450K SNP en fonction de la lignée étudiée. Ceci a également été illustré chez les bovins avec la puce HD de 777K (Illumina Inc, 2012a ; Pérez O'Brien et al., 2014). Le nombre d'individus porteurs de l'allèle mineur est faible, ce qui implique que les phénotypes associés à cet allèle sont également rares. Ces SNP manquent donc de puissance pour permettre d'estimer précisément leurs effets alléliques. Ceci peut être problématique pour les évaluations génomiques. Les SNP avec une faible MAF sont également plus sujet à des erreurs de génotypage (Marees et al., 2017) à cause de problèmes d'individus mal positionnés au niveau du cluster AB : soit ils sont positionnés dans le cluster AB alors qu'ils n'ont pas ce génotype (Figure 9a), soit ils sont en dehors du cluster AB alors qu'ils ont le génotype AB (Figure 9b) (Guo et al., 2014 ; Zhao et al., 2018).

Ils convient donc de fixer un seuil minimum de MAF à 5%. Là encore, en fonction des études et des espèces, les seuils peuvent varier. Dans le cas de grandes populations d'étude, ce seuil peut être abaissé (Marees et al., 2017).

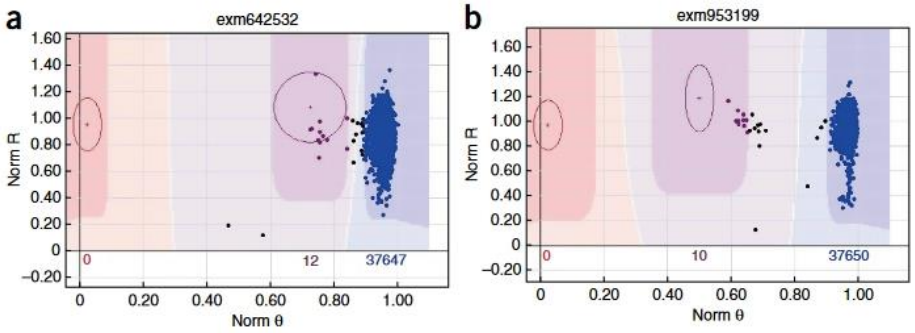


Figure 9. Exemple de SNP avec un problème de génotypage dû à des SNP à faible MAF. Dans le cas (a), les deux SNP à l'extrême droite du cluster AB sont considérés AB alors qu'ils devraient être BB ou non génotypé (entre deux clusters). Dans le cas (b), plusieurs SNP sont en dehors du cluster AB alors qu'ils sont AB. D'après Guo et al., 2014.

d) Contrôle de l'équilibre d'Hardy-Weinberg

Le quatrième point de contrôle permet de contrôler si les génotypes des SNP ayant passé les précédentes étapes du contrôle qualité vérifient l'équilibre d'Hardy-Weinberg. Cette théorie développée en 1908 suppose qu'au sein d'une population dite « idéale », il existe un équilibre des fréquences alléliques et génotypique d'un locus au cours des générations. En supposant un SNP avec deux allèles notés A et B, de fréquences respectives p et q, on doit alors vérifier que :

$$f(AA) = p^2 \; ; \; f(BB) = q^2 \; ; \; f(AB) = 2pq$$

Cette théorie s'applique si les hypothèses suivantes sont respectées :

- La population est de taille infinie.
- La population est constituée d'individus diploïdes à reproduction sexuée.
- La reproduction est panmictique (les gamètes s'associent au hasard).
- Il y a absence de migration, de mutation et de sélection.
- Les générations ne sont pas chevauchantes.

En pratique, ces conditions sont très rarement remplies… Un exemple tout simple est que nous étudions la sélection génomique !

Toutefois, Hosking et al. (2004) ont montré qu'une forte déviation de l'équilibre d'Hardy-Weinberg était généralement liée à des problèmes de génotypages du SNP chez certains individus et donc à un problème de clustering. Une p-value assez faible de 10^{-4} est généralement choisie pour supprimer les marqueurs présentant une déviation de l'équilibre.

Enfin, dans le cas d'une étude mélangeant différentes races ou lignées, Guo et al. (2014) précisent que ce test est spécifique à chaque population. Le contrôle de l'équilibre d'Hardy-Weinberg doit donc être réalisé intra-race ou intra-lignée.

e) Contrôle de l'hétérozygotie moyenne des individus

Le cinquième point de contrôle concerne le calcul du pourcentage de génotypage hétérozygote chez un individu. L'hétérozygotie moyenne d'un individu correspond au pourcentage du nombre de marqueurs génotypés hétérozygotes par rapport au nombre total de marqueurs génotypés. Un taux d'hétérozygotie moyen trop faible est signe d'un ADN de mauvaise qualité ou d'un problème de consanguinité. Une valeur trop élevée indique une possible contamination de l'échantillon d'ADN de l'individu par un autre (Guo et al., 2014 ; Zhao et al., 2018).
Marees et al. (2017) suggèrent d'enlever les individus déviant de plus de trois écart-types par rapport à la moyenne du taux d'hétérozygotie de la population.

f) Contrôle du pedigree et des problèmes d'incompatibilité

Le sixième et dernier point de contrôle permet de vérifier la compatibilité entre le pedigree et les génotypages. Ce point de contrôle peut être vérifié de deux manières différentes. La première passe par l'identification de problème de compatibilités entre paires d'individus. En supposant que les parents d'un individu sont AA et TT pour un SNP, le descendant est alors forcément AT pour ce SNP. Une différence au niveau du génotype du SNP du descendant permet de repérer une erreur de compatibilité. Si le nombre d'incompatibilités est supérieur à un certain seuil, il est alors possible d'identifier un problème de pedigree.
La deuxième manière passe par le calcul de l'identité par descendance (IBD) entre paire d'individus. Cet IBD caractérise la probabilité pour deux individus A et B que deux allèles d'un locus pris au hasard aient la même origine. En fonction des relations de parenté entre individus, l'IBD peut prendre différentes valeurs (Zhao et al., 2018) :

- IBD = 0 dans le cas d'une absence de relation génétique
- IBD > 0.125 dans le cas d'une relation de 3ème degré (cousins, etc.)
- IBD > 0.25 dans le cas d'une relation de 2nd degré (oncle, tante, demi-frère ou sœur, etc.)
- IBD > 0.5 dans le cas d'une relation de 1er degré (frère, sœur, descendants directs, etc.)
- IBD > 0.95 dans le cas d'échantillons dupliqués

En fonction du degré de parenté entre individus, une déviation trop importante par rapport à ces valeurs indique un problème de compatibilité entre les deux individus. Le dernier cas indique un problème de manipulation des échantillons d'ADN, l'ADN d'un individu n'étant en fait pas le bon et étant remplacé par celui d'un autre individu déjà analysé.
Cette étape de vérification de l'association entre génotypage et individu, et donc du génotypage avec les phénotypes associés, est une étape très importante afin d'éviter des erreurs dans l'estimation de l'effet des SNP sur différents caractères d'intérêt.

D. Le cas particulier de la ponte

1. Organisation de la sélection en filière pondeuse

La filière avicole est organisée autour de grands groupes qui intègrent l'amont (sélection et alimentation) et l'aval de la filière (abattage et transformation) de façon différente en fonction des groupes. Les différents maillons de la chaine peuvent donc être indépendants ou liés entre eux au sein d'un même groupe comme c'est souvent le cas entre les fabricants d'aliments et les organisations de production (Figure 10).

Dans cette filière, au niveau du maillon de la production, la grande majorité des éleveurs de poules pondeuses appartient à une organisation de production qui a en charge la gestion des plannings de production. Ces organisations gèrent la production et l'approvisionnement des intrants. Tout en amont de la filière se trouve le maillon de la sélection.

La sélection avicole est très différente de la sélection des ruminants et des porcins car il n'y a aucune organisation collective à l'échelle nationale. La sélection est gérée par des entreprises privées ayant chacune leurs propres stratégies, schémas de sélection et objectifs de production. En filière pondeuse, 4 groupes se partagent la totalité du marché (Figure 11) (communication interne, T. Burlot, Novogen) :

- Le groupe Erich Wesjohann avec les sociétés Lohmann tierzucht, Hyline et H&N
- Le groupe Hendrix avec la société Hendrix layers et ses différentes marques : ISA, Dekalb, Bovans, Babcock, Shaver, Hisex
- Le groupe Grimaud avec la société Novogen
- La société Tetra

Dans cette situation sans organisation collective, certains sélectionneurs avicoles français (en l'occurrence Novogen pour la sélection de pondeuse) ont choisi de s'appuyer sur les connaissances et les compétences du SYSAAF (Syndicat des Sélectionneurs Avicoles et Aquacoles Français). La principale mission du SYSAAF est une mission de conseil et d'appui aux sélectionneurs adhérents. Elle consiste à assister les sélectionneurs adhérents dans le calcul des valeurs génétiques, dans le choix des animaux reproducteurs et dans la réalisation des plans d'accouplements selon leurs besoins. Dans cette situation, c'est le sélectionneur qui indique ses objectifs au SYSAAF et qui prendra la décision finale en fonction des scénarii développés par le SYSAAF pour répondre aux différents objectifs. La deuxième mission du SYSAAF est un travail de recherche et de développement en collaboration avec l'INRA, l'ITAVI (Institut Technique de l'Aviculture) et les sélectionneurs adhérents pour répondre aux besoins futurs des adhérents et anticiper les attentes sociétales.

Enfin, du fait de la gestion de la sélection par les entreprises elles-mêmes, cela entraîne une très grande compétitivité entre les différentes sociétés et groupes, rendant de surcroit l'accès à l'information compliqué. En attestent les chiffres concernant les parts de marchés détenus par les différents groupes qui ne sont qu'estimés très vaguement (communication interne, T. Burlot, Novogen) ! L'accès compliqué à l'information peut-être un atout pour les différents groupes qui peuvent prendre une avance considérable en travaillant sur un sujet novateur sur lequel les autres sociétés ne se sont pas encore positionnées. Mais c'est également un frein pour les différents groupes lorsqu'il s'agit de mettre en place des projets de collaboration au niveau mondial ou européen. En atteste l'exemple de la création en 2013 de la puce à SNP HD commerciale de 600 000 marqueurs intervenue 9 ans après la publication du génome de référence de la poule pondeuse et après le développement en interne de différentes puces moyennes densités propres à chaque société, et donc non commerciales. En bovin, les puces moyenne (54K) et haute densité (777K) commerciales ont été développées respectivement 3 ans et 6 ans après la publication du génome de référence.

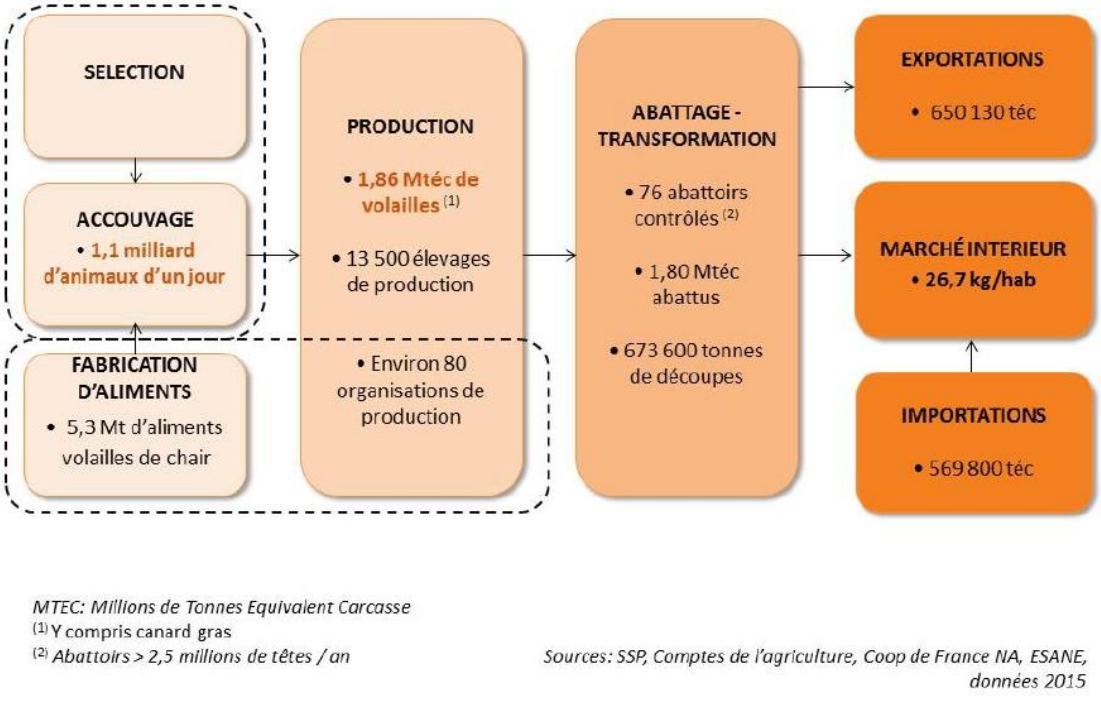


Figure 10. Organisation de la filière pondeuse.

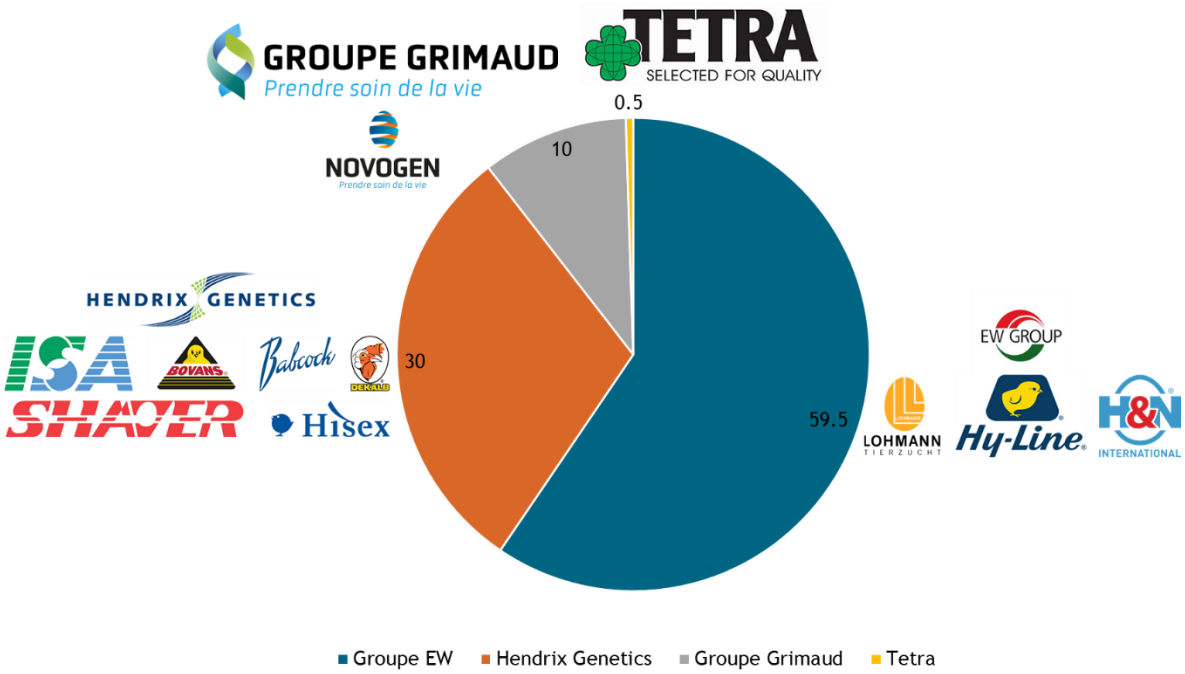


Figure 11. Parts du marché mondial occupées par les sociétés des différents groupes de sélection de la filière pondeuse (communication interne, T. Burlot, Novogen).

2. Mise en place de l'amélioration génétique et du schéma de sélection

a) Une organisation pyramidale du schéma de sélection

La filière avicole est organisée selon une structure pyramidale, avec un premier étage de sélection, deux étages de multiplication (grand-parentales et parentales), et un étage de production (pondeuses commerciales avec la production d'œufs) (Figure 12). Ce schéma pyramidal repose généralement sur le croisement de 4 lignées pures avec 2 lignées mâles (lignées A et B) et 2 lignées femelles (lignées C et D). La première étape de la multiplication est réalisée par les sélectionneurs qui produisent les individus grands parentaux mâles et femelles par croisement intra-lignées. La deuxième étape de la multiplication consiste ensuite à réaliser des croisements des grands parentaux mâles et femelles. Ainsi les individus de la lignée A sont croisés avec ceux de la lignée B, et les individus de la lignée C sont croisés avec ceux de la lignée D. Ces croisements permettent d'aboutir à la production d'individus parentaux mâles AB et femelles CD respectivement. Les sélectionneurs fournissent ensuite aux accouveurs ou au multiplicateurs les individus parentaux dont le croisement permet d'obtenir des poussins commerciaux qui forment la lignée commerciale. Ces individus de la lignée commerciale sont ensuite mis en place dans des élevages de production pour la production d'œufs de consommation.

Cette organisation permet une multiplication du nombre d'individus et une diffusion du progrès génétique réalisé à l'étage de sélection (Guéméné et al., 2011). L'utilisation du croisement dans

ce schéma pyramidal permet de bénéficier au niveau de la lignée commerciale de la complémentarité des lignées pures concernant certains caractères de reproduction ou de production. Il permet également d'exploiter l'effet d'hétérosis qui intervient lors du croisement d'individus de fonds génétique différents et qui permet d'obtenir, pour les individus croisés, des performances supérieures à la moyenne des performances des populations parentales. Enfin, dans un contexte de forte compétition entre entreprises de sélection, cette organisation permet de ne pas avoir à diffuser la génétique des lignées pures.

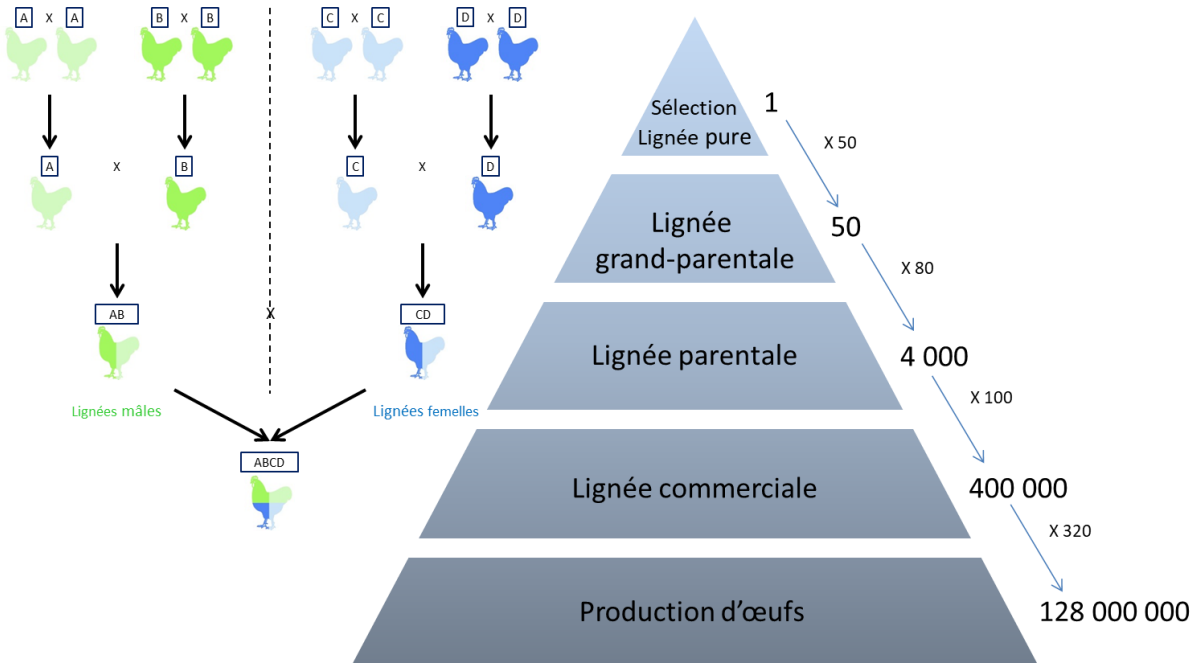


Figure 12. Organisation pyramidale de la filière ponte - Exemple de Novogen. Les lignées A et B correspondent aux lignées mâles, les lignées C et D correspondent aux lignées femelles. Les lignées grand-parentales sont issues des croisements intra-lignées et les lignées parentales sont issues des croisements entre lignées mâles (AB) et femelles (CD). La lignée commerciale est obtenue par croisements des lignées parentales mâles et femelles.

b) Les différents objectifs et critères de sélection

La mise en place d'une sélection des meilleurs individus au sein d'une lignée nécessite l'identification préalable des objectifs et critères de sélection. Les objectifs de sélection correspondent aux objectifs globaux d'amélioration des individus pour chaque lignée. Les objectifs nécessitent la prise en compte de tous les acteurs de la filière, une bonne adéquation entre spécificité de la race ou de la lignée et l'utilisation des individus, ainsi qu'une anticipation des évolutions du marché et des systèmes de production. Ces objectifs sont traduits en critères de sélection correspondant à des caractères ou aptitudes zootechniques mesurables. Ces critères

doivent être facilement mesurables sur le terrain, variables dans la population et héritables. Les objectifs de sélection en filière ponte sont multiples et peuvent donc être traduits en un grand nombre de critères mesurables (Tableau 1).

Tableau 1. Exemple d'objectifs et de critères de sélection pour la filière pondeuse.

Objectif de sélection	Critère de sélection
Production d'œufs	- Intensité de ponte (fonction de la période de production) - Persistance de ponte - Âge au pic de ponte
Qualité interne et externe des œufs	- Poids de l'œuf - Couleur de la coquille - Force de fracture de la coquille - Forme de la coquille - Consistance du blanc d'œuf - Taux de jaune d'œuf - Inclusions (tache de viande ou de sang)
Efficacité alimentaire	- Mesure de la quantité d'aliment ingérée - Indice de consommation - Nombre de repas
Viabilité des animaux	- Piquage des animaux et cannibalisme - Présence de fractures (ex : du bréchet) - Résistance aux maladies
Comportement des animaux	- Taux de ponte au nid (élevage au sol) - Nombre de nids explorés (élevage au sol)

3. Intérêt de la sélection génomique en filière pondeuse

a) Gain de progrès génétique

Comme présenté dans la section I.B.4.a, un gain de progrès génétique peut être obtenu en modifiant l'intensité de sélection, l'intervalle de génération et la précision des évaluations. Le sélectionneur avicole peut justement intervenir sur ces trois facteurs.

Prenons l'exemple d'un schéma de sélection classique de poules pondeuses avec 50 coqs et 200 poules. Ces animaux sont mis en reproduction pour produire 2000 mâles et 2000 femelles. Supposons ensuite que dans ce schéma, seul 1 mâle sur 10 soit conservé à l'éclosion, soit 1 fils

par mère. Ces individus sont ensuite conservés pendant environ 80 semaines, le temps d'obtenir les phénotypes de leurs sœurs. À ce moment, les individus peuvent être évalués et sélectionnés de façon à ne garder que les 50 meilleurs mâles, comme reproducteurs de la génération suivante. Avec l'application de la sélection génomique dans ce schéma de sélection, il peut être envisageable de conserver les 2000 mâles le temps d'obtenir leurs génotypes et de réaliser une évaluation génomique de ces individus. Les 50 meilleurs individus sont ensuite choisis et conservés pour les mettre en reproduction une fois la puberté atteinte, soit environ à 30 semaines. La mise en place de la sélection génomique permettrait donc en théorie, une augmentation de l'intensité de sélection en passant d'une sélection de 50 mâles parmi 200 à 50 mâles parmi 2000. Une diminution de l'intervalle de génération serait également possible en passant de 80 semaines à 30 semaines. En revanche, Picard Druet et al. (2019a) ont montré qu'il y avait une potentielle dégradation des évaluations génomiques pour certains caractères, comparativement aux précisions atteignables lorsque les mâles sélectionnés ont des performances.

La structure pyramidale peut également impliquer des différences d'environnements entre les élevages de sélection, multiplication et production. Ceci peut être problématique lorsqu'un individu est sélectionné dans un type d'environnement et qu'il doit exprimer son potentiel dans un environnement différent. Toutefois, ce problème d'interaction du génotype avec l'environnement peut être pris en compte avec la sélection génomique (Le Roy et al., 2014).

Enfin, il convient de noter que les objectifs de progrès génétique sont définis en race pure. Or le produit commercial est un individu issu de plusieurs croisements. Picard-Druet et al. (2019b) ont montré que l'inclusion des performances des individus croisés dans les évaluations génomiques d'individus de lignée pure permettait d'obtenir, en fonction des caractères étudiés, des précisions équivalentes ou meilleures qu'en incluant seulement les performances d'individus de la lignée pure.

b) Gain économique

En filière ponte, le coût total d'un phénotypage est complexe à calculer. Toutefois il est possible d'estimer un coût R&D hors génotypage à environ 60€ par individu. Ce coût comprend l'alimentation, l'utilisation du bâtiment, les charges R&D. En comparaison, le coût du génotypage avec une puce HD est bien plus élevé. Ce coût peut être un frein au développement de la sélection génomique en poules pondeuses. En effet, la puce à SNP HD coûte environ 150€. Le prix d'une poussin commercial est estimé entre 60 et 70 centimes par individu. Il faut multiplier par 1000 ce prix pour estimer le prix d'un individu reproducteur. Dans la réalité, ces individus ne sont pas vendus et n'ont donc pas de valeur marchande. Leur prix reste toutefois

très éloigné des 50 000$ cités par Schaeffer (2006) pour un taureau laitier au Canada ! Mais le gain économique d'un reproducteur au niveau du schéma de sélection ne dépend pas que de sa valeur génétique. Elle dépend également du nombre de fois où elle est exprimée dans sa descendance.

D'après la figure 12, un coq permet la production de plus de 400 000 poules pondeuses à l'étage de la production, et la production de plus de 128 millions d'œufs. Toutefois le progrès génétique est réalisé à l'étage de la sélection. Un coq transmettant en moyenne 50% de son information génétique à sa descendance et 3 générations séparant le coq des poules pondeuses de l'étage de production, le gain économique doit être divisé par 8. En supposant un gain d'un gramme au niveau du poids d'œuf, d'une valeur de 0,8 centimes d'euros, le gain économique sera alors de 128 000€ ! Ceci permet de rentabiliser les 150€ investis dans le génotypage HD du coq. Toutefois, il est à noter que, même si les reproducteurs sont vendus plus cher, les gains réalisés ne bénéficient pas totalement au sélectionneur qui a payé le génotypage… Une solution pour le sélectionneur est alors d'utiliser des puces à plus basse densité coûtant moins cher (environ 35€ pour une puce de 10K SNP) puis d'utiliser la technique de l'imputation pour remonter à l'information des puces HD.

II. Utilisation de l'imputation pour la sélection génomique

A. Intérêt de l'imputation

L'imputation consiste à pr_d_re les l_t_res ma_qu_nte_ d__s des m_ts ou d_s ph_a_es et r_pos_ s_r l'_tili__ti_n d_ dé_é_uil___e de l_ai_o_ (réponse page 62). En génétique, l'imputation est le remplacement du génotypage manquant à un SNP par son génotypage prédit.

1. Utilisation de l'imputation pour combler des données manquantes

Après un contrôle de la qualité des génotypages des différents individus, il est encore possible que certains SNP soit non génotypés pour différents individus ce qui génère des génotypages incomplets. Plutôt que de restreindre les génotypages aux données complètes, il est possible de combler les données manquantes.

La première façon, basique, de « remplir les trous » serait de s'appuyer sur les fréquences alléliques des SNP puis de remplacer le génotype manquant par le génotype le plus probable d'après le calcul des fréquences génotypiques sous l'équilibre d'Hardy Weinberg. D'après la figure 12, pour le SNP 1, les fréquences alléliques sont $f(A) = 0.4$ et $f(G) = 0.6$. Sous l'équilibre d'Hardy Weinberg, les fréquences génotypiques sont alors $p(AA) = 0.4^2 = 0.16$,

$p(AG) = 2 * 0.4 * 0.6 = 0.48$ et $p(GG) = 0.6^2 = 0.36$. Le génotype du SNP 1 de l'individu 4 est alors considéré comme AG. Le problème de cette méthode est qu'elle modifie le déséquilibre de liaison avec les marqueurs proches et qu'elle conduit à des biais dans l'estimation des effets des SNP sur les phénotypes d'intérêt. Il est alors plus intéressant de remplacer le génotype manquant par le génotype prédit sur la base des génotypes observés pour les SNP voisins, permettant ainsi de prendre en compte le déséquilibre de liaison du SNP manquant avec ses voisins. C'est ce que l'on peut voir avec la figure 13 où la première méthode, basée sur les fréquences alléliques, peut aboutir à 5 génotypes erronés (en rouge). La méthode utilisant la connaissance des génotypes observés aux SNP voisins, et donc le déséquilibre de liaison entre SNP, permet de remarquer pour les SNP 1 à 4 que les génotypages des individus 1, 2 et 5 sont respectivement les mêmes que ceux des individus 3, 4 et 6. De la même façon pour les SNP i à l, les génotypages des individus 1, 2 et 3 sont respectivement les mêmes que ceux des individus 6, 5 et 4. Enfin, il est également possible d'inclure l'information familiale pour prédire les génotypes manquants.

Les méthodes et les logiciels permettant de réaliser ces imputations sont présentés dans la section II.C.2.

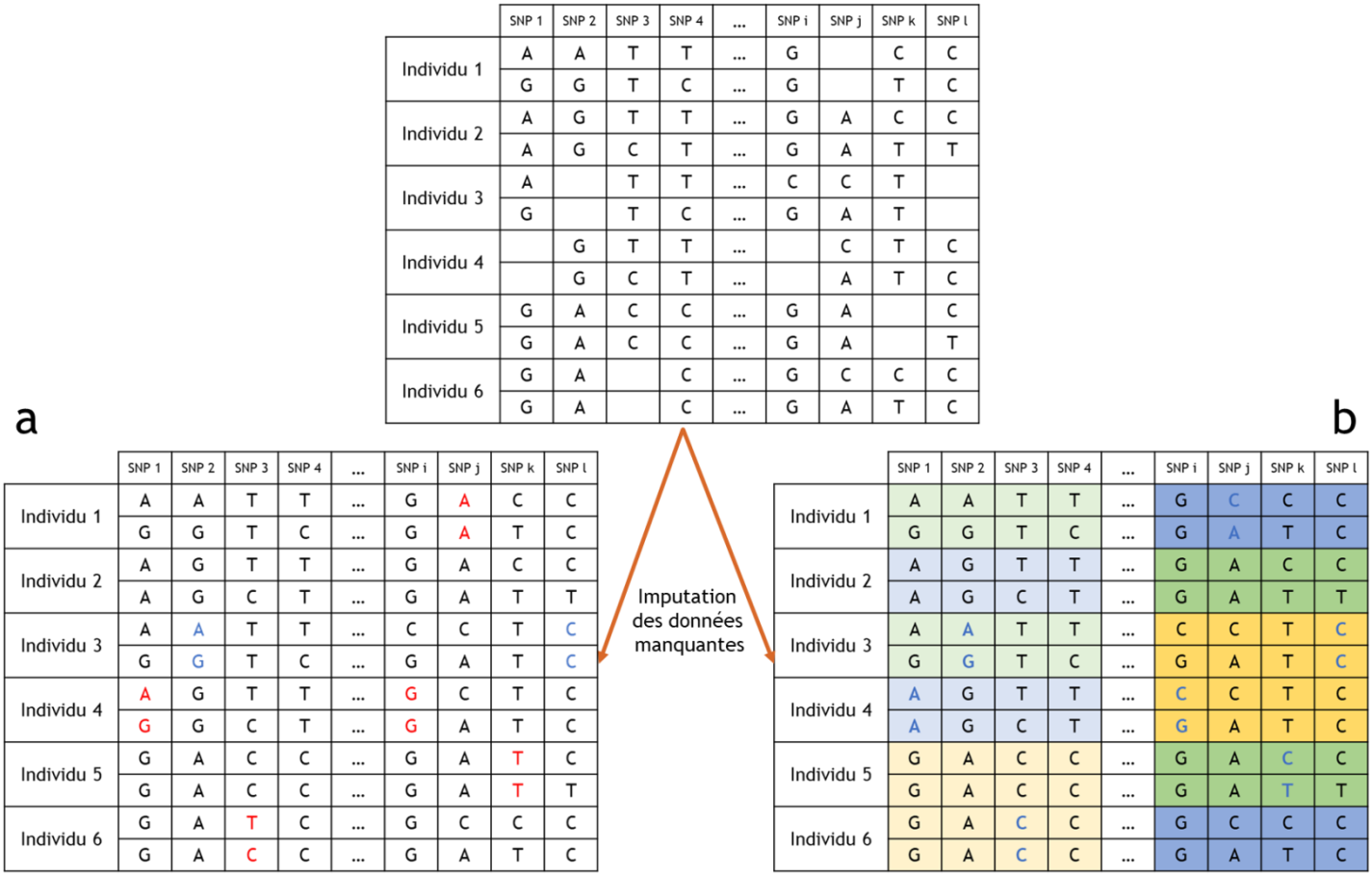


Figure 13. Imputation des données manquantes sur la base des fréquences alléliques (a) ou sur la base du déséquilibre de liaison (b).

2. Utilisation de l'imputation pour remonter à des densités plus élevées de génotypages

Comme vu précédemment, le coût des puces à SNP HD pour génotyper l'ensemble des candidats à la sélection peut s'avérer assez onéreux. La solution est alors de génotyper avec une puce à SNP de plus basse densité l'ensemble des candidats à la sélection. Puis en se basant sur une population de référence génotypée avec la puce à SNP HD, il est possible pour les candidats à la sélection de remonter à l'information des génotypages HD, ceci à la condition d'avoir un certain nombre de SNP en communs entre puces HD et BD comme illustré avec la Figure 14. Cette utilisation de l'imputation est largement utilisée chez les espèces d'élevages et est détaillée dans la section III.A. La première publication de cette thèse s'est également concentrée sur ce sujet (cf. Chapitre II).

L'imputation permettant de remonter à des densités plus élevées, elle peut donc, sur le même principe, permettre de remonter aux données de séquences (Hayes, 2011).

Enfin, cette méthode peut être utilisée pour mutualiser les génotypages de différents individus génotypés sur différentes puces, comme cela peut être le cas entre différents pays ou au sein d'un même pays comme en bovin avec des candidats mâles génotypés en MD et des femelles génotypées en BD. Cela peut également être le cas entre différents protocoles expérimentaux (Druet et al., 2010).

Les différentes méthodes permettant de remonter aux génotypages HD utilisent le déséquilibre de liaison et peuvent inclure l'information familiale. Ces méthodes seront détaillées dans la section II.C.

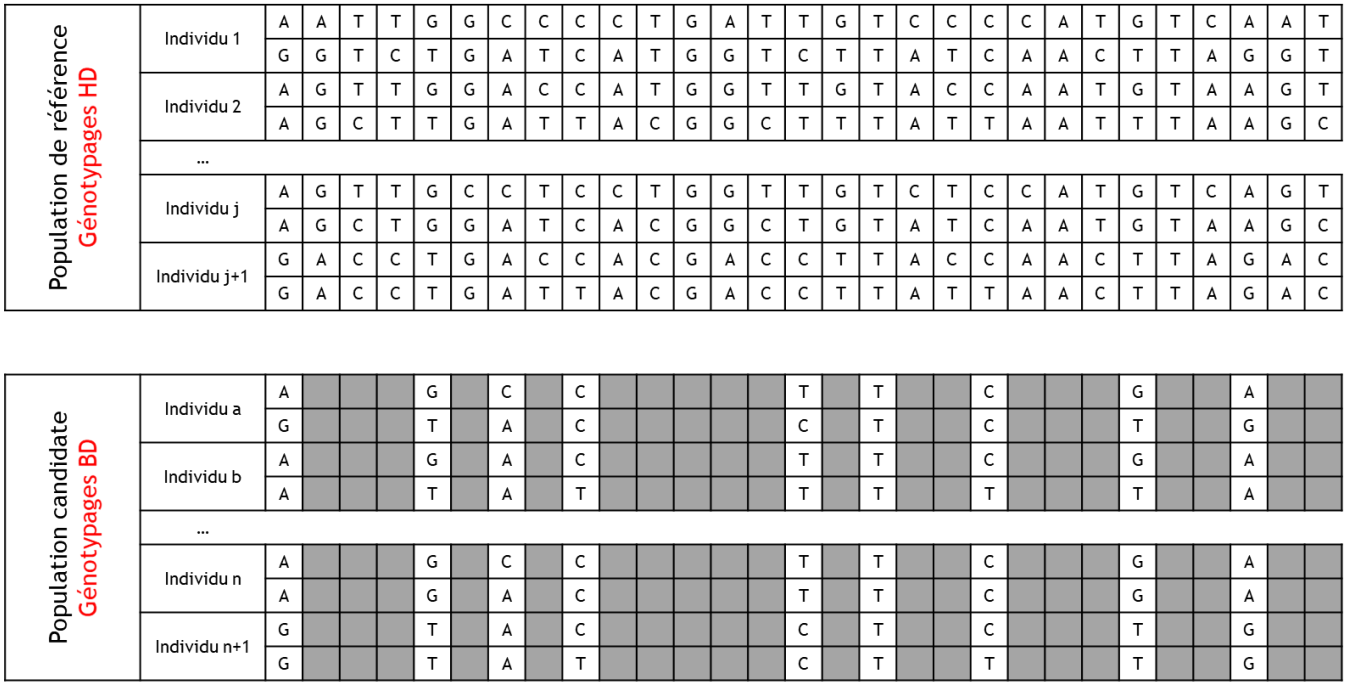


Figure 14. Imputation des génotypages BD de la population candidate à partir des génotypages HD de la population de référence.

B. Importance du déséquilibre de liaison

1. Définition du déséquilibre de liaison

a) Principe et propriétés du déséquilibre de liaison

Le déséquilibre de liaison (DL) se définit comme une association non aléatoire d'allèles à différents loci. Dans le cas de SNP, les loci sont bialléliques. Supposons deux SNP avec respectivement des allèles A/a et B/b. Le déséquilibre de liaison D se calcule selon la formule : $D = x_{AB} - p_A p_B$, avec x_{AB} la fréquence de l'haplotype AB, et p_A et p_B la fréquence respective des allèles A et B (Lewontin, 1964). Dans le cas d'un équilibre de liaison, $D = 0$. Dans le cas d'un déséquilibre de liaison, $D > 0$. Ces deux cas sont illustrés avec le tableau 2.

Un parent transmet à ses descendants un brin de chacun de ses chromosomes. Toutefois, des recombinaisons peuvent se produire entraînant un échange d'une portion de brin entre chromosomes parentaux. Le DL se maintient d'autant plus dans le temps, et donc au fil des générations, que les deux loci sont proches, et donc liés. Ardlie et al. (2002) ont montré que l'évolution du DL au fil des générations suivait la formule suivante : $D_t = (1 - r)^t D_0$ avec D_0 et D_t le DL aux générations 0 et t, et r le taux de recombinaison. En supposant deux loci proches, le taux de recombinaison est faible, et le DL peut se maintenir à travers les générations. En revanche, pour deux loci plus éloignés, le taux de recombinaison est élevé et le DL chute d'autant plus vite que le taux de recombinaison est élevé.

Tableau 2. Exemple d'une situation d'équilibre et de déséquilibre de liaison.

	Nombre d'haplotypes	
	Cas 1	Cas 2
	42	50
	28	20
	18	10
	12	20
Fréquences alléliques	$F(A) = 0.7 = p_A$ $F(a) = 0.3 = p_a$ $F(B) = 0.6 = p_B$ $F(b) = 0.4 = p_b$	
Fréquences haplotypiques théoriques	$F(AB) = 0.42$ $F(Ab) = 0.28$ $F(aB) = 0.18$ $F(ab) = 0.12$	
Fréquences haplotypiques observés	$x_{AB} = \frac{42}{100} = 0.42$ $x_{Ab} = \frac{28}{100} = 0.28$ $x_{aB} = \frac{18}{100} = 0.18$ $x_{ab} = \frac{12}{100} = 0.12$	$x_{AB} = \frac{60}{100} = 0.50$ $x_{Ab} = \frac{20}{100} = 0.20$ $x_{aB} = \frac{10}{100} = 0.10$ $x_{ab} = \frac{20}{100} = 0.20$
	Équilibre de liaison D = 0	Déséquilibre de liaison D ≠ 0

b) Les forces évolutives à l'origine du déséquilibre de liaison

Quatre forces évolutives sont à l'origine du DL : la mutation, la migration et le mélange de population, la dérive génétique et la sélection.

La mutation peut toucher un ou plusieurs nucléotides par changement, addition ou suppression d'un ou plusieurs nucléotides. Les mutations sont généralement rares, apparaissent ponctuellement et entraînent la création d'un nouvel haplotype. La mutation est donc uniquement associée à l'haplotype et aux allèles correspondants. Il y a alors un DL total entre le nouvel haplotype et les marqueurs adjacents. Cette mutation ainsi que le DL associé peuvent se transmettre aux descendants et devenir plus fréquents, soit par dérive génétique, soit par sélection.

La migration et le mélange de populations initialement en équilibre de liaison peuvent entrainer la création d'un DL à cause de fréquences alléliques différentes pour les deux populations. Ce DL peut ensuite chuter au fur et à mesure des générations pour revenir à un état d'équilibre.

La dérive génétique est également une cause de création du DL. Elle correspond aux variations aléatoires des fréquences alléliques liées à l'utilisation d'un nombre limité de reproducteurs pour produire la génération suivante. En supposant l'utilisation massive de 10 individus parmi 50 comme reproducteurs dont un individu avec un DL complet entre deux marqueurs, ce DL est transmis à sa descendance dans des proportions plus élevées que si l'ensemble des individus avaient été utilisés pour produire la génération suivante. La dérive génétique peut même aller jusqu'à la disparition aléatoire de certains allèles et haplotypes ce qui va diminuer la chute du DL au fil des générations (Ytournel, 2008).

Enfin, la sélection est également une des origines du DL. La sélection implique le choix et l'utilisation d'un nombre fini de reproducteurs pour produire la génération suivante (comme avec la dérive génétique) ce qui peut diminuer le nombre d'haplotypes présents dans la population, augmenter la fréquence des allèles ayant un effet positif sur les différents caractères d'intérêt et donc créer un DL entre les différents marqueurs soumis à la sélection. La sélection, diminue également la chute du DL entre certains marqueurs au cours des générations.

c) Importance du déséquilibre de liaison pour la sélection génomique

Comme expliqué précédemment, l'utilisation de marqueurs moléculaires pour la sélection génomique repose sur l'association, et donc sur le déséquilibre de liaison, entre allèle d'un marqueur et allèle d'un QTL d'intérêt. Dans le cas d'un DL entre allèle du marqueur et allèle du QTL, la connaissance de l'allèle au marqueur permet de prédire l'allèle au QTL (Boichard, 2013). Ainsi, il est possible d'extrapoler le génotype d'un individu à différents QTL en disposant du génotype de l'individu aux différents marqueurs moléculaires.

En se basant sur une population de référence génotypée et phénotypée, il est donc possible d'établir les équations permettant d'associer les effets des allèles des SNP aux différents phénotypes d'intérêt. En appliquant ces équations sur les candidats à la sélection pour lesquels seuls les génotypages sont disponibles, leur valeur génétique pour les caractères d'intérêt peut alors être prédite (Figure 15).

Par ailleurs, il est également important de disposer d'une densité de marqueurs suffisamment importante pour pouvoir établir correctement les équations de prédiction des valeurs génomiques. En effet, avec trop peu de marqueurs, la distance entre marqueurs et QTL d'intérêt peut être assez élevée et est donc beaucoup plus sujette à des recombinaisons faisant chuter le DL entre allèles aux marqueurs et aux QTL. Ceci explique pourquoi la première puce de 3K SNP développée en poule pondeuse ne permettait pas d'obtenir de résultats satisfaisant (Qanbari et al., 2010). À l'inverse, la puce commerciale HD de 600K SNP est suffisamment

dense pour disposer de marqueurs proches des QTL d'intérêt et ainsi exploiter le DL entre les allèles aux marqueurs et aux QTL (Megens et al., 2009).

Enfin, dans la partie précédente, il a été montré que ce DL évoluait au fil des générations à cause des recombinaisons. C'est pour cela qu'il est très important de renouveler la population de référence afin de ne pas déconnecter les effets prédits des allèles des SNP des phénotypes observés (Robert-Granié et al., 2011).

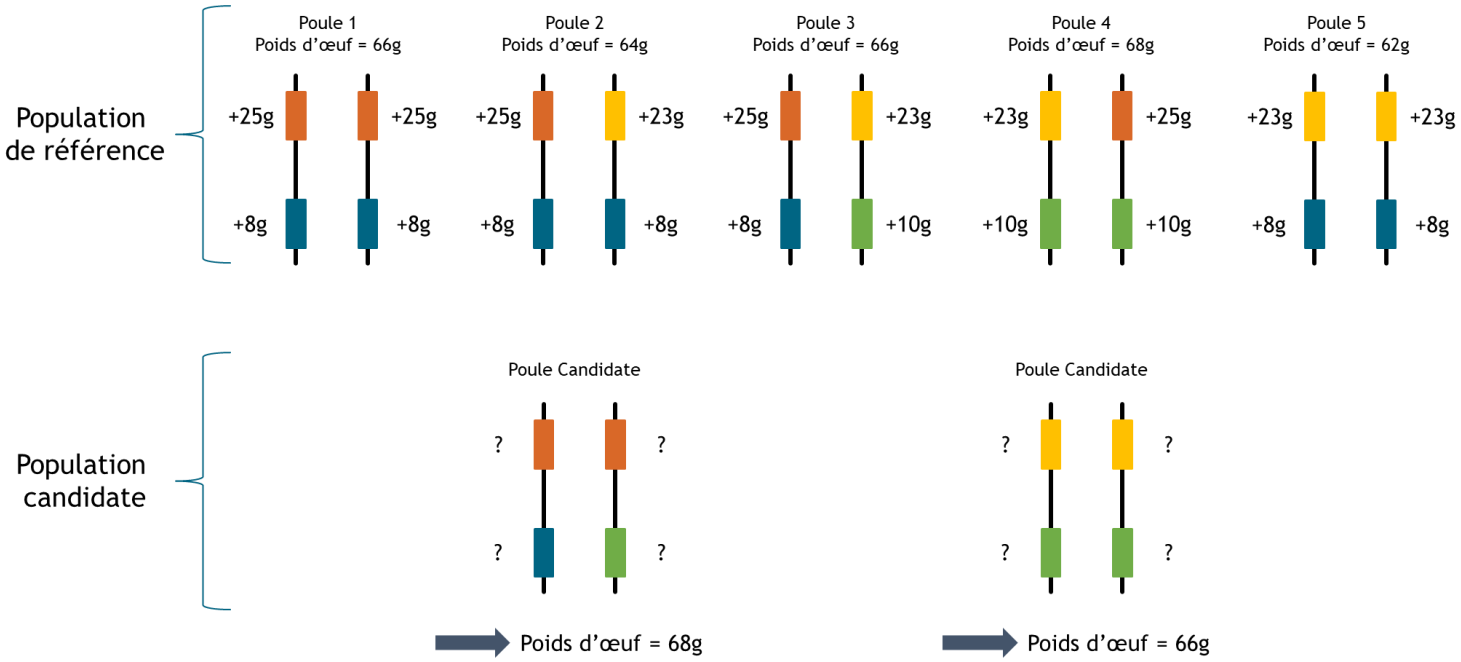


Figure 15. Détermination des effets des différents allèles sur le poids d'œuf.

2. Mesures du déséquilibre de liaison

La mesure du DL D présentée dans la partie précédente est dépendante des fréquences alléliques et haplotypiques. La valeur numérique de D ne permet pas de juger de la force de l'association entre les allèles de différents loci, ni de comparer les résultats entre eux. La mesure idéale du DL ne doit pas dépendre des fréquences alléliques (Ytournel, 2008). Il existe d'autres mesures du DL, toute dépendante du calcul du D de Lewontin, mais ayant des propriétés différentes et ne mesurant pas toujours la même chose (Ardlie et al., 2002).

La première mesure du DL est celle du déséquilibre normalisé D' (Lewontin, 1988). Cette mesure est calculée de la façon suivante :

$$D' = \frac{D}{D_{max}} \text{ avec } Dmax = \begin{cases} \min(p_A p_b, p_a p_B) \; si \; D > 0 \\ \min(p_A p_B, p_a p_b) \; si \; D < 0 \end{cases}$$

Cette mesure prend des valeurs comprises entre -1 et 1. Le cas où $D' = 1$ implique un déséquilibre de liaison total. Ceci indique alors que les associations entre allèles de deux marqueurs sont toujours les mêmes (par exemple, les allèles A et a du SNP 1 seront toujours

respectivement associés aux allèles B et b du SNP 2). En revanche, les valeurs intermédiaires sont moins interprétables et comparables entre SNP. Par ailleurs, le D' dépend du nombre d'individus utilisés et est fortement surévalué pour les SNP avec des fréquences alléliques faibles (Ardlie et al., 2002 ; Weiss et Clark, 2002), permettant d'observer un certain niveau de DL pour des marqueurs qui sont en réalité en équilibre.

Une autre mesure du DL est celle du coefficient de corrélation entre allèles r^2 (Hill et Robertson, 1968). Cette mesure est calculée en appliquant la formule :

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

Cette mesure prend des valeurs comprises entre 0 et 1. Le cas $r^2 = 1$ correspond à un cas de DL total. Cette mesure standardise la mesure du DL entre SNP en corrigeant des fréquences alléliques. Elle est donc moins sensible aux fréquences alléliques extrêmes (Du et al., 2007) et est également moins sensible à la taille de la population étudiée (Weiss et Clark, 2002). Enfin, cette mesure est utilisée comme mesure de référence dans les analyses de DL de populations animales basées sur des SNP. C'est donc cette mesure qui a été utilisée au cours de la thèse pour calculer le DL des différentes lignées utilisées.

Par ailleurs, un r^2 moyen supérieur à 0.3 est considéré comme un seuil de DL utile pour réaliser des études d'association ou pour la sélection génomique (Ardlie et al., 2002 ; Aerts et al., 2007). Le r^2 donne une indication du pouvoir de détection d'une association entre caractère d'intérêt et marqueurs et permet ainsi d'estimer le nombre d'individus ($1/r^2$) nécessaires pour détecter une association avec le polymorphisme causal directement. Une valeur de $r^2>0.3$ est la plus souvent choisie car elle est un bon compromis entre l'augmentation du nombre d'individus nécessaire pour détecter l'association avec le polymorphisme causal et le niveau de l'association (corrélation) entre marqueurs.

3. Déséquilibre de liaison dans les populations animales

Le déséquilibre de liaison et son évolution ont été étudiés chez de nombreuses espèces animales et pour de nombreuses races. Certaines études ont utilisé des marqueurs microsatellites pour estimer le DL moyen entre marqueurs, d'autres ont utilisé des marqueurs SNP, plus nombreux que les précédents marqueurs et permettant d'obtenir des mesures plus fines du DL. En effet, comme expliqué dans la section II.B.1.b, plus la densité de marqueurs est élevée, plus la distance entre marqueurs diminue, et plus le DL entre marqueurs adjacents augmente.

D'une manière générale, il est noté chez les différentes espèces d'élevages une chute rapide du DL avec une augmentation de la distance entre marqueurs. Pérez O'Brien et al. (2014) ont montré en Holstein et Nelore qu'avec respectivement près de 600K et 500K SNP, le DL moyen

passait respectivement de 0.5 et 0.35 à 0.17 et 0.16 puis à 0.08 et 0.06 pour des distances respectives entre marqueurs de 10kb, 100kb et 1Mb. De la même façon, Badke et al. (2012) ont montré, pour les races porcines Duroc et Landrace, qu'avec respectivement 34K et 40K SNP, le DL moyen passait respectivement de 0.36 et 0.27 à 0.19 et 0.15 pour une distance respective entre marqueurs de 100kb et 1Mb. Enfin en volaille, Qanbari et al. (2010) ont mis en évidence pour des races White Leghorn et New Hampshire qu'avec respectivement 20K et 37K SNP, le DL moyen passait respectivement de 0.51 et 0.27 à 0.19 et 0.12 pour des distances respectives de 100kb et 1Mb. Ces résultats sont d'ailleurs cohérents avec les résultats de la thèse présentés dans la Partie IV.

Dans les trois exemples précédents, il est également mis en évidence des différences d'évolution du DL entre races. Dans le cas de la poule, il a même été mis en évidence des différences d'évolution du DL entre chromosomes. D'après Qanbari et al. (2010), pour une distance allant jusqu'à 25kb, le DL moyen était de 0.42 pour les macro-chromosomes (1 à 5), 0.37 pour les chromosomes intermédiaires et 0.32 pour les micro-chromosomes. Il y a donc une persistance plus faible du DL avec une diminution de taille des chromosomes.

C. Principe et fonctionnement global de l'imputation / logiciels

Comme définit en début de partie (mais avec des trous), l'imputation consiste à prédire les lettres manquantes dans des mots ou des phrases et repose sur l'utilisation du déséquilibre de liaison.

Pour imputer les génotypes manquants, il est nécessaire de disposer d'une population de référence pour l'imputation avec des génotypages complets sans données manquantes, et d'une population candidate dont les génotypages sont à imputer. Il est important de disposer de marqueurs en commun entre les génotypages complets de la population de référence et les génotypages à imputer de la population candidate. Sans marqueur en commun, l'imputation est tout simplement impossible. Il est également possible d'inclure l'information familiale pour imputer les données manquantes, ce qui est particulièrement utile dans le cas des espèces d'élevages, les individus étant souvent apparentés et le pedigree connu.

1. Bases statistiques de l'imputation

L'utilisation des modèles de Markov cachés permet d'expliciter les bases statistiques de l'imputation. Un modèle de Markov est un modèle stochastique qui suppose que la distribution des probabilités conditionnelles des états futurs ne dépend que de l'état présent (Dassonneville,

2012). Pour rendre cette explication un peu plus claire avant d'y rajouter de la génétique, sortons de la génétique pour comprendre le fonctionnement de ces modèles.

a) Un exemple simple : jeu des sacs en papier

Supposons que l'on dispose de deux sacs A et B, le sac A contenant 18 jetons rouges R et 2 jetons noirs N, le sac B contenant 2 jetons rouges R et 3 jetons noirs N. Commençons maintenant par piocher un jeton au hasard dans le sac A. Si un jeton rouge est pioché, on continue de piocher dans le sac A. Si un jeton noir est pioché, on passe dans le sac B. À chaque tirage, on note la couleur du jeton tiré puis on le remet dans le sac. Puis le tirage recommence avec le sac en cours, jusqu'à ce que les tirages s'arrêtent (quand on veut).

Les probabilités de tirer dans un sac au tirage *n+1* sachant dans quel sac on tire au tirage *n*, peuvent se calculer facilement et sont résumés dans le tableau 3.

Tableau 3. Probabilité conditionnelle de tirage dans le sac n+1 sachant le tirage dans le sac n.

	Tirage n+1 dans le sac A	*Tirage n+1 dans le sac B*
Tirage n dans le sac A	0.9	0.1
Tirage n dans le sac B	0.4	0.6

En effectuant plusieurs tirages, il est donc possible d'obtenir différentes séquences de couleur de jetons : RRRNNRRRRN, RNRNNRRRRN, RRNRRRRRNN, etc.

Ce type de tirage peut être modélisé par une chaine de Markov (non cachée), chaque sac représentant un état observé, la couleur du jeton représentant une transition, et la proportion de jeton de chaque couleur représentant une probabilité de transition (Figure 16).

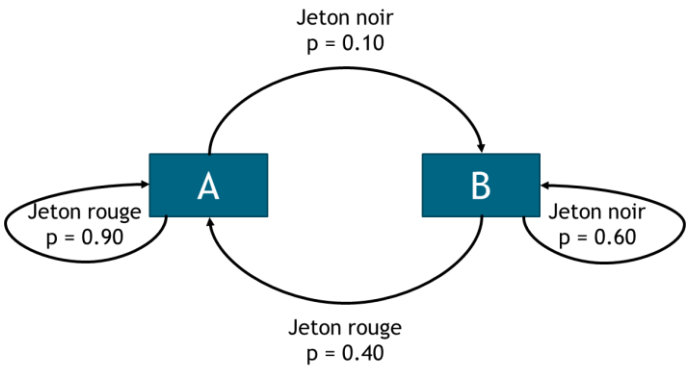


Figure 16. Schématisation d'une chaîne de Markov non cachée appliquée au jeu des sacs de papier.

Supposons que l'on rajoute deux sacs A' et B', le sac A' contenant 4 jetons bleus et 1 jeton vert, le sac B' contenant 1 jeton bleu et 4 jetons verts. Le sac A' est ensuite placé à côté du sac A et le sac B' est placé à côté du sac B.

Commençons maintenant avec le groupe de sac A/A'. Un premier jeton est tiré dans le sac A', sa couleur est notée puis il est remis dans le sac. Un jeton est ensuite tiré dans le sac A sans que sa couleur soit notée, puis il est remis dans le sac. Si un jeton rouge a été tiré, les deux prochains tirages se font dans le groupe de sacs A/A'. Si un jeton noir a été tiré, les deux prochains tirages se font dans le groupe de sacs B/B'. Les tirages sont ensuite renouvelés autant de fois que l'on souhaite. À chaque étape, un jeton est tiré dans un groupe de sac permettant d'obtenir l'information de la couleur bleue ou verte du jeton, mais pas d'information sur la couleur rouge ou noire du jeton.

La succession des tirages permet de générer :

- Une séquence de sortie ou suite d'observations, connue, étant la séquence des couleurs des jetons bleus ou verts des sacs A' et B'
- Une séquence des transitions, inconnue, étant la séquence des couleurs des jetons rouges ou noirs des sacs A et B.

Il est important de noter qu'il est possible d'obtenir des séquences de sorties différentes pour une même séquence de transition.

Ces types de tirages peuvent être modélisés par un modèle de Markov cachés. Le tout premier tirage (dans le sac A' dans l'exemple) correspond à l'état de départ et est associé à une probabilité de départ. Les groupes de sacs A et B représentent les états cachés et les sacs A' et B' les états observés, les tirages dans les sacs A et B (permettant de savoir dans quel groupe de sacs se fait le tirage suivant) représentent les transitions avec les probabilités de transitions associées, et les tirages dans les sacs A' et B' représentent les observations du modèle avec les probabilités d'émissions associées (Figure 17).

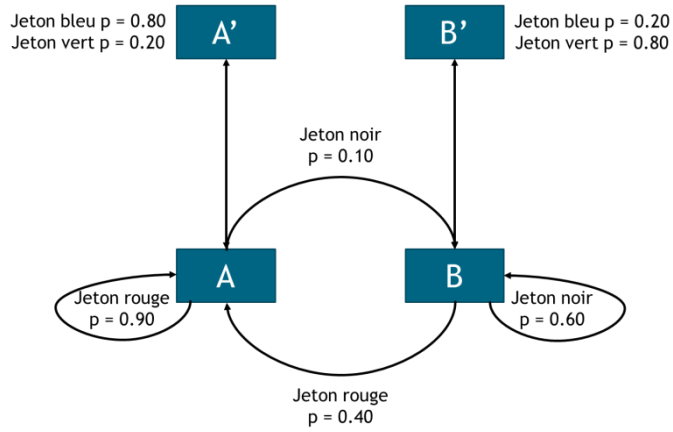


Figure 17. Schématisation d'une chaîne de Markov cachée appliquée au jeu des sacs de papier.

b) *Application à la génétique*

Appliqués à la prédiction des génotypes manquants, les modèles de Markov cachés sont utilisés pour mettre en relation des observations faites sur les génotypes (ou les haplotypes dans le cas d'un pré-phasage), parfois manquants, d'une population candidate, à des haplotypes de références. Suite aux événements de recombinaison et de mutations, les haplotypes des candidats sont modélisés comme une mosaïque (non-observée) des haplotypes de référence (Marchini et Howie, 2010).

En reprenant les termes de l'exemple précédent :

- La probabilité de départ correspond à la probabilité de démarrer la chaine à un haplotype de référence particulier.
- Les états cachés sont les haplotypes de la population de référence et de la population candidate.
- Les états observés sont les génotypes (ou les haplotypes dans le cas d'un pré-phasage), parfois manquants, des individus cibles.
- Les probabilités de transitions correspondent aux probabilités de passer d'un haplotype de référence à un autre lorsqu'un haplotype cible est modélisé comme une mosaïque d'haplotypes de référence. En l'absence de pré-phasage, les probabilités de transitions sont considérées pour des paires d'haplotypes.
- Les probabilités d'émissions correspondent aux probabilités d'observer un génotype / un allèle chez un candidat pour un marqueur, conditionnellement aux états cachés (aux haplotypes de référence constituant la mosaïque).

2. Les différentes méthodes d'imputation

a) *Application des modèles de Markov cachés pour l'imputation*

De nombreuses méthodes utilisent les modèles de Markov cachés pour imputer les génotypes manquants d'une population candidate : FastPHASE (Scheet and Stephens, 2006), IMPUTE2.0 (Howie et al., 2009), Minimac (Howie et al., 2012 ; Fuchsberger et al., 2014), Beagle (Browning et Browning, 2007). Ces logiciels diffèrent en fonction de leurs hypothèses sur les états cachés, les probabilités de transitions et d'émissions, et l'état de départ (Hayes, 2011).

Par exemple, FastPHASE (Scheet and Stephens, 2006) considère que sur des régions courtes, il est possible de créer des clusters d'haplotypes similaires. Il est toutefois nécessaire de spécifier le nombre de clusters d'haplotypes K qui correspondent aux états cachés. Ce logiciel permet de phaser (détermination de l'origine parentale des différents allèles) et d'imputer les données manquantes. Le génotype des individus candidats est modélisé selon un modèle de

Markov caché. Un algorithme espérance-maximisation est ensuite utilisé pour adapter le modèle et les génotypes manquants sont imputés conditionnellement aux paramètres estimés selon un algorithme forward-backward (algorithme permettant de calculer la probabilité d'une séquence de génotypes observés).

L'imputation selon le modèle d'IMPUTE2.0 (Howie et al., 2009) commence par un phasage de la population de référence dont l'ensemble des génotypes est connu, et un phasage de la population candidate dont les haplotypes sont modélisés comme une mosaïque d'haplotypes de référence. Un certain nombre d'itérations est ensuite réalisé selon la méthode de Monte-Carlo par chaînes de Markov. Ceci permet d'imputer les allèles des individus candidats, conditionnellement aux probabilités associées aux différents haplotypes estimés à partir de la population de référence et de la population candidate. L'incertitude des différentes phases est prise en compte en réalisant ces itérations. Après avoir imputé les allèles manquants pour les individus candidats, il est ensuite possible de déduire les génotypes manquants.

Enfin, Beagle (Browning et Browning, 2007) est également basé sur des modèles de Markov cachés. Le programme fonctionne en deux étapes successives. Un arbre des différents allèles aux différents marqueurs est tout d'abord construit à partir d'un set d'haplotypes ré-échantillonné itérativement, puis un poids est associé à chaque branche à l'issue de la dernière itération. Le poids correspond au nombre d'haplotypes par lesquels passe la branche. L'arbre est ensuite condensé en un graphe orienté acyclique (DAG). Cet arbre permet de donner les différentes probabilités de transition d'un état caché à la position d'un SNP à l'état caché du SNP suivant, ainsi que les différentes probabilités d'émissions. Le graphe développé avec le modèle de Beagle présente la propriété d'avoir peu ou beaucoup de « branches » dans les régions avec respectivement un faible ou un fort niveau de DL (Marchini et Howie, 2010). Ce modèle peut donc s'adapter à la diversité haplotypique locale des différents individus et est similaire au programme FastPHASE. La principale différence réside toutefois dans le nombre d'états cachés K qui peut varier pour chaque SNP.

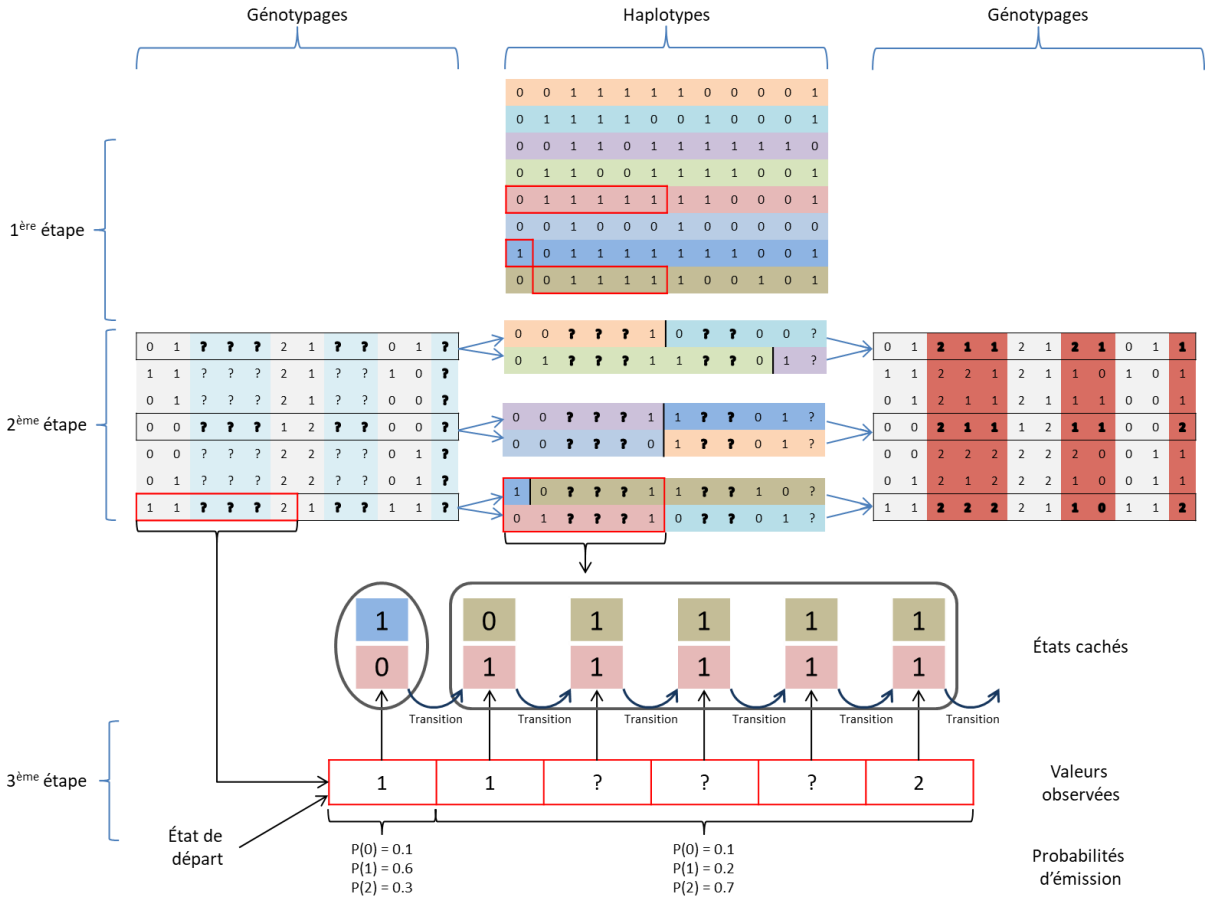


Figure 18. Exemple de phasage et d'imputation selon le principe d'IMPUTE2.0. 1) Phasage de la population de référence et création de la librairie d'haplotypes ; 2) Phasage de la population selon leurs haplotypes (états cachés) correspondants à des fragments d'haplotype de la population référence ; 3) Calcul des probabilités associées à chaque phasage possible en fonction de l'état de départ du génotypage observé, des probabilités d'émission et des probabilités de transition ; 4) Conservation du phasage avec la meilleure probabilité associée et imputation des génotypages manquants en fonction du phasage retenu. D'après Howie et al., 2009.

b) Utilisation de l'information de parenté

Initialement, ces modèles développés en humain n'utilisaient pas l'information généalogique pouvant s'avérer très utile pour phaser et imputer plus rapidement la population candidate. Lorsqu'un individu et plusieurs de ses apparentés sont génotypés, en utilisant les principes de ségrégation mendélienne, il est possible de déterminer rapidement les allèles reçus par l'individu. Supposons qu'un individu et ses deux parents sont génotypés et que l'on cherche à déterminer les allèles à un locus de l'individu. Si les parents sont homozygotes AA et GG, il est alors certain que le descendant est hétérozygote AG. Ces marqueurs homozygotes peuvent

donc servir de « point d'ancrage » pour permettre de phaser puis imputer plus rapidement les génotypes manquants.

Druet et George (2010) ont proposé une méthode permettant de combiner à la fois information populationnelle et information de parenté avec la suite de programme PHASEBOOK. Une mise à jour de Beagle (Browning et Browning, 2009) a permis depuis la prise en compte de l'information familiale pour imputer les données manquantes. Enfin elle est également prise en compte avec les méthodes et programmes développés dans les deux points suivants.

c) Méthode du phasage à grande distance et imputation d'haplotypes longs

Initialement proposée par Kong et al. (2008), cette méthode repose sur le fait que certains individus, même non apparentés, peuvent partager un même ancêtre commun, et donc partager de longs segments chromosomiques (Hayes, 2011). Ceci se traduit par l'identification de longs segments IBD (Identique par descendance) et repose sur l'idée que si des individus n'ont pas de génotype homozygote opposé (par exemple AA et TT) sur une longue suite de loci, alors ils ont en commun au moins un haplotype de grande taille, venant d'un ancêtre commun.

En pratique, à partir d'une population de référence, une librairie d'haplotypes longs est créée. Puis pour chaque individu de la population candidate, les vrais parents ou des parents « substituts » sont utilisés. Les parents « substituts » correspondent à des individus partageant un grand haplotype avec l'individu candidat et identifiés comme n'ayant pas, sur une longue distance, de génotype homozygote opposé (moins de 2%). Pour les marqueurs homozygotes, la phase de l'individu n'est pas compliquée à obtenir. Pour les marqueurs hétérozygotes, la phase de l'individu candidat est obtenue en remontant parmi les phases des parents ou « substituts » des parents. Dès qu'un individu est identifié comme homozygote pour le locus, il est utilisé pour phaser le candidat. Si aucun individu ne présente de marqueurs homozygotes au locus d'intérêt, les parents des parents ou substitut des parents ou substituts des substituts sont alors utilisés, et ainsi de suite jusqu'à trouver un individu avec un marqueur homozygote pour le locus étudié. Pour chaque marqueur non génotypé, leur phase est d'abord imputée à partir du phasage réalisé pour les marqueurs adjacents génotypés. Leur génotype est ensuite imputé sur la base des phases imputées.

Le logiciel AlphaImpute développé par Hickey et al. (2011) permet l'imputation des génotypes manquants à partir de cette approche. Toutefois, elle ne permet pas d'imputer la totalité des marqueurs non génotypés. Une option a été incluse dans le logiciel pour réaliser les imputations en deux temps : imputation des données manquantes à partir de la méthode développée ici, puis imputation des dernières données manquantes à partir des chaines de Markov cachées.

d) Méthode de la fenêtre glissante chevauchante

Similaire à la méthode précédente, cette méthode a été proposée par Sargolzaei et al. (2014) et mise en place dans le programme FImpute. Elle repose sur l'utilisation de l'information de pedigree et exploite les relations de parenté en recherchant des correspondances entre haplotypes de différents individus. Elle suppose également que des individus proches partagent des haplotypes de plus grandes tailles que des individus plus éloignées. Cette recherche de correspondance entre haplotypes se fait sur une grande fenêtre de SNP dont la taille peut varier. Une fois les correspondances trouvées, une nouvelle fenêtre chevauchant partiellement la fenêtre précédente est étudiée. C'est pourquoi on parle de fenêtre glissante chevauchante.

Dans le détail, cette méthode se déroule en six principales étapes :

- Première étape : sur la base du pedigree, repérage de paires d'haplotypes parents-descendants avec une fenêtre recouvrant tout le génome
- Deuxième étape : balayage de chaque chromosome sur une fenêtre de 1000 SNP pour construire une librairie d'haplotypes de référence basée sur les génotypages HD de la population de référence.
- Troisième étape : repérage, pour les individus candidats, d'haplotypes similaires (>99%) à ceux de la librairie d'haplotypes de référence.
- Quatrième étape : Si des haplotypes similaires sont bien retrouvés, les génotypes manquants peuvent être imputés. En cas d'échec, la taille de la fenêtre des SNP est réduite progressivement jusqu'à trouver des haplotypes similaires. La taille de la fenêtre peut être réduite à seulement 2 SNP.
- Cinquième étape : Une fois les haplotypes similaires retrouvés et les génotypages imputés, passage à la fenêtre suivante de 1000 SNP avec un chevauchement de 750 SNP sur la fenêtre précédente. Les étapes 3 et 4 sont ensuite reprises.
- Sixième étape : en cas de génotypes manquants après balayage de l'ensemble des chromosomes, imputation réalisée sur la base des fréquences alléliques de la population de référence.

Cette méthode présente l'avantage d'obtenir de très bons résultats d'imputation dans un temps bien plus faible que les autres programmes évoqués (Sargolzaei et al., 2014). C'est donc cette méthode qui a été principalement utilisée au cours de la thèse. La comparaison entre les logiciels Beagle, AlphaImpute et FImpute est réalisée dans le chapitre II.

D. Les mesures d'efficacité de l'imputation

1. Comment mesurer l'efficacité de l'imputation ?

L'efficacité de l'imputation peut se mesurer selon deux méthodes réelles ou simulées. La première consiste à disposer d'une population de référence génotypée avec une puce HD et d'une population candidate génotypée avec une puce HD et une puce BD. Les génotypages BD de la population candidates sont ensuite imputés à partir des génotypages HD de la population de référence pour remonter aux génotypages HD de la population candidate. Les génotypages HD imputés sont ensuite comparés aux vrais génotypages HD de la population candidate. Les inconvénients de cette méthode sont assez conséquents. En effet, elle nécessite de disposer de génotypages HD et BD pour la population candidate ce qui peut être très coûteux pour les sélectionneurs. Enfin, si l'on souhaite teste plusieurs puces BD, elle nécessite de génotyper autant de fois les candidats qu'il y a de puces.

La deuxième méthode consiste à disposer uniquement des génotypages HD de la population de référence et de la population candidate. Les génotypages BD de la population candidate sont ensuite simulés in silico par « effacement » de certains génotypages de la puce HD. Ceci permet alors de simuler l'utilisation d'une puce BD. Cette opération, qui ne coûte rien, peut être répétée autant de fois que l'on souhaite afin de tester diverses puces BD. L'imputation de la population candidate est réalisée à partir des génotypages BD simulés et des génotypages HD de la population de référence. Les génotypages HD imputés sont comparés aux vrais génotypages HD de la population candidate. Cette méthode a été utilisée tout au long de la thèse pour mesurer l'efficacité des imputations. Par ailleurs, c'est également la méthode la plus utilisée dans la littérature.

2. Critères de mesure de l'efficacité de l'imputation

a) Taux d'erreur génotypique

C'est le critère de mesure d'efficacité de l'imputation le plus simple à mesurer. Cette mesure indique la proportion de marqueurs mal imputés. Pour calculer ce taux d'erreur génotypique, il suffit de comparer, marqueurs par marqueurs, les génotypages imputés avec les vrais génotypages HD. Si l'on observe une différence sur l'un des deux allèles du marqueur, l'imputation est considérée comme fausse, et l'erreur est comptabilisée. Cette opération est répétée pour l'ensemble des marqueurs. Le nombre d'erreur est ensuite sommé puis divisé par le produit du nombre total de marqueurs à imputer et du nombre de candidats à la sélection. En multipliant ce résultat par 100, on obtient alors le taux d'erreur génotypique. Ce taux d'erreur a déjà été utilisé par Hayes et al. (2011) ou encore Ventura et al. (2014).

Le taux de concordance peut également être utilisé et correspond à la proportion de marqueurs bien imputés. Ce taux de concordance se calcule facilement à partir du taux d'erreur :

$$taux\ de\ concordance = 100 - taux\ d'erreur\ génotypique$$

Le taux de concordance a été utilisé dans les publications de Weigel et al. (2010a, 2010b).

b) Taux d'erreur allélique

Le taux d'erreur génotypique peut être affiné en calculant un taux d'erreur allélique. En effet, il est possible, pour un marqueur à imputer, qu'un des deux allèles soit correctement imputé, et l'autre non. Dans ce cas, il peut être intéressant de considérer un taux d'erreur allélique plutôt que génotypique pour mesurer l'efficacité de l'imputation.

Pour calculer le taux d'erreur allélique, il faut comparer, marqueur par marqueur, les génotypages imputés avec les vrais génotypages HD. Si l'on observe une différence sur l'un des deux allèles du marqueur, l'imputation est considérée comme à moitié bonne (ou à moitié fausse), et une demie-erreur est comptabilisée. Si les deux allèles sont différents du vrai génotypage, l'imputation est considérée comme fausse et une erreur est comptabilisée.

Cette opération est répétée pour l'ensemble des marqueurs. Le nombre d'erreur est ensuite sommé puis divisé par le produit du nombre total de marqueurs à imputer et du nombre de candidats à la sélection. En multipliant ce résultat par 100, on obtient alors le taux d'erreur allélique.

Weigel et al. (2010a, 2010b) considèrent le taux d'erreur allélique comme une façon de présenter de meilleurs résultats qu'avec le taux d'erreur génotypique. En effet, le taux d'erreur allélique correspond environ à la moitié du taux d'erreur génotypique, les erreurs d'imputation n'impactant principalement qu'un seul des deux allèles des marqueurs. En revanche, le taux d'erreur allélique est utilisé par Druet et al. (2010), Zhang et Druet (2010), Dassonneville et al. (2011), Dassonneville et al. (2012). Druet et al. (2010) et Dassonneville (2012) justifient l'utilisation du taux d'erreur allélique en expliquant que les génotypages imputés sont utilisés pour les évaluations génomiques qui reposent sur un modèle additif. Pour cette raison, dans le cas d'un marqueur bien imputé pour un seul des deux allèles, les résultats des évaluations génomiques seront alors à moitié vrais.

c) Corrélations

Hickey et al. (2012) et Calus et al. (2014) suggèrent de calculer des corrélations pour estimer la qualité des imputations. En effet, les taux d'erreur génotypique et allélique dépendent des fréquences alléliques, à l'inverse des corrélations. Plusieurs études (Hickey et al., 2012 ; Hozé

et al., 2013, Schrooten et al., 2014) ont montré que les taux d'erreurs augmentaient avec la MAF suggérant que l'imputation de marqueurs à faible MAF était plus facile. À l'inverse, d'autres études (Hickey et al., 2012 ; Carvalheiro et al., 2014 ; Heidaritabar et al., 2015) ont montré que les corrélations augmentaient avec la MAF, suggérant que l'imputation de marqueurs à faible MAF était plus complexe. Ces résultats, paradoxaux, résultent du fait que les taux d'erreur dépendent de la MAF et donnent un poids plus fort à une erreur d'imputation sur un marqueur à forte MAF. Les corrélations permettent donc de mieux comparer les résultats d'imputation pour tous les marqueurs en fonction de leur MAF.

Les corrélations de Pearson sont calculées marqueur par marqueur pour tous les candidats, puis la corrélation de Pearson moyenne est estimée sur le nombre total de marqueurs.

Ces corrélations ont été utilisées dans de nombreuses publications (Dassonneville et al., 2012 ; Carvalheiro et al., 2014 ; Bouquet et al., 2015 ; Wolc et al., 2016, Herry et al., 2018).

d) Impact sur les évaluations génomiques

Il est également possible de vérifier la qualité des imputations en regardant l'impact des imputations sur les évaluations génomiques. Ce critère n'est pas considéré dans la littérature comme un critère de mesure d'efficacité de l'imputation en tant que tel. Toutefois, pour un candidat à la sélection, une erreur d'imputation peut avoir un impact sur sa valeur génomique calculée à partir de la population de référence et de son génotype imputé. Il est également possible que certaines erreurs d'imputation soient plus importantes que d'autres si elles correspondent à des erreurs de génotypage sur des marqueurs à effets forts sur certains caractères d'intérêt. C'est pourquoi, nous avons considéré ici l'impact sur les évaluations génomiques comme un critère de mesure d'efficacité de l'imputation.

L'impact des erreurs d'imputation sur les évaluations génomiques peut être vérifié en comparant les résultats des évaluations génomiques des candidats à la sélection avec leur vrais génotypes HD ou leur génotypes HD imputés (Figure 19). Dans les deux cas, la population de référence est la même et est constituée d'individus avec des génotypes HD et des performances. Il est ensuite possible de calculer des corrélations entre les valeurs génomiques des candidats avec leur vrais génotypes HD ou leur génotypes HD imputés. Si l'on s'intéresse au classement des candidats, on calcule des corrélations de Spearman. Si l'on s'intéresse à la relation linéaire entre les valeurs génomiques, on calcule des corrélations de Pearson.

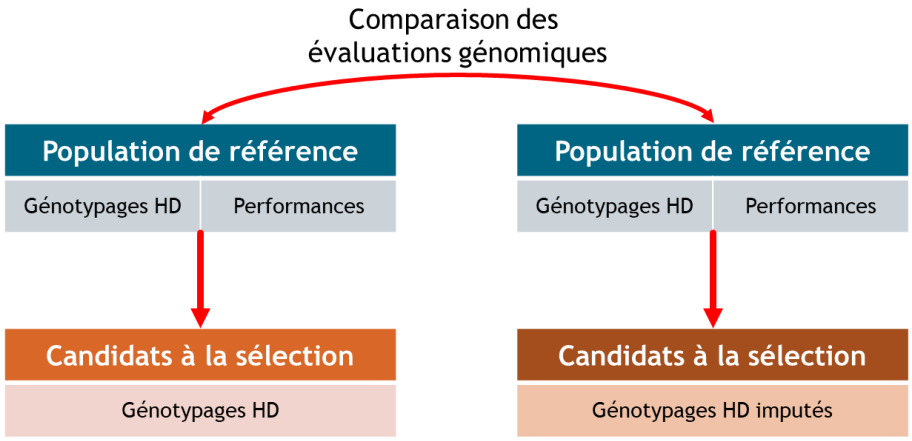


Figure 19. Étude de l'impact des erreurs d'imputation par comparaison des évaluations génomiques des candidats à la sélection avec leur vrais génotypes HD ou leur génotypes HD imputés.

Il est également possible de vérifier l'impact des erreurs d'imputation sur la précision des évaluations génomiques. La précision des évaluations se calcule comme la corrélation entre les vraies valeurs génétiques TBV (True Breeding Values) et les valeurs génomiques estimées GEBV (Genomic Estimated Breeding Values) (Meuwissen et al., 2001 ; Legarra et al., 2008). Toutefois, il n'est jamais possible d'avoir accès aux TBV avec des données réelles. En revanche, il est possible de les approcher en calculant pour les mâles des pseudo-phénotypes comme les Daughter Yield Deviation (DYD) correspondant aux moyennes des performances des filles, corrigées pour l'ensemble des facteurs non génétique du modèle d'évaluation génétique et pour la moitié de la valeur génétique de leur mère (Van Raden, 2008 ; Robert-Granié et al., 2011). Dans le cas d'évaluation de femelles, on peut calculer des Yield Deviation (YD) qui correspondent aux moyennes des performances des filles corrigées pour l'ensemble des facteurs non génétique du modèle. Enfin, Picard Druet et al. (2019a) ont montré en pondeuses que les corrélations entre DYD et des GEBV estimés sur descendance étaient très fortes pour les coqs reproducteurs. Il est donc possible d'estimer la précision des évaluations en calculant des corrélations entre des GEBV estimées sur descendance, servant de référence, et des GEBV estimées avec moins d'information (sur ascendance ou collatéraux). Dans le cas illustré avec la figure 20, à partir de toutes les informations disponibles (performances et génotypes HD), une évaluation sur descendance des candidats à la sélection est réalisée. Une évaluation sur ascendance des candidats à la sélection avec les génotypes HD imputés est ensuite réalisée. Puis les résultats des deux évaluations sont comparés avec des corrélations de Pearson. Pour évaluer la précision des évaluations, on compare donc deux évaluations avec des quantités de données différentes.

