



HAL
open science

Intégration d'une puce basse densité en sélection génomique ovine laitière

Helene Larroque, Jean-Michel Astruc

► **To cite this version:**

Helene Larroque, Jean-Michel Astruc. Intégration d'une puce basse densité en sélection génomique ovine laitière. [Contrat] 2017. hal-02785230

HAL Id: hal-02785230

<https://hal.inrae.fr/hal-02785230>

Submitted on 4 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



C N B L

Comité National
Brebis Laitières



AI FGE GOLD (Genomic Ovine Low Density)
Intégration d'une puce basse densité en sélection
génomique ovine laitière

RAPPORT FINAL

Convention 615fge043



Rapport scientifique final

Projet GOLD (Genomic Ovine Low Density)

Intégration d'une puce basse densité en sélection génomique ovine laitière

Rédacteurs : H. Larroque (INRA/GenPhySE) et J-M. Astruc (Idele)

NB : Les travaux menés dans le cadre de l'action innovante GOLD (FGE) ont bénéficié de l'apport du projet GENOPYR (région Aquitaine). Le présent rapport scientifique est la synthèse des travaux réalisés.

L'objectif du projet GOLD était d'intégrer des génotypages sur la puce basse densité (LD ou 15K) ovine (conçue par l'International Sheep Genomics Consortium (ISGC)) dans l'évaluation génomique des races ovines laitières françaises : Lacaune, Manech Tête Rousse (MTR), Manech Tête Noire (MTN), et Basco-Béarnaise (BB). La race Corse a également fait partie du projet. L'utilisation de génotypages sur puce LD pour les évaluations génomiques nécessite leur harmonisation avec ceux réalisés plus classiquement sur puce 54K (ou MD). Cette harmonisation n'est possible que par des techniques d'imputation capables de « prédire » les génotypages aux SNP absents de la puce LD.

Le projet a été décomposé en 2 grandes parties :

- d'une part la validation des outils : c'est-à-dire le choix du logiciel d'imputation et l'évaluation de la qualité de l'imputation dans les différentes races françaises, ainsi que l'intégration du processus d'imputation dans la chaîne d'évaluation génomique ovine laitière;
- d'autre part la mise au point d'une stratégie d'optimisation technico-économique d'utilisation de la puce LD par les schémas de sélection : c'est-à-dire la définition d'une stratégie de génotypage des reproducteurs en puce LD et/ou MD, la valorisation des gènes d'intérêt présents sur la puce LD, ainsi que l'élaboration d'une optimisation éventuelle de la puce LD par la définition d'un panel additionnel de SNP.

1. Validation des « outils »

1.1. Choix du logiciel d'imputation et validation de la puce LD

a. Choix du logiciel d'imputation

Au moment de l'écriture du projet d'action innovante plusieurs logiciels étaient utilisés par les équipes travaillant déjà sur le sujet en bovins (tels que Beagle, DAGPHASE, FastPHASE). Les plus performants de ces logiciels présentaient souvent l'inconvénient de temps de calcul très longs. Lorsque les travaux du projet GOLD ont débuté un large consensus dans la communauté scientifique avait établi que le logiciel Flmpute (Sargolzaei et al., 2014) récemment développé permettait une très bonne précision d'imputation dans de nombreuses populations avec des temps de calculs réduits. En France, l'UMT eBIS a testé et comparé 3 logiciels d'imputation DAGPHASE, Beagle et Flmpute pour 3 populations bovines allaitantes (Limousine, Charolaise et Blonde d'Aquitaine) (Saintilan et al., 2014). Pour ces populations, les précisions d'imputation étaient supérieures avec le logiciel Flmpute par rapport au logiciel DAGPHASE et comparables à celles obtenues avec le logiciel Beagle avec des temps de calculs 2 à 3 fois moindres, de plus dans 8 autres populations bovines françaises laitières et allaitantes Flmpute montrait des précisions d'imputation élevées (avec des % de snp correctement

imputés de plus de 99%) (R. Saintilan communication dans le cadre du projet SELGEN/INCoMINGS). De même, le développement de la puce LD ovine avait conduit une équipe de Nouvelle Zélande à procéder à des premiers tests d'imputation dans leurs populations en utilisant le logiciel Flmpute (Ventura et al., 2015). Les performances du logiciel Flmpute sont liées à l'utilisation simultanée d'une information populationnelle, en établissant une librairie d'haplotypes présents dans la population, et également familiale en utilisant judicieusement la connaissance des liens de parentés entre les individus. Suite aux tests de précision d'imputation, l'UMT eBIS a ensuite développé une chaîne d'imputation bovine spécialement adaptée pour le logiciel Flmpute (R. Saintilan communication dans le cadre du projet SELGEN/INCoMINGS).

Au démarrage des travaux du projet GOLD nous avons donc choisi d'utiliser et de tester en priorité le logiciel Flmpute.

b. Evaluation de la qualité de l'imputation

La qualité de l'imputation a été évaluée en fonction de 2 critères : la précision de l'imputation et son impact sur les index génomiques. Cette évaluation a été réalisée par simulation de génotypages sur la puce LD car la puce elle-même n'était pas encore disponible au démarrage du projet GOLD.

Principe général

Le principe général d'estimation de la précision d'imputation est présenté à la figure 1. Pour chaque race, la même procédure est appliquée en s'appuyant sur les typages 54K disponibles dans chaque population. Cependant, seuls les SNP retenus après contrôle qualité pour l'évaluation génomique servent aux décomptes, c'est-à-dire ceux avec un Call Freq (% d'animaux génotypés pour un SNP) supérieur à 0,97, une MAF (fréquence de l'allèle mineur) supérieure à 0,01, en équilibre de Hardy Weinberg, et non impliqués dans trop de problème de filiation. Leur nombre est différent en fonction de la race mais ils sont de l'ordre de 40 000 d'où le nom : typages 40K.

Les génotypages 54k disponibles dans chacune des races sont scindés en 2 populations: une population dite d'apprentissage et une population dite de validation. Seuls les SNP passant les contrôles qualités sont gardés, les génotypages sont donc dans chacune des populations transformés en génotypages 40K. Pour la population de validation, une deuxième transformation est opérée par la sélection des SNP de la puce LD en commun avec les SNP de la puce 54K, soit ici 12153 SNP d'où le nom génotypages 12K. Ces génotypages sont ensuite remis sous le même format que les 40K, c'est-à-dire qu'ils sont expansés : les typages des SNP non présents sont manquants (des trous).

Les génotypages 40K de la population d'apprentissage et 12K expansés de la population de validation sont mis en forme pour le logiciel d'imputation. Flmpute considère l'ensemble de ces typages pour réaliser l'imputation des SNP manquants de la population de validation. Les typages des SNP imputés et vrais sont ensuite comparés selon deux critères: le taux de concordance (pourcentage d'allèles correctement imputés) par animal puis par SNP, et les corrélations entre génotypages vrais et génotypages imputés (r^2 par SNP).

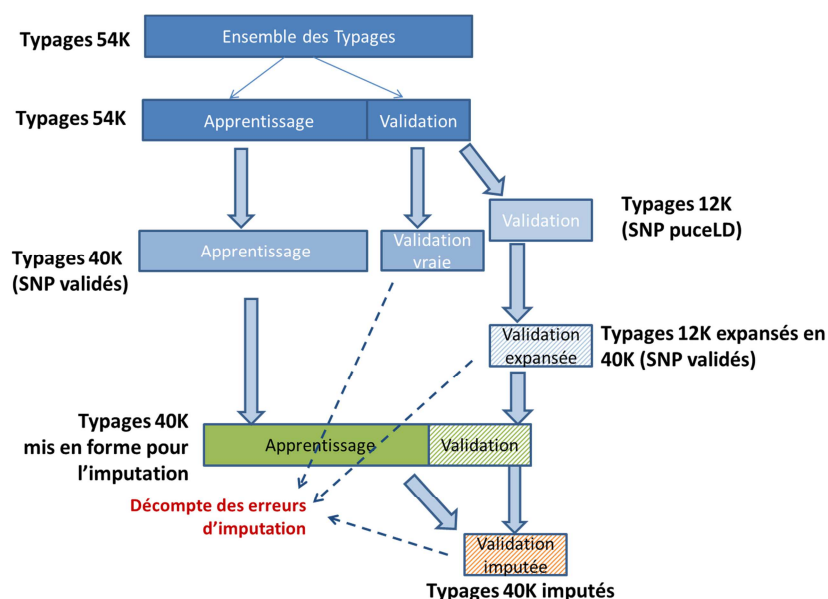


Figure 1 : Principe général d'estimation de la précision d'imputation par simulation

Afin d'estimer l'impact des erreurs d'imputation sur les évaluations génomiques, 2 évaluations ont ensuite été conduites en utilisant pour la population de validation les génotypes vrais ou les génotypes imputés, les valeurs génomiques de cette population dans les 2 cas ont ensuite été comparées.