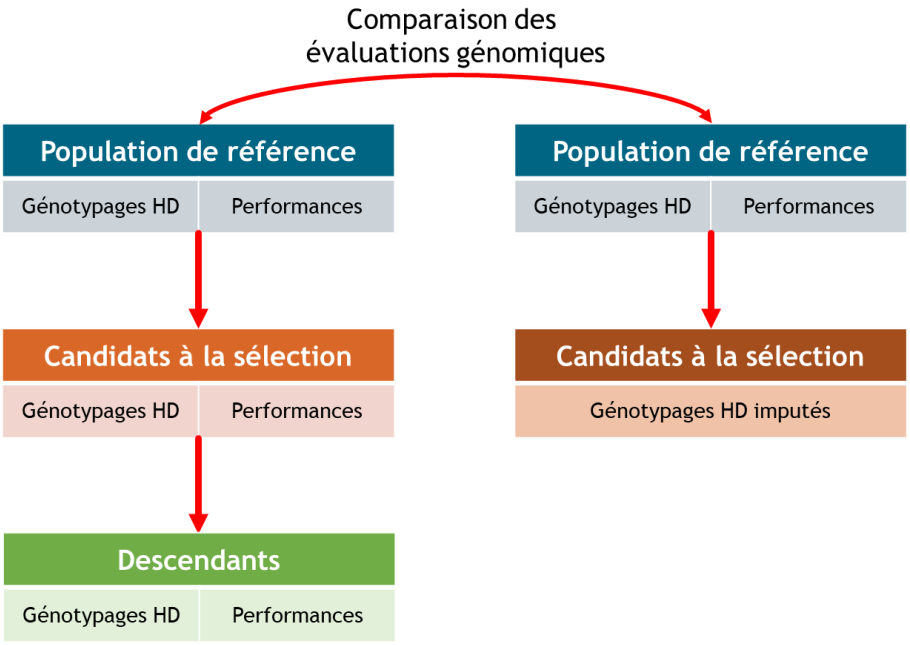


Figure 20. Étude de la précision des évaluations génomiques par comparaison de l'évaluation génomique sur descendance des candidats à la sélection avec leur vrais génotypes HD et de l'évaluation génomique sur ascendance des candidats à la sélection avec leur génotypes HD imputés.

E. Les facteurs influençant la qualité de l'imputations

La qualité des imputations peut être influencée par de nombreux facteurs que l'on peut classer en deux catégories : les facteurs liés au design des puces BD et les facteurs liés à la constitution de la population de référence. Les logiciels utilisés pour les imputations ont également un impact sur la qualité des imputations. Ce point particulier sera développé dans la partie II.

1. Facteurs liés aux puces à SNP

a) Densité de SNP

Un des facteurs influençant la qualité des imputations est la densité de marqueurs sur les puces BD. De façon assez intuitive, plus le nombre de marqueurs sur les puces BD est important, plus l'imputation est bonne. En effet, une augmentation du nombre de marqueurs sur les puces BD permet de disposer de plus de marqueurs pour identifier dans la population de référence le bon haplotype de référence et ainsi bien imputer les marqueurs manquants. À l'inverse, plus la densité de marqueurs diminue, plus le nombre de génotypages disponibles diminue, plus la longueur physique des haplotypes à retrouver dans la population de référence augmente. Avec les événements de recombinaisons ou de mutations et la longueur des haplotypes, il est donc possible que l'haplotype de référence ne corresponde pas à celui du candidat. Le risque

d'identifier par hasard un mauvais haplotype en commun entre les populations de référence et candidate augmente. De très nombreuses publications ont étudié ce facteur : Hayes, 2011 ; Hickey et al., 2012 ; Hozé et al., 2013 ; Carvalheiro et al., 2014 ; Aliloo et al., 2018.

b) *Fréquences alléliques des marqueurs*

La fréquence allélique des marqueurs à imputer est un des facteurs qui influence la qualité des imputations. Dans le cas d'un marqueur avec une faible MAF, la probabilité de retrouver l'haplotype correspondant dans la librairie d'haplotype de référence sera plus faible que pour un marqueur avec une MAF plus élevée. Plus la MAF est faible, plus le nombre d'individus de la population de référence ayant l'haplotype correspondant est faible. Si l'haplotype de référence n'est pas retrouvé, l'imputation peut alors être réalisée sur la base des fréquences alléliques de la population de référence, ce qui entraîne des erreurs d'imputation pour les marqueurs à faible MAF. Il a été montré dans la section II.D.2 qu'à l'inverse des taux d'erreurs, les corrélations ne dépendaient pas des fréquences alléliques. L'imputation de marqueurs à faible MAF entraîne donc des corrélations plus faibles sur les marqueurs à faible MAF que sur les marqueurs à MAF plus élevée (Hayes et al., 2012 ; Carvalheiro et al., 2014 ; Aliloo et al., 2018).

Par ailleurs, ces marqueurs à faible MAF sont plus susceptibles d'avoir un rôle important dans le déterminisme génétique de certains caractères et peuvent avoir un effet plus fort que certains marqueurs à MAF plus élevée (Hickey et al., 2012 ; Heidaritabar et al., 2014). L'imputation des marqueurs à faible MAF est donc très importante dans une optique de sélection. Pour pallier ce problème, il est important de disposer d'une population de référence suffisamment grande pour capter toute la diversité haplotypique des individus étudiés et ainsi réaliser de bonnes imputations. Ce point concernant la taille de la population de référence est discuté dans la partie suivante.

c) *Influence des chromosomes*

Les chromosomes et plus particulièrement leur taille peuvent également influencer la qualité des imputations. En effet, en bovins, la taille des chromosomes est comprise entre 159Mb et 51Mb pour les chromosomes 1 et 29 respectivement. Plusieurs publications rapportent des imputations moins bonnes avec une diminution de la taille des chromosomes (Sun et al., 2012 ; Picolli et al., 2014). De la même façon, en porcin, la taille des chromosomes est comprise entre 274Mb et 56Mb pour les chromosomes 1 et 18 respectivement. Grossi et al. (2018) ont obtenu de légères variations dans la qualité des imputations, la meilleure imputation étant pour le

chromosome 1 et la moins bonne pour le chromosome 18. La diminution de la qualité des imputations avec une diminution de taille des chromosomes peut s'expliquer par des imputations moins bonnes dans les parties terminales des chromosomes (Cleveland et Hickey, 2013 ; Wellmann et al., 2013 ; Picolli et al., 2014). Plus la taille du chromosome diminue, plus les erreurs d'imputation dans les parties terminales des chromosomes ont un poids fort.

La poule présente la particularité d'avoir des chromosomes de tailles très différentes (Figure 2). La persistance du DL diminue également avec la taille des chromosomes. La longueur des haplotypes diminue donc avec la taille des chromosomes. Combinée à des imputations moins bonnes dans les parties terminales des chromosomes, ceci peut donc également entraîner une diminution de la qualité des imputations avec une diminution de la taille des chromosomes. Vereijken et al. (2010) ont montré chez la poule une diminution de la qualité des imputations avec une diminution de la taille des chromosomes. En revanche, Heidaritabar et al. (2015) n'ont pas montré de différence entre chromosomes.

Ces résultats sont toutefois fortement dépendant de la méthodologie utilisée pour développer les puces BD. Ceci est détaillé dans le chapitre II avec l'article n°1.

d) Méthodologie utilisée pour développer les puces BD

Trois principales méthodologies peuvent être utilisées pour sélectionner le sous-ensemble de marqueurs d'une puce HD à inclure sur une puce BD. Ce sont des méthodes de sélection des SNP propres à chaque population ou lignée étudiée et dépendent des fréquences alléliques de la population ou de la lignée étudiée. La première méthode consiste à sélectionner des SNP à intervalles réguliers le long de chaque chromosome en fonction ou non de leur MAF (Habier et al., 2009 ; Weigel et al., 2009 ; Zhang et al., 2011 ; Cleveland et Hickey, 2013 ; Wang et al., 2013 ; Herry et al., 2018). Cette méthode est utilisée lorsque les nombreux caractères d'intérêt sont sous contrôle d'un grand nombre de QTL à effets faibles. La deuxième méthode consiste à sélectionner des SNP en fonction de leurs effets sur différents caractères d'intérêts (Weigel et al., 2009, Zhang et al., 2011). Cette méthode est donc intéressante en cas de QTL à effets forts sur certains caractères d'intérêts. Enfin, la troisième méthode est de sélectionner des SNP en fonction du DL entre marqueurs (Gualdrón Duarte et al., 2013 ; Herry et al., 2018). Cette méthode a notamment été utilisée dans le premier article pour utiliser la différence de persistance du DL entre type de chromosomes.

2. Facteurs liés aux populations utilisées

a) Taille de la population de référence

Un des principaux facteurs influençant la qualité des imputations est la taille de la population de référence. Plus la taille de la population de référence est grande, plus la librairie d'haplotypes de référence est grande, et plus la probabilité pour un candidat de retrouver le bon fragment d'haplotype dans la librairie de référence est grande (Dassonneville et al., 2011 ; Hozé et al., 2013). Si aucun haplotype correspondant ne peut être retrouvé dans la librairie de référence, l'imputation peut toutefois se faire sur la base des fréquences alléliques de la population de référence mais les risques d'erreurs d'imputation sont plus fréquents. Il existe aussi un seuil au-delà duquel l'inclusion d'individus supplémentaires dans la population de référence ne permet plus d'augmenter la précision. Ceci s'explique par une diversité haplotypique déjà entièrement capté par les individus présents dans la population de référence. Toutefois, cela est fortement dépendant de la diversité génétique de la population et de la taille efficace de la population d'étude. Plus la diversité génétique est grande, plus il est important de disposer d'une grande population de référence. Ventura et al. (2014) ont montré en bovin de race Angus que l'imputation de 146 individus devenait stable à partir d'une population de référence constituée de plus de 1000 individus.

Par ailleurs, l'augmentation de taille de la population de référence implique une augmentation du temps de calcul nécessaire pour réaliser les imputations… Inutile donc de cumuler indéfiniment des individus dans la population de référence.

Enfin, lorsque la taille de la population de référence est suffisamment grande et que le nombre de marqueurs de la population candidate est faible, il est plus intéressant d'essayer d'augmenter le nombre de marqueurs en commun pour obtenir de meilleurs imputations (Zhang et Druet, 2011).

b) Relations de parenté entre population de référence et population candidate

Les relations de parenté entre population de référence et population candidate sont également à prendre en compte pour réaliser de bonnes imputations. En effet, plus les relations de parenté entre les deux populations sont fortes, plus les individus ont en commun des fragments d'haplotypes de grandes tailles. À l'inverse, en diminuant les relations de parenté entre les deux populations, les individus ont en commun des fragments d'haplotypes de plus petite taille, ce qui ne facilite pas les imputations. De meilleures imputations sont ainsi obtenues en imputant les candidats à partir de leur parents directs, génotypés en HD, qu'à partir de leur grands-parents

ou ascendants plus anciens. Les recombinaisons au fur et à mesure des générations diminuent la taille des haplotypes en commun entre ascendants et candidats.

Druet et al. (2010) et Zhang et Druet (2011) ont ainsi pu calculer en bovin un score de parenté correspondant à la proportion du génome hérité des individus de référence. Ils ont montré que plus ce score était élevé, plus les imputations étaient bonnes. Hozé et al. (2013) montrent que la qualité des imputations diminue avec une diminution des relations de parenté pour différentes races bovines, mais que le faible apparentement entre population de référence et population candidate peut être compensé par une plus grande taille de population de référence. De même, Bouquet et al. (2015) ont montré en porcin que la qualité des imputations diminuait avec une diminution de l'apparentement entre population de référence et population candidate. Cette diminution était d'autant plus forte que la densité de SNP de la population candidate était faible.

c) Inclusion des mères dans la population de référence

Un autre facteur influençant la qualité des imputations est la présence ou non des mères des candidats dans la population de référence. Cet ajout dans la population de référence permet d'obtenir une amélioration des imputations. En effet, le génotype des candidats correspond à une combinaison des deux haplotypes parentaux. En ayant dans la population de référence les deux parents des candidats à la sélection et en prenant en compte le pedigree, les haplotypes paternels et maternels peuvent être rapidement retrouvés et ainsi permettre de très bonnes imputations. Toutefois, l'ajout des mères dans la population de référence entraîne une augmentation du nombre d'individus dans la population de référence ayant un apparentement fort avec la population candidate. Il est donc difficile de distinguer clairement quelle part d'amélioration de la qualité des imputations est due à l'augmentation de taille de la population de référence ou à l'ajout d'individus ayant un fort apparentement avec les individus candidats. Enfin, l'inclusion des mères dans la population de référence génère un coût de génotypage supplémentaire.

III. Optimisation des génotypages à l'échelle des schémas de sélection

A. Optimisation des génotypages des candidats à la sélection par l'utilisation des puces à SNP BD

Comme expliqué précédemment, le coût de la puce à SNP HD est trop élevé pour qu'un sélectionneur puisse génotyper avec une telle puce l'ensemble des candidats à la sélection. La solution est donc d'utiliser des puces à SNP basse densité, qui coûtent moins cher, pour génotyper l'ensemble des candidats à la sélection. En s'appuyant ensuite sur une population de référence génotypée avec la puce HD, il est possible d'imputer les génotypes manquants des candidats à la sélection pour remonter à une haute densité pour les candidats.

Ainsi de nombreuses puces BD ont été développées pour les différentes espèces d'élevages. Ces puces correspondent à des puces commerciales ou bien à des puces privées développées en interne par les entreprises de sélection. Par rapport aux puces privées, les puces commerciales présentent l'intérêt de représenter des volumes beaucoup plus importants ce qui permet de baisser les coûts du génotypage.

1. Conception des puces à SNP basse densité pour la sélection

Excepté pour les volailles, les puces HD utilisées en routine pour la sélection des espèces d'élevages correspondent à des puces moyenne densité de 50K à 70K SNPs. Des puces HD de plus de 100K SNP existent également mais ne sont pas utilisés en routine pour la sélection. Enfin, les puces BD correspondent à des puces de moins de 10K SNP. À ce jour, il n'existe que très peu de puces BD commerciales. Les puces BD utilisées par les entreprises correspondent très souvent à des puces à façon développées par et pour les entreprises.

Chez les bovins, la puce utilisée en routine pour la sélection correspond à la première puce MD BovineSNP50 Genotyping BeadChip (Illumina Inc, 2011b) de 50K évoquée dans la section I.A.5 après une mise à jour des SNP initiaux de la puce. Une première puce BD GoldenGate® Bovine3K Genotyping BeadChip (Illumina Inc, 2011c) issue d'un consortium international a été commercialisée par Illumina à partir de janvier 2010 puis testée pour la sélection génomique à partir de Septembre 2010 (Wiggans et al., 2012). Cette puce a été développée à partir des SNP présents sur la puce MD de 50K, présentant un espacement équidistant entre marqueurs et ayant une MAF élevée. Toutefois, la technologie utilisée pour génotyper les marqueurs de la puce entraînait des qualités de génotypage et de call rate plus faibles (Wiggans et al., 2013). Peu de races ont également été utilisées pour développer la puce (Jersiaise, Brown Suisse et Holstein)

rendant de fait la puce inadéquate pour un grand nombre de races. En Septembre 2011, une nouvelle puce basse densité BovineLD Genotyping BeadChip de 7K SNP a été conçue (Illumina Inc, 2011a ; Boichard et al., 2012a) dans le cadre d'un autre consortium international pour le même coût que la précédente puce et est toujours actuellement utilisée en routine (Wiggans et al., 2012). Cette puce a également été développée à partir des marqueurs sur la puce MD de 50K. Cinq critères ont été pris en compte : 1) SNP ayant une MAF élevée pour les différentes races testées, 2) espacement uniforme entre SNP avec une densification aux extrémités des chromosomes, 3) SNP permettant de contrôler le sexe des individus et de réaliser des assignations de parenté, 4) SNP permettant une bonne qualité de génotypage et peu de problèmes de compatibilités, 5) environ 2000 SNP en commun avec la puce 3K afin de permettre une compatibilité avec les précédents génotypages et les différentes études réalisées avec cette puce. La puce a depuis évolué en puce EuroG10K en gardant les 7K SNP précédents et en y ajoutant des panels de SNP privées pouvant évoluer en cours du temps. Il existe également une puce GeneSeek Genomic Profiler (GGP) Bovine 50K développée par Neogen en 2012 et composée des SNP les plus informatifs de la puce BovineSNP50 et optimisant le nombre de SNP en communs avec les différentes puces commerciales disponibles. Cette puce est toutefois moins intéressante pour les espèces *Bos indicus*. Une puce GGP *indicus* de 35K SNP a ainsi été développée à partir des SNP de la puce HD 777K en 2018 (Ferraz et al., 2018 ; Neogen, 2018). Enfin, différentes puces BD simulées à partir des puces HD ou MD existantes ont été testés dans de nombreuses études (Weigel et al., 2010a ; Zhang et Druet, 2010 ; Chen et al., 2014 ; Aliloo et al., 2018).

En porcin, il n'y a actuellement pas de puce BD commerciale disponible. En revanche, de nombreuses études ont développé in-silico des puces personnalisées pour les populations étudiées, avec des densités allant de 300 à environ 10K SNP (Gualdrón Duarte et al., 2013 ; Bouquet et al., 2015 ; Carillier-Jacquin et al., 2018 ; Grossi et al., 2018). Ces puces ont toutes été développées à partir de la puce MD Illumina Porcine60K BeadChip de 60K SNP (Ramos et al., 2009 ; Illumina Inc, 2009), en choisissant des SNP équidistants avec ou sans densification dans les extrémités des chromosomes. Carillier-Jacquin et al. (2018) ont tenu compte des SNP ayant des effets forts sur certains caractères d'intérêt. Pour constituer leur puce BD, Gualdrón Duarte et al. (2013) ont également utilisé le DL entre SNP afin de sélectionner le sous ensemble de marqueurs de la puce MD.

En ovins et en caprins, il n'y a également pas de puces BD commerciales disponibles. Des puces BD ont été développées in-silico à partir des puces MD. Bolormaa et al. (2015) ont développé une puce ovine BD de 12K SNP avec la même méthode que Boichard et al. (2012a). Raoul et al. (2017) ont développé des puces ovines de moins de 1K SNP équidistants.

Enfin, en volaille, une puce commerciale de 600K est disponible (Kranis et al., 2013). En revanche, hormis les puces développées en interne par les entreprises Cobb et Hendrix en 2008 (60K SNP) et Wesjohann en 2009 (42K SNP), il n'y avait jusqu'à aujourd'hui aucune puce commerciale de plus basse densité disponible. Liu et al. (2019) ont remédié à ce problème en développant une puce Affymetrix de 55K dont 24K SNP sont en communs avec les SNP de la puce HD de 600K SNP. Toutefois en parallèle, les entreprises comme Novogen n'ont pas attendu cette puce et ont développé en interne des puces BD ou MD en sélectionnant des sous-ensembles de SNP de la puce HD. Ces travaux sont détaillés dans les articles 1 et 2. À titre indicatif, une puce 55K différente de la puce Affymetrix est aujourd'hui utilisée en routine par Novogen pour génotyper l'ensemble de ses candidats à la sélection. La conception de cette puce est développée dans le chapitre IV.

La liste des différentes puces commerciales disponibles pour les différentes espèces d'élevages terrestres est présentée dans le tableau 4.

Tableau 4. Liste des différentes puces commerciales existantes pour les différentes espèces d'élevages.

Espèce	Puce HD (>100K SNP)	Puce MD (<100K SNP)	Puce BD (<12K SNP)
Bovins	BovineHD Genotyping BeadChip : 777K SNP Illumina Inc, 2012a	BovineSNP50 Genotyping BeadChip : 50K SNP Matukumalli et al., 2009 ; Illumina Inc, 2011b GeneSeek Genomic Profiler (GGP) Bovine : 50K SNP Neogen Genomics, 2012 GeneSeek Genomic Profiler (GGP) indicus : 35K SNP Ferraz et al., 2018 ; Neogen Genomics, 2018	GoldenGate® Bovine3K Genotyping BeadChip : 3K SNP Illumina Inc, 2011c BovineLD Genotyping BeadChip : 7K SNP Illumina Inc, 2011a ; Boichard et al., 2012
Porcins	Axiom® Porcine Genotyping Array : 660K SNP Affymetrix Inc, 2015a	Illumina Porcine60K BeadChip : 60K SNP Ramos et al., 2009 ; Illumina Inc, 2009 GeneSeek Genomic Profiler (GGP) Porcine : 52K SNP Neogen Genomics	
Poule	Affymetrix Axiom Chicken Array Kranis et al., 2013	Affymetrix Axiom Chicken Genotyping Array : 55K SNP Liu et al., 2019	
Equin	Axiom® Equine Genotyping Array : 670K SNP Schaefer et al., 2017 ; Affymetrix Inc, 2017	EquineSNP50 Genotyping BeadChip : 54K SNP McCue et al., 2012 ; Illumina Inc, 2012b	
Ovins	Illumina OvineHD BeadChip : 600K SNP International Sheep Genomics Consortium (non publié)	OvineSNP50 Genotyping BeadChip : 54K SNP Illumina Inc, 2008	
Caprins		Illumina GoatSNP50 BeadChip : 52K SNP Tosser-Klopp et al., 2014	
Canard	Puce SNP 600K Thermo Fisher Thébault et al., 2019		

2. Utilisation des puces à SNP basse densité dans les schémas de sélection

a) Des résultats intéressants pour les puces BD commerciales

En bovins, Boichard et al. (2012a) ont testé pour différentes races la puce BD de 7K SNPs en imputant des populations candidates à partir de populations de référence génotypées avec la puce MD de 50K SNP. Ils montrent ainsi que des taux de concordance de 98.1% et 98.5% sont obtenus pour les animaux français de race Holstein et Montbéliarde respectivement. En revanche, pour les animaux australiens de races Angus ou Jersey, les taux de concordance sont

respectivement de 92.3% et 94.9%. Ces taux de concordances plus faibles peuvent toutefois s'expliquer par une taille de population de référence faible (respectivement 200 et 454 animaux) par rapport aux nombre d'individus à imputer (respectivement 82 et 86 animaux). Comparativement, 3505 et 1170 individus ont permis d'imputer respectivement 966 Holstein et 222 Montbéliarde françaises. Chen et al. (2014) ont montré que l'imputation de 3232 taureaux laitiers Holstein génotypés avec la puce BD de 7K SNP à partir de 7077 taureaux génotypés avec la puce MD de 50K permettait d'obtenir un taux de concordance de 98.4%. Les analyses sont allées plus loin en étudiant l'impact de la puce BD et des imputations sur les évaluations génomiques des candidats imputés. Les évaluations génomiques ont été réalisées selon une méthode GBLUP. La précision des évaluations génomiques, calculée par des corrélations de Pearson entre les DGV (Direct Genomic Value) à partir des vrais génotypages 50K et les performances des taureaux, était de 0.61 pour le rendement laitier, et de 0.62 pour le taux de cellules. Avec des génotypages 50K imputés à partir de la puce BD, les corrélations étaient également de 0.61 pour le rendement laitier, et de 0.62 pour le taux de cellules, indiquant ainsi un impact nul des imputations sur la précision des évaluations génomiques.

b) Exemple de résultats de puces BD non commerciales pour la sélection génomique

Ces résultats très intéressants de la puce commerciale BD bovine sont également retrouvés pour d'autres espèces dans de nombreuses autres études se basant sur des puces BD non commerciales développées in-silico. Par exemple, en porcin, Grossi et al (2018) ont développé des puces BD in-silico par sélection de marqueurs équidistants à partir de la puce MD 60K. Pour des cochons Landrace et Yorkshire des taux de concordance supérieurs à 94% ont été obtenus entre les génotypages imputés à partir de plus de 1000 SNP et les génotypages MD. Ceci a également permis d'obtenir des précisions d'évaluations aussi élevées avec les génotypages imputés qu'avec les génotypages MD pour les différents caractères étudiés.

Toutefois, en diminuant de façon trop importante la densité de SNP sur une puce, le risque de diminuer la qualité des imputations et ainsi d'impacter la précision des évaluations génomiques est plus grand. Avec des imputations erronées, l'application des équations de prédictions des valeurs génomiques aux candidats à la sélection entraîne des résultats faux à cause des erreurs de génotypages. En supposant qu'un SNP avec un effet fort sur un caractère d'intérêt est mal imputé pour un individu, une valeur génétique plus faible sera alors estimée pour le candidat à cause de l'erreur d'imputation.

Chen et al. (2014) ont ainsi montré qu'un taux de concordance de 76.6% était obtenu pour une imputation à partir de seulement 384 SNP équidistants. La précision des évaluations

génomiques diminuait alors avec des corrélations de Pearson entre les DGV à partir des génotypages 50K imputés et les performances des taureaux de seulement 0.49 et 0.53 pour le rendement laitier et le taux de cellules respectivement.

De la même façon, Raoul et al. (2017) ont montré en ovins des taux de concordance de 87.3% et 96.1% entre les génotypages MD de 50K et des génotypages imputés à partir de 250 et 1000 SNP respectivement. La précision des évaluations des candidats mâles à la sélection, estimée comme la corrélation de Pearson entre les TBV et les GEBV avec les génotypages MD était de 0.71. Avec les génotypages imputés à partir de 250 et 1000 SNP, les précisions étaient de 0.38 et 0.63 respectivement. Ceci illustre bien la diminution conséquente de précision avec des densités de SNP trop faibles et donc avec de moins bonnes imputations.

Enfin, les résultats des puces BD sont dépendants de la méthodologie utilisée pour réaliser les évaluations, comme expliqué à la partie I.B.3.

Toutes ces questions de design de puce BD, qualité d'imputation et impact sur les évaluations génomiques ont été développées dans les chapitres II et III avec les articles 1 et 2.

B. Optimisation des génotypages des candidats à la sélection par l'utilisation du séquençage à basse profondeur

En parallèle aux méthodes utilisant les puces à SNP, les techniques de séquençage nouvelle génération (NGS) permettent de détecter et de génotyper simultanément un très grand nombre de marqueurs. Ces techniques ont été utilisées de façon croissante chez les espèces d'élevages. En travaillant sur la séquence des candidats à la sélection, il est possible de distinguer deux alternatives aux puces BD : séquencer l'ensemble du génome des candidats à de faibles profondeurs ou bien séquencer seulement une partie du génome des candidats à de plus hautes profondeurs. La population de référence étant génotypée en HD, il est important d'avoir des SNP en communs entre les SNP de la puce HD et les SNP détectés par les méthodes alternatives aux puces BD.

1. Utilisation du séquençage NGS basse profondeur comme alternatives aux puces BD

a) Principe du séquençage NGS

(1) Extraction et préparation de l'ADN

Pour pouvoir réaliser un séquençage NGS du génome de plusieurs individus, il faut tout d'abord extraire l'ADN des individus. L'ADN est ensuite fragmenté par méthode physique (coupure acoustique ou sonication) ou enzymatique (Head et al., 2014). Une ligation des fragments à

deux différents adaptateurs est ensuite réalisée au niveau des extrémités de chaque fragment : un adaptateur de type A aux extrémités 3' et un adaptateur de type B aux extrémités 5'. Les doubles brins d'ADN sont ensuite dénaturés.

(2) Fixation des fragments sur la flowcell et formation des clusters

Le support (flowcell) sur lequel vont être fixés les fragments d'ADN est constitué de canaux dans lesquels sont présents des amorces A' et B' complémentaires des adaptateurs A et B fixés aux extrémités des fragments d'ADN. Les monobrins d'ADN sont déposés sur la flowcell et les extrémités des brins d'ADN vont s'hybrider avec leur amorce complémentaire. Une enzyme polymérase permet la synthèse du brin complémentaire puis le double brin d'ADN est dénaturé par chaleur. Le monobrin d'ADN original est ainsi détaché puis éliminé par lavage. L'extrémité libre du brin d'ADN fixé contient le deuxième type d'adaptateur qui va s'hybrider avec une amorce complémentaire proche en formant un pont. L'enzyme polymérase permet ensuite la synthèse du brin complémentaire. Enfin l'ADN double brin formant un pont est dénaturé et permet d'obtenir deux monobrins d'ADN fixés à la flowcell. Ces étapes sont ensuite répétées simultanément pour tous les fragments d'ADN permettant une amplification clonale de tous les fragments d'ADN. Un fois ces étapes terminées, les brins anti-sens sont clivés puis la flowcell est lavée permettant de retenir seulement les brins sens. Toutes ces étapes sont résumées avec la figure 21. Le séquençage peut alors commencer.

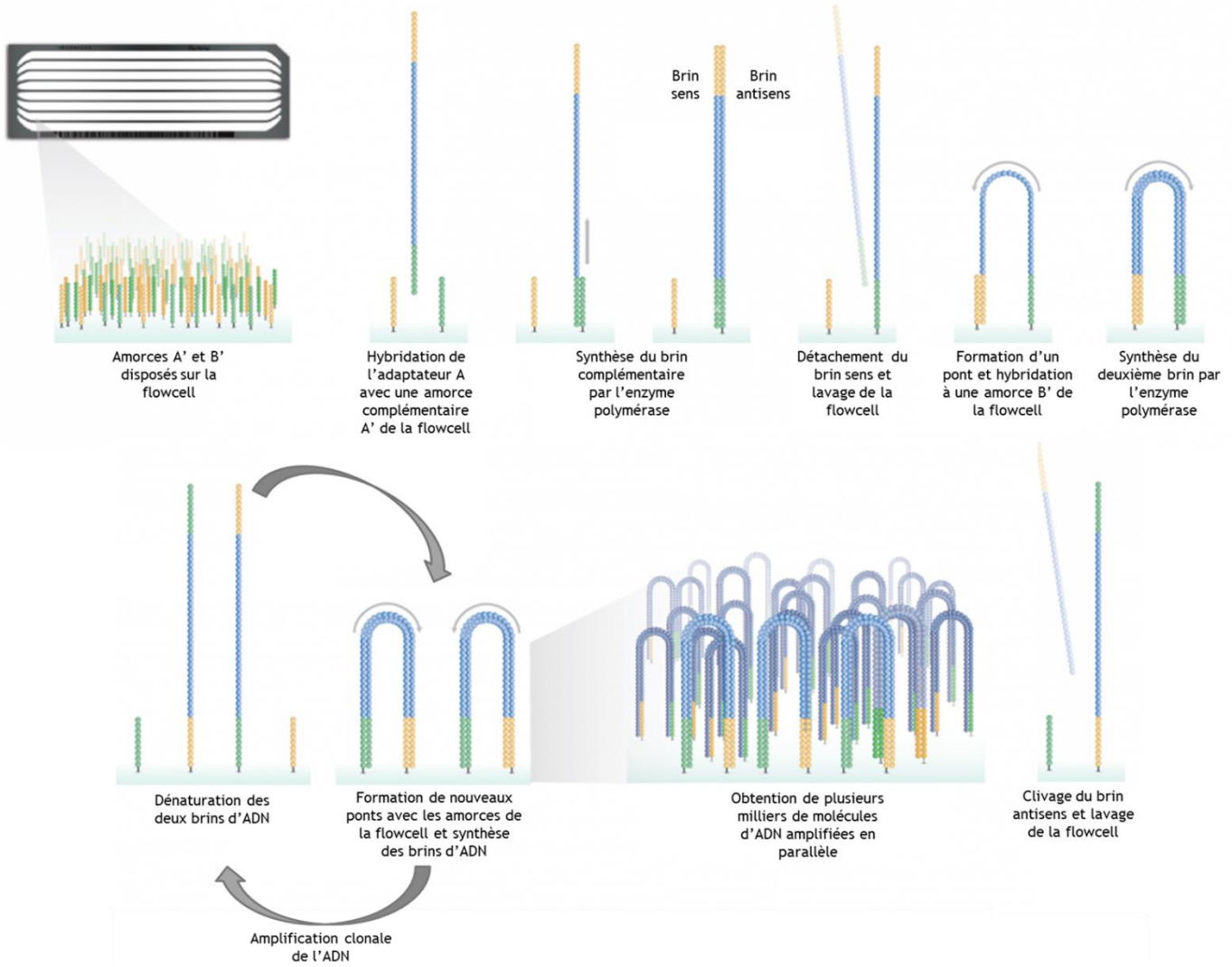


Figure 21. Fixation des fragments sur la flowcell et formation des clusters d'amplification. D'après Biofidal (2016) et Illumina Inc (2017).

(3) Séquençage des fragments

Une amorce va se fixer au niveau de l'extrémité libre des monobrins d'ADN. Une enzyme polymérase permet alors l'incorporation de nucléotides associés avec des molécules fluorescentes de différentes couleurs spécifiques à chaque type de nucléotide. À chaque ajout d'une base par l'enzyme polymérase, la fluorescence de la base ajoutée est captée permettant d'identifier la base qui a été ajoutée. La molécule fluorescente est ensuite clivée puis une nouvelle base est ajoutée. Ces cycles sont répétés entre 150 à 300 fois en fonction de la machine utilisée pour séquencer l'ADN. Une fois le premier cycle de séquençage du brin sens terminé, le monobrin d'ADN va reformer un pont avec l'amorce complémentaire adjacente, puis une enzyme polymérase va permettre la synthèse du brin complémentaire. L'ADN double brin formant un pont est ensuite dénaturé puis le brin sens original est clivé puis la flowcell est lavée. On se retrouve alors cette fois-ci avec le brin anti-sens. Le séquençage est réalisé de la même manière que le brin sens. Ces étapes sont résumées avec la figure 22. Les cycles de séquençage

sont ensuite répétés un certain nombre de fois ce qui permet de déterminer la profondeur de lecture. Elle correspond au nombre moyen de lectures qui se superposent. Un séquençage basse profondeur 1X veut dire que chaque fragment d'ADN a été lu en moyenne une fois. La profondeur dépend du nombre d'individus que l'on séquence en même temps sur une ligne de la flowcell (multiplexage), ainsi que de la quantité d'information séquençable sur une ligne. En poule, dont le génome mesure environ 1Gb, et en supposant une ligne de 80Gb, il est possible de séquencer 4 individus pour obtenir une profondeur moyenne 20X. Si l'on séquence 80 individus sur une même ligne, il est alors possible d'obtenir une profondeur moyenne 1X.

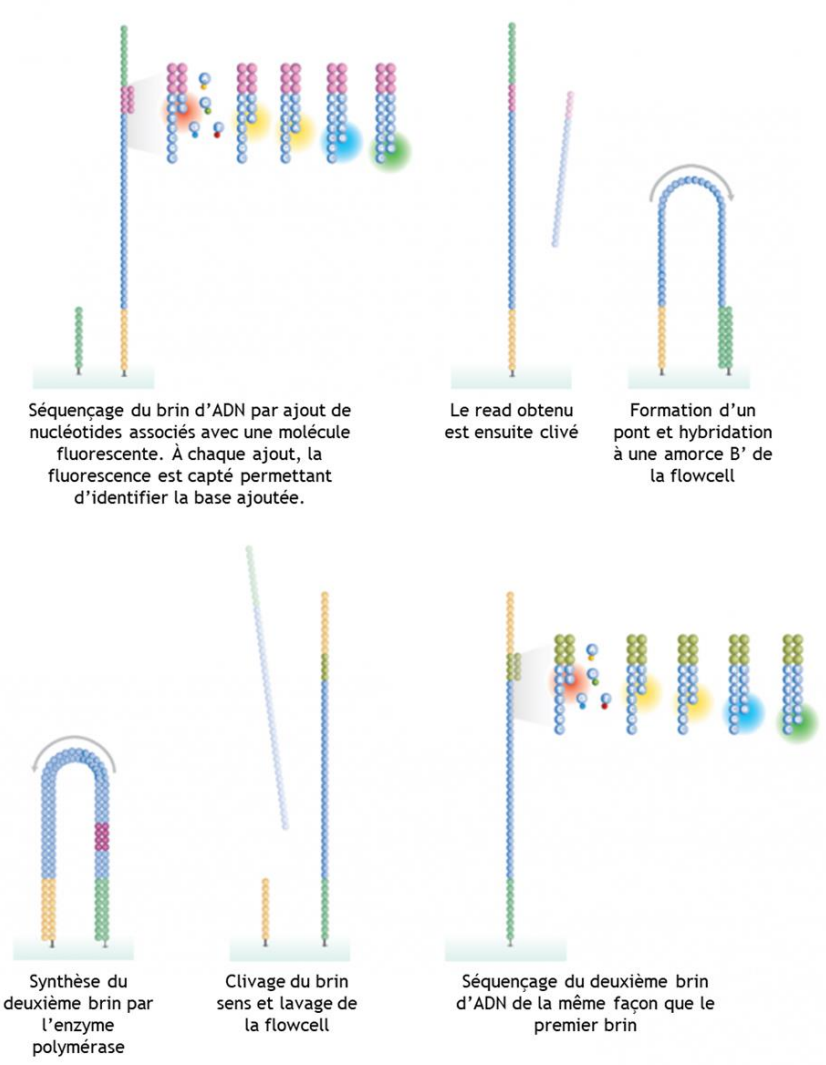


Figure 22. Séquençage des brins d'ADN. D'après Biofidal (2016) et Illumina Inc (2017).

b) Limite du séquençage NGS basse profondeur pour le schéma de sélection

La principale limite du séquençage NGS est son coût trop élevé pour un usage en routine dans les schémas de sélection. Ainsi par exemple, le séquençage d'une poule avec une profondeur 20X est estimée à 600€ avec un séquenceur Illumina HiSeq 3000. En diminuant la profondeur

de lecture, il est possible de séquencer plus d'individus sur une même ligne ce qui permet de diminuer le coût du séquençage de la ligne. En supposant ainsi un séquençage d'une poule avec une profondeur moyenne 1X, le coût serait de 30€. Toutefois, le coût des kits de préparation des librairies d'ADN représente une partie incompressible du montant total du coût du séquençage, quelle que soit la profondeur désirée. Le coût des kits est estimé à 100€ par individu. En rajoutant en plus le coût du séquençage, cette technique est bien trop onéreuse par rapport à une puce BD. Le séquençage NGS basse profondeur n'est donc pas amené à être plus compétitif qu'une puce BD.

Par ailleurs, un autre inconvénient est qu'un séquençage 1X implique que les régions du génome ne sont lues en moyenne qu'une seule fois. Certaines régions peuvent donc être lues une ou plusieurs fois, et d'autres, très nombreuses, aucune fois. Pour un individu homozygote à un marqueur, la lecture d'un seul des deux allèles ne pose pas de problème. En revanche, pour un individu hétérozygote à un marqueur, il est impossible de déterminer correctement son haplotype en ne lisant qu'un seul des deux allèles. Avec une profondeur moyenne de 2X, il est possible de déterminer le deuxième allèle. Mais il est également possible, dans 50% des cas, de lire à nouveau le même brin et donc le même allèle… Enfin, un séquençage basse profondeur 1X implique que certaines régions du génome des individus sont lues en moyenne une fois ou plus, et d'autres pas du tout. Ces régions non lues peuvent être variables entre individus. En plus d'un coût plus élevé, l'utilisation du séquençage NGS basse profondeur génère donc également de la variabilité entre génotypes des différents individus. De ce fait, cette technique n'est, à l'heure actuelle, pas une bonne alternative aux puces BD.

2. Utilisation des méthodes RAD-Seq

a) Définition des méthodes RAD-Seq

Initialement le RAD-Seq (Restriction site-Associated DNA Sequencing) correspondait à une méthode bien particulière de séquençage d'une fraction du génome par l'utilisation d'enzyme de restriction. De même, on retrouve le terme de GBS (Genotyping-By-Sequencing) qui désignait une méthode bien particulière. Aujourd'hui, les noms de ces méthodes ont été étendus à l'ensemble des techniques consistant à utiliser une enzyme de restriction pour couper l'ADN et à séquencer une partie du génome, sur la base des fragments d'ADN obtenus (Andrews et al., 2016). La principale différence avec le séquençage NGS se trouve au niveau de la préparation des librairies d'ADN. Les étapes d'amplification de fragments par PCR (Polymerase Chain Reaction), de fixation des fragments sur la flowcell, formation des clusters d'amplification et séquençage des fragments sont en revanche identiques.

b) *Les différentes méthodes de RAD-Seq et leurs principes*

Les techniques permettant le séquençage d'une fraction du génome par l'utilisation d'enzyme de restriction sont assez nombreuses. Ces techniques peuvent être classées en fonction de l'utilisation d'une ou deux enzymes pour couper l'ADN et de la présence ou non d'une sélection de taille des fragments d'ADN lors de la préparation des librairies d'ADN (Tableau 5).

Tableau 5. Récapitulatif des différentes méthodes RAD-Seq en fonction du nombre d'enzymes utilisées et de la présence d'une sélection de taille des fragments d'ADN.

Nombre d'enzymes	Sélection de taille des fragments d'ADN	Méthode
1	Oui	- RAD-Seq : Baird et al., 2008 - GBS : Beissinger et al., 2013 - Reduced Representation Libraries (RRL) : Van Tassell et al., 2008 - Multiplexed Shotgun Sequencing (MSG) : Andolfatto et al., 2011 - 2bRad : Wang et al., 2012
1	Non	- GBS : Elshire et al., 2011 - Genotyping by Genome Reducing and Sequencing (GGRS) : Chen et al., 2013
2	Oui	- Double-digest RAD-Seq (ddRAD-Seq) : Peterson et al., 2012 - GBS avec deux enzymes : Gardner et al., 2014
2	Non	- GBS avec deux enzymes : Poland et al., 2012 - Complexity Reduction of Polymorphic Sequences (CRoPS) : van Orsouw et al., 2007

Le but n'étant pas de détailler chaque méthode l'une après l'autre, il est toutefois possible de comprendre leurs principes en distinguant plusieurs points communs pour la préparation des librairies d'ADN (Andrews et al., 2016). Ces différentes méthodes commencent toutes par une digestion enzymatique avec une ou deux enzymes. Le choix des enzymes est très important car de ce choix dépendent le nombre de fragments totaux obtenus, leur répartition sur le génome et donc le nombre total de SNP en communs avec les SNP de la puce HD. Une étape de ligation par des adaptateurs est ensuite réalisée à chaque extrémité des fragments obtenus. De nombreuses méthodes se différencient sur ce point en choisissant des types d'adaptateurs et des constructions de barcodes différents. Les barcodes sont de courtes séquences de 4 à 8 nucléotides, différentes les unes des autres et permettant de relier les fragments d'ADN aux

individus dont ils sont issus. Ainsi, par exemple, la méthode RAD-Seq utilise un adaptateur contenant le barcode suivi d'une coupure aléatoire des fragments d'ADN puis un adaptateur divergent en forme de Y est rajouté. La méthode GGRS utilise le même adaptateur contenant le barcode pour chaque extrémité des fragments. La méthode GBS (une enzyme, sans sélection de taille des fragments) utilise un adaptateur commun et un adaptateur contenant le barcode. Au cours de la préparation de la librairie d'ADN, une sélection de taille des fragments peut également être réalisée. Ces méthodes permettent finalement de séquencer une portion différente du génome à différentes profondeurs : de moins de 5X par site par individu avec les méthodes GBS et MSG à plus de 20X par site par individu avec les méthodes de RAD-Seq et ddRAD-Seq (Chen et al., 2013). Toutefois, la profondeur de lecture des fragments dépend principalement du multiplexage envisagé. Les différentes étapes de préparation des librairies d'ADN des méthodes RAD-Seq, GGRS, GBS et ddRAD-Seq sont résumés avec la figure 23.

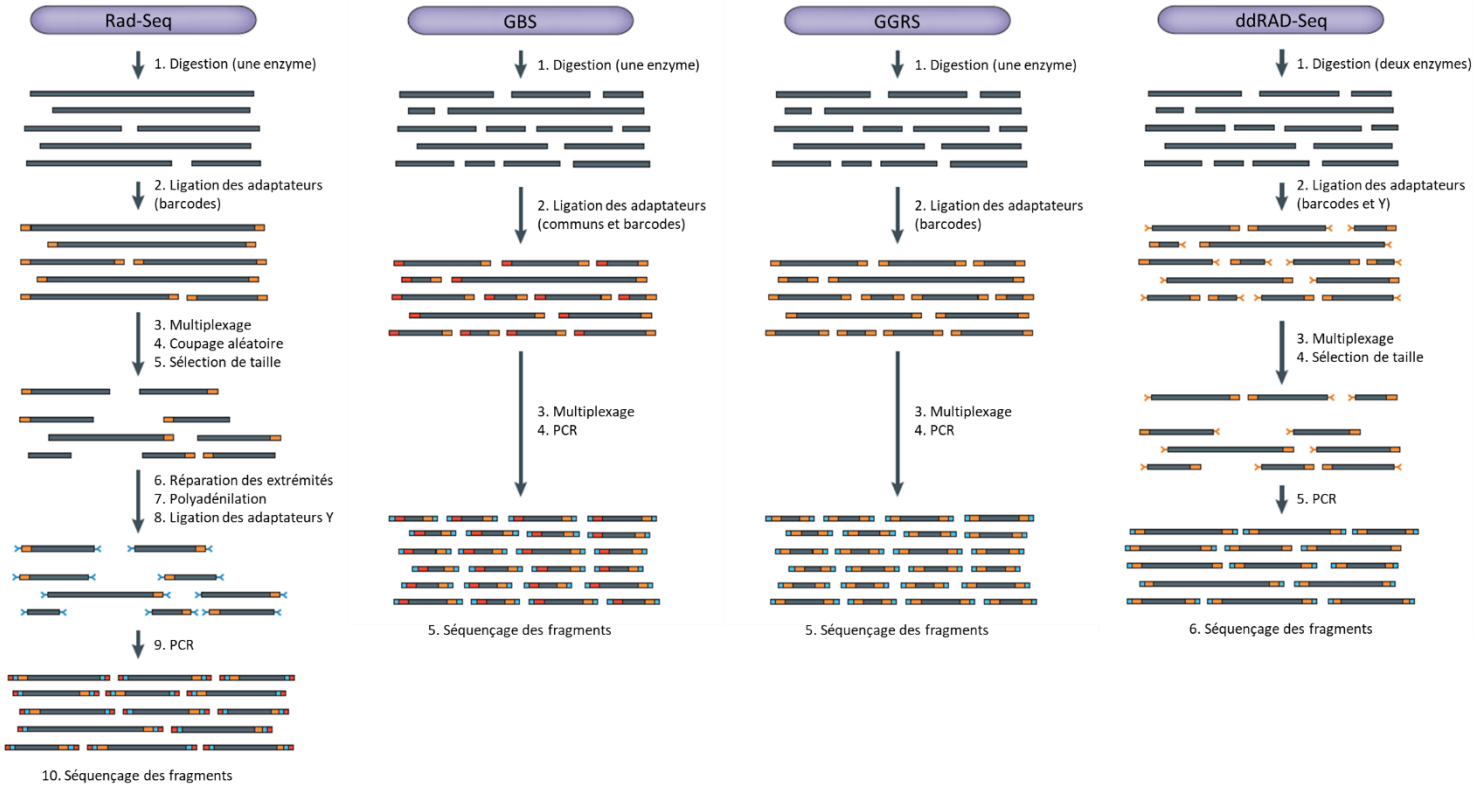


Figure 23. Différentes étapes de préparation des librairies d'ADN pour les méthodes RAD-Seq, GBS, GGRS et ddRAD-Seq. D'après Andrews et al. (2016).

c) Avantage et inconvénients par rapport au séquençage NGS basse profondeur

Par rapport au séquençage NGS basse profondeur, les méthodes de RAD-Seq permettent d'obtenir des profondeurs de lecture plus élevées augmentant ainsi la probabilité pour un individu hétérozygote à différents marqueurs de lire les deux brins d'ADN. Ces lectures seront

homogènes d'un individu à l'autre, à la différence du séquençage NGS basse profondeur où les lectures seront aléatoires entre individus. En revanche, même si avec le NGS basse profondeur, des zones du génome ne sont pas séquencées, il est possible que des zones beaucoup plus grandes ne le soient pas avec les méthodes RAD-Seq. De plus, comme avec le séquençage NGS basse profondeur, les méthodes RAD-Seq peuvent entraîner une certaine variabilité entre individus à cause de plusieurs facteurs impactant l'action des enzymes de restriction et qui sont détaillés dans l'article 3. Le premier est la sensibilité aux méthylations des enzymes de restriction. La présence d'une méthylation est très variable en fonction des individus. En cas de présence d'une méthylation au niveau d'un site de restriction d'une enzyme sensible aux méthylations, l'enzyme n'est pas capable de couper l'ADN au niveau du site. Les fragments obtenus pour certains individus peuvent donc être plus grands que les fragments obtenus par d'autres individus, non porteurs de la méthylation au niveau du même site de restriction. Cela peut entraîner une variabilité entre individus. Le deuxième facteur correspond aux polymorphismes dans les sites de restrictions pouvant également empêcher l'enzyme de restriction de couper le brin d'ADN présentant le polymorphisme. Cela peut également entraîner un biais dans le génotypage avec un phénomène d'élimination d'allèle (allele dropout en anglais) pour les individus hétérozygotes présentant un polymorphisme dans un site de restriction. Il en résulte que l'individu est considéré comme homozygote au niveau du marqueur. Le polymorphisme peut également permettre de définir un nouveau site de restriction Toutefois, la variabilité entre individus générée par ces différents facteurs peut se gérer sans une perte trop importante de précision grâce à l'imputation. De nombreux exemples sont ainsi détaillés dans l'article 3.

Le dernier point, non négligeable, est que les méthodes RAD-Seq sont moins coûteuses qu'un séquençage NGS basse profondeur. De nombreuses publications (Peterson et al., 2012 ; Liao et al., 2015 ; Pértille et al., 2016) estiment un coût inférieur à 50$ par individu. Tout ceci explique pourquoi les méthodes RAD-Seq seraient des alternatives plus intéressantes que le séquençage NGS basse profondeur aux puces BD.

3. Application des méthodes RAD-Seq pour les schémas de sélection

a) État des lieux de l'utilisation dans les schémas de sélection

À l'heure actuelle, les méthodes RAD-Seq peuvent être utilisées en routine dans les schémas de sélection d'espèces végétales par certains sélectionneurs mais chaque entreprise a sa propre stratégie de génotypages avec l'utilisation des puces à SNP (il en existe plus de 50 commerciales pour différentes espèces) ou des techniques de RAD-Seq (Rasheed et al. 2017 ;

Torkamaneh et al., 2018). En revanche, les méthodes RAD-Seq ne sont pas utilisées en routine dans les schémas de sélection d'espèces animales. La majeure partie des recherches a été tout d'abord menée sur des espèces végétales (Elshire et al., 2011 ; Poland et al., 2012) puis sur des espèces animales (De Donato et al., 2013 ; Chen et al., 2013 ; Zhai et al., 2015). En revanche, les méthodes de RAD-Seq permettent la détection de novo de SNP pour toutes les espèces, mêmes celles ne disposant pas de génome de référence. Les SNP détectés peuvent alors être utilisés pour développer des puces à SNP. Cela a été le cas pour le saumon d'Atlantique où de nombreux SNP ont été détectés par RAD-Seq et ont permis le développement d'une puce Affymetrix de 130K SNP (Houston et al., 2014). De la même manière, une puce Affymetrix de 57K SNP a été développée pour la truite arc-en-ciel à partir, entre autre, de SNP détectés par RAD-Seq (Palti et al., 2015 ; Affymetrix Inc, 2015b). Enfin, Gutierrez et al. (2017) ont développé une puce à façon Affymetrix de 55K SNP pour l'huître du Pacifique et l'huître Européenne. Les SNP concernant l'huître Européenne ont été identifiés par RAD-Seq.

b) Intérêt des méthodes RAD-Seq par rapport aux puces BD

Pour des raisons de coûts et de volumes de puces commandés par les sélectionneurs, les puces à SNP BD utilisées en routine servent généralement pour plusieurs lignées voire même plusieurs espèces. Avec les méthodes RAD-Seq, le prix n'est plus dépendant du volume désiré et des fabricants de puces à SNP, les coûts étant principalement dus aux kits de préparation des librairies d'ADN et au séquençage. C'est l'un des principaux avantages de cette méthode.

En revanche, si l'on suppose qu'une méthode RAD-Seq est utilisée comme alternative aux puces BD, il est nécessaire de détecter des SNP en commun avec les SNP de la puce HD car la population de référence est génotypée avec une puce HD. Ce nombre de SNP peut être variable et dépend fortement du choix de l'enzyme de restriction du fait de la répartition des sites de restriction et du nombre de fragments générés. L'imputation peut encore permettre de remonter aux génotypages HD avec des précisions intéressantes, à condition d'avoir suffisamment de SNP en commun. À notre connaissance, il n'y a pas de publication étudiant la qualité d'imputation à partir des marqueurs en commun et faisant le lien avec la précision des évaluations. L'article 3 se penche sur ce sujet mais uniquement avec des données RAD-Seq simulées.

Enfin, comme expliqué précédemment, les méthodes de RAD-Seq permettent la détection de novo de SNP pour toutes les espèces, mêmes celles ne disposant pas de génome de référence.

C. Optimisation de la sélection génomique en travaillant sur les reproducteurs

En plus d'une optimisation de la sélection génomique en travaillant au niveau des candidats à la sélection, il est également possible d'optimiser la sélection génomique en travaillant au niveau des reproducteurs, et donc de la population de référence. Deux axes principaux de travaux peuvent être envisagés : travailler sur les individus qui composent la population de référence ou travailler sur le nombre de marqueurs des individus de la population de référence en utilisant la séquence des individus.

1. Travailler sur le nombre d'individus dans la population de référence

a) Cumuler les individus et les performances

Le principe de la sélection génomique étant basé sur l'estimation des effets des SNP sur différents caractères à partir d'une population de référence, plus l'on dispose d'individus génotypés et de performances dans la population de référence, plus l'estimation des effets des SNP est précise. Plusieurs études ont montré que cumuler des individus génotypés dans la population de référence permettait d'augmenter la précision des évaluations. Pszczola et al. (2009) ont démontré, sur simulations en bovins laitiers, que la précision des évaluations des candidats (estimée comme la corrélation entre les TBV et les GEBV) pour un caractère avec une héritabilité de 0.3 était de 0.84 avec 1000 individus génotypés et de 0.90 avec 2000 individus génotypés. Ces études ont été confirmées avec des données réelles. Hozé et al. (2014) ont ainsi montré pour des bovins laitiers de race Normande que la précision des évaluations (estimées comme la corrélation entre GEBV et DYD) pour la production laitière était de 0.32 avec seulement 198 taureaux normands génotypés et de 0.48 avec 1597 taureaux normands génotypés. Par ailleurs, le niveau de l'amélioration de la précision des évaluations avec une augmentation de taille de la population dépend également de l'héritabilité du caractère. Plus le caractère est héritable, plus l'augmentation de taille de la population de référence est bénéfique pour la précision des évaluations. Pour un caractère avec une héritabilité de 0.05, Pszczola et al. (2009) ont obtenu une précision des évaluations de 0.70 et 0.79 avec respectivement 1000 et 2000 individus dans la population de référence. Enfin, la précision de l'évaluation dépend de la taille efficace N_e de la population. Plus la taille de la population efficace augmente, plus le DL entre marqueurs décroit rapidement et plus le nombre d'effets de SNP à estimer augmente. Il en résulte une diminution de la précision des évaluations avec une augmentation de la taille efficace de la population. En supposant une taille de population efficace élevée, il est donc

nécessaire d'augmenter considérablement la taille de la population de référence pour améliorer la précision des évaluations.

Toutefois, le coût du génotypage HD d'un individu étant élevé, cumuler des individus dans la population de référence n'est pas une solution très économique. Plutôt que d'augmenter de façon trop importante la taille de la population de référence, il est intéressant d'essayer d'optimiser le choix des individus constituant la population de référence.

b) Optimiser le choix des individus

De nombreuses études (Habier et al., 2007 ; Habier et al., 2010 ; Pszczola et al., 2012 ; Weng et al., 2016) ont montré que les relations de parenté au sein même de la population de référence et les relations de parenté entre population de référence et population candidate devaient être bien analysées afin d'obtenir une bonne précision des évaluations génomiques.

En effet, au niveau de la population de référence, il est préférable de choisir des individus qui ne sont pas apparentés entre eux afin de maximiser la diversité génétique et ainsi de capter la diversité haplotypique de la population (Pszczola et al., 2012). Ceci permet de limiter le risque qu'un allèle reçu par un candidat ne soit pas retrouvé parmi les individus de la population de référence. Une meilleure estimation des effets des SNP, et donc des valeurs génomiques, est alors possible. Pszczola et al. (2012) montrent ainsi qu'une population de référence composée de 2000 vaches laitières choisies aléatoirement permet d'obtenir une meilleure précision des évaluations qu'à partir de 2000 vaches issues de seulement 5 taureaux différents.

En revanche, il est préférable de maximiser les relations de parenté entre population de référence et population candidate. Ceci permet d'éviter qu'un allèle particulier soit présent dans la population candidate et absent dans la population de référence. Habier et al. (2010) ont montré, en bovin laitier, qu'utiliser une population de référence constituée des pères, des frères et demi-frères des pères des candidats à la sélection permettait d'obtenir de meilleures précisions des évaluations qu'avec une population de référence contenant le même nombre d'individus mais sans les pères, les frères et les demi-frères des pères des candidats à la sélection. Maximiser les relations de parenté entre populations de référence et candidate permet également de bénéficier du DL à longue distance. Clark et al. (2012) et Habier et al. (2013) ont montré que pour un individu donné, la précision de son évaluation génomique dépend du DL à courte distance observable sur l'ensemble de la population de référence, mais aussi du DL à longue distance observable au sein des individus d'une même famille. Ce DL à longue distance se traduit par de longs segments chromosomiques transmis d'une génération à une autre. Ces deux études suggèrent que le DL de la population de référence permet de définir une précision

d'évaluation minimale pour tout candidat non apparenté à la population de référence. Le DL intra-famille permet ensuite d'améliorer la précision de l'évaluation pour l'individu.

Il est donc important que la population de référence soit constituée d'individus avec le moins d'apparentement possible entre eux, mais avec le plus d'apparentement possible avec la population candidate. Ceci permet de maximiser le DL populationnel et le DL intra-famille.

Enfin, les recombinaisons au fil des générations peuvent réduire les niveaux de DL et faire disparaître les associations entre marqueurs et QTL. Le DL intra-famille étant un DL à longue distance, les recombinaisons vont faire chuter rapidement les niveaux de DL au fil des générations pour atteindre un niveau plus faible correspondant au DL populationnel. Celui-ci diminue également, mais plus lentement, au cours des générations (Bastiaansen et al., 2012). Comme expliqué à la section I.B.2, il est donc important de renouveler la population de référence au fur et à mesure des générations pour ne pas se tromper dans l'estimation des effets des SNP.

2. Travailler sur le nombre de marqueurs dans la population de référence

Plus la densité de marqueurs augmente, plus la distance entre marqueurs diminue et plus le DL observé entre marqueurs adjacents est fort. Ceci augmente la probabilité qu'un marqueur soit fortement associé avec un QTL. Une augmentation du nombre de marqueurs est possible en utilisant une puce à SNP HD si une telle puce existe et si l'on ne travaille déjà pas avec. Il est également possible d'utiliser le séquençage NGS ou des méthodes RAD-Seq haute-profondeur.

a) Utilisation de puces HD ou du séquençage NGS pour les individus de référence

Pour certaines espèces comme en bovin, une puce à SNP HD est disponible mais n'est pas utilisée en routine pour la sélection génomique. Or, de nombreuses publications (Chen et al., 2014 ; Raoul et al., 2017 ; Grossi et al., 2018) ont montré qu'augmenter le nombre de marqueurs jusqu'à une densité moyenne (environ 50K SNP) permettait d'améliorer de façon significative la précision des évaluations. En utilisant une haute densité de SNP, il peut alors être envisageable d'améliorer encore la précision des évaluations en augmentant la probabilité d'association des marqueurs avec les QTL. Cette hypothèse a été testée par Su et al. (2012) et VanRaden et al. (2013) qui ont comparé la précision des évaluations génomiques obtenues à partir de la puce bovine 50K et à partir de la puce HD 777K. Ces études ont montré qu'un gain faible de précision (environ 1%) était obtenu en passant de la puce 50K à la puce HD. Ces résultats ont également été retrouvés en poule pondeuse et sont détaillés dans le chapitre III et

l'article II. L'effet de la densité de marqueurs sur la précision des évaluations diminue lorsqu'une valeur seuil est dépassée. Toutefois, une telle densité de marqueurs entraîne une faible distance entre marqueurs et permet donc de rendre les estimations des effets des SNP plus résistantes à la chute de DL au fil des générations.

Par ailleurs, il est également possible d'utiliser du séquençage NGS haute-profondeur pour obtenir la séquence des individus qui constituent la population de référence. Utiliser la séquence des individus permet la détection de plusieurs millions de SNP et surtout des mutations causales. Il n'y a alors plus besoin du DL entre marqueurs et mutations causales puisqu'elles sont inclues dans la séquence (Bolormaa et al., 2019). Ceci renforce encore plus les estimations des effets des SNP. Druet et al. (2014) montrent que la présence des mutations causales avec une faible MAF (<1%) permet une amélioration de 28% de la précision des évaluations génomiques par rapport aux résultats obtenus avec une puce à SNP moyenne densité simulée avec une densité de 50kb entre marqueurs. En revanche, pour des mutations causales ayant des fréquences similaires aux autres marqueurs de la séquence, la séquence ne permet qu'une amélioration de 1.5% de la précision des évaluations par rapport aux résultats de la même puce simulée. Par ailleurs, Ni et al. (2017) ont montré en poule pondeuse que les gains de précisions obtenus à partir de séquences imputées étaient très faibles comparés aux précisions obtenues avec des puces HD. Ceci renforce les conclusions déjà mises en évidence concernant les faibles améliorations de précision d'évaluation obtenues avec une puce HD par rapport à une puce MD. L'intérêt du séquençage pour la sélection génomique peut donc s'avérer assez limité.

Enfin à l'heure actuelle, le séquençage NGS est très couteux. Il n'est pas envisageable de séquencer autant d'individus qu'avec une puce à SNP. Si le choix du séquençage est fait, il convient donc de réfléchir aux choix des individus à séquencer, le point le plus important développé dans la partie précédente étant de maximiser la diversité génétique entre individus à séquencer. La maximisation de l'apparentement entre individus à séquencer et population candidate n'est plus aussi importante qu'avec des puces à SNP grâce à la densité de marqueurs et au DL très fort entre marqueurs adjacents.

b) Utilisation des méthodes RAD-Seq pour les individus de référence

Enfin, lorsqu'une puce HD n'est pas disponible et que le séquençage NGS n'est pas envisageable pour une population, des méthodes RAD-Seq permettant une haute profondeur de lecture des fragments d'ADN peuvent être mises en place. Les travaux de Poland et al. (2012) et d'Elbasyoni et al. (2018) ont montré que de bonnes précisions d'évaluations génomiques peuvent être obtenues avec des méthodes RAD-Seq pour des densités de SNP équivalentes aux puces à SNP. Il y avait même une augmentation de la précision en fonction des caractères. Ces

résultats ont également été retrouvés par Gorjanc et al. (2015) avec des données simulées pour des espèces d'élevages. Ces méthodes peuvent donc être intéressantes dans le cas de populations de référence et candidates génotypées avec ces méthodes.

Toutefois, les individus composant les populations de référence sont habituellement génotypés avec des puces MD ou HD dont seule une petite proportion des SNP sont en commun avec les SNP issus des méthodes RAD-Seq. Torkamaneh et Belzile (2015) ont montré qu'il était possible de combiner les SNP issus de puces et ceux issus de méthodes RAD-Seq puis d'imputer les données manquantes (sur la base des SNP en commun entre les deux panels de SNP) pour obtenir une population de référence génotypée en haute densité avec des SNP issus de puces et de méthode RAD-Seq. La combinaison des deux types de panels SNP peut contribuer à augmenter la précision des évaluations, même si comme expliqué précédemment, au-delà d'une certaine densité, l'intérêt de l'augmentation du nombre de marqueurs peut être assez limité. Enfin, à l'heure actuelle, il s'agit d'une méthode plus coûteuse que celle reposant sur des génotypages HD pour la population de référence et des génotypages BD pour la population candidate.

D. Optimisation du génotypage en considérant un génotypage BD ou MD pour les populations de référence et candidate

Un dernier point concernant l'optimisation des génotypages est de considérer la même densité de génotype BD ou MD pour la population de référence et la population candidate, et de ne pas utiliser d'imputation. Cela peut permettre de diminuer de façon conséquente les coûts liés au génotypage des individus reproducteurs en ne génotypant plus en HD ces individus. L'étude de Moghaddar et al. (2015) a permis d'étudier l'effet du remplacement des génotypages MD par des génotypages BD pour des moutons de race Mérinos. En considérant 1000 moutons dans la population de référence dont les 175 plus anciens constituant la population de validation, la précision des évaluations à partir de génotypages 50K est de 0.446 pour le poids à la naissance et de 0.219 pour la profondeur de la noix de côte à la naissance. À partir de génotypages 12K, les précisions pour ces deux caractères sont respectivement de 0.412 et 0.205. Il y a donc une diminution de la précision des évaluations avec une diminution du nombre de marqueurs. Le remplacement des génotypages MD par des génotypages BD pour l'ensemble des individus peut donc avoir des conséquences sur la précision des évaluations génomiques.

En revanche, d'autres études ont montré que l'utilisation de génotypages MD plutôt que HD pour l'ensemble des individus n'avait qu'un impact faible sur la précision des évaluations génomiques. VanRaden et al. (2011) ont montré sur des simulations en bovins que le passage

d'un génotypage de 500K à 50K SNP pour plus de 33 000 individus Holstein n'entrainait qu'une diminution de 1.6% de la précision des évaluations. De même, Su et al. (2012) ont étudié la précision des évaluations estimée comme la corrélation au carré entre les valeurs génomiques directes et les valeurs dérégressées divisée par la fiabilité des valeurs dérégressées pour des individus Holstein avec environ 3000 individus dans la population de référence et 1400 individus dans la population candidate. Avec des génotypages 777K ou 54K pour l'ensemble des individus, la précision estimée était respectivement de 0.429 et 0.425 pour le taux de protéine et de 0.413 et 0.404 pour la fertilité.

Ces évolutions sont attendues d'après les sections I.B.2 et III.C.2.a. Plus le nombre de marqueurs augmente, plus la précision des évaluations augmente. Toutefois, l'effet de la densité de marqueur sur la précision des évaluations diminue lorsqu'une valeur seuil est dépassée. C'est pourquoi il est intéressant d'essayer d'optimiser ce nombre de marqueurs en diminuant la densité de SNP sans trop diminuer la précision des évaluations. Par ailleurs, à notre connaissance, aucune étude ne s'est penchée sur ces questions en poules pondeuses. Or, dans le schéma de sélection génomique classique en pondeuse, la puce HD de 600K est utilisée pour génotyper les reproducteurs car elle est la seule puce commerciale disponible jusqu'à très récemment. D'après les études de VanRaden et al. (2011) et Su et al. (2012), il peut donc être intéressant d'étudier l'intérêt de ne plus utiliser d'imputation en se servant directement des génotypages MD ou BD pour l'ensemble des individus. Toutefois, il convient de noter que les niveaux de précisions atteints en Holstein sont bien plus élevés qu'en poule pondeuse (cf. section I.B.4.c). Ainsi, un impact sur la précision des évaluations, aussi faible soit-il pour les Holstein, sera plus conséquent sur la précision des évaluations génomiques des pondeuses. Il a également été montré que la précision des évaluations génomiques diminuait au fil des générations et nécessitait un renouvellement de la population de référence. En supposant une faible dégradation des résultats pour les premières générations, l'impact peut devenir bien plus conséquent au bout d'un certain nombre de générations en cumulant les erreurs réalisées pour chaque génération… Plutôt que d'utiliser des génotypages BD pour l'ensemble des individus, les génotypages MD pourrait peut-être être un compromis. Ces questions ont été abordées et sont largement détaillées dans le chapitre III et l'article II.

Chapitre II. Qualité d'imputation des génotypes obtenus à partir de puces basse densité en poule pondeuse

I. Introduction

Les coûts de génotypages avec une puce HD restent élevés pour une utilisation en routine sur un grand nombre de candidats à la sélection. Un des enjeux de la sélection génomique est de développer des puces basse densité coûtant moins cher. Puis à partir des génotypes haute densité d'une population de référence, qui est très souvent la population parentale, il est possible avec les différentes techniques d'imputation de déduire les génotypes HD des candidats à la sélection. Toutefois, de nombreux facteurs peuvent influencer la qualité des imputations et doivent être pris en compte lors du développement des puces BD. Ces facteurs développés dans le premier chapitre sont la densité de SNP sur les puces BD (Hickey et al., 2012 ; Hozé et al., 2013), la fréquence allélique des SNP à imputer (Hayes et al., 2012 ; Heidaritabar et al., 2014), la méthodologie utilisée pour développer les puces BD (Weigel et al., 2009 ; Aliloo et al., 2018), la taille de la population de référence (Dassonneville et al., 2011 ; Hozé et al., 2013), les relations d'apparentement entre populations de référence et candidate (Druet et al., 2010 ; Bouquet et al., 2015). En revanche, à notre connaissance, les particularités du génome aviaire notamment en matière de structure de déséquilibre de liaison n'ont pas été totalement explorées pour développer des puces BD. L'objectif de cette étude est donc de déterminer la stratégie de construction de puce basse densité la mieux adaptée à l'espèce poule.

II. Article I : Design de puces basse densité pour l'imputation de génotypes chez la poule pondeuse

Article paru dans BMC Genetics

Herry F, Hérault F, Picard Druet D, Varenne A, Burlot T, Le Roy P, Allais S. 2018.
Design of low density SNP chips for genotype imputation in layer chicken.
BMC Genetics. 19:108-121.

Herry et al. BMC Genetics (2018) 19:108
https://doi.org/10.1186/s12863-018-0695-7

BMC Genetics

RESEARCH ARTICLE Open Access

Design of low density SNP chips for genotype imputation in layer chicken

Florian Herry[1,2], Frédéric Hérault[2], David Picard Druet[2], Amandine Varenne[1], Thierry Burlot[1], Pascale Le Roy[2] and Sophie Allais[2*]

Abstract

Background: The main goal of selection is to achieve genetic gain for a population by choosing the best breeders among a set of selection candidates. Since 2013, the use of a high density genotyping chip (600K Affymetrix® Axiom® HD genotyping array) for chicken has enabled the implementation of genomic selection in layer and broiler breeding, but the genotyping costs remain high for a routine use on a large number of selection candidates. It has thus been deemed interesting to develop a low density genotyping chip that would induce lower costs. In this perspective, various simulation studies have been conducted to find the best way to select a set of SNPs for low density genotyping of two laying hen lines.

Results: To design low density SNP chips, two methodologies, based on equidistance (EQ) or on linkage disequilibrium (LD) were compared. Imputation accuracy was assessed as the mean correlation between true and imputed genotypes. The results showed correlations more sensitive to false imputation of SNPs having low Minor Allele Frequency (MAF) when the EQ methodology was used. An increase in imputation accuracy was obtained when SNP density was increased, either through an increase in the number of selected windows on a chromosome or through the rise of the LD threshold. Moreover, the results varied depending on the type of chromosome (macro or micro-chromosome). The LD methodology enabled to optimize the number of SNPs, by reducing the SNP density on macro-chromosomes and by increasing it on micro-chromosomes. Imputation accuracy also increased when the size of the reference population was increased. Conversely, imputation accuracy decreased when the degree of kinship between reference and candidate populations was reduced. Finally, adding selection candidates' dams in the reference population, in addition to their sire, enabled to get better imputation results.

Conclusions: Whichever the SNP chip, the methodology, and the scenario studied, highly accurate imputations were obtained, with mean correlations higher than 0.83. The key point to achieve good imputation results is to take into account chicken lines' LD when designing a low density SNP chip, and to include the candidates' direct parents in the reference population.

Keywords: Imputation accuracy, Low density chip, Layer chickens, SNP density, Linkage disequilibrium, MAF, Degree of kinship

Background

In 2001, Meuwissen et al. [1] proposed a method known as "genomic selection", consisting in using dense molecular markers such as single nucleotide polymorphisms (SNPs), to predict the genomic value of individuals without information regarding their phenotype. Since 2013, a high density (HD) genotyping SNP chip for chicken (600K Affymetrix® Axiom® HD genotyping array) [2] has enabled the implementation of genomic selection in layer and broiler breeding. When the genotypes and phenotypes of a reference population are known, it is possible to estimate the genomic value of a genotyped individual. The main objective is to choose, among the selection candidates of generation N, the best breeders for one or more traits. The selected breeders will then produce the individuals of generation N + 1.

However, the genotyping costs induced by the HD SNP chip remain high for a routine use on a large

* Correspondence: sophie.allais@agrocampus-ouest.fr
[2]PEGASE, INRA, Agrocampus Ouest, 16 Le Clos, 35590 Saint-Gilles, France
Full list of author information is available at the end of the article

number of selection candidates. It has therefore been deemed interesting to develop a low cost genotyping approach, which can be achieved through the development of a low density genotyping chip. In that perspective, a set of SNP markers has to be selected to enable the imputation (prediction) of missing genotypes on a high density SNP chip. Imputation involves predicting the high density genotyping of selection candidates from their low density genotyping and from the high density genotyping of the reference population [3]. This approach relies on the Mendelian Laws of Inheritance and on linkage disequilibrium (LD).

To date, many studies on genotype imputation have been led in the bovine, porcine, ovine and poultry sectors. These studies have been conducted using different software such as FImpute [4], Beagle [5] or AlphaImpute [6]. Imputation accuracy can be calculated by comparing imputed genotype with true HD genotype, for each SNP.

Several factors influencing imputation accuracy have been studied in the literature. These factors need to be taken into account when designing a low density SNP chip, in order to get accurate imputation. The SNP density of low density SNP chips [7], the effect of linkage disequilibrium threshold [8], the effect of minor allele frequencies (MAF) of imputed SNPs [9, 10], the size of the reference population [11], and the degree of kinship between reference population and candidate population [8, 10] have all been identified in the literature as factors influencing imputation accuracy. These factors also have an impact on genomic evaluations [12–14]. However, the specificities of the *Gallus gallus* genome [15], especially with regard to the particular structure of the avian linkage disequilibrium [16–18] have not yet been fully investigated.

In this study, several factors affecting imputation accuracy were therefore investigated. These factors are: SNP density, LD threshold, MAF of imputed SNPs, chromosome size, the methodology used to design the low density SNP chip (based on physical equidistant intervals or on LD), the composition of the reference population in terms of size and degree of kinship, as well as the effect of using female genotypes. Various in silico analysis were conducted in order to choose the best strategy to achieve low density genotyping of two different laying hen lines.

Methods

Animals

The populations studied were comprised of two different commercial pure lines of Rhode Island (RI) and Leghorn (L) laying hens. Each line was created and selected by Novogen (Plédran, France). The RI line was comprised of 2370 chickens split in four generations. The L line was comprised of 1483 chickens split in two generations. For both lines, each generation was divided in three batches and a new batch was produced every 6 months from 2010 to 2015 (Fig. 1). The selection objectives were the same for each batch and for each line, and remained the same over time. Animals were firstly selected on bird weight and egg quality, and secondly on egg quality and egg production. In addition, for each batch, the theoretical selection numbers defined were 50 males and 200 females and each sire was mated with 4 females.

For the RI line and for each batch, all male breeders were genotyped. Female breeders were genotyped as well, starting at the third generation (G2). From generation G3, born in November 2014, genomic selection was routinely implemented, by genotyping only the breeders. This enabled to reduce the generation interval for sires from 90 to 30 weeks. This accounts for the lower number of sires genotyped in November 2014 and May 2015.

For the L line and for each batch, all male breeders were genotyped. Female breeders were genotyped starting from the last batch of the first generation.

Genotyping

Blood was taken from the brachial veins of the animals. DNA was extracted and hybridized using the 600K Affymetrix® Axiom® HD genotyping array [2]. Genotyping was performed by Ark-Genomics (Edinburgh, UK) for the first two generations of the RI line, and by the high-throughput genotyping platform Gentyane (Clermont-Ferrand, France), for the rest. Both platforms used Axiom Analysis Suite to create genotype files, which were received in PED format [19]. In order to check genotype consistency, PLINK identified 870 SNPs with visible strand inversion. Genotype harmonizer [20] identified 148 SNPs with invisible strand inversion. PLINK used both lists to flip the strand for the SNPs with strand inversion. Finally, the different sets of genotypes were successfully merged with PLINK.

Each individual was genotyped for 580,961 SNPs. According to the fifth annotation release of *Gallus gallus* genome [21], these SNPs were distributed over macrochromosomes (1 to 5), intermediate chromosomes (6 to 10), micro-chromosomes (11 to 28 and 33), one linkage group (LGE64), two sexual chromosomes Z and W, as well as a group of 3724 SNPs with unknown location.

Genotypes were filtered through six successive steps with classical thresholds (Table 1), including individual call rate (> 95%), MAF (< 0.05), SNP call rate (> 95%), and Hardy-Weinberg equilibrium ($P < 10^{-4}$). Subsequently, the SNPs with unknown location or located on sexual chromosome W were removed, as well as the animals showing pedigree incompatibilities. Most of the SNPs had to be removed because they showed zero MAF. This was to be expected, since the HD SNP chip was designed for both layers and broilers and for a ratio of 1:2.

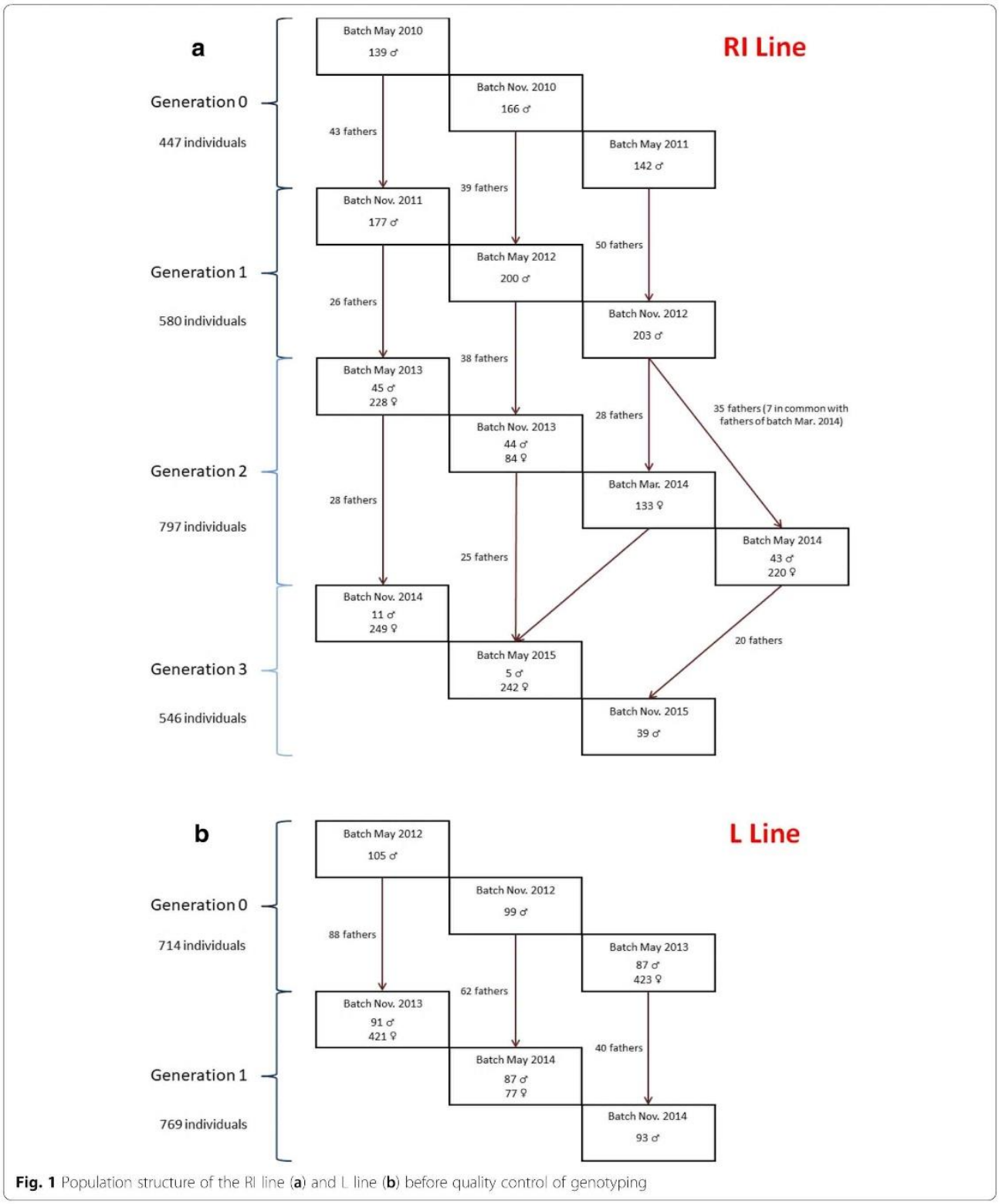


Fig. 1 Population structure of the RI line (**a**) and L line (**b**) before quality control of genotyping

For the RI line and for the L line respectively, 300,351 SNPs and 2362 individuals and 245,669 SNPs and 1474 individuals remained available for the analyses.

Low density SNP chips design

From the remaining SNPs of the RI and L lines, several in silico low density SNP chips were designed by

Table 1 Summary of the different steps of quality control

Genotypes filtration	RI Line	L Line
Individual Call Rate (> 95%)	8	3
MAF (= 0)	204,122	228,452
MAF (0 < X < 0.05)	54,650	99,000
SNP Call Rate (> 95%)	7541	2530
Hardy-Weinberg equilibrium ($P < 10^{-4}$)	12,538	3857
SNP with unknown location or on W	1759	1453
Pedigree Incompatibility issues	0	6
SNP retained for analyses	**300,351**	**245,669**
Animals retained for analyses	**2362**	**1474**

In bold are the total number of SNPs and animals retained for analyses

selecting a subset of SNPs (Table 2). Both lines were studied independently of each other. Low density genotyping was generated by masking all markers, except those corresponding to the in silico selected SNP panel for each line.

Many studies, mainly in non-avian sectors [7–9, 11, 22], focused on low density SNP chips designed according to the equidistance methodology (EQ), by choosing SNPs at regular physical intervals (in pb) along chromosomes. This methodology was therefore chosen. More precisely, for each interval, the SNP with the highest MAF, or the one located furthest on the left, in case of equivalent MAF, was chosen as representative of the interval. This way, 11 low density "equi" SNP chips designed according to this method were studied for each line, with SNP density ranging from 2K to 50K SNPs.