Moyens mis en œuvre

Les travaux menés dans le cadre du projet GOLD ont fait l'objet du stage de 4 mois (03/04/2017 au 31/07/2015) d'Ostiane de Bourguisson en master 1 « Bioinformatique et Génomique » de l'Université de Rennes 1, et du CDD de 6 mois (du 01/01/2016 au 30/06/2016) de Marjorie Chassier, financé dans la prestation de recherche de l'INRA, et prolongé de 2 mois sur fonds CNBL(01/07/2016 au 31/08/2016).

Lors du stage d'Ostiane de Bourguisson, une première comparaison des puces 54K et LD a été menée sur la base d'une première liste de SNP de la puce LD fournie par l'ISGC : nombre de SNP communs, répartition le long des chromosomes, vérification des MAF dans les populations françaises. Le langage de programmation AWK a été choisi pour la manipulation des génotypes.

Lors du CDD de Marjorie Chassier, la conception générale de la procédure d'estimation de la précision d'imputation a été établie, chaque élément de base a été programmé, un script global et paramétrable de lancement de toute la procédure a été développé permettant une automatisation complète jusqu'au décompte des erreurs d'imputation quel que soit la race. De plus, les fichiers nécessaires aux tests d'évaluation génomique ont été générés, ainsi que les fichiers de paramétrage nécessaires au traitement des génotypes LD par la chaîne de traitement développée par l'UMT eBIS (cf 1.2).

L'action GOLD a été réalisée dans le cadre de l'UMT GGPR et a mobilisé notamment Hélène Larroque (INRA GenPhySE) et Jean-Michel Astruc (Idele). Des contrats de prestation de recherche ont été élaborés entre le CNBL et l'INRA d'une part et entre le CNBL et l'Institut de l'Élevage d'autre part.

Par ailleurs, les responsables et les personnes en charge des schémas de sélection des 5 races ovines laitières ont été mobilisés pour définir avec eux la stratégie d'intégration de la puce LD dès l'année 2017 pour le choix des futurs reproducteurs au 1^{er} semestre. Environ 10 jours de travail par entreprise de sélection ont été consacrés à l'action au cours des années 2015 à mi-2017.

Matériel et méthode

Conception d'une chaîne d'évaluation de la précision d'imputation:

Chaque action sur les génotypages a fait l'objet d'un script élémentaire paramétrable :

- Sélection des génotypages d'une liste d'animaux
- Sélection des typages SNP en fonction d'une liste de SNP choisis
- Expansion des typages d'une petite puce en typages sur une plus grande puce (avec des trous)
- Découpage des génotypages par chromosome
- Mise en forme des fichiers de typages pour Flmpute
- Décompte des erreurs d'imputation
- Production de statistiques par chromosome et tout génome

Le script d'enchaînement a été écrit en bash, les scripts élémentaires ont été principalement écrits en AWK, les statistiques sont produites par des scripts R, la mise en forme pour Flmpute fait intervenir un programme en fortran90.

Critères de mesure de la précision de l'imputation:

Les critères retenus pour estimer la qualité de l'imputation sont ceux traditionnellement calculés dans beaucoup de publications :

- La corrélation de Pearson par SNP au carré entre typages imputés et typages vrais: r^2
- Le taux d'erreur allélique par animal = nombre d'erreurs alléliques / (2 X nombre de SNP imputés)
- Le taux d'erreur allélique par SNP = nombre d'erreurs alléliques / (2 X nombre d'animaux)

dont on peut déduire :

- le taux de concordance par animal = 1- taux d'erreur allélique par animal (concordance rate : CR animal)
- le taux de concordance par SNP = 1-taux d'erreur allélique par SNP (concordance rate : CR SNP)

Les critères par SNP (CR et r^2) ont été moyennés par chromosome, puis sur tout le génome.

Caractéristiques de la puce LD:

La liste des SNP gardés après contrôles qualité pour les évaluations génomiques a été transmise à l'ISGC afin que les SNP les plus informatifs pour les populations Françaises, notamment parce qu'ils avaient une MAF minimum dans ces populations, soient identifiés pour constituer le panel de la puce LD. L'ISGC a établi et communiqué en septembre 2015 un panel pour la puce LD à partir duquel nous avons réalisé les simulations : il comprenait 16301 SNP dont 12399 SNP présents sur la puce 54K (54241 SNP) soit 22% et 3902 SNP particuliers pour l'assignation de parenté et la localisation de mutations causales. Les SNP étaient présents sur tous les chromosomes dont 15992 sur les autosomes 1 à 26 retenus pour l'évaluation génomique.

Pour les races Françaises, en moyenne les SNP de la puce LD ont des MAF plus élevées que les SNP de la puce 54K avant ou après contrôle qualité (Tableau 1). L'analyse de la répartition des SNP a montré une répartition relativement homogène des SNP de la puce LD le long des chromosomes. L'imputation sera faite intra race sur la base des SNP de la puce 54K passant les contrôles de qualité et de filiation, c'est-à-dire 40K en enlevant les SNP pouvant être monomorphes dans une des populations (apprentissage ou validation) mais pas dans l'autre. Le nombre de SNP 40K est le plus faible en race Lacaune (37143) et le plus élevé en race MTN (39128). Les SNP à imputer constituent de 72,78 à 73,56% des SNP de la puce 40K (Tableau 1).

Populations	MAF 50K	MAF 40K	MAF LD	Nombre de SNP 40K	Nombre de SNP LD et 40K	Nombre de SNP à imputer	% de SNP imputés
Lacaune	0,28	0,28	0,34	37143	9822	27321	73,56
Manech Tête Rousse	0,27	0,29	0,32	38156	10362	27794	72,84
Manech Tête Noire	0,26	0,28	0,31	39051	10628	28423	72,78
Basco Béarnais	0,26	0,28	0,31	39128	10638	28490	72,81
Corse	0,28	0,29	0,33	38934	10578	28356	72,83

Tableau 1 : Caractéristiques de la puce LD

Caractéristiques des populations:

Chaque population a été scindée en 2: la population de validation comprend les génotypages des mâles les plus récents (soit de 17% à 25% des génotypages disponibles), celle d'apprentissage comprend les génotypages des mâles les plus anciens parents des mâles de la population de validation (Tableau 2). En race Lacaune et en races ovines laitières de Pyrénées (ROLP) nous avons choisi comme population de validation des populations de mâles relativement anciennes ayant des performances de filles et permettant ainsi par la suite de comparer la précision de validation génomique. En race Lacaune, la population de validation est constituée des mâles nés en 2012 qui ont tous été génotypés (cf. expérimentation génomique du projet Roquefort'in), tous les génotypages antérieurs constituent la population d'apprentissage. En ROLP la population de validation est constituée des mâles génotypés et gardés après testage comme reproducteurs, nés de 2010 à 2012. En race Corse, la population de validation est constituée des génotypages des mâles les plus récents nés en 2016, tous les autres constituent la population d'apprentissage. Les races Lacaune et Manech Tête Rousse possèdent les plus grandes populations de validation mais aussi d'apprentissage. Les autres populations sont beaucoup plus modestes.

Populations	Apprentissage		Validation		% Validation /(Val+App)	% avec pères dans Apprentissage
	Millésimes	Effectifs	Millésimes	Effectifs		
Lacaune	<2012	4648	2012	1144	19,75%	93,30%
Manech Tête Rousse	<2010	1311	2010→2012	430	24,70%	93,70%
Manech Tête Noire	<2010	390	2010→2012	80	17,02%	96,30%
Basco Béarnais	<2010	416	2010→2012	140	25,18%	92,90%
Corse	<2016	492	2016	117	19,21%	100,00%

Tableau 2 : Caractéristiques des populations

Résultats

Précision de l'imputation

En fonction des races, la précision moyenne tout génome de l'imputation varie de 96,81% en race Corse à 99,12% en race Lacaune pour le CR par SNP, et de 82,14% en race Corse à 94,95% en race Lacaune pour le r^2 par SNP. Par animal, le CR varie de 96,69% en race Corse à 99,05% en race Lacaune (Tableau 3). Globalement les niveaux de CR par animal sont très proches de ceux par SNP, par contre les r^2 par SNP sont eux moins élevés.

Populations	CR_anim (%)	CR_SNP (%)	r^2 _SNP
Lacaune	99,05	99,12	94,95
Manech Tête Rousse	98,86	98,95	93,57
Manech Tête Noire	98,49	98,58	90,69
Basco Béarnais	98,72	98,81	92,25
Corse	96,56	96,81	82,14

Tableau 3 : Précision de l'imputation pour chaque race

Quel que soit le chromosome les taux de CR par SNP sont très proches pour les races Lacaune et MTR (Figure 2) et sont de l'ordre de 99%. Pour les races MTN, BB et Corse les fluctuations sont plus importantes entre chromosomes mais ne dépassent pas 1% par rapport à la moyenne tout génome. Clairement la qualité d'imputation en race Corse entre 95,9% et 97,3% est inférieure à celles des autres races qui dépassent pour presque tous les chromosomes 98%. Comme l'illustre la figure 3, certains SNP sont plus difficiles à imputer que d'autres même dans la race présentant les meilleurs taux d'imputation moyens comme la Lacaune. Dans cette race le plus mauvais CR pour un SNP est de 69%, il est de 65% en race Corse.

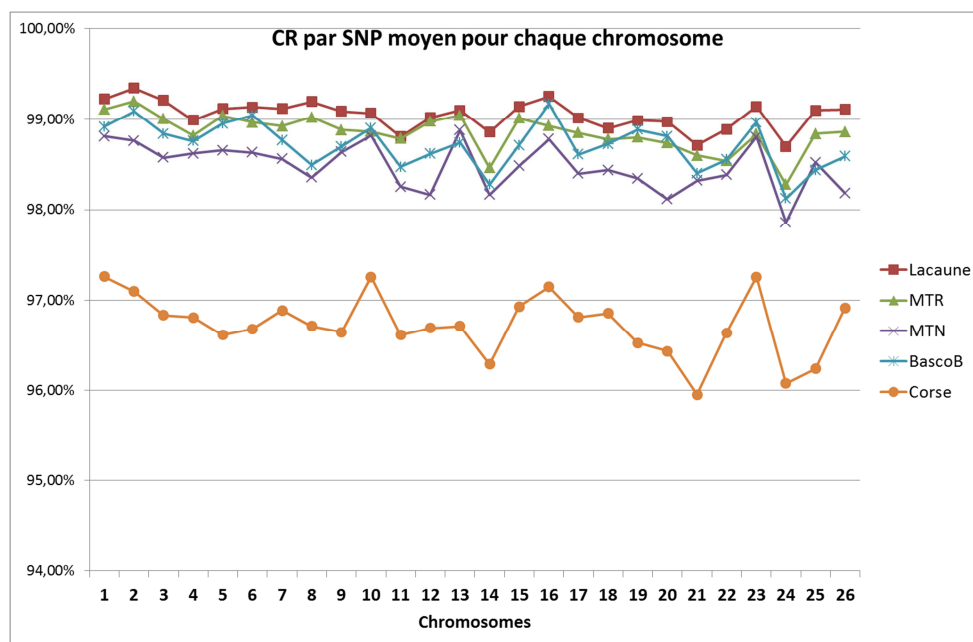


Figure 2 : CR par SNP moyen pour chaque chromosome et par race

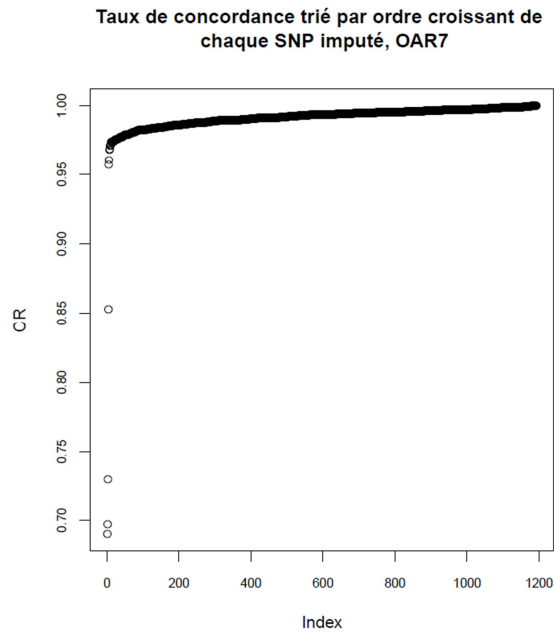


Figure 3 : CR par SNP trié par ordre croissant race Lacaune Chromosome 7

Dans chaque population de validation, les CR par animal sont de même ordre que les CR par SNP. En race Lacaune, le CR d'un seul animal est en-deçà de 98%, alors qu'en race Corse, peu d'animaux ont un CR qui dépasse ce seuil (Figure 4). Pour les races ROLP, la plupart des CR par animal dépassent 98%. Ce seuil de 98% est considéré en France pour les races bovines comme la limite en deçà de laquelle la qualité de l'imputation est insuffisante pour une prise en compte en sélection génomique (R. Saintilan communication personnelle).

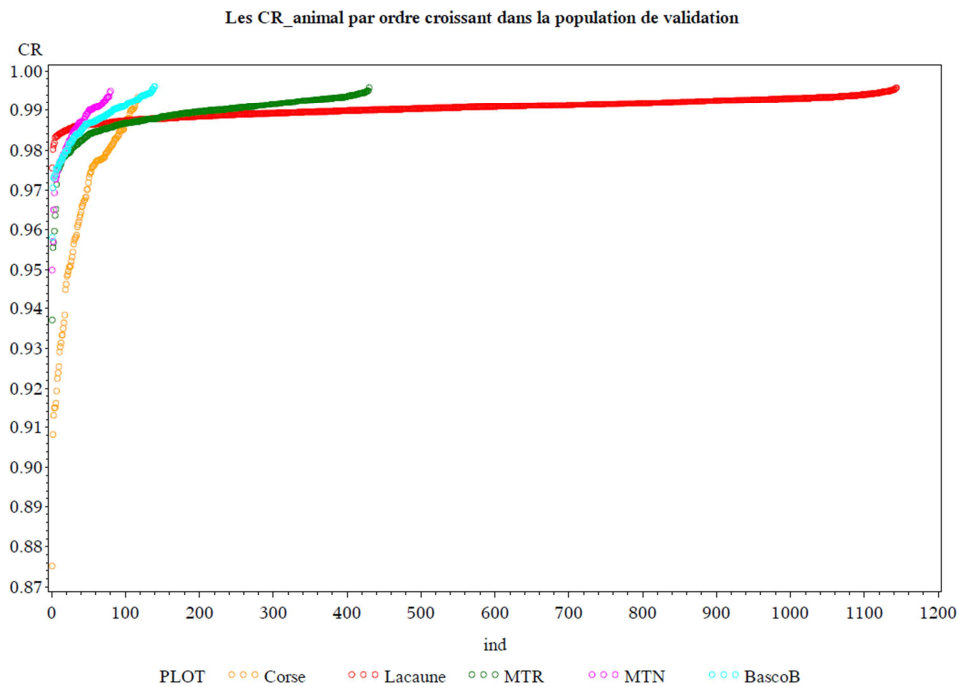


Figure 4 : CR par animal (moyenné par chromosome puis sur tout le génome) en ordre croissant dans les populations de validation

Facteurs influençant la qualité de l'imputation

En dehors du logiciel d'imputation deux facteurs influencent la qualité de l'imputation :

Tout d'abord le choix des marqueurs :

Le nombre de SNP de la puce LD est un facteur important : s'il y a trop peu de SNP il sera plus difficile de « combler les trous » adjacents, s'il y en a trop le prix de la puce sera trop important. Dans l'espèce bovine des puces de basse et très basse densité ont été développées (3K, 6K, 10K). Dans cette espèce les CR par SNP peuvent atteindre les 98% avec des puces 3K et 6K pour les races laitières françaises (Dassonneville et al., 2012) pour lesquelles le déséquilibre de liaison est élevé (en lien avec de forts taux de consanguinité), ce qui n'est pas le cas pour des races bovines allaitantes comme la Blonde d'Aquitaine. Pour la puce LD ovine, le nombre de SNP à imputer utilisé en sélection génomique est de l'ordre de 10K dans nos populations ce qui semble relativement important.