Table 2 Summary of the different low density SNP chips simulated

Methodology	SNP Chip	Number of SNP	
		RI Line	L Line
Equidistance	50Kequi	49,636	50,307
	40Kequi	40,160	39,838
	30Kequi	29,970	30,075
	20Kequi	19,910	19,948
	15Kequi	14,963	14,955
	10Kequi	**10,001**	**9966**
	7.5Kequi	7527	7496
	5Kequi	4991	4996
	4Kequi	4023	4000
	3Kequi	2992	3003
	2Kequi	2013	2003
Linkage Disequilibrium	LD0.8	21,717	18,052
	LD0.7	16,615	13,696
	LD0.6	13,214	**10,736**
	LD0.5	**10,711**	8626
	LD0.4	8521	6944
	LD0.3	6875	5578
	LD0.2	5371	4330
	LD0.1	3935	3232
	LD0.05	3205	2624

SNP chips in bold are SNP chips having an equivalent SNP density of 10 K SNPs

However, considering the heterogeneous structure of chicken linkage disequilibrium (LD) [18], it would have been better to design the low density SNP chips according to the intra-chromosomes LD, i.e. choosing tag SNPs at regular genetical intervals [23]. The low density SNP chips were therefore designed using the SS4I software [24]. This method made it possible to get clusters of SNPs according to a chosen LD threshold. For each cluster, the SNP with the highest MAF was then kept and used as representative of this cluster. Nine "LD SNP" chips designed with this method were studied, with LD threshold ranging from 0.05 to 0.8.

Population scenarios

Eight population scenarios were set up and differed depending on the reference and candidate populations (Table 3). The individuals with the simulated low density genotyping were called "candidate population". The reference population, in this study, refers to the individuals having high density genotyping and used to impute the candidates.

Scenarios (A), (B), (D_1) and (D_2) correspond to cases where the individuals of the candidate population are directly related to the reference population, because of the presence of their sires and/or dams in the reference population. In scenarios (C), (E) and (F), the individuals of the candidate population are also directly related to the reference population, but this time the size of the reference population was increased by adding individuals from previous generations. Finally, scenarios (G), (H) and (I) correspond to cases where the reference population does not include the sires of the candidate population, as a generation gap was introduced between the reference population and the candidate population.

The scenario (A) concerned RI and L lines, and the others concerned only the RI line.

Imputation accuracy studies

Based on the low density SNP chips designed and on the population scenarios simulated, seven different parameters were studied, in order to investigate their influence on imputation accuracy.

The first four parameters were studied on scenario (A), for both lines, and concerned the low density SNP chips used. The lines were studied independently of each other.

Table 3 Summary of the different scenarios depending on reference and candidate populations

	A	B	C	D_1	D_2	E	F	G	H	I
Reference Population	G0	G1	G0 + G1	G2(♂)	G2(♂ + ♀)	G1 + G2(♂)	G0 + G1 + G2(♂)	G0	G1(♂)	G0(♂)
Number of individuals	447	580	1027	73	735	653	1100	447	120	132
Selection Candidates	G1	G2	G2	G3	G3	G3	G3	G2	G3	G3
Number of individuals	580	794	794	541	541	541	541	794	541	541

♂ indicates that only male breeders are used in the reference population
♂ + ♀ indicates that both male and female breeders are used in the reference population

1) The first parameter studied was the effect of SNP density on the low density SNP chips. This study was conducted with the 11 low density SNP chips designed with the EQ methodology as well as with the 9 low density SNP chips designed with the LD methodology.
2) Secondly, the effect of the LD threshold used to design the 9 low density SNP chips was investigated.
3) Thirdly, the effect of minor allele frequencies of imputed SNP on imputation accuracy was studied. This study was done using the low density SNP chips of 3K and 10K SNPs, designed according to the two methodologies, i.e. EQ and LD.
4) Fourth, the effect of the type of chromosome (micro, intermediate, macro or Z) was studied for both the equi and the LD chips, with a density of 3K and 10K SNPs.

The remaining last three parameters were studied at an equivalent SNP density of 10K SNPs, i.e. 10Kequi and LD0.5 low density SNP chips, and of 3K SNPs, i.e. 3Kequi and LD0.05 low density SNP chips. This was meant to focus on the effects of population structure and was done for the RI line only. The number of generations for the L line was insufficient and so did not enable to study the effects of population structure.

5) The effect of the size of reference population on imputation accuracy was studied by comparing scenarios (B) - (C) and scenarios (D) - (E) - (F), adding individuals from previous generations in the reference population in some of the scenarios.
6) The study of the effect of the degree of kinship between reference population and candidate population was conducted by comparing scenarios (B) - (C) - (G) and scenarios (D) - (F) - (H) - (I).
7) Finally, the effect of the presence or absence of the dams in the reference population was investigated on scenario (D), by taking into account or not taking into account the dams in the reference population.

Software

FImpute V2.2 [4] was used to impute the high density genotyping of the selection candidates. FImpute is a software which was developed for livestock species and which uses pedigree information. It relies on overlapping sliding windows methodology to achieve imputations. Others imputation software like Beagle or AlphaImpute, which have also been reported in the literature, were tested on scenario (A), together with FImpute. However, given the relatively long execution time of Beagle (half a day) and AlphaImpute (1 week), in comparison to FImpute, only FImpute was subsequently used.

Moreover, because only the sires were present in the reference population, FImpute was used with the option *"turnoff_fam"* activated. This option enabled the software to turn off family imputation and to use the whole range of haplotypes of the reference population to achieve imputation. The information brought by the sire was not used. There was however one exception, which concerned scenario D_2, where both sires and dams were present in G2. In that case, the analysis was carried out without activating the *"turnoff_fam"* option.

Imputation accuracy

Following the suggestion of Hickey et al. [25] and Calus et al. [26], imputation accuracy was assessed as the mean correlation between true and imputed genotypes. Indeed, for one SNP, the correlation was not dependent on MAF and could be used to assess imputation accuracy, rather than using genotype and allelic imputation error rates. Correlations were calculated one SNP at a time for all the candidates, as suggested in Pearson's method. Mean correlation was then estimated on 300,351 correlations for the RI line, and on 245,669 correlations for the L line. The mean correlations obtained were subsequently compared for the different low density SNP chips and/or scenarios, using Student tests with type 1 error rate of 0.1%.

Results

The influence of the parameters, i.e. marker density, LD threshold, MAF of imputed SNPs, chromosome type, and composition of the reference population through the cumulative use of generations in the reference population, degree of kinship between reference population and candidate population and the effect of using dams' genotypes in the reference population, was investigated.

Influence of marker density

The evolution of mean correlations between true and imputed genotypes according to the number of SNPs on the low density SNP chips was studied on scenario (A), for both lines and both methodologies (Fig. 2).

For both lines and for both methodologies, an increase in mean correlations was observed when the number of SNPs on the low density SNP chips was increased. Regarding the RI line and the EQ methodology, the mean correlation was 0.875 for 2992 SNPs and 0.973 for 19,910 SNPs. Regarding the same line and the LD methodology, the mean correlation was 0.893 for 3205 SNPs and 0.977 for 16,615 SNPs. An inflexion point was also noticed between 5000 SNPs and 10,000 SNPs.

In addition, for both methodologies, the growth rate of the mean correlation was 0.004 for 3000 SNPs, which means that adding 100 SNPs on a low density SNP chip of 3000 SNPs would significantly increase the mean correlation of 0.004. Furthermore, the growth rate of the mean correlation was $7.0*10^{-5}$ for 20,000 SNPs (Fig. 3), which was not significant.

Finally, given the fact that the inflexion point was between 5000 and 10,000 SNPs, a density of 10K SNPs enabled to reach steady and good imputation accuracy and was subsequently used throughout the rest of the present study. A density of 3K SNPs was also considered, in order to investigate the consequences that would result from a deteriorated, but nonetheless correct, imputation accuracy with mean correlations above 0.870.

Influence of LD threshold

The evolution of imputation accuracy as a function of LD threshold was studied on scenario (A) for both lines (Fig. 4). When the LD threshold used for the selection of representative SNP was increased, an increase in imputation accuracy was observed. For the RI line, the LD thresholds of 0.05, 0.5 and 0.8, respectively resulted in mean correlations of 0.893, 0.966 and 0.981. For the L line, the mean correlations were respectively 0.902, 0.978 and 0.990.

Influence of MAF of imputed SNPs

The influence of minor allele frequencies of imputed SNPs was studied on scenario (A), using both methodologies. The 10Kequi and the LD0.5 SNP chips were used for the RI line and the 10Kequi and the LD0.6 SNP chips were used for the L line. The influence of minor allele frequencies was also studied for a density of 3K SNPs, using the 3Kequi and LD0.05 SNP chips for the RI line (LD0.1 for the L line). The results for the density of 10K SNPs are the only ones shown, since they are similar to those obtained with the density of 3K SNPs. The same results were obtained for both lines and the results for the RI line are illustrated in Fig. 5.

With the 10Kequi SNP chip, an increase in imputation accuracy was noticed when the MAF of imputed SNPs was increased (Fig. 5a). Comparatively, more steady correlations were observed when MAF was increased using the LD methodology (Fig. 5c). Moreover, mean correlations were higher with the LD SNP chip than they were with the 10Kequi SNP chip. The variability of mean correlations according to MAF was also higher with the 10Kequi SNP chip than with the LD SNP chip.

Finally, by looking the MAF distribution of the SNPs of the different low density SNP chips, the SNPs of the 10Kequi SNP chip had mostly high MAF (Fig. 5b), whereas the SNPs of the LD0.5 SNP chip had both low and high MAF (Fig. 5d).

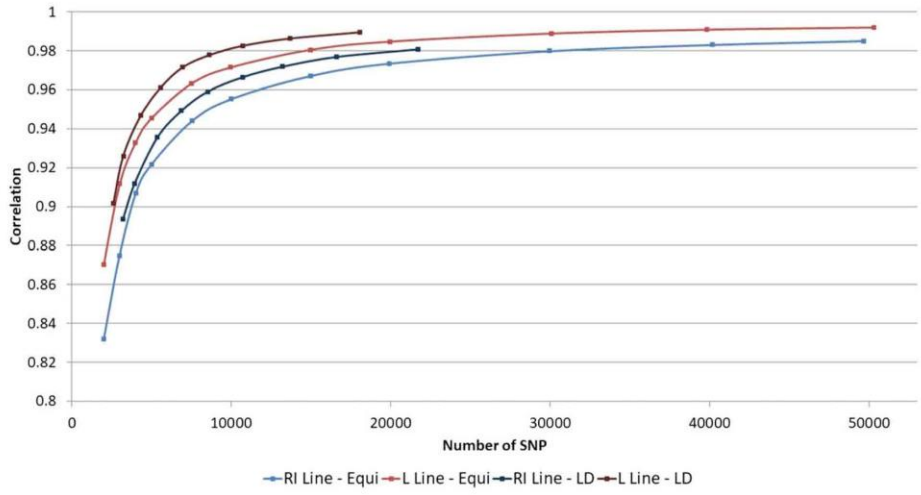


Fig. 2 Evolution of mean correlations between true and imputed genotypes according to the number of SNPs on the low density SNP chips for both methodologies and both lines

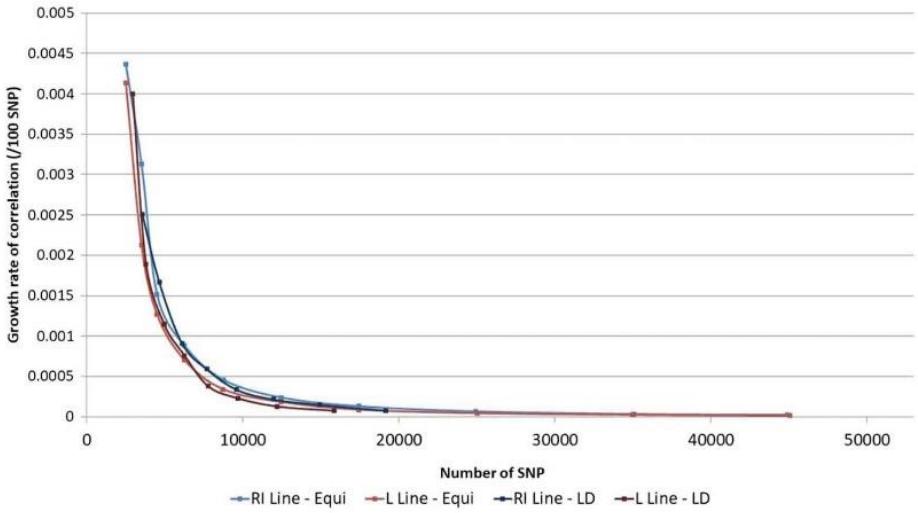


Fig. 3 Evolution of the growth rate of mean correlations (/100 SNP) according to the number of SNPs on low density SNP chips for both methodologies and both lines

Influence of the type of chromosome

The influence of the type of chromosome was studied on scenario (A), for both lines, with the two low density 10K SNPs, i.e. 10Kequi and LD0.5 chips for the RI line (LD0.6 for the L line). The influence of the type of chromosome was also studied at a density of 3K SNPs, using the 3Kequi and LD0.05 chips, for the RI line (LD0.1 for the L line). The results for the density of 10K SNPs are the only ones shown, since they are similar to those obtained with the density of 3K SNPs. Chromosomes were split into four different groups: macro-chromosomes (1 to 5), intermediate chromosomes (6 to 10), micro-chromosomes (11 to 28 and 33), and sexual chromosome Z [15].

Regarding the RI line (Fig. 6a) in conjunction with the EQ methodology (10Kequi SNP chip), mean correlations varied depending on the type of chromosome. When using the 10Kequi SNP chip, the mean correlations were 0.963 for macro-chromosomes, 0.953 for intermediate chromosomes, and 0.893 for micro-chromosomes. The differences in mean correlation were significant. As regards the LD0.5 SNP chip, the mean correlations were 0.963 for macro-chromosomes, 0.965 for intermediate chromosomes and 0.968 for micro-chromosomes. These differences in mean correlation were also significant and, except for macro-chromosomes, the differences between the 10Kequi SNP chip and the LD0.5 SNP chip were significant. As far as the LD0.5 SNP chip is concerned,

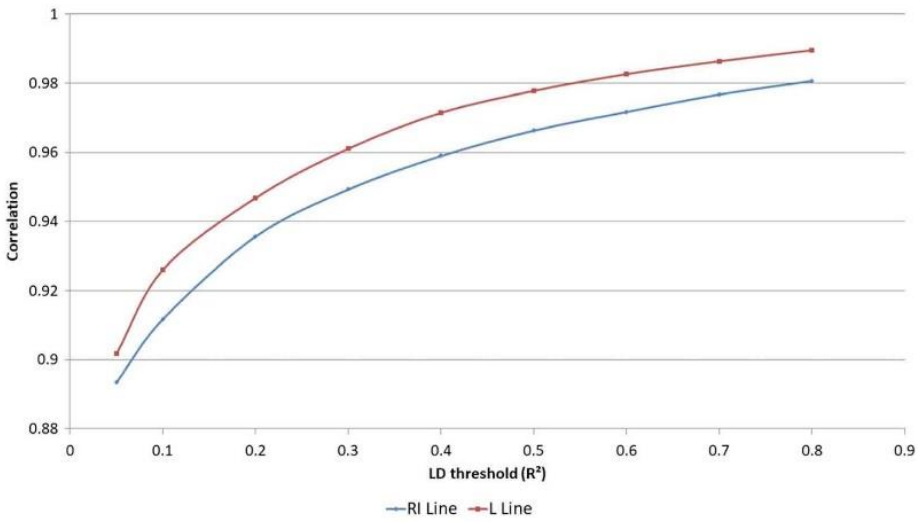


Fig. 4 Evolution of mean correlations of true and imputed genotypes according to the LD threshold used to design the low density SNP chips, for both lines

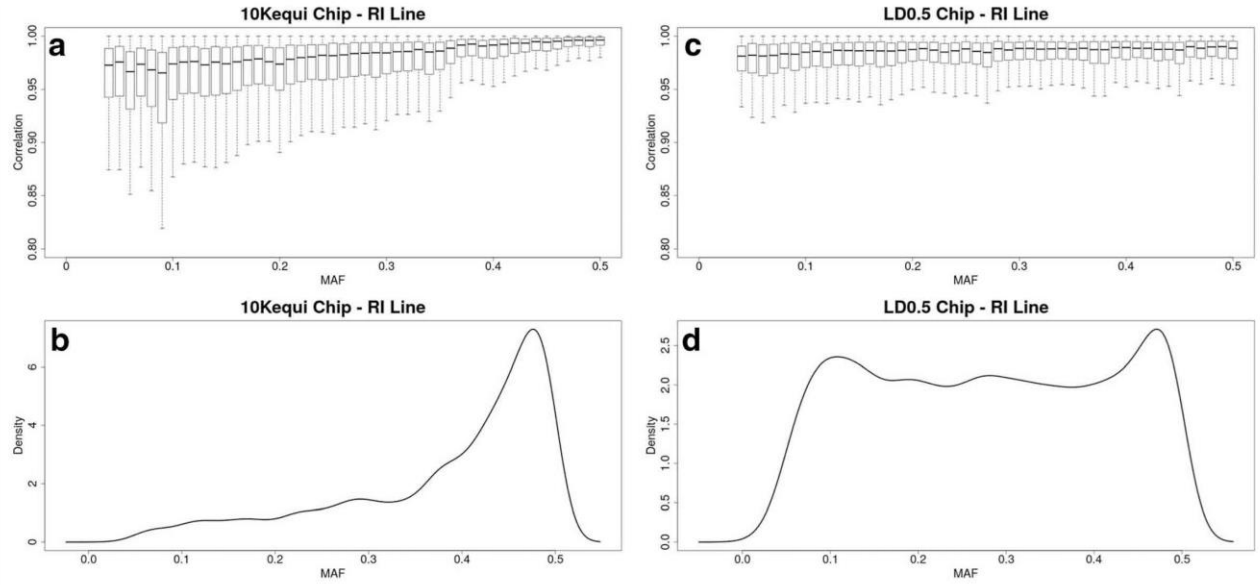


Fig. 5 For the RI line, evolution of the correlations between true and imputed genotypes according to the MAF of imputed SNPs on the 10Kequi (**a**) and LD0.5 (**c**) SNP chips and distribution of the SNPs on the 10Kequi (**b**) and LD0.5 (**d**) SNP chips according to MAF

imputation accuracy was rather steady regardless of the type of chromosome. Another point is that the standard error varied greatly depending on the type of chromosome, when the 10Kequi chip was used. This was not the case with the LD0.5 SNP chip, which showed stable and low variance for all types of chromosome. Finally, the results for sexual chromosome Z proved better with the EQ methodology than with the LD methodology. As far as the L line is concerned, similar observations were made regarding the

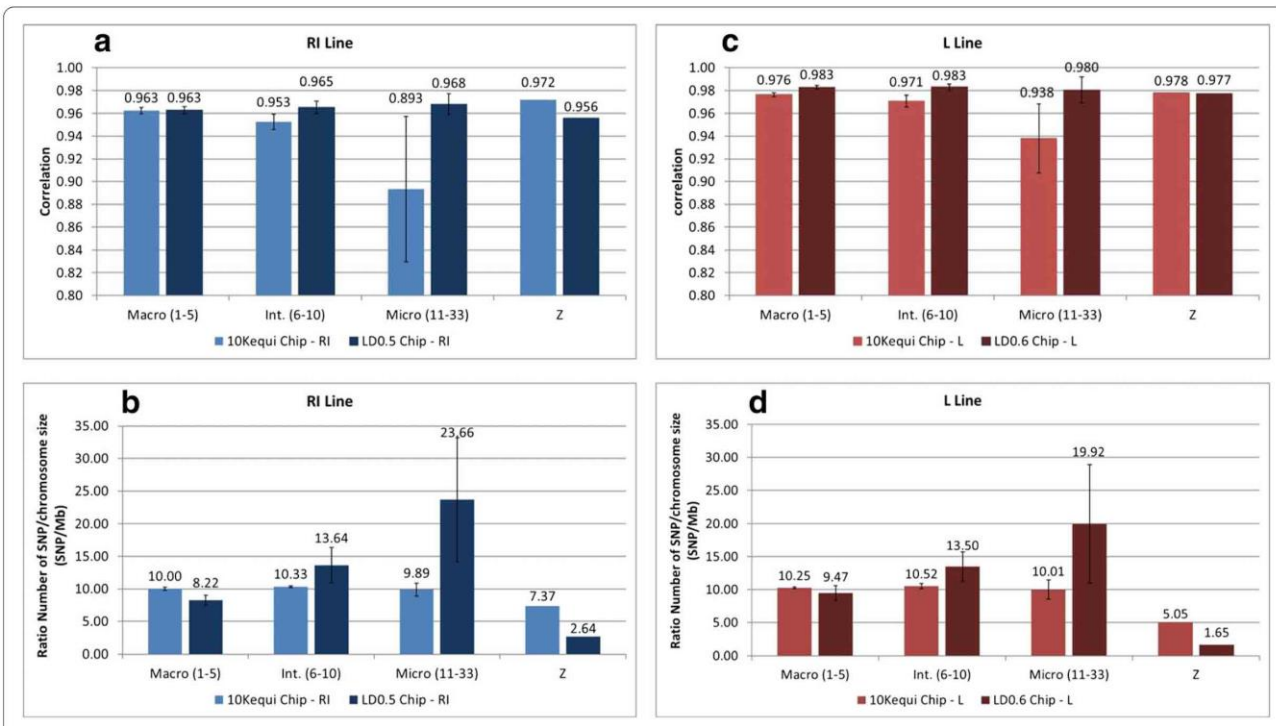


Fig. 6 Evolution of mean correlations between true and imputed genotypes depending on the type of chromosome, for the RI line (**a**) and the L line (**c**), and evolution of the ratio number of SNP/Size of chromosome ($SNP.Mb^{-1}$) depending on the type of chromosome, for the RI line (**b**) and the L line (**d**)

EQ and LD methodologies, except for sexual chromosome Z (Fig. 6c).

To understand the performances of each low density SNP chip, and in turn to understand the evolution of mean correlations on each low density SNP chip, the ratio of selected SNPs on the low density SNP chips, per chromosome size ($SNP.Mb^{-1}$), was studied for each type of chromosome. For the RI line (Fig. 6b) in conjunction with the 10Kequi SNP chip, the ratios were steady regardless of the type of chromosome, with a ratio of 10.00 ± 0.23 $SNP.Mb^{-1}$ for macro-chromosomes, 10.33 ± 0.13 $SNP.Mb^{-1}$ for intermediate chromosomes and 9.89 ± 1.00 $SNP.Mb^{-1}$ for micro-chromosomes. For the L line (Fig. 6d) in conjunction with the 10Kequi SNP chip, the ratios were also steady and close to those of the RI line. Finally, in the case of sexual chromosome Z, the ratio was 7.37 $SNP.Mb^{-1}$ for the RI line and 5.05 $SNP.Mb^{-1}$ for the L line.

Regarding the LD0.5 SNP chip (for the RI line) and the LD0.6 SNP chip (for the L line), the ratio was 8.22 ± 0.76 $SNP.Mb^{-1}$ and 9.47 ± 1.18 $SNP.Mb^{-1}$ for macro-chromosomes, 13.64 ± 2.70 $SNP.Mb^{-1}$ and 13.50 ± 2.23 $SNP.Mb^{-1}$ for intermediate chromosomes, and 23.66 ± 9.51 $SNP.Mb^{-1}$ and 19.92 ± 8.94 $SNP.Mb^{-1}$ for micro-chromosomes. As regards sexual chromosome Z, the ratio was 2.64 $SNP.Mb^{-1}$ for the RI line and 1.65 $SNP.Mb^{-1}$ for the L line.

Influence of the size of the reference population

The influence of the size of the reference population on imputation accuracy was studied by adding individuals from previous generations in the reference population. The study was conducted for the RI line, using both methodologies. Two different cases were considered, with the imputation of G2 and G3 generations respectively as candidate populations (Table 4). The different scenarios were studied at an equivalent SNP density of 3K SNPs, for the 3Kequi and the LD0.05 SNP chips, and at an equivalent SNP density of 10K SNPs, for the 10Kequi and the LD0.5 SNP chips. In the case of the imputation of G2, the influence of the size of the reference population was studied by comparing between 580 sires of G1 and 1027 sires of (G0 + G1) as the reference population (scenarios (B) and (C)). In the case of the imputation of G3, the influence of the size of the reference population was studied by comparing between 73 sire breeders of G2, 653 sires of (G1 + male breeders of G2) and 1100 sires of (G0 + G1 + male breeders of G2) as the reference population (scenarios (D_1), (E) and (F)).

From scenario (B) to scenario (C), at an equivalent SNP density of 3K SNPs, the mean correlations increased from 0.884 to 0.914 for the 3Kequi SNP chip, and from 0.899 to 0.921 for the LD0.05 SNP chip. At an equivalent SNP density of 10K SNPs, the mean correlations increased from 0.961 to 0.973 for the 10Kequi SNP chip, and from 0.975 to 0.981 for the LD0.5 SNP chip. Similarly, from scenario (D_1), to (E) and (F), the increase in imputation accuracy was significant and went from 0.868 to 0.896 to 0.912 for the 3Kequi SNP chip, from 0.953 to 0.965 to 0.974 for the 10Kequi SNP chip, from 0.892 to 0.914 to 0.929 for the LD0.05 SNP chip and from 0.973 to 0.978 to 0.983 for the LD0.5 SNP chip.

Influence of the degree of kinship between reference population and candidate population

The influence of the degree of kinship on imputation accuracy was studied on the RI line, for both methodologies and in two different cases corresponding to the imputation of G2 and G3 (Table 5). The different scenarios were studied at an equivalent SNP density of 3K SNPs, for the 3Kequi and the LD0.05 SNP chips, and at an equivalent SNP density of 10K SNPs, for the 10Kequi and the LD0.5 SNP chips. Regarding G2 imputation, a decrease in the degree of kinship was achieved from scenario (B), with 580 sires of G1 as the reference population, to scenario (G), with 447 sires of G0 as the reference population. A gap of one generation was thereby created between the reference population and the candidate population in scenario (G). As far as G3 imputation is concerned, a decrease in the degree of kinship was achieved, starting from scenario (D_1), with 73 male breeders of G2 as the reference population, to scenario (H), with 120 male breeders of G1 as the reference population, and scenario (I), with 132 male breeders of G0 as the reference population. A gap of one generation was created between the reference population and the candidate population in

Table 4 Evolution of mean correlations between true and imputed genotypes according to the size of the reference population, with the imputation of G2 and G3 as candidate population, for the RI line

Ref. pop.	G2 Imputation				G3 Imputation					
	G1 (B)		G0G1 (C)		G2♂ (D1)		G1G2♂ (E)		G0G1G2♂(F)	
SNP chip	Corr.	SE	Corr.	SE	Corr.	SE	Corr.	SE	Corr.	SE
10Kequi	0.961	0.063	0.973	0.047	0.953	0.076	0.965	0.056	0.974	0.043
3Kequi	0.884	0.113	0.914	0.085	0.868	0.136	0.896	0.103	0.912	0.087
LD0.5	0.975	0.045	0.981	0.038	0.973	0.047	0.978	0.039	0.983	0.032
LD0.05	0.899	0.057	0.921	0.083	0.892	0.107	0.914	0.081	0.929	0.068

Table 5 Evolution of the mean correlations between true and imputed genotypes according to the degree of kinship between reference population and candidate population, with the imputation of G2 and G3 as candidate population

Ref. pop.	G2 Imputation				G3 Imputation					
	G1 (B)		G0 (G)		G2♂ (D1)		G1♂ (H)		G0♂(I)	
SNP chip	Corr.	SE	Corr.	SE	Corr.	SE	Corr.	SE	Corr.	SE
10Kequi	0.961	0.063	0.952	0.072	0.953	0.076	0.936	0.098	0.940	0.089
3Kequi	0.884	0.113	0.841	0.149	0.868	0.136	0.811	0.172	0.821	0.167
LD0.5	0.975	0.045	0.973	0.046	0.973	0.047	0.965	0.057	0.967	0.051
LD0.05	0.899	0.057	0.872	0.119	0.892	0.107	0.852	0.131	0.856	0.123

scenario (H). A gap of two generations was created in scenario (I).

Regarding G2 imputation, which goes from scenario (B) to scenario (G), mean correlations decreased from 0.884 to 0.841 for the 3Kequi SNP chip, and from 0.961 to 0.952 for the 10Kequi SNP chip. For the LD0.05 SNP chip, imputation accuracy decreased from 0.899 to 0.872, whereas for the LD0.5 SNP chip, it decreased from 0.975 to 0.973. The differences in mean correlations were therefore significant.

Regarding G3 imputation, which goes from scenario (D_1) to scenario (H), imputation accuracy decreased from 0.868 to 0.811 for the 3Kequi chip and from 0.953 to 0.936 for the 10Kequi SNP chip. For the LD0.05 and LD0.5 SNP chips, imputation accuracy respectively decreased from 0.892 to 0.852 and from 0.973 to 0.965. The differences in mean correlations were significant. However, by further increasing the gap to two generations (scenario (I)), imputation accuracy became a little bit higher, compared to the scenarios with a gap of one generation (H). The differences in mean correlations were significant, with imputation accuracy values of 0.821 for the 3Kequi SNP chip, 0.940 for the 10Kequi SNP chip, 0.856 for the LD0.05 SNP chip, and 0.967 for the LD0.5 SNP chip.

Influence of dams genotyping

The influence of the information regarding dams on imputation accuracy was studied on the RI line, using both methodologies and by comparing scenario (D_1) and scenario (D_2) (Table 6). The different scenarios were studied at an equivalent SNP density of 3K SNPs, for the 3Kequi and the LD0.05 SNP chips, and at an equivalent SNP density of 10K SNPs, for the 10Kequi and the LD0.5 SNP chips. In these scenarios, G3 was the candidate population. As far as the reference population is concerned, only the 73 male breeders of G2 were taken into account for scenario (D_1). For scenario (D_2), the 73 male breeders and 662 female breeders of G2 were taken into account. In addition, in scenario (D_2), imputation was done without activating the "*turnoff_fam*" option, because of the presence of sires and dams in the reference population. This enabled FImpute to use both pedigree and haplotype diversity to achieve imputations.

Regarding scenario (D_1), in which dams' genotype was not taken into consideration, imputation accuracy was 0.868 for the 3Kequi SNP chip, 0.953 for the 10Kequi SNP chip, 0.892 for the LD0.05 SNP chip and 0.973 for the LD0.5 SNP chip. When adding dams' genotype (D_2), imputation accuracy increased to 0.946 for the 3Kequi SNP chip, 0.983 for the 10Kequi SNP chip, 0.953 for the LD0.05 SNP chip and 0.989 for the LD0.5 SNP chip. The differences in mean correlations were significant.

Discussion

Influence of marker density

For both lines and for both methodologies, an increase in mean correlations was observed when the number of SNPs on the low density SNP chips was increased. It was also concluded to better imputation accuracy with the LD methodology, which required less SNPs compared to the EQ methodology (for instance 16,615 SNPs versus 19,910 SNPs) for similar correlation. An inflexion point was also noticed between 5000 SNPs and 10,000 SNPs.

These results are in line with those found in the literature [7, 27], where better imputations were achieved when the number of SNPs was increased. This greater number of SNPs on the low density SNP chips, results in an increased number of genotypes present to identify the corresponding reference haplotypes. As a consequence, the probability of randomly identifying haplotypes common to the reference and candidate populations decreases.

Table 6 Evolution of the mean correlations between true and imputed genotypes depending on the presence or absence of dams in the reference population

Ref. pop.	G3 Imputation			
	G2♂ (D1)		G2 (D2)	
SNP chip	Corr.	SE	Corr.	SE
10Kequi	0.953	0.076	0.983	0.036
3Kequi	0.868	0.136	0.946	0.068
LD0.5	0.973	0.047	0.989	0.026
LD0.05	0.892	0.107	0.953	0.024

Influence of LD threshold

For both lines, an increase in mean correlation was observed when the LD threshold increased. The rise in mean correlations, linked to the increase in the LD threshold, can be explained by the way the selection of SNPs was done, that is by clustering the whole set of high density SNPs according to their pairwise LD. When the LD threshold is high (0.8 for instance), the number of clusters of SNPs is also high, because few pairs of SNPs are in very strong LD with each other. A great number of SNPs are therefore present on the low density SNP chips (Fig. 4). Conversely, when the LD threshold is lower (0.5 for instance), the clusters previously formed (on the basis of a LD threshold of 0.8) may become aggregated. This in turn reduces the number of SNP clusters and, as a consequence, the number of SNPs on the low density SNP chips. However, the number of SNPs on the low density SNP chip was not proportional to the LD threshold. In addition, as previously outlined, imputation accuracy decreases when the number of SNPs on the low density SNP chips is reduced.

Finally, when observing the evolution of imputation accuracy as a function of LD threshold, the inflexion point was much less distinct than it was when observing the evolution of imputation accuracy as a function of the number of SNPs on the low density SNP chips.

Influence of MAF of imputed SNPs

Regarding the EQ methodology, the increase in imputation accuracy related to MAF was expected [25, 26]. Correlations between true and imputed genotypes are more sensitive to false imputation for SNPs with low MAF than for SNPs with high MAF. Because of the way the chip was built, the SNPs of the 10Kequi SNP chip had mostly high MAF (Fig. 5b), whereas the SNPs of the LD0.5 SNP chip had both low and high MAF (Fig. 5d), which favored a better imputation of haplotypes with a low MAF when the LD methodology was used.

Influence of the type of chromosome

The results showed that better imputation accuracy was obtained with the LD0.5 SNP chip than with the 10Kequi SNP chip, except for the macro-chromosomes (no significant differences) and for sexual chromosome Z, where the ranking of the SNP chips was reversed. However, the results for sexual chromosome Z proved better with the EQ methodology than with the LD methodology.

For each chromosome, and consequently for each type of chromosome, the results needed to be related to the ratio of selected SNPs on the low density SNP chips, per chromosome size (SNP.Mb^{-1}). This enabled to understand the performances of each low density SNP chip, and in turn to understand the evolution of mean correlations on each low density SNP chip.

For the RI line (Fig. 6b) in conjunction with the 10Kequi SNP chip, the ratios were steady and consistent with the methodology used and with the size of the genome (approximately 1 Gb). Indeed, once the distance between the SNPs was defined, the ratio (SNP.Mb^{-1}) was stable and the number of SNPs on the SNP chip was proportional to the size of the chromosomes. However, in the specific case of sexual chromosome Z, the significant differences with the expected ratio of 10 SNP.Mb^{-1} are due to the number of SNPs kept on chromosome Z after quality control. This number was 38% for the RI line and only 18% for the L line. In addition, the distribution of the SNPs kept on chromosome Z was non-homogeneous (particularly on the L line), which resulted in large intervals devoid of SNPs (Fig. 7).

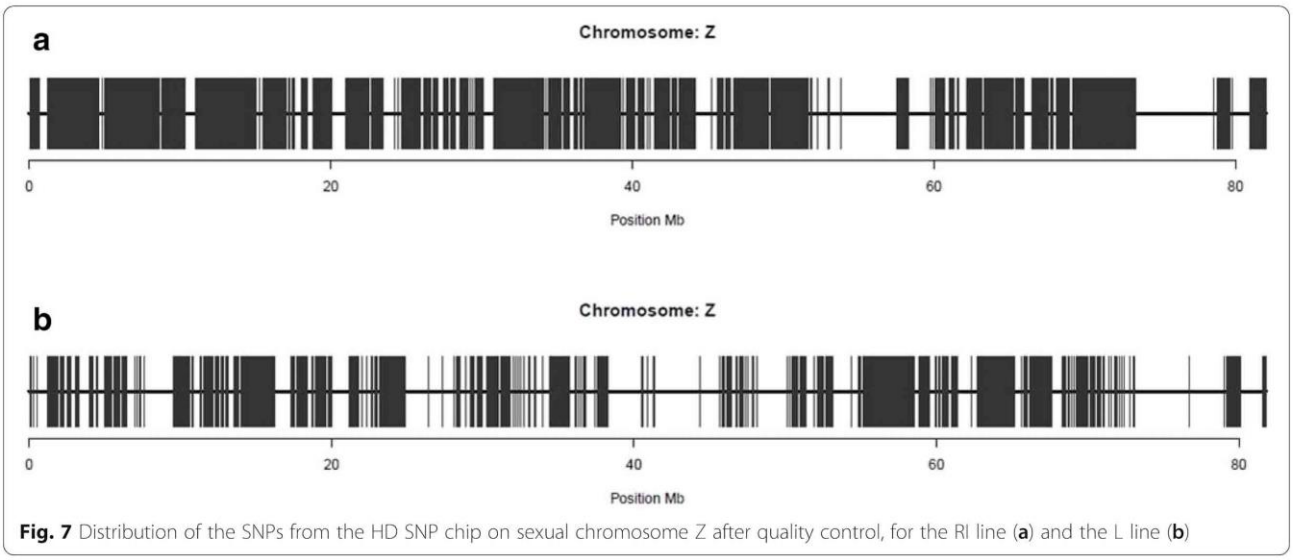


Fig. 7 Distribution of the SNPs from the HD SNP chip on sexual chromosome Z after quality control, for the RI line (**a**) and the L line (**b**)

Regarding the LD0.5 SNP chip (for the RI line) and the LD0.6 SNP chip (for the L line), the ratio varied depending on the type of chromosome, and the variation was higher than the variation observed with the equidistance based approach.

With the LD methodology, the specific structure of the LD in layer chickens was taken into account and the number of SNPs on each chromosome was not proportional to the size of the chromosomes. According to Robert et al. [18], for a fixed LD threshold, the extent of LD is higher on macro-chromosomes than it is on micro-chromosomes. Because of this high extent of LD on macro-chromosomes, few SNPs are needed to cover macro-chromosomes. Comparatively, the lower extent of LD on micro-chromosomes results in more SNPs necessary to cover micro-chromosomes. The LD methodology therefore enables to decrease the ratio on macro-chromosomes, in order to further optimize the number of SNPs on the low density SNP chips. Conversely, it enables to increase the ratio on intermediate and micro-chromosomes, in order to further densify the number of SNPs on the low density SNP chips. In addition, the recombination process breaks the LD and creates new haplotypes that occur more frequently on micro-chromosomes than on macro-chromosomes [28]. Overall, this resulted in better imputations with the LD SNP chips than with the EQ SNP chips for all type of chromosome and for both lines.

Finally, for both lines, better imputation accuracies were obtained with the LD methodology than with the EQ methodology. In addition, higher imputation accuracies were obtained for the L line, compared to the RI line. This was expectable because of the number of SNPs retained and distributed over the genome after quality control in both lines. The number of SNPs was greater for the RI line (300,351 SNPs) than for the L line (245,669 SNPs). The L line has therefore less polymorphic markers and can be better imputed, compared to the RI line [29].

Influence of the size of the reference population

The increase in the size of the reference population, achieved by adding individuals from previous generations, resulted in an increase in mean correlations. Indeed, increasing the size of the reference population led to an increase in the size of the library of reference haplotypes. The probability of finding haplotype fragments of the candidate in the library of reference haplotypes was therefore increased. These results are consistent with those found in the literature [10, 11, 13].

Finally, the ranking of the methodologies remained unchanged despite the increase in the size of the reference population: the LD0.05 and the LD0.5 SNP chips respectively achieved better results than the 3Kequi and the 10Kequi SNP chips.

Influence of the degree of kinship degree between reference population and candidate population

The gap of one generation introduced in scenarios (G) and (H), as well as the gap of two generations introduced in scenario (I), led to a decrease in the degree of kinship between the reference population and the candidate population. The downward trend in imputation accuracy in scenarios including a gap of one generation (G2 and G3 imputation) was due to a decrease in the degree of kinship between reference population and candidate population. This in turn led to a decrease in the size of haplotype fragments that are common to the reference population and the candidate population. The decrease in the size of haplotype fragments can be explained by the recombination process that occurs over the generations. Selection candidates have therefore smaller haplotype fragments in common with the reference population. The probability of mistakenly identifying a haplotype fragment common to reference population and candidate population is consequently increased, which results in a lower number of good imputations. Moreover, the decrease in imputation accuracy was higher at a density of 3K SNPs than it was at a density of 10K SNPs. Indeed, in the case of G2 imputation, there was a decrease of respectively 0.043 and 0.027 for the 3Kequi and the LD0.05 SNP chips, and a decrease of respectively 0.009 and 0.002 for the 10Kequi and the LD0.5 SNP chips. Likewise, in the case of G3 imputation, there was a decrease of respectively 0.057 and 0.040 for the 3Kequi and the LD0.05 SNP chips and a decrease of respectively 0.017 and 0.008 for the 10Kequi and the LD0.5 SNP chips. The influence of the generation gap between the reference population and the candidate population was more important at a density of 3K SNPs than it was at a density of 10K SNPs. The greater decrease in imputation accuracy at a density of 3K SNPs can be explained by the fact that the SNP density was not sufficient to compensate for the loss of imputation accuracy caused by the generation gap.

Moreover, in the case of G3 imputation, which goes from scenario (D_1) to scenario (H), one can notice that the increase in the size of the reference population did not enable to get better imputation, in spite of the generation gap. For scenario (D_1) the reference population was comprised of only 73 sires. In scenario (H), it was comprised of 120 sires from G1. Therefore, the loss of the information brought by the direct sires was not counterbalanced by the increase in the size of the reference population.

This observation still held true when the generation gap was further increased to two generations (scenario (I)). In that case, for both methodologies and for each low density SNP chip, a significant increase in mean correlations was noticed, compared to scenario (H). This improvement in imputation accuracy was due to the

increase in the size of the reference population. However, these results were still lower than the mean correlations obtained with scenario (D_1). With 132 sires from G0 (scenario (I)) and 120 sires from G1 (scenario (H)), the increase in the size of the reference population did not counterbalance the loss of information brought by the direct sires. Consequently, a key point to get good imputation accuracy is to include the direct parents, or at least the direct sires, of the candidate population.

Finally, the ranking of the methodologies remained unchanged, despite the decrease in the degree of kinship: the LD0.05 and LD0.5 SNP chips respectively achieved better results than the 3Kequi and 10Kequi SNP chips. In addition, in the cases of G2 and G3 imputation, and regardless of the SNP density, the decrease in imputation accuracy was less important with the LD methodology than it was with the EQ methodology. LD methodology is less sensitive to the degree of kinship, since LD does not drop very quickly through generations.

Influence of dams genotyping

The contribution of the presence of dams in the reference population led to very high imputation accuracy. Indeed, by having both the direct sire and the direct dam of a selection candidate in the reference population, paternal and maternal haplotypes of the candidate will show in the haplotypes library. This in turn increases the probability of getting the complete genotyping of the candidate. However, it is difficult to know precisely whether the increase in imputation accuracy is due to the increase in the size of the reference population or to the presence of the dams in the reference population. As previously seen, one of the key points in imputation is to include direct sires in the reference population. One can go further by saying that it is important to have both direct sires and direct dams in the reference population, in order to get good imputation.

In addition, the results for the 3Kequi and the LD0.05 SNP chips were higher than those obtained using the same low density SNP chips, but with three generations included in the reference population (F). Therefore, when the SNP density is very low, a better alternative would be to genotype both dams and sires to achieve good imputation accuracy, rather than genotyping individuals from previous generations.

Finally, once again, the ranking of the methodologies remained unchanged when the dams were included in the reference population: the LD0.05 and the LD0.5 SNP chips still respectively achieved better results than the 3Kequi and the 10Kequi SNP chips.

Conclusions

The above studies showed that, whatever the SNP chip used, the methodology and the scenario studied, highly accurate imputations were obtained, with mean correlations higher than 0.83. These studies also highlighted two key points allowing for good imputation results. The first one, related to SNP chip factors, is the necessity to take into consideration the particular structure of the LD of chicken species. Indeed, each time the two methodologies were compared, better results were obtained with the LD methodology. In particular, when studying the type of chromosome, except for sexual Z chromosome (for both lines), better imputation accuracies were obtained with the LD methodology. More precisely, this methodology enabled to optimize the number of SNPs on macro-chromosomes and to densify the number of SNPs on intermediate and micro-chromosomes. The second key point, related to the influence of population structure, is to include the direct parents, or at least the direct sires, of the candidate population in the reference population. Indeed, it was shown that the contribution of the direct parents (or sires) was more important than the contribution of the size of the reference population. For an equivalent quantity of information, the 10K SNPs chips achieved better results than the 3K SNPs chips. However, the results proved that the loss of imputation accuracy noticed in the case of 3K SNPs (compared to the results obtained with 10K SNPs) could be largely compensated by genotyping both the dams and the sires of the candidate population. Consequently, the choice of a very low density SNP chip will have to be considered, if new technologies are implemented with a reduction of the cost of this type of SNP chip.

Finally, the objective of genetic selection is to choose the most suitable individuals for the traits studied. The results of the genomic evaluations from all the different imputations strategies will be studied, in order to determine and finalize the best strategy to implement for low density genotyping of laying hen lines.

Abbreviations

EQ: Equidistance; HD: High density; L: Leghorn; LD: Linkage disequilibrium; MAF: Minor allele frequency; RI: Rhode Island; SNP: Single nucleotide polymorphism

Funding

This research project was supported by the French national research agency ANR, within the framework of project ANR-10-GENOM_BTV-015 UtOpIGe. FlH is a PhD fellow supported by the poultry breeding company Novogen, and taking part in a CIFRE thesis N°2016/0804 shared between PEGASE INRA's unit, Agrocampus Ouest and Novogen.

Availability of data and materials

The datasets used and/or analysed throughout the present study are available from the corresponding author on reasonable request.

Authors' contributions

FlH filtered the genotype data and performed the imputation analyses. AV and TB supervised animal management and production. All authors conceived the study. FrH conceived the program of SNP selection based on LD (SS4I). FlH drafted the manuscript. FrH, PLR and SA proofread the manuscript. All authors contributed to the ideas and methods. All authors read and approved the final manuscript.

Ethics approval
All the blood samples analyzed in this study were taken from the brachial veins of the animals. These animals, and the scientific investigations described herein, are therefore not to be considered as experimental animals per se, as defined in EU directive 2010/63 and subsequent national application texts. As a consequence, we did not seek ethical review and approval of this study as one including the use of experimental animals. All animals were reared in compliance with national regulations pertaining to livestock production and according to procedures approved by the French Veterinary Services.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]NOVOGEN, 5 rue des Compagnons, Secteur du Vau Ballier, 22960 Plédran, France. [2]PEGASE, INRA, Agrocampus Ouest, 16 Le Clos, 35590 Saint-Gilles, France.

Received: 6 March 2018 Accepted: 14 November 2018
Published online: 04 December 2018

References
1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819.
2. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013;14:59.
3. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007;39:906.
4. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478.
5. Browning BL, Browning SR. Genotype imputation with millions of reference samples. Am J Hum Genet. 2016;98:116.
6. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, Van Der Werf JHJ. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet Sel Evol. 2011;43:12.
7. Dassonneville R, Fritz S, Ducrocq V, Boichard D. Short communication: imputation performances of 3 low density marker panels in beef and dairy cattle. J Dairy Sci. 2012;95:4136.
8. Hozé C, Fouilloux MN, Venot E, Guillaume F, Dassonneville R, Fritz S, et al. High-density marker imputation accuracy in sixteen French cattle breeds. Genet Sel Evol. 2013;45:33.
9. Hayes BJ, Bowman PJ, Daetwyler HD, Kijas JW, Van Der Werf JHJ. Accuracy of genotype imputation in sheep breeds: genotype imputation in sheep. Anim Genet. 2012;43:72.
10. Heidaritabar M, Calus MPL, Vereijken A, Groenen MAM, Bastiaansen JWM. Accuracy of imputation using the most common sires as reference population in layer chickens. BMC Genet. 2015;16:101.
11. Ventura RV, Lu D, Schenkel FS, Wang Z, Li C, Miller SP. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. J Anim Sci. 2014;92:1433.
12. Dassonneville R, Brøndum RF, Druet T, Fritz S, Guillaume F, Guldbrandtsen B, et al. Effect of imputing markers from a low density chip on the reliability of genomic breeding values in Holstein populations. J Dairy Sci. 2011;94:3679.
13. Heidaritabar M, Calus MPL, Vereijken A, Groenen MAM, Bastiaansen JWM. High imputation accuracy in layer chicken from sequence data on a few key ancestors. In: Proceedings of the 10th World Congress on Genetics Applied to Livestock Production. Vancouver; 2014.
14. Wolc A, Kranis A, Arango J, Settar P, Fulton JE, O'Sullivan N, et al. Applications of genomic selection in poultry. In: Proceedings of the 10th World Congress on Genetics Applied to Livestock Production. Vancouver; 2014.
15. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004;432:695.
16. Megens H, Crooijmans RP, Bastiaansen JW, Kerstens HH, Coster A, Jalving R, et al. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genet. 2009;10:86.
17. Qanbari S, Hansen M, Weigend S, Preisinger R, Simianer H. Linkage disequilibrium reveals different demographic history in egg laying chickens. BMC Genet. 2010;11:103.
18. Robert R, Hérault F, Romé H, Varenne A, Chapuis H, Vignal A, et al. A linkage disequilibrium study in a layer chicken population. Tuusula: On proceedings of the 9th European symposium on poultry genetics; 2015.
19. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559.
20. Deelen P, Bonder MJ, van der Velde KJ, Westra HJ, Winder E, et al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Res Notes. 2014;7:901.
21. Warren WC, Hillier LDW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A new chicken genome assembly provides insight into avian genome structure. G3. 2017;7:109.
22. Bouquet A, Fève K, Riquet J, Larzul C. Précision de l'imputation de génotypages haute densité à partir de puces basse densité pour des individus de race pure et croisés Piétrain. Journées Rec Porcine. 2015;47:1.
23. Zhang K, Deng M, Chen T, Waterman MS, Sun F. A dynamic programming algorithm for haplotype block partitioning. Proc Natl Acad Sci USA. 2002;99:7335.
24. Hérault F, Yon J, Herry F, Allais S, Le Roy P. SS4I: select SNP subset for imputation. 2016 (in French). https://prodinra.inra.fr/record/375448.
25. Hickey JM, Crossa J, Babu R, De Los Campos G. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 2012;52:654.
26. Calus MPL, Bouwman AC, Hickey JM, Veerkamp RF, Mulder HA. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. Animal. 2014;8:1743.
27. Carvalheiro R, Boison SA, Neves HHR, Sargolzaei M, Schenkem FS, Utsunomiya YT, et al. Accuracy of genotype imputation in Nelore cattle. Genet Sel Evol. 2014;46:69.
28. Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens HJ, Crooijmans RPMA, et al. A high-density SNP-based linkage map of the chickens reveals sequence features correlated with recombination rate. Genome Res. 2009;19:510.
29. Hérault F, Herry F, Varenne A, Burlot T, Picard-Druet D, Recoquillay J, et al. A linkage disequilibrium study in layer and broiler commercial chicken populations. In: Proceedings of the 11th World Congress on Genetics Applied to Livestock Production. Auckland; 2018.

III. Discussion

A. Choix du critère de mesure de l'efficacité de l'imputation

Bien que les corrélations aient été utilisées tout au long de la thèse pour mesurer la qualité des imputations, les taux d'erreurs génotypiques et alléliques ont également été calculés dans le cas du scénario A pour la lignée RI. Les résultats sont présentés pour les puces DL0.5 et 10Kequi dans le tableau 6. Il est constaté pour les deux puces que le taux d'erreur génotypique est 1.98 fois supérieur au taux d'erreur allélique avec de meilleurs résultats pour la puce DL0.5. La majorité des erreurs d'imputations (98%) n'impacte qu'un seul des deux allèles du génotype. Or une erreur sur un seul des deux allèles est comptabilisée comme une demie-erreur avec le taux d'erreur allélique. Ceci explique donc les valeurs des deux différents taux.

En comparaison, les corrélations aboutissent aux mêmes conclusions que les taux d'erreurs mais sont moins discriminantes avec une variance plus réduite. En revanche, d'après les recommandations de Hickey et al. (2012) et de Calus et al. (2014), les corrélations sont indépendantes des fréquences alléliques des SNP à imputer. C'est pourquoi elles ont été privilégiées tout au long de la thèse et des différents travaux.

Tableau 6. Résultats des différents critères de mesure de qualité de l'imputation sur le scénario (A) pour la lignée RI.

	Taux d'erreur génotypique	Taux d'erreur allélique	Corrélation (± SE)
Puce DL0.5	2.38	1.20	0.9664 ± 0.06
Puce 10Kequi	3.01	1.52	0.9553 ± 0.04

B. Impact de la méthode d'imputation

Les logiciels FImpute (Sargolzaei et al., 2014), Beagle (Browning et Browning, 2007) et AlphaImpute (Hickey et al., 2011) ont été testés et comparés sur le scénario (A) pour les deux méthodologies avec les puces DL0.5 (Figure 24a) et 10Kequi (Figure 24b). Quelle que soit la puce étudiée, les corrélations sur l'ensemble du génome sont plus élevées avec Beagle qu'avec les autres logiciels. Pour la puce DL0.5, Beagle permet d'obtenir une corrélation moyenne de 0.974. FImpute permet ensuite d'obtenir de meilleurs résultats qu'AlphaImpute avec des corrélations respectives de 0.966 et 0.957. En revanche, pour la puce 10Kequi, il n'y a pas de différence significative entre FImpute et AlphaImpute avec des corrélations de 0.955 (0.968 pour Beagle). Dans le détail des chromosomes, Beagle permet également d'obtenir les

meilleures corrélations pour les deux puces. Pour la puce DL0.5, les corrélations obtenues avec FImpute sont plus élevées qu'avec AlphaImpute avec des corrélations respectives de 0.963 et 0.954 pour les macro-chromosomes, 0.965 et 0.958 pour les chromosomes intermédiaires, 0.968 et 0.939 pour les micro-chromosomes et 0.9556 et 0.944 pour le chromosomes Z. En revanche, pour la puce 10Kequi, AlphaImpute tient la comparaison avec FImpute sur les macro-chromosomes et chromosomes intermédiaires, mais les corrélations se dégradent plus fortement pour les micro-chromosomes avec des corrélations de 0.893 et 0.876 respectivement pour FImpute et AlphaImpute.

La dégradation plus forte des corrélations observée sur les micro-chromosomes pour les deux puces avec AlphaImpute s'explique par la méthodologie d'imputation elle-même. AlphaImpute est basée sur la méthode de phasage longue distance et d'imputation d'haplotypes longs. Les micro-chromosomes sont caractérisés par une persistance plus faible du DL ce qui peut entraîner une diminution de taille des haplotypes présents sur les micro-chromosomes. Il est alors plus compliqué pour le programme d'identifier de longs haplotypes dans sa librairie d'haplotypes de référence pour réaliser les imputations. En comparaison, la taille des fenêtres utilisées pour identifier les haplotypes peut varier avec FImpute ce qui peut expliquer les meilleurs résultats obtenus avec FImpute sur les micro-chromosomes. Bien que les résultats sur l'ensemble du génome soient équivalents, ceci rend FImpute plus intéressant qu'AlphaImpute en filière volaille.

Enfin, compte tenu des résultats obtenus sur l'ensemble des chromosomes, Beagle pourrait s'avérer le plus intéressant par rapport à FImpute. Toutefois, Beagle réalise les imputations des 580 candidats à la sélection à partir des 447 individus de référence en 24h. FImpute réalise les mêmes imputations en seulement moins de 4 minutes. AlphaImpute quant à lui, impute les données manquantes en… 1 semaine. Ces résultats sont cohérents avec la littérature. Sargolzaei et al. (2014) ont montré que l'imputation de 2000 candidats bovins à la sélection génotypés avec la puce Illumina GoldenGate 3K à partir de 10 000 individus de référence génotypés avec la puce Illumina BovineSNP50 (50K SNP) était réalisée avec FImpute en moins de 5 minutes contre plus de 25h avec Beagle. De même, Ventura et al. (2014) ont montré que l'imputation de 146 candidats bovins génotypés avec la puce BD de 7K SNP à partir de 4886 individus de référence génotypés avec la puce bovine 50K prenait moins de 6 minutes avec FImpute et plus de 7 heures avec Beagle. Enfin, Ma et al. (2013) ont montré, toujours en bovins, que l'imputation avec FImpute, Beagle et AlphaImpute du chromosome 1 de 971 individus génotypés avec la puce 3K à partir de 2931 individus génotypés avec la puce 50K était réalisée respectivement en 1min51, 7h19 et 21h20. Étant donné les temps d'exécution des différents logiciels et les corrélations obtenues, c'est FImpute qui a été utilisé tout au long de la thèse.

C’est par ailleurs le logiciel qui est utilisé en routine pour les évaluations génomiques bovines ou en volailles par le SYSAAF.

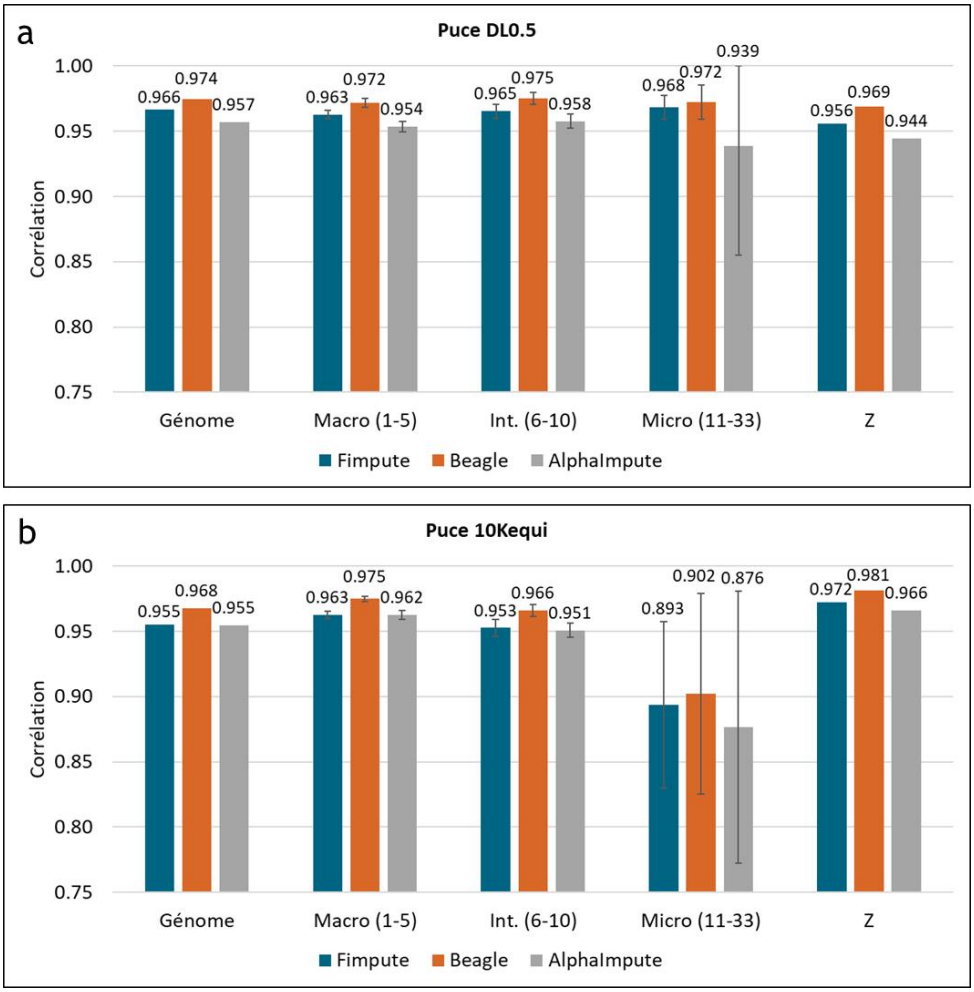


Figure 24. Corrélations entre les vrais génotypes HD et les génotypes HD imputés en fonction des types de chromosomes sur le scénario (A) pour la lignée RI et pour les différents logiciels testés. Les résultats sont présentés pour la puce DL0.5 (a) et la puce 10Kequi (b).

IV. Bilan

L’objectif de cette étude était de déterminer la stratégie d’imputation la plus adaptée pour la filière pondeuse. Cette étude a montré que, quels que soient la méthodologie et le scénario étudiés, des bonnes imputations peuvent être obtenues avec des corrélations supérieures à 0.83 à partir de seulement de 2K SNP pour déduire les génotypes 300K. Il a également été mis en évidence deux points clés pour obtenir de bonnes imputations. Le premier est la nécessité de prendre en considération la structure particulière du DL des pondeuses. Pour chaque facteur étudié, les corrélations étaient plus élevées avec la méthodologie DL. Dans le détail des chromosomes, hormis le chromosome sexuel Z, c’est également cette méthodologie qui s’avère

la plus intéressante. Ceci s'explique par la méthodologie qui permet d'optimiser le nombre de SNP sélectionnés pour les macro-chromosomes et de densifier le nombre de SNP sur les micro-chromosomes. Le deuxième point clé, relatif à la structure de la population de référence, est l'importance d'inclure dans la population de référence les parents directs ou au minimum les pères directs des candidats à la sélection. En effet, il a été montré que l'inclusion des parents directs (ou des pères) dans la population de référence était plus intéressante pour les imputations que le cumul d'individus de générations antérieures dans la population de référence. Ainsi pour une puce de 3K SNP, il a été mis en évidence que la diminution des corrélations peut être compensée par le génotypages des pères et des mères des candidats à la sélection plutôt que par l'ajout des individus des générations antérieures à la population de référence dans cette même population de référence. En conséquence, le choix d'une puce très basse densité (moins de 5K SNP) pourra devenir une option intéressante pour la sélection génomique si le coût du génotypage à cette densité continue de diminuer.

Toutefois, au-delà d'obtenir de bonnes imputations, l'objectif de la sélection génétique est de choisir les meilleurs reproducteurs pour différents caractères pour produire la génération suivante. En considérant les candidats à la sélection comme les individus de la génération G1, des évaluations génomiques peuvent être réalisées à partir des différentes puces BD utilisées ici afin de déterminer et finaliser la meilleure stratégie à mettre en place pour le génotypage BD des poules pondeuses. Ces travaux sont l'objet du chapitre suivant et de l'article 2.

Chapitre III. Intérêt de l'utilisation des génotypes issus de puces basse densité, avec ou sans imputation, pour les évaluations génomiques en poule pondeuse

I. Introduction

L'étude précédente s'est concentrée sur les facteurs influençant la qualité des imputations. Mais l'objectif de chaque sélectionneur est d'obtenir de bons résultats d'évaluations génomiques afin de pouvoir sélectionner les reproducteurs de la génération suivante. Il convient donc de mettre en relation les différentes puces BD développées avec les évaluations génomiques pour vérifier si de bonnes imputations sont synonymes de bonnes évaluations génomiques. Cette relation entre qualité d'imputation et évaluations génomiques des candidats à la sélection a été étudiée dans de nombreux travaux. En théorie, à cause des erreurs d'imputation, la précision des évaluations génomiques des candidats avec leurs génotypages HD imputés est attendue plus faible qu'avec leurs vrais génotypages HD. Ceci est confirmé dans la littérature pour des imputations à partir de puces très basse densité (de quelques SNP à 3K SNP) avec une diminution de la précision des évaluations génomiques avec une diminution, parfois limitée, de la précision des imputations (Weigel et al., 2009 ; Weigel et al., 2010a ; Mulder et al., 2012 ; Cleveland & Hickey, 2013 ; Raoul et al., 2017). En revanche, à partir de puces de densité plus élevée (entre 6K et 20K SNP), d'autres études montrent que l'impact des erreurs d'imputation est très limité (Weigel et al., 2010a ; VanRaden et al., 2012 ; Moghaddar et al., 2015). Enfin, à notre connaissance, peu d'études faisant le lien entre qualité des imputations et évaluations génomiques ont été réalisées en volaille (Wang et al., 2013). Peu d'études se sont également penchées sur les conséquences d'une utilisation directe des génotypages BD sans imputation sur la précision des évaluations des candidats à la sélection (Weigel et al., 2009 ; Su et al., 2012 ; Moghaddar et al., 2015).

L'objectif de cette étude est donc de juger de l'intérêt de l'imputation des génotypages BD pour l'évaluation génomique des coqs de la lignée RI, population étudiée dans le premier article. La lignée L n'a pas pu être étudiée car seulement deux générations ont été utilisées dans l'article précédent. Le dispositif ne permettait pas de disposer d'une génération de référence, d'une génération candidate et d'une génération de descendants avec performances de la génération candidate. Il n'était donc pas possible d'étudier la précision des évaluations génomiques pour cette lignée.

II. Article II : Intérêt de l'utilisation de l'imputation pour les évaluations génomiques en poule pondeuse

Article soumis dans Poultry Science

Herry F, Picard Druet D, Hérault F, Varenne A, Burlot T, Le Roy P, Allais S. 2019.

Design of low density SNP chips for genotype imputation in layer chicken.

Poultry Science.

IMPUTATION FOR GENOMIC EVALUATION IN LAYERS

Interest of using imputation for genomic evaluation in layer chicken

Florian Herry,[*,†] David Picard Druet,[†] Frédéric Hérault,[†] Amandine Varenne,[*] Thierry Burlot,[*] Pascale Le Roy,[†] and Sophie Allais[†,1]

[*]*NOVOGEN, Mauguerand 22800 Le Foeil, France;* †*PEGASE, INRA, Agrocampus Ouest, 16 Le Clos 35590 Saint-Gilles, France*

[1]Corresponding author: sophie.allais@agrocampus-ouest.fr

GENETICS AND GENOMICS

ABSTRACT

With the availability of the 600K Affymetrix® Axiom® high-density (HD) single nucleotide polymorphism (SNP) chip, genomic selection has been implemented in broiler and layer chicken. However, the cost of this SNP chip is too high to genotype all selection candidates. A solution is to develop low density SNP chip, at a lower price, and to impute all missing markers. But to routinely implement this solution, the impact of imputation on genomic evaluation accuracy must be studied. It is also interesting to study the consequences of the use of low density SNP chips on genomic evaluation accuracy. In this perspective, the interest of using imputation in genomic selection was studied in a pure layer line.

Two low density SNP chip design were compared: an equidistant (EQ) methodology and a methodology based on linkage disequilibrium (LD). Egg weight, egg shell color, egg shell strength and albumen height were evaluated with single-step GBLUP methodology. The impact of imputation errors or the absence of imputation on the ranking of the male selection candidates was assessed with a genomic evaluation based on ancestry. Thus, genomic estimated breeding values (GEBV), with imputed HD genotypes or low density genotypes, were compared to GEBV obtained with the HD SNP chip. The relative accuracy of GEBV was also investigated by considering as reference GEBV estimated on offspring.

A limited reordering of the breeders, selected on a multi-trait index, was observed. Spearman correlations between GEBV on HD genotypes and GEBV on low density genotypes (with or without imputation) were always higher than 0.94 with more than 3K SNPs. For the genetically closer top 150 individuals for a specific trait, with imputation, the reordering was reduced with correlation higher than 0.94 with more than 3K SNPs. Without imputation the correlations remained below 0.85 with less than 3K and 16K SNPs for EQ and LD methodology, respectively. The differences in GEBV correlations between both methodologies never were significant. The conclusions were the same for all studied traits.

Key words: Genomic selection, layer chicken, low density panel, imputation accuracy, genomic evaluation accuracy

INTRODUCTION

The availability of single nucleotide polymorphisms (SNP) enabled the development of high-throughput genotyping technologies leading to the use of the 600K Affymetrix® Axiom® high density (HD) genotyping array, a high-density genotyping chip developed by Kranis et al. in 2013, in layer and broiler breeding. Genomic selection as described by Meuwissen et al. (2001) has then been implemented in many livestock species with different statistical methods like genomic best linear unbiased prediction methods (GBLUP) (Legarra et al., 2009; Goddard et al., 2011) or Bayesian methods (Meuwissen et al., 2001; Xu, 2003; Habier et al., 2009). From a reference population with genotypes and phenotypes, it is possible to estimate the genomic value of the genotyped selection candidates with or without phenotype. The main objective is to choose among the selection candidates of generation N, the best breeders for one or more traits to produce the individuals of the generation N+1. In addition, compared to a genetic selection, genomic selection may increase the genetic gain through the decrease in generation interval, most particularly for species with high generation interval, through the increase in selection intensity by genotyping many selection candidates and through the increase in evaluation accuracy.

However, the high cost of such high density (HD) SNP chip is still a problem for all livestock species. To reduce the cost of genomic selection, low density SNP chips can be developed. The idea is to select a subset of markers from the HD SNP chip and to impute the genotypes at missing markers. Three main methods to select the marker panel have been developed: (1) selection of a subset of SNPs chosen at regular intervals along each chromosome taking into account or not the MAF of the selected SNPs (Habier et al., 2009; Weigel et al., 2009; Zhang et al., 2011; Cleveland & Hickey, 2013; Wang et al., 2013; Herry et al., 2018), (2) selection of a subset of SNPs having high effects on different traits of interest (Weigel et al., 2009, Zhang et al., 2011), or (3) selection of a subset of SNPs based on linkage disequilibrium (LD) between markers (Herry et al., 2018). This latter method was studied because of the particularities of the Gallus gallus genome (International Chicken Genome Sequencing Consortium, 2004) and the particular structure of the avian linkage disequilibrium (Megens et al., 2009; Qanbari et al., 2010; Hérault et al., 2018).

Factors influencing imputation accuracy are well documented as well as the relation between imputation accuracy and genomic evaluation of the selection candidates. Theoretically, due to imputation errors, genomic evaluation accuracy with imputed genotypes is expected to be lower than a genomic evaluation done with HD genotypes. The literature confirms it for very low density SNP chip (from few SNPs to 3K SNPs) with a decrease in genomic evaluation accuracy with a decrease, sometimes limited, in imputation accuracy

(Weigel et al., 2009; Weigel et al., 2010; Mulder et al., 2012; Cleveland & Hickey, 2013, Raoul et al., 2017). But concerning intermediate low density SNP chip (between 6K and 20K SNPs), other studies showed that the impact of imputation errors was very limited (Weigel et al., 2010; VanRaden et al., 2011; VanRaden et al., 2012; Moghaddar et al., 2015; Wang et al., 2016). However, few studies about the impact of imputation on genomic evaluation have been led on chickens (Wang et al, 2013).

In addition, several studies showed that for traits affected by few large QTL, genomic evaluations are more sensitive to imputation errors. This was shown by Habier et al. (2009) and Zhang et al. (2011) in simulation studies and confirmed by Chen et al. (2014) on real data. They showed, in Holstein bulls, that the accuracy of direct genomic value (DGV) for milk fat percentage, a trait affected by few large QTL, decreased by 34% via GBLUP using imputed genotypes. Conversely, they showed that the accuracy of DGV for the somatic cell score, a trait affected by many small QTL, decreased only by 15%. In layer chickens, most of studied traits are affected by many small QTL. This could indicate that genomic evaluation would not be severely impacted by imputation errors.

Finally, most studies investigated the impact of imputation on genomic evaluation accuracy, but only few studies focused on the impact of the use of medium density SNP chip (Su et al., 2012; Moghaddar et al., 2015) or low density SNP chip (Weigel et al., 2009; Harris & Johnson, 2010) without imputation on genomic evaluation.

The main objective of a company is to select their breeders and to describe the consequences on the loss of selection response and on genetic progress by investigating if the ranking of their best candidates would be modified with the use of low density SNP chip. Thus, focusing on four generations of a pure line of laying hens, the first objective of this study was to investigate the impact of imputation errors on genomic evaluation with an evaluation based on ancestry of the candidates of the second generation with true HD genotyping or imputed HD genotyping. The second objective was to study the impact of a direct use of low density SNP chips, without imputation, on genomic evaluation. To do so, a comparison was done between the same previous genomic evaluation of the candidates based on ancestry with true HD genotyping or with low density genotyping without imputation. Then, to get closer to the true breeding values of the candidates, their genomic estimated breeding values (GEBV) was estimated with a genomic evaluation with optimal information (phenotypes on descendants). Thus, the third objective was to assess the relative accuracy of genomic evaluation by comparing the GEBV of the candidates of the second generation with optimal information (phenotypes on their descendants of the third and fourth generations) and their GEBV based on ancestry with imputed HD genotyping. Finally, imputed HD genotyping of the candidates were

replaced by their low density genotyping without imputation. Therefore, the fourth objective was to assess the relative accuracy of genomic evaluation of the candidates without imputation.

MATERIAL AND METHODS

Ethics Statement

All blood samples were carried out as part of the commercial and selection activities of Novogen. These animals studied and the scientific investigations described herein are therefore not to be considered as experimental animals per se, as defined in EU directive 2010/63 and subsequent national application texts. As a consequence, we did not seek ethical review and approval of this study as one inclusing the use of experimental animals. All animals were reared in compliance with national regulations pertaining to livestock production and according to procedures approved by the French Veterinary Services.

Animals

All animals studied were detailed in Herry et al. (2018). They consisted in a commercial pure line of Rhode Island (RI) laying hens. This line was created and selected by Novogen (Plédran, France). The population studied was comprised of 21,475 chickens split in four generations. Each generation was divided in three batches and a new batch was bred every six months from 2010 to 2015 (Figure 1).

Concerning the laying hens, phenotypic data were recorded from 60 to 90 weeks of age, when birds where bred in individual cages. Each data collected was associated with a laying hen. There were 75,121 measures recorded for 7983 birds. Finally, the sires were bred in individual cages.

Genomic selection was implemented in 2015 on males of this line. However, females were still selected based on pedigree and performances, and not with genomic selection. Thus, this study concerned male selection candidates. In addition, among the different parameters studied and detailed in a next section, the relative accuracy of genomic selection was investigated. To calculate this relative accuracy, it is necessary to have a set of male selection candidates with information on their offspring. These male selection candidates were the 67 male breeders of the generation G1.

Genotyping

Genotyping are briefly described because detailed in Herry et al. (2018). 2370 animals were genotyped for 580,961 SNPs using the 600K Affymetrix® Axiom® HD genotyping array (Kranis et al., 2013).

Based on the fifth annotation release of Gallus gallus genome (Warren et al., 2017), these SNPs were distributed on macro-chromosomes (1 to 5), intermediate chromosomes (6 to 10), micro-chromosomes (11 to 28 and 33), one linkage group (LGE64), two sexual chromosomes Z and W, as well as a group of 3,724 SNPs with unknown location.

Genotypes were filtered through six successive steps (Table 1) including individual call rate (<95%), MAF (<0.05), SNP call rate (<95%) and Hardy-Weinberg equilibrium ($P < 10\text{-}4$). SNPs with unknown location or located on sexual chromosome W were removed, as well as the animals showing pedigree incompatibilities. Most of the SNPs had to be removed because they showed zero MAF. Finally, 300,351 SNPs and 2362 individuals remained available for the analyses.

Low Density SNP Chips Design

Several low density SNP chips were previously designed in silico by selecting a subset of SNPs (Herry et al., 2018) from the HD SNP chip.

An equidistant (EQ) methodology was studied by selecting SNPs at regular physical intervals (in pb) along each chromosome. In addition, for each interval, the SNP with the highest MAF, or the one located furthest on the left, in case of equivalent MAF, was selected. 12 low density "equi" SNP chips were designed according to this method with different SNP densities: 1K, 2K, 3K, 4K, 5K, 7.5K, 10K, 15K, 20K, 30K, 40K and 50K SNPs.

A linkage disequilibrium (LD) methodology was studied considering the particular structure of the chicken linkage disequilibrium (Robert et al., 2015). Low density SNP chips were designed using the SS4I software (Hérault et al., 2016). This software enabled to obtain clusters of SNPs according to a chosen LD threshold. For each cluster, the SNP with the highest MAF was selected and used as representative of this cluster. 9 low density "LD" SNP chips were designed with different LD thresholds: 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 and 0.8.

Imputation Accuracy

In our study, the selection candidates were the 580 sires of the second generation (G1) with simulated low density genotyping. The selection candidates were imputed from the high density genotyping of the 447 sires of the first generation (G0). These 447 individuals were the

fathers or the fathers' half-brothers of the selection candidates. Thus, the selection candidates were directly related to them.

For each low density SNP chip designed, imputation accuracy of the selection candidates was previously assessed as the mean correlation between true and imputed genotypes (Herry et al., 2018). Correlations were calculated one SNP at a time for all the candidates, as suggested in Pearson's method. The mean correlation was then estimated on 300,351 correlations. The mean correlations obtained were subsequently compared for the different low density SNP chips and/or scenarios, using Student tests with type 1 error rate of 0.1%.

Measurement of Traits

Four distinct traits were studied in this paper. They are named according to Animal Trait Ontology for Livestock (Atol Ontology, 2012). From 60 to 75 weeks, egg production was recorded each day for all individuals. There were individual data. 75,121 eggs concerning 7983 birds were measured from (G0) to (G3).

One egg was collected per layer and per week, between 60 and 75 weeks, for all layers. These eggs were then transferred at Zootests (Ploufragan, France) to study egg quality traits. The first step was to measure Egg Weight (EW, in g). Then, three traits concerning egg shell color were estimated with a Minolta Chroma Meter: redness (a*), yellowness (b*) and lightness (L*) of egg shell. Egg Shell Color (ESC) was then calculated as $ESC = 100 - (L^* - a^* - b^*)$. The next step consisted in measuring Egg Shell Strength (ESS, in N) by using a compression machine to evaluate the shell static stiffness. ESS corresponded to the maximum force recorded before fracturing the shell. Finally, each egg was broken and Albumen Height (AH) was measured using a tripod.

Genomic Evaluation Strategies

EW, ESC, ESS and AH were evaluated with single-step GBLUP methodology (Legarra et al., 2009) using BLUPF90 programs (Misztal et al., 2002).

The first part aimed to investigate the impact of imputation errors on genomic evaluations (Figure 2a). To do so, a genomic evaluation based on ancestry "Anc_HD" was done using true HD genotyping of the 447 G0 sires and selection candidates (G1), and phenotypes of the first generation (G0). A second genomic evaluation based on ancestry "Anc_Imputed" was done using the same data for the 447 G0 sires and imputed HD genotyping of the selection candidates (G1) from simulated low density SNP chips previously designed. For each low

density SNP chip and for each trait, Spearman correlations, that enabled to estimate the reordering of the selection candidates, were calculated between true "Anc_HD" Genomic Estimated Breeding Value (GEBV) and "Anc_Imputed" GEBV. Spearman correlations were calculated for the top 150 individuals from G1 according to each trait. Spearman correlations were limited to the top 150 males to better describe the consequences of imputation errors on the reordering of these individuals, and thus to better describe the consequences on the loss of selection response and on genetic progress. The objective was to identify the good candidates and to successfully rank them among themselves. We did not focus on the ranking of the less good candidates. There were also calculated for the 67 breeders from G1 having at least 10 offspring in G2.

Then, concerning the second objective, imputed HD genotyping of the candidates were replaced by their low density genotyping without imputation, allowing to simulate the impact of the direct use of the different low density SNP chips without imputation (Figure 2b). This part also implied the use of low density genotyping without imputation for the reference population. For each low density SNP chip and for each trait, Spearman correlations were calculated between the same previous true "Anc_HD" GEBV and "Anc_Not_Imputed" GEBV obtained with low density genotyping (without imputation). These correlations were calculated for the same 67 breeders of G1 and the top 150 individuals from G1 according to each trait.

The third objective was to study the attainable relative accuracy with imputation (Figure 2c). To calculate this relative accuracy, it is necessary to have a set of male selection candidates with information on their offspring. On one hand, males don't have own phenotypes and only a few of them have daughter records. Thus, information from them is limited. On the other hand, Generation 2 had 662 genotyped females with own performances and some of them with progeny records. They would provide a more reliable validation set with GEBVs using all available information fairly close to the true breeding values. However, females were still selected based on pedigree and performances, and not with genomic selection. Thus, this study focused on male selection candidates. To get closer to the true breeding values for the males, a genomic evaluation "Full_HD" of the G1 candidates was done with all available information (phenotypes and genotypes) from (G0) to (G3). These "Full_HD" GEBV leaded to closer to the true breeding values of the G1 candidates which cannot be calculated. These "Full_HD" GEBV represented the maximum of relative accuracy attainable regarding this genomic evaluation with all information and were calculated only for the 67 G1 breeders which had at least 10 offspring in G2. Then, these "Full_HD" GEBV were compared by Pearson correlations with the previous GEBV based on ancestry "Anc_Imputed" with imputed HD genotyping of the breeders, for each simulated low density SNP chip.

Finally, imputed HD genotyping of the candidates were replaced again by their low density genotyping without imputation. The "Full_HD" GEBV of the 67 G1 breeders were compared by Pearson correlations with their GEBV obtained with low density genotyping without imputation ("Anc_Not_Imputed" GEBV). The fourth objective was thus to investigate the impact of a direct use of low density SNP chips without imputation on relative accuracy of genomic evaluation (Figure 2d).

The four traits were jointly estimated according to a classical multi-trait animal model: $Y = 1\mu + X\beta + Zu + \varepsilon$. Y is a vector of the four traits of each individual, μ is the vector of means of each trait, β is a vector of fixed effects including batches, battery and position in the battery, u is a vector of genomic breeding values and ε is a vector of random residual effects. X and Z are design matrixes relating respectively phenotypes to fixed effects and phenotypes to genomic breeding values (u). It is assumed that $u \sim N(0, H \otimes W)$ where H is the genetic relationship matrix combining SNP information and pedigree data (Legarra et al., 2009) and W is the matrix of variance and covariance of the genomic breeding values of the four traits. Finally, $\varepsilon \sim N(0, I \otimes R)$ where I is the identity matrix and R is the matrix of residual variance and covariance of the four traits.

Software

FImpute V2.2 (Sargolzaei et al., 2014) was used to impute the selection candidates with low density genotyping to high density genotyping from the individuals of G0 with high density genotyping.

The scenario with all available information (Full_HD) was used to estimate the genetic parameters of the model. Remlf90 (Misztal et al., 2002) was used to estimate the genetic and residual variance components. Once fixed, all different genomic evaluations based on ancestry were performed with Blupf90. The variance components were compared to components estimated with a pedigree based model using all phenotypes. They were highly correlated (Picard Druet et al., 2019).

RESULTS AND DISCUSSION

Imputation Accuracy

All the results concerning imputation accuracy were presented in Herry et al. (2018) but the evolution of the mean correlations between true and imputed genotypes for the two different methodologies were recalled in Figure 3. For both methodologies, there was an increase in

mean correlation with an increase in the number of SNPs on the different low density SNP chips. Better imputation accuracies were obtained with the LD methodology at an equivalent SNP density. The differences observed in mean correlation between the two methodologies were all significant. In addition, for the EQ methodology at a very low density of 1K SNPs, the mean correlation was 0.7098 indicating a quite deteriorated imputation accuracy. This corresponded to a genotyping imputation error rate of 18.5%.

These results were consistent with those found in the literature (Dassonneville et al., 2012; Carvalheiro et al., 2014) where an increase in the number of SNPs on low density SNP chip led to better imputations.

Impact of Imputation Errors

The impact of imputation errors was investigated by comparing the results of a genomic evaluation based on ancestry, with true HD genotyping or with imputed HD genotyping. Only the results for Egg Weight (EW) were shown to simplify the reading and because of the similarity of the results for the other traits.

Results For The Top 150 Individuals.

For both methodologies (Figure 4a), there was an increase in Spearman correlations between "Anc_HD" GEBV and "Anc_imputed" GEBV with an increase in SNP density. Indeed, for the LD0.05 and LD0.8 SNP chips, the mean correlations were respectively 0.8661 and 0.9931. For the 3Kequi and 20Kequi SNP chips, there were respectively 0.9045 and 0.9885. These results are in agreement with imputation accuracies obtained with the different low density SNP chips. There was an increase in mean correlation concerning the evaluations with an increase in imputation accuracy which is consistent with the literature. Moghaddar et al. (2015) showed, for Merino sheep, that the mean correlations between GEBV based on true genotypes (50K) and GEBV based on imputed genotypes (50K imputed from 12K) increased with imputation accuracies.