Dassonneville et al. (2012) a aussi montré que la précision d'imputation était supérieure, à même nombre de SNP, lorsque les SNP étaient choisis avec les MAF les plus élevées dans les populations à imputer et en optimisant l'espacement entre SNP. Bien que le nombre de SNP et leur MAF moyenne pour les populations ovines françaises soient relativement assez élevés, nous avons regardé comment se comportaient les CR par SNP et les r^2 le long des chromosomes et en fonction des MAF. Excepté en race Lacaune où les r^2 sont élevés pratiquement quel que soit la MAF, pour les autres races les r^2 sont faibles pour les MAF les plus petites et atteignent un plateau maximum quand les MAF sont de l'ordre de 0,20-0,25 et au-delà. A l'inverse les CR sont plus élevés pour les faibles MAF : il est plus facile de prédire les typages de ces SNP, puis ils diminuent jusqu'à des MAF de 0,25-0,30 où leur diminution s'infléchit mais continue (Figure 5). Comme le montre la figure 6, il existe des zones de moins bonne imputation. Nous les avons observées principalement aux extrémités des chromosomes mais également au centre de certains chromosomes résultant d'une fusion ancienne de 2 chromosomes.

Evolution du CR moyen et du r^2 moyen en fonction de la MAF, tout génome

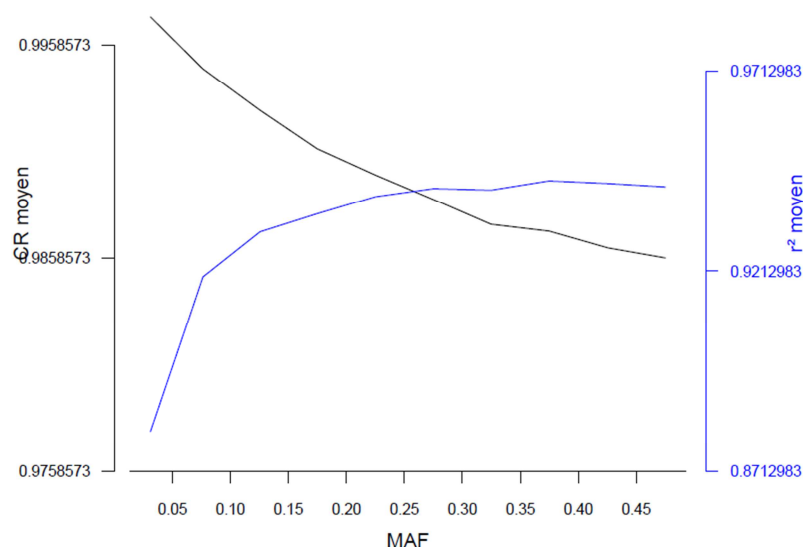


Figure 5 : CR et r^2 par SNP en fonction de la MAF, tout génome, Race Manech Tête Rousse

Estimation de l'exactitude de l'imputation par SNP, via le CR et le r^2 , OAR1

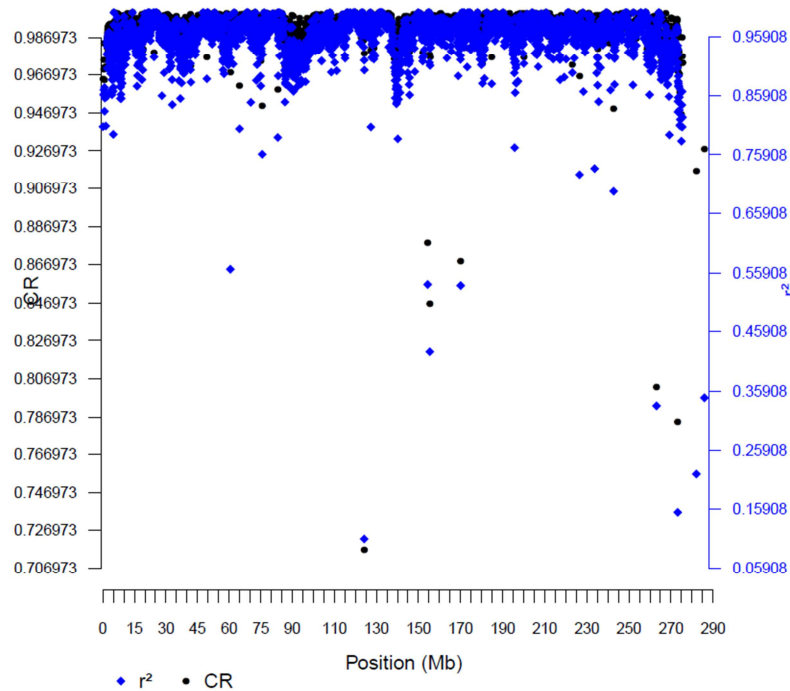


Figure 6 : race Lacaune chromosome 1, CR et r^2 par SNP le long du chromosome

Le deuxième facteur influençant la qualité de l'imputation concerne les caractéristiques des populations: la taille de la population d'apprentissage génotypée en 54K, le niveau de diversité génétique de la population globale, le niveau d'apparentement entre la population d'apprentissage et celle à imputer.

Les tailles des populations d'apprentissage sont fonction des génotypages disponibles eux-mêmes fonction des tailles des schémas de sélection : avec 440 et 140 mâles testés chaque année dans leur schéma classique les races Lacaune et MTR ont génotypé pour le passage à la sélection génomique un grand nombre de mâles et constituer ainsi leurs populations de référence, ce qui nous a permis ici de bénéficier de grandes populations d'apprentissage et de validation pour ces 2 races. Pour les autres races les effectifs sont pour les mêmes raisons beaucoup moins importants.

Une étude sur la diversité génétique (Danchin_Burge et al., 2011) avait montré que les niveaux de diversité génétique sont plus élevés en ovins qu'en bovins en France : cette diversité est la plus faible en race MTN avec une taille efficace de 89, un nombre de fondateurs de 73 et un taux de consanguinité de 2,8%. Ces chiffres doivent être cependant relativisés notamment en regard de la race Holstein dont la taille efficace était de 62, le nombre de fondateurs de 20,7 et la consanguinité de 4,3%. Cette situation, favorable à la diversité génétique ovine, n'est pas favorable à l'imputation, en effet plus les individus d'une population se ressemblent plus il est « facile » de prédire leurs génotypes.

Un niveau d'apparentement élevé entre populations d'apprentissage et de validation augmente la précision d'imputation des génotypes de la population de validation. Dans nos populations, pratiquement tous les jeunes mâles de validation ont un père génotypé dans la population d'apprentissage, de 93% en race Basco Béarnaise à 100% en race Corse (Tableau2). Nous avons

regardé dans quelle mesure l'intégration des pères et grands-pères paternels (GPP) ou maternels (GPM) dans la population d'apprentissage impactait la précision d'imputation de nos populations. En races Lacaune et MTR les pourcentages de pères d'animaux de la population de validation génotypés sont de l'ordre de 93% dans les 2 cas (Tableau 2). En race Lacaune les grands-pères génotypés sont plus fréquents et 55,6% des mâles de validation ont GPM et GPP génotypés contre 38,8% en race MTR (Figures 7a et 7b). En race Lacaune, les animaux de validation sans père génotypé ont les CR par animal parmi les plus faibles (Figure 8a). Ceci n'est pas constaté en race MTR (Figure 8b) où ces animaux se retrouvent dans les différentes classes de CR, mais dans cette race seuls 3 individus n'ont aucun ancêtre génotypés contre 21 en race Lacaune. On peut constater dans les 2 races que les animaux avec père GPP et GPM génotypés sont proportionnellement les plus nombreux dans la catégorie de CR les plus élevés. La connaissance des génotypes des pères et grands-pères jouent donc un rôle dans la précision de l'imputation. Les résultats obtenus en race Corse semblent un peu atypiques : les CR par animal et SNP sont faibles alors que tous les pères sont génotypés. Cependant, dans cette race la diversité génétique semble plus importante que dans les autres races. D'autres études devront être menées dans cette race pour approfondir ces résultats.

Impact sur les évaluations génomiques

Nous avons choisi de réaliser deux évaluations génomiques en utilisant soit les génotypages vrais soit les génotypages imputés des mâles de validation, en nous replaçant à l'automne 2011 en races Lacaune et ROLP lorsque ces animaux avaient un génotypage comme jeunes reproducteurs mais n'avaient pas encore de filles. Cependant, pour les races ROLP la population de validation comprenant les millésimes de mâles nés de 2010 à 2012, en nous plaçant en 2011 certains collatéraux des mâles nés en 2010 pouvaient avoir des filles à l'automne 2011. En race Corse, les mâles de validation étant des mâles très récents nous avons réalisés les évaluations génomiques à partir des dernières performances connues au printemps 2016.

A partir de ces évaluations nous avons mesuré l'impact de l'imputation sur les évaluations génomiques de 2 manières :

- d'une part en comparant les index génomiques « jeunes reproducteurs » des mâles de validations avec typages vrais ou typages imputés. Pour cela des corrélations de Pearson ont été calculées entre index (Tableau 4) ainsi qu'entre coefficients de détermination (CD) pour chaque index (Tableau 5a). Quel que soit la race et le caractère, les corrélations entre index sont très fortes de 0,99 et au-delà montrant que les index sont très peu impactés par l'imputation (Tableau 4).

Caractères	Lacaune n=1127	MTR n=430	MTN n=80	BB n=140	Corse n=117
Lait	0,992	0,998	0,992	0,999	0,998
MG	0,990	0,997	0,999	0,998	
MP	0,989	0,997	0,999	0,998	
TB	0,997	0,998	0,999	0,999	
TP	0,997	0,998	0,999	0,999	
CCS	0,996				
ANG	0,996				
SIL	0,995				
DPLJ	0,996				

Tableau 4 : Coefficients de corrélation de Pearson entre index génomiques sur génotypages vrais et sur génotypages imputés pour les 5 caractères de productions laitières et pour 3 caractères de morphologie mammaire (ANG : angle du trayon ; SIL : profondeur du sillon ; DPLJ : distance plancher jarret).

Les CD sont légèrement plus élevés avec typages vrais que typages imputés, leurs corrélations sont très élevées mais on constate qu'en race Lacaune elles sont légèrement plus faibles surtout pour les caractères de production laitière (Tableau 5a).

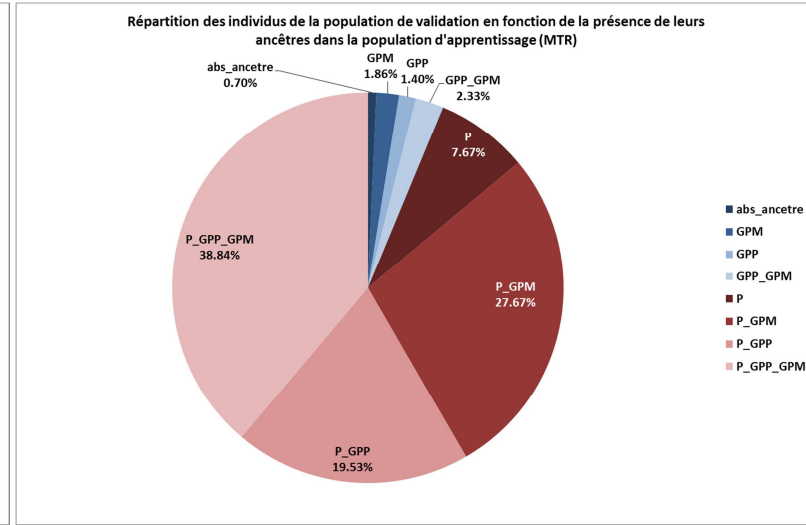
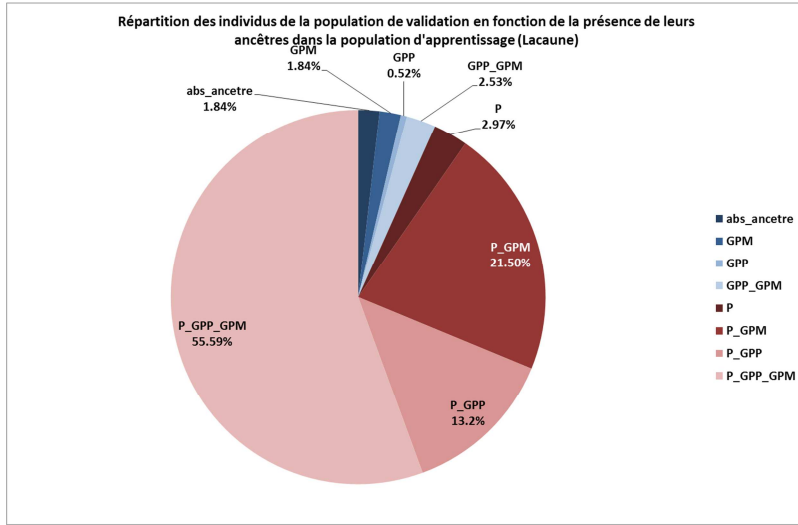
Caractères	Lacaune n=1127	MTR n=430	MTN n=80	BB n=140	Corse n=117
Lait	0,975	0,997	0,999	0,997	0,998
MG	0,975	0,997	0,999	0,997	
MP	0,975	0,997	0,999	0,997	
TB	0,964	0,997	0,998	0,997	
TP	0,939	0,995	0,996	0,993	
CCS	0,995				
ANG	0,978				
SIL	0,985				
DPLJ	0,992				

Tableau 5a : Coefficients de corrélation de Pearson entre coefficients de détermination des index sur génotypages vrais et sur génotypages imputés pour les 5 caractères de production laitière et pou 3 caractères de morphologie mammaire (ANG : angle du trayon ; SIL : profondeur du sillon ; DPLJ : distance plancher jarret)

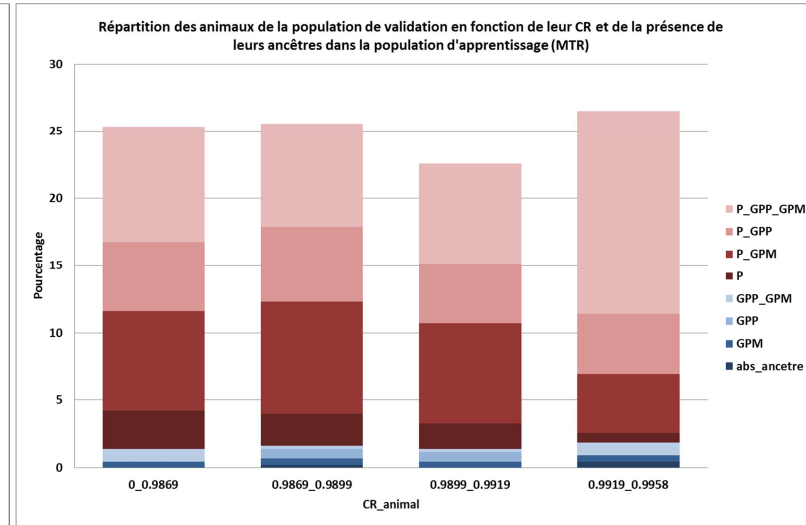
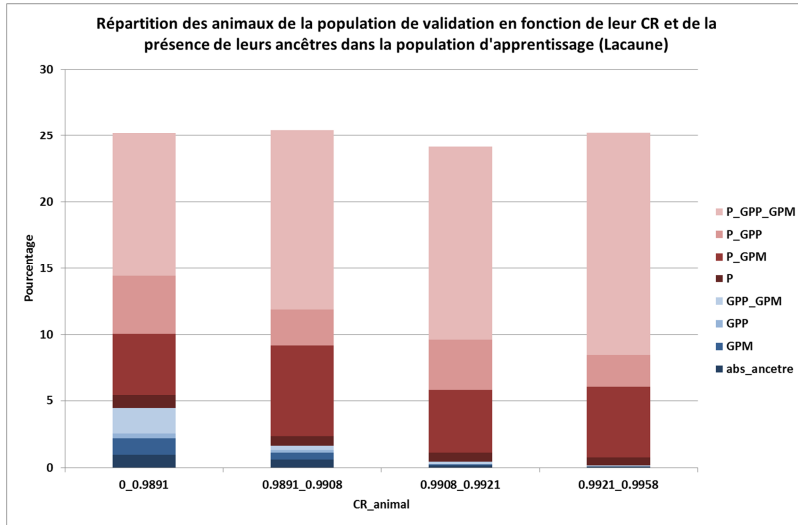
- d'autre part en race Lacaune et MTR, en comparant les précisions de validation génomique c'est-à-dire en comparant les corrélations entre les index « jeunes reproducteurs » avec génotypages (vrais ou imputés) et les « pseudo-performances » pour ces mâles en tant que reproducteurs confirmés (Tableau 5b). Ces « pseudo performances » sont les *daughters yield deviation* (DYD) ou performances moyennes des filles corrigées des effets d'environnement et du niveau génétique de la mère. Les corrélations entre DYD et index génomiques sur typages vrais ou entre DYD et index génomiques sur typages imputés ne montrent pas de différence significative.