It was noticed that for both methodologies, with more than 5K SNPs, the mean correlations were above 0.90 indicating a re-ranking rather reduced of the best individuals for EW. However, for the 1Kequi SNP chip, the mean correlation was 0.7833 indicating a reordering quite important of the best individuals for egg weight.

Finally, at equivalent SNP density of 3K SNPs, the EQ methodology seemed to present higher results than the LD methodology with mean GEBV correlations of respectively 0.9045 and 0.8661 for the 3Kequi and LD0.05. But the differences were not significant since the

standard errors were ±0.04 for both SNP chips. At a density of 20K SNPs, both methodologies were equivalent with mean GEBV correlations of respectively 0.9885 and 0.9931 for the 20Kequi and LD0.8. However, as seen previously, the LD methodology appeared to be better to get good imputation accuracies. Thus, higher imputation accuracies with the LD methodology were not synonymous of better mean correlations between GEBV compared to the EQ methodology. This could be due to the methodology itself. Indeed, Harris and Johnson (2010) and Weigel et al. (2010) said that an equidistant methodology was better to get good genomic evaluation results for traits controlled by many small QTL, which is the case for the four traits studied. On the contrary, genomic evaluations concerning traits controlled by few large QTL were more sensitive to equidistant methodology which was consequently not the most appropriated methodology. Moreover, ssGBLUP methodology considers a same variance for each SNP (Legarra et al., 2009) and consequently would favor the EQ methodology. Finally, another reason could be due to the errors done with imputation. Some imputation errors from LD SNP chips could degrade more the GEBV estimation than imputation errors from equidistant SNP chips. The EQ methodology would be more robust than the LD methodology in case of imputation errors.

Results For The Breeders.

Spearman correlations between "Anc_HD" GEBV and "Anc_Imputed" GEBV were also calculated for the 67 G1 breeders having at least 10 offspring in the next generation G2. For both methodologies (Figure 4b), there was an increase in Spearman correlations with an increase in SNP density. Indeed, for the LD0.05 and LD0.08 SNP chips, the mean GEBV correlations were respectively 0.9777 and 0.9979. For the 3Kequi and the 20Kequi SNP chips, the results were respectively 0.9771 and 0.9972. Thus, the results were higher compared to the results for the top 150 individuals. This is due to the distribution of the 67 breeders which were not the best breeders of G1 for EW, but the best for a set of selection criteria. This was confirmed by plotting the normal distribution of HD GEBV estimated on ancestry with true HD genotyping for all G1 candidates (Figure 5). The 67 breeders (in red on the plot) were well distributed among the 580 individuals of G1 which reduced the reordering of the individuals.

The results also showed that even with a SNP density superior to 2K SNPs, good mean correlations (superior to 0.95) could be obtained indicating a very reduced re-ranking of the individuals. With only 5K SNPs imputed to the HD SNP chips, mean correlations above 0.98 could be reached.

However, with the 1Kequi SNP chip, the mean GEBV correlation was under 0.95. This decrease in correlation was also illustrated by Cleveland and Hickey (2013) in pig. They used

only 450 SNPs imputed to the Illumina PorcineSNP60 BeadChip which resulted in a decrease in correlation to 0.866 (for an imputation accuracy of 0.914). Thus, by decreasing to much the SNP density, the reduced imputation accuracies can have negative consequences on genomic evaluations.

Finally, our results did not show any difference between EQ and LD methodologies.

Impact of the Absence of Imputation

Given the good results of genomic evaluations with imputed genotyping, the impact of the absence of imputation was studied. Only the results for Egg Weight (EW) were shown to simplify the reading and because of the similarity of the results for the other traits.

Results For The Top 150 Individuals.

For the top 150 individuals for both methodologies (Figure 6a), there was an increase in Spearman correlation between "Anc_HD" GEBV and "Anc_Not_Imputed" GEBV with an increase in SNP density. Indeed, the mean correlations for the 3Kequi and the 20Kequi SNP chips were respectively 0.8507 and 0.9379. For the LD0.05 and the LD0.8 SNP chips there were respectively 0.7816 and 0.8658. Zhang et al. (2011) showed in simulation studies that compared to the results of a genomic evaluation done with HD SNP chip, the results of genomic evaluations done with low density SNP chips without imputation also decreased. With an effective population size of 100, heritability of 0.5, 241 QTL, and a SNP chip of 10K markers, the relative accuracy of the GBLUP evaluation decreased from 0.88 with 5K markers to 0.69 with only 200 markers.

For both methodologies, there was a consequent decrease in mean correlations compared to the results of the genomic evaluations done with imputed HD genotyping. For the 1Kequi and the 50Kequi SNP chips, both imputed, the results were respectively 0.7833 and 0.9964. Without imputation, the results were respectively 0.6261 and 0.9503. Likewise, for the LD0.05 and the LD0.8 SNP chips with imputation, the results were respectively 0.8661 and 0.9931. Without imputation, the results decreased respectively to 0.7816 and 0.8658. Furthermore, from 20K SNPs, the results for the EQ methodology seemed to reach a mean correlation threshold of 0.95 whereas with imputation the mean correlations were above 0.99. Thus, imputations enabled to increase significantly the mean correlations, mainly for very low density SNP chips. In addition, these results indicate that the ranking of the best 150 individuals of G1 for EW obtained without imputation was quite different from the ranking obtained with

HD genotyping. The lower results obtained for very low SNP density indicated that using few SNPs could not be sufficient to accurately rank individuals having very close genomes.

Finally, at equivalent SNP density, a tendency to get higher results with the EQ methodology was observed. Indeed, at 3K SNPs, the difference in mean correlation between 3Kequi and LD0.05 SNP chips was equal to 0.07. The same difference was obtained between 20Kequi and LD0.8 SNP chips. Such differences were higher than with imputation but were not significant. However, we can note that the correlations remained always below 0.90 for the top 150 individuals whatever the SNP density with the LD methodology without imputation. The differences between methodologies are consistent with the genetic determinism of the four traits as explained in the previous part (Harris and Johnson, 2010; Weigel et al., 2010). In addition, the EQ methodology enabled a covering of all chromosomes more optimal than the LD methodology (Herry et al., 2018). With the LD methodology, there were some gaps on chromosomes without SNPs selected on low density SNP chips. With the EQ methodology, the number of gaps was decreased, or at least their size was lower.

Results For The Breeders.

Spearman correlations between "Anc_HD" GEBV and "Anc_Not_Imputed" GEBV were also calculated for the 67 breeders (Figure 6b). For both methodologies, there was an increase in Spearman correlations with an increase in SNP density. At equivalent SNP density, the results for the 3Kequi and 20Kequi SNP chips were respectively 0.9484 and 0.9802. For the LD0.05 and LD0.8 the results were respectively 0.9349 and 0.9665. Compared to the results for the top 150 individuals, the results were better for the 67 breeders as shown previously in the scenario with imputation. Finally, for a SNP density higher than 3K, the mean correlations were above 0.94 for both methodologies, indicating a reordering rather reduced of the 67 breeders. In bovine, Weigel et al. (2009) showed that compared to the top 500 bulls selected from progeny testing, 306 were truly selected with 32K SNPs chosen from the Illumina BovineSNP50 Bead Chip. With 2K equally spaced SNPs, 292 bulls were chosen. With only 500 equally spaced SNPs, 247 bulls were chosen. This illustrates that compared to the HD SNP chip, the re-ranking was limited and that even with few SNPs, the reordering of the individuals was limited.

Compared to the results obtained with imputation, there was a slight decrease in correlations with "Anc_HD" GEBV. Indeed, for the 1Kequi and 50Kequi SNP chips, the results were respectively 0.9316 (±0.0451) and 0.9983 (±0.0072) with imputation, and 0.8718 (±0.0608) and 0.9815 (±0.0238) without imputation. Likewise, for the LD0.05 and the LD0.8, the results were respectively 0.9777 (±0.0261) and 0.9979 (±0.0080) with imputation, and

0.9349 (±0.0440) and 0.9665 (±0.0318) without imputation. Thus, the differences observed for both methodologies were not significant and the results were still high whatever the SNP chip used. These results were rather different from those obtained by Aliloo et al. (2018). They showed in bovine, for 1034 individuals, that correlations between HD GEBV (on 777K genotypes) and GEBV based on imputed HD genotyping were significantly higher than without imputation. Indeed, according to their MAFI (Minor Allele Frequency within Interval) method which was the closest to our EQ methodology, using 4013 and 25,410 SNPs imputed to 777K SNPs resulted respectively in correlations of 0.9398 and 0.9927. These results decreased dramatically without imputation with correlations of respectively 0.6485 and 0.8598. Such a large decrease was not observed in our study.

Finally, the differences observed between the two methodologies were also not significant. Consequently, the simpler EQ methodology seems to be sufficient to get good genomic evaluation results for traits controlled by many small QTL, which is the case for the four traits studied.

Impact Of Imputation On Relative Accuracy Of Genomic Evaluation

The impact of imputation on the attainable relative accuracy of genomic evaluations was studied by comparing a genomic evaluation "Full_HD" of the 67 G1 breeders using all available information (phenotypes and genotypes) from generation (G0) to (G3) and GEBV of the G1 breeders based on ancestry with imputed HD genotyping ("Anc_Imputed" GEBV), for each low density SNP chip. Only the results for Egg Weight (EW) were shown to simplify the reading and because of the similarity of the results for the other traits.

It was noticed (Figure 7) for the EQ methodology a slight increase in Pearson correlations from very low density SNP chips to 20K SNPs. Indeed, for the 1Kequi and the 20Kequi SNP chips, the mean correlations were respectively 0.4472 and 0.4854. But for the LD methodology, the results were rather stable with mean correlations of respectively 0.4917 and 0.4875 for the LD0.05 and LD0.8 SNP chips. For both methodologies, the results varied slightly up to 20K SNPs. They became steady for the EQ methodology from 20K to higher SNP densities. Finally, for both methodologies, the correlations of "Anc_Imputed" GEBV with "Full_HD" GEBV were not significantly different from those obtained by comparison between true HD GEBV on ancestry and "Full_HD" GEBV. The mean correlation was 0.4848 and corresponded to a theoretical maximum value attainable. The standard error for each low density SNP chip was ± 0.11 indicating that there was no difference with the theoretical maximum value. For information purposes, the mean correlations for ESC, ESS and AH were 0.2618 ± 0.12, 0.4027 ± 0.11 and 0.4802 ± 0.11. This is consistent with the previous results

showing a very slight impact of imputations errors on GEBV estimations of the 67 breeders on ascendance. For both methodologies, from a density of 5K SNPs imputed to the HD SNP chip, the mean correlations were above 0.98 between "Anc_HD" GEBV and "Anc_Imputed" GEBV. These results are also in agreement with the literature. Indeed, Harris and Johnson (2010) showed that in bovine, from 5K to 1000K SNPs, the increase in correlations between true phenotypes and predicted phenotypes was very limited (0.62 to 0.65). VanRaden et al. (2012) showed that, for 28 traits tested in bovine, in average, the estimated genomic reliability was 61.1% with 300K SNPs and decreased to only 60.7% when they used 45K SNPs. In the study of Wellman et al. (2013), 768 SNPs imputed to the Illumina PorcineSNP60 BeadChip (60K SNPs) led to a negligible loss in genomic evaluation accuracy. Likewise, Chen et al. (2014) estimated in bovine that the accuracy of genomic prediction with observed 50K or imputed 50K (from 6K) genotypes was 0.61 for milk yield and 0.62 for somatic cell score (SCS).

However, a decrease in relative accuracy was observed with the 1Kequi SNP chip with a mean correlation of 0.4472. The highest decrease was observed for albumen height (AH) where the mean correlation for the 1Kequi SNP chip was 0.4045 (±0.11) and the theoretical maximum value was 0.4802. One cannot conclude about the significance of this difference but this decrease was also expected because the results regarding the impact of imputation accuracies showed a mean correlation of 0.9316 for the 1Kequi SNP chip. Other studies showed that decreasing to much the SNP density has consequences on genomic evaluation accuracies. Raoul et al. (2017) illustrated this point in Merino sheep where using only 500 or 250 SNPs imputed to the Illumina OvineSNP50 BeadChip resulted respectively in a decrease in accuracies from 0.53 (with HD SNP chip) to 0.45 and 0.38. Wellman et al. (2013) showed that 384 SNPs imputed to the Illumina PorcineSNP60 BeadChip led to a loss of 3% in genomic evaluation accuracy. Likewise, Chen et al. (2014) showed that the accuracy of genomic prediction decreased from 0.61 to 0.49 for milk yield and from 0.62 to 0.53 for SCS with imputed 50K genotypes from 384 SNPs.

Consequently, we can conclude that the effects of imputation errors on GEBV relative accuracies were very limited even if slightly more important for very low densities.

Impact of the Direct Use of Low Density SNP Chips Without Imputation on Relative Accuracy of Genomic Evaluation

The impact of the direct use of low density SNP chips on relative accuracy of genomic evaluation was studied by comparing the "Full_HD" GEBV of the G1 and GEBV of the G1 breeders on ancestry with low density genotyping without imputation ("Anc_Not_Imputed"

GEBV), for each low density SNP chip. For both methodologies, only the results for Egg Weight (EW) were shown to simplify the reading and because of the similarity of the results for the other traits.

Both methodologies were rather stable with slight variations in Pearson correlations up to 20K SNPs (Figure 8). The results for the 3Kequi and 20Kequi SNP chips were respectively 0.4471 and 0.4675. For the LD0.05 and LD0.8 the correlations were respectively 0.4583 and 0.4888. However, the standard errors associated to these results were ±0.11 and the correlation between the "Full HD" GEBV and the HD GEBV based on ancestry was 0.4848. This indicates that the differences observed between each low density SNP chip, and consequently between the two methodologies, were not significant. These results are in agreement with the previous results showing a very slight impact of the absence of imputation on GEBV estimation of the 67 breeders on ascendance. However, the results for the 1Kequi was 0.4018 (±0.11). This lower but non-significant result was also expected because the correlation between "Anc_HD" GEBV and "Anc_Not_Imputed" GEBV was lower (0.8718 ±0.0608) than those obtained with higher SNP densities. This was the case for all traits studied.

The results found in the literature are contrasted. Moghaddar et al. (2015) showed in Merino sheep, that the accuracy of genomic prediction based on observed 50K genotypes was 0.446 for post-weaning weight (PWW) and 0.219 for post-weaning eye muscle depth (PW_EMD). Based on genotypes imputed from 12K to 50K genotypes, with imputation accuracy comprised between 0.88 and 0.99, the accuracy of genomic prediction was 0.443 for PWW and 0.219 for PW_EMD. Based on observed 12K genotypes, the accuracy was 0.412 for PWW and 0.205 for PW_EMD. Thus, the results were slightly better with imputation compared to a direct use of the 12K without imputation, but in both cases, there was not a dramatic decrease in genomic prediction accuracy despite a significant gap of SNP density between HD and low density chips. Weigel et al. (2009) had a gap of SNP density closer to our but the results were rather different. The correlation between the results from progeny testing and the genomic result with a HD SNP chip of 32K was 0.612. With 300, 1K and 2K equally spaced SNPs, the results were respectively 0.253, 0.422 and 0.539. Contrary to the results of Moghaddar et al. (2015), there was a significant decrease in their results with the use of low density SNP chips without imputation. In 2010, they showed that their results were better with imputation.

Finally, for a SNP density higher than 3K, using low density SNP chips without imputation leaded to results as good as those obtained with the HD SNP chip itself.

CONCLUSIONS

This study showed a very limited reordering of the breeders, selected on a multi-traits index, with low density genotyping (with or without imputation) instead of HD genotyping. Indeed, Spearman correlations between GEBV on HD genotyping and GEBV on low density genotyping were always higher than 0.94 with more than 3K SNP. For the top 150 individuals, who are genetically closer than the breeders, the reordering was a bit more important. Thus, the correlations between GEBV with HD genotyping and GEBV with low density genotyping remained below 0.85 with less than 3K SNP with the EQ methodology and less than 16K SNP (LD0.6) with the LD methodology. The differences in GEBV correlations between the two methodologies were never significant but seemed to indicate that the simpler EQ methodology was sufficient to obtain similar results.

Thus, using directly low density SNP chips designed with the EQ methodology with more than 5K SNPs could enable to get good results of genomic evaluation and could be a cost effective solution for genomic selection. However, only four traits were studied. These four traits were controlled by many small QTL, which explained why the equidistant methodology was more appropriated to realize genomic evaluation with ssGBLUP than the LD methodology, whereas the results on imputation accuracies were inverted. Further investigations on other traits with different genetic architectures should be conducted.

Finally, as shown by Habier et al. (2009), there could be a decrease in genomic evaluation accuracy over the generations with low density genotyping. This would require to genotype at higher density birds selected at each generation to avoid a decrease in genomic evaluation accuracy which could be prejudicial for genomic selection. In addition, in our study, only the males were genotyped but having both parents genotyped could lead to higher genomic evaluation accuracies..

ACKNOWLEDGMENTS

This research project was partly supported by the French national research agency "ANR" within the framework of project ANR-10-GENOM_BTV-015 UtOpIGe. FlH is a PhD fellow supported by the poultry breeding company Novogen as part of a CIFRE thesis between PEGASE INRA's unit, Agrocampus Ouest and Novogen.

REFERENCES

Aliloo, H., R. Mrode, A. M. Okeyo, G. Ni, M. E. Goddard, and J. P. Gibson. 2018. The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa. J. Dairy Sci. 101:1-20.

Atol Ontology. INRA 2012. [http://www.atol-ontology.com]. Accessed 11 Mar 2015.

Carvalheiro, R., S. A. Boison, H. H. R. Neves, M. Sargolzaei, F. S. Schenkel, Y. T. Utsunomiya, A. M. P. O'Brien, J. Sölkner, J. C. McEwan, C. P. Van Tassell, T. S. Sonstegard, and J. F. Garcia. 2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol. 46:69-79.

Chen, L., C. Li, M. Sargolzaei, and F. Schenkel. 2014. Impact of genotype imputation on the performance of GBLUP and Bayesian methods for genomic prediction. PLoS One. 9:e101544.

Cleveland, M. A., and J. M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J. Anim. Sci. 91:3583-3592.

Dassonneville, R., S. Fritz, V. Ducrocq, and D. Boichard. 2012. Short communication: Imputation performances of 3 low density marker panels in beef and dairy cattle. J. Dairy Sci. 95:4136-4140.

Deelen, P., M. J. Bonder, K. J. van der Velde, H. J. Westra, E. Winder, D. Hendriksen, L. Franke, and M. A. Swertz. 2014. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Res. Notes. 7:901-904.

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed Genet. 128:409-421.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics. 182:343-353.

Harris, B. L., and D. L. Johnson. 2010. The impact of high density SNP chips on genomic evaluation in dairy cattle. Interbull Bull. 42:40-43.

Hérault, F., J. Yon, F. Herry, S. Allais, and P. Le Roy. 2016. SS4I: select SNP susbset for imputation. (in French). https://prodinra.inra.fr/record/375448.

Hérault, F., F. Herry, A. Varenne, T. Burlot, D. Picard-Druet, J. Recoquillay, C. Macé, F. Fagnoul, S. Allais, and P. Le Roy. 2018. A linkage disequilibrium study in layer and broiler commercial chicken populations. Proc. 11th World Congr. Genet. Appl. Livest. Prod., Auckland, New-Zealand.

Herry, F., F. Hérault, D. Picard Druet, A. Varenne, T. Burlot, P. Le Roy, and S. Allais. 2018. Design of low density SNP chips for genotype imputation in layer chicken. BMC Genetics. 19:108-121.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 432:695-716.

Kranis, A., A. A. Gheyas, C. Boschiero, F. Turner, L. Yu, S. Smith, R. Talbot, A. Pirani, F. Brew, P. Kaiser, P. M. Hocking, M. Fife, N. Salmon, J. Fulton, T. M. Strom, G. Haberer, S. Weigend, R. Preisinger, M. Gholami, S. Qanbari, H. Simianer, K. A. Watson, J. A. Woolliams, and D. W. Burtet. 2013. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 14:59-71.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656-4663.

Megens, H. J., R. P. M. A. Crooijmans, J. W. M. Bastiaansen, H. H. D. Kerstens, A. Coster, R. Jalving, A. Vereijken, P. Silva, W. M. Muir, H. H. Cheng, O. Hanotte, and M. A. M. Groenen. 2009. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genetics. 10:86-96.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819-1829.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D. H. Lee. 2002. BLUPF90 and related programs (BGF90). Proc. 7th. World Congr. Genet. Appl. Livest. Prod., Montpellier, France.

Moghaddar, N., K. P. Gore, H. D. Daetwyler, B. J. Hayes, and J. H. J. van der Werf. 2015. Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. Genet. Sel. Evol. 47:97-108.

Mulder, H. A., M. P. L. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95:876-889.

Picard Druet, D., A. Varenne, F. Herry, F. Hérault, S. Allais, T. Burlot and P. Le Roy. 2019. Properties of genomic evaluation for egg quality traits in layers. In submission in Genet. Sel. Evol.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81:559-575.

Qanbari, S., M. Hansen, S. Weigend, R. Preisinger, and H. Simianer. 2010. Linkage disequilibrium reveals different demographic history in egg laying chickens. BMC Genetics. 11:103-115.

Raoul, J., A. A. Swan, and J. M. Elsen. 2017. Using a very low-density SNP panel for genomic selection in a breeding program for sheep. Genet. Sel. Evol. 49:76-87.

Robert, R., F. Hérault, H. Romé, A. Varenne, H. Chapuis, A. Vignal, T. Burlot, and P. Le Roy. 2015. A linkage disequilibrium study in a layer chicken population. Proc. 9th. Eur. Symp. Poult. Genet., Tuusula, Finland.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 15:478-489.

Su, G., R. F. Brøndum, P. Ma, B. Guldbrandsten, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J. Dairy Sci. 95:4657-4665.

VanRaden, P. M., J. R. O'Connell, G. R. Wiggans, and K. A. Weigel. 2011. Genomic evaluation with many more genotypes. Genet. Sel. Evol. 43:10-20.

VanRaden, P. M., J. D. Null, M. Sargolzaei, G. R. Wiggans, M. E. Tooker, J. B. Cole, T. S. Sonstegard, E. E. Connor, M. Winters, J. B. C. H. M. van Kaam, A. Valentini, B. J. Van Doormaal, M. A. Faust, and G. A. Doak. 2012. Genomic imputation and evaluation using high-density Holstein genotypes. J. Dairy Sci. 95:1-11.

Wang, C., D. Habier, B. L. Peiris, A. Wolc, A. Kranis, K. A. Watson, S. Avendano, D. J. Garrick, R. L. Fernando, S. J. Lamont, and J. C. M. Dekkers. 2013. Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide polymorphism panel in broiler chickens. Poult. Sci. 92:1712-1723.

Wang, Y., G. Lin, C. Li, and P. Stothard. 2016. Genotype imputation methods and their effects on genomic predictions in cattle. Springer Sci. Rev. 4:79-98.

Warren, W. C., L. D. W. Hillier, C. Tomlinson, P. Minx, M. Kremitzki, T. Graves, C. Markovic, N. Bouk, K. D. Pruitt, F. Thibaud-Nissen, V. Schneider, T. A. Mansour, C. T. Brown, A. Zimin, R. Hawken, M. Abrahamsen, A. B. Pyrkosz, M. Morisson, V. Fillon, A. Vignal, W. Chow, K. Howe, J. E. Fulton, M. M. Miller, P. Lovell, C. V. Mello, M. Wirthlin, A. S. Mason, R. Kuo, D. W. Burt, J. B. Dodgson, and H. H. Cheng. 2017. A new chicken genome assembly provides insight into avian genome structure. G3. 7:109-117.

Weigel, K. A., G. de los Campos, O. González-Recio, H. Naya, X. L. Wu, N. Long, G. J. M. Rosa, and D. Gianola. 2009. Predictive ability of direct genomic values for lifetime net

merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. J. Dairy Sci. 92:5248-5257.

Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. van Tassell. 2010. Accuracy of direct genomic values dervied from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93:5423-5435.

Wellmann, R., S. Preuβ, E. Tholen, J. Heinkel, K.Wimmers, and J. Bennewitz. 2013. Genomic selection using low density marker panels with application to a sire line in pigs. Genet. Sel. Evol. 45:28-38.

Xu, S. 2003. Estimating polygenic effects using markers of the entire genome. Genetics. 163:789-801.

Zhang, Z., X. Ding, J. Liu, Q. Zhang, and D. J. de Koning. 2011. Accuracy of genomic prediction using low-density marker panels. J. Dairy Sci. 94:3642-3650.

Table 1. Summary of the different steps of quality control.

Genotypes filtration	**RI Line**
Individual Call Rate (<95%)	8
MAF (=0)	204,122
MAF (<0.05)	54,650
SNP Call Rate (<95%)	7541
Hardy-Weinberg equilibrium ($P<10^{-4}$)	12,538
SNP with unknown location or on chromosome W	1759
Pedigree Incompatibility problem	0
SNP retained for analyses	**300,351**
Animals retained for analyses	**2362**

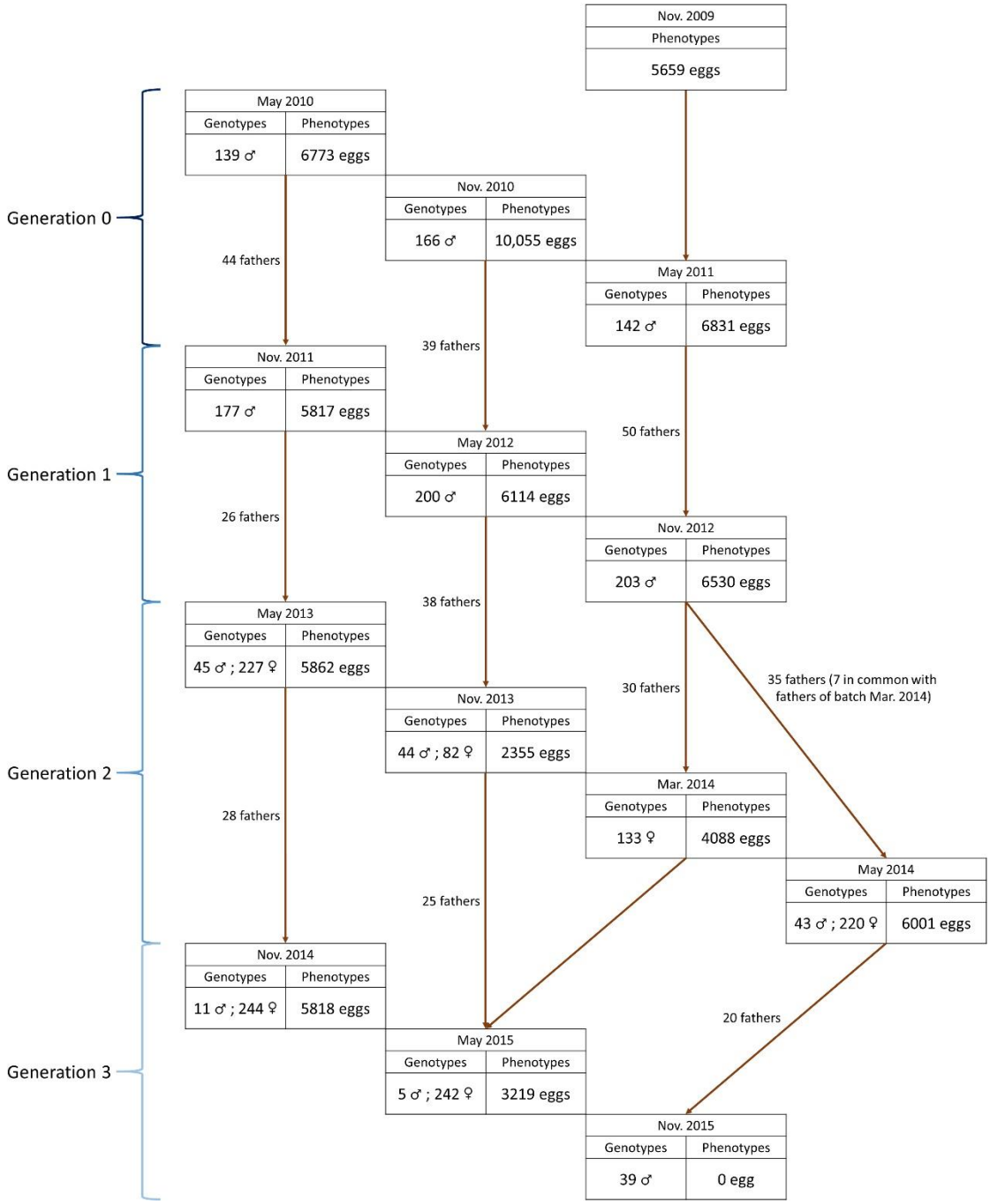


Figure 1. Population structure of the RI line.

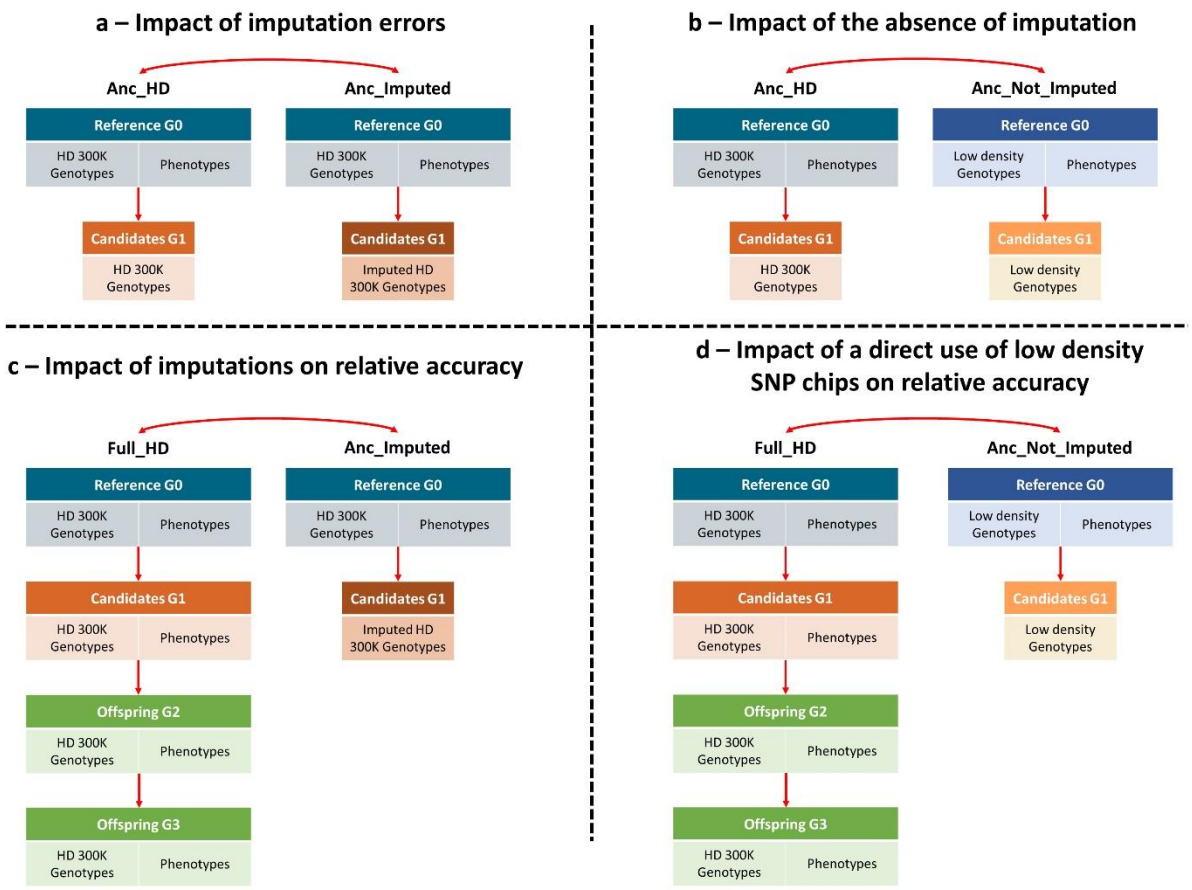


Figure 2. Summary of all different genomic evaluation strategies studied.

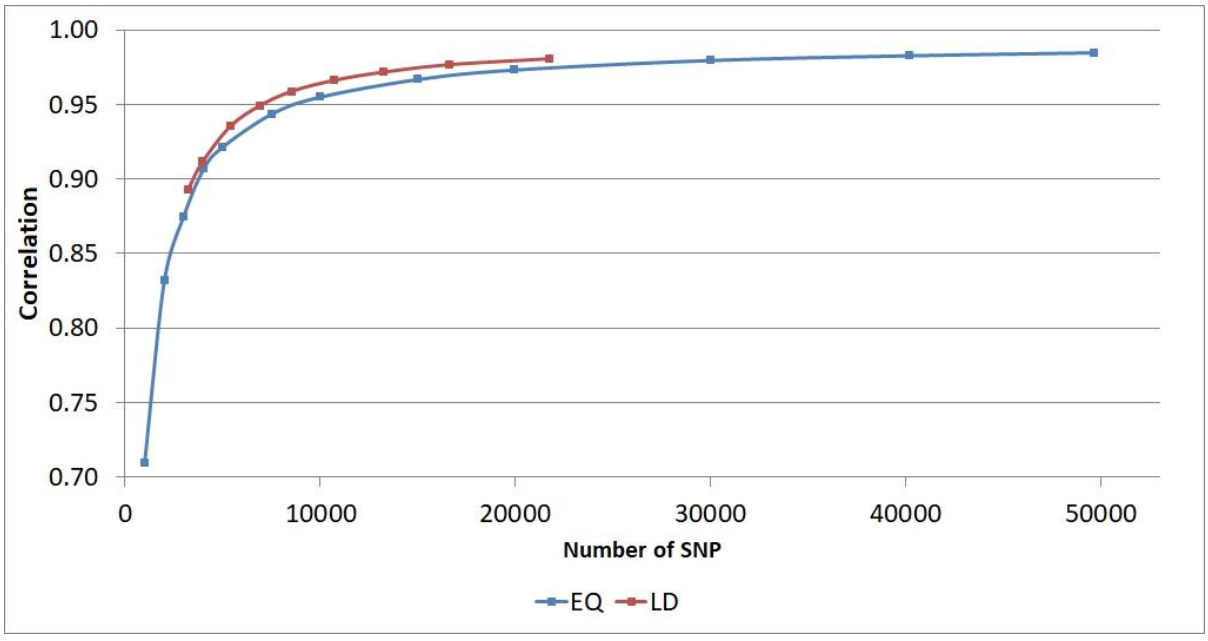


Figure 3. Mean correlations between true and imputed genotypes according to the number of SNPs on low density SNP chips for EQ and LD methodologies.

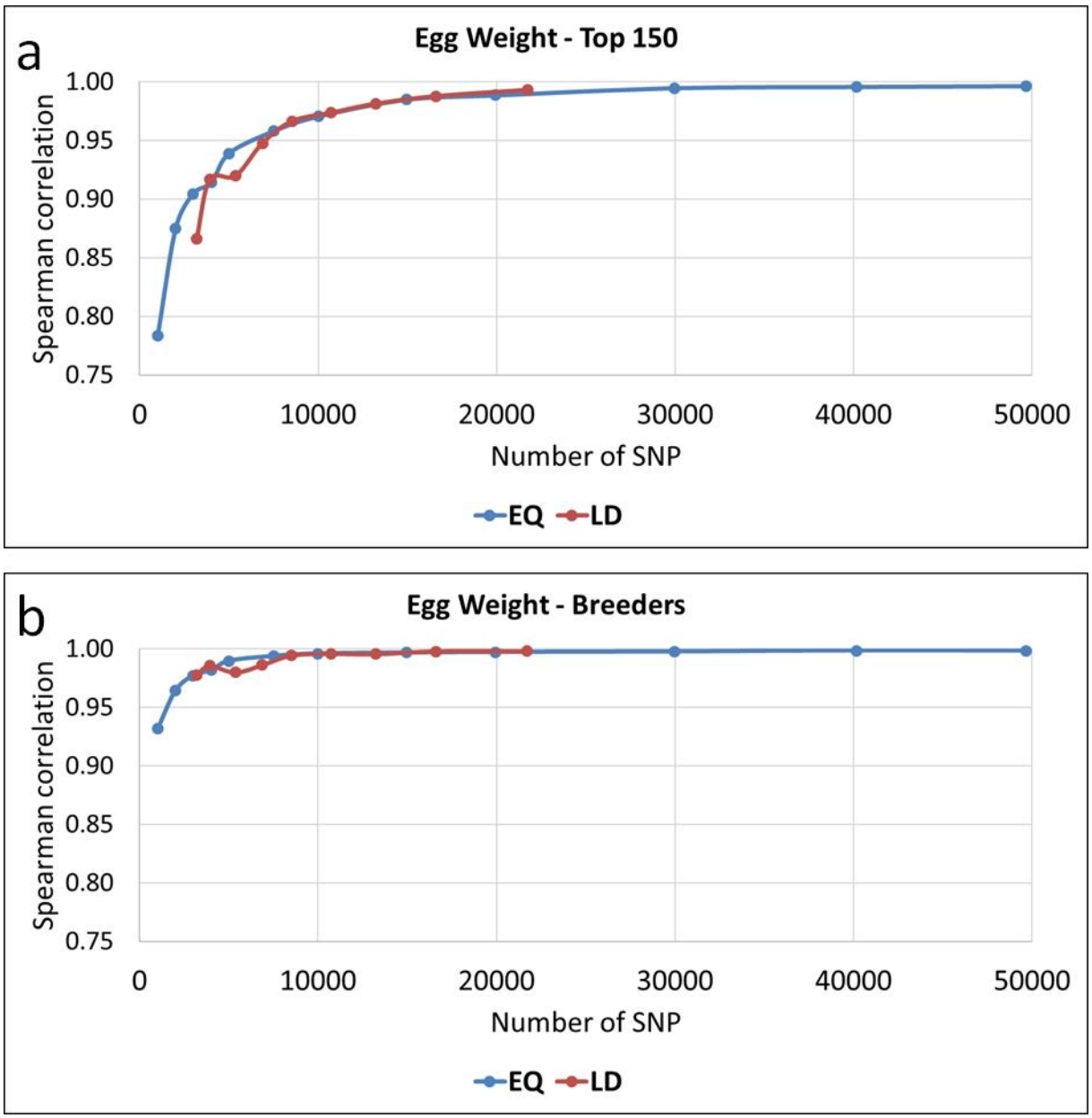


Figure 4. Spearman correlations between GEBV based on ancestry obtained with true HD genotyping and GEBV based on ancestry obtained with imputed HD genotyping. Results are shown for egg weight and for the top 150 individuals (a) or the 67 breeders (b) according to the number of SNPs on low density SNP chip for both methodologies.

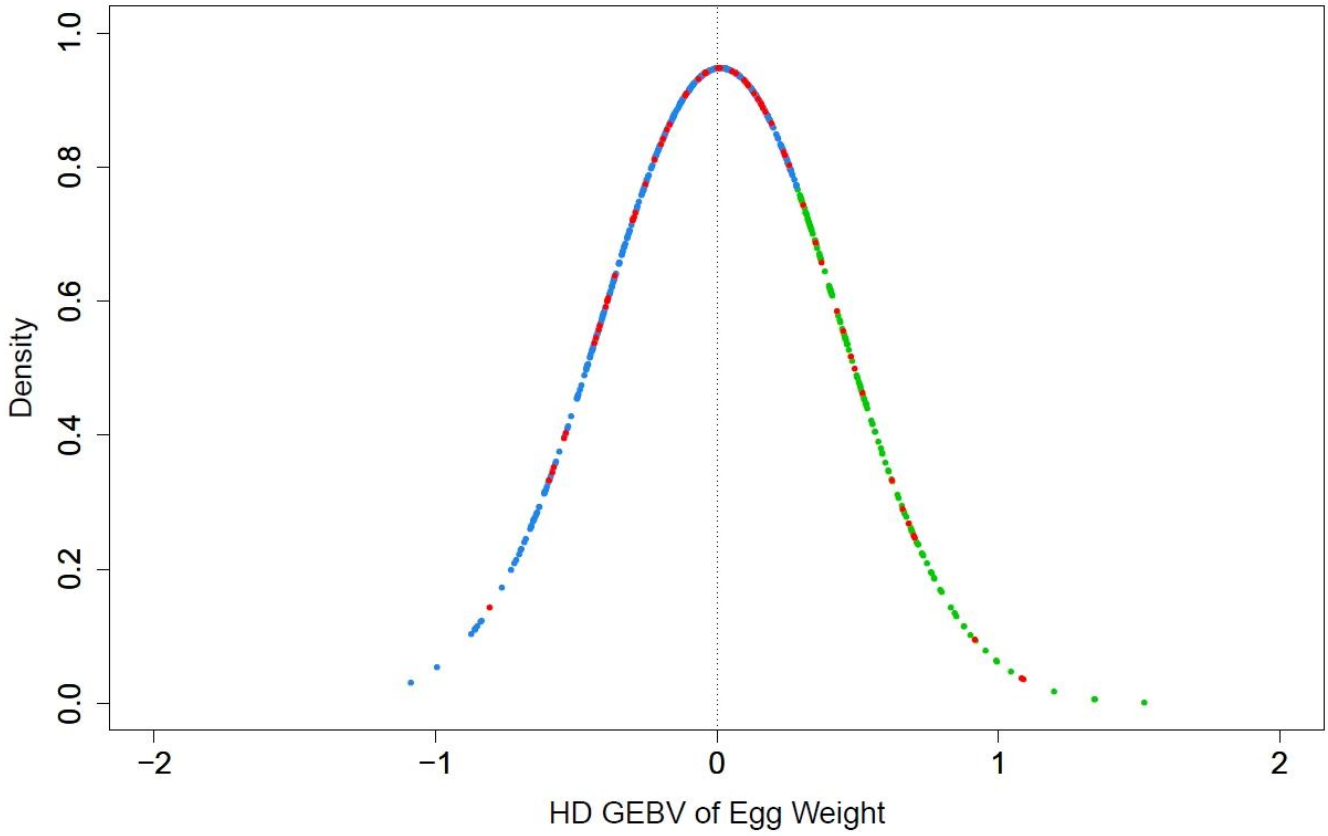


Figure 5. Normal distribution of all G1 selection candidates according to their HD GEBV of Egg Weight estimated on ancestry with true HD genotyping. Red dots represent the 67 G1 breeders, green dots represent the top 150 individuals for EW, and blue dots represent the other selection candidates.

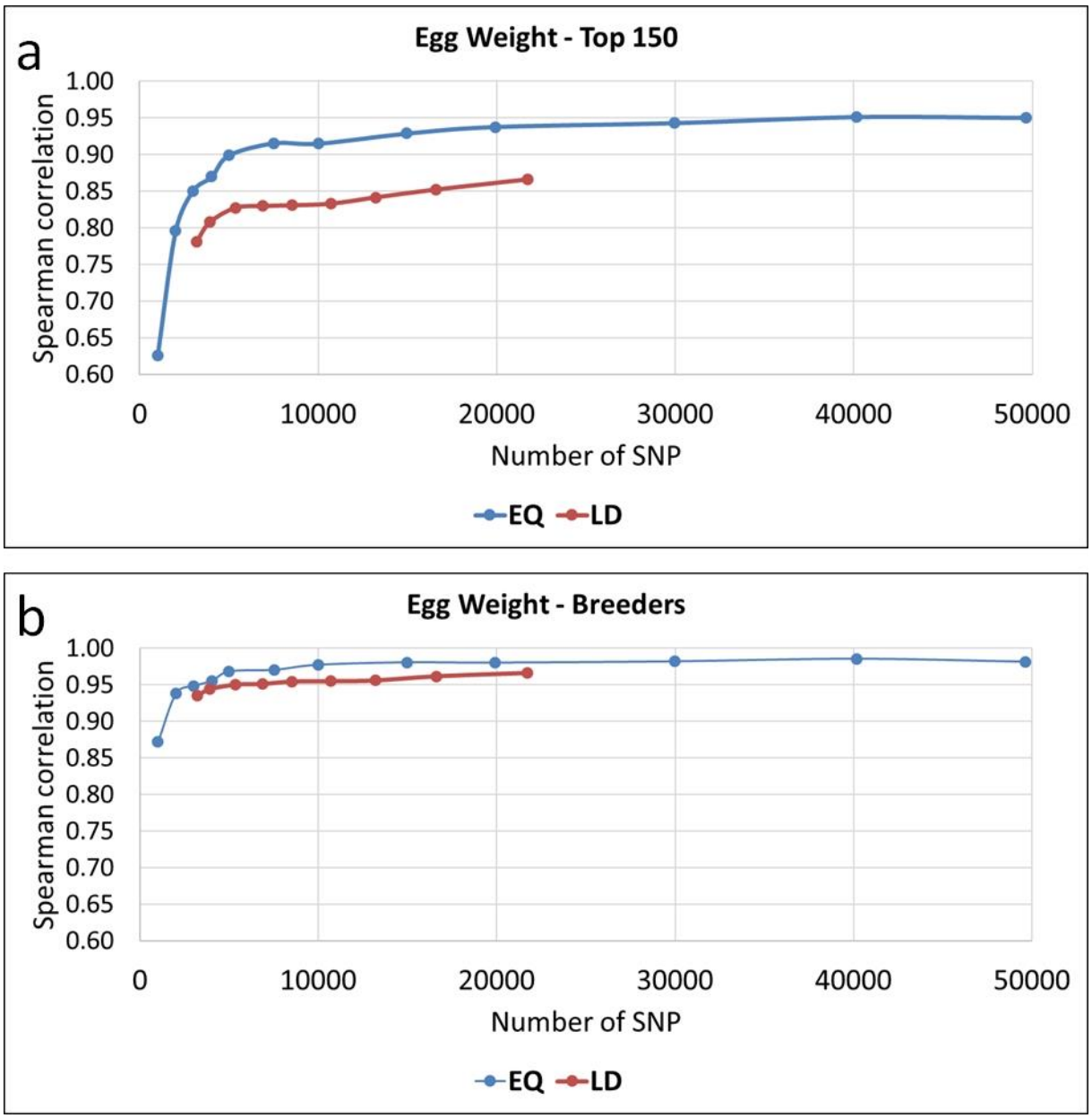


Figure 6. Spearman correlations between GEBV based on ancestry obtained with true HD genotyping and GEBV based on ancestry obtained with low density genotyping (without imputation). Results are shown for egg weight and for the top 150 individuals (a) or the 67 breeders (b) according to the number of SNPs on low density SNP chip for both methodologies.

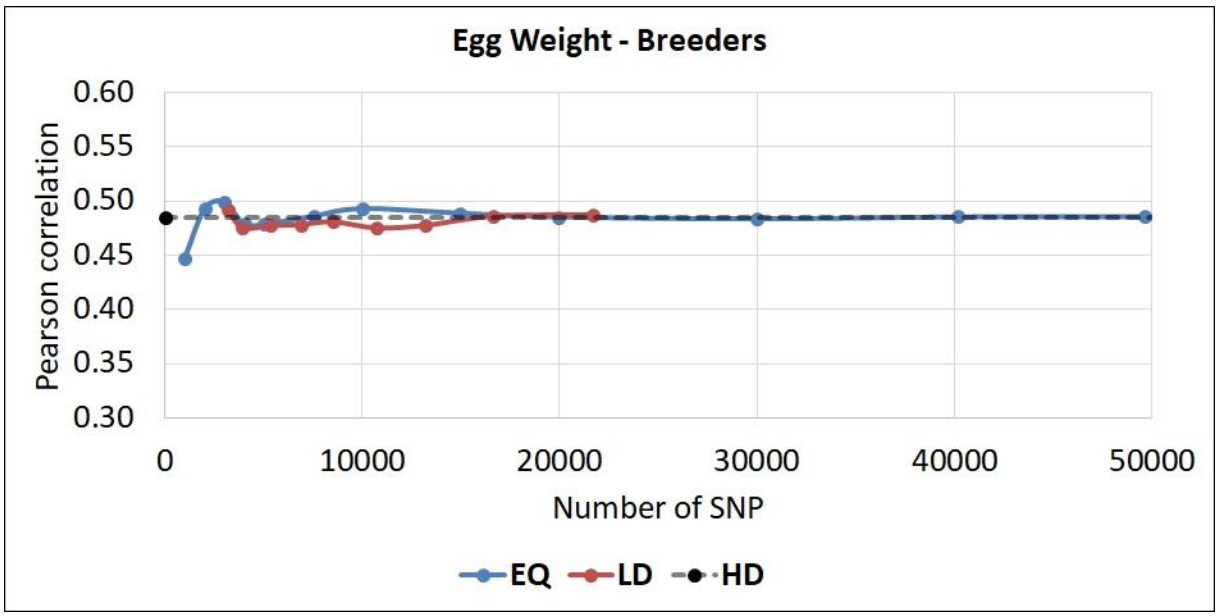


Figure 7. Pearson correlations between "Full_HD" GEBV based on offspring with true HD genotyping and GEBV based on ancestry with imputed HD genotyping. Results are shown for egg weight and for the 67 G1 breeders according to the number of SNPs on low density SNP chip for both methodologies.

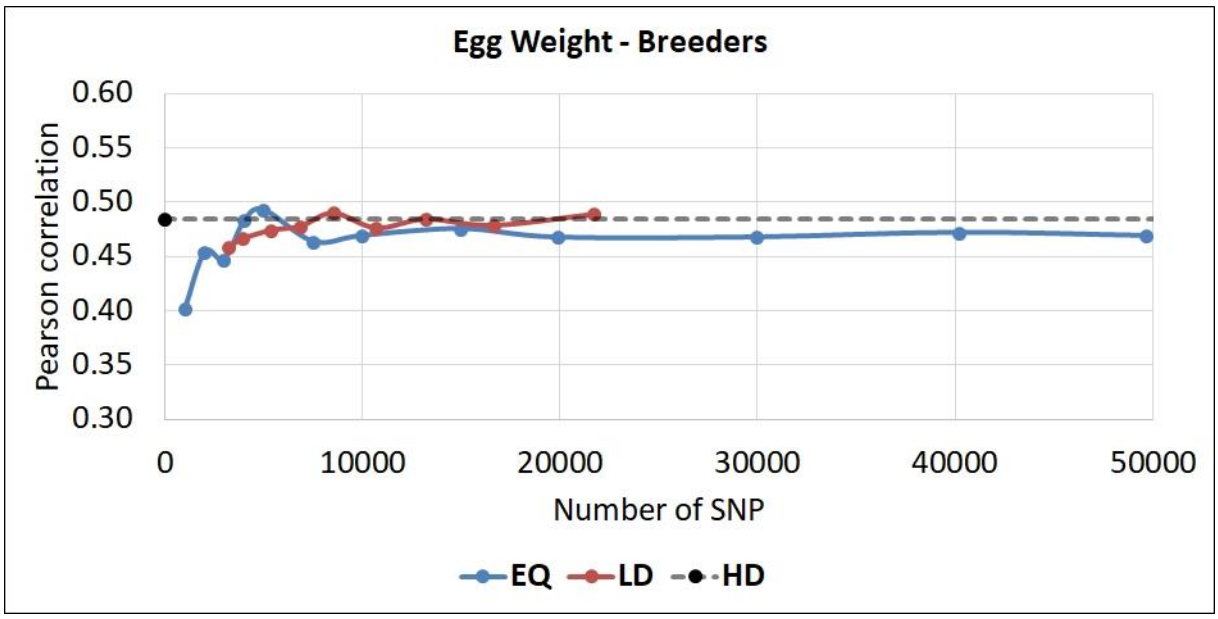


Figure 8. Pearson correlations between "Full_HD" GEBV based on offspring with true HD genotyping and GEBV based on ancestry with low density genotyping (without imputation). Results are shown for egg weight and for the 67 G1 breeders according to the number of SNPs on low density SNP chip for both methodologies.

III. Discussion

A. Extension des études à la période d'élevage en cage collective

1. Données

Le premier article s'est concentré sur les évaluations génomiques des coqs à partir des performances des poules pondeuses en cages individuelles de 60 à 90 semaines d'âge. Cependant, des analyses similaires ont également été réalisées pour la période d'élevage en cage collective des femelles, à savoir la période de 18 à 60 semaines d'âge. Durant cette période, les pondeuses sont élevées en cages collectives de 5 pleines sœurs. En conséquence, chaque mesure collectée est associée à une cage et donc, à une famille. Il n'est pas possible d'associer la mesure à une poule spécifique de la cage. 27 970 mesures ont ainsi été collectées pour 19 212 poules pondeuses. Parmi les 580 candidats mâles à la sélection de la génération G1, 172 ont des filles (en génération G2) en cages collectives et 67 parmi les 172 ont des filles (en génération G2) en cages individuelles.

2. Intérêt de l'imputation quant au classement des individus

L'impact des erreurs d'imputation ou de l'absence d'imputation sur le classement des individus a été étudié de la même façon que dans l'article 2. Pour cela, les GEBV de l'évaluation génomique sur ascendance des candidats à la sélection avec leurs vrais génotypes HD « Anc_HD » ont été comparés avec les GEBV des évaluations génomiques sur ascendance des candidats avec leurs génotypes HD imputés « Anc_Imputed » ou leurs génotypes BD sans imputation « Anc_Not_Imputed ». Les résultats ne sont présentés que pour le poids d'œuf à cause de la similarité des résultats concernant les autres caractères étudiés.

a) Études avec imputation

Les résultats concernant l'impact des erreurs d'imputation sur le reclassement des individus montrent, pour les meilleurs individus pour un caractère (top 150) comme pour les reproducteurs (ainsi que pour les deux méthodologies), une augmentation des corrélations de Spearman avec une augmentation de la densité de SNP (Figure 25). Avec plus de 3K SNP, les corrélations sont toutes supérieures à 0.90 ce qui indique un reclassement plutôt réduit des différents individus. En revanche, avec seulement 1K SNP obtenus avec la méthodologie EQ, le reclassement des meilleurs individus pour le poids d'œuf est plus impacté avec une corrélation de 0.80. Par ailleurs, ce reclassement est plus faible pour les reproducteurs que pour les 150 meilleurs individus pour le poids d'œuf. Toutefois, des différences significatives entre

les corrélations des reproducteurs et celles des meilleurs individus par caractère, pour les deux méthodologies, sont seulement observées pour moins de 5K SNP. Ceci s'explique par les mêmes raisons que celles évoquées dans l'article, à savoir une répartition des 172 reproducteurs dans l'ensemble du classement des 580 individus pour le poids d'œuf.

D'une manière générale, les valeurs obtenues pour la période d'élevage en cage collective sont légèrement plus élevées que celles obtenues pour la période d'élevage en cage individuelle. Cela peut s'expliquer par le fait que le modèle en cage individuelle est plus précis car chaque mesure est associée à une poule pondeuse et non à une famille. Le nombre de données collectées est également plus important en CI. De plus, les individus du top 150 ne sont pas les mêmes en cages collectives et en cages individuelles.

Enfin, les résultats concernant la méthodologie EQ semblent légèrement supérieurs à ceux de la méthodologie DL pour les 150 meilleurs individus pour des densités inférieures à 15K SNP. Toutefois, aucune différence significative sur le reclassement des meilleurs individus et des reproducteurs n'est notée entre les deux méthodologies.

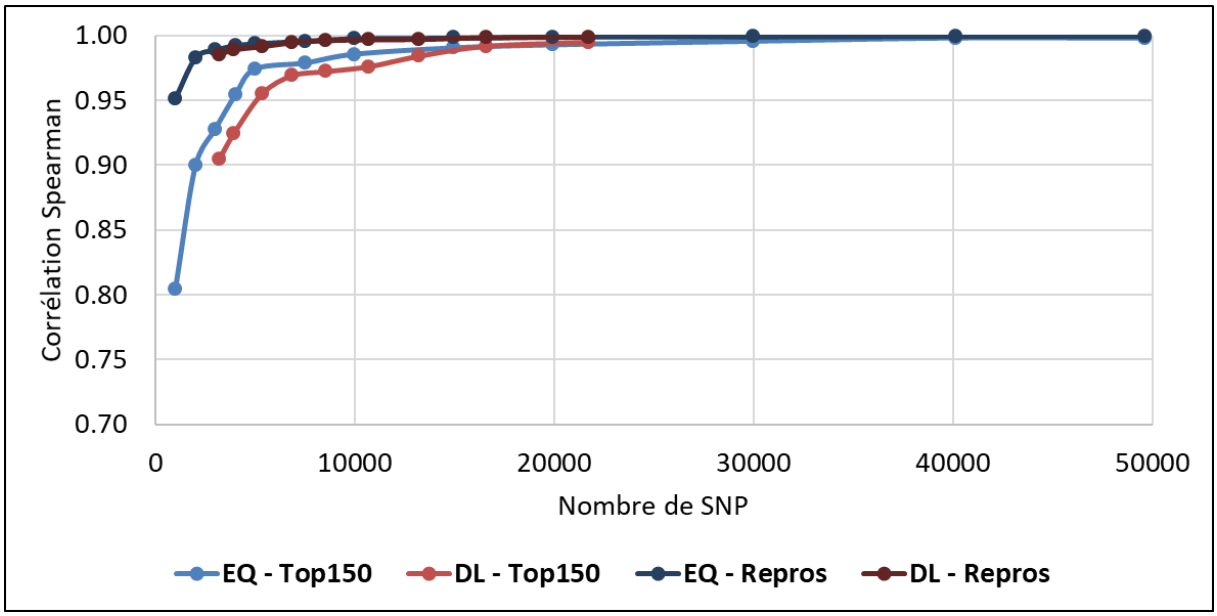


Figure 25. Corrélation de Spearman entre les GEBV estimés sur ascendance avec les vrais génotypes HD ou avec les génotypes HD imputés en fonction du nombre de SNP pour les deux méthodologies, à partir des performances des femelles en cages collectives. Les résultats sont présentés pour le poids d'œuf pour les 150 meilleurs individus et les 172 individus reproducteurs.

b) *Études sans imputation*

Les résultats concernant l'absence d'imputation sur le reclassement des individus montrent, pour les meilleurs individus comme les reproducteurs ainsi que pour les deux méthodologies, une augmentation des corrélations de Spearman avec une augmentation de la densité de SNP (Figure 26). Pour les deux méthodologies et pour les reproducteurs comme les 150 meilleurs individus par caractère, il y a une diminution des corrélations en comparaison avec celles obtenues avec imputation. Par exemple, pour les 150 meilleurs individus et les puces 1Kequi et 50Kequi, les corrélations sont de 0.8047 et 0.9979 avec imputation. Sans imputation, les corrélations sont de 0.7244 et 0.9420 (Figure 27a). De même pour les puces DL0.05 et DL0.8, avec imputation les corrélations sont de 0.9047 et 0.9947. Sans imputation, les corrélations sont de 0.7581 et 0.8590 (Figure 27b). Par ailleurs, la diminution des corrélations est plus forte pour les meilleurs individus par caractère que pour les reproducteurs. Le reclassement des meilleurs individus par caractère est donc plus impacté sans imputation. Dans le cas des meilleurs individus par caractère, cette diminution des corrélations est significative, sauf pour la méthodologie EQ pour des densités de SNP inférieures à 5K. Dans le cas des reproducteurs, cette diminution des corrélations n'est significative que pour la méthodologie DL.

En outre, comme observé avec les études en cages individuelles, la méthodologie EQ semble atteindre un seuil de corrélation de 0.95 et 0.99 au-delà de 20K SNP pour les meilleurs individus par caractère et les reproducteurs respectivement.

Enfin, comme pour les études en cages individuelles, bien que les différences entre méthodologies ne soient pas significatives, des corrélations plus élevées tendent à être obtenues avec la méthodologie EQ. Dans le cas des 150 meilleurs individus pour les puces DL0.05 et 3Kequi, la différence des corrélations est de 0.09. Pour les puces DL0.8 et 20Kequi, la différence est de 0.07. Dans le cas des 172 individus reproducteurs, la différence entre les mêmes puces sont de 0.03. Ces différences sont également plus fortes sans imputation qu'avec imputation.

Tout ceci confirme les observations faites pour les études en cages individuelles.

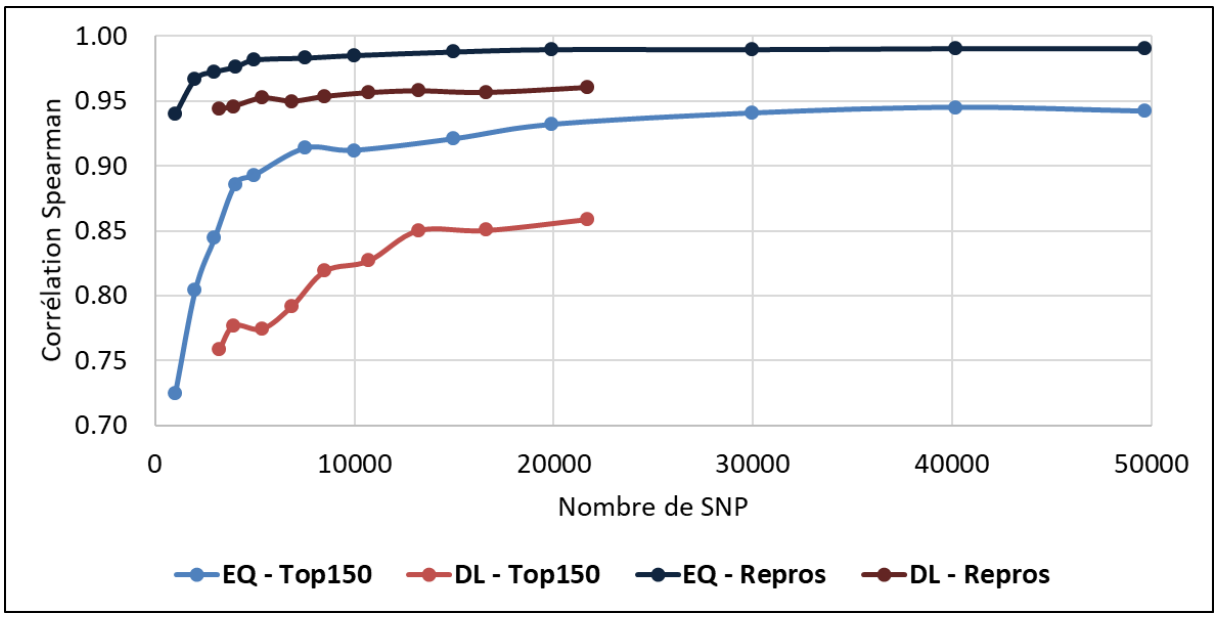


Figure 26. Corrélation de Spearman entre les GEBV estimés sur ascendance avec les vrais génotypes HD ou avec les génotypes BD non imputés en fonction du nombre de SNP pour les deux méthodologies, à partir des performances des femelles en cages collectives. Les résultats sont présentés pour le poids d'œuf pour les 150 meilleurs individus et les 172 individus reproducteurs.

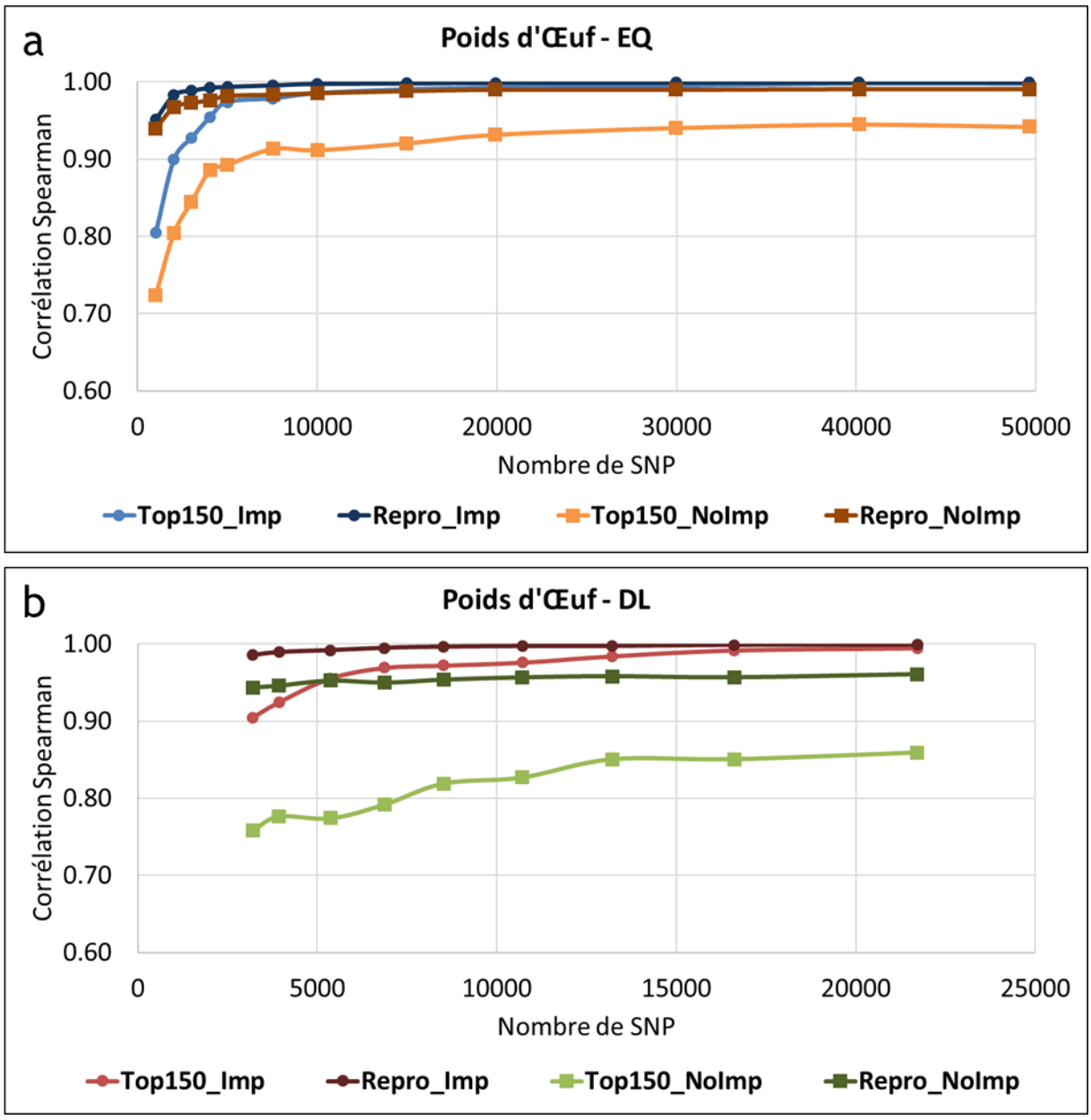


Figure 27. Comparaison des corrélations de Spearman obtenus pour les méthodologies EQ (a) et DL (b) avec ou sans imputation en fonction du nombre de SNP, à partir des performances des femelles en cages collectives. Les résultats sont présentés pour le poids d'œuf pour les 150 meilleurs individus et les 172 individus reproducteurs.

3. Intérêt de l'imputation quant à la précision relative des évaluations

L'impact des imputations ou de l'absence d'imputation sur la précision des évaluations génomiques a été étudié de la même façon que dans l'article 2. Pour cela, les GEBV « Full_HD » de l'évaluation génomique sur descendance des candidats à la sélection avec leurs vrais génotypes HD ont été comparés avec les GEBV des évaluations génomiques sur ascendance des candidats avec leurs génotypes HD imputés « Anc_Imputed » ou leur génotypes

BD sans imputation « Anc_Not_Imputed ». Les résultats ne sont présentés que pour le poids d'œuf à cause de la similarité des résultats concernant les autres caractères étudiés.

a) Études avec imputation

Les résultats illustrés avec la figure 28 montrent une légère augmentation des corrélations de Pearson pour la méthodologie EQ jusqu'à une densité de 20K SNP. Les corrélations sont ensuite stables, atteignant un palier de corrélation à 0.35. Les résultats concernant la méthodologie DL suivent également la même tendance. Par ailleurs, la corrélation obtenue par comparaison de l'évaluation génomique « Full_HD » des candidats avec l'évaluation génomique sur ascendance « Anc_HD » des candidats est de 0.35. L'erreur standard associée aux différentes puces est de ±0.07. Comme dans le cas des cages individuelles, les corrélations obtenues pour les deux méthodologies ne sont donc pas significativement différentes de la précision relative maximale. Il est cependant noté une tendance à la baisse des résultats avec une densité très faible de SNP comme illustré avec la puce 1Kequi qui ne permet d'obtenir qu'une corrélation de 0.3031.

Ceci confirme donc les résultats observés lors des études en cages individuelles.

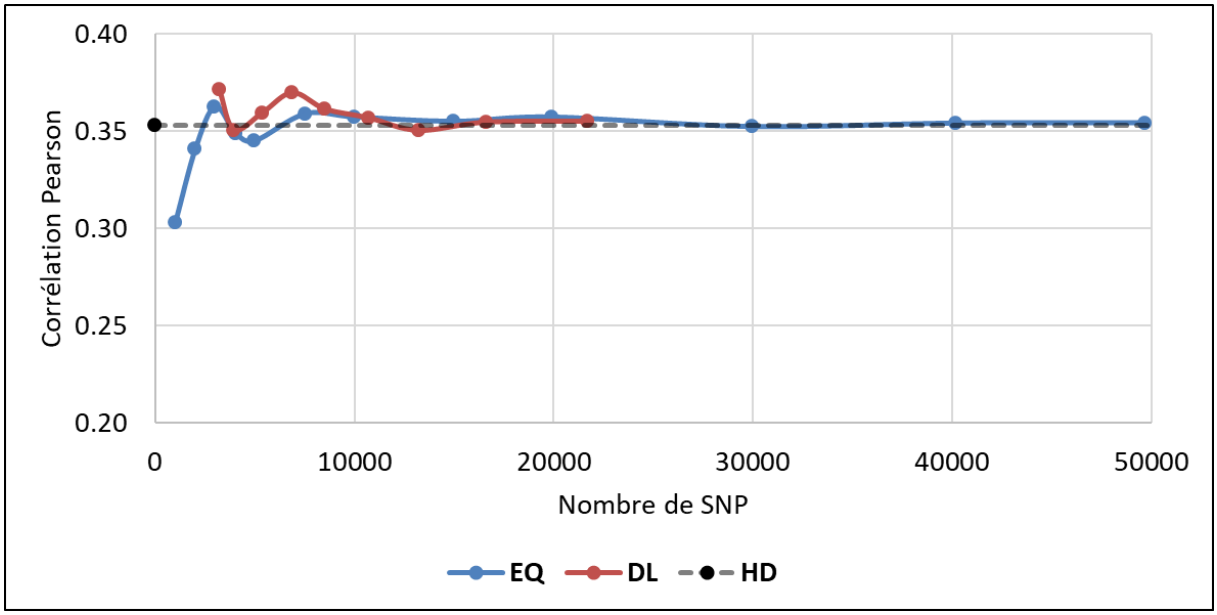


Figure 28. Corrélation de Pearson entre les GEBV "Full_HD" estimés sur descendance avec les vrais génotypes HD et les GEBV estimés sur ascendance avec les génotypes HD imputés en fonction du nombre de SNP pour les deux méthodologies, à partir des performances des femelles en cages collectives. Les résultats sont présentés pour le poids d'œuf pour les 172 individus reproducteurs.

b) *Études sans imputation*

Sans imputation, les résultats illustrés avec la figure 29 montrent de légères variations dans les corrélations pour les deux méthodologies jusqu'à une densité de 20K SNP. Puis pour des densités supérieures, les corrélations obtenues avec la méthodologie EQ se stabilisent autour d'un seuil de corrélation à 0.37. Toutefois, comme dans le cas précédent avec imputation, l'erreur standard associée aux différentes puces est de ±0.07. Les corrélations obtenues pour les deux méthodologies ne sont donc pas significativement différentes de la précision relative maximale qui est de 0.35. Ceci confirme donc encore les résultats observés lors des études en cages individuelles.

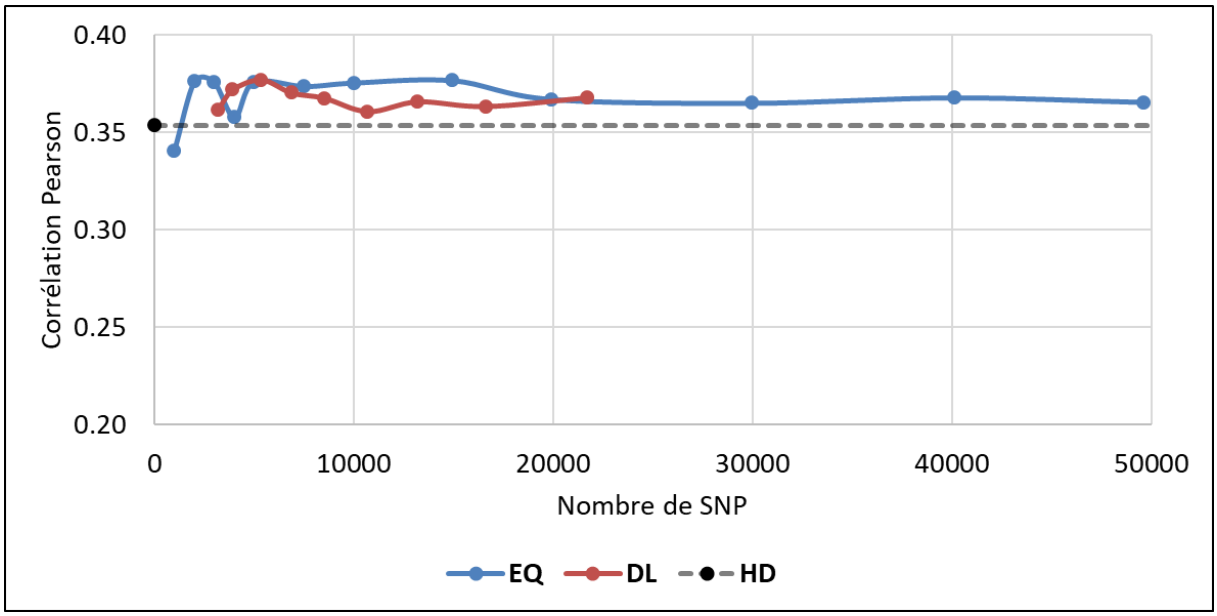


Figure 29. Corrélation de Pearson entre les GEBV "Full_HD" estimés sur descendance avec les vrais génotypes HD et les GEBV estimés sur ascendance avec les génotypes BD non imputés en fonction du nombre de SNP pour les deux méthodologies, à partir des performances des femelles en cages collectives. Les résultats sont présentés pour le poids d'œuf pour les 172 individus reproducteurs.

B. Limite des études

1. Impact de la qualité des imputations liée à la constitution de la population de référence sur évaluations génomiques des candidats

Le chapitre précédent a permis de présenter les différents facteurs influençant la qualité des imputations des génotypes obtenus à partir de puces basse densité. Parmi eux, la densité de SNP sur la puce BD ainsi que la méthodologie utilisée pour développer les puces (équidistante ou basée sur le DL) ont été étudiées dans le précédent chapitre, permettant ainsi de faire le lien

entre qualité des imputations liée à ces facteurs et qualité des évaluations génomiques des candidats.

En revanche, les facteurs concernant la structure de la population de référence n'ont pas pu être analysés dans ce chapitre. En effet, les questions du cumul d'individus dans la population de référence, de l'apparentement entre population de référence et population candidate ainsi que de la présence des mères dans la population de référence n'ont été étudiées dans le chapitre précédent qu'en considérant les générations G2 ou G3 comme population candidate. Or il est nécessaire de disposer de suffisamment de performances sur les individus des générations G3 ou G4 pour pouvoir étudier la précision des évaluations génomiques en considérant les générations G2 ou G3 comme population candidate. À la date où nous avons décidé de fixer les effectifs pour les différentes études de la thèse, nous ne disposions pas de suffisamment de performances pour les individus des générations G3 ou G4. C'est pourquoi il n'a pas été possible de faire le lien entre la qualité des imputations liée à la constitution de la population de référence et évaluations génomiques des candidats.

2. Taille réduite du nombre de reproducteurs

Dans les études précédentes, seuls 67 individus ont été sélectionnés à l'issue de la période d'élevage en cage individuelle et ont eu des descendants G2 avec performance. Il en résulte que seuls ces individus ont pu être utilisés pour étudier l'impact des imputations ou de l'absence d'imputation sur la précision relative des évaluations. Ceci a entraîné des erreurs standards élevées ne permettant pas de conclure à des différentes significatives entre les deux méthodologies ou même par rapport à la précision relative maximale.

Deux solutions permettant d'augmenter le nombre d'individus reproducteurs considérés dans la population candidate étaient toutefois possible. La première était de considérer les individus reproducteurs ayant 8 descendants ou plus dans la génération G2 (et non 10 comme cela a été le cas dans les résultats présentés). Ceci a permis de considérer 78 individus reproducteurs dans la population candidate. Aucune différence n'a été observée avec les résultats précédents. La deuxième solution était de considérer la génération G0 ainsi que les deux premiers lots de la génération G1 comme la population de référence, et le dernier lot de la génération G1 et les deux suivants de la génération G2 comme la population candidate (Picard et al., 2019). Dans ce cas, il était possible d'obtenir 87 individus reproducteurs dans la population candidate. Cette solution n'a toutefois pas été testée au cours de la thèse. Les résultats de Picard et al. (2019) montrent que des précisions plus élevées, mais non significativement différentes, sont obtenues par rapport au cas où seules les générations G0 et G1 sont respectivement utilisées en populations de référence et candidate.

IV. Bilan

Ces études ont permis de montrer que le remplacement des génotypages HD des individus reproducteurs, sélectionnés selon un index multi-caractères, par leurs génotypages BD avec ou sans imputation était possible avec un impact très faible sur le reclassement des individus. En effet, les corrélations de Spearman entre les GEBV des candidats avec leurs génotypages HD ou leurs génotypages BD sont de 0.94 avec plus de 3K SNP. Les 150 meilleurs individus pour un caractère étudié sont eux génétiquement plus proches car souvent issus des mêmes familles. Sans imputation leur reclassement est plus important avec des corrélations de Spearman entre les GEBV des candidats avec leurs génotypages HD ou leurs génotypages BD sans imputation inférieures à 0.85 avec moins de 3K SNP pour la méthodologie EQ et moins de 16K SNP (DL0.6) pour la méthodologie DL. Bien que les différences entre les corrélations des GEBV obtenues avec les deux méthodologies ne soient pas significativement différentes, des résultats plus élevés ont tendance à être obtenus avec la méthodologie EQ. Par ailleurs, pour les deux méthodologies avec ou sans imputation, les études ont montré que l'impact de l'utilisation des puces BD en remplacement des puces HD sur la précision des évaluations génomiques n'était pas significatif. Toutefois ce calcul de précision n'a pu être effectué que pour les individus reproducteurs, les individus du top 150 par caractère n'ayant généralement pas de descendance. On peut penser que la précision des GEBV des meilleurs individus par caractère serait plus affecté par l'utilisation des génotypages BD.

En conclusion, nous pouvons tout de même dire que les puces développées selon une méthodologie EQ avec plus de 5K SNP permettent d'obtenir de bonnes évaluations génomiques des candidats à la sélection pour des coûts de génotypages intéressants. Il est également à noter que les quatre caractères étudiés sont contrôlés par de nombreux petits QTL. Couplé à une méthodologie ssGBLUP plus intéressante que des méthodes bayésiennes pour ce type de caractère, ceci peut expliquer pourquoi la méthodologie EQ est plus intéressante que la méthodologie DL. Il reste tout de même à comprendre plus en profondeur pourquoi les meilleures imputations avec la méthodologie DL ne sont pas synonymes de meilleurs résultats d'évaluations génomiques comparés à ceux obtenus avec la méthodologie EQ. Des éléments de réponses ont été trouvés au cours de la thèse et sont développés dans le chapitre V.

Chapitre IV. De l'analyse de diversité génétique à l'optimisation du design d'une puce BD pour plusieurs lignées

I. Introduction

Les années 2000 ont été marquées par la découverte de marqueurs moléculaires tels que les Single Nucleotide Polymorphism (SNP) correspondant à une variation d'une base nucléotidique de l'ADN. L'utilisation et l'intérêt portés à ces marqueurs ont débouché sur la création d'une puce commerciale haute densité (HD) de 600 000 marqueurs développée par Kranis et al. (2013). Cette puce a été développée à partir de 24 lignées différentes (lignées expérimentales et commerciales, pontes et chairs) permettant de génotyper un grand nombre de lignées différentes tout en permettant son utilisation en sélection génomique, analyses d'associations et cartographie fine de QTL dans les filières ponte et chair. Toutefois, le nombre de SNP informatifs pour une lignée précise est compris entre 100K et 450K SNP en fonction de la lignée étudiée, illustrant ainsi le fait qu'un grand nombre de SNP de cette puce restent non informatifs pour une lignée précise. Ce point a également été illustré chez les bovins où les 777K SNP de la puce HD ne sont pas tous informatifs en fonction des races étudiées (Illumina Inc, 2012 ; Pérez O'Brien et al., 2014). Enfin, les coûts d'une telle puce HD restent élevés. Il peut donc être intéressant d'essayer de développer, à moindre coût, une puce basse densité (BD) en sélectionnant un sous-ensemble de marqueurs de la puce HD utile pour l'ensemble des lignées à étudier. Grâce à l'imputation, il est possible de remonter à l'information de la puce HD.

Il existe trois principales méthodes pour sélectionner le sous-ensemble de marqueurs de la puce HD. Ce sont des méthodes de sélection des SNP propre à chaque population ou lignée étudiée et sont donc dépendantes des fréquences alléliques de la population ou de la lignée étudiée. La première consiste à sélectionner des SNP à intervalles réguliers le long de chaque chromosome en fonction ou non de leur fréquence allélique mineure (MAF) (Habier et al., 2009; Weigel et al., 2009; Zhang et al., 2011; Cleveland & Hickey, 2013; Wang et al., 2013; Herry et al., 2018). La deuxième consiste à sélectionner des SNP en fonction de leurs effets sur différents caractères d'intérêts (Weigel et al., 2009, Zhang et al., 2011). Enfin, la troisième est de sélectionner des SNP en fonction du déséquilibre de liaison (DL) entre marqueurs (Herry et al., 2018).

Par ailleurs, l'efficacité d'une puce dépend également de sa capacité à exploiter le déséquilibre de liaison du génome (Sargolzaei et al., 2008). Le DL correspond à l'association non-aléatoire entre allèles de deux loci. Ce DL est indispensable pour les études d'associations entre allèles à un locus marqueur et un locus impliqué dans la variation d'un caractère quantitatif, ainsi que pour la cartographie fine de QTL (Quantitative Trait Loci). De nombreuses études ont montré les particularités du génome avicole avec des chromosomes de tailles différentes, ainsi qu'une structure du DL particulière entre chromosomes mais aussi entre lignées Rhode Island et Leghorn (Megens et al., 2009 ; Qanbari et al., 2010). En effet, les macro-chromosomes sont caractérisés par une forte persistance du DL alors que les micro-chromosomes présentent une persistance plus faible du DL. Enfin, les lignées Rhode Island ont plus de marqueurs polymorphes et une persistance du DL plus faible sur l'ensemble des chromosomes que les lignées Leghorn ayant moins de marqueurs polymorphes. Ainsi, avec un DL s'étendant sur de longues distances, il sera plus facile de retrouver une association significative entre caractère d'intérêt et certains marqueurs. En revanche, la qualité de la cartographie sera plus faible lorsque le DL s'étend sur de longues distances pour identifier précisément le polymorphisme causal associé avec le caractère d'intérêt (Fu et al., 2015). Il est donc important de bien comprendre le DL des différentes lignées avant de vouloir développer une puce à SNP BD, de façon à exploiter au mieux le DL des différentes lignées. L'étude du DL et de la diversité génétique des lignées utilisées sont donc nécessaires pour construire une puce BD multi-lignée. Megens et al. (2009) ont préconisé une densité d'au moins 100K pour exploiter au mieux le DL des différentes lignées avicoles. Toutefois, plusieurs études (Dassonneville et al., 2011 ; Bouquet et al., 2015 ; Herry et al., 2018) ont montré pour diverses puces BD la possibilité de remonter à l'information de la puce HD grâce à l'imputation, avec un impact faible sur les évaluations génomiques avec plus de 5K SNP (Herry et al., 2019).