	Lacaune n=389		MTR n=427	
	Vrais	Imputés	Vrais	Imputés
lait	0,45	0,45	0,43	0,43
QMG	0,39	0,40	0,36	0,36
QMP	0,37	0,37	0,41	0,41
TB	0,55	0,55	0,57	0,57
TP	0,68	0,68	0,60	0,59
CCS	0,37	0,37		
ANG	0,51	0,51		
SIL	0,47	0,48		
PLJ	0,40	0,40		

Tableau 5b : Coefficients de corrélation de Pearson pondérés par les EDC (*effective daughter contribution*, fonction du nombre de filles) entre DYD (en tant que reproducteur confirmé) et index génomique (en tant que jeune reproducteur) soit avec typages 54K imputés soit avec typages 54K vrais.



Figures 7a et 7b : Races Lacaune et MTR – ancêtres génotypés des mâles de la population de validation



Figures 8a et 8b : Races Lacaune et MTR – classes de CR par animal en fonction des ancêtres génotypés des mâles de la population de validation

1.2. Intégration du logiciel d'imputation dans la chaîne d'évaluation génomique

a. Validation des premiers génotypages LD

Les premiers génotypages sur la puce LD ont été réalisés en Janvier 2017, avec une version de puce légèrement différente de la version de notre étude comportant 1369 SNP de moins et 68 SNP en plus. Afin de valider la lecture des SNP de cette nouvelle puce, Labogena DNA, laboratoire chargé du génotypage, a souhaité qu'une centaine d'animaux déjà typés sur la puce 54K soient génotypés sur la puce LD afin de vérifier que les SNP communs avaient la même lecture de typage.

Un panel de 95 animaux ont été génotypés sur la puce LD. Ce panel était constitué de béliers (15 Corse de millésime 2016, 20 Lacaune de millésime 2016, et 20 béliers de millésimé 2012 pour chacune des races Basco-Béarnaise, MTN et MTR) choisis avec des pères différents et des index variés sur les principaux caractères évalués. Grâce aux scripts développés dans le projet leurs génotypages LD ont été expansés en 54K. Puis un script spécifique a été développé afin de comparer SNP par SNP les typages en 54K et en LD. Sur 11342 SNP présents dans les 2 types de génotypages, 19 présentaient des discordances de lecture de typage entre les 2 puces (sur les 95 individus), dont 4 avaient plus de 4% de discordance: pour 2 d'entre eux leur lecture a été revue, pour les 2 autres leur lecture a été considérée comme trop mauvaise, leur génotypage a donc été annulé.

Le nombre de SNP avec discordance a été considéré comme très faible et donc globalement la lecture de la puce LD comme bonne.

b. Traitement des génotypages : chaîne BDIR_Phase

Les résultats du projet GOLD jusqu'ici ont montré que la puce LD pouvait être utilisée avec une bonne précision d'imputation dans les races ovines Française Lacaune et ROLP pour la sélection génomique. Il restait cependant à intégrer le traitement du contrôle qualité, du contrôle de filiation et de l'imputation de ces génotypages dans la chaîne d'évaluation génomique afin d'homogénéiser ces typages avec ceux sur la puce 54K. Le système de traitement des génotypages 54K déjà en place pouvait être difficilement adapté. De plus, nous devons prendre en considération que dans l'avenir d'autres puces pourraient être développées, il nous fallait donc un système évolutif.

Lors de notre participation au projet INCoMINGS (INRA/Selgen) nous avons pu constater que l'UMT eBIS de Jouy-en-Josas avait développé pour les races bovines une chaîne de traitement combinant contrôle qualité, vérification de filiation et imputation, paramétrable, pouvant gérer plusieurs types de puce et plusieurs races. Romain Saintilan (Alicia/UMT eBIS) après nous avoir expliqué le fonctionnement de la chaîne a proposé de l'adapter à d'autres espèces. Cette chaîne est directement connectée à la base de génotypages SIGENO (gérée par Valogène et hébergée au CTIG), elle est lancée par le CTIG et assure : le contrôle qualité des typages, un contrôle de filiation, la mise en cohérence des typages multiples pour un même animal, l'imputation et le phasage des génotypages, la mise en forme des génotypages pour l'évaluation génomique et des projets de recherche, le suivi des typages dans un fichier résumant pour chaque typage les étapes de son traitement, et elle permettra dans l'avenir l'assignation de parenté. La chaîne SIGENO adaptée aux ovins fournit, pour chacune des races, les fichiers demandés dans la BDIR :

-1 fichier de génotypes au format d'entrée de l'indexation génomique (n° animal, nombre de SNP, liste des génotypes aux SNP (codées 0,1,2)

-26 fichiers de phase (1 par autosome) avec les génotypes sous forme de lettres (A/C/G/T). 2 lignes par individu, une ligne avec les allèles provenant du père, une ligne avec les allèles provenant de la mère.

-26 fichiers de phase (1 par autosome) avec les génotypes sous forme de chiffres (1/2). 2 lignes par individu, une ligne avec les allèles provenant du père, une ligne avec les allèles provenant de la mère. Pour que cette chaîne puisse intégrer les génotypages ovins nous avons établi la liste des SNP de toutes les puces soit 54241 marqueurs de la 54K (dont 11342 communs avec la LD) et les 3658 marqueurs de la LD que l'on ne retrouve pas sur la 54K, et identifiés les SNP retenus pour la sélection génomique c'est-à-dire ceux ayant passé les contrôles qualité lors des 5 dernières évaluations (de Mars 2015 à Juin 2016) soit 38696 SNP. De plus, parmi ces SNP les 227 identifiés pour le contrôle de filiation ont été repérés.

Finalement, cette chaîne a été utilisée en routine à partir de début 2017. Elle a notamment permis d'intégrer les génotypages LD réalisés sur les 2780 jeunes béliers Lacaune et 834 jeunes béliers ROLP de millésime 2017 (Tableau 7).

2. Stratégie d'optimisation technique et économique des schémas de sélection OL intégrant la puce LD

2.1. Stratégie d'intégration en routine

a. Stratégie de génotypage en routine

La question posée était comment articuler les génotypages LD des jeunes béliers sur lesquels on veut appliquer la pression de sélection génomique à 4 mois et la nécessité d'entretenir une population d'animaux avec des génotypages MD afin de garantir la qualité de l'imputation des futures générations.

Le scénario proposé est le suivant : les jeunes béliers candidats, les plus nombreux, sont génotypés avec la puce LD. Les béliers retenus pour l'IA, beaucoup moins nombreux, sont ensuite re-génotypés en puce MD.

Nous avons vu que l'imputation de génotypages LD vers des génotypages MD était précise sur l'ensemble des races françaises. Quel est le gain économique du scénario proposé ?

La sélection génomique est effective en race Lacaune lait depuis 2015 et en races ROLP depuis 2017. Les effectifs de jeunes béliers génotypés sont présentés dans le tableau 7.

Races	Millésime de béliers	Nombre de jeunes béliers candidats génotypés	Nombre de béliers d'IA retenus
Lacaune	2015	1740	270
	2016	2250	270
	2017	2780	270
ROLP	2017	834	270

Tableau 7 : Effectif des béliers candidats à la sélection puis retenus pour diffusion

Les millésimes 2015 et 2016 Lacaune ont été génotypés en puce MD, la puce LD n'étant pas disponible. Le millésime 2017 a été, aussi bien en Lacaune qu'en ROLP génotypé en puce LD.

Nous avons calculé le coût de génotypage pour les 4 situations (Lacaune 2015, 2016, 2017 ; ROLP 2017) dans la situation génotypage puce MD pour tous et la dans la situation génotypages sur puce LD pour les candidats puis sur puce MD pour les retenus. Le bilan économique est présenté dans le tableau 8. Les coûts de génotypage retenus sont les suivants :

- Puce MD : 83 € incluant les différents coûts (puce, génotypage, prestations...)
- Puce LD : 39,50 € incluant les différents coûts (puce, génotypage, prestations...)

Races	Millésime de béliers	Coût génotypage MD (k€)	Coût scénario LD+MD (k€)	Gain économique scénario LD+MD (k€)
Lacaune	2015	144	91	53 (-37%)
	2016	187	111	75 (-40%)
	2017	230	132	99 (-43%)
ROLP	2017	69	55	14 (-20%)

Tableau 8 : Comparaison des scénarios génotypages tout MD ou LD avec re-génotypage MD, bilan économique

En Lacaune, le gain économique apporté par la puce LD a été de 43% en 2017. Si la puce LD avait été disponible en 2015 et 2016, le gain aurait été respectivement de 37 et 40%. Il y a donc un gain économique d'environ 40% avec les effectifs génotypés actuels. En 2017, ce gain a été en partie réinvesti en un nombre supérieur de génotypages, c'est-à-dire en une pression de sélection génomique supérieure.

En ROLP, le gain économique est plus modeste. En effet le nombre de béliers génotypés est inférieur (834 en 2017), alors que le nombre de béliers retenus (270), au moins dans un premier temps, est identique au chiffre Lacaune. Néanmoins, le gain observé en 2017 est de 20% par rapport au coût d'un scénario en puce MD.

Il y a donc un réel bénéfice pour les Entreprises de Sélection à génotyper en puce LD les agneaux candidats.

b. Valorisation des gènes d'intérêt

La puce LD est constituée de 11342 SNP également présents dans la puce MD (dont 9643, soit 85%, sont conservés pour l'évaluation). Ces SNP sont des SNP neutres, dédiés à la sélection génomique « sans connaître les gènes ». Par ailleurs, la puce LD comprend 3658 SNP non présents sur la puce MD. Ces SNP soit appartiennent à un panel de vérification de la filiation ou d'assignation de la parenté, soit sont des marqueurs ou des mutations causales de gènes ou de caractères d'intérêt. Parmi ces gènes d'intérêt, 2 intéressaient vivement la filière ovine laitière : le gène PrP et le gène SOCS2.

Le gène PrP

Le gène PrP est porté par le chromosome 13. La puce LD contient 23 SNP permettant de donner le génotype au gène PrP. Beaucoup concernent des mutations rares ou qui ne concernent pas les races françaises. Rappelons les mutations importantes qui concernent les populations françaises :

-A136V permettant de distinguer ARQ et VRQ

-R154H permettant de distinguer ARQ et AHQ

-Q171R permettant de distinguer ARQ et ARR

L'allèle sauvage est ARQ. Les autres allèles sont constitués d'une seule mutation (de A ou R ou Q).

Malheureusement, selon les informations données par McEwan (AgResearch, Nouvelle-Zélande et animateur de l'ISGC) :

-le SNP 15K_OAR_46225630 (A136V) est bien lu,

-le SNP 15K_OAR13_46225714 (R154H) n'est pas bien lu (il fait l'objet de compression importante)

-le SNP 15K_OAR13_46225765 (R171Q) n'est pas bien lu (il fait l'objet de compression et n'est pas utilisable). Notons qu'il y a 2 SNP liés dans la puce LD : 15K_OAR13_46225764 et 15K_OAR13_46225765. C'est peut-être cette proximité qui génère des interférences et rend impossible le génotypage.

Or ce dernier SNP est stratégique car il permet de distinguer ARR et ARQ.

Nous avons toutefois essayé de vérifier sur les 95 génotypes dont on disposait à la fois de génotype MD et de génotype LD ce qu'il en était. En effet, nous savions d'après le génotypage PrP que 5 parmi les 95 ovins du panel de vérification était hétérozygotes ARR/ARQ, les autres étant homozygotes ARR/ARR. Or lorsqu'on vérifie l'état sur le SNP 15K_OAR13_46225765 (permettant de distinguer ARR et ARQ), sur les 95 ovins, 1 n'est pas lisible et 94 sont hétérozygotes aux SNP. Les résultats lus avec la puce LD sont donc incohérents avec le génotype PrP des animaux. Il ne semble donc pas possible à ce stade d'utiliser la puce SNP pour prédire le génotype PrP des animaux.

Il ne semble donc pas possible, avec la puce LD de déterminer le génotype PrP des ovins.

Une démarche de la part de GenPhySE a été entreprise auprès de John McEwan pour essayer de remédier à ce problème. Pas de réponse positive à ce jour, l'objectif du consortium étant de concevoir une nouvelle puce de moindre densité (et donc moins chère) plutôt que de « corriger » la puce LD commercialisée. Une démarche professionnelle doit être entreprise.

Le gène SOCS2

Le gène SOCS2 est un gène porté par le chromosome 3 dont une mutation est responsable de comptages cellulaires élevées (CCS) associées à une augmentation de la taille, de la croissance et de la production laitière des brebis Lacaune (cf. projet ANR REIDSOCS 2017-2019). Le gène SOCS2 explique 22% de la variance du caractère CCS, il présente 2 allèles C (sauvage) et T (muté). L'allèle T augmente les taux de CCS, sa fréquence dans la population Lacaune a été estimée à 22% dans le programme SheepSNPQTL (il semblerait que la fréquence de l'allèle T ait diminué de moitié au cours des dernières années, probablement sous l'effet de la sélection pour diminuer les CCS, passant de 22% au début des années 2000 à 11% au milieu des années 2010).

Sur la puce LD, la mutation SOCS2 est identifiée par le **SNP 15k_OAR3_129722200**. L'allèle sauvage du SNP lu dans le génotype est G (complémentaire du C). L'allèle muté du SNP lu dans le génotype est A (complémentaire du T). Dans la suite, nous appellerons C l'allèle sauvage et T l'allèle muté.

Nous avons pu étudier le génotype SOCS2 des jeunes béliers de millésime 2017 génotypés en puce LD. Au total, 2780 béliers Lacaune et 832 béliers ROLP ont été génotypés en puce LD. Les résultats sont présentés dans le tableau 9.

Races	Lacaune	ROLP
Nombre de béliers génotypés	2780	832
Nombre de béliers indexables (paternité confirmée et qualité génotypage suffisante) et de génotype SOCS2 connu	2638	791
Nombre de béliers CC	2148	791
Nombre de béliers CT	479	0
Nombre de béliers TT	9	0
Fréquence allélique de T	9,40%	0%

Tableau 9 : Fréquence des génotypes pour le gène SOCS2 dans les races Lacaune et ROLP pour les mâles du millésime 2017

Ces résultats confirment l'absence de la mutation dans les ROLP.

Ils montrent également une fréquence de l'allèle muté en diminution par rapport aux précédentes estimations (21,7% pour les 884 béliers SheepSNPQTL, 13,1% pour les 440 béliers d'IA 2013, 11% pour les 321 béliers d'IA 2015).

Les résultats ont été diffusés aux maîtres d'œuvre de la sélection Lacaune en juin 2017. Désormais, en routine, et sous réserve que le génotypage direct sur SOCS2 prévu prochainement soit cohérent avec le génotype obtenu à partir de la puce LD, les génotypes SOCS2 pourront être diffusés en temps réel, à l'arrivée des génotypes, de façon à retenir les béliers pour l'IA en connaissance du génotype SOCS2.

2.2. « Optimisation » de la puce LD

Lors de l'écriture du projet GOLD il avait été évoqué la possibilité de définir un panel additionnel de SNP à ajouter dans la partie « custom » de la puce afin d'améliorer les résultats d'imputation. Etant donné les très bonnes précisions d'imputation obtenues, cette modification de la puce ne semble plus nécessaire.