Ainsi, le premier objectif de cette étude est d'analyser le DL et la diversité génétique de 5 lignées pures de poules pondeuses de souches Rhode Island et Leghorn. À partir des résultats obtenus, le deuxième objectif est de réaliser une puce BD (moins de 50K SNP) pour l'ensemble des lignées étudiées puis de valider la puce BD en étudiant la qualité de l'imputation et l'impact de la puce, imputée ou non, sur les évaluations génomiques.

II. Matériels et méthodes

A. Populations d'études

Les animaux étudiés sont issus de trois lignées pures commerciales de poules pondeuses Rhode Island (RI) et de deux lignées pures commerciales de poules pondeuses Leghorn (L). Ces

lignées ont été créées et sélectionnées par Novogen. Chaque génération est constituée de trois lots. Les lignées RI1 et L2 correspondent à celles étudiées dans les travaux de Herry et al. (2018). Les effectifs de chaque lignée sont détaillés dans le tableau 7.

Tableau 7. Récapitulatif des effectifs des différentes lignées avant contrôle qualité

Lignées	RI1	RI2	RI3	L1	L2
Données	2370 individus	301 individus	650 individus	674 individus	1483 individus
Génération G0	447♂	100♂ ; 201♀	250♂ ; 400♀	261♂ ; 413♀	291♂ ; 423♀
Génération G1	580♂				271♂ ; 498♀
Génération G2	132♂ ; 665♀				
Génération G3	55♂ ; 491♀				

B. Génotypages

Des échantillons de sang sont prélevés au niveau de la veine brachiale des animaux. L'ADN est extrait puis hybridé en utilisant la puce 600K Affymetrix® Axiom® HD. Les génotypages des deux premières générations d'individus RI1 sont obtenus par le laboratoire Ark-Genomics (Édimbourg, Royaume-Uni). Les génotypages des générations suivantes et des autres lignées sont obtenus par la plateforme de génotypage et séquençage à haut-débit Gentyane (Clermont-Ferrand, France). Le traitement des génotypages issus de deux laboratoires différents a été détaillé dans Herry et al. (2018).

L'ensemble des animaux est génotypé pour 580 961 SNP. Conformément au cinquième assemblage du génome Gallus gallus (Warren et al., 2017), ces SNP sont répartis sur les macro-chromosomes (1 à 5), chromosomes intermédiaires (6 à 10), micro-chromosomes (11 à 28 et 33), un groupe de liaison (LGE64), deux chromosomes sexuels Z et W, ainsi qu'un groupe de 3724 SNP avec une localisation inconnue.

Un premier contrôle qualité a été réalisé sur chaque lignée indépendamment les unes des autres avec le logiciel Plink V1.9 (Chang et al., 2015). Les génotypes sont filtrés en 6 étapes successives selon le call rate individu (<95%), la MAF (<0.05), le call rate SNP (<95%), l'équilibre de Hardy-Weinberg ($P<10^{-4}$). Les SNP restant avec une position inconnue ou localisés sur le chromosome W sont supprimés. Enfin, pour les lignées RI1 et L2, les individus avec des incompatibilités de pedigree sont supprimés. L'identification de problème de compatibilité de pedigree n'est pas possible pour les lignées RI2, RI3 et L1 car une seule génération d'individus a été génotypée pour ces lignées.

Enfin, un dernier contrôle qualité a été réalisé en considérant les génotypages des 5 lignées comme les génotypages d'une seule et même lignée (« Full »). Les génotypes ont été filtrés

selon les mêmes étapes que précédemment. Toutefois, la filtration selon l'équilibre de Hardy-Weinberg a été réalisée en considérant uniquement l'intersection des 5 listes de SNP déviant de l'équilibre de Hardy-Weinberg obtenues lors des contrôles qualités propre à chaque lignée. Ce contrôle qualité permet ainsi d'obtenir des génotypages pour 483 749 SNP pour l'ensemble des 5450 individus restants. Le récapitulatif des différentes étapes de contrôle qualité est présenté dans le tableau 8.

Tableau 8. Récapitulatif des différentes étapes de contrôle qualité

	RI1	RI2	RI3	L1	L2	Full
Call Rate Individu (<95%)	8	0	1	16	3	17
MAF (=0)	204 122	224 020	170 319	199 808	228 452	10 680
MAF (<0.05)	54 650	50 860	64 494	162 360	99 000	79 828
Call Rate SNP (<95%)	7541	4044	1922	2078	2530	3459
Équilibre de Hardy Weinberg ($P<10^{-4}$)	12 538	2990	3333	3831	3857	51
SNP avec position inconnue ou sur le chromosome W	1759	1713	2028	1283	1453	3194
Incompatibilité pedigree	0				6	6
SNP retenus pour les analyses	300 351	299 660	338 865	211 601	245 667	483 749
Animaux retenus pour les analyses	2362	301	649	658	1474	5450

C. Mesure des caractères (en CI)

Concernant la lignée RI1, des mesures de performances sur le poids d'œuf (PO), la force de fracture (FF), la couleur de la coquille des œufs (Lab) ainsi que la hauteur d'albumen (HA) ont été collectées entre 60 et 90 semaines. Ces données correspondent à des données individuelles. Les caractères sont nommés selon la classification Animal Trait Ontology for Livestock (Atol Ontology, 2012). Au fil des générations G0 à G3, 75 121 œufs concernant 7983 poules pondeuses ont été mesurés pour cette période d'élevage. Ces caractères n'ont pas été étudiés dans les autres lignées à cause d'un nombre insuffisant de générations disponibles pour réaliser des évaluations génomiques.

L'ensemble des œufs produits durant cette période ont été collectés puis transférés à Zootests (Ploufragan, France) afin de réaliser les mesures de qualité d'œufs. Le poids des œufs (PO, en g) est mesuré en premier. Trois caractères concernant la couleur de la coquille sont ensuite mesurés avec un chromamètre Minolta : la couleur rouge (a*), la couleur jaune (b*) et la clarté (L*). Le Lab est calculé en appliquant la formule $Lab = 100 - (L^* - a^* - b^*)$. La troisième étape consiste à mesure la force de fracture de l'œuf (FF, en N) en utilisant une machine de

compression qui permet de déterminer la valeur de la force minimale à appliquer pour fracturer la coquille. Enfin, chaque œuf est cassé et la hauteur d'albumen (H) est mesuré avec un tripode.

D. Analyse du déséquilibre de liaison

Le déséquilibre de liaison a été étudié dans l'ensemble des 5 lignées en utilisant les génotypages obtenus après contrôle qualité propres à chaque lignée. Le DL est calculé en utilisant le r^2 qui permet d'estimer la corrélation entre chaque paire de SNP de chaque chromosome (Hill et Robertson, 1968 ; Pritchard et Przeworski, 2001). Le D' de Lewontin (1988) peut également être utilisé pour estimer le DL mais plusieurs études ont montré les limites de cette mesure. En effet, le D' dépend du nombre d'animaux utilisés et est fortement surévalué pour les SNP avec des fréquences alléliques faibles (Ardlie et al., 2002 ; Weiss et Clark, 2002 ; Aerts et al., 2007 ; Sargolzaei et al., 2008), permettant ainsi d'observer un certain niveau de DL pour des marqueurs qui sont en réalité en équilibre. Le r^2 est moins sensible aux fréquences alléliques et est donc une mesure plus robuste du DL (Du et al., 2007). Le r^2 est calculé comme suit :

$$r^2 = \frac{D^2}{f(A) * f(a) * f(B)f(b)}$$

avec $D = f(AB)f(ab) - f(Ab)f(aB)$. $f(AB)$ représente pour la population étudiée la fréquence de l'haplotype AB, $f(ab)$ la fréquence de l'haplotype ab, $f(Ab)$ la fréquence de l'haplotype Ab et $f(aB)$ la fréquence de l'haplotype aB. $f(A)$, $f(B)$, $f(a)$ et $f(b)$ représentent respectivement les fréquences des allèles A, B, a et b.

La mesure du r^2 pour chaque paire de SNP est estimée avec la fonction --r² de Plink. Il est également indiqué le numéro et la position (en kb) du SNP le plus éloigné sur le chromosome étudié avec les fonctions --ld-window et --ld-window-kb respectivement, de façon à calculer le DL pour chaque paire de SNP, quelle que soit la distance entre SNP. L'étendue du DL est ensuite estimée pour chaque population et chaque chromosome (ou type de chromosome) en calculant le r^2 moyen pour différents intervalles (de 25kb à 10Mb). Dans le cas particulier du chromosome sexuel Z, seuls les mâles ont été utilisés pour analyser le chromosome.

E. F-Statistiques

Les coefficients de consanguinité F_{IS} sont calculés en utilisant les génotypages de chaque lignée après un contrôle qualité spécifique à chaque lignée. Le coefficient de consanguinité F_{IS} caractérise le déficit en hétérozygote par rapport à la valeur attendue. Ce coefficient est calculé selon la formule $F_{IS} = 1 - \frac{H_O}{H_E}$ avec H_O et H_E caractérisant respectivement les taux

d'hétérozygoties observé et attendu. Le coefficient F_{IS} est directement calculé pour chaque individu par Plink V1.9 avec la fonction --het avec une formule légèrement différente :

$$F_{IS} = \frac{Nb\ marqueurs\ homozygotes\ observés - Nb\ marqueurs\ homozygotes\ attendus}{Nb\ marqueurs\ totaux - Nb\ marqueurs\ homozygotes\ attendus}$$

Le coefficient F_{IS} moyen de la population est ensuite calculé en moyennant les F_{IS} calculés pour chaque individu. Les différences de F_{IS} entre lignées et par rapport à un F_{IS} nul sont testées avec des tests de Student au seuil de première espèce α = 1%.

L'indice de différenciation F_{ST} de Wright (1951) est calculé en utilisant les génotypages « Full » obtenus après contrôle qualité. L'indice de différenciation F_{ST} permet d'estimer le niveau de différenciation entre paires de populations. Plus la valeur de l'indice est élevée, plus les populations sont génétiquement différentes. Cet indice de différenciation F_{ST} est calculé par paires de population sur les SNP génotypés dans les deux populations, ce qui explique l'utilisation des génotypages « Full » après contrôle qualité. L'indice F_{ST} est également calculé par Plink V1.9 avec la fonction --fst d'après la méthode développée par Weir et Cockerham (1984).

F. Analyse en Composantes Principales

L'analyse en Composantes Principales (ACP) permet d'agréger l'information provenant d'un grand nombre de marqueurs dans un nombre réduit de variables synthétiques et d'obtenir une représentation spatiale de la diversité génétique des différentes lignées.

L'ACP est réalisée en utilisant les génotypages « Full » obtenus après contrôle qualité, permettant d'avoir les mêmes marqueurs génotypés pour tous les individus. Les coordonnées de la projection de chaque individu selon les 20 premières composantes de la variance sont obtenues avec Plink V1.9 grâce à la fonction --pca. Le pourcentage d'inertie expliqué par chaque axe est obtenu en divisant la valeur propre associée à chaque axe par la somme des valeurs propres. Les graphiques représentant la position de chaque individu selon les deux principaux axes sont réalisés sous R (R Core Team, 2017).

G. Analyse de structuration des différentes lignées

Une analyse de structuration des différentes lignées est réalisée avec le logiciel ADMIXTURE (Alexander et al., 2009). À partir de données de génotypages, le logiciel calcule simultanément des proportions d'ascendance et des corrélations entre fréquences alléliques de population, et permet d'estimer le nombre de populations ancestrales *K* en utilisant la méthode du maximum de vraisemblance.

Les analyses sont réalisées avec les génotypages « Full » obtenus après contrôle qualité.
Le logiciel est utilisé pour un nombre de population ancestrales allant de $K = 2$ à $K = 9$. Le nombre optimal de clusters, et donc de populations ancestrales, est déterminé pour le nombre de clusters K ayant l'erreur de validation croisée la plus faible. Les graphiques représentant les proportions ancestrales pour chaque individu sont réalisés sous R (R Core Team, 2017).

H. Design des puces BD

À partir des génotypages obtenus après contrôle qualité, plusieurs puces basse densité ont été développées in-silico en sélectionnant un sous-ensemble de SNP. Cinq densités différentes ont été testées : 10K, 20K, 30K, 40K et 50K SNP. Suite aux précédentes études (Herry et al., 2018 ; Herry et al., 2019), la sélection des SNP est faite selon une méthodologie équidistante. Deux stratégies de sélection des SNP sont possibles en fonction des résultats concernant l'analyse du DL et la diversité génétique sur l'ensemble des individus (Figure 30) :

- La première stratégie « Indep » consiste à considérer K groupes de population et à sélectionner un nombre équivalent de SNP par groupe de population après un contrôle qualité des génotypages intra-lignée. Les SNP sont sélectionnés à intervalles réguliers le long de chaque chromosome, en choisissant pour chaque intervalle le SNP avec la MAF la plus élevée. Dans le cas d'un intervalle avec plusieurs SNP en compétition pour la sélection, le SNP le plus à gauche de l'intervalle est sélectionné.
- La deuxième stratégie « Multi » consiste à considérer un seul groupe de population et à sélectionner le nombre total de SNP à partir des génotypages « Full » obtenus après un contrôle qualité commun aux différentes lignées. Les SNP sont également sélectionnés à intervalles réguliers le long de chaque chromosome, en choisissant pour chaque intervalle le SNP avec la MAF moyenne des 5 lignées la plus élevée et la variance de MAF des 5 lignées la plus faible. Dans le cas d'un intervalle avec plusieurs SNP en compétition pour la sélection, le SNP le plus à gauche de l'intervalle est sélectionné.

I. Qualité des imputations

Une fois les puces développées, elles sont testées in-silico pour les lignées indépendamment les unes des autres. Les imputations sont ensuite lignées spécifiques.
Dans cette étude, seules les lignées RI1 et L2 ont permis de réaliser des imputations grâce à la présence de plusieurs générations d'individus. Pour les deux lignées, les candidats à la sélection sont les individus de la génération G1 avec des génotypages BD et imputés à partir des génotypages HD des individus de la génération G0. Pour la lignée RI1, ces individus G0 sont

les pères ou les demi-frères des pères des candidats à la sélection G1. Pour la lignée L2, ce sont les pères, les mères, ou les collatéraux des parents des candidats à la sélection G1. Enfin, les génotypages HD correspondent aux génotypages obtenus après contrôle qualité, propres à chaque lignée et non aux génotypages « Full » (Figure 1).

D'après les recommandations de Hickey et al. (2012) et Calus et al. (2014), la qualité des imputations est estimée en calculant la corrélation de Pearson moyenne par SNP entre les vrais génotypages HD et les génotypages HD imputés. Les corrélations moyennes obtenues sont comparées pour chaque puce basse densité et chaque stratégie en réalisant des tests de Student au seuil de première espèce $\alpha = 0.1\%$.

FImpute V2.2 (Sargolzaei et al., 2014) est utilisé pour imputer les génotypages BD des candidats à la sélection à partir des génotypages HD des individus de la génération G0.

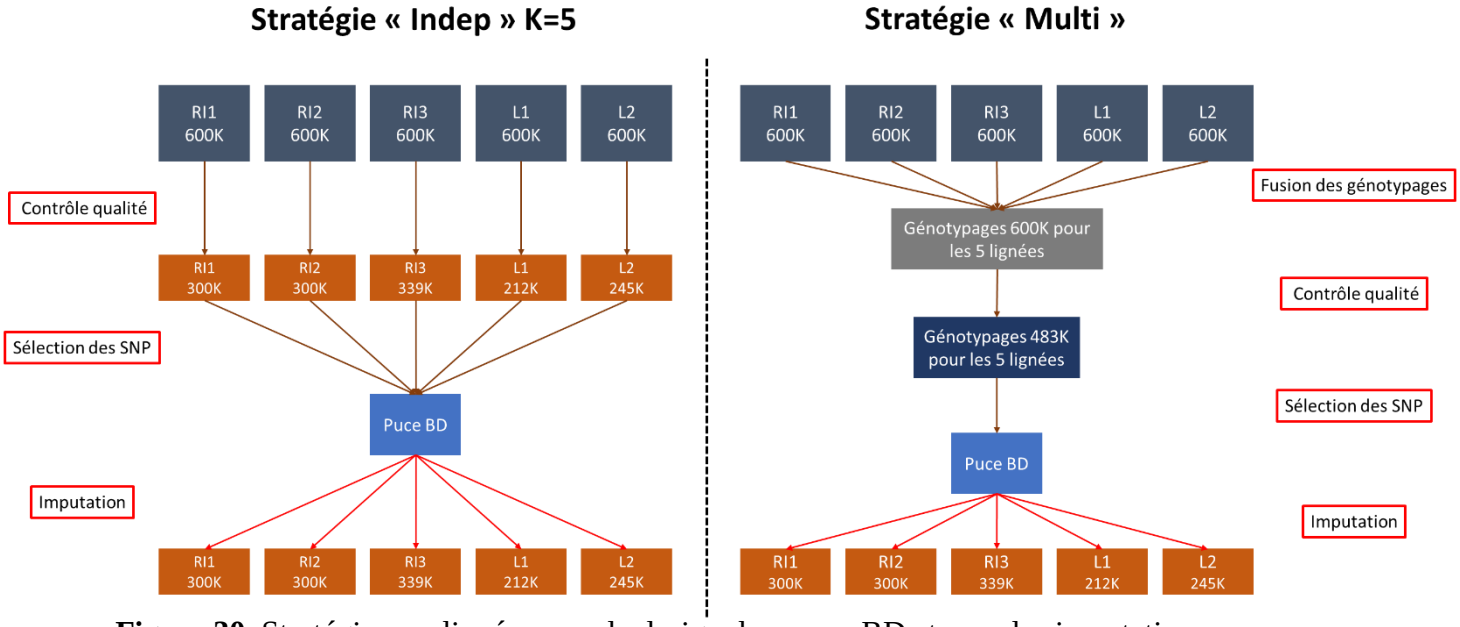


Figure 30. Stratégies appliquées pour le design des puces BD et pour les imputations

J. Évaluations génomiques (top150 et 67 mâles repros)

Au-delà d'une bonne qualité d'imputation, l'objectif pour tout sélectionneur est d'obtenir de bons résultats d'évaluations génomiques à partir des puces BD développées. Les travaux de Herry et al. (2019) ont montré un impact faible des erreurs d'imputations sur les évaluations génomiques des candidats à la sélection. Ils ont aussi mis en évidence une possible utilisation directe, sans imputation, des puces BD pour des densités supérieures à 2K SNP. La dernière étape de validation des puces BD développées est donc de tester, avec ou sans imputation, leur impact sur les évaluations génomiques des candidats à la sélection.

Le poids d'œuf (PO), la force de fracture (FF), la couleur de la coquille des œufs (Lab) ainsi que la hauteur d'albumen (AH) ont été évalués avec la méthode du Single Step GBLUP

(Legarra et al., 2009). Ces caractères sont évalués simultanément selon un modèle animal multi-caractère. Les évaluations ne concernent que la lignée RI1. En effet, en considérant les individus G1 comme candidats à la sélection, avec la présence d'ascendants (G0) et de descendants (G2 et G3) avec performances, il est possible d'estimer la vraie valeur génétique de chaque candidat à la sélection G1 et ainsi de comparer cette valeur avec les résultats des évaluations génomiques. Une évaluation génomique sur descendance « Full_HD » des individus G1 avec descendance est réalisée en utilisant l'ensemble des informations disponibles (génotypages et performances des ascendants, collatéraux et descendants) pour estimer les paramètres génétiques du modèle. Remlf90 (Misztal et al., 2002) est utilisé pour estimer les composantes de la variance, génétique et résiduelle. Une fois les composantes fixées, les différentes évaluations génomiques sont réalisées avec Blupf90.

Le premier objectif est d'étudier l'impact des erreurs d'imputation sur les évaluations génomiques sur ascendance. Une évaluation génomique sur ascendance « Asc_HD » est réalisée avec les génotypages HD des mâles G0 et des candidats, ainsi que les phénotypes des femelles G0. Une deuxième évaluation « Asc_Imp » est réalisée en remplaçant pour les candidats les génotypages HD par les génotypes BD imputés. Pour chaque puce BD, des corrélations de Spearman sont calculées entre les vraies GEBV HD et les GEBV BD imputées pour 67 mâles reproducteurs G1 avec au moins 10 descendantes mesurées ainsi que pour les 150 meilleurs individus G1 selon leur GEBV HD, en fonction du caractère d'étude. En effet, les 67 individus reproducteurs sont choisis selon un index-multi-caractère et sont donc différents des 150 meilleurs individus pour un caractère.

Le deuxième objectif vise à étudier l'impact des imputations sur la précision des évaluations génomiques. Pour cela, une évaluation génomique sur descendance « Full_HD » des individus G1 avec descendance est réalisée en utilisant l'ensemble des informations disponibles (génotypages et performances des ascendants, collatéraux et descendants). Pour chaque puce BD, des corrélations de Pearson sont calculées entre les GEBV « Full_HD » et les GEBV BD imputées estimées sur ascendance pour les 67 mâles reproducteurs G1.

Enfin, il est également intéressant d'étudier les possibles conséquences d'une utilisation directe des génotypages BD sans imputation sur les évaluations génomiques. Pour cela, les deux objectifs précédents sont repris en utilisant directement les génotypages BD de l'ensemble des animaux.

III. Résultats et discussion

A. Analyse du déséquilibre de liaison

L'étendue du DL a été étudiée pour les cinq lignées après contrôle qualité des génotypages propres à chaque lignée. Le r^2 moyen en fonction de la distance entre marqueurs (Mb) a été mesuré pour les 5 lignées sur l'ensemble du génome et pour les quatre types de chromosomes (Figure 31). Il est noté, pour chaque lignée, pour chaque type de chromosome et donc sur l'ensemble du génome, une diminution du DL avec une augmentation de la distance entre marqueurs. En effet, pour des distances inférieures à 20kb, le DL moyen est compris entre 0.48 pour la lignée RI3 et 0.67 pour la lignée L1. Pour des distances supérieures à 5Mb, le DL moyen est compris entre 0.01 et 0.02 pour les 5 lignées. Enfin, le DL moyen calculé pour les lignées RI est inférieur ou égal à celui calculé pour les lignées Leghorn. Ceci est cohérent avec les résultats de Qanbari et al. (2010). Pour une distance entre marqueurs inférieure à 25kb, le DL moyen calculé est de 0.32 pour les individus de souche Rhode Island/White Rock contre 0.73 pour les individus de souche Leghorn. Pour une distance supérieure à 5Mb, le DL moyen chute à 0.01 et 0.03 pour les Rhode Island/White Rock et Leghorn respectivement. Il y a donc une première distinction entre les souches RI et L. Cette chute de DL avec la distance entre SNP est également observée chez les poulets de chair. En effet, Hérault et al. (2018) ont calculé un DL moyen de 0.34 pour des distances entre marqueurs inférieures à 25kb et de 0.01 pour des distances entre marqueurs supérieures à 5Mb. Il est toutefois noté un niveau de DL plus faible que celui obtenu pour les souches RI et L entre marqueurs proches (<25kb). Ces chutes de DL sont fortement corrélées aux taux de recombinaison, aux méthylations et à la présence d'îlots CpG (Pengelly et al., 2016). De la même façon, chez les porcins, Badke et al. (2012) ont calculé un DL moyen compris entre 0.27 et 0.36 pour des distances entre marqueurs inférieures à 100kb pour des individus Landrace et Duroc, respectivement. Pour des distances supérieures à 5Mb, le DL moyen chute à 0.06 pour les deux races. Enfin, Pérez O'Brien et al. (2014) montrent, pour différentes races de bovins, une chute du DL moyen compris entre 0.39 à 0.59 pour des distances entre marqueurs inférieures à 5kb à un DL moyen compris entre 0.04 à 0.06 pour des distances supérieures à 5Mb.

Des différences de r^2 moyen sont également observées entre type de chromosomes. Quelle que soit la lignée étudiée, le DL moyen est plus élevé sur les macro-chromosomes, que sur les chromosomes intermédiaires, présentant eux-mêmes un DL moyen plus élevé que sur les micro-chromosomes. Ces résultats sont également retrouvés chez les poulets de chair (Fu et al., 2015) ainsi que pour des poulets de races traditionnelles ou des poulets villageois (Wragg et al., 2012 ; Khanyile et al., 2015). L'International Chicken Genome Sequencing Consortium (2004) a

montré que la taille des chromosomes était inversement corrélée au taux de recombinaisons, aux méthylations et à la présence d'îlots CpG. Ces travaux ont été complétés par ceux de Groenen et al. (2009) montrant la présence de points chauds de recombinaison principalement sur les micro-chromosomes. La chute du DL étant corrélée à tous ces éléments, cela explique les différences d'étendue du DL observée entre type de chromosomes. Ces différences observées entre chromosomes sont également retrouvées chez les bovins (Sargolzaei et al., 2008 ; Lu et al., 2012) mais dans une moindre mesure, le rapport de taille entre les chromosomes les plus extrêmes étant de 3.5, quand il est de 200 en poule pondeuse.

Le cas du chromosome Z est particulier avec un DL moyen supérieur à celui observé pour les macro-chromosomes, quelle que soit la lignée étudiée. La taille du chromosome Z est de 82Mb et un faible nombre de SNP est retenu à l'issu du contrôle qualité (entre 4954 SNP pour la lignée L2 et 10 113 SNP pour la lignée RI1). Ceci peut s'expliquer par le nombre réduit de mâles reproducteurs utilisés pour chaque lignée et des taux de recombinaison proche de ceux des macro-chromosomes (Wahlberg et al., 2007). Il y a donc finalement peu de nouveaux haplotypes ce qui peut entraîner, au fur et à mesure des années de sélection, de grands intervalles sans aucun SNP informatif. Ceci explique le DL moyen supérieur à celui des macro-chromosomes pour les 5 lignées étudiées.

Enfin, il est d'usage de considérer un r^2 moyen supérieur à 0.3 comme un seuil de DL utile pour réaliser des études d'association ou pour la sélection génomique (Ardlie et al., 2002 ; Aerts et al., 2007). Le r^2 donne une indication du pouvoir de détection d'une association entre caractère d'intérêt et marqueurs et permet ainsi d'estimer le nombre d'individus ($1/r^2$) nécessaires pour détecter une association avec le polymorphisme causal directement. Une valeur de $r^2>0.3$ est la plus souvent choisie car elle est un bon compromis entre augmentation du nombre d'individus nécessaire pour détecter l'association avec le polymorphisme causal directement et force de la corrélation entre marqueurs pour la distance correspondante au r^2.

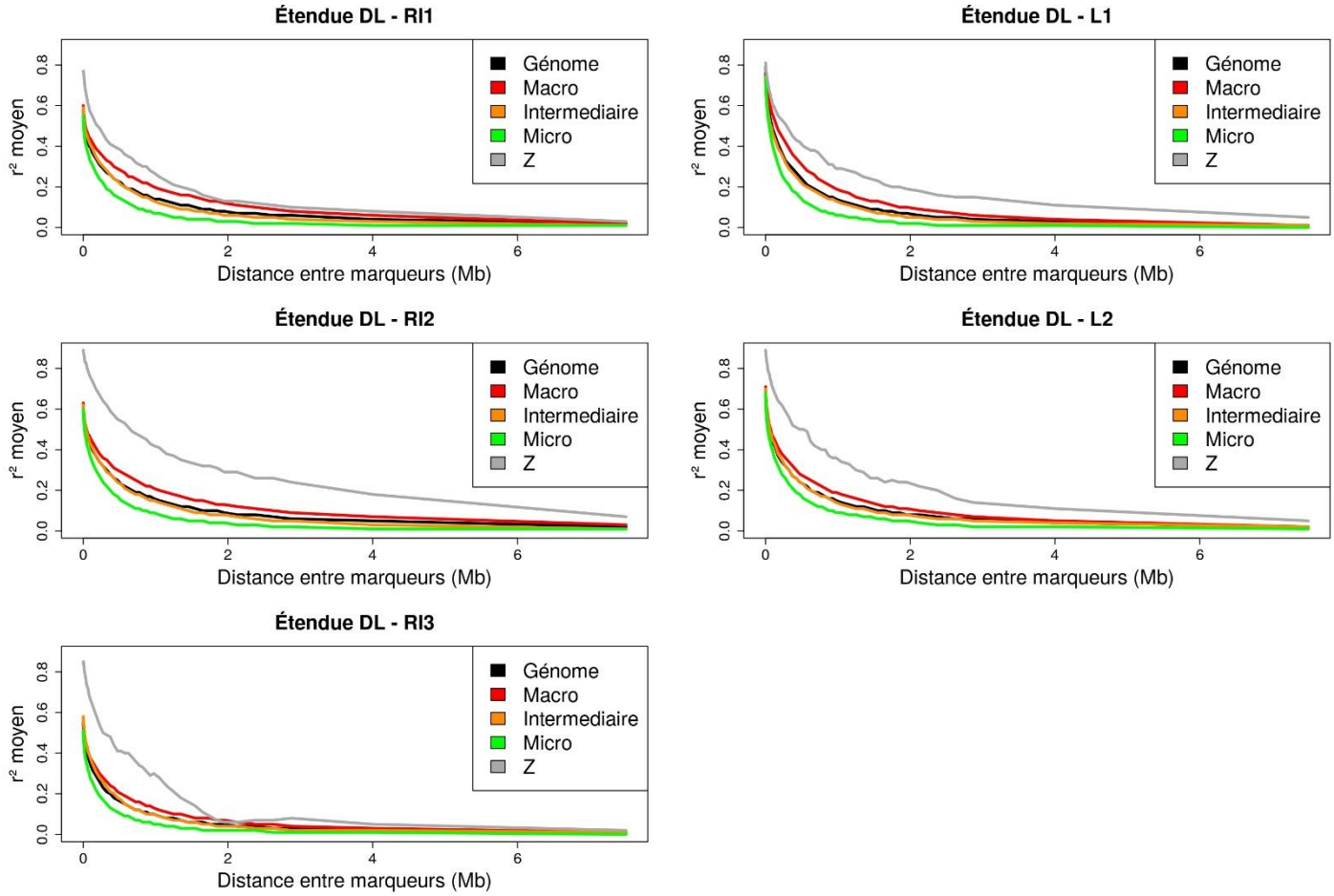


Figure 31. Étendue du DL à l'échelle du génome et des différents types de chromosomes, pour l'ensemble des 5 lignées étudiées.

L'étendue du DL utile (r^2>0.3) pour le génome et les différents types de chromosomes, pour les 5 lignées étudiées est présentée dans le tableau 9. À l'échelle du génome, l'étendue du DL utile est inférieure à 250kb pour les lignées RI1 et RI3 et supérieure à 300kb pour les lignées RI2, L1 et L2. De même, hormis le chromosome Z, on note les mêmes distinctions entre les lignées RI1 et RI3 et les lignées RI2, L1 et L2. Ces résultats sont en accord avec ceux de Qanbari et al. (2012) et Hérault et al. (2018) montrant des étendues de DL utile plus grandes pour les lignées Leghorn que les lignées Rhode Island. Toutefois, dans notre étude, la lignée RI2 semble plus proche des lignées Leghorn que des lignées RI1 et RI3. Le chromosome Z est encore particulier avec les lignées RI1, RI3 et L1 ayant une étendue du DL utile inférieure à 1000kb et les lignées RI2 et L2 ayant une étendue du DL utile supérieure à 1200kb. La distinction précise des souches observée précédemment ainsi que dans les études Qanbari et al. (2012) et Hérault et al. (2018) n'est donc plus aussi tranchée.

Tableau 9. Étendue du DL utile (r^2>0.3) pour les 5 lignées

	RI1	RI2	RI3	L1	L2
Génome	200-250 kb	300-350 kb	100-150 kb	300-350 kb	300-350 kb
Macro	400-450 kb	450-500 kb	200-250 kb	500-550 kb	400-450 kb
Intermédiaire	250-300 kb	250-300 kb	150-200 kb	250-300 kb	300-350 kb
Micro	100-150 kb	150-200 kb	80-90 kb	150-200 kb	150-200 kb
Z	850-900 kb	1800-1900 kb	950-1000 kb	900-950 kb	1200-1300 kb

B. Diversité génétique dans les populations

Après un contrôle qualité des génotypages de chaque lignée, il est constaté pour les trois lignées Rhode Island un nombre de marqueurs polymorphes plus élevé que pour les deux lignées Leghorn. En effet, pour les lignées RI1, RI2 et RI3, respectivement 258 772, 274 880 et 234 813 SNPs sont supprimés à cause d'une MAF < 0.05. Pour les lignées L1 et L2, respectivement 362 168 et 327 452 SNP sont supprimés à cause d'une MAF < 0.05. Ceci montre une distinction entre les souches Rhode Island et Leghorn. Cette distinction est en accord avec la littérature. Groeneveld et al. (2010) ont résumé les résultats de nombreux travaux ayant montré que les Leghorn ont moins de marqueurs polymorphes que les Rhode Island. Qanbari et al. (2012) ont montré que pour des animaux génotypés sur une puce Lohmann de 36 455 SNP, les Rhode Island/White Rick présentaient 11 800 marqueurs monomorphes quand les Leghorn présentaient 22 815 marqueurs monomorphes. Les lignées RI sont donc moins fixées que les lignées L.

Les taux d'hétérozygotie moyens observés et attendus, ainsi que le coefficient de consanguinité F_{IS} sont présentés dans le tableau 10. Les cinq différentes lignées présentent des taux d'hétérozygotie moyens observés similaires mais avec des différences significatives entre lignées. En effet, les lignées RI ont un taux d'hétérozygotie moyen observé similaire entre elles mais significativement différent du taux moyen observé pour les lignées L. Les deux lignées L1 et L2 présentent également entre elles des taux d'hétérozygoties moyens observés significativement différents, et respectivement plus élevée et plus faible que celles des RI. Les valeurs observées pour les souches Leghorn et Rhode Island sont en accord avec celles de Qanbari et al. (2012).

Toutefois en calculant le coefficient de consanguinité F_{IS} pour chaque lignée, il est constaté que le coefficient est négatif pour l'ensemble des lignées et est significativement différent de zéro pour les trois lignées RI et la lignée L2. Il y a donc pour ces lignées un excès d'hétérozygotes par rapport à la valeur attendue et les lignées ne sont pas consanguines. Les 5 lignées

correspondent à des lignées pures commerciales et présentent des valeurs similaires aux lignées commerciales étudiées par Qanbari et al. (2012). Enfin, hormis la lignée RI2 qui présente un F_{IS} plus faible que les autres lignées RI, aucune différence significative n'est notée entre les différentes lignées.

Tableau 10. Estimation des taux d'hétérozygotie moyens observés et attendus et du coefficient de consanguinité moyens pour les 5 différentes lignées

Lignée	H_O ± SD	H_E ± SD	F_{IS} ± SD
RI1	0.365 ± 0.017[a]	0.364 ± 0.000	-0.002 ± 0.047[a*]
RI2	0.366 ± 0.018[a]	0.362 ± 0.000	-0.010 ± 0.049[b*]
RI3	0.364 ± 0.014[a]	0.362 ± 0.000	-0.004 ± 0.037[a*]
L1	0.368 ± 0.019[b]	0.367 ± 0.000	-0.003 ± 0.052[a]
L2	0.352 ± 0.030[c]	0.350 ± 0.000	-0.007 ± 0.087[ab*]

H_0 : taux d'hétérozygotie moyen observé ; H_E : taux d'hétérozygotie moyen attendu ; F_{IS} : coefficient de consanguinité ; les lettres a, b et c indiquent des différences significatives entre lignées d'après des tests de Student au seuil de première espèce α=1% ; le symbole « * » indique une différence significative par rapport à une valeur nulle d'après des tests de Student au seuil de première espèce α=1%.

C. Diversité génétique entre les populations

Une Analyse en Composantes Principales a été menée sur l'ensemble des 5 lignées avec les génotypages « Full » (Figure 32). Les deux principaux axes représentent 49.33% et 18.23% de la variance totale. Le premier axe sépare clairement les souches Rhode Island (à droite) des souches Leghorn (à gauche). Le deuxième axe sépare les lignées RI1, RI2, L2 avec des valeurs positives, des lignées L1 et RI3 avec des valeurs négatives. Les lignées RI1, RI2 et L2 correspondent à des lignées mâles alors que les lignées RI3 et L1 correspondent à des lignées femelles. Ceci pourrait expliquer la séparation observée selon le second axe. Enfin, il est à noter la présence de trois groupes (RI1, RI2), (RI3) et (L1, L2) selon les deux principaux axes. Toutefois, il est à noter que chaque lignée reste clairement distincte des autres.

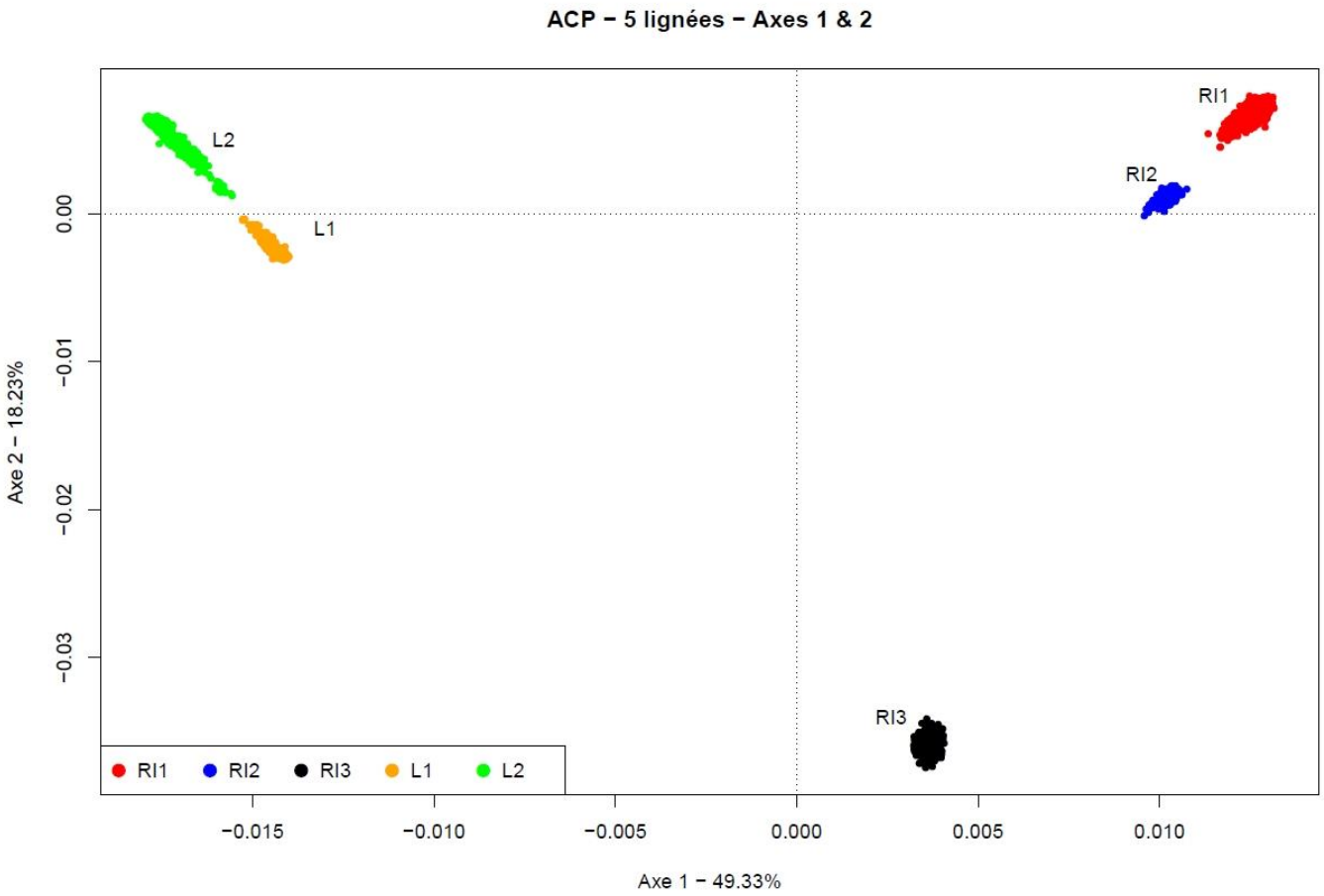


Figure 32. Analyse en Composantes Principales pour l'ensemble des individus des 5 lignées à partir des génotypages "Full", selon les deux principaux axes.

D. Analyse des relations et de la différenciation entre populations

Les indices de différenciation F_{ST} ont été calculés par paires de lignées à partir des génotypages « Full ». Les indices sont compris entre 0.154 pour la paire RI1-RI2 et 0.434 pour la paire RI2-L2. D'après Wright (1978), les valeurs de F_{ST} peuvent être divisées en 4 groupes indiquant différents niveaux de différenciation :

- $0 < F_{ST} < 0.05$: différenciation faible,
- $0.05 < F_{ST} < 0.15$: différenciation modérée,
- $0.15 < F_{ST} < 0.25$: différenciation importante,
- $F_{ST} > 0.25$: différenciation très importante.

Hormis les lignées RI1 et RI2 qui présentent entre elles un indice F_{ST} de 0.154, les autres lignées sont toutes différentes les unes des autres (tableau 11). Le niveau de différenciation plus faible observé pour les lignées RI1 et RI2 est en accord avec les résultats de l'ACP selon les deux principaux axes positionnant les individus RI1 proches des individus RI2. De même, les individus L1 et L2 ont un indice F_{ST} de 0.271 important mais toutefois plus faible que celui

observé pour les autres paires de lignées. D'après les résultats de l'ACP, les individus L1 sont également proches des individus L2. Enfin, quelle que soit les autres paires de lignées, l'indice de différenciation F_{ST} est supérieur à 0.295 et cohérent avec les résultats de l'ACP montrant des lignées très éloignées les unes des autres.

Tableau 11. Indice de différenciation F_{ST} pour les 5 différentes lignées

Lignée	RI1	RI2	RI3	L1
RI1				
RI2	0.154			
RI3	0.295	0.305		
L1	0.388	0.431	0.393	
L2	0.379	0.434	0.398	0.271

Enfin, une analyse de structuration a été réalisée pour les 5 lignées à partir des génotypages « Full » afin de d'estimer le nombre de populations ancestrales K et ainsi déterminer une potentielle origine commune entre lignée. Les analyses ont été menées pour un nombre de populations ancestrales allant de $K = 2$ à $K = 9$ (Figure 33). Le nombre optimal de clusters est déterminé par l'erreur de validation croisée la plus faible. La figure 34 montre que $K = 5$ serait le nombre optimal de clusters permettant de déterminer le nombre de populations ancestrales et une éventuelle origine commune entre lignée. L'erreur de validation croisée est de 0.29 et n'est significativement pas différente pour un nombre de cluster $K \geq 5$. En étudiant les résultats du clustering, il est noté pour $K = 5$ que chaque lignée est encore une fois bien distincte des autres lignées. Ce résultat est donc cohérent avec l'ensemble des résultats précédents. En étudiant les $K = 6$ et $K = 7$, les 5 mêmes clusters sont retrouvés permettant d'identifier les 5 lignées étudiées. Toutefois, la lignée L2 pourrait avoir plusieurs origines qui ne seraient que faiblement partagées avec la lignée L1 (couleur rouge du clustering). Pour $K = 8$, les mêmes remarques sont faites pour la lignée L2 (couleur cyan). Les individus de la lignée RI1 semblent également présenter un génome issu de trois populations ancestrales. En effet, pour cette lignée, le clustering fait apparaitre une couleur orange et deux couleurs vertes dans des proportions sensiblement différentes en fonction des individus. Cela s'explique par le fait que chaque génération est constituée de 3 lots éclos tous les 6 mois sans mélange d'individus entre lots. Les mêmes observations sont faites pour le $K = 9$ que pour les $K = 6, 7$ et 8.

Enfin, de façon intéressante, il est noté un minimum local avec une erreur de validation croisée plus faible pour $K = 2$ (0.43) que pour $K = 3$ (0.59). Le $K = 2$ agrège les lignées L1 et L2 et les lignées RI1, RI2 et RI3. Il est également observé une origine commune pour les lignées RI2 et RI3 avec les lignées L1 et L2, la lignée RI3 présentant 31% d'origine L contre seulement

7.5% pour la lignée RI2. Ce clustering sépare les lignées Leghorn des lignées Rhode Island mais fait donc apparaitre une origine commune pour les lignées RI2 et RI3 avec les lignées Leghorn. Ces résultats sont cohérents avec les résultats de l'ACP selon le premier axe positionnant les individus de la lignée RI3 entre les lignées L1 et L2 et les lignées RI1 et RI2.

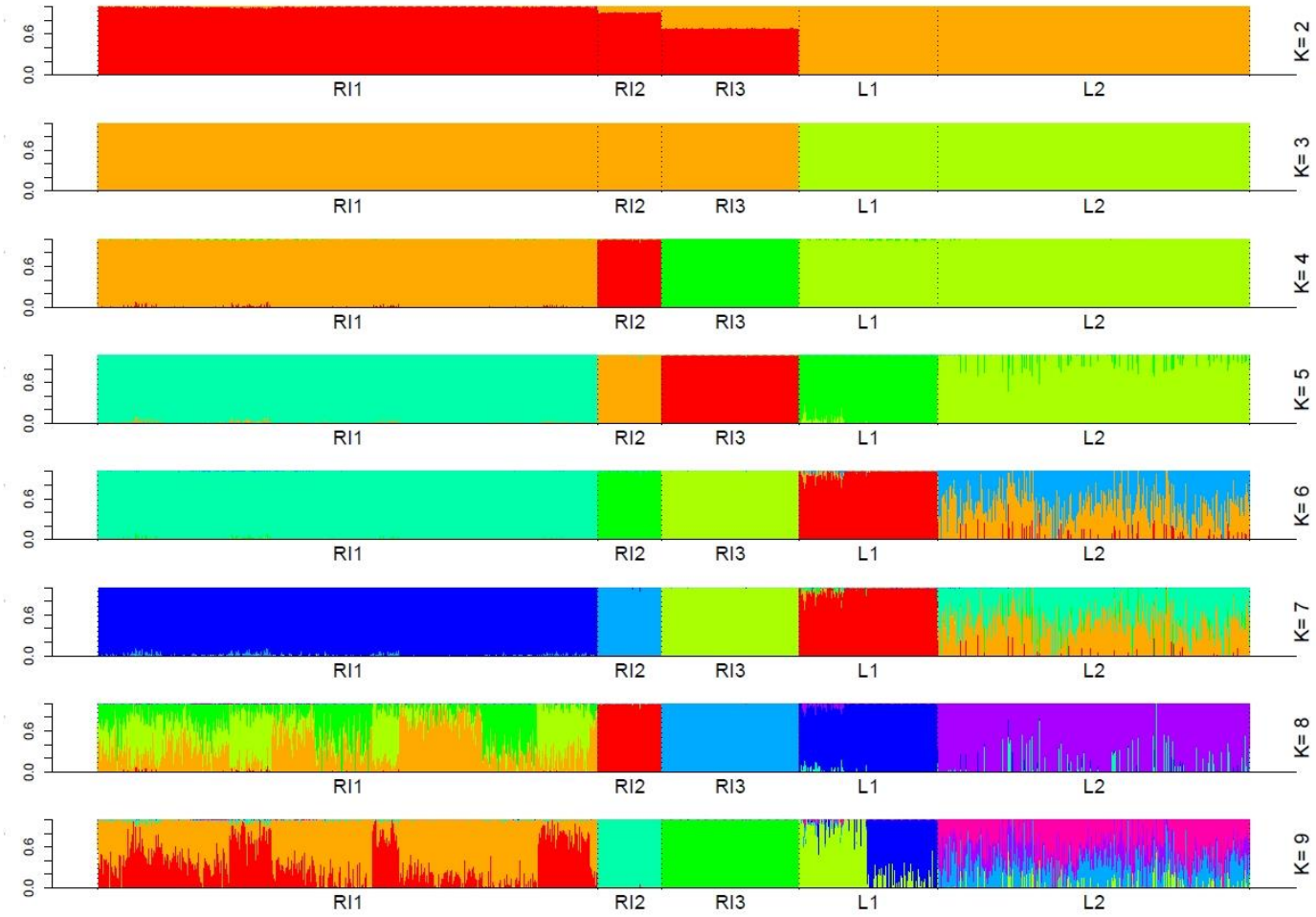


Figure 33. Analyse de structuration pour l'ensemble des individus des 5 lignées étudiées à partir des génotypages "Full".

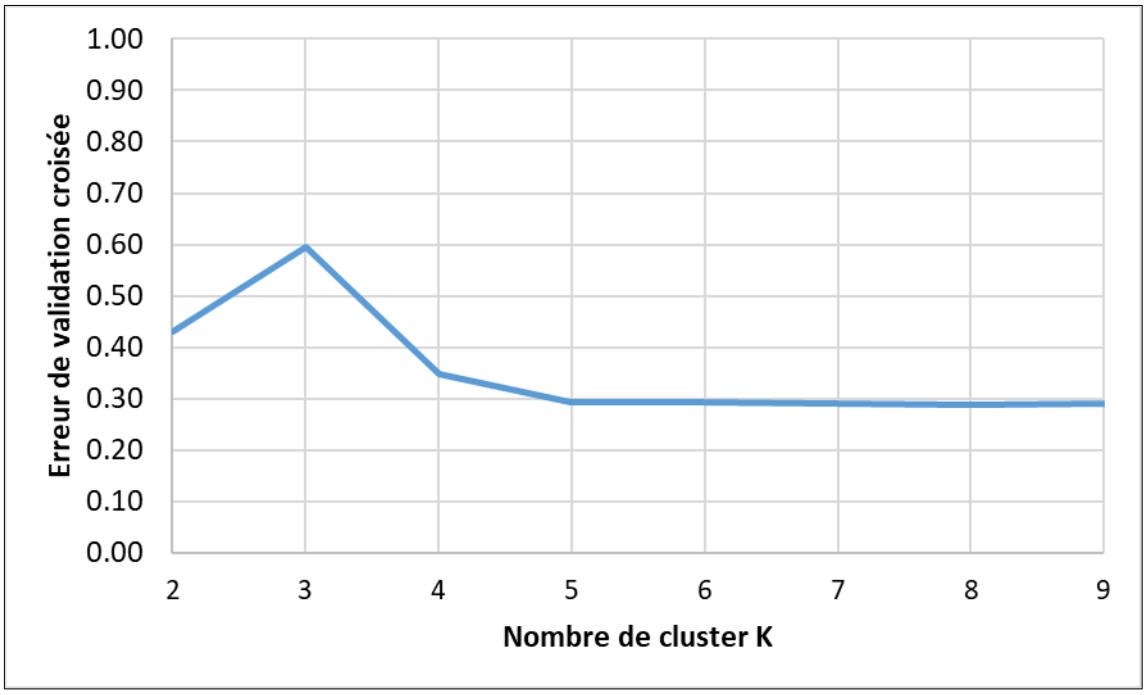


Figure 34. Évolution de l'erreur de validation croisée en fonction du nombre de cluster.

E. Design des puces basse densité

À l'issue des analyses de déséquilibre de liaison et de diversité génétique et à la vue des fortes différences entre les 5 lignées, il apparaît difficile de vouloir agréger une lignée avec une autre. Il est donc nécessaire de considérer distinctement les 5 lignées.

Comme décrit dans la partie matériels et méthodes, la première stratégie de sélection « Indep » consiste alors à considérer les 5 lignées indépendamment les unes des autres et à sélectionner un nombre équivalent de SNP propres à chaque lignée à partir des génotypages obtenus après contrôle qualité. Il faut donc sélectionner 2K, 4K, 6K, 8K et 10K SNP par lignée afin d'étudier les puces 10K, 20K, 30K, 40K et 50K respectivement. La deuxième stratégie « Multi » consiste à considérer les génotypages des 5 lignées en même temps et à sélectionner le nombre total de SNP à partir des génotypages « Full » (Figure 30). La stratégie « Indep » en considérant le $K = 2$ de l'analyse de structuration avec d'une part les génotypages des individus Rhode Island, et d'autre part les génotypages des individus Leghorn a également été testée. Toutefois, les résultats issus de ces puces ne seront pas développés car elles permettent d'obtenir uniquement des résultats intermédiaires aux résultats des stratégies « Indep » avec $K = 5$ et « Multi » présentés par la suite.

Une fois les puces développées, le nombre de SNP informatifs pour à chaque lignée est déterminé en étudiant le recouvrement des SNP de la puce BD avec les SNP retenus après contrôle qualité pour chaque lignée. Les puces BD sont ensuite testées en simulant in silico les génotypages des individus de la génération G1 des lignées RI1 et L2. Les imputations des

individus des générations G1 sont ensuite réalisées à partir des génotypages HD des individus de la génération G0. Les imputations sont réalisées pour les deux lignées indépendamment les unes des autres. Les génotypages HD correspondent aux génotypages obtenus après contrôle qualité, propres à chaque lignée et non aux génotypages « Full ».

Cinq densités différentes ont ainsi été testées pour les deux stratégies : 10K, 20K, 30K, 40K et 50K SNP sur les lignées RI1 et L2 (tableau 12). Le nombre de SNP avec une MAF supérieure à 0.05 pour chaque lignée a également été calculé. Il correspond au nombre de SNP informatifs pour la lignée considérée. Ainsi, pour la puce 10Kindep, sur les 10 011 SNP de la puce, 6768 et 5719 SNP ont une MAF supérieure à 0.05 pour les lignée RI1 et L2 respectivement. Pour la puce 10Kmulti, sur les 9999 SNP de la puce, 8884 et 7912 SNP ont une MAF supérieure à 0.05 pour les lignées RI1 et L2 respectivement. D'une manière générale, la stratégie « Multi » permet d'obtenir plus de SNP informatifs que la stratégie « Indep » pour les deux lignées étudiées.

Tableau 12. Récapitulatif des différentes puces BD simulées.

Stratégie	**Puce**	**Nombre total SNP**	**RI1**	**L2**
			SNP informatifs	SNP informatifs
Indep	10Kindep	10 011	6768	5719
	20Kindep	20 031	13 584	11 468
	30Kindep	30 016	20 352	17 209
	40Kindep	40 047	27 227	22 942
	50Kindep	49 880	33 873	28 566
Multi	10Kmulti	9999	8884	7912
	20Kmulti	19 999	17 440	15 042
	30Kmulti	30 036	25 894	21 817
	40Kmulti	39 961	34 059	28 189
	50Kmulti	49 950	42 159	34 395

F. Qualité des imputations

La qualité des imputations est estimée en calculant la corrélation de Pearson moyenne par SNP entre les vrais génotypages HD et les génotypages HD imputés à partir des différentes puces BD développées et pour les deux lignées RI1 et L2.

Pour les deux lignées et les deux stratégies, une augmentation des corrélations moyennes est observée avec l'augmentation du nombre de SNP informatifs sur les puces BD, avec des différences significatives entre toutes les puces BD (Figure 35). En effet, pour la lignée RI1 avec 6768 SNP (puce 10Kindep) et 8884 SNP (puce 10Kmulti), les corrélations moyennes sont

respectivement de 0.951 et 0.963 et augmentent jusqu'à 0.984 et 0.985 pour 33 873 SNP (puce 50Kindep) et 42 159 SNP (puce 50Kmulti) Les mêmes observations sont faites pour la lignée L2. Cette augmentation est cohérente avec les résultats observés dans les travaux de Herry et al. (2018).

Un deuxième point observé est que pour les deux lignées, d'après le nombre de SNP informatifs sur les puces BD, la stratégie « Multi » présente de meilleurs résultats que la stratégie « Indep » jusqu'à 10K SNP. Au-delà de 10K SNP, il n'y a pas de différence significative entre les deux stratégies. En étudiant le nombre de SNP informatifs sur les puces BD, la stratégie « Multi » peut donc s'avérer intéressante pour développer une puce multi-lignée de moins de 10K SNP. Toutefois, rapportés au nombre de SNP total des puces BD, les résultats sont différents (Figure 36). En effet, pour les deux lignées, quelle que soit la densité testée, la stratégie « Multi » présente de meilleurs résultats que la stratégie « Indep », avec des différences significatives entre toutes les puces BD. Pour la lignée RI1 avec les puces 10Kmulti et 50Kmulti, les corrélations moyennes sont respectivement de 0.963 et 0.985. Avec les puces 10Kindep et 50Kindep, les corrélations moyennes sont respectivement de 0.951 et de 0.984. Les mêmes observations sont faites pour la lignée L2. Ainsi, pour une densité totale donnée, la stratégie « Multi » permet de sélectionner un nombre plus important de SNP informatifs que la stratégie « Indep » pour l'ensemble des lignées. Dans une optique d'optimisation du nombre de SNP sélectionnés et informatifs pour l'ensemble des lignées, la stratégie « Multi » semble donc être la stratégie la plus intéressante pour obtenir de bons résultats d'imputations.

Dans le détail, l'impact des deux stratégies sur l'imputation des différents types de chromosomes a également été étudié pour les deux lignées pour une densité de 10K (Figure 37a) et 50K (Figure 37b). Pour les deux lignées, et pour les deux stratégies, il est noté de meilleurs imputations sur les macro-chromosomes que sur les chromosomes intermédiaires, eux-mêmes mieux imputés que les micro-chromosomes. En effet, pour la lignée RI1 et pour les stratégies « Indep » et « Multi » avec une densité de 10K, les corrélations moyennes sont respectivement de 0.963 et 0.970 pour les macro-chromosomes, 0.949 et 0.962 pour les chromosomes intermédiaires, 0.882 et 0.915 pour les micro-chromosomes. Les mêmes observations sont faites pour la lignée L2, ainsi qu'à une densité de 50K. Ces diminutions des corrélations des macro-chromosomes aux micro-chromosomes sont cohérentes avec les résultats de Herry et al. (2018).

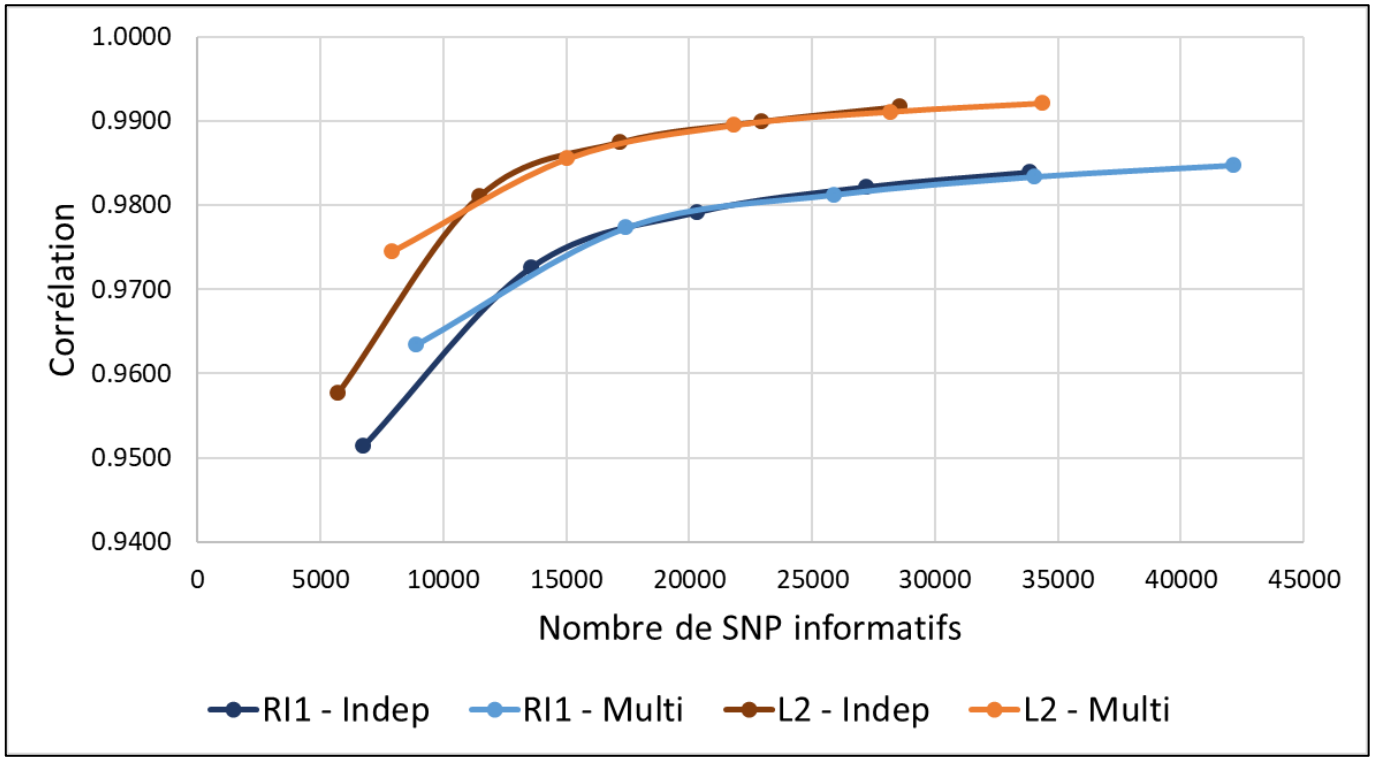


Figure 35. Évolution des corrélations moyennes entre vrais génotypages et génotypages imputés en fonction du nombre de SNP informatifs sur les puces BD pour les deux stratégies et pour les lignées RI1 et L2.

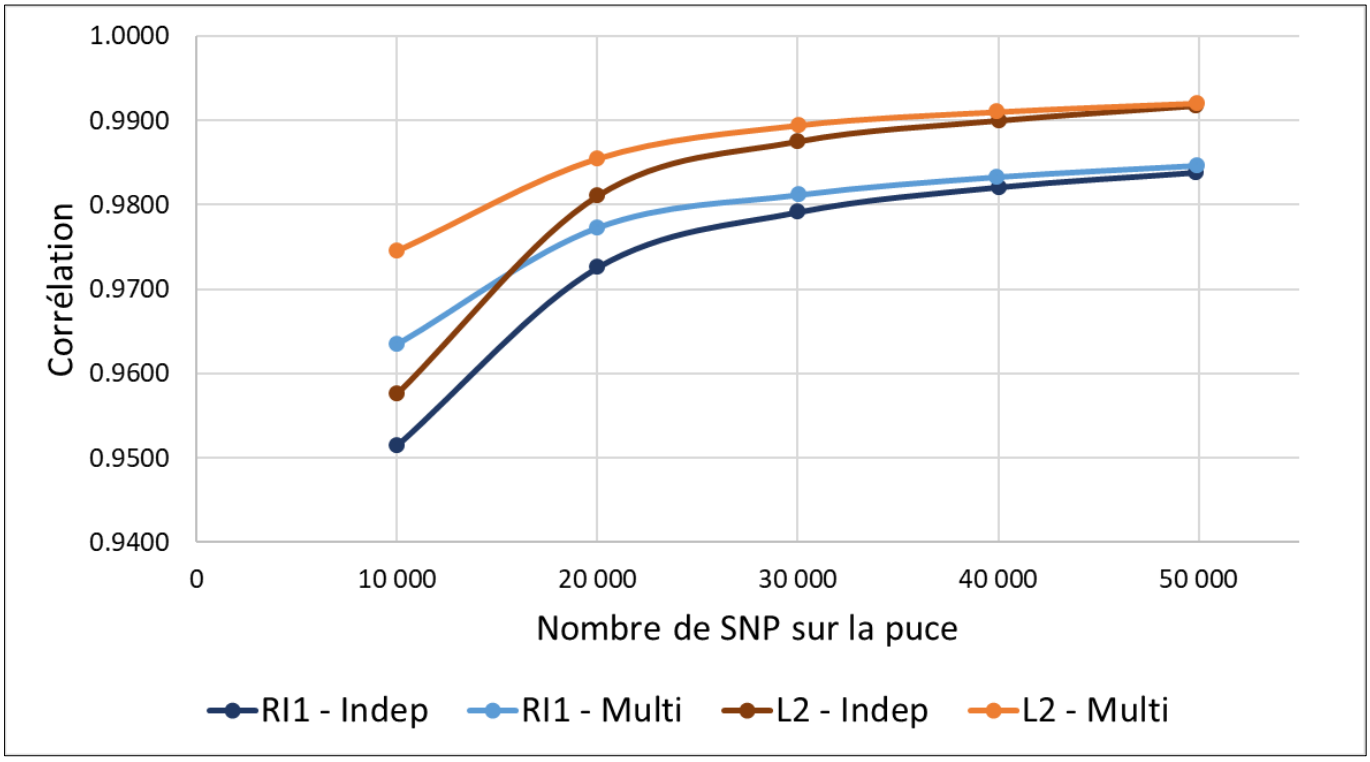


Figure 36. Évolution des corrélations moyennes entre vrais génotypages et génotypages imputés en fonction du nombre de SNP sur les puces BD pour les deux stratégies et pour les lignées RI1 et L2.

Il est également observé de meilleures imputations avec la stratégie « Multi » qu'avec la stratégie « Indep », quel que soit le type de chromosome. En effet, pour la lignée RI1, quelle que soit la densité des puces, des résultats significativement meilleurs sont obtenus pour la stratégie « Multi » que la stratégie « Indep », excepté pour les chromosomes intermédiaires à une densité de 50K où il n'y a pas de différence significative. Les mêmes observations sont faites pour la lignée L2, exception faite des micro-chromosomes et du chromosome Z pour une densité de 50K où la stratégie « Indep » permet d'obtenir de meilleurs résultats.

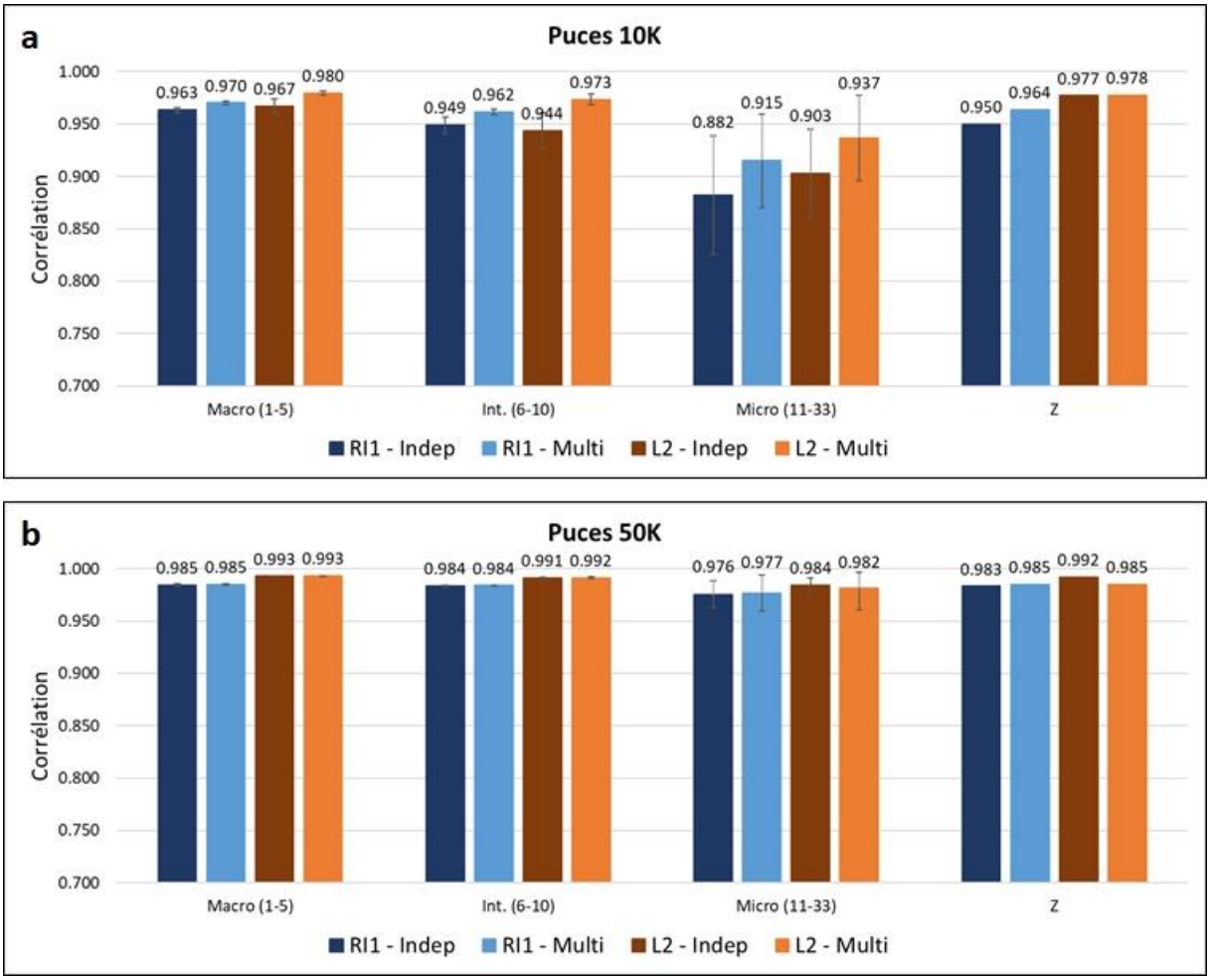


Figure 37. Évolution des corrélations entre vrais génotypages et génotypages imputés en fonction du type de chromosome pour les deux lignées et pour une densité de 10K (a) ou (50K) SNP sélectionnés selon deux stratégies.

G. Impact sur les évaluations génomiques (top150 et 67 mâles repros)

1. Avec imputation

a) Impact des erreurs d'imputations

L'impact des erreurs d'imputation a été étudié en comparant les résultats d'une évaluation génomique sur ascendance des candidats G1 de la lignée RI1 avec leurs vrais génotypages HD

ou avec leurs génotypages BD imputés. Des corrélations de Spearman sont calculées entre les GEBV « Asc_HD » et les GEBV « Asc_Imp » pour les deux stratégies et pour chaque puce BD.

Concernant les 150 meilleurs individus (Top150) en fonction du caractère d'étude, la figure 38 montre, pour chaque caractère et pour les deux stratégies, une augmentation des corrélations de Spearman avec une augmentation du nombre de SNP sur les puces BD. Pour le poids d'œufs (PO) et les puces 10Kindep et 10Kmulti, les corrélations sont respectivement de 0.9764 et 0.9857. Pour les puces 50Kindep et 50Kmulti, les corrélations sont respectivement de 0.9980 et 0.9983. Les mêmes observations sont faites pour les autres caractères. Ceci est en accord avec la littérature. Aliloo et al. (2018) ont montré en bovin que les corrélations entre les GEBV estimés à partir de la puce HD (Illumina BovineHD BeadChip) avec 777K SNP et les GEBV estimés à partir de génotypages HD imputés à partir de 4013 et 25 410 SNP étaient respectivement de 0.9398 et 0.9927. Les résultats sont également en accord avec les précédents travaux de Herry et al. (2019).

Enfin, quels que soient la stratégie et le caractère étudiés, le reclassement des meilleurs individus reste faible. En effet, la corrélation la plus faible est obtenue pour la puce 10Kindep et pour le Lab avec une corrélation de 0.9459. La puce 10Kmulti permet d'obtenir une corrélation de 0.9777. Toutefois, les erreurs standard associées à ces deux puces sont respectivement de ±0.03 et ±0.02. Bien que les résultats soient plus faibles avec la stratégie « Indep », il n'y a donc qu'une tendance non significative à obtenir de meilleurs résultats avec la stratégie « Multi ».

Concernant les 67 reproducteurs G1, en fonction du caractère d'étude, la figure 38 montre, pour chaque caractère et pour les deux stratégies, une augmentation des corrélations de Spearman avec une augmentation du nombre de SNP sur les puces BD. En effet, pour le PO et les puces 10Kindep et 10Kmulti, les corrélations sont respectivement de 0.9955 et 0.9967. Pour les puces 50Kindep et 50Kmulti, les corrélations sont respectivement de 0.9988 et 0.9991. Les résultats sont également supérieurs à ceux obtenus pour les 150 meilleurs individus en fonction du caractère d'étude. Cela s'explique par une meilleure distribution des 67 individus reproducteurs dans le classement des 580 individus G1. Les résultats sont en accord avec les précédents travaux de Herry et al. (2019). Les résultats sont également très bons quels que soient le caractère et la stratégie étudiés. En effet, hormis pour le Lab et la puce 10Kindep qui permet d'obtenir une corrélation de 0.9857, les corrélations sont toutes supérieures à 0.99 indiquant un reclassement très faible des reproducteurs. Enfin, les erreurs standard associées aux puces sont de ±0.01. Il n'y a donc pas de différence significative entre les résultats des différentes puces, que ce soit entre puces d'une même stratégie ou entre puces de différentes stratégies.

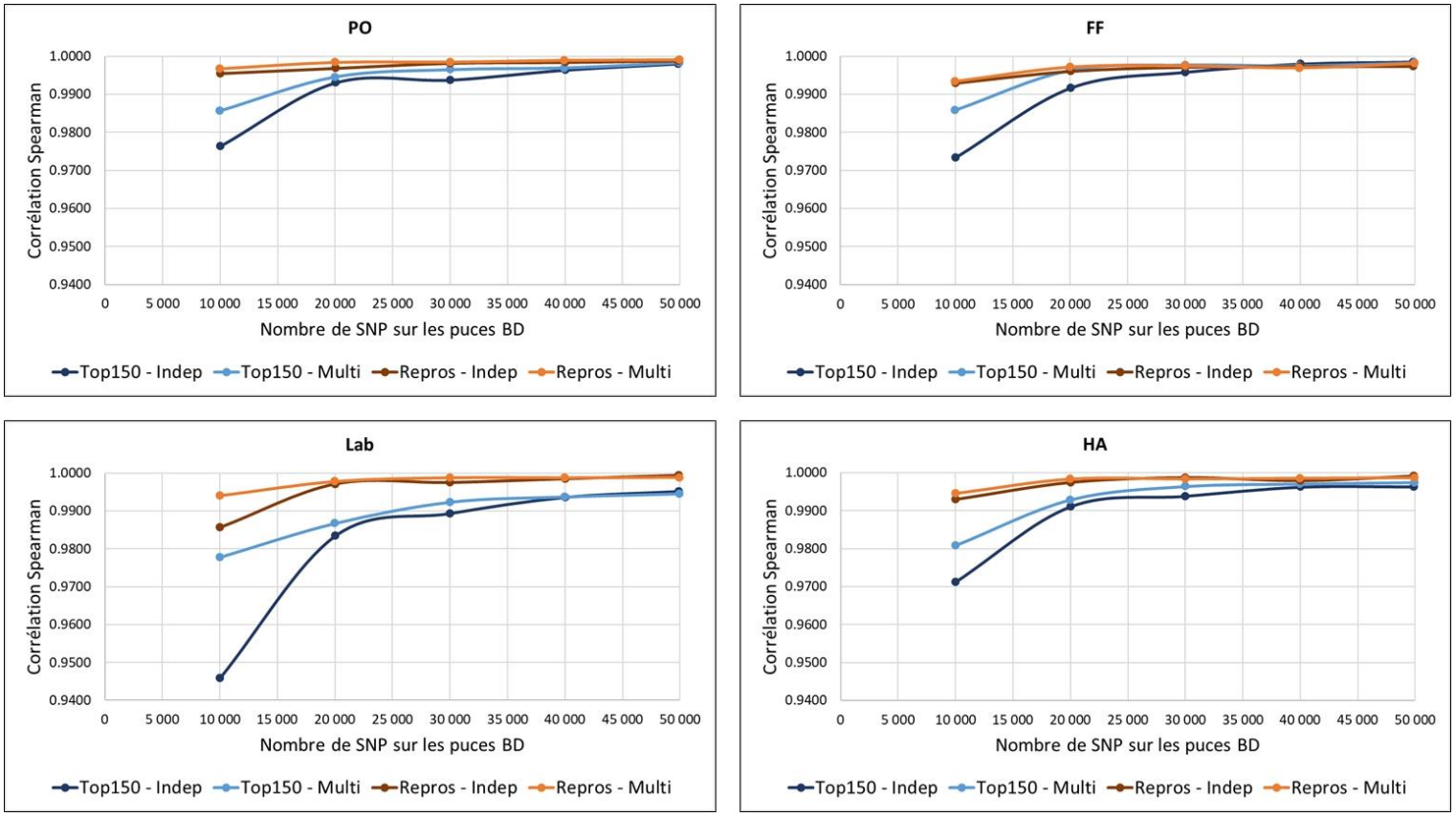


Figure 38. Évolution des corrélations de Spearman en fonction du nombre de SNP sur les puces BD pour les deux stratégies et pour les 150 meilleurs individus selon le caractère d'étude et les 67 reproducteurs G1. Les résultats sont présentés pour le poids d'œuf (PO), la couleur de la coquille des œufs (Lab), la force de fracture (FF) et la hauteur d'albumen (HA) pour des évaluations génomiques sur ascendance avec imputation des génotypages BD.

b) Impact sur la précision des évaluations génomiques

L'impact des différentes stratégies sur la précision des évaluations a été étudié en comparant les résultats de l'évaluation génomique sur descendance « Full_HD » des candidats G1 de la lignée RI1 avec leurs vrais génotypages HD, avec une évaluation génomique sur ascendance des mêmes individus avec leur génotypages BD imputés. Des corrélations de Spearman sont calculées pour les 67 mâles reproducteurs G1 entre les GEBV « Full_HD » et les GEBV « Asc_Imp » pour les deux stratégies et pour chaque puce BD.

Pour les deux stratégies, les résultats sont similaires quelle que soit la densité de la puce DB. En effet, pour le PO et les puces 10Kindep et 10Kmulti, les corrélations sont respectivement de 0.4845 et 0.4880. Pour les puces 50Kindep et 50Kmulti, les corrélations sont respectivement de 0.4849 et 0.4931. Toutefois, la corrélation entre les GEBV « Full_HD » et les GEBV « Asc_HD » est de 0.4848. Cette corrélation représente la valeur théorique maximale atteignable en réalisant une évaluation sur ascendance avec la puce HD (au lieu des puces BD). Les corrélations associées aux différentes puces ne sont significativement pas différentes de celle

obtenue avec la puce HD. En effet l'erreur standard pour chaque puce est de ±0.11. Ceci est en accord avec les résultats précédents montrant un impact faible des erreurs d'imputation sur le reclassement des reproducteurs. En effet, quelles que soient la densité et la stratégie utilisées, les corrélations entre les GEBV HD estimées sur ascendance et GEBV BD imputées estimées sur ascendance sont supérieures à 0.99. Les mêmes observations sont faites pour les autres caractères. Ces résultats sont cohérents avec la littérature et les précédents travaux de Herry et al. (2019). Chen et al. (2014) ont montré chez les bovins que la précision des évaluations génomiques calculée par des corrélations de Pearson entre les DGV (Direct Genomic Value) à partir des vrais génotypages 50K et les performances des taureaux était de 0.61 pour le rendement laitier, et de 0.62 pour le taux de cellules. Avec des génotypages 50K imputés à partir de 6K SNP, les corrélations étaient également de 0.61 pour le rendement laitier, et de 0.62 pour le taux de cellules.

2. Sans imputation

a) Impact de l'absence d'imputation

Suite aux bons résultats obtenus avec imputations pour les deux stratégies et les différentes puces, les évaluations génomiques sur ascendance ont été réalisées directement avec les génotypages BD de l'ensemble des individus des générations G0 et G1 (référence et candidats). L'impact de l'absence d'imputation a été étudié en comparant les résultats d'une évaluation génomique sur ascendance des candidats G1 de la lignée RI1 avec leurs vrais génotypages HD ou avec leurs génotypages BD sans imputation. Des corrélations de Spearman sont calculées entre les GEBV « Asc_HD » et les GEBV « Asc_BD » pour les deux stratégies et pour chaque puce BD.

Concernant les 150 meilleurs individus (Top150) en fonction du caractère d'étude, la figure 39 montre, pour chaque caractère et pour les deux stratégies, une augmentation des corrélations de Spearman avec une augmentation du nombre de SNP sur les puces BD. Pour le PO et les puces 10Kindep et 10Kmulti, les corrélations sont respectivement de 0.9030 et 0.9196. Pour les puces 50Kindep et 50Kmulti, les corrélations sont respectivement de 0.9516 et 0.9451. Lorsque ces résultats sont comparés à ceux obtenus avec imputation, il est noté une diminution des résultats pour l'ensemble des puces étudiées. Ainsi, pour le PO et les puces 10Kindep et 10Kmulti, les corrélations baissent respectivement de 0.0734 et de 0.0661. Pour les puces 50Kindep et 50Kmulti, les corrélations baissent respectivement de 0.0464 et de 0.0532. Les différences observées sont significatives. Les mêmes observations sont faites pour les autres caractères. Enfin, excepté le Lab et la puce 10Kindep pour lesquels la corrélation est de 0.8985, les

corrélations sont toutes supérieures à 0.90. L'utilisation directe des puces BD entraîne donc un reclassement des individus plus important que celui observé avec imputation, tout en restant cependant modéré. Les erreurs standard associées aux puces sont comprises entre ±0.03 et ±0.04. Il n'y a donc pas de différence significative entre les résultats des différentes puces, que ce soit entre puces d'une même stratégie ou entre puces de différentes stratégies.

Concernant les 67 reproducteurs G1, en fonction du caractère d'étude, la figure 39 montre, pour chaque caractère et pour les deux stratégies, une augmentation des corrélations de Spearman avec une augmentation du nombre de SNP sur les puces BD. Pour le PO et les puces 10Kindep et 10Kmulti, les corrélations sont respectivement de 0.9775 et 0.9806. Pour les puces 50Kindep et 50Kmulti, les corrélations sont respectivement de 0.9861 et 0.9879. Les résultats sont encore supérieurs à ceux obtenus pour les 150 meilleurs individus en fonction du caractère d'étude. Cela est dû à une meilleure distribution des 67 individus reproducteurs dans le classement des 580 individus G1. De même, en comparant avec les résultats obtenus avec imputation, les résultats pour les deux stratégies sont également plus faibles que ceux obtenus avec imputation pour les mêmes densités, mais les différences ne sont en revanche pas significatives. Toutefois, quels que soient le caractère et la puce étudiés, les corrélations sont toutes supérieures à 0.96. Le reclassement des reproducteurs est donc plus important qu'avec utilisation de l'imputation mais il reste toutefois très modéré. Enfin, les erreurs standards associées aux puces sont comprises entre ±0.02 et ±0.03. Il n'y a donc là encore pas de différence significative entre les résultats des différentes puces.

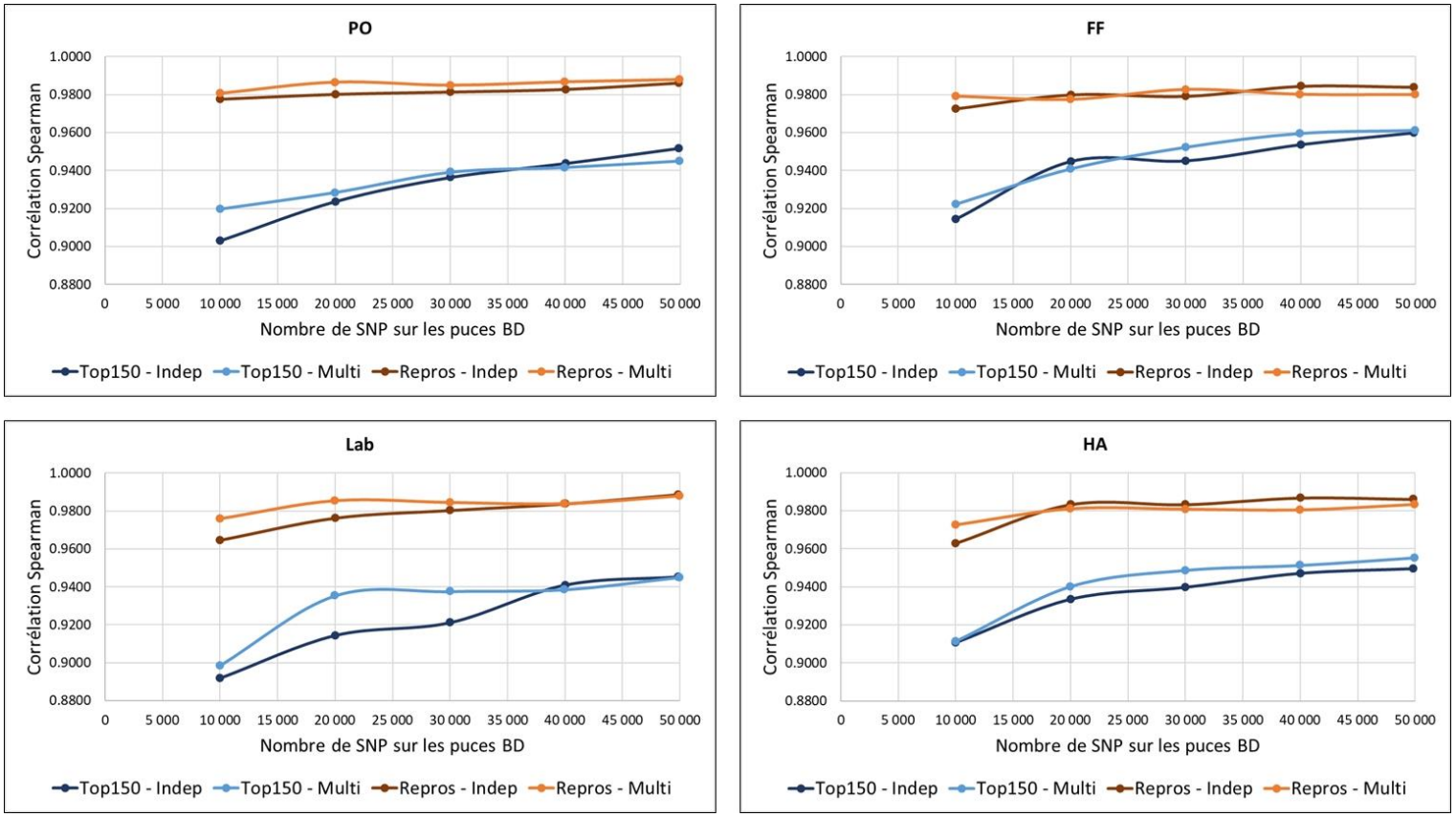


Figure 39. Évolution des corrélations de Spearman en fonction du nombre de SNP sur les puces BD pour les deux stratégies et pour les 150 meilleurs individus selon le caractère d'étude et les 67 reproducteurs G1. Les résultats sont présentés pour le poids d'œuf (PO), la couleur de la coquille des œufs (Lab), la force de fracture (FF) et la hauteur d'albumen (HA) pour des évaluations génomiques sur ascendance sans imputation des génotypages BD.

b) Impact sur la précision des évaluations génomiques

L'impact de l'utilisation directe des puces BD sans imputation sur la précision des évaluations a été étudié en comparant les résultats de l'évaluation génomique sur descendance « Full_HD » des candidats G1 de la lignée RI1 avec une évaluation génomique sur ascendance des mêmes individus avec les génotypages BD pour les individus G0 et G1. Des corrélations de Spearman sont calculées pour les 67 mâles reproducteurs G1 entre les GEBV « Full_HD » et les GEBV « Asc_Imp » pour les deux stratégies et pour chaque puce BD.

Pour les deux stratégies, les résultats sont similaires quelle que soit la densité de la puce DB. En effet, pour le PO et les puces 10Kindep et 10Kmulti, les corrélations sont respectivement de 0.4803 et 0.4923. Pour les puces 50Kindep et 50Kmulti, les corrélations sont respectivement de 0.4707 et 0.4830. L'erreur standard pour chaque puce est de ±0.11. La corrélation entre les GEBV « Full_HD » et les GEBV « Asc_HD » est de 0.4848. Il n'y a donc pas de différence significative entre les résultats des différentes puces ainsi qu'avec ceux de la

puce HD. Ceci est cohérent avec les résultats précédents montrant un impact léger de l'absence d'imputation sur le reclassement des reproducteurs.

Il est toutefois noté pour le Lab une diminution des résultats pour une densité de 10K. Avec la puce 10Kindep et 10K multi, les corrélations sont respectivement de 0.2167 et 0.2377. La corrélation entre les GEBV « Full_HD » et les GEBV « Asc_HD » est de 0.2618. Toutefois, les erreurs standards associées aux puces est de ±0.12. Ces résultats sont en accord avec ceux de Herry et al. (2019). Ainsi, une utilisation directe des puces BD sans imputation pourrait permettre d'obtenir des résultats aussi bons que ceux obtenus avec imputation ou qu'avec la puce HD elle-même. Toutefois, la diminution observée des corrélations pour le Lab et pour une densité de 10K montre que l'utilisation directe de puces à faible densité sans imputation pourrait avoir des impacts sur la précision des évaluations de certains caractères.

IV. Conclusion

Dans le but d'optimiser une puce BD multi-lignée en poule pondeuse, une analyse du déséquilibre de liaison et de la diversité génétique des différentes lignées Novogen concernées par la puce peut être intéressante. Ceci peut permettre, in fine, d'éviter pour les lignées étudiées des biais de vérification trop importants et pouvant avoir des conséquences sur les différentes études liées aux fréquences alléliques. L'analyse du DL a permis d'observer une chute du DL avec une augmentation de la distance entre marqueurs. En effet, pour des distances inférieures à 20kb, et en fonction des lignées, le DL moyen est compris entre 0.48 et 0.67. Pour des distances supérieures à 5Mb, le DL moyen est compris entre 0.01 et 0.02. L'analyse a également permis de noter des différence d'étendue du DL entre type de chromosome, avec une étendue qui diminue des macro-chromosomes aux chromosomes intermédiaires et aux micro-chromosomes. En étudiant le DL utile (r^2>0.3) sur le génome et pour les différents types de chromosomes, les résultats sont également différents en fonction de chaque lignée. Les études de diversité génétique dans les populations ont montré que les 5 lignées n'étaient pas consanguines et qu'il n'y avait pas de différence significative entre lignées. Les études de diversité entre populations avec l'ACP ont permis de montrer que les lignées RI1 et RI2 étaient éloignées de la lignée RI3, ces trois lignées étant elles-mêmes distantes des lignées L1 et L2. Enfin, l'analyse des relations et de la différenciation entre population a permis de faire émerger deux cas possibles pour la construction des puces BD. Le premier cas consiste à considérer deux groupes d'individus en séparant les souches Leghorn et Rhode Island ($K = 2$ de l'analyse de structuration). Le deuxième cas, en accord avec l'ensemble des résultats d'analyse du DL et

de diversité génétique, consiste à considérer les 5 lignées ($K = 5$ de l'analyse de structuration). Le design des puces BD selon ce deuxième cas s'est ensuite fait selon deux stratégies « Indep » ou « Multi ».

En étudiant la qualité d'imputation à partir des puces BD développées, il est constaté que, pour une densité totale donnée, la stratégie « Multi » permet de sélectionner un nombre plus important de SNP informatifs que la stratégie « Indep » pour l'ensemble des lignées. Dans une optique d'optimisation du nombre de SNP sélectionnés et informatifs pour l'ensemble des lignées, la stratégie « Multi » semble donc être la stratégie la plus intéressante pour obtenir de bons résultats d'imputations.

Enfin, l'objectif des sélectionneurs est d'obtenir de bons résultats d'évaluations génomiques des candidats à la sélection. En étudiant l'impact des puces sur les évaluations, des différences non significatives sont observées entre les deux stratégies, avec toutefois des résultats qui tendent à être meilleurs avec la stratégie « Multi ». Il est également noté un impact faible des deux stratégies sur le reclassement des meilleurs individus et des reproducteurs ainsi que sur la précision des évaluations génomiques. Sans imputation, l'utilisation directe des puces BD a montré que des résultats moins corrélés étaient obtenus comparés à ceux obtenus avec imputation. Là encore, des différences non significatives sont observées entre les deux stratégies, la stratégie « Multi » tendant à de meilleurs résultats que la stratégie « Indep ». Il y a encore un impact faible sur le reclassement des meilleurs individus et des reproducteurs, ainsi que sur la précision des évaluations génomiques. Combiné à l'imputation, il est donc encore possible de diminuer la densité de la puce BD du fait d'un certain nombre de SNP informatifs pour plusieurs lignées. Il convient néanmoins de noter que l'utilisation directe de puces à faible densité (10K) sans imputation pourrait avoir des impacts sur la précision des évaluations de certains caractères.

Chapitre V. Utilisation des techniques RAD-Seq comme alternatives aux puces basse densité

I. Introduction

Le coût des puces HD étant trop élevé pour génotyper l'ensemble des candidats à la sélection, l'utilisation des puces BD s'avère une solution intéressante pour la sélection génomique chez de nombreuses espèces d'élevages. Toutefois, pour des raisons de coûts et de volume de puces commandées par les sélectionneurs, les puces BD sont généralement développées pour un ensemble de races ou de lignées.

En parallèle aux puces basse densité, les techniques de séquençage nouvelle génération (NGS) sont de plus en plus utilisées pour les espèces d'élevages. Ces techniques permettent à la fois de détecter et de génotyper des SNP mais restent trop onéreuses pour pouvoir être utilisées en routine pour la sélection génomique. Il est en revanche possible d'utiliser des techniques de RAD-Seq par l'utilisation d'enzymes de restriction coupant l'ADN au niveau de sites de restriction puis par le séquençage d'une partie des différents fragments obtenus. Toutefois, la qualité des génotypes obtenus avec ces méthodes est plus variable que celle obtenue avec des puces à SNP à cause d'une profondeur de séquençage plus faible. La sensibilité de l'enzyme de restriction aux méthylations et le taux de polymorphismes dans les sites de restriction sont également d'autres facteurs qui vont provoquer une certaine variabilité entre les génotypes des différents individus. De nombreuses études ont cependant montré qu'il était possible de réduire la variabilité entre individus grâce à l'imputation des données manquantes (Poland et al., 2012b ; Torkamanek et Belzile, 2015 ; Brouard et al., 2017, Elbasyoni et al., 2018).

Enfin, après avoir correctement géré la variabilité entre individus, il est nécessaire d'étudier le recoupement entre les SNP détectés et génotypés avec les techniques RAD-Seq et les SNP issus de la puce HD. En effet, dans une optique de remplacement des puces BD par les technologies RAD-Seq, seuls les SNP en commun avec les SNP de la puce HD vont permettre de déduire les génotypes HD des candidats. Il en résulte que seul un nombre réduit de SNP détectés et génotypés par les méthodes RAD-Seq serviront pour les imputations. Enfin, avec les méthodes RAD-Seq, le prix n'est plus dépendant du volume désiré et des fabricants de puces à SNP. Les coûts sont principalement dus aux kits de préparation des librairies d'ADN et au séquençage.

L'objectif de cette étude est de simuler deux différentes méthodes de RAD-Seq, à savoir le génotypage par réduction du génome et séquençage (GGRS) (Chen et al., 2013 ; Liao et al.,

2015 ; Pértille et al., 2016) et la double digestion RAD-Seq (ddRAD-Seq) (Peterson et al., 2012), puis d'identifier les SNP en commun avec ceux de la puce HD. Des imputations sont ensuite testées à partir des SNP en communs. Enfin, l'impact des imputations sur les évaluations génomiques à partir des données imputées est étudié afin de conclure sur l'intérêt de ces méthodes RAD-Seq pour la sélection génomique en poule pondeuse.

II. Article III : Intérêt des technologies RAD-Seq comme alternatives aux puces basse densité pour la sélection génomique en poule pondeuse

Article soumis dans Genetic Selection Evolution

Herry F, Hérault F, Picard Druet D, Bardou P, Varenne A, Burlot T, Le Roy P, Allais S. 2019. Interest of using restriction site-associated DNA sequencing as an alternative to low density SNP chips for genomic selection in layer chicken.

Genetic Selection Evolution.

INTEREST OF RESTRICTION SITE-ASSOCIATED DNA SEQUENCING TECHNOLOGIES AS AN ALTERNATIVE TO LOW DENSITY SNP CHIPS FOR GENOMIC SELECTION IN LAYER CHICKEN: IN SILICO RESULTS

Florian Herry[1,2,*], Frédéric Hérault[2], David Picard-Druet[2], Philippe Bardou[3], Amandine Varenne[1], Thierry Burlot[1], Pascale Le Roy[2] and Sophie Allais[2]

[1] NOVOGEN, 5 rue des Compagnons, Secteur du Vau Ballier, 22960 Plédran, France,

[2]PEGASE, INRA, Agrocampus Ouest, 16 Le Clos, 35590 Saint-Gilles, France

[3]SIGENAE, GenPhySE, Université de Toulouse, INRA, ENVT, 24 chemin de Borde-Rouge - Auzeville Tolosane, 31326 Castanet Tolosan, France

Abstract

Background: To reduce the cost of genomic selection, low density single nucleotide polymorphism (SNP) chip can be used in combination with imputation for genotyping the selection candidates instead of using high density SNP chip. Concurrently, next-generation sequencing (NGS) techniques have been increasingly used in livestock species. Nevertheless, they remain expensive to be routinely used for genomic selection. An alternative and cost-efficient solution is to use restriction site-associated DNA sequencing (RADseq) techniques to sequence only a fraction of the genome by using restriction enzymes. In this perspective, the interest of RADseq techniques as alternative to low density SNP chip for genomic selection was studied in a pure layer line.

Results: Two RADseq approaches were simulated from sequences of 1027 individuals. A genotyping by genome reducing and sequencing (GGRS) approach was simulated using four enzymes (EcoRI, TaqI, AvaII and PstI). A double-digest RADseq (ddRADseq) method was also simulated using TaqI and PstI. Imputation accuracy was assessed as the mean correlation between true and imputed genotypes. Egg weight, egg shell color, egg shell strength and albumen height were evaluated with single-step GBLUP methodology. The impact of imputation errors on the ranking of the selection candidates was assessed by comparing a genomic evaluation based on ancestry using true HD or imputed HD genotyping. The relative accuracy of genomic estimated breeding values (GEBV) was also investigated by considering as reference the GEBV estimated on offspring. With a GGRS approach with AvaII or PstI and a ddRADseq with TaqI and PstI, more than 10K SNPs were detected in common with the HD SNP chip resulting in imputation accuracy higher than 0.97. The impact of imputation errors on genomic evaluation of the breeders was reduced with Spearman correlation higher than 0.99. Finally, the relative accuracy of GEBV was also equivalent.

Conclusions: GGRS and ddRADseq approaches can be interesting alternatives to low density SNP chip for genomic selection. With more than 10K SNPs in common with the SNPs of the HD SNP chip, good imputation and genomic evaluation results can be obtained. However, with real data, heterogeneity between individuals with missing data has to be taken into account.

Keywords: Genomic selection, layer chicken, low density panel, imputation accuracy, genomic evaluation accuracy, NGS, genotyping-by-sequencing

Background

Genomic selection, as described in 2001 by Meuwissen et al. [1], has been implemented in layer and broiler breeding through the use of the 600K Affymetrix® Axiom® high density (HD) genotyping array, developed by Kranis et al. in 2013 [2]. This HD SNP chip is based on single nucleotide polymorphisms (SNP) corresponding to variations of a single nucleotide base of the DNA, frequent along the DNA. The principle of genomic selection is to estimate the genomic values of the genotyped selection candidates with or without phenotype from a reference population with phenotypes and genotypes. This allows to choose the best breeders for one or more traits to produce the individuals of the next generation.

However, the cost of such HD SNP chip is still a problem for all livestock species. But it is possible to reduce the cost of genomic selection through the use of low density SNP chip by selecting a subset of markers from the HD SNP chip and to impute the genotypes at missing markers. This is a very common method used in many livestock species like cattle [3, 4, 5, 6, 7], pig [8, 9, 10], sheep [11, 12, 13] or poultry [14, 15, 16]. Nevertheless, depending on the number of individuals used to designed the genotyping array, it may result in a skewing of the distribution of allele frequency towards common alleles [17]. This ascertainment bias is due to the SNPs genotyped from genotyping array which may not be all representatives of the genotyped individuals. According to the diversity level or the population structure, this can lead to biased conclusions.

In parallel to this SNP chip methods, next-generation sequencing (NGS) techniques to simultaneously detect and genotype SNPs have been increasingly used in livestock species. Nevertheless, they remain expensive to be routinely used for genomic selection. An alternative and cost-efficient solution is to sequence only a fraction of the genome by using restriction enzymes. This solution was first named as restriction site-associated DNA sequencing (RADseq) or Genotyping-By-Sequencing (GBS) but now refers to a large range of techniques relying on the use of restriction enzymes to detect and genotype SNPs [18]. Subsequently, the term RADseq will be used in this study to refer to the different RADseq approaches. These techniques can be categorized depending on the use of one or two restriction enzymes and the absence or presence of a size selection of the DNA fragments during library preparation [18, 19, 20]:

- One enzyme with size selection: RADseq [21], GBS [22], Reduced Representation Libraries (RRL) [23], Multiplexed Shotgun Sequencing (MSG) [24], 2bRad [25].
- One enzyme without size selection: GBS [26], Genotyping by Genome Reducing and Sequencing (GGRS) [27].

- Two enzymes with size selection: Double-digest RADseq (ddRADseq) [28], GBS with two enzymes [29].
- Two enzymes without size selection: GBS with two enzymes [30], Complexity Reduction of Polymorphic Sequences (CRoPS) [31].

These techniques were first tested on plant [26, 30, 32] with or without reference genome, and then on cattle [33], pig [27], goat [34] and poultry [35, 36, 37]. RADseq methods also enable de novo detection of SNPs for all species, even those without reference genome [18].

As stated by Andrews et al. [18], these techniques have several steps in common to prepare the sequencing libraries. They all start with an enzymatic digestion with one or two enzymes followed by a ligation of adaptors on both sides of the fragments obtained. Depending on the technique, adaptors can contain barcodes which are short sequences of 4 to 8 nucleotides, all different from each other, allowing to identify each sample sequenced. A size selection of DNA fragments can also occur during library preparation. Finally, the sequencing depth mainly depends on the considered multiplexing. But for a given multiplexing capacity, they all allow to sequence at different sequencing depth (less to 5X per site per individual with GBS and MSG methods to more than 20X per site per individual with RADseq) a fraction of the genome [27]. However, the quality of genotypes obtained with RADseq methods is lower than that obtained with genotyping array due to lower sequencing depth. It is possible to increase sequencing depth, and thus the quality of genotypes, but it will increase sequencing costs per individual. For outbred populations like livestock species and with low sequencing depth, these techniques are not easily applicable for SNP identification and genotyping due to high level of heterozygosity and phase ambiguity in the haplotypes. This leads to many missing genotypes at a specific locus for different individuals and introduces variability between individuals. Consequently, one of the major drawbacks of these techniques is the management of missing data and variability between individuals. Many studies have shown the possibility to impute accurately this missing data to reduce the variability between individuals in plants [32, 38, 39, 40] but also in cattle [41]. Finally, after having accurately handled this variability between individuals, SNPs in common with RADseq methods and genotyping array can be identified. In the study of Torkamanek and Belzile [39], only 2,975 SNPs were in common with the 42,508 SNPs of the soy SNP chip. These SNPs obtained with RADseq methods have been used for genomic selection. This was mainly implemented in plant [32, 38, 40] due to the lower level of heterogeneity and phase ambiguity in the haplotypes of the different species studied compared to that of livestock species. Thus, the implementation of RADseq methods for genomic selection in livestock species is far from being used in routine. To our knowledge, there is just

one publication focusing on simulated data illustrating the potential of RADseq methods for genomic selection in livestock species [42].

To date, RADseq [35], GGRS [36, 37], and double digest GBS [43] have been successfully used in poultry to detect and genotype SNPs. However, RADseq is expensive and library preparation is rather complex and labor-intensive. In addition, the random shearing step before size selection of the fragment introduces a variability in the fragments that can be obtained for the different individuals. Concerning the Double digest GBS, the protocol is more simple but there is also a variability in the different sequenced fragments since there is no size selection of the fragment before the PCR step. Conversely, the simplification of the protocol, the design of adapters and barcodes and the removing of several clean-up steps to reduce the variation of fragment number between individuals make GGRS a simple and highly reproducible method [27]. It is also the case for ddRADseq [28]. Thus, focusing on true HD genotyping and on real and simulated sequence data on a pure layer line, the first objective was to simulate GGRS and ddRADseq approaches and to identify the SNPs in common between those obtained from HD SNP chip and from the two different RADseq approaches. Based on these SNPs, the second objective was to impute the genotypes at missing markers to go back to the genotypes obtained with the HD SNP chip. Finally, the third objective was to investigate the impact of imputation errors on genomic evaluation.

Methods

Ethics approval

All blood samples were carried out as part of the commercial and selection activities of Novogen. These animals studied and the scientific investigations described herein are therefore not to be considered as experimental animals per se, as defined in EU directive 2010/63 and subsequent national application texts. Consequently, we did not seek ethical review and approval of this study as regarding the use of experimental animals. All animals were reared in compliance with national regulations pertaining to livestock production and according to procedures approved by the French Veterinary Services.

Animals

All animals studied consisted in a commercial pure line of Rhode Island laying hens and were detailed in Herry et al. [16]. This line was created and selected by Novogen (Plédran, France). The population was comprised of 21,475 chickens distributed in four generations and each

generation was constituted by three batches with the breeding of a new batch every six months from 2010 to 2015. (Figure 1).

Genotyping

2370 animals were genotyped for 580,961 SNPs using the 600K Affymetrix® Axiom® HD genotyping array [2]. Genotyping acquisition was detailed in Herry et al. [16].

In accordance with the fifth annotation release of Gallus gallus genome [44], these SNPs were distributed on macro-chromosomes (1 to 5), intermediate chromosomes (6 to 10), micro-chromosomes (11 to 28 and 33), one linkage group (LGE64), two sexual chromosomes Z and W, as well as a group of 3,724 SNPs with unknown locations.

Genotypes were filtered in six successive steps (Table 1) with Plink V1.9 [45] including individual call rate (<95%), MAF (<0.05), SNP call rate (<95%) and Hardy-Weinberg equilibrium ($P < 10^{-4}$). SNPs with unknown location, or located on linkage group LGE64 or on sexual chromosome W were not included to assure consistency with sequence data used in this study. SNPs located on chromosome 16 and 33 were also removed since some enzymes used in this study did not detect more than 2 SNPs in common with the SNPs of the HD SNP chip, thus preventing the possibility to impute these chromosomes. Pedigree incompatibility problems were also checked. Finally, 300,028 SNPs and 2362 individuals were used in this study.

Sequencing

Among the individuals genotyped with the HD SNP chip, 90 individuals of the first generation (G0) were also sequenced with the Illumina HiSeq2000 technology with a target coverage of 20X at the Genomics and Transcriptomics platform GeT-PlaGe (Toulouse, France). Sequences of these individuals were obtained in two different times. Firstly, 50 individuals were sequenced as part of UtOpIGe project. These individuals were chosen to best represent haplotype diversity of the chicken of the first generation (G0). Among these 50 individuals, 13 individuals were breeders of the second generation (G1), and 37 were collaterals of the breeders. Secondly, 40 individuals were sequenced within the framework of project OptiSeq. These individuals were chosen among the 120 remaining breeders from the first generation (G0) and they best represented haplotype diversity of the chicken of this generation.

Data were aligned to the fifth annotation release of the chicken reference genome with Burrows-Wheeler Aligner V0.7.15 [46] with default parameters for paired-end alignment. SNP calling was done with GATK V3.7 [47] with default parameters. After filtering, 8 213 876 SNP remained and were distributed on chromosome 1 to 28, 33 and sexual chromosome Z. The 90

whole-genome sequenced individuals were used as reference to impute up to the sequence the HD genotyping of the 357 remaining individuals of the first generation (G0) and the 580 individuals of the second generation (G1). FImpute V2.2 [48] was used to impute these 937 individuals.

The correlation between the true whole-genome sequence and the imputed whole-genome sequence of an individual was higher than 0.98, indicating a very good imputation accuracy to the HD genotyping up to the sequence level.

Enzyme selection and simulations of GGRS and ddRADseq

Four distinct restriction enzymes were used to simulate in silico digestion of DNA. According to the two papers of Liao et al. [36] and Pértille et al. [37], two different enzymes, respectively AvaII and PstI, were used to digest their DNA. In addition, EcoRI and TaqI were suggested by the Genomics and Transcriptomics platform GeT-PlaGe. The different sequence patterns and their sensitivity to methylation are described in Table 2.

In addition, a double digestion of DNA was also simulated by using simultaneously TaqI and PstI. In silico digestion of DNA with these different enzymes was realized with R using the Bioconductor packages [50]: Biostrings, BSgenome.Ggallus.UCSC.galGal5, plyr, ggplot2, reshape2 and scales. The R script was done to identify all restriction sites on the chicken reference genome according to the enzyme used. It also counted the number of DNA fragment in accordance to the size of the fragments. Concerning the double digestion of DNA with TaqI and PstI, the R script identified all fragments obtained thanks to the action of the two enzymes. A fragment between two restriction sites of the same enzyme cannot be used in the ddRADseq method. Then, the fragments ranging from 200 to 500bp were selected since previously identified as the appropriate length for sequencing fragments with HiSeq Illumina sequencing system [51]. From the reduced list of fragments and to simulate the paired-end sequencing, windows of 150bp after start position of the restriction site and 150bp before start position of a second restriction site were selected. A bed file was then created containing, for each fragment ranging from 200 to 500bp, two sequences of 150bp obtained with paired-end sequencing. This bed file was used with Plink to extract from the 1027 imputed sequenced individuals all SNPs located on the 150bp windows according to their physical positions.

Finally, among the list of SNPs extracted from the 1027 imputed sequenced individuals, the SNPs in common with the HD genotyping after quality control were identified. The HD genotyping of the 1027 individuals were then reduced to the SNPs previously identified, thus allowing to simulate GGRS and ddRADseq approaches.

Imputation accuracy

In this study, the selection candidates were the 580 individuals of the second generation (G1) with simulated low density genotyping obtained through GGRS and ddRADseq simulations. These candidates were imputed from the HD genotyping of the 447 individuals of the first generation (G0). The selection candidates were directly related to the 447 individuals of the first generation which were the fathers or the fathers' half-brothers of the selection candidates. For each simulated RADseq approaches, imputation accuracy was estimated as the mean correlation between true and imputed genotypes [16]. Correlations were calculated, SNP by SNP, for all candidates according to Pearson's method. Mean correlation was then estimated, respectively on 300,028 correlations. Mean correlations obtained were compared between each case with Student tests with a type 1 error rate of 0.1%.

Phenotypes

The four traits studied in this paper were named according to Animal Trait Ontology for Livestock [49]. Measures concerning Egg Weight (EW), Egg Shell Color (ESC), Egg Shell Strength (ESS) and Albumen Height (AH) were recorded between 60 and 90 weeks of age. These measures corresponded to individual measures collected in individual cages. 75,121 eggs concerning 7983 birds were measured from (G0) to (G3).

During this period, all eggs were collected and transferred at Zootests (Ploufragan, France) to study egg quality traits. Analyses started by measuring Egg Weight (EW, in g). Then, a Minolta Chroma Meter was used to estimate three traits concerning egg shell: redness (a*), yellowness (b*) and lightness (L*) of egg shell. Egg Shell Color (ESC) was then calculated as $ESC = 100 - (L^* - a^* - b^*)$. Next, a compression machine was used to evaluate the shell static stiffness and a measure concerning Egg Shell Strength (ESS, in N) was done. ESS corresponded to the maximum force recorded before fracturing the shell. Finally, each egg was broken and Albumen Height (AH) was measured with a tripod.

Genomic evaluation strategies

One of the major point of interest for any breeders is to get good genomic evaluations from the low density genotyping simulated based on GGRS and ddRADseq approaches. The previous work of Herry el al. [52] showed a low impact of imputation errors on genomic evaluations of the selection candidates by using low density SNP chips with more than 3K SNPs. Consequently, the next step was to validate the simulated in silico GGRS and ddRADseq

methods by investigating the impact of imputed HD genotyping of the selection candidates on genomic evaluations.

EW, ESC, ESS and AH were evaluated with single-step GBLUP methodology [53] using BLUPF90 programs [54]. The four traits were jointly estimated according to a classical multi-trait animal model. A genomic evaluation "Full_HD" of the G1 selection candidates was done by using all available information (phenotypes and HD genotypes of ancestry, collaterals and progeny). This evaluation allowed to estimate the relative maximum genomic breeding value of each selection candidate G1 and to compare this value to the results of the different genomic evaluations. The "Full_HD" evaluation was also used to estimate the genetic parameters of the model. The genetic and residual variance components were estimated with remlf90 [54]. After fixed, the different genomic evaluations were carried out with blupf90.

The first objective was to investigate the impact of imputation errors on genomic evaluation based on ancestry. A genomic evaluation based on ancestry was done with phenotypes of the individuals of the first generation (G0) and with HD genotyping of the 1027 sires of the two first generations (G0+G1). A second genomic evaluation was done by replacing the HD genotyping of the selection candidates G1 by their imputed HD genotyping, obtained with the two RADseq approaches. For each simulated GGRS and ddRADseq methods and for each trait, the reordering of the selection candidates was estimated with Spearman correlations calculated between true HD Genomic Estimated Breeding Value (GEBV) and imputed HD GEBV. Spearman correlations were calculated for the 67 breeders from G1 having at least 10 offspring in G2. They were also calculated for the top 150 individuals from G1 according to each trait. The 67 breeders from G1 were not all included in the top 150 individuals from G1 because the 67 breeders were selected according to a multi-trait index, whereas the top 150 individuals were different following the trait studied.

The second objective was to study the attainable relative accuracy with imputation. The "Full_HD" GEBV represented the maximum of relative accuracy attainable regarding this genomic evaluation with all information. Thus, the results of the "Full_HD" genomic evaluation of the selection candidates G1 were compared with those from the genomic evaluation based on ancestry with imputed HD genotyping of the selection candidates G1. For each simulated RADseq methods and for each trait, Pearson correlations were calculated for the 67 breeders of G1 between "Full_HD" GEBV and GEBV based on ancestry with imputed HD genotyping of the selection candidates G1.

Results

Fragment size distribution

The enzymatic digestion of the genome was simulated with the four different enzymes, as well as with the double digestion with TaqI and PstI. The evolution of the number of DNA fragments obtained using the different enzymes according to the size of the fragments is summarized in Figure2. The total number of DNA fragment and the number of fragment ranging between 200 and 500bp is presented in Table 3.

In silico digestion of DNA showed, for each enzyme, a similar pattern for the evolution of the number of fragment according to their length. EcoRI was the enzyme generating the lower number of fragments (270,629) whereas PstI and AvaII generated the higher amount of fragments (respectively 829,382 and 869,482). The results were similar for the number of fragment ranging between 200 and 500 bp with a lower number of fragment for EcoRI (21,267) and a higher number for PstI and AvaII (respectively 165,804 and 178,980) The number of fragment obtained with the double-digest TaqI and PstI (530,105) did not correspond to the sum of the fragments obtained with TaqI and PstI separately, since the two restriction sites framing a fragment could not be restriction sites for only one enzyme. 128,823 fragments ranging between 200 and 500 were obtained with the double digestion.

Finally, PstI and AvaII, and the double digestion of DNA with TaqI and PstI generated the higher proportion of DNA fragments comprised between 200 and 500 bp with a proportion of respectively 19.99%, 20.58% and 24.30%. In contrast, EcoRI produced only 7.85% of fragments of interesting sizes.

For each enzyme, the distribution of the number of fragment according to the type of chromosome was studied (Figure 3a). For each enzyme, the number of fragment decreased going from macro-chromosomes to micro-chromosomes. Concerning sexual chromosome Z, this number of fragment was close to the number obtained for macro-chromosomes. In addition, as previously seen in Table 3 with the total number of fragment of interesting sizes (200 - 500 bp) with the different enzymes, the distribution of the fragments was different for each type of chromosome according to the enzyme used. EcoRI enabled to get respectively 2726 ± 1009, 585 ± 142, 171 ± 125 and 1810 fragments for macro-chromosomes, intermediate chromosomes, micro-chromosomes and sexual chromosome Z. The use of AvaII to digest DNA enabled to get respectively 18,873 ± 8315, 4792 ± 792, 2645 ± 809 and 15,694 fragments for macro-chromosomes, intermediate chromosomes, micro-chromosomes and sexual chromosome Z.

For each enzyme, the number of fragment of interesting sizes (200 – 500 bp) per megabase (Figure 3b) was also different according to the type of chromosome. Indeed, for EcoRI, there

was a decrease in the ratio from macro-chromosomes (22.6 ± 0.5) to intermediate chromosomes (20.1 ± 0.5) to micro-chromosomes (15.0 ± 4.5). For the 3 other enzymes and the double use of TaqI and PstI, the ratio increased. The most extreme case was for PstI with an increase in the ratio from 126.1 ± 16.1 to 189.9 ± 22.8 to 346.5 ± 79.4 for macro-chromosomes, intermediate chromosomes and micro-chromosomes, respectively. Finally, excepted for AvaII, the results concerning sexual chromosome Z were similar to those obtained for macro-chromosomes.

Covering between high density SNP chip and simulated data

Based on their physical positions, the fragments ranging between 200 and 500 bp were used to produce the bed file used by Plink to extract for the 1027 imputed sequenced individuals all SNPs located on the windows covered by the fragments. This allowed to extract from 46,568 SNPs with EcoRI to 427,141 and 470,425 SNPs with PstI and AvaII respectively. The double digestion of DNA with TaqI and PstI also enabled to get 318,408 SNPs.

The covering between the number of SNPs detected with the GGRS approach and the SNPs that can be genotyped with the HD SNP chip was then studied. It showed that the number of SNPs in common was quite reduced for EcoRI with only 1797 SNPs. For TaqI, AvaII and PstI, this number of SNPs increased respectively from 4126 to 12,453 and to 14,390 SNPs. The number of SNPs in common between ddRADseq and the HD SNP chip was 11,193 SNPs. Finally, among all the SNPs detected with the GGRS and ddRADseq approaches, the proportion of SNPs in common with the HD SNP chip was comprised between only 2.65% and 3.86% for AvaII and EcoRI respectively.

Distribution of SNP on chromosomes

Before studying imputation accuracy based on the SNPs in common between the HD SNP chip and the SNPs detected with the RADseq methods, the distribution of these SNPs according to the type of chromosome was studied (Figure 4). From macro-chromosomes to micro-chromosomes, the ratio of SNP per megabase was rather stable for EcoRI with a slight increasing trend from 1.6 ± 0.2 SNP.Mb^{-1} to 2.3 ± 0.7 SNP.Mb^{-1}. However, for the other enzymes and the double-digestion of DNA with TaqI and PstI, the results showed an increase in the ratio going from macro-chromosomes to micro-chromosomes. Indeed, for TaqI and PstI, the ratio increase respectively from 3.0 ± 0.3 SNP.Mb^{-1} and 8.9 ± 1.4 SNP.Mb^{-1} to 9.4 ± 3.5 SNP.Mb^{-1} and 42.6 ± 16.3 SNP.Mb^{-1}. The ratios for the double digestion with TaqI and PstI were comprised between the ratios for TaqI and PstI. The case of sexual chromosome Z was particular with a lower ratio than what could be observed for the other type of chromosome, for each enzyme tested. The ratio was only 0.8 SNP.Mb^{-1} for EcoRI and 4.4 SNP.Mb^{-1} for PstI.

Imputation accuracy

Imputation accuracy of the 580 selection candidates of the second generation (G1) with simulated low density genotyping obtained with the RADseq approaches was studied for each enzyme used. These individuals were imputed from the HD genotypes of the 447 individuals of the first generation (G0).

The mean correlation between true and imputed HD genotyping was 0.7906 from 1797 SNPs with EcoRI, 0.9121 from 4126 SNPs with TaqI, 0.9691 from 11,193 SNPs with TaqI and PstI, 0.9699 from 12,453 SNPs with AvaII and 0.9735 from 14,390 SNPs with PstI (Table 4). There was an increase in mean correlations with an increase in the number of SNPs used to impute the selection candidates.

The influence of the type of chromosome on mean correlations, for each enzyme used, was also studied (Figure 5). EcoRI presented a decrease in mean correlations going from macro-chromosomes (0.81 ± 0.03) to intermediate chromosomes (0.77 ± 0.02) to micro-chromosomes (0.66 ± 0.11). On the contrary, the other enzymes as well as the double use of TaqI and PstI showed results that were rather stable. Indeed, for TaqI the mean correlations were 0.91 ± 0.01, 0.92 ± 0.01 and 0.91 ± 0.02 for macro-chromosomes, intermediate chromosomes and micro-chromosomes respectively. Similarly, for PstI the mean correlations were 0.97 ± 0.00, 0.97 ± 0.00 and 0.98 ± 0.01 for macro-chromosomes, intermediate chromosomes and micro-chromosomes respectively.

Finally, EcoRI was the enzyme which presented the lower results for all type of chromosomes and AvaII and PstI, as well as the double digestion with TaqI and PstI, showed the higher results for all type of chromosomes, with slightly higher results for PstI compared to AvaII and the double digestion. However, there was no significant difference for macro-chromosomes between AvaII and PstI. Likewise, there was no significant difference for intermediate and micro-chromosomes with AvaII and the double use of TaqI and PstI.

The results concerning sexual chromosome Z were significantly lower than mean correlations obtained for all type of chromosomes and for each enzyme, excepted for EcoRI with no significant difference with the mean correlation of intermediate chromosomes.

Impact on genomic evaluations

Impact of imputation errors

For each enzyme used, the impact of imputation errors was studied by comparing with Spearman correlations the results of a genomic evaluation based on ancestry with true HD genotyping or imputed HD genotyping of the selection candidates G1. This enabled to estimate the reordering of the selection candidates. Spearman correlations were calculated for the best

150 individuals of G1 for each trait studied and for the 67 breeders of G1 having at least 10 offspring in the next generation G2 (Table 5).

Concerning the top 150 individuals of G1 for each trait studied, the results were significantly lower for EcoRI than the results obtained for the other enzymes. With the enzyme EcoRI, Spearman correlations were 0.8430, 0.6481, 0.8226 and 0.8427 for EW, ESC, ESS and AH respectively. In addition, among all correlations for the different traits and the different enzymes, the lowest was obtained using EcoRI for ESC. For TaqI, the results were significantly higher than those obtained with EcoRI. Highly correlated results were obtained for each trait with correlations higher than 0.97 with ddRADseq and with AvaII or PstI. The results for these 3 last cases were not significantly different from each other for the different traits studied.

Concerning the 67 breeders of G1, the results for each trait were also lower for EcoRI compared to the results obtained for the ddRADseq and for the simple digestions with TaqI, AvaII or PstI. Spearman correlations were 0.9450, 0.8854, 0.9255 and 0.9096 for EW, ESC, ESS and AH respectively with the enzyme EcoRI. However, excepted for ESC, the results were not significantly different from the results obtained using TaqI. The results concerning the ddRADseq or the use of AvaII or PstI were not significantly different from each other and all enabled to get correlations above 0.99, thus indicating a very reduced reordering of the 67 breeders.

Finally, with EcoRI or TaqI, the results obtained for the top 150 individuals for each trait were significantly lower than those obtained for the 67 breeders with the exception of AH with EcoRI and ESS with TaqI. For AvaII, PstI and the ddRADseq, the results concerning the top 150 individuals for each trait were also lower than those obtained for the 67 breeders but the differences were not significant.

Impact on relative accuracy of genomic evaluation

The impact on relative accuracy of genomic evaluation was studied by comparing with Pearson correlations the results of the "Full_HD" genomic evaluation and the results of the different genomic evaluations based on ancestry with imputed HD genotyping of the selection candidates of G1. Pearson correlations were calculated for the 67 breeders of G1 having at least 10 offspring in the next generation G2 (Table 6). These results were compared to the Pearson correlation between true "Full_HD" GEBV and true HD GEBV based on ancestry for the 67 G1 breeders. This represented the maximum of relative accuracy attainable with HD information.

With EcoRI, excepted for ESC, Pearson correlations for the different traits were all lower than those obtained with the other enzymes or the HD SNP chip. Indeed, the correlations were

0.3774, 0.3420 and 0.4261 for EW, ESS and AH, respectively. With the HD SNP chip, the correlations for EW, ESS and AH were respectively 0.4713, 0.3940 and 0.4802. On the contrary, the results were higher for ESC with a correlation of 0.2962 with EcoRI and 0.2460 with the HD SNP chip. However, these differences were not significant. With EcoRI, the standard errors were ±0.11 for EW and AH and ±0.12 for ESC and ESS. Concerning the HD SNP chip, the standard errors were ±0.11 for EW and ±0.12 for ESC, ESS and AH.

The results concerning the other enzymes were closer to the maximum of relative accuracy attainable with HD information but the differences observed were also not significant.

Discussion

Choice of the restriction enzyme

The major key point of a good restriction enzyme for genotyping is to cut DNA into a number of fragment of appropriate size and also to avoid frequently cutting DNA sequences which could led to a large number of too small fragments [55]. The diversity of restriction enzyme concerning the length, the position of the cut site, the AT or GC content in their recognition site and their sensitivity to methylation are factors than can impact this major key point [19]. Thus, the choice of a suitable restriction enzyme has to be consistent with the study and the species studied.

In this study, to determine the choice of the restriction enzyme, the total number of fragment to be sequenced per sample to enable a sufficient covering between SNP detected with the different RADseq methods and SNP on the HD SNP chip has to be calculated. According to Liao et al. [36], the total number of fragment can be calculated by dividing the size of the chicken genome by the extent of the linkage disequilibrium. This extent was proven very variable between breeds, lines and chromosomes [56, 57, 58]. Thus, concerning the genome, the extent of useful LD (r^2>0.3) [59, 60] for the line was 200-250kb. Concerning macro-chromosomes and micro-chromosomes, the extent of useful LD was respectively 400-450kb and 100-150kb. The most extreme case was for sexual chromosome Z with an extent of useful LD of 850-900kb. In addition, the International Chicken Genome Sequencing Consortium [61] showed that the size of the chromosome was inversely correlated to the recombination rate, to the methylations and to the gene density. Thus, it could be useful to densify the number of fragment on the micro-chromosomes. Assuming a useful LD of 0.3 and an extent of 100kb, a total fragment number of 10,000 of interesting size would be sufficient.

EcoRI generated 21,267 fragments of interesting sizes but only 1797 SNPs were found in common with the SNPs detected with the GGRS approach and the SNPs on the HD SNP chip.

To ensure a better covering between the SNPs of the two approaches, the total number of fragment was multiplied by 10 or 20, to obtain 100,000 or 200,000 fragments. The ddRADseq with TaqI and PstI enabled the production of 128,823 fragments and 11,193 SNPs in common with the SNPs on the HD SNP chip. AvaII and PstI produced respectively 178,980 and 165,804 fragments of interesting size and respectively 12,453 and 14,390 SNPs were detected in common with the SNP on the HD SNP chip.

By looking at the distribution of the fragments (Figure 3) and the SNPs in common with the HD SNP chip (Figure 4) on the different type of chromosomes, the results showed that DNA digestion with AvaII and PstI, or the double digestion with TaqI and PstI enabled to densify the fragments and the SNPs on micro-chromosomes. This was particularly due to their GC-rich recognition sites. Since micro-chromosomes were proven to have a higher GC content [61] and thus a higher SNP density, this explained the differential distribution between the different type of chromosomes. Conversely, EcoRI had a recognition site less rich in GC which led to a lower densification of the fragments and a lower number of SNPs on micro-chromosomes. The AT or GC content of the recognition site of the restriction enzyme is thus of major importance when choosing the most adapted restriction enzyme for the specie studied.

Finally, the sensitivity to the methylation has to be taken into account to choose the restriction enzyme. Indeed, if a type of methylation (CpG, *dam* or *dcm*) occurs in a restriction site, the methylation sensitive enzyme will not be able to cleave the DNA. This would lead to the loss of the restriction site and the creation of a longer fragment which may not be sequenced if the size exceeds 500 bp. However, the use of methylation sensitive restriction enzyme can also avoid to cut some repetitive regions allowing to target lower copy region with a higher efficiency [62]. In this study, only PstI was not sensitive to methylations whereas EcoRI was sensitive to CpG methylations, TaqI to *dam* methylations and AvaII to CpG and *dcm* methylations. This sensitivity was not simulated in this study but has to be taken into account with real data since this would lead to more variability between the reads for the different individuals.

Management of heterogeneity between individuals

No matter the restriction enzyme used, a major drawback of the RADseq methods is the management of the variability between individuals. Indeed, as seen previously, the sensitivity to methylations can lead to variability between reads for the different individuals. Assuming a methylated nucleotide in a restriction site for an individual A, the restriction enzyme will not be able to cut the DNA. For the same nucleotide, not methylated, in a restriction site for an individual B, the restriction enzyme will be able to cleave the DNA. The reads obtained will

thus be different for these two individuals. Another factor that will increase the variability between individuals is the polymorphism occurring in restriction sites. This can lead to allele dropout [18]. This phenomenon occurs for heterozygous individuals with a polymorphism in the restriction site, resulting in a failure to cut the DNA. The missing allele will therefore not be sequenced (null allele), heterozygous individuals will be considered as homozygous for this SNP and thus, it will introduce genotyping errors. By failing to cut the DNA, this phenomenon can also create longer fragments which will not have an interesting size for the study and that will not be sequenced. The polymorphism rate on a restriction site is variable depending on the enzyme used. Theoretically, the polymorphism rate is less important for a 4-cutter enzyme than for a 6-cutter enzyme. The polymorphism rate was also not modeled in this in-silico study but has to be taken into account with real data.

Finally, the read depth variability between loci is a factor that can influence the quality of the reads and thus the quality of the genotypes [19]. The read depth does not create genotyping errors but is a factor that strongly impacts the number of variants that can be detected after filtering the reads. In addition, with only one read it is impossible to correctly call heterozygous individual since the read informs on just one allele. With two reads it is possible to call correctly heterozygous individual but the probability to read the same allele is 50% thus leading to an inaccurate calling. Gorjanc et al. [42] showed that the calling of an heterozygous individual from n sequence reads can be represented by n draws from a Bernoulli distribution with a probability of $1 - (\frac{2}{2^n})$. The read depth variability can be created during PCR amplification with the preferential amplification of fragment that are rich in GC content, or with the preferential amplification of short fragment compared to the long fragments [18].

However, many studies have shown that this variability can be handled without a big loss of accuracy. Brouard et al. [41] studied in bovine the raw reads with different levels of read depth, including or not a filter concerning the genotype quality. They also focused on the minimum call rate (CR) of the SNPs and the minimum minor allele frequency (MAF) and thanks to the imputation they filled missing data. Based on the different filters, their imputation accuracy with FImpute ranged between 70% and 83.3%. The maximum of accuracy was obtained for a minimum read depth of 4, a minimum call rate of 0.4 and a minimum MAF of 0.02. For the same parameters but with a minimum call rate of 0.2, their imputation accuracy was 73.3%. They also showed that filtering for a minimum MAF of 0.02 enabled to get slightly larger and better imputed datasets than filtering for a minimum MAF of 0.05. In addition, Torkamaneh and Belzile [39], showed for Canadian soybean lines that their imputation accuracy ranged between 86% to 94% for a maximum amount of missing data of 20% and 80% respectively.

Conversely to the previous study, they noted that increasing the maximum amount of missing data (thus lowering the minimum call rate) up to 80% led to better imputed missing data than with a maximum amount of missing data of 20%. The difference in the conclusion of the two studies may be due to the shorter size of the haplotype and the lower LD extent of cattle compared to those of soybean. Indeed, the average distance for having a LD extent corresponding to half of its maximum value was comprised between 75kb and 150kb for wild or cultivated soybean, whereas it was less than 10kb for cattle. The chicken LD extent for the line studied was comprised between 250kb and 300kb. In addition, as seen on Figure 6, for each enzyme, the whole SNPs detected with the different in-silico GGRS and ddRADseq approaches were separated by a distance much lower than 250kb. Indeed, for EcoRI, the distance between SNPs was 19.63 ± 0.43kb, 21.09 ± 2.46kb, 30.32 ± 9.47kb and 46.76kb respectively for macro-chromosomes, intermediate chromosomes, micro-chromosomes and sexual chromosome Z. For PstI, the distance was 2.38 ± 0.20kb, 2.18 ± 0.18kb, 1.45 ± 0.63kb and 4.80kb respectively for macro-chromosomes, intermediate chromosomes, micro-chromosomes and sexual chromosome Z. Thus, we could expect as well that increasing the maximum amount of missing data would result in the adding of markers with a higher extent of LD than what observed in cattle. These markers would be helpful to impute accurately the neighboring markers and would lead to higher imputation accuracy of missing data. Finally, a rather low loss of detected SNPs with the different RADseq methods on real data could be observed compared to the expected number of SNPs with the in-silico studies.

Imputation accuracy

The results showed an increase in mean correlations between true and imputed genotypes with an increasing number of SNP in common with the SNP HD ship. Indeed, the mean correlation was 0.7906 with 1797 SNPs for EcoRI, 0.9691 with 11,193 SNPs for the double digestion TaqI and PstI, and 0.9735 with 14,390 SNPs for PstI. This increase in mean correlations was consistent with the results found in the literature [5, 7, 8, 16]. To realize imputations, a higher number of SNPs to go back to the HD genotypes resulted in an increased number of genotypes available to identify the corresponding reference haplotypes in the haplotype reference library. Thus, the probability to identify a wrong haplotype for the selection candidate decreased.

In addition, these results can be compared to the results presented in Herry et al. [16] were several low density SNP chips were designed according to an equidistant methodology or methodology based on linkage disequilibrium. Indeed, the same individuals were studied in this study as well as in the previous study. The results obtained with the different RADseq methods (Table 4) can be compared to the results of four low density SNP chips designed with the

equidistant methodology (Table 7). With EcoRI, the results were significantly lower than those obtained with a low density of 2K SNPs. This was partly due to the difference in SNP density (216 SNPs). For very low density, an increase of few SNPs can impact mean correlations.

With TaqI, the results were significantly higher than the results for the low density SNP chip of 4K SNPs. This was also partly due to the difference in SNP density (103 SNPs). For AvaII, PstI, and the association of TaqI and PstI, the results were significantly higher than those obtained with equidistant low density SNP chip of 15K SNPs, with in addition less SNPs than on the low density SNP chip. In details, the equidistant methodology used in Herry et al. [16] resulted in a decrease in the mean correlation from macro-chromosomes to micro-chromosomes. In this study, only EcoRI led to a decrease in mean correlations from macro-chromosomes to micro-chromosomes. The other enzymes as well as the double use of TaqI and PstI led to rather stable results among chromosomes.

To understand these differences, the results needed to be studied according to the type of chromosome. With the equidistant methodology used in Herry et al. [16], the SNP distribution per megabase was assumed to be the same for each type of chromosome and a decrease in mean correlation was observed with a decrease in chromosome size. For instance, the 10Kequi was assumed to have 10 SNP.Mb^{-1} for each type of chromosome and mean correlations decreased with the chromosome size. With EcoRI, the ratio was only 1.6 SNP.Mb^{-1} for macro-chromosomes, 2.1 SNP.Mb^{-1} for intermediate chromosomes and 2.3 SNP.Mb^{-1} for micro-chromosomes and mean correlations decreased as well with the size of the chromosomes. However, compared to the close and expected ratio of 2 SNP.Mb^{-1} for the 2Kequi, the significantly lower results obtained with EcoRI could be explain by the distribution of the SNPs on chromosomes which were not equidistant. On the extreme opposite, with PstI, the ratio was 8.9 SNP.Mb^{-1} for macro-chromosomes, 18.8 SNP.Mb^{-1} for intermediate chromosomes and 42.7 SNP.Mb^{-1} for micro-chromosomes and mean correlations were rather stable according to the type of chromosome. For the 15Kequi, the ratio was expected to be around 15 SNP.Mb^{-1} for each type of chromosome with a decrease in mean correlations with chromosome size. The use of PstI enabled to decrease the number of SNPs detected on macro-chromosomes and to highly increase the number of SNPs detected on micro-chromosomes. Thus, in addition to the non-uniform distribution of the SNPs, there was an optimization in the number of SNPs on macro-chromosomes and a densification on micro-chromosomes. All these factors can explain the higher results obtained for enzymes enabling to optimize the number of SNPs on macro-chromosomes and to densify this number on micro-chromosomes compared to an equidistant methodology. The use of restriction enzyme that enabled to densify the Gallus gallus micro-chromosomes was thus of major importance to get high imputation accuracy.

The case of sexual chromosome Z was particular for each enzyme tested with a lower imputation accuracy and a SNP distribution per megabase lower than what was observed for the other type of chromosome. With the HD SNP chip, among the 26,867 SNPs of Z chromosome, only 10,113 SNPs were informative for the line studied. Among the 10,113 SNPs, only 63, 113, 314, 316 and 360 SNPs were in common with those obtained with EcoRI, TaqI, TaqI and PstI, AvaII and PstI, respectively. However, by looking at the ratio of the fragment number of interesting sizes per megabase (Figure 3b), the values concerning sexual chromosome Z were close to the ratio of macro-chromosomes, with higher values for AvaII and PstI. Thus, the lower SNP distribution per megabase was explained by a lower number of SNPs in common with the SNPs of the HD SNP chip after quality control which was quite reduced. This led to a lower imputation accuracy observed for chromosome Z whatever the enzyme.

Impact on genomic evaluation

Impact of Imputation errors

Concerning the top 150 individuals of G1 and the 67 G1 breeders, the results showed that Spearman correlations increased for all traits studied with an increase in the number of SNPs in common with the HD SNP chip from EcoRI (1797 SNPs) to AvaII and PstI (respectively 12,453 and 14,390 SNPs). This increase in mean correlations was consistent with the literature. Aliloo et al. [7] showed in bovine that the correlations between GEBV estimated with the HD SNP chip (Illumina BovineHD BeadChip with 777K SNPs) and GEBV estimated with imputed low density SNP chip of 4013 and 25,410 SNP were respectively 0.9398 and 0.9927. The results can also be compared to the results presented in Herry et al. [51] were several low density SNP chips were designed according to an equidistant methodology or a methodology based on linkage disequilibrium. The same individuals were studied in this study and the impact of the use of these low density SNP chip on genomic evaluations was studied. Thus, the results obtained with the different RADseq approaches (Table 5) can be compared to the results of four low density SNP chips designed with the equidistant methodology (Table 8).

With EcoRI, the results were lower than those obtained with a low density of 2K SNPs for all traits studied. The differences were not significant excepted for ESC of the top 150 individuals for which the correlations were 0.6481 with EcoRI and 1797 SNPs, and 0.7956 with an equidistant low density SNP chip of 2013 SNPs. The results concerning TaqI were higher but not significantly different from the results obtained with an equidistant low density SNP chip of 4K SNPs. Finally, the results for AvaII and PstI and the association of TaqI and PstI were also higher but not significantly different from the results obtained with an equidistant low

density SNP chip of 15K SNPs. However, the number of SNPs used with AvaII or PstI, and with TaqI and PstI was lower than the 14,963 SNPs of the 15Kequi SNP chip. Thus, higher results were obtained with lower SNPs with these two enzymes. These differences can be explained by the SNP distribution according to the type of chromosome. As seen previously, the gene density is higher on micro-chromosomes than on macro-chromosomes [61]. Combined to the optimization in the number of SNPs on macro-chromosomes and the densification on micro-chromosomes, this can explain why higher but not significantly different correlations were obtained with TaqI, TaqI and PstI, AvaII and PstI than what could be obtained with equidistant low density SNP chips.

Impact on relative accuracy of genomic evaluation

The impact on relative accuracy of genomic evaluations by comparing the results of the "Full_FD" genomic evaluations and genomic evaluations based on ancestry with imputed HD genotyping showed that the results were lower for EcoRI, excepted for ESC, compared to the results obtained with the other enzymes or the HD SNP chip. The differences observed were not significant. The results concerning the other enzymes were closer to the maximum of relative accuracy attainable with HD information but the differences observed were also not significant. The results were consistent with the previous results showing a low impact of imputation errors on genomic evaluations of the 67 breeders. Indeed, excepted for EcoRI, Spearman correlations were higher than 0.98 for each trait studied by comparing the GEBV based on ancestry obtained with true HD genotyping or imputed HD genotyping. These results were also in agreement with the literature. Chen et al. [63] showed in bovine that the accuracy of genomic evaluation calculated with Pearson correlations between direct genomic values with true 50K genotyping and bull proofs were 0.61 for milk yield and 0.62 for somatic cell score. With imputed 50K genotyping from 6K SNPs, Pearson correlations were also 0.61 for milk yield and 0.62 for somatic cell score. Likewise, Herry et al. [51] showed that the use of equidistant low density SNP chips with a SNP density higher than 2K SNPs led to no difference in Pearson correlations with what could be obtained by using the HD SNP chip. Indeed, for the 2Kequi SNP chip, the correlations were 0.4929 for EW, 0.3126 for ESC, 0.3775 for ESS and 0.4157 for AH. For the 15Kequi SNP chip, the correlations were 0.4889 for EW, 0.2574 for ESC, 0.3955 for ESS and 0.4762 for AH. There was no significant difference with the correlations obtained with the HD SNP chip.

Therefore, the use of restriction enzymes to realize GGRS or ddRADseq approaches and the use of the SNPs detected in common with the SNPs of the HD SNP chip could be an interesting

alternative to low density SNP chip without a decrease in relative accuracy of genomic evaluation of the selection candidates.

Balance between covering and read depth

As seen previously, the use of different restriction enzyme led to different number of fragment of interesting size and finally a different number of SNPs detected in common with the SNPs of the high density SNP chip. The sequencing costs depend on the number of fragment, the number of individuals and the depth of each fragment for each individual. In the previous study using the GGRS approach in chicken [36, 37], a sequencing depth between 5X and 7X was expected. With a ddRADseq approach, a sequencing depth higher than 7X was expected [28]. Assuming this depth, increasing the number of fragment to be sequenced would lead to a decrease in the number of individuals that could be sequenced together in a lane and finally to an increase in sequencing costs. In the opposite, decreasing the number of fragments to be sequenced would lead to an increase in the number of individuals that could be sequenced together in a lane. But a decrease in the number of fragments would result in fewer SNPs detected in common with the SNPs of the HD SNP chip. Given the heterogeneity expected between individuals with the different RADseq approaches, using EcoRI or TaqI could be too optimistic since the number of SNPs detected in common with the SNPs of the HD SNP chip would be lower than expected. On the other hand, with PstI, AvaII and the association of TaqI and PstI, given the number of fragments and the number of SNPs detected in common with the SNPs of the HD SNP chip, imputation accuracy, the results of genomic evaluation, and the expected costs of less than 50$ per individual, the use of these enzymes could be an interesting alternative to low density SNP chips.

Finally, the availability of sequencer with higher performances allowing the sequencing of more individuals in a single lane of a flowcell with increasing read density are now plummeting the sequencing costs for the RADseq approaches. However, the costs of the different kits to prepare DNA libraries can still be expensive.

Combining RADseq data and SNP chip for genomic selection

Among the total number of SNPs detected with the GGRS and ddRADseq methods, only 2.64% for AvaII to 3.85% for EcoRI were in common with the SNPs of the HD SNP chip. However, the remaining SNPs can be useful for imputation and genomic prediction. Indeed, Brouard et al. [41] showed that using both SNP chip and GBS panel enabled better imputation accuracy of missing data than using only the GBS panel. Indeed, the addition of SNP chip panel led to an increased number of high quality genotypes available to identify the corresponding reference

haplotypes in the haplotype reference library. This resulted in better imputation accuracy of missing data than using only GBS panel.

Likewise, Torkamaneh and Belzile [39] focused on the imputation of untyped loci in soybean. They showed that combining the GBS and SNP chip panel for their candidates and imputing them to a WGS level, from a set of reference sample, resulted in imputation accuracy of 88.1%. With only GBS panel, imputation accuracy was 80%. Thus, the combination of SNP chip and GBS approach could be useful to get high imputation accuracy of untyped markers.

Finally, Poland et al. [38] and Elbasyoni et al. [40] dealt with this topic in depth by studying the impact of the use of GBS or SNP chip panels on genomic evaluations of wheat. Genomic prediction accuracies were assessed as the correlation between GEBV and phenotypic values. At equivalent SNP density, they showed that significantly higher correlations were obtained for different traits with GBS panel compared to the results obtained with SNP chip. Increasing the number of SNP in the GBS panel by increasing the percentage of missing data led to correlation not significantly different from the results obtained with the previous GBS panel.

Thus, applied to chicken, combining SNP chip and RADseq panel could be helpful to accurately impute missing data among RADseq dataset. But combining SNP chip and RADseq panel to impute untyped loci would need a reference population with higher SNP density than what could be obtained with RADseq methods and SNP chip separately, to ensure a covering of both panel. This is currently still too expensive for a routine use in a selection scheme. Finally, the use of RADseq techniques instead of SNP chip could also be interesting to get higher genomic evaluation accuracy. But it currently needs big changes in genotyping strategies by creating reference populations genotyped with a RADseq approach.

Conclusion

A common use to reduce the cost of genomic selection is to use low density SNP chip in combination with imputation for genotyping the selection candidates instead of using a HD SNP chip. Depending on the SNP density, this results in genomic evaluation as good as genomic evaluation done with HD genotyping. But nowadays, restriction site-associated DNA sequencing (RADseq) or Genotyping-By-Sequencing (GBS) techniques are alternatives and cost-efficient solutions allowing to sequence only a fraction of the genome and finally to simultaneously detect and genotype SNPs. The above studies showed that the use of restriction enzymes enabled the detection and genotyping of thousands of SNPs. However, to use these techniques as alternative to low density SNP chip, a certain amount of SNPs in common with the SNPs on the HD SNP chip was needed. Indeed, without any SNP in common with the SNPs of the HD SNP chip, imputation to go back to the HD SNP chip level was impossible since the

reference population was genotyped with the HD SNP chip. EcoRI showed that only 1797 SNPs were in common with the HD SNP chip leading to an imputation accuracy lower than 0.80. On the other hand, the use of AvaII or PstI, or the association of TaqI and PstI enabled the detection of more than 10K SNPs in common with the HD SNP chip leading to imputation accuracy higher than 0.97. In addition, excepted for EcoRI, imputation accuracies obtained with the different enzymes were all higher than what could be obtained with equidistant low density SNP chips. This was due to the non-uniform distribution of SNPs detected with the RADseq techniques. There was an optimization in the number of SNPs on macro-chromosomes and a densification on micro-chromosomes. However, the objectives of the breeders are to get good genomic evaluations. By linking imputation accuracy and impact on genomic evaluation, the different studies showed that the impact on the reordering of the top 150 individuals for each trait studied was low with AvaII, PstI and ddRADSeq with Spearman correlations higher than 0.97 for each trait. On the contrary, EcoRI led to lower Spearman correlations, the most extreme case concerning ESC with a correlation of 0.6481. Concerning the 67 breeders, the results followed the same trends with higher correlations. Then, by investigating the impact on relative accuracy of genomic evaluation, the only differences were observed for EcoRI and, excepted for ESC, lower but not significant results were obtained compared to the results obtained with the other enzymes or the HD SNP chip. The results concerning the other enzymes were not significantly different from the maximum of relative accuracy attainable with HD information. Finally, the use of AvaII or PstI to realize a GGRS approach, or TaqI and PstI to realize a ddRADseq approach could be good compromises between enzyme properties, the total fragment number that can be obtained, the final number of SNPs in common with the HD SNP chip, and imputation and genomic evaluation results. Given the expected and decreasing costs of less than 50$ per individual, the use of these enzymes could be an interesting alternative to low density SNP chips.

However, it is important to emphasize that these are just results from simulated data which did not allow to take into account the heterogeneity between individuals. Indeed, methylation sensitivity of enzymes, and polymorphism rate occurring in restriction site would result in a restriction enzyme that would not be able to cut the DNA and thus introducing variability between individuals. In addition, the read depth variability between loci is also a factor that can influence the quality of reads and thus the quality of the genotypes. But to some extent, it is possible to impute accurately missing data.

Finally, if the costs are too high to get a sufficient number of SNPs with high genotype quality in common with the SNPs of the HD SNP chip, the use of restriction enzyme could however

be seen as a new methodology to design low density SNP chip by selecting SNPs to include on the low density SNP chip.

References

1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819.
2. Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013;14:59.
3. Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, et al. Design of a bovine low-density SNP array optimized for imputation. PLoS One. 2012;7:e34130
4. Dassonneville R, Brøndum RF, Druet T, Fritz S, Guillaume F, Guldbrandtsen B, et al. Effect of imputing markers from a low density chip on the reliability of genomic breeding values in Holstein populations. J Dairy Sci. 2011;94:3679.
5. Dassonneville R, Fritz S, Ducrocq V, Boichard D. Short communication: Imputation performances of 3 low density marker panels in beef and dairy cattle. J Dairy Sci. 2012;95:4136.
6. VanRaden PM, Null JD, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, et al. Genomic imputation and evaluation using high-density Holstein genotypes. J Dairy Sci. 2012;95:1.
7. Aliloo H, Mrode R, Okeyo AM, Ni G, Goddard ME, Gibson JP. The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa. J Dairy Sci. 2018;101:1.
8. Bouquet A, Fève K, Riquet J, Larzul C. Précision de l'imputation de génotypages haute densité à partir de puces basse densité pour des individus de race pure et croisés Piétrain. Journées Recherche Porcine. 2015;47:1.
9. Cleveland MA, Hickey JM. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J Anim Sci. 2013;91:3583.
10. Grossi DA, Brito LF, Jafarikia M, Schenkel FS, Feng Z. Genotype imputation from various low-density SNP panels and its impact on accuracy of genomic breeding values in pigs. Animal. 2018;12:2235.
11. Hayes BJ, Bowman PJ, Daetwyler HD, Kijas JW, Van Der Werf JHJ. Accuracy of genotype imputation in sheep breeds: Genotype imputation in sheep. Anim Genet. 2012;43:72.

12. Moghaddar N, Gore KP, Daetwyler HD, Hayes BJ, van der Werf JHJ. Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. Genet Sel Evol. 2015;47:97.
13. Raoul J, Swan AA, Elsen JM. Using a very low-density SNP panel for genomic selection in a breeding program for sheep. Genet Sel Evol. 2017;49:76.
14. Heidaritabar M, Calus MPL, Vereijken A, Groenen MAM, Bastiaansen JWM. Accuracy of imputation using the most common sires as reference population in layer chickens. BMC Genetics. 2015;16:101.
15. Wang C, Habier D, Peiris BL, Wolc A, Kranis A, Watson KA, et al. Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide polymorphism panel in broiler chickens. Poult Sci. 2013;92:1712.
16. Herry F, Hérault F, Picard Druet D, Varenne A, Burlot T, Le Roy P, et al. Design of low density SNP chips for genotype imputation in layer chicken. BMC Genetics. 2018;19:108.
17. Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol. 2010;27:2534.
18. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17:81.
19. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12:499.
20. Jiang Z, Wang H, Michall JJ, Zhou X, Liu B, Soldberg Woods LC, et al. Genome wide sampling sequencing for SNP genotyping: Methods, challenges and future development. Int J Biol Sci. 2016;12:100.
21. Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One. 2008;3:e3376.
22. Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, et al. Marker Density and Read Depth for Genotyping Populations Using Genotyping-by-Sequencing. Genetics. 2013;193:1073.
23. Van Tassell CP, Smith TP, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Methods. 2008;5:247.

24. Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, et al. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res. 2011;21:610.
25. Wang S, Meyer E, McKay JK, Matz MV. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat Methods. 2012;9:808.
26. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6e19379.
27. Chen C, Ma Y, Yang Y, Chen Z, Liao R, Xie X, et al. Genotyping by genome reducing and sequencing for outbred animals. PLoS One. 2013;8:e67500.
28. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One. 2012;7:e37135.
29. Gardner KM, Brown P, Cooke TF, Cann S, Costa F, Bustamante C, et al. Fast and cost-effective genetic mapping in apple using next-generation sequencing. Genetics. 2014;4:1681.
30. Poland JA, Brown PJ, Sorrells ME, Jannink JL. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One. 2012;7:e32253.
31. van Orsouw NJ, Hogers RC, Janssen A, Yalcin F, Snoeijers S, Verstege E, et al. Complexity Reduction of Polymorphic Sequences (CRoPS™): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. PLoS One. 2007;2:e1172.
32. Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, et al. Genomic prediction in maize breeding populations with genoyping-by-sequencing. G3. 2013;3:1903.
33. De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG. Genotyping by-sequencing (GBS): A novel, efficient and cost-effective genotyping method for cattle using next-generation sequencing. PLoS One. 2013;8:e62137.
34. Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans R, Dong Y, et al. Design and Characterization of a 52K SNP Chip for Goats. PLoS One. 2014;9:e86227.
35. Zhai Z, Zhao W, He C, Yang K, Tang L, Liu S, et al. SNP discovery and genotyping using restriction-site-associated DNA sequencing in chickens. Anim Genet. 2015;46:216.
36. Liao R, Wang Z, Chen Q, Tu Y, Chen Z, Wang Q, et al. An efficient genotyping method in chicken based on genome reducing and sequencing. PLoS One. 2015;10:e0137010.

37. Pértille F, Guerrero-Bosagna C, da Silva VH, Boschiero C, Nunes JDD, Ledur MC, et al. High-throughput and Cost-effective Chicken Genotyping Using Next-Generation Sequencing. Sci Rep. 2016;6:26929.
38. Poland J, Endelman J, Dawson J, Rutkoski J, Shuangye W, Manes Y, et al. Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen. 2012;5:103.
39. Torkameneh D, Belzile F. Scanning and filling: Ultra-dense SNP genotyping combining genotyping-by-sequencing, SNP array and whole-genome resequencing data. PLoS One. 2015;10:e0131533.
40. Elbasyoni IS, Lorenz AJ, Guttieri M, Frels K, Baenziger PS, Poland J, et al. A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. Plant Sci. 2018;270:123.
41. Brouard JS, Boyle B, Ibeagha-Awemu EM, Bissonnette N. Low-depth genotyping-by-sequencing (GBS) in bovine population: strategies to maximize the selection of high quality genotypes and the accuracy of imputation. BMC Genetics. 2017;18:32.
42. Gorjanc G, Cleveland MA, Houston RD, Hickey JM. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet Sel Evol. 2015;47:12.
43. Wang Y, Cao X, Zhao Y, Fei J, Hu X, Li N. Optimized double-digest genotyping by sequencing (ddGBS) method with high density SNP markers and high genotyping accuracy for chickens. PLoS One. 2017;12:e0179073.
44. Warren WC, Hillier LDW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A new chicken genome assembly provides insight into avian genome structure. G3. 2017;7:109.
45. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.
46. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009;25:1754.
47. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297.
48. Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478.
49. Atol Ontology. INRA 2012. [http://www.atol-ontology.com]. Accessed 11 Mar 2015.
50. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12:115.

51. Quail MA, Gu Y, Swerdlow H, Mayho M. Evaluation and optimisation of preparative semi-automated electrophoresis systems for Illumina library preparation. Electrophoresis. 2012;33:3521.
52. Herry F, Picard Druet D, Hérault F, Varenne A, Burlot T, Le Roy P, et al. Interest of using imputation for genomic evaluation in layer chicken. Poult Sci. In submission.
53. Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 2009;92:4656.
54. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH. BLUPF90 and related programs (BGF90). In Proceedings of the 7th World Congress of Genetics Applied to Livestock Production. Montpellier; 2002.
55. Gurgul A, Miksza-Cybulska A, Szmatoła T, Jasielczuk I, Piestrzyńska-Kajtoch A, Fornal A, et al. Genotyping-by-sequencing performance in selected livestock species. Genomics. 2019;111:186.
56. Megens HJ, Crooijmans RPMA, Bastiaansen JWM, Kerstens HHD, Coster A, Jalving R, et al. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genetics. 2009;10:86.
57. Qanbari S, Hansen M, Weigend S, Preisinger R, Simianer H. Linkage disequilibrium reveals different demographic history in egg laying chickens. BMC Genetics. 2010;11:103.
58. Hérault F, Herry F, Varenne A, Burlot T, Picard-Druet D, Recoquillay J, et al. A linkage disequilibrium study in layer and broiler commercial chicken populations. In Proceedings of the 11th World Congress of Genetics Applied to Livestock Production. Auckland; 2018.
59. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nat Rev Genet. 2002;3:299.
60. Aerts J, Megens HJ, Veenendaal T, Ovcharenko I, Crooijmans R, Gordon L, et al. Extent of linkage disequilibrium in chicken. Cytogenet Genome Res. 2007;117:338.
61. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004;432:695.
62. Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz EZ, Hurwitz BL, et al. A first-generation haplotype map of maize. Science. 2009;326:1115.
63. Chen L, Li C, Sargolzaei M, Schenkel F. Impact of genotype imputation on the performance of GBLUP and Bayesian methods for genomic prediction. PLoS One. 2014;9:e101544.

Table 1. Summary of the different steps of quality control.

Genotypes filtration	**RI Line**
Individual Call Rate (<95%)	8
MAF (=0)	204,122
MAF (<0.05)	54,650
SNP Call Rate (<95%)	7541
Hardy-Weinberg equilibrium ($P<10^{-4}$)	12,538
SNP with unknown location	1748
SNP located on chromosome 16, 33, W or on linkage group LGE64	334
Pedigree Incompatibility problem	0
SNP retained for analyses	**300,028**
Animals retained for analyses	**2362**

Table 2. Summary of the different restriction enzymes used.

Enzyme	Recognition sequence	Methylation sensitivity
AvaII	GGWCC	CpG and *dcm* methylation
EcoRI	GAATTC	CpG methylation
PstI	CTGCAG	Not sensitive
TaqI	TCGA	*dam* methylation

W denotes A or T

Table 3. Summary of the fragments number obtained, the number of SNPs detected based on the fragments and the 1027 simulated sequences and the covering between the HD SNP chip and SNPs detected with the different RADseq methods.

	EcoRI	TaqI	TaqI+PstI	AvaII	PstI
Total Fragment number	270,629	425,312	530,105	869,482	829,382
Fragment number between 200 and 500 bp	21,267	51,163	128,823	178,980	165,804
Percentage of fragment between 200 and 500 bp	7.86%	12.03%	24,30%	20.58%	19.99%
Number of SNPs detected	46,568	122,248	318,408	470,425	427,141
Covering between SNPs detected and HD SNP chip	1797	4126	11,193	12,453	14,390

Table 4. Summary of the mean correlations of true and imputed genotypes obtained for the different enzymes for the 580 selection candidates of the second generation (G1).

	EcoRI	TaqI	TaqI_PstI	AvaII	PstI
Number of SNPs	1797	4126	11,193	12,453	14,390
Mean correlation	0.7906	0.9121	0.9691	0.9699	0.9735

Table 5. Evolution of Spearman correlations between true HD GEBV and imputed HD GEBV, according to each enzyme used for Egg Weight (EW), Egg Shell Color (ESC), Egg Shell Strength (ESS) and Albumen Height (AH), for genomic evaluations based on ancestry. Results are shown for the top 150 individuals for each trait and for the 67 breeders of G1.

	Number of SNPs	EW		ESC		ESS		AH	
		Top150	Breeders	Top150	Breeders	Top150	Breeders	Top150	Breeders
EcoRI	1797	0.8430	0.9450	0.6481	0.8854	0.8226	0.9255	0.8427	0.9096
TaqI	4126	0.9388	0.9914	0.9012	0.9833	0.9501	0.9847	0.9088	0.9813
TaqI_PstI	11,193	0.9913	0.9971	0.9779	0.9938	0.9904	0.9957	0.9859	0.9973
AvaII	12,453	0.9899	0.9975	0.9737	0.9949	0.9879	0.9951	0.9867	0.9958
PstI	14,390	0.9937	0.9980	0.9781	0.9959	0.9907	0.9943	0.9848	0.9949

Table 6. Evolution of Pearson correlations between true "Full_HD" GEBV and imputed HD GEBV based on ancestry for the 67 G1 breeders, according to each enzyme used for Egg Weight (EW), Egg Shell Color (ESC), Egg Shell Strength (ESS) and Albumen Height (AH).

	Number of SNPs	EW	ESC	ESS	AH
EcoRI	1797	0.3774	0.2962	0.3420	0.4261
TaqI	4126	0.4476	0.2453	0.3906	0.4478
TaqI_PstI	11,193	0.4740	0.2442	0.3869	0.4684
AvaII	12,453	0.4681	0.2430	0.3859	0.4794
PstI	14,390	0.4664	0.2450	0.3953	0.4689
HD SNP chip	300,028	0.4713	0.2460	0.3940	0.4802

The line HD SNP chip corresponds to the Pearson correlation between true "Full_HD" GEBV and true HD GEBV based on ancestry for the 67 G1 breeders.

Table 7. Summary of the mean correlations of true and imputed genotypes obtained for different low density SNP chips designed in Herry et al. [16] for the 580 selection candidates of the second generation (G1).

	2Kequi	4Kequi	10Kequi	15Kequi
Number of SNPs	2013	4023	10,001	14,963
Mean correlation	0.8319	0.9070	0.9553	0.9670

Table 8. Evolution of Spearman correlations between true HD GEBV and imputed HD GEBV, according to different low density SNP chips designed in Herry et al. [16, 51] for Egg Weight (EW), Egg Shell Color (ESC), Egg Shell Strength (ESS) and Albumen Height (AH), for genomic evaluations based on ancestry. Results are shown for the top 150 individuals for each trait and for the 67 breeders of G1.

	Number of SNPs	EW		ESC		ESS		AH	
		Top150	Breeders	Top150	Breeders	Top150	Breeders	Top150	Breeders
2Kequi	2013	0.8752	0.9643	0.7956	0.9579	0.8663	0.9490	0.8730	0.9577
4Kequi	4023	0.9143	0.9818	0.8826	0.9731	0.9380	0.9752	0.9165	0.9820
10Kequi	10,001	0.9705	0.9959	0.9596	0.9930	0.9768	0.9894	0.9766	0.9950
15Kequi	14,963	0.9849	0.9967	0.9704	0.9964	0.9704	0.9950	0.9828	0.9962

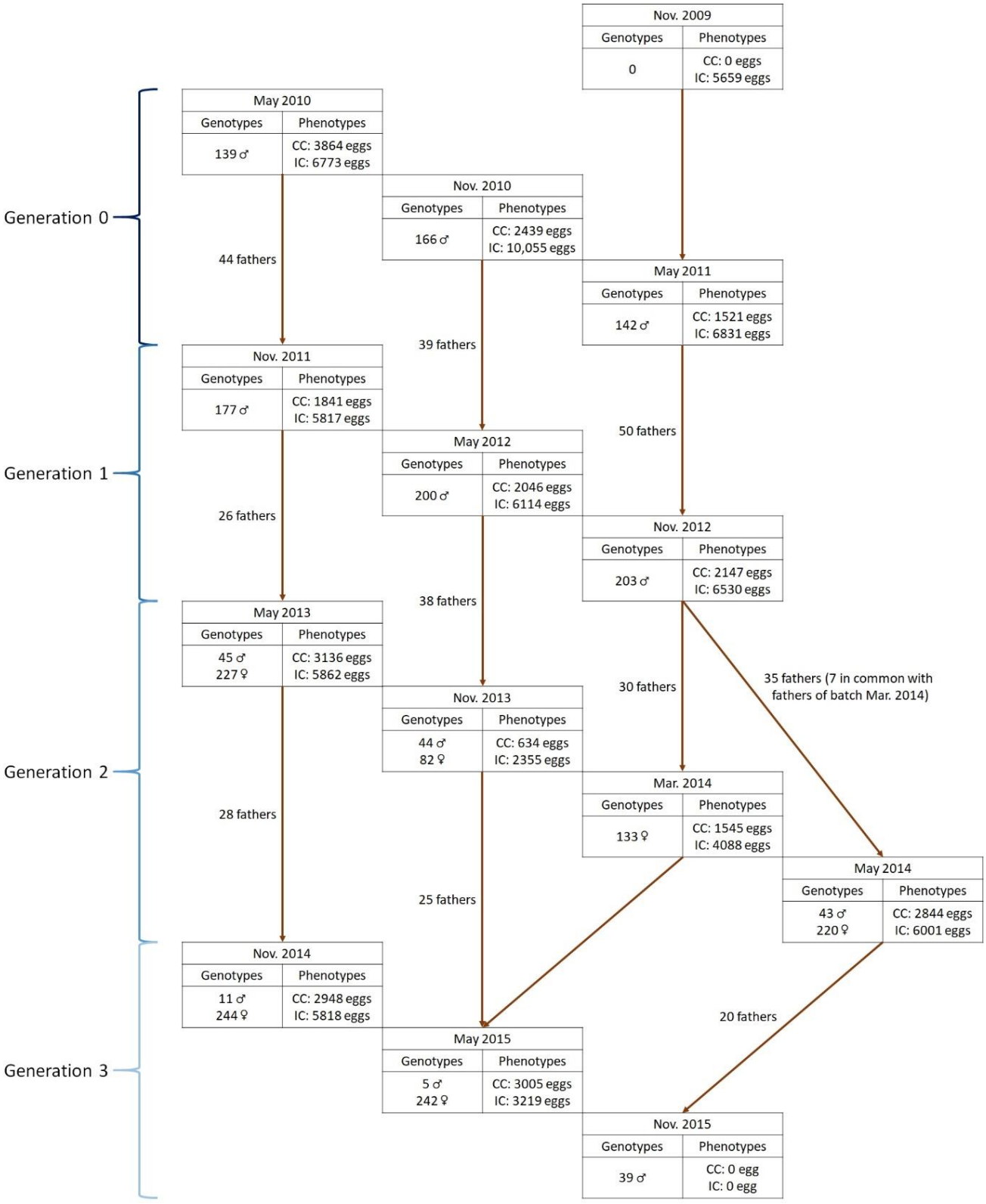


Figure 1. Population structure of the Rhode Island line.

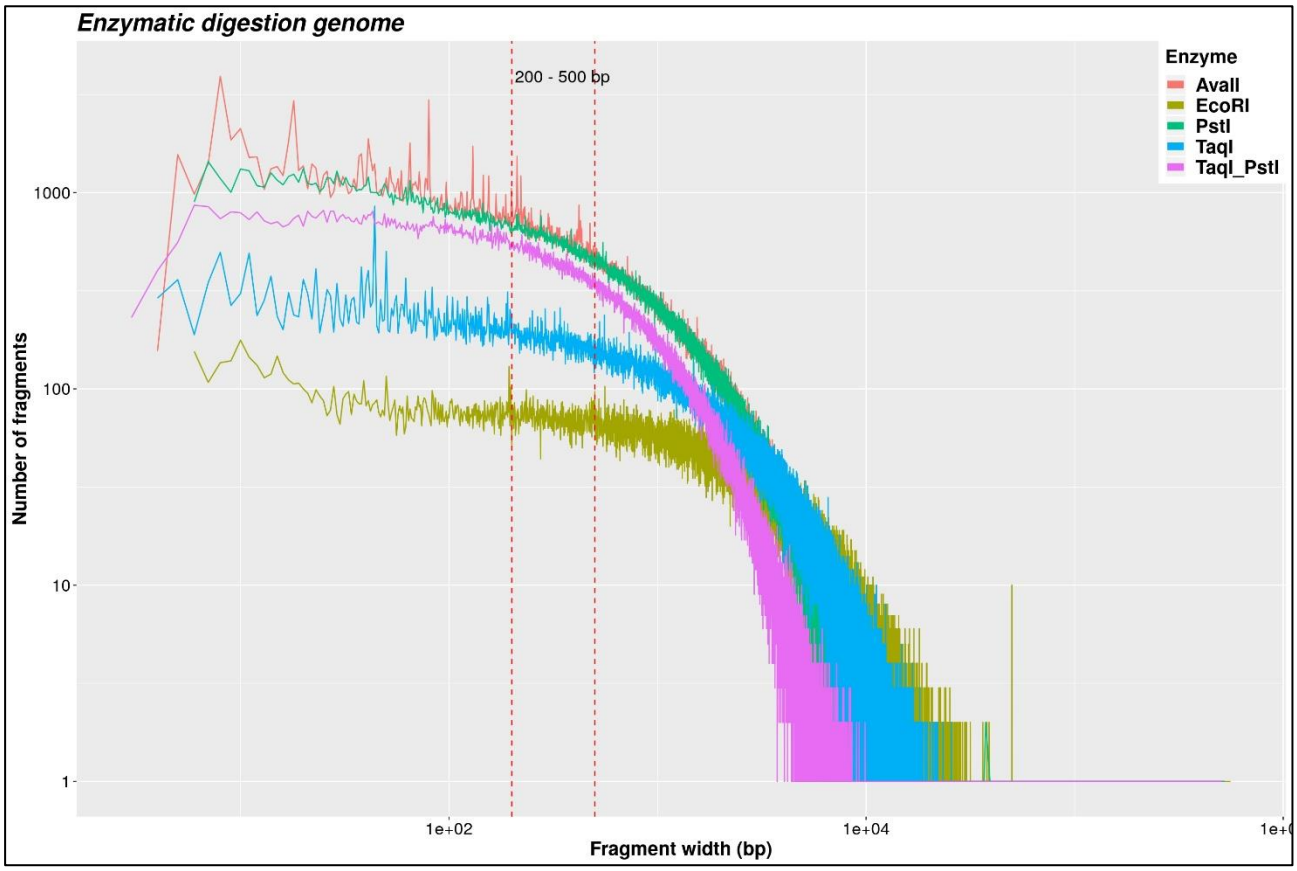


Figure 2. Enzymatic digestion pattern using AvaII, EcoRI, PstI, TaqI or the double association of TaqI and PstI.

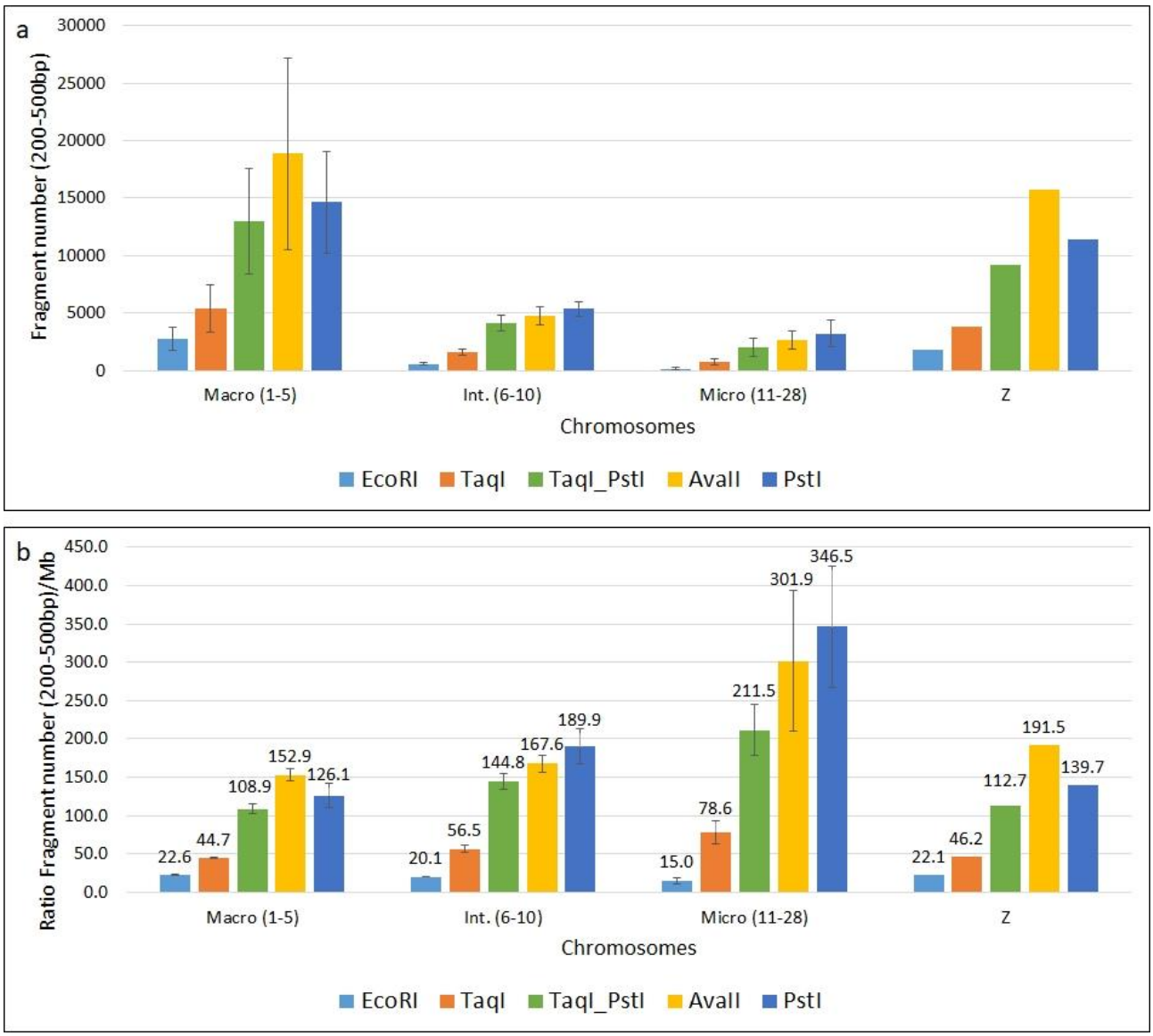


Figure 3. Distribution of the number of fragment of interesting sizes (200-500bp) (a) and number of fragment of interesting sizes (200-500bp) per Mb (b) obtained for each enzyme according to the type of chromosome.

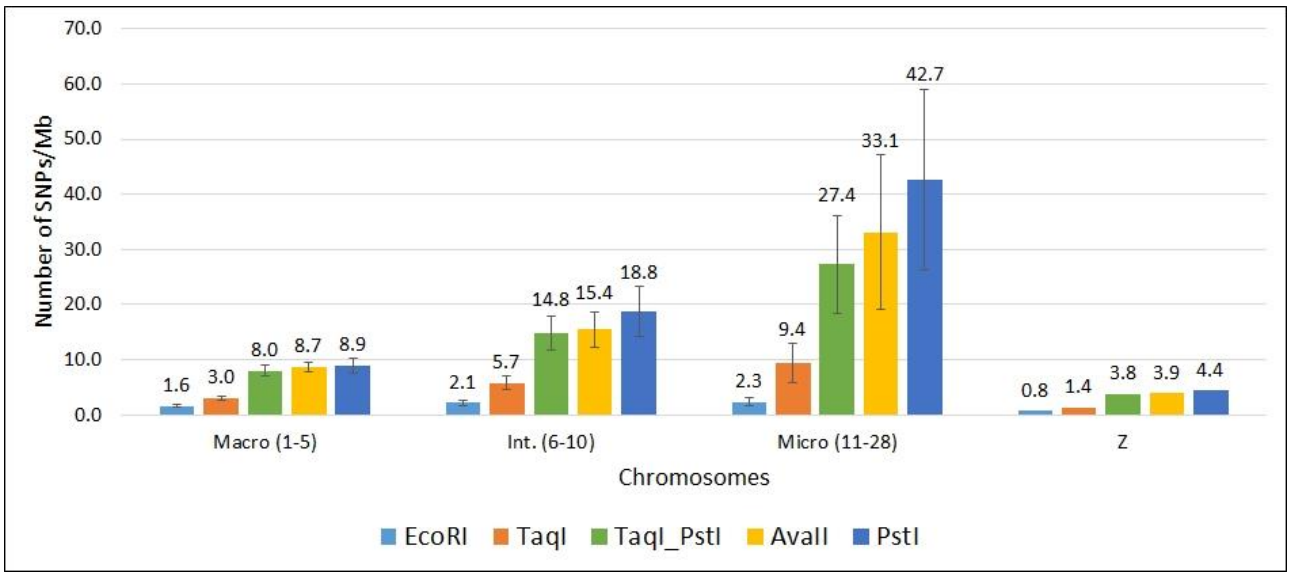


Figure 4. Evolution of the ratio number of SNPs/Mb depending on the type of chromosome, for each enzyme.

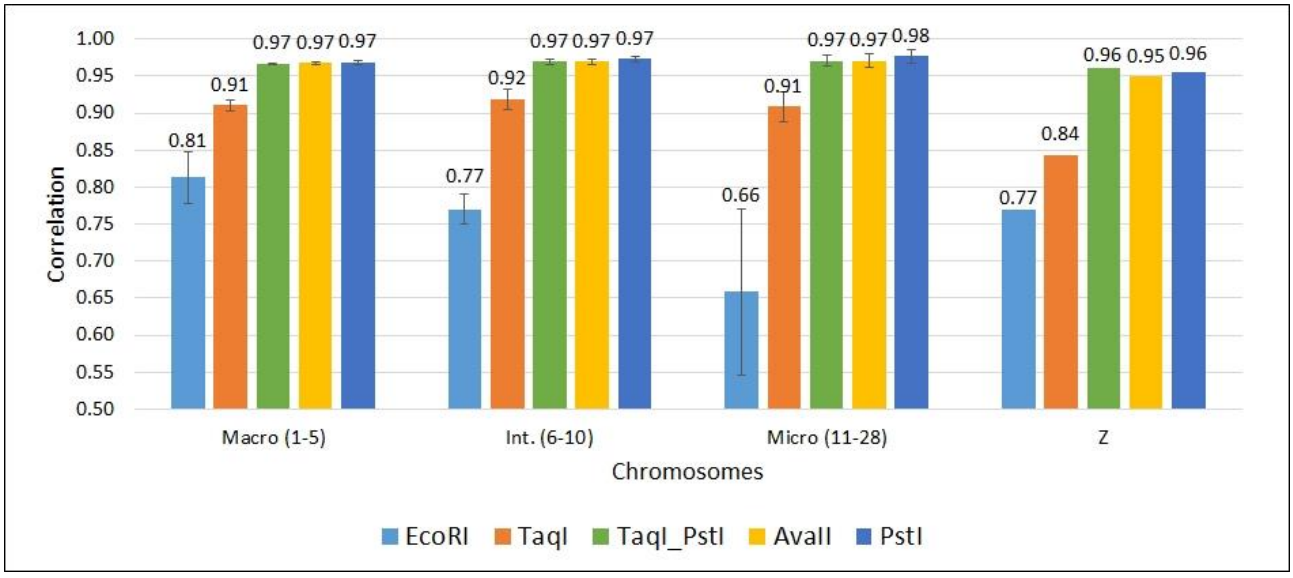


Figure 5. Evolution of the mean correlations between true and imputed genotypes according to the type of chromosome, for each enzyme.

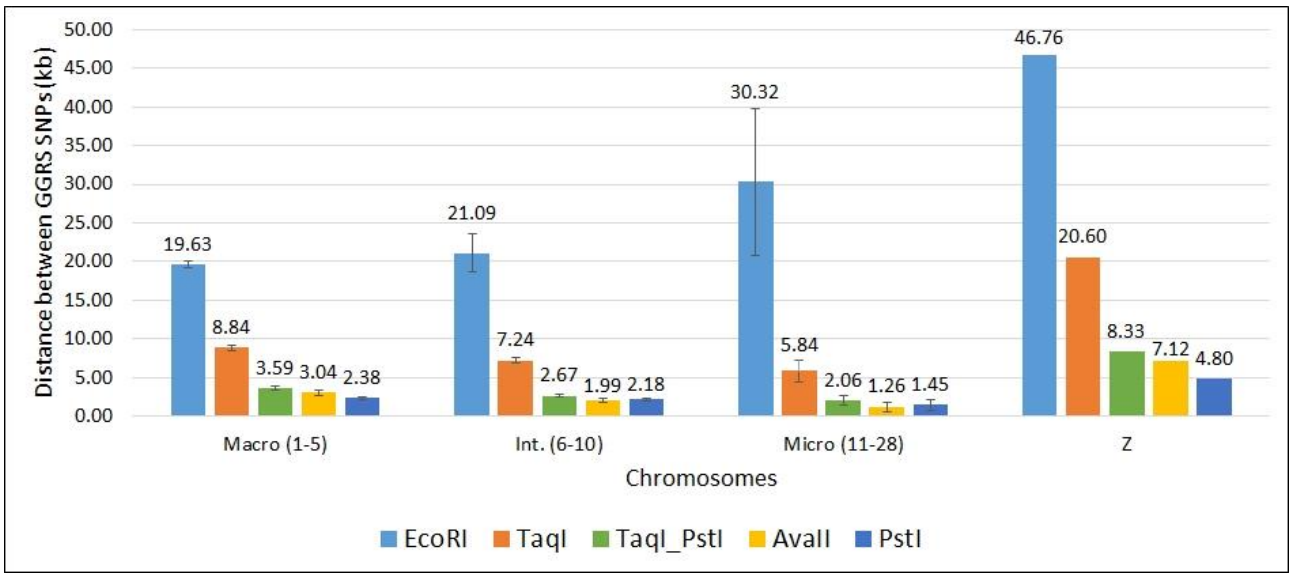


Figure 6. Distance between SNPs detected with GGRS and ddRADseq approaches (in kb) according to the type of chromosome, for each enzyme.

III. Discussion

Ces études ont permis de mettre en évidence que ces simulations de RAD-Seq pouvaient permettre d'obtenir des résultats d'imputations et d'évaluations génomiques tout aussi voire plus intéressants que les puces à SNP basse densité développées selon la méthodologie EQ. Les résultats des enzymes PstI et AvaII ainsi que ceux de la double digestion avec TaqI et PstI s'expliquent par la répartition des SNP sur les chromosomes. En effet, comparativement aux puces équidistantes de densité de SNP équivalente, les méthodes RAD-Seq entraînent une réduction du nombre de SNP sur les macro-chromosomes et une densification sur les micro-chromosomes. Cela n'est pas sans rappeler ce qui se passait pour la méthodologie DL !
Le tableau 13 récapitule les corrélations obtenues pour les imputations à partir des SNP des méthodes RAD-Seq avec AvaII, PstI et l'association TaqI et PstI et pour les imputations à partir des puces BD EQ et DL de densité équivalente.

Tableau 13. Récapitulatif des corrélations moyennes entre les vrais génotypes HD et les génotypes HD imputés à partir des méthodes RAD-Seq en utilisant les enzymes AvaII, PstI et la double association TaqI et PstI, ainsi que pour les puces EQ et DL de densité équivalente.

	TaqI_PstI	AvaII	PstI	10Kequi	15Kequi	DL0.5	DL0.6
Nombre de SNP	11 193	12 453	14 390	10 001	14 963	10 711	13 214
Corrélation moyenne	0.9691	0.9699	0.9735	0.9553	0.9670	0.9664	0.9717

Les résultats de RAD-Seq montrent que pour AvaII, PstI et le ddRADseq, des résultats d'imputations supérieurs sont obtenus par rapport à la puce 15Kequi, avec moins de SNP sur les puces BD, et équivalents aux résultats des puces DL. Pour rappel, avec la méthodologie DL, des résultats d'imputation légèrement supérieurs étaient obtenus par rapport aux résultats des puces EQ. En revanche, il n'y avait pas de différence significative en termes de reclassement des individus. C'est pourquoi il était intéressant d'étudier l'impact des imputations des génotypes issus des méthodes RAD-Seq sur le reclassement des individus selon leur GEBV estimés sur ascendance.
Les résultats concernant l'impact des imputations sur les évaluations génomiques pour les différentes enzymes et puces sont détaillés dans le tableau 14.

Tableau 14. Corrélation de Spearman entre les GEBV estimés sur ascendance avec les vrais génotypes HD ou avec les génotypes HD imputés à partir des méthodes RAD-Seq en utilisant les enzymes AvaII, PstI et la double association TaqI et PstI, ainsi que pour les puces EQ et DL de densité équivalente. Les résultats sont présentés pour les 150 meilleurs individus pour les différents caractères étudiés et les 67 individus reproducteurs en cage individuelle.

	Nombre de SNP	Poids d'œuf		Couleur de la coquille		Force de Fracture		Hauteur d'Albumen	
		Top150	Repros	Top150	Repros	Top150	Repros	Top150	Repros
TaqI_PstI	11 193	0.9913	0.9971	0.9779	0.9938	0.9904	0.9957	0.9859	0.9973
AvaII	12 453	0.9899	0.9975	0.9737	0.9949	0.9879	0.9951	0.9867	0.9958
PstI	14 390	0.9937	0.9980	0.9781	0.9959	0.9907	0.9943	0.9848	0.9949
10Kequi	10 001	0.9705	0.9959	0.9596	0.9930	0.9768	0.9894	0.9766	0.9950
15Kequi	14 963	0.9849	0.9967	0.9704	0.9964	0.9704	0.9950	0.9828	0.9962
DL0.5	10 711	0.9735	0.9957	0.9548	0.9930	0.9733	0.9889	0.9599	0.9921
DL0.7	13 214	0.9878	0.9978	0.9758	0.9953	0.9888	0.9938	0.9818	0.9957

Dans le cas des meilleurs individus pour un caractère ou des 67 individus reproducteurs, les résultats montrent que les différentes méthodes RAD-Seq permettent d'obtenir des corrélations qui sont supérieures mais non significativement différentes de celles obtenues avec les puces BD développées selon les méthodologies EQ et DL. Même si les résultats obtenus avec les méthodes RAD-Seq ne sont pas significativement meilleurs, ils apparaissent toutefois plus logiques que ceux obtenus avec la méthodologie DL. La répartition des SNP sur les différents types de chromosomes est un élément d'explication. En effet, les méthodes RAD-Seq permettent principalement de densifier encore plus le nombre de SNP sur les micro-chromosomes que ne le faisait la méthodologie DL. Par exemple, avec la puce DL0.5, il y avait 8.2 ± 0.8 SNP.Mb^{-1} et 23.7 ± 9.5 SNP.Mb^{-1} sur les macro-chromosomes et les micro-chromosomes respectivement. La double association de TaqI et PstI permet d'obtenir 8.0 ± 1.0 SNP.Mb^{-1} et 27.4 ± 8.9 SNP.Mb^{-1} sur les macro-chromosomes et les micro-chromosomes respectivement. AvaII et PstI permettent une densification encore plus forte du nombre de SNP sur les micro-chromosomes avec respectivement 33.1 ± 14.1 SNP.Mb^{-1} et 42.7 ± 16.3 SNP.Mb^{-1}.

Dans l'article II, plusieurs hypothèses identifiées dans la littérature ont été proposées pour expliquer les différences entre les deux méthodologies EQ et DL. Harris et Johnson (2010) et Weigel et al. (2010a) ont expliqué qu'une méthodologie équidistante est plus adaptée pour

obtenir de bonnes évaluations génomiques de caractères contrôlés par de nombreux petits QTL, ce qui est le cas ici. Par ailleurs, la méthodologie ssGBLUP suppose une même variance pour tous les SNP et favorise donc une méthodologie équidistante. Enfin, la méthodologie EQ serait plus robuste que la méthodologie DL en cas d'erreur d'imputation, certaines erreurs d'imputations pour des puces DL pouvant avoir plus d'importance que les erreurs réalisées avec les puces EQ.

Dans le cas des méthodes RAD-Seq, les SNP détectés et génotypés ne dépendent pas du DL entre SNP et se rapprochent plus d'une méthodologie EQ avec une densification en SNP différente en fonction du type de chromosome mais des zones dépourvues de SNP moins grandes qu'avec la méthodologie DL. Ceci peut donc expliquer les niveaux de corrélations obtenus avec les méthodes RAD-Seq.

IV. Bilan

L'utilisation des technologies RAD-Seq permettent de détecter et de génotyper simultanément un grand nombre de SNP. Toutefois, le nombre de SNP communs avec les puces HD est limité et variable selon les enzymes utilisées. Or, dans le cas d'une substitution des puces BD par du RAD-seq, ce sont ces SNP qui vont permettre de réaliser les imputations de la population candidate génotypée en RAD-Seq à partir de la population de référence génotypée avec la puce HD. Deux méthodologies de génotypage par réduction du génome et séquençage (GGRS) et de double digestion RAD-Seq (ddRAD-Seq) ont été simulées. En fonction de l'enzyme utilisée, il est possible d'identifier entre 4K et 14K SNP en commun avec les SNP de la puce HD avec TaqI et PstI respectivement. Ces mêmes enzymes permettent d'obtenir des imputations de bonnes qualités avec des corrélations moyennes supérieures à 0.91 et 0.97. La double digestion avec ces deux enzymes permet de génotyper 11K SNP en commun avec les SNP de la puce HD permettant d'obtenir une corrélation moyenne de 0.97. En revanche, l'utilisation de l'enzyme EcoRI ne permet que le génotypage de 1797 SNP en commun avec les SNP de la puce HD, entraînant une corrélation moyenne d'imputation de 0.79. Par ailleurs en comparant avec les résultats obtenus avec la puce 15Kequi, des corrélations supérieures sont obtenues avec AvaII, PstI et le ddRADseq avec en plus moins de SNP que sur la puce 15Kequi. En comparant avec les puces DL, des corrélations équivalentes sont obtenues.

En faisant le lien entre qualité d'imputation et impact sur les évaluations génomiques, les études ont montré que l'impact sur le reclassement des meilleurs individus pour un caractère était

faible avec des corrélations de Spearman supérieures à 0.97 pour chaque caractère étudié pour les enzymes AvaII, PstI et la double digestion TaqI et PstI. L'impact sur le classement des reproducteurs est encore plus réduit avec des corrélations moyennes supérieures à 0.99. Enfin, avec ces mêmes enzymes, il n'est pas noté de baisse significative de la précision des évaluations comparée à l'utilisation d'une puce HD pour réaliser les évaluations. Les résultats issus de RAD-Seq s'expliquent par une densification importante du nombre de SNP sur les micro-chromosomes. Cette densification est plus importante qu'avec la méthodologie DL. Toutefois, dans le principe, les coupures de l'ADN avec les enzymes n'étant pas basées sur le DL, les techniques RAD-Seq se rapprochent plus d'une méthodologie EQ, avec une densification des SNP différentes en fonction des chromosomes. Avec la méthodologie DL, de meilleurs résultats d'imputations ne se traduisaient pas par des résultats d'évaluations génomiques significativement meilleurs. Pour les techniques RAD-Seq, la distribution des SNP sur les chromosomes peut expliquer les niveaux de corrélations obtenus.

Il est toutefois important de noter que ce ne sont là que des résultats issus de simulations qui restent donc à être confirmés sur données réelles. En effet, ces simulations ne tiennent pas comptent de l'hétérogénéité entre individus. La sensibilité aux méthylations et le taux de polymorphismes dans les sites de restrictions sont deux facteurs qui introduisent de la variabilité entre individus. Enfin, la profondeur moyenne de lecture peut également influencer la qualité du génotypage. On peut toutefois espérer que l'imputation permette de combler de façon satisfaisante les données manquantes générées.

Par ailleurs, en supposant un génotypage de la population de référence et de la population candidate avec une méthode RAD-Seq et une bonne gestion de la variabilité entre individus, il peut être possible d'utiliser directement les génotypages RAD-Seq sans imputation. De cette façon, les SNP situés dans les régions chromosomiques sans génome de référence, et donc non présents sur les puces à SNP peuvent être inclus dans les évaluations génomiques. Des gains de précisions d'évaluations génomiques peuvent alors être espérés pour certains caractères contrôlés par des QTL situés dans ces régions chromosomiques mal connues.

Chapitre VI : Discussion générale et perspectives

I. Bilan des études

Les parties précédentes ont permis de discuter des questions concernant l'optimisation de l'utilisation des génotypages au niveau des candidats à la sélection. Nous avons vu qu'il est possible de génotyper les candidats à la sélection avec des puces hautes ou basses densités. Il est également possible de séquencer ces mêmes candidats en haute ou en basse profondeur avec les techniques de séquençage NGS ou de RAD-Seq.

Dans les chapitres II, III et IV, nous nous sommes intéressés au design de puces BD pour le génotypage des candidats à la sélection. En effet, la valeur marchande d'un individu est bien inférieure au coût de génotypage avec la puce HD (150€ par individu). Il n'est donc pas possible pour le sélectionneur de réaliser sur le long terme un génotypage de l'ensemble des candidats à la sélection avec une puce HD. L'utilisation des puces BD et de la méthode de l'imputation pour déduire les génotypages HD des candidats à partir d'une population de référence génotypée en HD s'est avérée être une technique intéressante d'un point de vue économique, une puce BD de 10K SNP coutant environ 30€ par individu. Concernant la méthode de design des puces BD, il est apparu que la méthodologie équidistante, plus simple que la méthodologie DL, est finalement la méthodologie la plus adéquate pour obtenir de bons résultats d'imputation et d'évaluation génomiques, dans la lignée de poules pondeuses étudiée.

Dans le chapitre V, nous nous sommes penchés sur les alternatives aux puces BD pour génotyper les candidats à la sélection, et notamment sur les technologies de RAD-Seq. Ces techniques, moins cher que le séquençage NGS basse profondeur, consistent à utiliser une ou des enzymes de restriction coupant l'ADN au niveau de sites de restriction puis à séquencer une partie des différents fragments obtenus. Parmi les différentes méthodes, nous nous sommes concentrés sur les méthodes GGRS et double RAD-Seq. Les résultats ont montré que le choix de l'enzyme est très important pour obtenir un nombre suffisant de SNP en communs avec les SNP de la puce HD ainsi que pour obtenir de bonnes qualités d'imputations. C'est ce qui a été noté avec les enzymes AvaII et PstI dans le cadre du GGRS, et avec TaqI et PstI dans le cadre du ddRAD-Seq, permettant de détecter plus de 10K SNP en commun avec les SNP de la puce HD. Enfin, l'impact sur les évaluations génomiques est également réduit avec des résultats qui ne sont pas significativement différents des résultats obtenus avec les puces équidistantes. Ces résultats restent toutefois à être confirmés sur données réelles.

À l'issue de ces différentes études, il apparait donc que les puces BD ou les technologies de RAD-Seq sont de bonnes alternatives aux puces HD pour génotyper à moindre coût l'ensemble des candidats à la sélection.

Enfin, il a également été noté la possibilité d'utiliser les puces BD directement sans imputation. Par extension, il est également possible d'utiliser les génotypages issus de RAD-Seq sans imputation. Mais cela implique toutefois que la population de référence soit génotypée avec une puce BD ou avec une méthode RAD-Seq. Il est aussi possible de séquencer en haute-profondeur la population de référence pour augmenter la précision des évaluations génomiques. Ainsi, l'ensemble des questions qui se posent concernant les stratégies de génotypages des candidats à la sélection se posent également pour la population de référence et les individus reproducteurs.

II. Optimisation des génotypages au niveau des reproducteurs

A. Choix de l'utilisation du génotypages BD, MD ou des séquences

1. Diminuer la densité de génotypage

Une première possibilité concernant la population de référence et les individus reproducteurs est d'obtenir les génotypages avec une puce BD. De cette façon, il est possible pour le sélectionneur de diminuer les coûts liés au génotypage de la population de référence.

Les études réalisées dans les chapitre III et IV ont justement permis d'étudier l'intérêt d'un génotypage BD au niveau la population de référence. En effet, comme expliqué précédemment, l'utilisation des puces BD sans imputation implique que la population de référence soit également génotypée en BD. Dans le chapitre III, il a été mis en évidence que la méthodologie EQ avec plus de 3K SNP non imputés permet d'obtenir de bons résultats d'évaluations génomiques avec des corrélations de Spearman supérieures à 0.94 pour les individus reproducteurs. En revanche, pour les meilleurs individus par caractère, l'utilisation des génotypages BD sans imputation peut avoir plus de conséquences sur leur reclassement. En effet, avec moins de 5K SNP, les corrélations de Spearman restent inférieures à 0.90.

Toutefois, le prix d'une puce BD est quasiment le même pour une densité de 10K ou de 50K SNP. Il est donc dans l'intérêt du sélectionneur de maximiser la densité de SNP étant donné les gains économiques réalisés sur le génotypage de la population de référence. Pour la puce 50Kequi, les corrélations de Spearman sont supérieures à 0.95 et 0.98 pour les meilleurs

individus par caractères et pour les reproducteurs respectivement. Enfin, il n'est pas noté de diminution significative de la précision des évaluations.

Dans le chapitre IV, la question de l'intérêt d'un génotypage BD ou MD au niveau la population de référence a été une nouvelle fois posée dans le cadre du développement de la puce multi-lignée. En supposant une puce 50K développée selon la stratégie « Multi », 42 159 SNP sont informatifs pour la lignée RI1. Pour cette puce, les corrélations de Spearman sont supérieures à 0.94 et 0.98 pour les meilleurs individus par caractères et pour les reproducteurs respectivement. Il n'y a également pas de diminution significative de la précision des évaluations.

Ces résultats sont en accord avec les résultats de la littérature présentés dans le chapitre I, section III.D. Les études de VanRaden et al. (2011) et Su et al. (2012) ont montré que le passage d'un génotypage HD à un génotypage MD pour l'ensemble des individus n'avait qu'un impact faible sur la précision des évaluations génomiques. En revanche, l'étude de Moghaddar et al. (2015) montre que le passage d'un génotypage MD à un génotypage BD peut avoir plus de conséquences sur la précision des évaluations génomiques.

Ainsi, l'utilisation des génotypages MD pour la population de référence et les individus reproducteurs peut être une solution intéressante pour le sélectionneur pour diminuer les coûts liés au génotypage de la population de référence. Il faut toutefois faire attention à ne pas trop diminuer la densité des SNP sous peine de dégrader la précision des évaluations génomiques.

2. Augmenter la densité de génotypage

Une deuxième possibilité concernant la population de référence et les individus reproducteurs est de génotyper la population de référence avec les méthodes RAD-Seq. De cette façon, il est possible de s'affranchir de la nécessité de recouvrement entre les SNP de la puce HD et les SNP issus de RAD-Seq. Les résultats de Poland et al. (2012) et Elbasyoni et al. (2018) montrent que pour des densités de SNP équivalentes aux densités de puces, il est possible d'obtenir une amélioration de la précision des évaluations pour certains caractères.

Par ailleurs, dans le chapitre V, nous avons exclu les fragments localisés dans les régions chromosomiques sans génome de référence. Ainsi, pour EcoRI et AvaII, nous avons retenu respectivement 21 357 et 183 012 fragments de 200 à 500bp sur les 25 988 et 243 474 répartis sur l'ensemble du génome. Nous avions ensuite vu qu'il était possible de détecter pour les enzymes EcoRI et AvaII respectivement 46 568 et 470 425 SNPs répartis sur les chromosomes 1 à 28, 33 et le chromosome sexuel Z. En n'utilisant plus de puce à SNP, il est donc possible d'inclure les SNP localisés dans les régions chromosomiques sans génome de référence. Des

gains de précisions d'évaluations génomiques peuvent alors être espérés pour certains caractères contrôlés par des QTL situés dans ces régions chromosomiques mal connues.

Avec le génotypage RAD-Seq, en fonction de la méthode et des enzymes utilisées, et de la gestion de la variabilité entre individus, il peut être possible d'utiliser un nombre plus conséquent de SNP que celui sur les puces à SNP. Les résultats retrouvés dans les publications de Liao et al. (2015) et Pértille et al. (2016) montrent que des coûts inférieurs à 50$ par individu peuvent être attendus. Ces coûts peuvent même être amenés à décroitre avec l'augmentation des capacités de séquençage. Les coûts de préparation des librairies représentent toutefois des coûts incompressibles qui, eux, sont moins sujet à diminuer dans le temps.

Une troisième possibilité concernant la population de référence et les individus reproducteurs est de séquencer en haute-profondeur les individus. Avec une séquence NGS haute-profondeur, le DL entre marqueurs et mutations causales n'est plus nécessaire puisque celles-ci sont inclues dans la séquence (Bolormaa et al., 2019). Il est donc possible d'améliorer la précision des évaluations de certains caractères. Bien que des séquences soient disponibles sur 90 individus de la génération G0 de la lignée RI1, l'intérêt de la séquence des individus reproducteurs pour les évaluations génomiques n'a pas encore pu être étudié. Toutefois, en pondeuse, Heidaritabar et al. (2016a) n'ont obtenu qu'un gain de précision d'environ 1% pour le poids d'œufs en passant de la puce 60K (Groenen et al., 2011) à des séquences imputées (plus de 3.9 millions de SNP) pour les individus de référence. De la même façon, Ni et al. (2017) ont montré que les gains de précisions obtenus à partir de séquences imputées (plus de 5.2 millions de SNP) étaient très faibles comparés aux précisions obtenues avec la puce HD 600K. Par ailleurs, comme expliqué dans le chapitre I, section III.B.1.b et section III.C.2.a, les coûts du séquençage NGS sont bien trop élevés avec notamment un coût incompressible des kits estimé à 100€ par individu. À ce jour, il faut encore compter environ 600€ pour un séquençage haute profondeur 20X en poule. Au regard de la précision des évaluations obtenues dans les deux études précédentes, l'intérêt de la séquence NGS haute-profondeur parait pour le moment très limité pour une utilisation en routine dans le schéma de sélection pour les individus de référence.

B. Le renouvellement de la population de référence

1. Pourquoi renouveler la population de référence ?

Pour des raisons économiques, il peut être intéressant de se poser la question du renouvellement de la population de référence. Par exemple, s'il est possible de modifier la population de référence seulement toutes les deux ou trois générations, il est alors possible de réaliser de grosses économies sur le génotypage des individus de référence.

Toutefois, au cours des chapitres précédents, il a été expliqué qu'il était important de renouveler la population de référence et d'éviter de réaliser une sélection génomique sur une population candidate trop éloignée de la population de référence (chapitre I, section I.B.2). Il a également été expliqué que le DL chute à cause des recombinaisons intervenant au fur et à mesure des générations. Ces recombinaisons pouvant casser le DL entre allèle d'un marqueur et allèle d'un QTL, le système d'équation développé sur la population de référence n'est alors plus adapté si on l'applique à une population candidate trop éloignée de la population de référence. Meuwissen et al. (2001) ont ainsi montré, sur des simulations en bovins, que la corrélation entre les GEBV et les TBV des générations candidates 1003 à 1008 estimés à partir des générations de référence 1001 et 1002 diminuait de 0.848 pour la génération 1003 à 0.804 pour la génération 1004 à 0.718 pour la génération 1008. De la même façon, Wolc et al. (2011) ont montré en pondeuses une diminution de la précision des évaluations génomiques au fur et à mesure que les générations candidates s'éloignaient des générations de référence (Figure 6). Les résultats montrent également que la diminution de la précision est aussi dépendante de la méthode utilisée pour réaliser les évaluations.

2. Les solutions pour limiter la chute de la précision

a) Construire la population de référence sur plusieurs générations

La question du renouvellement de la population de référence dépend fortement de la façon dont la population de référence a été construite. Une solution proposée par Muir (2007) est de cumuler dans la population de référence des individus de générations différentes. À taille de population de référence équivalente, les résultats montrent une diminution plus faible de la précision des évaluations en cumulant les quatre premières générations dans la population de référence pour évaluer les individus des générations suivantes plutôt qu'en cumulant les deux premières générations dans la population de référence (Figure 40). Ces résultats sont également retrouvés dans les travaux de Bastiaansen et al. (2012) qui montrent, à taille de population de référence équivalente, que la chute de précision des évaluations des candidats à la sélection éloignés de deux ou trois générations de la population de référence est moins forte en ayant une population de référence constituée de plusieurs générations plutôt que d'une seule génération. En revanche, pour un éloignement supérieur, la chute de précisions des évaluations est similaire dans les deux cas. En considérant une population de référence peu profonde constituée de tous les candidats à la sélection de la première génération, le système d'équation associant les génotypes de différents marqueurs aux phénotypes de différents individus peut être assez précis en utilisant l'information de DL entre marqueurs et la diversité haplotypique reflétant

l'information de pedigree. D'après les auteurs, cette information de pedigree contribue fortement à la précision obtenue pour les premières générations et contribue moins à mesure que la distance entre population de référence et candidate augmente. En revanche, en considérant une population de référence profonde constituée de 20% des candidats à la sélection des cinq premières générations, la distance génétique entre individus de référence et candidat augmente, ce qui réduit la chute de précision des évaluations sur les premières générations. C'est aussi ce qui a été expliqué dans le chapitre I, section III.C.1.b.

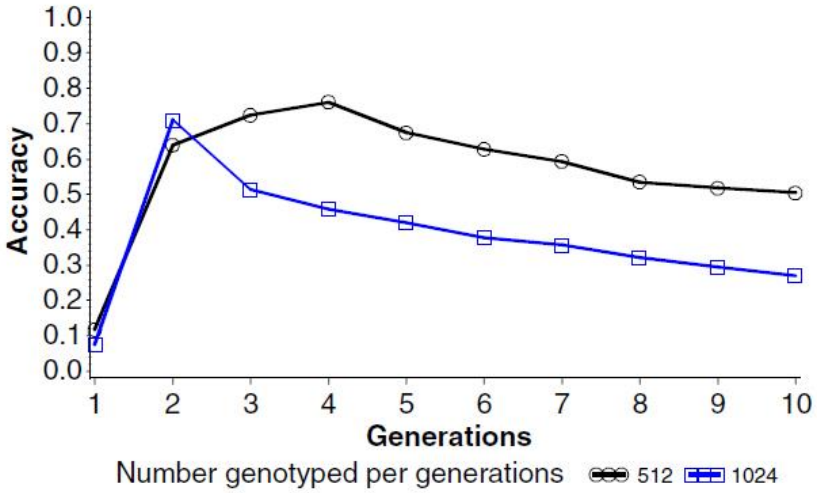


Figure 40. Effet du nombre d'individus génotypés par génération constituant la population de référence (512 individus pour les quatre premières générations ou 1024 individus pour les deux premières générations) sur la précision des évaluations des individus des différentes générations pour un caractère avec une héritabilité de 0.1. D'après Muir, 2007.

b) Influence de la méthodologie d'évaluation et de l'architecture génétique des caractères

Goddard (2009) explique que si la sélection génomique est menée sur plusieurs générations, ce n'est pas l'effet des marqueurs qui change mais la proportion de variance génétique expliquée par ces marqueurs qui diminue à cause du DL devenant trop faible entre allèles des marqueurs et allèles des QTL.

Pour faire face à la chute de précision des évaluations avec une diminution des relations entre populations de référence et candidate, Solberg et al. (2009) ont étudié l'effet de l'ajout d'une composante polygénique aux évaluations. Cette composante permet de capturer une partie de la variance génétique qui n'est plus capturée à cause du DL devenant trop faible entre allèle du marqueur et allèle du QTL. Les résultats montrent que l'inclusion de l'effet polygénique ne permet pas de diminuer la chute de la précision des évaluations par rapport à un cas sans ajout d'un effet polygénique. Toutefois, avec un effet polygénique, les évaluations sont moins

biaisées, et ce d'autant moins que le nombre de QTL considérés est faible. Cette étude montre également que plus le nombre de marqueurs est faible, plus la chute de la précision des évaluations est forte. En considérant un nombre faible de marqueurs, la distance entre marqueurs et QTL augmente tout comme la probabilité de recombinaison entre marqueurs et QTL. Cette recombinaison casse le DL et l'effet estimé du segment de chromosome est alors faux. C'est pourquoi si l'on ne ré-estime pas les effets des marqueurs au fur et à mesure des générations, il est nécessaire de disposer d'un nombre suffisant de marqueurs pour éviter une chute de DL trop importante entre générations.

L'étude de Bastiaansen et al. (2012) montre également que la réponse à la sélection génomique sur le long terme est dépendante de l'architecture génétique des différents caractères étudiés. Pour des caractères contrôlés par un faible nombre de QTL avec une variance différente, la précision de l'évaluation passe de 0.63 pour la première génération à 0.10 pour la 10ème génération. Pour des caractères contrôlés par un faible nombre de QTL avec une même variance, ou bien par un grand nombre de QTL avec une même variance ou non, la précision de l'évaluation passe de 0.63 pour la première génération à des valeurs comprises entre 0.12 et 0.16. L'étude de Muir (2007) se place dans le cas d'un petit nombre de QTL considérés avec une variance différente. Les auteurs justifient principalement la diminution de la précision par une diminution de la variance génétique expliqué par ces QTL et considèrent que les changements de niveaux de DL ne jouent ici qu'un rôle mineur. En revanche, dans les schémas de sélection actuels, la diminution de la variance génétique est faible et c'est bien la chute du DL qui explique la diminution de la précision des évaluations au fil des générations.

Par ailleurs, la méthodologie d'évaluation utilisée (GBLUP ou méthode Bayésienne) n'a qu'un impact très faible sur la précision des évaluations à la 10ème génération.

Ainsi, il est compliqué d'estimer à quel moment il est intéressant de renouveler la population de référence. Les différents travaux montrent que le choix est dépendant des espèces, du taux de recombinaison et de l'évolution du DL au fil des générations, et de l'architecture génétique des caractères évalués.

C. Optimiser le choix des individus de la population de référence

Dans la partie précédente, nous avons vu qu'il était important de renouveler régulièrement la population de référence au risque de déconnecter la population candidate de la population de référence avec la chute du DL au fil des générations. Pour les sélectionneurs, il n'est pas envisageable de génotyper en HD tous les candidats à la sélection pour que ceux-ci soient inclus dans la population de référence qui servira ensuite pour l'évaluation génomique des candidats

à la sélection de la génération suivante. L'optimisation du choix des individus à génotyper en HD et à inclure dans la population de référence est donc une question importante pour le sélectionneur.

D'après le chapitre I, section III.C.1.b, deux points sont à prendre en compte pour constituer une bonne population de référence. Le premier point concerne les relations de parenté entre individus de la population de référence. Ces relations doivent être les plus éloignées possible afin de maximiser la diversité génétique entre individus et ainsi de capter la diversité haplotypique de la population (Pszczola et al., 2012). Le deuxième point concerne les relations de parenté entre individus de la population de référence et de la population candidate. Ces relations doivent être cette fois-ci les plus fortes pour éviter qu'un allèle particulier soit présent dans la population candidate et absent dans la population de référence. Cela permet également de bénéficier du DL intra-famille qui correspond à un DL à longue distance et qui se traduit par de longs segments chromosomiques transmis d'une génération à une autre. Les études de Clark et al. (2012) et de Habier et al. (2013) suggèrent ainsi que le DL de la population de référence permet de définir une précision d'évaluation minimale pour tout candidat non apparenté à la population de référence. Le DL intra-famille permet ensuite d'améliorer la précision de l'évaluation pour l'individu. La pratique idéale serait donc d'inclure dans la population de référence les reproducteurs des candidats à la sélection qui permettent d'avoir une bonne représentation de la diversité haplotypique de la population sélectionnée et de maximiser les relations de parenté entre population de référence et population candidate.

III. Ré-estimation des effets des SNP et renouvellement des puces basse densité

La question de la ré-estimation des effets des SNP et du renouvellement des puces BD est très fortement liée à la question du renouvellement de la population de référence. Différentes puces BD ont été obtenues à partir de la puce HD 600K et des génotypages HD obtenus pour une population de référence. Si l'on souhaite modifier les SNP des puces BD pour les adapter à une nouvelle génération, il est nécessaire de disposer de génotypages HD pour cette nouvelle génération. Le renouvellement de la population de référence nécessite également de génotyper en HD les individus de la nouvelle génération. Il est donc clair que, pour des raisons économiques, le renouvellement de la population de référence et le renouvellement des puces BD doivent être réalisés en même temps.

D'après Habier et al. (2009), en fonction de la méthodologie utilisée pour développer les puces BD, il ne sera pas toujours nécessaire de renouveler les puces BD. Si les puces BD sont développées selon une méthodologie qui ne prend pas en compte le DL ou la fréquence allélique des SNP alors il n'y a pas besoin de renouveler les puces BD. C'est par exemple le cas d'une méthodologie équidistante avec sélection de SNP équidistants sans prise en compte de la MAF des SNP. Dans ce cas, il n'y a aucune raison de renouveler la puce puisque la position des SNP ne change pas au fur et à mesure des générations. En revanche, pour toutes les méthodologies de développement de puces BD utilisant la fréquence allélique des SNP, le DL entre SNP ou encore l'effet des SNP sur différents caractères d'intérêts, il sera nécessaire de renouveler la puce BD. En effet, la sélection modifie les fréquences alléliques de certains marqueurs pouvant aller jusqu'à fixer des zones entières sur certains chromosomes. En utilisant une méthodologie équidistante qui tient compte de la MAF des SNP, il est donc nécessaire de renouveler régulièrement la puce BD en ré-estimant les fréquences alléliques des SNP. De même, la chute de DL au fur et à mesure des générations implique que les SNP sélectionnés sur une génération selon la méthodologie DL ne seront plus forcément les mêmes pour les générations suivantes. Toutefois, à notre connaissance, il n'y a pas de publication qui traite du sujet du renouvellement des puces BD.

Dans les faits, les différentes entreprises qui développent des puces à SNP proposent de renouveler les puces BD tous les 2-3 ans voire tous les ans comme avec la puce EuroG10K bovine. La partie des SNP ne correspondant pas à la puce BD de 7K SNP peut en effet être réactualisée chaque année (Boichard et al., 2018).

IV. Intérêt du développement d'évaluation multi-lignées basée sur une puce multi-lignée

A. Intérêt de l'évaluation multi-lignée

Après avoir développé dans le chapitre IV une puce multi-lignée, il peut être légitime de se poser la question des évaluations multi-lignées. En effet, les SNP de la puce multi-lignée ont été choisis de façon équidistante en maximisant la MAF moyenne des SNP pour l'ensemble des lignées considérées. Pour une puce de 50K SNP équidistants selon la stratégie « Multi », il est possible de sélectionner 42 159 SNP et 34 395 SNP informatifs pour les lignées RI1 et L2 respectivement. Les évaluations génomiques des candidats de la lignée RI1 génotypés avec une telle puce montrent également que de bons résultats peuvent être obtenus.

Par ailleurs, pour chaque lignée, des individus supplémentaires aux reproducteurs ont été génotypés avec la puce HD de façon à constituer une population de référence suffisamment conséquente par lignée pour développer les puces BD. Or en routine, les générations candidates vont s'éloigner de cette population de référence initiale. Au fil des générations, les individus reproducteurs seront génotypés avec la puce HD et pourront être intégrés dans la population de référence. Les individus les plus anciens devront en revanche être enlevés de cette population (Weng et al., 2016). En conséquence, la taille de la population de référence sera réduite en routine. Pour faire face à cette réduction de taille, il peut donc être intéressant de mutualiser les génotypages et performances des individus des différentes lignées pour obtenir une seule et conséquente population de référence, puis de réaliser des évaluations multi-lignées.

B. Exemple d'application chez les espèces d'élevages

De nombreuses études, principalement chez l'espèce bovine, se sont penchées sur l'intérêt des évaluations multiraciales. Hayes et al. (2009) ont ainsi étudié l'intérêt d'une évaluation multiraciale pour les races Holstein et Jersey. Olson et al. (2012) se sont posés les mêmes questions pour les races Holstein, Jersey et Brown Suiss. Ces deux différentes études mettent en évidence qu'avec une population de référence multiraciale il est possible d'obtenir des précisions d'évaluations plus élevées pour les races à petits effectifs (Jersey, Brown Suiss) plutôt qu'en considérant seulement les individus de la race à petit effectif dans la population de référence. En revanche, l'intérêt d'une population de référence multiraciale pour des races à grands effectifs est plus limité. Ils montrent également que les résultats sont variables en fonction des caractères étudiés. Dans l'étude de Erbe et al. (2012), l'ajout de 2257 taureaux Holstein testés sur descendance dans une population de référence constituée de 540 taureaux Jersey entraîne même une diminution de la précision des évaluations des candidats Jersey pour certains caractères. Cette observation est aussi retrouvée dans l'étude de Hayes et al. (2009).
Les travaux de Hozé et al. (2014a) mettent également en évidence que le génotypage HD 777K SNP de la population de référence multiraciale permet d'obtenir une précision des évaluations plus élevée qu'avec des génotypages MD 50K SNPs. Ainsi, plus la population de référence multiraciale est génotypée pour un grand nombre de marqueurs, plus la précision des évaluations est élevée. Par ailleurs, les gains de précisions des évaluations sont seulement très faibles en passant d'une évaluation mono-race à une évaluation multiraciale pour une population de référence génotypée en MD. Les auteurs montrent aussi que si la précision des évaluations est déjà élevée avec une population de référence mono-race, l'ajout d'individus de

races différentes dans la population de référence ne permet pas d'augmenter significativement la précision des évaluations.

Toutefois, la suite des études de Hozé et al. (2014b) montrent, dans le cas de deux races bovines proches génétiquement (Montbéliarde et Simmental), qu'il est possible d'augmenter la précision des évaluations des individus de race Simmental en ajoutant dans la population de référence des individus génotypés en MD 50K de race Montbéliarde. Cette étude n'est valable que pour des caractères de production. Ces résultats sont cohérents avec ceux de Legarra et al. (2014) en race ovine. À partir de 6 populations ovines génotypées avec la puce MD 50K, le regroupement des populations de deux races proches (Manech Tête Rousse + Latxa Cara Negra Euskadi et Manech Tête Noire + Latxa Cara Negra Navarre) permet d'atteindre la même précision d'évaluation que lorsque tous les animaux des différentes races sont inclus dans la population de référence. Il semble donc plus intéressant de faire des évaluations multiraciales à partir de sous-groupes de races génétiquement proches plutôt qu'en combinant l'ensemble des races dans une seule population de référence. On pourrait ainsi imaginer de transposer le même type d'études aux différentes lignées de souche Rhode Island (au moins les lignées RI1 et RI2 d'après l'ACP figure 32) et de souche Leghorn.

Finalement, constituer une population de référence multiraciale peut présenter un intérêt pour les races à faibles effectifs en bénéficiant de l'apport des races de grands effectifs ou pour augmenter la précision des évaluations lorsque les individus sont tous génotypés en HD. Elles peuvent également présenter un intérêt dans le cas de races proches génétiquement.

C. Limites des évaluations multi-populations

Il existe toutefois de nombreuses limites au développement des évaluations multiraciales ou multi-lignées. La première question à se poser est celle de la conservation du DL entre races ou lignées. En effet, si le DL ne se conserve pas bien entre les races à évaluer, les associations entre allèles aux QTL et allèles aux marqueurs ne sont pas les mêmes pour les différentes races. Avec la puce MD bovine, la distance entre marqueurs adjacents est de 70kb. Pour cette distance, seuls 10% des marqueurs présentent un niveau de DL supérieur à 0.5 (Hozé, 2015). La distance entre QTL et marqueurs peut donc être assez élevée et si le DL ne se conserve pas suffisamment entre races, il est possible qu'une association détectée dans une race ne le soit pas dans une autre. L'utilisation de races proches génétiquement peut toutefois s'avérer une alternative assez intéressante pour réaliser des évaluations génomiques multiraciales. Une autre solution est d'utiliser une puce HD qui permet d'obtenir un niveau de DL plus fort entre marqueurs adjacents et qui permet donc une meilleure conservation du DL entre les différentes races. Cela

représente toutefois un coût supérieur de génotypage de la population de référence par rapport à l'utilisation d'une puce MD.

La deuxième question à se poser concerne la présence de QTL en commun dans les différentes races à évaluer (Hozé, 2015). En effet, si des QTL sont fixés dans une population A et pas dans une population B, ces QTL ne peuvent pas servir pour évaluer les QTL de la population B. Ces QTL ne peuvent servir qu'à l'évaluation intra-population de la population A (Tribout, 2011). L'étude de Guillaume (2009) montre justement qu'il n'y a qu'une concordance inférieure à 30% entre les QTL influençant les caractères laitiers détectés chez les trois grandes races laitières. De la même façon, l'étude de Allais (2011) illustre le fait qu'un nombre limité de QTL influençant la qualité de la viande sont en commun entre différentes races de bovins allaitants. Toutefois, Goddard et Hayes (2009) précisent que la puissance de détection des QTL dépend du DL entre les marqueurs et les QTL, de l'effet du QTL sur le caractère d'intérêt, de la fréquence des allèles aux QTL et du nombre d'animaux dans la population de référence. Ceci est également dépendant des objectifs de production fixés pour les différentes races ou lignées. Il est donc possible que pour un effet et un niveau de DL équivalent, un QTL soit détecté dans une race et pas dans une autre à cause d'une fréquence allélique ou d'une taille de population de référence trop faible.

La troisième question à se poser et découlant de la précédente est celle des effets des QTL entre races. En effet, si des QTL sont bien détectés en commun entre différentes races, il est possible que ces QTL n'aient pas les mêmes effets sur le caractère d'intérêt. Sans trop rentrer dans les détails, des stratégies de sélection génomique multiraciales ont déjà pu être développées. D'après Hozé (2015), elles peuvent être classées en deux grandes catégories. La première stratégie suppose que les QTL et leurs effets sont identiques entre races. On cherche alors à augmenter le niveau de DL entre marqueurs et QTL et à améliorer la conservation du DL entre races. Cela peut passer par l'utilisation de puces HD voire même des données de séquences permettant de supprimer la barrière de la conservation du DL entre races, la mutation causale étant cette fois présente dans le jeu de données. Cela n'est toutefois pas envisageable aujourd'hui d'un point de vue économique. La deuxième stratégie suppose qu'il existe des QTL spécifiques à chaque race. Des méthodes visant à identifier et distinguer les QTL communs entre races et les QTL spécifiques à chaque race ont ainsi pu être développées et reposent sur des méthodes Elastic Net ou Bayes Cπ (Hozé et al., 2014b). Druet et al. (2014) précisent que les données de séquences peuvent là aussi permettre d'améliorer la détection de QTL à faible fréquence allélique spécifique à chaque race.

V. Utilisation du génotype des femelles pour la sélection génomique

Au cours des parties et travaux précédents, nous nous sommes principalement penchés sur la sélection génomique sur voie mâle. L'apport du génotype des femelles n'a été que brièvement abordé dans le chapitre II avec l'intérêt du génotype des femelles (en tant que mères) pour les imputations des génotypes des mâles. De façon générale, le génotype des femelles peut contribuer à une meilleure évaluation génomique sur voie mâle. Mais il peut également être tout aussi intéressant d'étudier l'intérêt du génotype des femelles pour la sélection génomique sur voie femelle. Dans cette partie, nous distinguerons donc bien ces deux types d'intérêts.

A. Intérêt pour la sélection génomique des mâles

Un premier point noté dans le chapitre II avec le premier article est que la disponibilité des génotypages des femelles, combinée aux génotypages des mâles, peut permettre d'atteindre de très bonnes qualités d'imputations des candidats à la sélection génotypés avec une puce BD.

Par ailleurs, excepté chez les principales races bovines laitières, la taille de la population de référence peut être assez réduite pour les différentes espèces d'élevages et pour les différentes lignées sélectionnées par les sélectionneurs. Comme expliqué précédemment, la taille de la population de référence est un des facteurs limitant la précision des évaluations. Or, pour ces espèces ou lignées sélectionnées, de nombreuses femelles avec performances et génotypages peuvent être disponibles. L'inclusion du génotype des femelles dans les évaluations pourrait donc s'avérer une solution intéressante afin d'augmenter la précision des évaluations des candidats mâles.

De nombreuses études, principalement chez les bovins, se sont penchées sur l'intérêt de l'inclusion des femelles dans les populations de référence pour réaliser une sélection génomique sur les mâles. Ainsi en bovins Holstein, Pryce et al. (2012) ont montré que l'inclusion de plus de 10 000 femelles dans une population de référence d'environ 3000 mâles permet d'augmenter la précision des évaluations des jeunes candidats mâles entre 4 et 8% en fonction des caractères étudiés. Ces gains de précisions sont finalement assez faibles au regard de l'information apportées par ces nombreuses femelles. Ces résultats sont également retrouvés dans les travaux de Bapst et al. (2013) qui montrent qu'il n'y a pas de gain de précision en en ajoutant 1236 femelles à une population de référence de 4085 taureaux Brown Suiss. Il n'y a donc pas d'intérêt à ajouter les femelles dans la population de référence lorsque celle-ci est déjà suffisamment grande. Par ailleurs, l'étude de Carillier-Jacquin et al. (2013) chez les caprins précise les

résultats (Figure 41). Pour une population de référence constituée seulement de 67 mâles, la précision des évaluations de 148 jeunes boucs est comprise entre 0.09 et 0.12 pour l'angle des trayons et la teneur en matière grasse respectivement. En ajoutant à cette population de référence 1985 chèvres génotypées avec performances, la précision des évaluations des jeunes boucs est comprise entre 0.28 et 0.34 pour les deux mêmes caractères. En revanche, pour une population de référence constituée de 677 mâles et d'aucune femelle, la précision des évaluations est de 0.52 et 0.58 pour les deux mêmes caractères. L'ajout des 1985 chèvres dans la population de référence ne permet alors qu'un gain faible de précision des évaluations. Ceci illustre bien le fait que l'ajout des femelles dans la population de référence peut être intéressant pour améliorer la précision des évaluations lorsque la taille des populations de référence est faible.

Enfin, en porcin, les travaux de Lillehammer et al. (2011) montrent qu'avec 1800 mâles génotypés dans la population de référence, la précision des évaluations pour les candidats mâles est de 0.34. En rajoutant 1200 femelles génotypés, la précision des évaluations des candidats mâles est de 0.56 et 0.72. L'ajout des femelles permet donc une amélioration de la précision des évaluations génomiques des mâles.

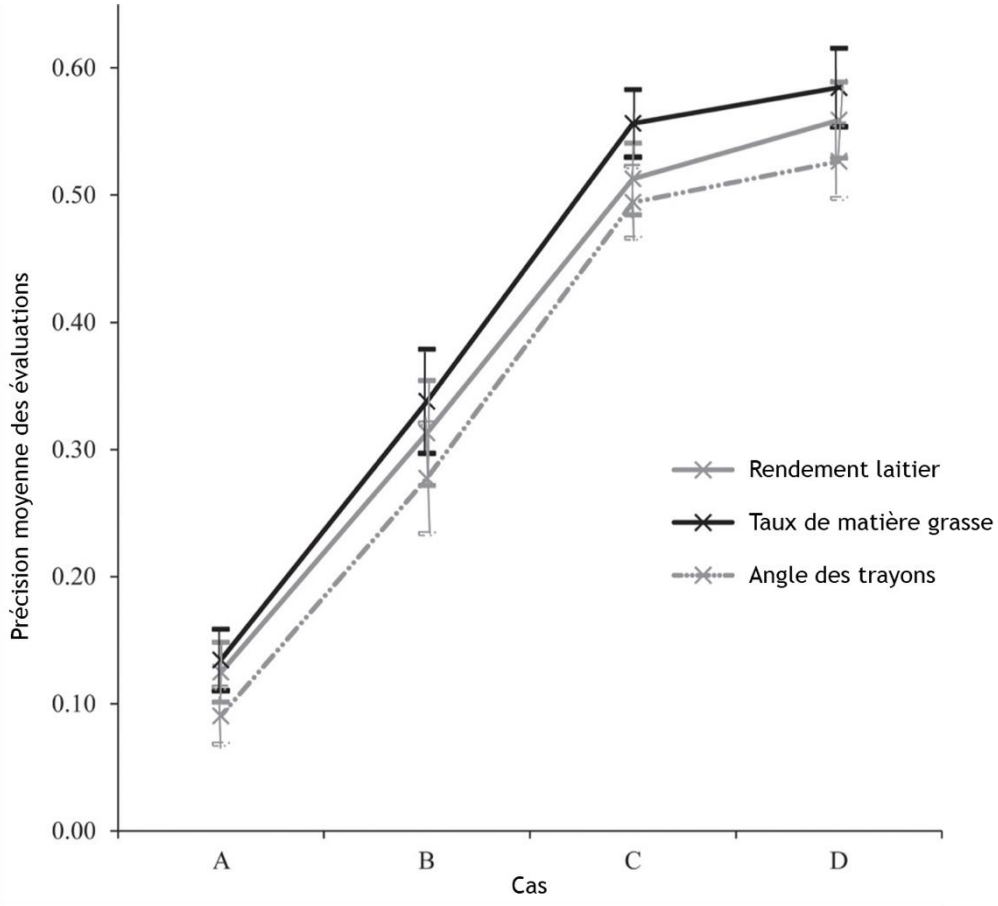


Figure 41. Précision moyenne des évaluations pour les 148 candidats mâles pour le rendement laitier, le taux de matière grasse et l'angle des trayons. Cas A : 67 mâles dans la population de

référence ; Cas B : 67 mâles et 1985 femelles dans la population de référence ; Cas C : 677 mâles dans la population de référence ; Cas D : 677 mâles et 1985 femelles dans la population de référence. D'après Carillier-Jacquin et al. (2013).

B. Intérêt pour la sélection génomique des femelles

En bovins, le premier intérêt du génotypage et de la sélection génomique des femelles était d'ouvrir le service de sélection aux éleveurs afin que ceux-ci puissent sélectionner plus précisément leurs femelles et puissent mieux gérer les accouplements (Boichard et al., 2012b). Ce service a été initialement proposé en 2011 en France (et 2009 aux États-Unis) avec la puce bovine 50K. Toutefois, les coûts de génotypages étaient encore trop élevés pour les éleveurs pour génotyper un grand nombre de femelles. Une solution a pu être mise en place à partir de 2012 grâce à la création de la puce bovine BD 7K. Couplé à l'imputation, il est alors possible pour les femelles génotypées en BD de déduire leur génotype MD. À titre d'exemple, les coûts de génotypages en BD sont aujourd'hui tels que le nombre de femelles génotypées et inclues dans les modèles d'évaluations génomiques des États-Unis augmente exponentiellement depuis 2009 (Figure 42). Ainsi pour les principales espèces bovines laitières, le nombre de femelles génotypées est passé de quelques milliers en janvier 2009 à plus d'un million en avril 2016.

Par ailleurs, quelques études, bien moins nombreuses, se sont penchées sur l'intérêt de l'utilisation du génotype des femelles sur la précision des évaluations génomiques des candidates femelles. Calus et al. (2013) ont ainsi montré qu'à partir d'une population de référence constituée de 296 taureaux Holstein, les précisions des évaluations des femelles sont de 0.173 et 0.156 pour les rendements en matière grasse et en protéine respectivement. En ajoutant à cette population de référence 1609 vaches génotypées avec performances, la précision des évaluations des femelles augmente à 0.273 et 0.221 pour les deux mêmes caractères. Ainsi, l'ajout des femelles génotypées avec performances peut être intéressant pour améliorer la précision des évaluations des femelles.

De même, en porcin, les travaux de Lillehammer et al. (2011) ont montré qu'avec 1800 mâles génotypés dans la population de référence, la précision des évaluations pour les candidates femelles est de 0.56. En rajoutant 1200 femelles génotypés, la précision des évaluations des candidates femelles est de 0.72. L'ajout des femelles permet donc également une amélioration de la précision des évaluations génomiques des femelles.

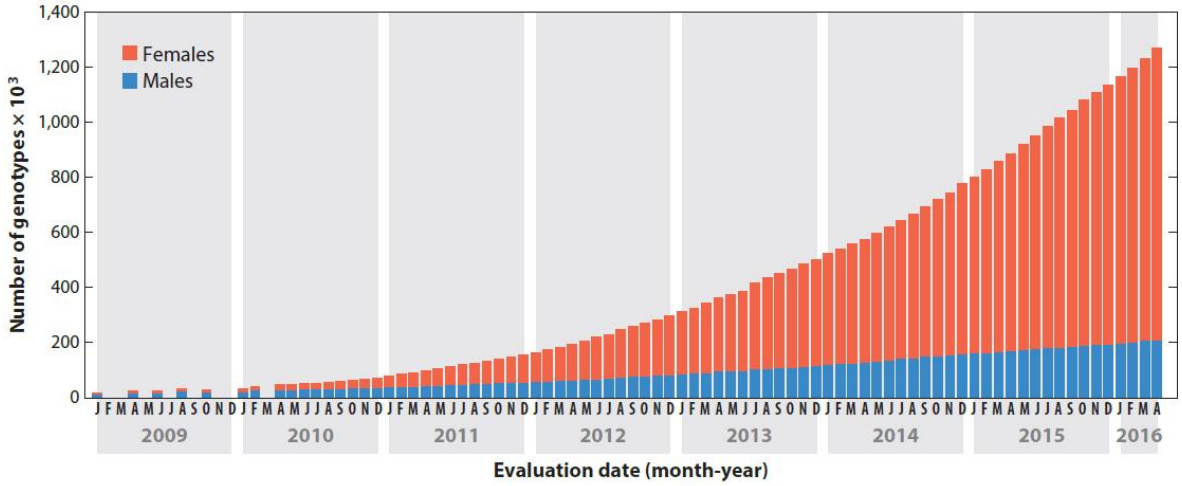


Figure 42. Nombre de bovins laitiers génotypés inclus dans les évaluations génomiques des États-Unis depuis Janvier 2009. D'après Wiggans et al. (2017).

C. Possibilité de prise en compte des effets de dominance dans les modèles d'évaluations

Avec l'ajout des femelles génotypées avec performances propres dans la population de référence se pose la question de la prise en compte des effets non-additifs de la variance génétique. En effet, des familles complètes avec des individus génotypées et avec performances sont disponibles. Il est donc possible de prendre en compte les effets de dominance (non additifs) dans les modèles d'évaluations génomiques. Ceux-ci ne prenaient jusque-là en compte que les effets additifs qui correspondent aux effets de substitution des allèles des marqueurs. La dominance correspond à l'interaction entre allèles d'un même locus et se mesure comme l'écart entre le phénotype hétérozygote et la moyenne des phénotypes homozygotes opposés. L'héritabilité faible des caractères mesurées sur une population de femelles peut s'expliquer par le nombre plus faible de mesure par individus, mais également par la présence d'effets non additifs pouvant introduire d'importantes variations sur ces caractères. La prise en compte des effets additifs et des effets de dominance dans les modèles d'évaluations génomiques peuvent théoriquement permettre d'atteindre des niveaux de précisions plus élevées qu'en prenant seulement en compte les effets additifs. Différentes études se sont intéressées à la prise en compte des effets de dominance dans les modèles d'évaluations pour essayer d'augmenter la précision des évaluations génomiques. Sun et al. (2014) montrent ainsi que pour deux races bovines Holstein et Jersey, l'inclusion d'un effet de dominance dans le modèle d'évaluation génomique ne permet qu'une augmentation de 4 à 5% de la précision des évaluations pour des caractères de rendement (lait, protéine et matière grasse) par rapport à un modèle sans effet de dominance. En revanche, pour les autres caractères testés (durée de production de la vache

avant réforme, taux de cellules, taux de réussite à l'insémination), il n'y a pas d'amélioration significative de la précision des évaluations génomiques. Ces résultats contrastent quelque peu avec ceux d'Aliloo et al (2016) qui ne retrouvent une amélioration de la précision des évaluations en incluant un effet de dominance pour seulement le rendement en matière grasse. Enfin pour des bovins de races Fleckvieh, Ertl et al. (2014) ne notent pas d'amélioration significative de la précision des évaluations pour 9 caractères de productions de lait et de conformation en incluant un effet de dominance dans le modèle d'évaluation. C'est également le cas en porcins avec l'étude de Nishio et Satoh (2014) pour deux caractères et en pondeuses avec l'étude de Heidaritabar et al. (2016b) pour 8 caractères de production et de qualité des œufs. Ces résultats quelque peu décevants peuvent s'expliquer par la complexité de l'estimation précise des effets de dominance, par un nombre d'individus restreints constituant des familles complètes et par l'utilisation d'une densité faible ou moyenne de génotypage. Avec des densités supérieures, il est possible d'obtenir un DL supérieur entre allèles aux marqueurs et allèles aux QTL, permettant de mieux estimer les effets de dominance (Wellmann et Bennewitz, 2012).
Les effets de dominance sont toutefois attendus plus forts pour des individus croisés. Il est en effet plus probable que la variance génétique soit plus fortement affectée par des effets de dominance car les fréquences alléliques sont supposées, en moyenne, être proches d'un niveau intermédiaire. Les marqueurs seront alors plus souvent hétérozygotes. Varona et Misztal (1999) supposent alors qu'il est plus probable d'augmenter la précision des évaluations en tenant compte des effets de dominance pour des individus croisés. Toutefois, les publications de Xiang et al. (2016) en porcin et de Moghaddar et van der Werf (2017) en ovin montrent qu'il n'y a pas d'amélioration significative de la précision des évaluations génomiques en prenant en compte un effet de dominance dans les modèles d'évaluations. Ceci s'explique par l'estimation de la variance génétique additive qui est similaire à celle estimée avec un modèle simplement additif. Dans ces cas, la variance génétique additive est donc déjà bien capturée avec un modèle additif.

D. Limites de l'apport du génotype des femelles pour la sélection génomique des mâles

Une première limite concerne la moindre informativité des femelles par rapport aux mâles. Boichard et al. (2015) et de Roos (2011) ont montré qu'atteindre de hauts niveaux de précisions d'évaluations génomiques avec une population de référence composée d'individus avec performances propres nécessite un nombre conséquent d'animaux génotypés, plus particulièrement pour les caractères à faible héritabilité. La solution mise en place chez les

bovins est de génotyper en BD les femelles puis d'imputer les données manquantes pour déduire leurs génotypes MD.

Par ailleurs, chez les bovins, de nombreux articles montrent que l'inclusion des femelles dans les populations de référence peut entraîner un risque de biais dans les évaluations génomiques à cause d'un problème de traitement préférentiel des femelles. Ce traitement préférentiel se définit comme des pratiques d'élevages qui modifient la production laitière et biaisent les valeurs génétiques des vaches concernées (Kuhn et al., 1994). Par exemple, il est possible qu'un éleveur favorise certaines vaches au détriment d'autres vaches en leur donnant un peu plus d'aliment ou en leur permettant l'accès à un meilleur logement (Hozé, 2015). Ceci peut donc améliorer les performances des individus favorisés et induire un biais dans l'estimation des effets des marqueurs en utilisant ces performances. Dassonneville et al. (2012) montrent ainsi que le biais lié au traitement préférentiel des individus peut contrebalancer l'amélioration de la précision des évaluations génomiques normalement attendue en augmentant la taille de la population de référence avec l'ajout des femelles génotypées. Toutefois, en incluant des femelles choisis aléatoirement dans une population commerciale, donc théoriquement sans traitement préférentiel, ce biais peut disparaitre. Il est quand même possible de corriger les phénotypes des femelles pour limiter l'impact du traitement préférentiel sur les évaluations génomiques (Wiggans et al., 2011). Ce problème de traitement préférentiel est toutefois nettement moins probable en volaille.

Enfin, Calus et al. (2013) montrent également que l'inclusion des femelles génotypées avec des performances propres dans la population de référence peut conduire à un autre biais dans l'évaluations génomiques des candidats. En effet, il y a un risque de prendre en compte deux fois les phénotypes des femelles avec leurs performances propres et avec les performances assignées aux mâles correspondant à la moyenne des performances des filles. Si l'on souhaite inclure les femelles génotypées avec performances dans la population de référence, une solution optimale serait de calculer la performance moyenne du mâle en ne prenant en compte que les performances des filles non génotypées.

E. Perspectives en filière ponte

En poule pondeuse, il peut être envisageable de réaliser une sélection génomique sur les femelles. Toutefois, Picard-Druet et al. (2019) ont comparé les précisions atteintes dans le cas d'une évaluation génétique sur collatérale des femelles (ce qui est réalisé actuellement) et dans le cas d'une évaluation génomique sur ascendance. Ils ont montré que la précision atteinte dans le cas de l'évaluation génétique est supérieure à 0.90 pour les 5 caractères étudiés. En revanche,

dans le cas de l'évaluation génomique sur ascendance, la précision des évaluations pour les mêmes caractères est comprise entre 0.44 et 0.6. Par ailleurs chez Novogen, les femelles sont actuellement élevées en cage multiple puis en cage individuelle, pour un intervalle de génération de 90 semaines. La période d'élevage en cage multiple peut être réduite mais reste nécessaire pour sélectionner les individus sur le comportement. La réduction de l'intervalle de génération est donc limitée. Ainsi, les gains de progrès génétiques théoriques ne sont pas aussi élevés que les gains qui peuvent être obtenus avec les mâles.

Conclusion générale

Le développement d'une puce à SNP commerciale HD de 600 000 marqueurs en 2013 (Kranis et al., 2013) a permis le développement de la sélection génomique dans les filières poules de chair et ponte. En parallèle, le développement des nouvelles techniques de séquençage (Next Generation Sequencing) permet dès à présent d'envisager des solutions autres que les puces à SNP pour réaliser cette sélection génomique. Toutefois, la sélection génomique coûte encore cher, surtout pour une espèce comme la poule où la valeur marchande du reproducteur est très faible. En effet, les coûts de génotypages avec la puce HD restent élevés (150€). Les prix sont même supérieurs concernant le séquençage haute-profondeur. Ces techniques ne sont donc pas utilisables en routine par les sélectionneurs pour un grand nombre de candidats à la sélection. Pour faire face à ce problème, un des enjeux de la sélection génomique est de développer des outils de génotypage ou de séquençage des candidats à la sélection à moindre coût tout en optimisant la précision des évaluations génomiques.

Une première solution concerne le développement de puces BD coûtant moins cher (environ 30€ pour une puce de 10K SNP) et l'utilisation de la technique de l'imputation pour déduire, pour l'ensemble des individus génotypés en BD leur génotype HD à partir d'une population de référence elle-même génotypée en HD. Les travaux menés au cours de cette thèse nous ont permis de tester différentes méthodologies de construction de puces BD avec une méthodologie de sélection des SNP équidistante et une méthodologie basée sur le déséquilibre de liaison. La question de la constitution de la population de référence a également été investiguée. Dans un second temps, l'impact des imputations ou de l'absence d'imputation sur les évaluations génomiques des candidats à la sélection a été étudié. À la lumière des résultats, nous avons vu que la méthodologie équidistante, pour une densité supérieure à 5K SNP, est une bonne méthodologie pour obtenir à la fois de bons résultats d'imputation et une bonne précision des évaluations génomiques. De plus, sans imputation avec cette même méthodologie, les évaluations génomiques sont capables de bien distinguer les bons individus des moins bons (les reproducteurs). Mais lorsqu'il s'agit de classer plus finement des individus qui sont proches génétiquement (les meilleurs individus par caractère), l'absence d'imputation peut s'avérer assez préjudiciable. L'impact de la constitution de la population de référence n'a en revanche pas pu être étudié, même s'il est apparu clairement que la présence des deux parents directs (ou au moins les pères) des candidats à la sélection dans la population de référence est importante pour obtenir de bonnes imputations. Les résultats que nous avons retrouvés dans la littérature

abondent également en ce sens concernant l'impact de la constitution de la population de référence sur les évaluations génomiques des candidats.

Suite aux résultats obtenus, des analyses de déséquilibre de liaison et de diversité génétique des différentes lignées Novogen ont été réalisées afin de développer une puce BD multi-lignée. L'analyse des relations et de la différenciation entre population nous a amené à considérer principalement les différentes lignées séparément pour construire des puces BD équidistantes selon une stratégie « Indep » ou « Multi ». Cette dernière stratégie permet d'optimiser le nombre de SNP sélectionnés et informatifs pour l'ensemble des lignées et semble être la stratégie la plus intéressante pour obtenir de bons résultats d'imputations. La stratégie « Multi » permet également d'obtenir des résultats d'évaluations génomiques, avec ou sans imputation, qui tendent à être meilleurs que ceux obtenus avec la stratégie « Indep » sans toutefois obtenir des différences significatives. Il est à noter que l'utilisation de puces à faible densité (10K) multi-lignée sans imputation pourrait avoir des impacts sur la précision des évaluations de certains caractères.

La deuxième opportunité pour diminuer les coûts de génotypage concerne l'utilisation du séquençage basse profondeur des candidats à la sélection et l'identification des SNP en communs avec les SNP de la puce HD. Ces SNP servent ensuite à imputer les candidats à la sélection à partir d'une population de référence génotypée en HD. Pour cela, nous avons étudié des techniques de RAD-Seq consistant à utiliser une ou des enzymes de restriction coupant l'ADN au niveau de sites de restriction puis à séquencer une partie des différents fragments obtenus. Deux différentes méthodes de RAD-Seq, à savoir le génotypage par réduction du génome et séquençage (GGRS) et la double digestion RAD-Seq (ddRAD-Seq) ont été simulées. Les résultats montrent que le choix de l'enzyme est très important pour obtenir un nombre suffisant de SNP en commun avec les SNP de la puce HD ainsi que pour obtenir de bonnes qualités d'imputations. Ce sont d'ailleurs les enzymes AvaII, PstI et la double utilisation de TaqI et PstI qui permettent d'identifier plus de 10K SNP en commun avec les SNP de la puce HD et d'obtenir une bonne qualité d'imputation avec des corrélations supérieures à 0.97. Ces enzymes permettent une densification importante du nombre de SNP sur les micro-chromosomes. Cette densification est plus importante qu'avec la méthodologie DL. Toutefois, dans le principe, les coupures de l'ADN avec les enzymes n'étant pas basées sur le DL, les techniques RAD-Seq se rapprochent plus d'une méthodologie équidistante avec une densification différente des SNP en fonction des chromosomes. En faisant le lien entre qualité d'imputation et impact sur les évaluations génomiques, les études ont montré que l'impact sur le reclassement des différents individus (reproducteurs et meilleurs individus par caractère) est faible avec des corrélations de Spearman supérieures à 0.97. Il n'y a également pas de baisse

significative de la précision des évaluations génomiques par rapport à une évaluation génomique réalisée avec des génotypes HD pour l'ensemble des individus.

Par ailleurs, toutes les questions qui se sont posées au niveau des candidats à la sélection peuvent également se poser au niveau de la population de référence. Il est ainsi possible d'optimiser l'utilisation des génotypages de ces individus en choisissant de les génotyper en MD et de ne plus utiliser les méthodes d'imputation pour la population candidate. Cela permet aussi aux sélectionneurs de réaliser des économies en ne génotypant plus en HD les individus de référence. Une autre possibilité est de génotyper avec des méthodes RAD-Seq les individus de référence. Ceci pourrait permettre, en fonction de l'enzyme sélectionnée, d'augmenter la densité de génotypage et de travailler avec les SNP détectés sur les chromosomes sans génome de référence. En revanche, l'utilisation du séquençage NGS haute-profondeur est plus discutable par rapport aux gains de précisions obtenus et aux coûts à engager. Par ailleurs, il est important de se poser la question du renouvellement de la population de référence au fil des générations de façon à ne pas dégrader la précision des évaluations génomiques. Enfin, l'optimisation du choix des individus à considérer dans la population de référence est également à prendre en compte, en essayant de minimiser les relations entre individus de la population de référence et de maximiser les relations de parenté entre individus des populations de référence et candidate.

Il est également important de se poser la question du renouvellement des puces BD dont la construction dépend des fréquences alléliques ou du DL entre marqueurs, deux paramètres qui diffèrent en fonction des générations. Par ailleurs, une puce multi-lignée ayant été développée, il peut être intéressant de réaliser des évaluations multi-lignées. Les résultats issus de la littérature montrent que la constitution d'une population de référence multiraciale peut avoir son intérêt pour les races à faibles effectifs en bénéficiant de l'apport des races de grands effectifs ou pour augmenter la précision des évaluations lorsque les individus sont tous génotypés en HD. Elles peuvent également présenter un intérêt dans le cas de races proches génétiquement. Il existe néanmoins de nombreuses limites du fait d'une possible conservation différente du DL entre races ou lignées, d'un nombre potentiellement limité de QTL en communs entre les différentes races ou lignées et des effets des QTL différents en fonction des races ou des lignées.

Enfin, la question de l'utilisation du génotype des femelles dans la sélection génomique est également une question importante. Il est possible d'utiliser le génotype des femelles pour la sélection génomique sur voie mâle. Le nombre de femelles étant généralement supérieur au nombre de mâles reproducteurs, il est possible d'utiliser ces femelles pour augmenter la taille de la population de référence. En fonction des espèces et des études, il est alors possible

d’améliorer la précision des évaluations génomiques des candidats mâles, à condition que la taille de la population de référence ne soit pas déjà trop élevée. Par ailleurs, il est également possible d’utiliser le génotype des femelles pour la sélection génomique sur voie femelle. En bovins laitiers, l’accès au génotypage des femelles a permis d’ouvrir le service de sélection aux éleveurs. Ceci a été rendu possible grâce à la disponibilité d’une puce BD et à l’utilisation de l’imputation pour déduire les génotypes MD des femelles afin de leur permettre de sélectionner plus précisément leurs femelles et de mieux gérer les plans d’accouplements. Il est également possible d’améliorer la précision des évaluations des femelles. La disponibilité des femelles génotypées avec performances permet également de prendre en compte les effets non-additifs dans les modèles d’évaluations génomiques. Enfin, en poules pondeuses, par rapport aux précisions obtenues avec une évaluation génétique classique, les premières évaluations génomiques sur ascendance des femelles ne montrent qu’un intérêt limité pour le sélectionneur. La réduction de l’intervalle de génération est également pour le moment limitée avec la nécessité d’élever les poules en cage multiple puis en cage individuelle selon le schéma Novogen. Il en résulte des gains de progrès génétiques théoriques qui ne sont pas aussi élevés que les gains obtenus pour les mâles avec la sélection génomique.

Références

Affymetrix. Axiom® Genome-Wide Chicken Genotyping Array. 2013. Consulté le 31 juillet 2019 sur : https://genomics.neogen.com/pdf/prodinfo/axiom_chicken_array_plate_datasheet.pdf

Affymetrix. Axiom® Porcine Genotyping Array. 2015. Consulté le 30 juillet sur : https://genomics.neogen.com/pdf/prodinfo/axiom_porcine_genotyping_array_datasheet.pdf

Affymetrix. Axiom® Trout Genotyping Array. 2015. Consulté le 2 août 2019 sur : https://genomics.neogen.com/pdf/prodinfo/axiom_trout_genotyping_array_datasheet.pdf

Affymetrix. Axiom® Equine Genotyping Array. 2017. Consulté le 31 juillet 2019 sur : https://genomics.neogen.com/pdf/prodinfo/axiom_equine_genotyping_array_hr_universal_datasheet.pdf

Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science. 2010;93:743-752.

Aerts J, Megens HJ, Veenendaal T, Ovcharenko I, Crooijmans R, Gordon L, et al. Extent of linkage disequilibrium in chicken. Cytogenetic and Genome Research. 2007;117:338-345.

Albrechtsen A, Nielsen FC, Nielsen R. Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Molecular Biology and Evolution. 2010;27:2534-2547.

Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009;19:1655-1664.

Aliloo H, Mrode R, Okeyo AM, Ni G, Goddard ME, Gibson JP. The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa. Journal of Dairy Science. 2018;101:9108-9127.

Aliloo H, Pryce JE, González-Recio O, Cocks BG, Hayes BJ. Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. Genetics Selection Evolution. 2016;48:8-18.

Allais S. Détection et validation de marqueurs génétiques impliqués dans la qualité de la viande bovine. Thèse de Doctorat, AgroParisTech-ABIES. 2011.

Andolfatto P, Davison D, Erezyilmaz D, Hu TT, Mast J, Sunayama-Morita T, et al. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Research. 2011;21:610-617.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nature Reviews Genetics. 2016;17:81-92.

Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics. 2002;3:299-309.

Atol Ontology. INRA. 2012. Consulté le 12 février 2019 sur : http://www.atol-ontology.com

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. GenABEL: an R library for genome-wide association analysis. Bioinformatics. 2007;23:1294-1296.

Badke YM, Bates RO, Ernst CW, Schwab C, Steibel JP. Estimation of linkage disequilibrium in four US pig breeds. BMC Genomics. 2012;13:24-33.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, et al. Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE. 2008;3:e3376.

Bapst B, Baes C, Seefried FR, Bieber A, Simianer H, Gredler B. Effect of cows in the reference population: First results in Swiss Brown Swiss. Interbull Bulletin. 2013;:187-191.

Bartholomé J, Van Heerwaarden J, Isik F, Boury C, Vidal M, Plomion C, et al. Performance of genomic prediction within and across generations in maritime pine. BMC Genomics. 2016;17:604-617.

Bastiaansen JWM, Coster A, Calus MPL, van Arendonk JAM, Bovenhuis H. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genetics Selection Evolution. 2012;44:3-15.

Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, et al. Marker Density and Read Depth for Genotyping Populations Using Genotyping-by-Sequencing. Genetics. 2013;193:1073-1081.

Biofidal. MiSeq Illumina sequencing system at Lyon. 2016. Consulté le 5 sept 2019 sur : http://www.biofidal-lab.com/details-miseq+illumina+sequencing+system+at+lyon-50.html

Boichard D, Le Roy P, Levéziel H, Elsen JM. Utilisation des marqueurs moléculaires en génétique animale. INRA Productions Animales. 1998;11:67-80.

Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, et al. Design of a Bovine Low-Density SNP Array Optimized for Imputation. PLoS ONE. 2012a;7:e34130.

Boichard D, Guillaume F, Baur A, Croiseau P, Rossignol MN, Boscher MY, et al. Genomic selection in French dairy cattle. Animal Production Science. 2012b;52:115-120.

Boichard D. La sélection génomique, une opportunité pour améliorer la santé des animaux d'élevage. Bulletin de l'Académie Vétérinaire de France. 2013;166:25-31.

Boichard D, Ducrocq V, Fritz S. Sustainable dairy cattle selection in the genomic era. Journal of Animal Breeding and Genetics. 2015;132:135-143.

Boichard D, Ducrocq V, Croiseau P, Fritz S. Genomic selection in domestic animals: Principles, applications and perspectives. Comptes Rendus Biologies. 2016;339:274-277.

Boichard D, Boussaha M, Capitan A, Rocha D, Hozé C, Sanchez MP, et al. Experience from large scale use of the EuroGenomics custom SNP chip in cattle. In: Proceedings, 11th World Congress of Genetics Applied to Livestock Production. Auckland. 2018.

Bolormaa S, Gore K, van der Werf JHJ, Hayes BJ, Daetwyler HD. Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. Animal Genetics. 2015;46:544-556.

Bolormaa S, Chamberlain AJ, Khansefid M, Stothard P, Swan AA, Mason B, et al. Accuracy of imputation to whole-genome sequence in sheep. Genetics Selection Evolution. 2019;51:1-17.

Bouquet A, Fève K, Riquet J, Larzul C. Précision de l'imputation de génotypages haute densité à partir de puces basse densité pour des individus de race pure et croisés Piétrain. Journées Recherche Porcine. 2015;47:1-6.

Brouard JS, Boyle B, Ibeagha-Awemu EM, Bissonnette N. Low-depth genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the selection of high quality genotypes and the accuracy of imputation. BMC Genetics. 2017;18:32-45.

Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. The American Journal of Human Genetics. 2007;81:1084-1097.

Browning BL, Browning SR. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. The American Journal of Human Genetics. 2009;84:210-223.

Calus MPL, de Haas Y, Veerkamp RF. Combining cow and bull reference populations to increase accuracy of genomic prediction and genome-wide association studies. Journal of Dairy Science. 2013;96:6703-6715.

Calus MPL, Bouwman AC, Hickey JM, Veerkamp RF, Mulder HA. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. Animal. 2014;8:1743-1753.

Carillier C, Larroque H, Palhière I, Clément V, Rupp R, Robert-Granié C. A first step toward genomic selection in the multi-breed French dairy goat population. Journal of Dairy Science. 2013;96:7294-7305.

Carillier-Jacquin C, Bouquet A, Labrune Y, Brenaut P, Riquet J, Larzul C. Using 1K SNP panel for genomic selection in 3 French pig breeds: Accuracy of Imputation and estimation of genomic breeding values using 1K SNP panel, designed for several breeds in French pig populations. In: Proceedings, 11th World Congress of Genetics Applied to Livestock Production. Auckland. 2018.

Carvalheiro R, Boison SA, Neves HHR, Sargolzaei M, Schenkel FS, Utsunomiya YT, et al. Accuracy of genotype imputation in Nelore cattle. Genetics Selection Evolution. 2014;46:69-79.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7-22.

Chen L, Li C, Sargolzaei M, Schenkel F. Impact of Genotype Imputation on the Performance of GBLUP and Bayesian Methods for Genomic Prediction. PLoS ONE. 2014;9:e101544.

Chen Q, Ma Y, Yang Y, Chen Z, Liao R, Xie X, et al. Genotyping by Genome Reducing and Sequencing for Outbred Animals. PLoS ONE. 2013;8:e67500.

Clark SA, Hickey JM, Daetwyler HD, van der Werf JHJ. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genetics Selection Evolution. 2012;44:4-12.

Cleveland MA, Hickey JM. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation1. Journal of Animal Science. 2013;91:3583-3592.

Coppieters W, Riquet J, Arranz JJ, Berzi P, Cambisano N, Grisart B, et al. A QTL with major effect on milk yield and composition maps to bovine chromosome 14. Mammalian Genome. 1998;9:540-544.

Croué I. Évaluation génétique et génomique de nouveaux caractères en bovins laitiers. Thèse de Doctorat, AgroParisTech-ABIES. 2017.

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics. 2010;185:1021-1031.

Dassonneville R, Brøndum RF, Druet T, Fritz S, Guillaume F, Guldbrandtsen B, et al. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. Journal of Dairy Science. 2011;94:3679-3686.

Dassonneville R, Fritz S, Ducrocq V, Boichard D. Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle. Journal of Dairy Science. 2012;95:4136-4140.

Dassonneville R. Sélection génomique des vaches laitières. Thèse de Doctorat, AgroParisTech-ABIES. 2012.

Dassonneville R, Baur A, Fritz S, Boichard D, Ducrocq V. Inclusion of cow records in genomic evaluations and impact on bias due to preferential treatment. Genetics Selection Evolution. 2012;44:40-47.

De Donato M, Peters SO, Mitchell SE, Hussain T, Imumorin IG. Genotyping-by-Sequencing (GBS): A Novel, Efficient and Cost-Effective Genotyping Method for Cattle Using Next-Generation Sequencing. PLoS ONE. 2013;8:e62137.

De Roos APW. Genomic selection in dairy cattle. Thèse de Doctorat, Wageningen University. 2011.

Druet T, Georges M. A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype Reconstruction and Quantitative Trait Locus Fine Mapping. Genetics. 2010;184:789-798.

Druet T, Schrooten C, de Roos APW. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. Journal of Dairy Science. 2010;93:5443-5454.

Druet T, Macleod IM, Hayes BJ. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity. 2014;112:39-47.

Du FX, Clutter AC, Lohuis MM. Characterizing linkage disequilibrium in pig populations. International Journal of Biological Sciences. 2007;3:166-178.

Elbasyoni IS, Lorenz AJ, Guttieri M, Frels K, Baenziger PS, Poland J, et al. A comparison between genotyping-by-sequencing and array-based scoring of SNPs for genomic prediction accuracy in winter wheat. Plant Science. 2018;270:123-130.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE. 2011;6:e19379.

Ensembl. Ensembl Variation - Data sources. Consulté le 2 octobre 2019 sur : https://www.ensembl.org/info/genome/variation/species/sources_documentation.html#gallus_gallus

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, et al. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science. 2012;95:4114-4129.

Ertl J, Legarra A, Vitezica ZG, Varona L, Edel C, Emmerling R, et al. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genetics Selection Evolution. 2014;46:40-49.

Ferraz JBS, Wu X, Li H, Xu J, Ferretti R, Simpson B, et al. Design of a low-density SNP chip for Bos indicus: GGP indicus technical characterization and imputation accuracy to higher density SNP genotypes. In: Proceedings, 11th World Congress of Genetics Applied to Livestock Production. Auckland. 2018.

Fisher RA. The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh. 1918;52:399-433.

Fu W, Dekkers JCM, Lee WR, Abasht B. Linkage disequilibrium in crossbred and pure line chickens. Genetic Selection Evolution. 2015;47:11-22.

Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. Bioinformatics. 2015;31:782-784.

Gardner KM, Brown P, Cooke TF, Cann S, Costa F, Bustamante C, et al. Fast and Cost-Effective Genetic Mapping in Apple Using Next-Generation Sequencing. G3; Genes|Genomes|Genetics. 2014;4:1681-1687.

Goddard ME, Hayes BJ. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics. 2009;10:381-391.

Gorjanc G, Cleveland MA, Houston RD, Hickey JM. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genetics Selection Evolution. 2015;47:12-24.

Groenen MAM, Cheng HH, Bumstead N, Benkel BF, Briles WE, Burke T, et al. A consensus linkage map of the chicken genome. Genome Research. 2000;10:137-147.

Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens HJ, Crooijmans RPMA, et al. A high-density SNP-based linkage map of the chicken genome reveals features correlated with recombination rate. Genome Research. 2008;19:510-519.

Groenen MAM, Megens HJ, Zare Y, Warren WC, Hillier LDW, Crooijmans RPMA, et al. The development and characterization of a 60K SNP chip for chicken. BMC Genomics. 2011;12:274-282.

Groeneveld LF, Lenstra JA, Eding H, Toro MA, Scherf B, Pilling D, et al. Genetic diversity in farm animals – a review. Animal Genetics. 2010;41:6-31.

Grosclaude F, Mahé MF, Brignon G, Di Stasio L, Jeunet R. A Mendelian polymorphism underlying quantitative variations of goat αs1-casein. Genetics Selection Evolution. 1987;19:399-412.

Grossi DA, Brito LF, Jafarikia M, Schenkel FS, Feng Z. Genotype imputation from various low-density SNP panels and its impact on accuracy of genomic breeding values in pigs. animal. 2018;12:2235-2245.

Gualdrón Duarte J, Bates RO, Ernst CW, Raney NE, Cantet RJC, Steibel JP. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC Genetics. 2013;14:38-50.

Guillaume F. Intégration de l'information moléculaire dans l'évaluation génétique. Thèse de Doctorat, AgroParisTech-ABIES. 2009.

Guo Y, He J, Zhao S, Wu H, Zhong X, Sheng Q, et al. Illumina human exome genotyping array clustering and quality control. Nature Protocols. 2014;9:2643-2662.

Gutierrez AP, Turner F, Gharbi K, Talbot R, Lowe NR, Peñaloza C, et al. Development of a Medium Density Combined-Species SNP Array for Pacific and European Oysters (*Crassostrea gigas* and *Ostrea edulis*). G3; Genes|Genomes|Genetics. 2017;7:2209-2218.

Habier D, Fernando RL, Dekkers JCM. The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. Genetics. 2007;177:2389-2397.

Habier D, Fernando RL, Dekkers JCM. Genomic Selection Using Low-Density Marker Panels. Genetics. 2009;182:343-353.

Habier D, Tetens J, Seefried FR, Lichtner P, Thaller G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution. 2010;42:5-16.

Habier D, Fernando RL, Garrick DJ. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics. 2013;194:597-607.

Haley CS, Visscher PM. Strategies to Utilize Marker-Quantitative Trait Loci Associations. Journal of Dairy Science. 1998;81:85-97.

Hardy GH. Mendelian proportions in a mixed population. Science. 1908;28:49-50.

Harris BL, Johnson DL. The impact of high density SNP chips on genomic evaluation in dairy cattle. Interbull Bulletin. 2010:40–43.

Hayes BJ, Goddard ME. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution. 2001;33:209-229.

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science. 2009;92:433-443.

Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K, Goddard ME. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution. 2009;41:51-59.

Hayes BJ. Course Notes: Genomic Selection. Toulouse; 2011.

Hayes BJ, Bowman PJ, Daetwyler HD, Kijas JW, van der Werf JHJ. Accuracy of genotype imputation in sheep breeds: Genotype imputation in sheep. Animal Genetics. 2012;43:72-80.

Hayes BJ, MacLeod IM, Daetwyler HD, Bowman PJ, Chamberlain AJ, Vander Jagt CJ, et al. Genomic Prediction from Whole Genome Sequence in Livestock: the 1000 Bull Genomes Project. In: Proceedings, 10th World Congress of Genetics Applied to Livestock Production. Vancouver. 2014.

Hazel LN. The genetic basis for constructing selection indexes. Genetics. 1943;28:476–490.

Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: Overviews and challenges. BioTechniques. 2014;56:61-77.

Heidaritabar M, Calus MPL, Vereijken A, Groenen MAM, Bastiaansen JWM. High imputation accuracy in layer chicken from sequence data on a few key ancestors. In: Proceedings, 10th World Congress of Genetics Applied to Livestock Production. Vancouver. 2014.

Heidaritabar M, Calus MPL, Vereijken A, Groenen MAM, Bastiaansen JWM. Accuracy of imputation using the most common sires as reference population in layer chickens. BMC Genetics. 2015;16:101-114.

Heidaritabar M, Calus MPL, Megens HJ, Vereijken A, Groenen MAM, Bastiaansen JWM. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. Journal of Animal Breeding and Genetics. 2016a;133:167-179.

Heidaritabar M, Wolc A, Arango J, Zeng J, Settar P, Fulton JE, et al. Impact of fitting dominance and additive effects on accuracy of genomic prediction of breeding values in layers. Journal of Animal Breeding and Genetics. 2016b;133:334-346.

Henderson CR. Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics. 1975;31:423-447.

Henderson CR. A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. Biometrics. 1976;32:69-83.

Hérault F, Herry F, Varenne A, Burlot T, Picard-Druet D, Recoquillay J, et al. A linkage disequilibrium study in layer and broiler commercial chicken populations. In: Proceedings, 11th World Congress of Genetics Applied to Livestock Production. Auckland. 2018.

Herry F, Hérault F, Picard Druet D, Varenne A, Burlot T, Le Roy P, et al. Design of low density SNP chips for genotype imputation in layer chicken. BMC Genetics. 2018;19:108-121.

Herry F, Picard Druet D, Hérault F, Varenne A, Burlot T, Le Roy P, et al. Interest of using imputation for genomic evaluation in layer chicken. 2019. Submitted to Poultry Science. 2019

Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, van der Werf JHJ. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genetics Selection Evolution. 2011;43:12-24.

Hickey JM, Crossa J, Babu R, de los Campos G. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Science. 2012;52:654-663.

Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics. 1968;38:226-231.

Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, et al. Detection of genotyping errors by Hardy-Weinberg equilibrium testing. European Journal of Human Genetics. 2004;12:395-399.

Houston RD, Taggart JB, Cézard T, Bekaert M, Lowe NR, Downing A, et al. Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar). BMC Genomics. 2014;15:90-102.

Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genetics. 2009;5:e1000529.

Howie BN, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics. 2012;44:955-959.

Hozé C, Fouilloux MN, Venot E, Guillaume F, Dassonneville R, Fritz S, et al. High-density marker imputation accuracy in sixteen French cattle breeds. Genetics Selection Evolution. 2013;45:33-45.

Hozé C. Développement d'évaluations génomiques multiraciales chez les bovins laitiers. Thèse de Doctorat, AgroParisTech-ABIES. 2014.

Hozé C, Fritz S, Phocas F, Boichard D, Ducrocq V, Croiseau P. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. Journal of Dairy Science. 2014a;97:3918-3929.

Hozé C, Fritz S, Phocas F, Boichard D, Ducrocq V, Croiseau P. Genomic evaluation using combined reference populations from Montbéliarde and French Simmental breeds. In: Proceedings, 10th World Congress of Genetics Applied to Livestock Production. Vancouver. 2014.

Illumina. OvineSNP50 Genotyping BeadChip. 2008. Consulté le 30 juillet 2019 sur : https://genomics.neogen.com/pdf/prodinfo/illumina_ovine_snp50_beadchip_datasheet.pdf

Illumina. Illumina Porcine60K BeadChip. 2009. Consulté le 30 juillet 2019 sur : https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_porcinesnp60.pdf

Illumina. BovineLD v2.0 Genotyping BeadChip. 2011a. Consulté le 30 juillet 2019 sur : https://www.illumina.com/Documents/products/datasheets/datasheet_bovineLD.pdf

Illumina. BovineSNP50 Genotyping BeadChip. 2011b. Consulté le 30 juillet 2019 sur : https://www.illumina.com/Documents/products/datasheets/datasheet_bovine_snp5O.pdf

Illumina. GoldenGate® Bovine3K Genotyping BeadChip. 2011c. Consulté le 30 juillet 2019 sur : https://www.illumina.com/Documents/products/datasheets/datasheet_bovine3k.pdf

Illumina. BovineHD Bead Chip datasheet. 2012a. Consulté le 30 juillet 2019 sur : http://www.illumina.com/documents/products/datasheets/datasheet_bovineHD.pdf

Illumina. EquineSNP50 Genotyping BeadChip. 2012b. Consulté le 31 juillet 2019 sur : https://genomics.neogen.com/pdf/prodinfo/illumina_equine_snp50_beadchip_datasheet.pdf

Illumina. An introduction to Next-Generation Sequencing Technology. 2017. Consulté le 5 septembre 2019 sur : https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature. 2004;432:695-716.

International Chicken Polymorphism Map Consortium. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature. 2004;432:717-722.

ITAVI. Situation du marché des oeufs et ovoproduits - Édition avril 2018. Note de conjoncture Poules Pondeuses. 2018.

Khanyile KS, Dzomba EF, Muchadeyi FC. Population genetic structure, linkage disequilibrium and effective population size of conserved and extensively raised village chicken populations of Southern Africa. Frontier in Genetics. 2015;6:1-11.

Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. Nature Genetics. 2008;40:1068-1075.

Kranis A, Gheyas AA, Boschiero C, Turner F, Yu L, Smith S, et al. Development of a high density 600K SNP genotyping array for chicken. BMC Genomics. 2013;14:59-71.

Kuhn MT, Boettcher PJ, Freeman AE. Potential Biases in Predicted Transmitting Abilities of Females from Preferential Treatment. Journal of Dairy Science. 1994;77:2428-2437.

Laloë D. La genèse et le développement des concepts de l'évaluation génétique classique. INRA Productions Animales. 2011;24:323-330.

Lande R, Thompson R. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics. 1990;124:743-756.

Legarra A, Robert-Granié C, Manfredi E, Elsen JM. Performance of Genomic Selection in Mice. Genetics. 2008;180:611-618.

Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science. 2009;92:4656-4663.

Legarra A, Baloche G, Barillet F, Astruc JM, Soulas C, Aguerre X, et al. Within- and across-breed genomic predictions and genomic relationships for Western Pyrenees dairy sheep breeds Latxa, Manech, and Basco-Béarnaise. Journal of Dairy Science. 2014;97:3200-3212.

Legarra A, Christensen OF, Aguilar I, Misztal I. Single Step, a general approach for genomic selection. Livestock Science. 2014;166:54-65.

Le Roy P, Naveau J, Elsen JM, Sellier P. Evidence for a new major gene influencing meat quality in pigs. Genetical Research. 1990;55:33-40.

Le Roy P, Chapuis H, Guemene D. Sélection génomique : quelles perspectives pour les filières avicoles ?. INRA Production Animales. 2014;27:331–336.

Lewontin RC. The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics. 1964;49:49-67.

Lewontin RC. On measures of gametic disequilibrium. Genetics. 1988;120:849-852.

Liao R, Wang Z, Chen Q, Tu Y, Chen Z, Wang Q, et al. An Efficient Genotyping Method in Chicken Based on Genome Reducing and Sequencing. PLOS ONE. 2015;10:e0137010.

Lillehammer M, Meuwissen THE, Sonesson AK. Genomic selection for maternal traits in pigs. Journal of Animal Science. 2011;89:3908-3916.

Liu R, Xing S, Wang J, Zheng M, Cui H, Crooijmans RPMA, et al. A new chicken 55K SNP genotyping array. BMC Genomics. 2019;20:410-421.

Liu Z, Seefried FR, Reinhardt F, Rensing S, Thaller G, Reents R. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genetics Selection Evolution. 2011;43:19-27.

Lourenco DAL, Tsuruta S, Fragomeni BO, Masuda Y, Aguilar I, Legarra A, et al. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. Journal of Animal Science. 2015;93:2653-2662.

Lu D, Sargolzaei M, Kelly M, Li C, Vander Voort G, Wang Z, et al. Linkage disequilibrium in Angus, Charolais and Crossbred beef cattle. Frontier in Genetics. 2012;3:1-10.

Ma P, Brøndum RF, Zhang Q, Lund MS, Su G. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. Journal of Dairy Science. 2013;96:4666-4677.

Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature Reviews Genetics. 2010;11:499-511.

Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, et al. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. International Journal of Methods in Psychiatric Research. 2018;27:e1608.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, et al. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE. 2009;4:e5350.

McCue ME, Bannasch DL, Petersen JL, Gurr J, Bailey E, Binns MM, et al. A High Density SNP Array for the Domestic Horse and Extant Perissodactyla: Utility for Association Mapping, Genetic Diversity, and Phylogeny Studies. PLoS Genetics. 2012;8:e1002451.

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH. BLUPF90 and related programs (BGF90). In: Proceedings, 7th World Congress of Genetics Applied to Livestock Production. Montpellier. 2002.

Megens HJ, Crooijmans RPMA, Bastiaansen JWM, Kerstens HHD, Coster A, Jalving R, et al. Comparison of linkage disequilibrium and haplotype diversity on macro- and microchromosomes in chicken. BMC Genetics. 2009;10:86-96.

Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819-1829.

Meuwissen T, Hayes B, Goddard M. Genomic selection: A paradigm shift in animal breeding. Animal Frontiers. 2016;6:6-14.

Moghaddar N, Gore KP, Daetwyler HD, Hayes BJ, van der Werf JHJ. Accuracy of genotype imputation based on random and selected reference sets in purebred and crossbred sheep populations and its effect on accuracy of genomic prediction. Genetics Selection Evolution. 2015;47:97-108.

Moghaddar N, van der Werf JHJ. Genomic estimation of additive and dominance effects and impact of accounting for dominance on accuracy of genomic evaluation in sheep populations. Journal of Animal Breeding and Genetics. 2017;134:453-462.

Muir WM. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters: Comparison of BLUP and GEBV selection. Journal of Animal Breeding and Genetics. 2007;124:342-355.

Muir WM, Wong GK, Zhang Y, Wang J, Groenen MAM, Crooijmans RPMA, et al. Review of the initial validation and characterization of a 3K chicken SNP array. World's Poultry Science Journal. 2008;64:219-226.

Mulder HA, Calus MPL, Druet T, Schrooten C. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. Journal of Dairy Science. 2012;95:876-889.

Neogen. GeneSeek® Genomic Profiler™ Bovine 50K. 2012. Consulté le 30 juillet 2019 sur : https://genomics.neogen.com/pdf/ag311_ggp_bovine50k_brochure.pdf

Neogen. GeneSeek ® Genomic Profiler™ indicus. 2018. Consulté le 13 août 2019 sur : https://genomics.neogen.com/pdf/ag359_ggp_indicus_brochure.pdf

Neogen. GeneSeek® Genomic Profiler Porcine. Consulté le 30 juillet 2019 sur : https://genomics.neogen.com/pdf/slicks/ag284_ggp_porcine.pdf

Ni G, Cavero D, Fangmann A, Erbe M, Simianer H. Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. Genetics Selection Evolution. 2017;49:8-21.

Nishio M, Satoh M. Including Dominance Effects in the Genomic BLUP Method for Genomic Evaluation. PLoS ONE. 2014;9:e85792.

Olson KM, VanRaden PM, Tooker ME. Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. Journal of Dairy Science. 2012;95:5378-5383.

Owen JJT. Karyotype studies on Gallus domesticus. Chromosoma. 1965;16:601-608.

Palti Y, Gao G, Liu S, Kent MP, Lien S, Miller MR, et al. The development and characterization of a 57K single nucleotide polymorphism array for rainbow trout. Molecular Ecology Resources. 2015;15:662-672.

Patry C, Ducrocq V. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. Journal of Dairy Science. 2011;94:1011-1020.

Pengelly RJ, Gheyas AA, Kuo R, Mossotto E, Seaby EG, Burt DW, Ennis S, Collins A. Commercial chicken breeds exhibit highly divergent patterns of linkage disequilibrium. Heredity. 2016;117:375-382.

Pérez O'Brien AM, Mészáros G, Utsunomiya YT, Sonstegard TS, Garcia JF, Van Tassell CP, et al. Linkage disequilibrium levels in Bos indicus and Bos taurus cattle using medium and high density SNP chip data and different minor allele frequency distributions. Livestock Science. 2014;166:121-132.

Pértille F, Guerrero-Bosagna C, Silva VH, Boschiero C, Nunes JRS, Ledur MC, et al. High-throughput and Cost-effective Chicken Genotyping Using Next-Generation Sequencing. Scientific Reports. 2016;6:1-12.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE. 2012;7:e37135.

Picard Druet D, Varenne A, Herry F, Hérault F, Allais S, Burlot T, et al. Relevance of genomic evaluation for egg quality traits in layers. BMC Genetics (in press). 2019.

Piccoli ML, Braccini J, Cardoso FF, Sargolzaei M, Larmer SG, Schenkel FS. Accuracy of genome-wide imputation in Braford and Hereford beef cattle. BMC Genetics. 2014;15:157-171.

Poland JA, Brown PJ, Sorrells ME, Jannink JL. Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. PLoS ONE. 2012;7:e32253.

Poland J, Endelman J, Dawson J, Rutkoski J, Wu S, Manes Y, et al. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. The Plant Genome Journal. 2012;5:103-113.

Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. American Journal of Human Genetics. 2001;69:1-14.

Pryce JE, Hayes BJ, Goddard ME. Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. In: Proceedings, 38th ICAR Session. 2012;28.

Pszczola M, Mulder HA, Calus MPL. Effect of enlarging the reference population with (un)genotyped animals on the accuracy of genomic selection in dairy cattle. Journal of Dairy Science. 2011;94:431-441.

Pszczola M, Strabel T, Mulder HA, Calus MPL. Reliability of direct genomic values for animals with different relationships within and to the reference population. Journal of Dairy Science. 2012;95:389-400.

Qanbari S, Hansen M, Weigend S, Preisinger R, Simianer H. Linkage disequilibrium reveals different demographic history in egg laying chickens. BMC Genetics. 2010;11:103-115.

R Core Team. R: A language and environment for statistical computing. Vienne. 2017.

Ramos AM, Crooijmans RPMA, Affara NA, Amaral AJ, Archibald AL, Beever JE, et al. Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. PLoS ONE. 2009;4:e6524.

Raoul J, Swan AA, Elsen JM. Using a very low-density SNP panel for genomic selection in a breeding program for sheep. Genetics Selection Evolution. 2017;49:76-87.

Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney RK, et al. Crop Breeding Chips and Genotyping Platforms: Progress, Challenges, and Perspectives. Molecular Plant. 2017;10:1047-1064.

Robert-Granié C, Legarra A, Ducrocq V. Principes de base de la sélection génomique. INRA Productions Animales. 2011;24:331-340.

Sargolzaei M, Schenkel FS, Jansen GB, Schaeffer LR. Extent of linkage disequilibrium in Holstein Cattle in North America. Journal of Dairy Science. 2008;91:21062117.

Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. BMC Genomics. 2014;15:478-489.

Schaefer RJ, Schubert M, Bailey E, Bannasch DL, Barrey E, Bar-Gal GK, et al. Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. BMC Genomics. 2017;18:565-582.

Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics. 2006;123:218-223.

Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet. 2006;78:629-644.

Schrooten C, Dassonneville R, Ducrocq V, Brøndum RF, Lund MS, Chen J, et al. Error rate for imputation from the Illumina BovineSNP50 chip to the Illumina BovineHD chip. Genetics Selection Evolution. 2014;46:10-18.

Sitzenstock F, Ytournel F, Sharifi AR, Cavero D, Täubert H, Preisinger R, et al. Efficiency of genomic selection in an established commercial layer breeding program. Genetics Selection Evolution. 2013;45:29-39.

Solberg TR, Sonesson AK, Woolliams JA, Meuwissen THE. Genomic selection using different marker types and densities. Journal of Animal Science. 2008;86:2447-2454.

Solberg TR, Sonesson AK, Woolliams JA, Ødegard J, Meuwissen THE. Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. Genetics Selection Evolution. 2009;41:53-60.

Su G, Brøndum RF, Ma P, Guldbrandtsen B, Aamand GP, Lund MS. Comparison of genomic predictions using medium-density (∼54,000) and high-density (∼777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of Dairy Science. 2012;95:4657-4665.

Sun C, Wu XL, Weigel KA, Rosa GJM, Bauck S, Woodward BW, et al. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genetics Research. 2012;94:133-150.

Sun C, VanRaden PM, Cole JB, O'Connell JR. Improvement of Prediction Ability for Genomic Selection of Dairy Cattle by Including Dominance Effects. PLoS ONE. 2014;9:e103934.

Thébault N, Riquet J, Diot C, Brard-Fudulea S, Guéméné D, Alletru B, et al. Développement d'une puce de génotypage haute densité 600K pour le canard commun et le canard de barbarie. In: Treizièmes Journées de la Recherche Avicole et Palmipèdes à Foie Gras. 2019 ; 75-79.

Torkamaneh D, Belzile F. Scanning and Filling: Ultra-Dense SNP Genotyping Combining Genotyping-By-Sequencing, SNP Array and Whole-Genome Resequencing Data. PLOS ONE. 2015;10:e0131533.

Torkamaneh D, Boyle B, Belzile F. Efficient genome-wide genotyping strategies and data integration in crop plants. Theoretical and Applied Genetics. 2018;131:499-511.

Tosser-Klopp G, Bardou P, Bouchez O, Cabau C, Crooijmans RPMA, Dong Y, et al. Design and Characterization of a 52K SNP Chip for Goats. PLoS ONE. 2014;9:e86227.

Tribout T. Perspectives d'application de la sélection génomique dans les schémas d'amélioration génétique porcins. INRA Productions Animales. 2011;24:369.

Tribout T, Bidanel JP, Phocas F, Schwob S, Guillaume F, Larzul C. La sélection génomique : principe et perspectives d'utilisation pour l'amélioration des populations porcines. Journées Recherche Porcine. 2011;13–25.

Van Eenennaam AL, Weigel KA, Young AE, Cleveland MA, Dekkers JCM. Applied Animal Genomics: Results from the Field. Annual Review of Animal Biosciences. 2014;2:105-139.

van Orsouw NJ, Hogers RCJ, Janssen A, Yalcin F, Snoeijers S, Verstege E, et al. Complexity Reduction of Polymorphic Sequences (CRoPS^{TM}): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. PLoS ONE. 2007;2:e1172.

VanRaden PM. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science. 2008;91:4414-4423.

VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA. Genomic evaluations with many more genotypes. Genetics Selection Evolution. 2011;43:10-20.

VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, et al. Genomic imputation and evaluation using high-density Holstein genotypes. Journal of Dairy Science. 2012;96:668-678.

Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, et al. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods. 2008;5:247-252.

Varona L, Misztal I. Prediction of Parental Dominance Combinations for Planned Matings, Methodology, and Simulation Results. Journal of Dairy Science. 1999;82:2186-2191.

Ventura RV, Lu D, Schenkel FS, Wang Z, Li C, Miller SP. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle1. Journal of Animal Science. 2014;92:1433-1444.

Vereijken ALJ, Albers GAA, Visscher J. Imputation of SNP genotypes in chicken using a reference panel with phased haplotypes. In: Proceedings, 9th World Congress of Genetics Applied to Livestock Production. Leipzig. 2010.

Vignal A. État de la carte de la poule. INRA Productions Animales. 2000;Hors-série « Génétique moléculaire : principes et applications aux populations animales »:113-114.

Wang C, Habier D, Peiris BL, Wolc A, Kranis A, Watson KA, et al. Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide polymorphism panel in broiler chickens. Poultry Science. 2013;92:1712-23.

Wang H, Misztal I, Aguilar I, Legarra A, Muir WM. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics Research. 2012;94:73-83.

Wang S, Meyer E, McKay JK, Matz MV. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nature Methods. 2012;9:808-810.

Warren WC, Hillier LW, Tomlinson C, Minx P, Kremitzki M, Graves T, et al. A New Chicken Genome Assembly Provides Insight into Avian Genome Structure. G3. 2017;7:109-117.

Weigel KA, de los Campos G, González-Recio O, Naya H, Wu XL, Long N, et al. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. Journal of Dairy Science. 2009;92:5248-5257.

Weigel KA, de los Campos G, Vazquez AI, Rosa GJM, Gianola D, Van Tassell CP. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. Journal of Dairy Science. 2010;93:5423-5435.

Weigel KA, Van Tassell CP, O'Connell JR, VanRaden PM, Wiggans GR. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. Journal of Dairy Science. 2010;93:2229-2238.

Weir BS, Cockerham CC. Estimating F-Statistics for the analysis of population structure. Evolution. 1984;38-1358-1370.

Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. Trends in Genetics. 2002;18:19-24.

Wellmann R, Bennewitz J. Bayesian models with dominance effects for genomic evaluation of quantitative traits. Genetics Research. 2012;94:21-37.

Wellmann R, Preuß S, Tholen E, Heinkel J, Wimmers K, Bennewitz J. Genomic selection using low density marker panels with application to a sire line in pigs. Genetics Selection Evolution. 2013;45:28-38.

Weng Z, Saatchi M, Schnabel RD, Taylor JF, Garrick DJ. Recombination locations and rates in beef cattle assessed from parent-offspring pairs. Genetics Selection Evolution. 2014;46:34-47.

Weng Z, Wolc A, Shen X, Fernando RL, Dekkers JCM, Arango J, et al. Effects of number of training generations on genomic prediction for various traits in a layer chicken population. Genetics Selection Evolution. 2016;48:22-31.

Wiggans GR, Cooper TA, VanRaden PM, Cole JB. Technical note: Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. Journal of Dairy Science. 2011;94:6188-6193.

Wiggans GR, Cooper TA, VanRaden PM, Olson KM, Tooker ME. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. Journal of Dairy Science. 2012;95:1552-1558.

Wiggans GR, Cooper TA, Van Tassell CP, Sonstegard TS, Simpson EB. Technical note: Characteristics and use of the Illumina BovineLD and GeneSeek Genomic Profiler low-density bead chips for genomic evaluation. Journal of Dairy Science. 2013;96:1258-1263.

Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. Genomic Selection in Dairy Cattle: The USDA Experience. Annual Review of Animal Biosciences. 2017;5:309-327.

Wolc A, Arango J, Settar P, Fulton JE, O'Sullivan NP, Preisinger R, et al. Persistence of accuracy of genomic estimated breeding values over generations in layer chickens. Genetics Selection Evolution. 2011;43:23-30.

Wolc A, Zhao HH, Arango J, Settar P, Fulton JE, O'Sullivan NP, et al. Response and inbreeding from a genomic selection experiment in layer chickens. Genetics Selection Evolution. 2015;47:59-70.

Wolc A, Kranis A, Arango J, Settar P, Fulton JE, O'Sullivan NP, et al. Implementation of genomic selection in the poultry industry. Animal Frontiers. 2016;6:23-31.

Wragg D, Mwacharo JM, Alcalde JA, Hocking PM, Hanotte O. Analysis of genome-wide structure, diversity and fine mapping of Mendelian traits in traditional and village chickens. Heredity. 2012;109:6-18.

Wright S. The genetical structure of populations. Annals of Eugenics. 1951;15:323-354.

Wright S. Evolution and the Genetics of Populations.Vol.4. Variability Within and Among Natural Populations. University of Chicago Press. Chicago. 1978.

Xiang T, Christensen OF, Vitezica ZG, Legarra A. Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genetics Selection Evolution. 2016;48:92-105.

Ytournel F. Déséquilibre de liaison et cartographie de QTL en population sélectionnée. Thèse de Doctorat, AgroParisTech-ABIES. 2008.

Zhai Z, Zhao W, He C, Yang K, Tang L, Liu S, et al. SNP discovery and genotyping using restriction-site-associated DNA sequencing in chickens. Animal Genetics. 2015;46:216-219.

Zhang X, Lourenco D, Aguilar I, Legarra A, Misztal I. Weighting Strategies for Single-Step Genomic BLUP: An Iterative Approach for Accurate Calculation of GEBV and GWAS. Frontiers in Genetics. 2016;7:743-756.

Zhang Z, Druet T. Marker imputation with low-density marker panels in Dutch Holstein cattle. Journal of Dairy Science. 2010;93:5487-5494.

Zhang Z, Ding X, Liu J, Zhang Q, de Koning D-J. Accuracy of genomic prediction using low-density marker panels. Journal of Dairy Science. 2011;94:3642-3650.

Zhao S, Jing W, Samuels DC, Sheng Q, Shyr Y, Guo Y. Strategies for processing and quality control of Illumina genotyping arrays. Briefings in Bioinformatics. 2018;19:765-775.

Annexes

Poster 12èmes Journées de la Recherche Avicole – 5-6 Avril 2017 – Tours

CRÉATION D'UNE PUCE BASSE-DENSITÉ POUR LA SÉLECTION GÉNOMIQUE EN POULE PONDEUSE

Herry Florian[1,2], Hérault Frédéric[2], Varenne Amandine[1], Burlot Thierry[1], Le Roy Pascale[2] et Allais Sophie[2]

[1]NOVOGEN, 22800 Le Foeil, [2]PEGASE, INRA, Agrocampus Ouest, 35590 Saint-Gilles

INTRODUCTION

Depuis 2013, une puce commerciale de génotypages à haute densité (HD) de 600 000 SNP pour l'espèce poule est disponible et permet la mise en place de la sélection génomique dans cette espèce. Toutefois, les coûts de génotypages avec cette puce restant élevés, seuls les reproducteurs peuvent être génotypés en routine sur cette puce. Un génotypage sur puce basse densité (BD), à coût réduit, doit être envisagé sur les candidats à la sélection qui sont très nombreux. Si les marqueurs sont bien choisis, l'imputation permet ensuite de déduire les génotypes manquants sur puce HD.

L'objectif de cette étude est de choisir la stratégie de génotypages basse densité la mieux adaptée à la lignée de poule pondeuse considérée, afin d'optimiser à la fois précision des évaluations génomiques des candidats et coût du schéma de sélection.

MATÉRIELS ET MÉTHODES

Population d'étude :

Étude de deux générations de coqs d'une lignée de poules pondeuses de la société Novogen du groupe Grimaud, génotypés sur puce HD (Figure 1). Après contrôle qualité, 282 928 SNP sont retenus et répartis sur les macro-chromosomes (1 à 5), les chromosomes intermédiaires (6 à 10), les micro-chromosomes (11 à 33) et le chromosome Z.

Puce basse densité :

Simulation de 8 puces basse densité (Tableau 1) selon deux méthodologies intra-chromosomes.

- Méthodologie « équidistante » : Sélection de SNP à intervalles réguliers.
- Méthodologie « DL » : Sélection de SNP en fonction du DL entre SNP, par clustering hiérarchique.

Stratégies d'imputation : utilisation du logiciel Fimpute (Sargolzaei et al., 2014)

À partir des puces, étude de l'effet de la densité des SNP sur puces BD, du seuil de DL utilisé pour construire les puces, et de la méthodologie utilisée sur le taux d'erreur génotypique.

Évaluations génomiques : utilisation de la méthode du BLUP Single Step (Legarra et al., 2009)

- Évaluations génomiques des candidats sur trois caractères bien distincts : intensité de ponte, poids d'œuf et couleur de la coquille des œufs (Lab).
- Comparaison des classements des 150 meilleurs individus obtenus avec les génotypages HD et les génotypages imputés.

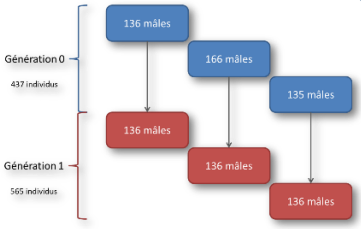


Figure 1 : Structure de la population d'étude

Méthodologie	Puce	Nombre de SNP
Équidistante	20Kequi	18 196
	10Kequi	9 352
	3Kequi	3 337
Déséquilibre de liaison (DL)	DL 0.8	18 359
	DL 0.5	9 820
	DL 0.2	5 224
	DL 0.1	3 988
	DL 0.05	3 357

Tableau 1 : Récapitulatif des puces étudiées

RÉSULTATS ET DISCUSSION

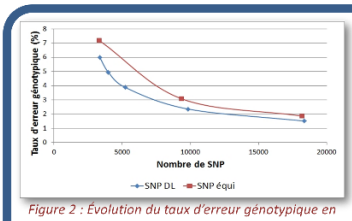


Figure 2 : Évolution du taux d'erreur génotypique en fonction du nombre de SNP sur puces BD

Influence de la densité de marqueurs :

- Diminution du taux d'erreur génotypique avec une augmentation du nombre de SNP (Figure 2).
- Valable pour les deux méthodologies.

Influence du seuil de déséquilibre de liaison :

- Diminution du taux d'erreur génotypique avec une augmentation du seuil de DL (Figure 3).
- En augmentant le seuil de DL, le nombre de SNP sur puce BD augmente. L'augmentation du seuil de DL permet également de choisir un SNP encore plus représentatif de son groupe.

Choix de la méthodologie :

- Méthodologie équidistante : nombre de SNP retenus sur puce LD proportionnel à la taille du chromosome (Figure 4).
- Méthodologie DL : Prise en compte de la structure particulière du DL de l'espèce poule et de la persistance du DL plus faible sur les micro-chromosomes que sur les macro-chromosomes.
  - → Un plus grand nombre de SNP est nécessaire sur les micro-chromosomes pour couvrir tout le chromosome (Figure 4).
- Méthodologie DL qui semble la plus adaptée pour obtenir des bonnes imputations.

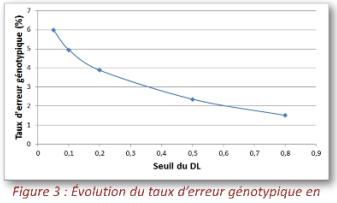


Figure 3 : Évolution du taux d'erreur génotypique en fonction du seuil de DL

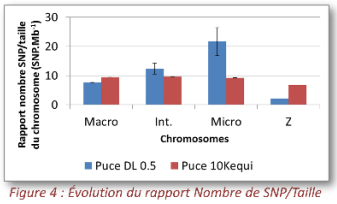


Figure 4 : Évolution du rapport Nombre de SNP/Taille du chromosome en fonction du type de chromosome pour les puces DL 0.5 et 10Kequi

Impact sur les évaluations génomiques :

- Le but de l'évaluation génomique est de classer les individus les uns par rapport aux autres.
- Pour les trois caractères étudiés, diminution significative des corrélations avec une diminution du nombre de SNP sur puces BD ainsi qu'avec une diminution du seuil de DL.
- À densité de SNP équivalente, résultats meilleurs significativement pour la puce 10Kequi comparée à la puce DL 0.5 pour la couleur de coquille (0.9695 contre 0.9460) et le poids d'œuf (0.9855 contre 0.9741).
- De même, résultats significativement meilleurs pour la couleur de coquille avec la puce 3Kequi (par rapport à la puce DL 0.05).

- **Meilleurs résultats d'imputation obtenus avec les puces DL.**
- **Résultats des évaluations génomiques, à densité de SNP équivalente, meilleurs pour certains caractères avec les puces équidistantes.**
- **Nécessité d'étudier l'Influence de la méthodologie d'évaluation sur ces résultats.**

GENOTYPING STRATEGIES FOR GENOMIC SELECTION IN LAYER CHICKENS

Florian HERRY[1,2], Frédéric Hérault[2], Amandine Varenne[1], Thierry Burlot[1], Pascale LE ROY[2]* & Sophie ALLAIS[2]*

[1]NOVOGEN, 22800 Le Foeil, [2]PEGASE, INRA, Agrocampus Ouest, 35590 Saint-Gilles

*Supervisors

CONTEXT

Since 2013, a commercial high-density SNP chip of 600 000 SNP for chicken is available and enables the implementation of genomic selection in layers production. However, genotyping costs still remain high for a routine use on a large number of selection candidates. Combining genotyping on low-density SNP chip, at a lower cost, and genotype imputation should be considered on a large number of selection candidates. Thus, the definition of SNP panel is the milestone of this approach.

Concurrently, the development of Next Generation Sequencing (NGS) enables as of now to consider others solutions than SNP chips, at the level of breeders and selection candidates, at different costs and accuracies, to do genomic selection.

THESIS OBJECTIVES

The main goal of the thesis will be, focusing on two laying hens lines selected by Novogen, to study different genotyping strategies of breeders and selection candidates to :

- Optimize the accuracy of genomic evaluations
- Minimize selection costs

MATERIALS AND METHODS

Study populations: Study of two different lines of *Rhode Island* and *Leghorn* from Novogen, genotyped on HD SNP chip (Table 1).

Line	Rhode Island	Leghorn
Number	1469	1102
Sires	323 ♂	189 ♂
Dams	1146 ♀	913 ♀
Nb of Generations	4	2

Table 1: Study populations

Low-density SNP chips:
Simulation of low-density SNP chip according to two intra-chromosomes methodologies:
- « Equidistant » methodology: Selection of SNP at regular intervals.
- « LD » methodology: Selection of SNP according to the Linkage Disequilibrium (LD) between SNP.

Sequences:
- 90 sequences at high sequencing depth (20X) of the first generation of *Rhode Island*.
- Forecast of 200-300 sequences at low sequencing depth obtained by RAD-Seq (Restriction-site-Associated DNA Sequencing) of the second generation of *Rhode Island.*

Imputation strategies: Use of Fimpute (Sargolzaei et al., 2014) and Beagle (Browning and Browning, 2016).

Genomic evaluations: Use of BLUP Single Step (Legarra et al., 2009) and Weighted Single Step (Zhang et al., 2016) methods.
- Genomic evaluations of candidates on 3 distinct characters: laying intensity, egg weight and egg shell color (Lab).
- Comparison of the rankings of the top 150 individuals obtained with HD genotyping and imputed genotyping.

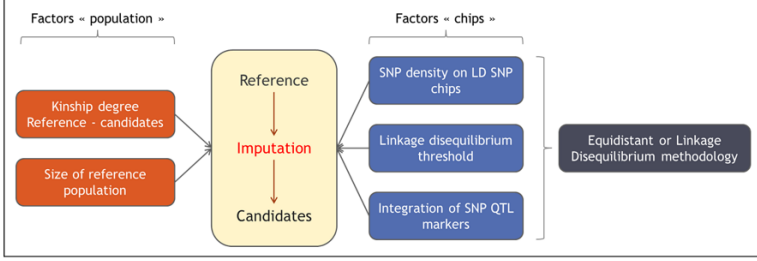
OBJECTIVES

1st objective

Study of different factors influencing imputation accuracy in chicken species for two lines, from low density SNP chips.
Comparison of three criteria of imputation accuracy :
- Genotyping error rate
- Allelic error rate
- Genotypes correlation

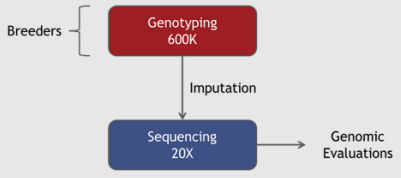
2nd objective

Based on previous studies :
- Creation of a multi-line chicken LD SNP chip
- Study of imputation accuracy
- Impact on genomic evaluations

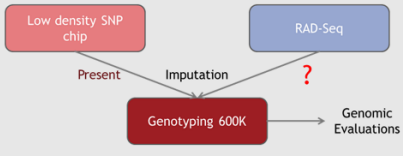


3rd objective

Study of the interest of sequences :
- Regarding breeders



- Regarding selection candidates



Funding : CIFRE contract with Novogen
Cursus : Double degree - Agrocampus Ouest (Rennes)
- Ingénieur Spécialité Ingénieur Agronome, spécialisation Ingénierie Zootechnique
- MASTER Sciences, Technologies, Santé, mention Agronomie, Biologie, Alimentation, spécialité Sciences de l'Animal pour l'Élevage de Demain

DESIGN OF A LOW DENSITY SNP CHIP FOR GENOMIC SELECTION IN LAYER CHICKENS

Herry Florian[1,2], Hérault Frédéric[2], Varenne Amandine[1], Burlot Thierry[1], Le Roy Pascale[2] & Allais Sophie[2]

[1]NOVOGEN, 22800 Le Foeil, [2]PEGASE, INRA, Agrocampus Ouest, 35590 Saint-Gilles

INTRODUCTION

Since 2013, a commercial high-density SNP chip of 600 000 SNP for chicken is available and enables the implementation of genomic selection in layers production. However, genotyping costs still remain high for a routine use on a large number of selection candidates. Combining genotyping on low-density SNP chip, at a lower cost, and genotype imputation should be considered on a large number of selection candidates. Thus, the definition of SNP panel is the milestone of this approach.

The main objective of this study is to choose the best strategy for low density genotyping of laying hen lines in order to optimize selection scheme.

MATERIALS AND METHODS

Study populations: Study of two different lines of *Rhode Island (RI)* and *Leghorn (L)* from Novogen, genotyped on HD SNP chip (Table 1). After quality control, respectively 300 351 and 245 667 SNP are retained and distributed for *Rhode Island* and *Leghorn* lines on the genome.

Line	Rhode Island	Leghorn
Number	1027	1474
Sires	1027 ♂	561 ♂
Dams	0 ♀	913 ♀
Generation 0 (G0)	447	711
Generation 1 (G1)	580	763

Table 1 : Study populations

Low density SNP chips:
Simulation of low-density SNP chip according to two intra-chromosomes methodologies (Table 2):

- « Equidistant » methodology: Selection of SNP at regular intervals.
- « Linkage Disequilibrium » methodology: Selection of SNP according to the LD between SNP.

Imputation strategies : Use of Fimpute (Sargolzaei et al., 2014) to impute G0 from G1. From low density SNP chips designed, study of the effect of:

- SNP density,
- LD threshold used to designed low density SNP chips,
- the type of chromosome (macro-chromosomes (1 to 5), intermediate chromosomes (6 to 10), micro-chromosomes (11 to 33) and sexual chromosome Z,
- MAF (Minor Allelic Frequence) of SNP,
- the methodology used to designed low-density SNP chips.

Methodology	SNP Chip	Number of SNP	
		Rhode Island	Leghorn
Equidistant	50Kequi	49636	50307
	40Kequi	40160	39838
	30Kequi	29970	30075
	20Kequi	19910	19948
	15Kequi	14963	14955
	10Kequi	10001	9966
	7.5Kequi	7527	7496
	5Kequi	4991	4996
	4Kequi	4023	4000
	3Kequi	2992	3003
	2Kequi	2013	2004
Linkage Disequilibrium	DL0.8	21717	18052
	DL0.7	16615	13696
	DL0.6	13214	10736
	DL0.5	10711	8626
	DL0.4	8521	6944
	DL0.3	6875	5578
	DL0.2	5371	4330

Table 2 : Summary of SNP chips studied

RESULTS AND DISCUSSION

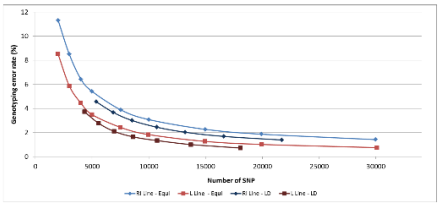


Figure 2 : Evolution of genotyping error rate as a function of the number of SNP on low density SNP chip

Influence of SNP density:

- Decrease of genotyping error rate with an increase in the number of SNP (Figure 2).
- Valid for both methodologies.

Influence of LD threshold:

- Decrease of genotyping error rate with an increase in LD threshold (Figure 3).
- By increasing LD threshold, the number of SNP on low density SNP chip increases. The increase of LD threshold also makes it possible to choose a SNP still more representative of its group.

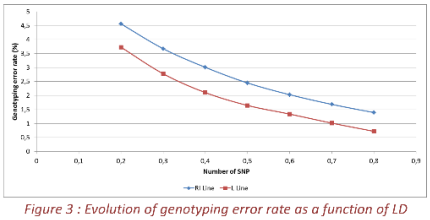


Figure 3 : Evolution of genotyping error rate as a function of LD threshold

Choice of the methodology:

- « Equidistant » methodology: Number of SNP proportional to the size of chromosome (Figure 4).
- « LD » methodology: Consideration of the particular structure of chicken species' LD and of the lower persistence of LD on micro-chromosomes than on macro-chromosomes.
  → Necessity of a greater number of SNP on micro-chromosome to cover the whole chromosome (Figure 4).
- « LD » methodology seems to be the most appropriated to get good imputations.

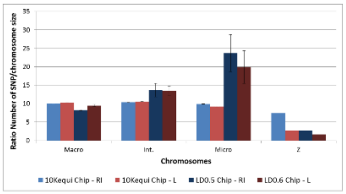


Figure 4 : Evolution of the ratio Number of SNP/Chromosome size as a function of the type of chromosome.

Influence of Minor Allelic Frequencies (MAF):

- Variability of genotyping error rate higher with equidistant methodology than with LD methodology (Figure 5).
- Lower genotyping error rates obtained with LD methodology, except for SNP with high MAF.
- Decrease of genotyping error rates for SNP high MAF with equidistant methodology.
- Valid for both lines.

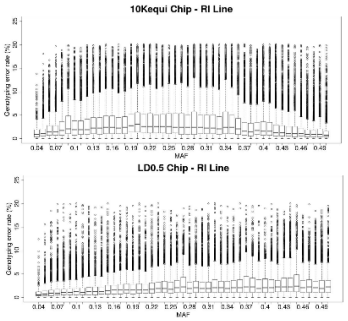


Figure 5 : Evolution of genotyping error rate according to the MAF for Rhode Island Line.

- **Consideration of the particular structure of chicken species' LD with LD methodology.**
- **Better results of imputation obtained with low density SNP chip based on LD.**
- **According to the MAF, lower genotyping error rates obtained with LD methodology, excepted for SNP with high MAF.**

CRÉATION D'UNE PUCE BASSE DENSITÉ POUR LA SÉLECTION GÉNOMIQUE EN POULE PONDEUSE

Herry Florian[1,2], Hérault Frédéric[2], Picard-Druet David[2], Varenne Amandine[1], Burlot Thierry[1], Le Roy Pascale[2] & Allais Sophie[2]

[1]NOVOGEN, 22800 Le Foeil, [2]PEGASE, INRA, Agrocampus Ouest, 35590 Saint-Gilles

INTRODUCTION

Depuis 2013, une puce commerciale de génotypages haute densité (HD) de 600 000 SNP pour l'espèce poule est disponible et permet la mise en place de la sélection génomique pour la filière avicole. Toutefois, les coûts de génotypages avec cette puce restent élevés pour une utilisation en routine sur un grand nombre de candidats à la sélection. Combiner génotypages sur puce basse densité (BD), à un coût plus faible, et imputation des génotypages doit donc être considéré sur un grand nombre de candidats à la sélection. Le choix du panel de SNP BD est une étape importante de ce processus.

L'objectif principal de cette étude est de choisir la stratégie de génotypages BD la mieux adaptée à deux lignées différentes de poules pondeuses.

MATÉRIELS ET MÉTHODES

Populations d'étude: Étude de 2 lignées différentes de *Rhode Island (RI)* et *Leghorn (L)* créées et sélectionnées par Novogen, génotypées sur des puces à SNP HD (Tableau 1). Après contrôle qualité, respectivement 300 351 et 245 667 SNP sont retenus et distribués sur le génome pour les lignées *Rhode Island* et *Leghorn*.

Line	Rhode Island	Leghorn
Nombre	1027	1474
Mâles	1027 ♂	561 ♂
Femelle	0 ♀	913 ♀
Génération 0 (G0)	447	711
Génération 1 (G1)	580	763

Tableau 1 : Populations d'étude

Puces à SNP basse densité:
Simulation de puces à SNP basse densité selon deux méthodologies intra-chromosomes (Tableau 2):

- Méthodologie « Équidistante »: Sélection de SNP à intervalles réguliers.
- Méthodologie « Déséquilibre de Liaison »: Sélection de SNP selon le DL entre SNP.

Stratégies d'imputation: Utilisation de FImpute (Sargolzaei et al., 2014) pour imputer G1 à partir de G0. À partir des puces à SNP BD créées, étude de l'effet:

- De la densité de SNP,
- Du seuil de DL utilisé pour créer les puces à SNP basse densité,
- Du type de chromosome (macro-chromosomes (1 à 5), chromosomes intermédiaires (6 à 10), micro-chromosomes (11 à 33) et chromosome sexuel Z),
- De la MAF (Fréquence Allélique Mineure) des SNP,
- De la méthodologie utilisée pour développer les puces à SNP basse densité.

Méthodologie	Puce à SNP	Nombre de SNP	
		Rhode Island	Leghorn
Equidistante	50Kequi	49636	50307
	40Kequi	40160	39838
	30Kequi	29970	30075
	20Kequi	19910	19948
	15Kequi	14963	14955
	10Kequi	10001	9966
	7.5Kequi	7527	7496
	5Kequi	4991	4996
	4Kequi	4023	4000
	3Kequi	2992	3003
	2Kequi	2013	2003
Déséquilibre de liaison	DL0.8	21717	18052
	DL0.7	16615	13696
	DL0.6	13214	10736
	DL0.5	10711	8626
	DL0.4	8521	6944
	DL0.3	6875	5578
	DL0.2	5371	4330
	DL0.1	3935	3232
	DL0.05	3205	2624

Tableau 2 : Résumé des puces à SNP étudiées

RÉSULTATS ET DISCUSSION

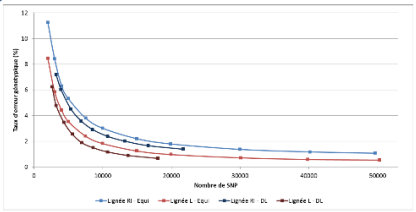


Figure 2 : Évolution du taux d'erreur génotypique en fonction du nombre de SNP sur les puces BD

Influence de la densité de SNP:

- Diminution du taux d'erreur génotypique avec une augmentation du nombre de SNP (Figure 2).
- Valide pour les deux méthodologies.

Influence du seuil de DL:

- Diminution du taux d'erreur génotypique avec une augmentation du seuil de DL (Figure 3).
- En augmentant le seuil de DL, le nombre de SNP sur les puces BD augmente. Cette augmentation du seuil de DL permet également de choisir un SNP bien représentatif de son groupe de SNP.

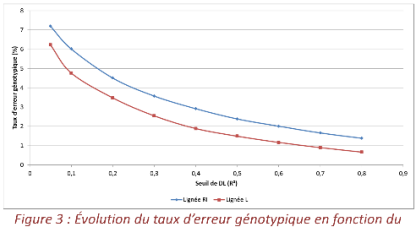


Figure 3 : Évolution du taux d'erreur génotypique en fonction du seuil de DL

Choix de la méthodologie:

- Méthodologie « Équi » : Nombre de SNP proportionnel à la taille du chromosome (Figure 4).
- Méthodologie « DL »: Prise en compte de la structure particulière du DL chez l'espèce poule et de la plus faible persistance du DL sur les micro-chromosomes que sur les macro-chromosomes.
  → Nécessité d'une plus forte densité de SNP sur les micro-chromosomes pour couvrir tout le chromosome (Figure 4).
- La méthodologie « DL » semble être la plus appropriée pour obtenir de bonnes imputations.

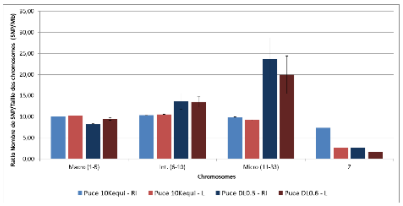


Figure 4 : Évolution du ratio Nombre de SNP/Taille du chromosome en fonction du type de chromosome

Influence de la fréquence allélique mineure (MAF):

- Plus forte variabilité du taux d'erreur génotypique avec la méthodologie « Équi » qu'avec la méthodologie « DL » (Figure 5).
- Taux d'erreur génotypique plus faible avec la méthodologie « DL », sauf pour les SNP à forte MAF.
- Diminution du taux d'erreur génotypique pour les SNP à forte MAF qui sont favorisés avec la méthodologie « Équi ».
- Valide pour les deux lignées.

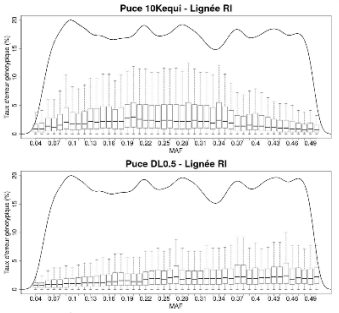


Figure 5 : Évolution du taux d'erreur génotypique et de la distribution des SNP imputés en fonction de la MAF

- **Prise en compte de la structure particulière du DL de l'espèce poule avec la méthodologie DL.**
- **Meilleurs résultats d'imputation avec les puces basse densité développées selon le DL.**
- **Selon la MAF, taux d'erreurs génotypiques plus faible obtenus avec la méthodologie DL, sauf pour les SNP à fortes MAF favorisés avec la méthodologie équidistante.**

DESIGN OF A LOW DENSITY SNP CHIP FOR GENOTYPE IMPUTATION IN LAYER CHICKENS

HERRY F.[1,2], HÉRAULT F.[2], PICARD-DRUET D.[2], VARENNE A.[1], BURLOT T.[1], LE ROY P.[2] & ALLAIS S.[2]

INTRODUCTION

Since 2013, a commercial high-density SNP chip of 600 000 SNP for chicken is available and enables the implementation of genomic selection in layers production. However, genotyping costs still remain high for a routine use on a large number of selection candidates. Combining genotyping on low-density SNP chip, at a lower cost, and genotype imputation should be considered on a large number of selection candidates. Thus, the definition of SNP panel is the milestone of this approach.

The main objective of this study is to choose the best strategy for low density genotyping of laying hen lines in order to optimize selection scheme.

MATERIALS AND METHODS

Study populations: Study of two different lines of *Rhode Island (RI)* and *Leghorn (L)* from Novogen, genotyped on HD SNP chip (Table 1). After quality control, respectively 300 351 and 245 667 SNP are retained and distributed for *Rhode Island* and *Leghorn* lines on the genome.

Line	Rhode Island	Leghorn
Number	1027	1474
Sires	1027 ♂	561 ♂
Dams	0 ♀	913 ♀
Generation 0 (G0)	447	711
Generation 1 (G1)	580	763

Table 1 : Study populations

Low density SNP chips:
Simulation of low-density SNP chip according to two intra-chromosomes methodologies (Table 2):

- « Equidistant » methodology: Selection of SNP at regular intervals.
- « Linkage Disequilibrium » methodology: Selection of SNP according to the LD between SNP.

Imputation strategies: Use of Fimpute (Sargolzaei et al., 2014) to impute G0 from G1. From low density SNP chips designed, study of the effect of:

- MAF (Minor Allelic Frequence) of SNP,
- SNP density,
- LD threshold used to designed low density SNP chips,
- the type of chromosome (macro-chromosomes (1 to 5), intermediate chromosomes (6 to 10), micro-chromosomes (11 to 33) and sexual chromosome Z,
- the methodology used to designed low-density SNP chips.

Methodology	SNP Chip	Number of SNP	
		Rhode Island	Leghorn
Equidistant	50Kequi	49636	50307
	40Kequi	40160	39838
	30Kequi	29970	30075
	20Kequi	19910	19948
	15Kequi	14963	14955
	10Kequi	10001	9966
	7.5Kequi	7527	7496
	5Kequi	4991	4996
	4Kequi	4023	4000
	3Kequi	2992	3003
	2Kequi	2013	2004
Linkage Disequilibrium	DL0.8	21717	18052
	DL0.7	16615	13696
	DL0.6	13214	10736
	DL0.5	10711	8626
	DL0.4	8521	6944
	DL0.3	6875	5578
	DL0.2	5371	4330

Table 2 : Summary of SNP chips studied

RESULTS AND DISCUSSION

Influence of Minor Allelic Frequencies (MAF):

- Variability of mean correlations higher with equidistant methodology than with LD methodology (Figure 1).
- Increase in mean correlations according to the MAF with equidistant methodology.
- Steady mean correlations according to the MAF with LD methodology.
- Valid for both lines.

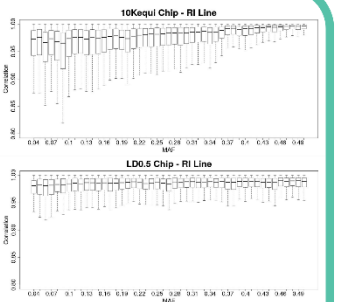


Figure 1 : Evolution of mean correlations according to the MAF for Rhode Island Line.

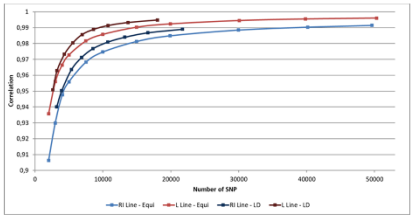


Figure 2 : Evolution of mean correlations between true and imputed SNPs according to the number of SNP on low density SNP chip

Influence of SNP density:

- Increase of mean correlations with an increase in the number of SNP (Figure 2).
- Valid for both methodologies.

Influence of LD threshold:

- Increase of mean correlations with an increase in LD threshold (Figure 3).
- By increasing LD threshold, the number of SNP on low density SNP chip increases. The increase of LD threshold also makes it possible to choose a SNP still more representative of its group.

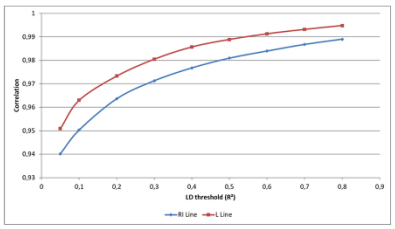


Figure 3 : Evolution of of mean correlations between true and imputed SNPs according to the LD threshold

Choice of the methodology:

- « Equidistant » methodology: Number of SNP proportional to the size of chromosome (Figure 4).
- « LD » methodology: Consideration of the particular structure of chicken species' LD and of the lower persistence of LD on micro-chromosomes than on macro-chromosomes.
  - → Necessity of a greater number of SNP on micro-chromosome to cover the whole chromosome (Figure 4).
- « LD » methodology seems to be the most appropriated to get good imputations.

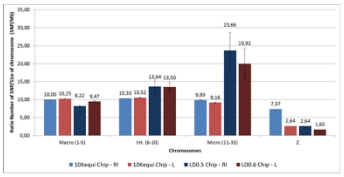


Figure 4 : Evolution of the ratio Number of SNP/Chromosome size as a function of the type of chromosome.

- **According to the MAF, increase in mean correlations with equidistant methodology and steady mean correlations with LD methodology.**
- **Consideration of the particular structure of chicken species' LD with LD methodology.**
- **Better results of imputation obtained with low density SNP chip based on LD.**

XI[th] World Congress on Genetics Applied to Livestock Production – from 11/02/2018 to 16/02/2018 - Auckland

1 NOVOGEN SAS, Les Châtelets, Secteur du Vau Ballier, 5 Rue des Compagnons,22960 PLEDRAN, France
www.novogen-layer.com
2 PEGASE, INRA, Agrocampus Ouest, 16 Le Clos, 35590 Saint Gilles, France
www.rennes.inra.fr

IMPACT OF THE SIZE OF THE REFERENCE POPULATION AND KINSHIP DEGREE ON LOW DENSITY GENOTYPING STRATEGIES FOR GENOTYPE IMPUTATION IN LAYER CHICKENS

BURLOT T.(1), HERRY F.(1,2), HÉRAULT F.(2), PICARD-DRUET D.(2), VARENNE A.(1), LE ROY P.(2) & ALLAIS S.(2)

INTRODUCTION

Since 2013, a commercial high-density SNP chip of 600 000 SNP for chicken is available and enables the implementation of genomic selection in layers production. However, genotyping costs still remain high for a routine use on a large number of selection candidates. Combining genotyping on low-density SNP chip, at a lower cost, and genotype imputation should be considered on a large number of selection candidates. Thus, the definition of SNP panel is the milestone of this approach. The size of the reference population and kinship degree between reference and candidate population are factors influencing imputation accuracy.

The main objective of this study is to choose the best strategy for low density genotyping of laying hen lines in order to optimize selection scheme.

MATERIALS AND METHODS

Study populations: Study of a line of *Rhode Island (RI)* from Novogen, genotyped on HD SNP chip. After quality control, 300 351 SNPs are retained and distributed on the genome. The RI line was constituted of 2362 chickens distributed in four generation (Table 1).

Generation	Total number of individuals	Number of ♂	Number of ♂ breeders	Number of ♀
G0	447	447	132	0
G1	580	580	120	0
G2	794	132	73	662
G3	541	55		486

Table 1: Summary of the RI line

Low density SNP chips:
Simulation of 2 low-density SNP chip according to two intra-chromosomes methodologies:

- « Equidistant » methodology: Selection of SNP at regular intervals (10Kequi SNP chip).
- « Linkage Disequilibrium » methodology: Selection of SNP according to the LD between SNP (LD0.5 SNP chip).

Population scenarios:
Study of nine scenarios with different sizes and kinship degree. The generations constituting reference and candidate populations are detailed on Table 2.

Scenario	(A)	(B)	(C_1)	(C_2)	(D)	(E)	(F)	(G)	(H)
G0		Ref. 1027 ♂				Ref. 1100 ♂	Ref. 447 ♂		Ref. 132 ♂
G1	Ref. 580 ♂				Ref. 653 ♂			Ref. 120 ♂	
G2	Cand. 132 ♂ + 662 ♀	Cand. 132 ♂ + 662 ♀	Ref. 73 ♂	Ref. 73 ♂ + 662 ♀			Cand. 132 ♂ + 662 ♀		
G3			Cand. 55 ♂ + 486 ♀	Cand. 55 ♂ + 486 ♀	Cand. 55 ♂ + 486 ♀	Cand. 55 ♂ + 486 ♀		Cand. 55 ♂ + 486 ♀	Cand. 55 ♂ + 486 ♀

Table 2: Summary of the population scenarios studied

Imputation strategies: Use of Fimpute (Sargolzaei et al., 2014) to impute selection candidates from reference population. Imputation accuracy was assessed as the mean correlations between true and imputed SNPs.
From the 2 low density SNP chips designed, study of the effect of:

- the size of the reference population,
- the kinship degree between reference and candidate populations,
- The presence of dams in the reference population.

RESULTS AND DISCUSSION

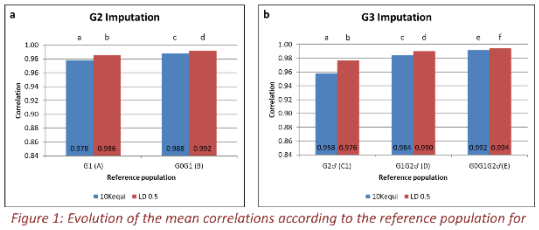


Figure 1: Evolution of the mean correlations according to the reference population for G2 (a) and G3 (b) imputation for both methodologies.

Influence of the size of the reference population:

- Increase in imputation accuracy with an increase in the size of the reference population by cumulating individuals from previous generations (Figure 1).
- Better results with the LD0.5 SNP chip than with the 10Kequi SNP chip.

Influence of the kinship degree between reference and candidate population:

- For G2 and G3 imputations, decrease in imputation accuracy with a decrease of kinship degree (Figure 2).
- Increase in the size of the reference population going from (C1) to (G) did not enable to get better imputations and did not counterbalance the amount of information brought by the direct sires.
- Better results with a gap of two generations (H) than with a gap of one generation (G).
  → Only due to the size of the reference population
  → Results are still lower than results obtained in (C_1)
- Better results with the LD0.5 SNP chip than with the 10Kequi SNP chip.

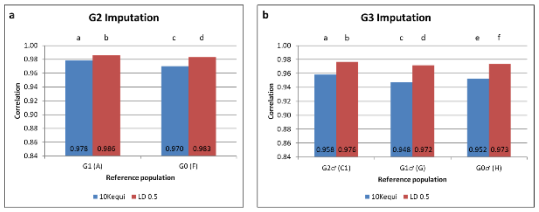


Figure 2: Evolution of the mean correlations according to the reference population for G2 (a) and G3 (b) imputation for both methodologies.

Influence of the presence of dams in the reference population:

- Increase in imputation accuracy with the presence of dams in the reference population (Figure 3).
- Better results with the LD0.5 SNP chip than with the 10Kequi SNP chip.

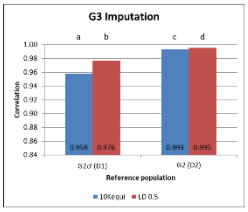


Figure 3: Evolution of the mean correlations with the presence or not of dams in the reference population for both methodologies.

- **An essential key point to get good imputation results was to have in the reference population the direct parents, or at least the direct sires, of the candidate population.**
- **Contribution of the direct parents (or sires) more important than the contribution of the size of the reference population.**
- **LD methodology enabled to get better results than equidistant methodology.**

XI[th] World Congress on Genetics Applied to Livestock Production – from 11/02/2018 to 16/02/2018 - Auckland

1 NOVOGEN SAS, Les Châtelets, Secteur du Vau Ballier, 5 Rue des Compagnons,22960 PLEDRAN, France
www.novogen-layer.com
2 PEGASE, INRA, Agrocampus Ouest, 16 Le Clos, 35590 Saint Gilles, France
www.rennes.inra.fr

INTEREST OF GENOTYPING-BY-SEQUENCING TECHNOLOGIES AS AN ALTERNATIVE TO LOW-DENSITY SNP CHIPS FOR GENOMIC SELECTION IN LAYER CHICKEN: IN-SILICO RESULTS

HERRY F.[1,2], HÉRAULT F.[2], PICARD-DRUET D.[2], BARDOU P.[3], BURLOT T.[1], VARENNE A.[1], LE ROY P.[2] & ALLAIS S.[2]

INTRODUCTION

To reduce the cost of genomic selection, low density single nucleotide polymorphism (SNP) chip can be used in combination with imputation for genotyping the selection candidates instead of using high density SNP chip. Concurrently, next-generation sequencing (NGS) techniques have been increasingly used in livestock species. Nevertheless, they remain expensive to be routinely used for genomic selection. An alternative and cost-efficient solution is to use genotyping-by-sequencing (GBS) techniques to sequence only a fraction of the genome by using restriction enzymes.

The main objective was to study the interest of GBS techniques as an alternative to low density SNP chip for genomic selection in a pure layer line.

MATERIALS AND METHODS

Study population: One line of *Rhode Island (RI)* from Novogen, genotyped on 600K HD SNP chip and with simulated sequences. The RI line was constituted of 1027 individuals distributed in two generations (Table 1).

Genotyping-By-Sequencing technologies simulated:

- Double-digest restriction site-associated DNA sequencing (ddRAD-Seq)
  - ✓ Use of TaqI and PstI
- Genotyping by Genome Reducing and Sequencing (GGRS)
  - ✓ Use of EcoRI, TaqI, AvaII and PstI

Strategies:

- Use of Fimpute (Sargolzaei et al., 2014) to impute the selection candidates from the reference population:
  - ✓ Imputation accuracy assessed as the mean correlation between true HD and imputed HD genotyping.
- Genomic evaluation on ancestry of the selection candidates for egg weight:
  - ✓ Spearman correlation between the results of the selection candidates with true HD or imputed HD genotyping.

Objectives:

- Study of imputation accuracy from SNPs detected with GBS technologies.
- Study of the impact of imputation errors on genomic evaluations.
  - ✓ For the best 150 individuals for egg weight
  - ✓ For the 67 G1 breeders

Reference G0 447 ♂ individuals		Candidates G1 580 ♂ individuals	
HD 300K Genotyping	Phenotypes	HD 300K Genotyping	GBS Genotyping

Table 1 : Data available on the study population

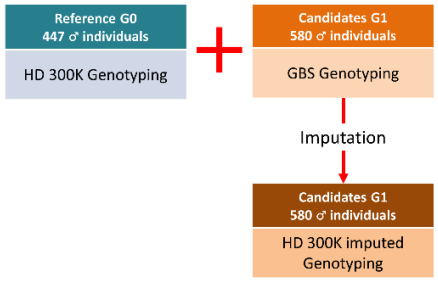


Figure 1: Imputation strategy of the candidate population

RESULTS AND DISCUSSION

	EcoRI	TaqI	TaqI-PstI	AvaII	PstI
Detected SNP	21,267	51,136	128,823	178,980	165,804
SNP in common with the HD SNP chip	1797	4126	11,193	12,453	14,390
Correlation	0.7906	0.9121	0.9691	0.9699	0.9735

Table 2 : Summary of the number of SNP detected and imputation accuracy according to the enzyme

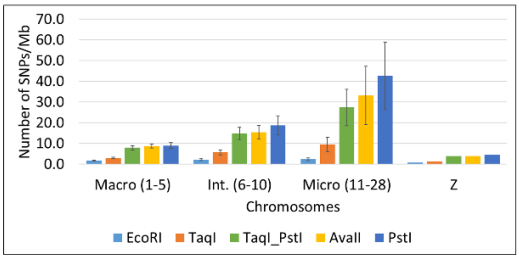


Figure 2: Ratio number of SNPs/Mb depending on the type of chromosome, for each enzyme

Importance of the choice of the restriction enzyme:

- Highly variable number of detected SNP according to the restriction enzyme used (Table 2).
- Reduced number of SNP in common with the HD SNP chip.

Impact of the restriction enzyme on imputation accuracy:

- With EcoRI, low quality imputation with a correlation of 0.7906 (Table 2).
- With ddRAD-Seq, AvaII and PstI, imputation accuracy higher than 0.97.

Distribution of the number of SNP.Mb^{-1} according to the type of chromosome:

- With EcoRI, stable ratio going from 1.6 ± 0.2 SNP.Mb^{-1} to 2.3 ± 0.7 SNP.Mb^{-1} for the macro-chromosomes and micro-chromosomes, respectively (Figure 2).
- For the ddRAD-Seq, increase in the ratio from 8.0 ± 1.0 SNP.Mb^{-1} to 27.4 ± 8.9 SNP.Mb^{-1} for the macro-chromosomes and micro-chromosomes, respectively.
- For PstI, increase in the ratio from 8.9 ± 1.4 SNP.Mb^{-1} to 42.6 ± 16.3 SNP.Mb^{-1} for the macro-chromosomes and micro-chromosomes, respectively.

→ **Excepted for EcoRI, densification of the number of SNP on micro-chromosomes.**

Impact of imputation errors on the ranking of the selection :

- For AvaII, PstI and the ddRAD-Seq, Spearman correlations higher than 0.99 for the results of genomic evaluations based on ancestry of the best individuals and the breeders (Table 3).
- For EcoRI, significantly lower Spearman correlations for the results of genomic evaluations based on ancestry of the best individuals.

- **Promising results with detection of more than 10K SNPs in common with the HD SNP chip with AvaII, PstI and the ddRAD-Seq and good imputation accuracy (correlation higher than 0.97)**
- **Low impact of imputation errors on the ranking of the selection candidates with AvaII, PstI and the ddRAD-Seq.**
- **But these are just results from an in-silico study !**

Enzyme	Number of SNPs	Egg Weight	
		Top150	Breeders
EcoRI	1797	0.8430	0.9450
TaqI	4126	0.9388	0.9914
TaqI-PstI	11,193	0.9913	0.9971
AvaII	12,453	0.9899	0.9975
PstI	14,390	0.9937	0.9980

Table 3: Spearman correlations between the results of genomic evaluations based on ancestry of the selection candidates with true HD or imputed HD genotyping for each enzyme used and for egg weight.

XI[th] European Symposium on Poultry Genetics – from 23/10/2019 to 25/10/2019 - Prague

1 NOVOGEN, 5 Rue des Compagnons, Secteur du Vau Ballier, 22960 PLEDRAN, France
www.novogen-layer.com
2 PEGASE, INRA, Agrocampus Ouest, 16 Le Clos, 35590 Saint Gilles, France
www.rennes.inra.fr
3 SIGENAE, GenPhySE, Université de Toulouse, INRA, ENVT, 24 chemin de Borde-Rouge, Auzeville Tolosane, 31326 Castanet Tolosan, France

Titre : Stratégies de génotypage pour la sélection génomique chez la poule pondeuse

Mots clés : Sélection génomique, puce basse densité, RAD-Seq, NGS, qualité d'imputation, précision d'évaluation génomique

Résumé : Le développement d'une puce à SNP commerciale haute densité (HD) de 600 000 SNP en 2013 a permis la mise en place de la sélection génomique dans les filières ponte et chair. En parallèle, de nouvelles techniques de séquençage NGS permettent d'envisager des solutions autres que les puces à SNP pour la sélection génomique. Toutefois, les coûts de génotypage avec ces outils restent élevés pour les sélectionneurs et ne sont donc pas utilisables en routine pour un grand nombre de candidats à la sélection. Un des enjeux de la sélection génomique est de développer des outils de génotypage ou de séquençage des candidats à la sélection à moindre coût. Puis à partir des génotypes HD d'une population de référence, il est possible avec des méthodes d'imputation de déduire les génotypes HD des candidats. Un premier travail a consisté à étudier l'impact de différents facteurs concernant le développement des puces à SNP basse densité (BD) ou la constitution de la population de référence sur l'efficacité de l'imputation. L'impact de l'utilisation ou non de l'imputation sur l'évaluation génomique des candidats à la sélection a également été étudié. Les résultats montrent qu'une méthodologie équidistante, pour une densité supérieure à 5K SNP, est adaptée pour obtenir de bons résultats d'imputation et une bonne précision d'évaluation génomique. Pour une densité supérieure à 5K SNP, il est également possible d'utiliser les puces BD sans imputation.
Les génotypages BD ont ensuite été remplacés par des génotypages issus de méthodes RAD-Seq. En fonction de l'enzyme de restriction utilisée, les études ont montré que les méthodes RAD-Seq pouvaient être une alternative intéressante aux puces BD.

Title: Genotyping strategies for genomic selection in layer chicken

Keywords: Genomic selection, low density panel, RAD-Seq, NGS, imputation accuracy, genomic evaluation accuracy

Abstract: The development of a commercial high density (HD) SNP chip of 600,000 SNPs enabled the implementation of genomic selection in layer and broiler. Concurrently, new sequencing technologies (NGS) allow to consider other solutions than SNP chip for genomic selection. However, genotyping costs with such tools still remain high for the breeders and cannot be used in routine on a large number of selection candidates. One of the main goal of genomic selection is to develop less expensive tools for genotyping or sequencing the selection candidates. Then, from the HD genotypes of a reference population, it is possible with different imputation methods to deduce the HD genotypes of the candidates. Firstly, the aim was to investigate the impact of different factors concerning the development of low density SNP chips or the constitution of the reference population on imputation accuracy. The impact of the use or not of imputation on genomic evaluation was also studied. The results showed that an equidistant methodology, for a SNP density higher than 5K, is suitable to get good imputation accuracy and good genomic evaluation accuracy. For an SNP density higher than 5K, it is also possible to use low density SNP chips without imputation.
Low density genotypes were then replaced by genotypes from RAD-Seq methods. Depending on the restriction enzyme used, studies have shown that RAD-Seq methods could be an interesting alternative to low density SNP chips.