Conclusion

Les principaux objectifs du projet GOLD ont été atteints :

- la qualité d'imputation des génotypes LD en 54K a été vérifiée dans les races ovines laitières françaises permettant d'ouvrir la voie à l'utilisation en routine de cette puce pour le génotypage des jeunes reproducteurs ;
- les outils de traitement de ces génotypes en routine pour les évaluations génomiques ont été mis au point et complètement intégrés à la chaîne d'évaluation génomique ;
- la stratégie de génotypage en LD et en MD a été définie permettant d'assurer une bonne précision d'imputation ainsi qu'un gain économique non négligeable pour les schémas de sélection ;
- la puce LD semble opérationnelle pour la gestion du gène SOCS2. Le seul problème concerne la gestion du gène PrP qui ne semble pas possible par cette puce en lien avec sa conception.

Au vu des résultats d'imputation nous avons conclu qu'il n'était pas nécessaire d'ajouter des SNP à cette puce. Par contre, il nous a semblé intéressant de poursuivre nos investigations sur les zones des chromosomes pour lesquelles la qualité d'imputation était moins bonne. En effet, la superposition des résultats de qualité d'imputation et des cartes génétiques le long de chaque chromosome montre que certaines zones moins bien imputées sont également des zones où le taux de recombinaison est le plus élevé (aux extrémités chromosomiques notamment). Un travail complémentaire est actuellement mené par Morgane Petit, dans sa thèse, afin de vérifier l'existence de cette relation et d'établir de quelle manière l'utilisation des cartes génétiques peut permettre un choix plus judicieux des SNPs d'une puce basse densité, afin de maximiser la qualité de l'imputation, tout en diminuant éventuellement le nombre de SNPs sur ce type de puce. Pour mener cette étude les programmes développés dans le cadre du projet GOLD sont utilisés.

Références


- Danchin-Burge, C. (2011). Compte rendu n°001172004, Idele, collection résultats, juin 2011.
- Dassonneville et al., 2012. J. Dairy Sci. 95 : 4136–4140
- Saintilan R. et al., 2014. Journées 3R, 241-244
- Sargolzaei M. et al., 2014. BMC Genomics, 15:478
- Ventura R.V. et al., 2015. Proc. Assoc. Advmt. Breed. Genet. 21: 302-305

Communications

Le projet et ses résultats ont été présentés à différentes reprises :

- Présentation du projet à l'équipe GeSPR – 09/05/2016
- Assemblée générale du CNBL – 28/09/2016
- Communication dans le cadre du projet INCoMINGS (INRA/Selgen) – 10/10/2016
- Poster pour les journées de l'unité INRA/GenPhySE – 17&18/10/2016*
- Conseil administration CNBL – 19/10/2016
- Note de synthèse pour l'UMT GGPR – 05/12/2016*
- COPIL projet GENOPYR/CDEO – 31/03/2017
- FGE- Commission ovine du 29/06/2017
- Présentation au congrès EAAP 2017 Tallinn – 28/08/2017*


* : voir en annexes



Intégration d'une puce basse densité en sélection génomique ovine laitière

Marjorie Chassier¹, Jean-Michel Astruc², Hélène Larroque¹
¹ INRA – UMR1388 GenPhySE, F-31326 Castanet-Tolosan, France
² Institut de l'élevage, F-75595, Paris, France

Introduction



Population de référence:
Mâles génotypés et connus sur descendance
Infos génomique + phénotypique des filles

Evaluation génomique

Population des jeunes candidats
à la sélection génotypés
Index génomique plus précis

Les puces SNP l'outil indispensable de la sélection génomique:

- Mieux estimer la valeur génétique (Index) des futurs reproducteurs dès leur naissance
- Puce actuelle: 54 000 SNP → Outil performant mais relativement coûteux
- Développement d'une puce 16 000 SNP (16K, LD= low density)


Objectif du projet GOLD: intégrer des génotypages 16K dans l'évaluation génomique des 5 races ovines laitières

- Harmoniser les typages 54k et 16k → tout en 54k par imputation
- Vérifier la précision de l'imputation
- Quantifier l'impact des erreurs d'imputation sur l'indexation génomique

Vérification de la précision de l'imputation

- **Précision de l'imputation fonction:**
 - Du choix des marqueurs (nombre, répartition)
 - De la population génotypée en 54K (taille, niveau de parenté, relation génétique avec la population en 16K)
 - Du logiciel
- **Principe: simuler une situation réelle**
- ① Scinder les génotypages en 2 populations: Apprentissage (les + anciens) / Validation (les + jeunes)
- ② Simuler des génotypages 16K de la population de validation (à partir des 54K)
- ③ Imputer les génotypages 16K simulés (Logiciel: Fimpute)
- ④ Comparer les 54K vrais et les 54K imputés

	Millésimes	Effectifs	Millésimes	Effectifs
Lacaune	<2012	4648	2012	1144
Manech Tête Rousse	<2010	1311	2010 à 2012	430
Manech Tête Noire	<2010	390	2010 à 2012	80
Basco Béarnais	<2010	416	2010 à 2012	140
Corse	<2016	492	2016	117



① Scinder la population génotypée 54K

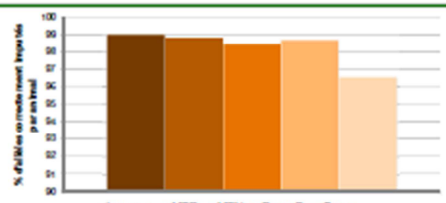
② Simulation des génotypes 16K

③ Imputation des SNP manquants

④ Comparaison

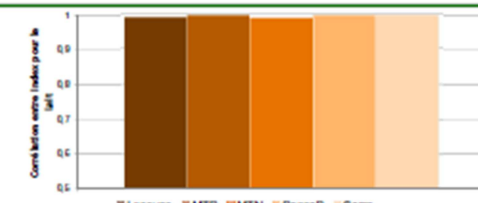
Résultats

Précision de l'imputation:
comparaison des génotypages 54K vrais et 54K imputés
% d'allèles correctement imputés



% d'allèles correctement imputés compris entre 96,68% et 99,12%


Impact des erreurs d'imputation sur l'indexation génomique:
Corrélations entre index sur génotypes vrais et sur génotypes imputés (exemple l'index Lait)





Corrélations du même ordre pour tous les caractères indexés de production ou de morphologie

Conclusion - Perspectives

- Des précisions d'imputation correctes (>98%, seuil utilisé en bovins laitiers) pour les populations ovines laitières de France
- Très faible impact des erreurs d'imputation sur l'indexation génomique
- Perspectives: évolution de la chaîne de traitement des génotypages pour l'indexation, en intégrant l'imputation



Toulouse research center
www.toulouse.inra.fr

Séminaire GenPhySE, Najac 17 et 18 Octobre 2016

Résumé – EAAP 2017

Imputation accuracy from a low density SNP panel in 5 dairy sheep breeds in France
Larroque H. *, Chassier M. *, Saintilan R. ***, Astruc J-M. **

* GenPhySE, Université de Toulouse, INRA, INPT, ENVT, 31326 Castanet-Tolosan, France

** Institut de l'Élevage, 31326 Castanet-Tolosan, France

*** Alice, GABI, domaine de Vilvert, 78350 Jouy-en-Josas, France

In France, genomic selection (GS) has been implemented in dairy sheep in Lacaune breed (LA) and more recently in Pyrenean breeds: Red-Faced Manech (RFM), Black-Faced Manech (BFM) and Basco-Béarnaise (BB). Genomic estimated breeding values (GEBV) are calculated on the basis of SNP50 Bead-Chip (50K) genotypes. Compared to dairy cattle, the relatively higher cost of genotyping limits the cost-effectiveness of GS in dairy sheep. However, a significant reduction of genotyping cost is forecast with the design of a low density (LD) chip by the International Sheep Genomics Consortium. Use of this chip requires harmonizing genotypes on LD panel and those existing on 50K panel by imputing missing SNPs. The aim of the study was to evaluate quality of imputation in 5 French dairy sheep breeds (with Corsica breed: CO) and its impact on GEBV.

Genotypes of rams available in each breed (5,792 in LA, 1,741 in RFM, 470 in BFM, 556 in BB and 609 in CO) were split in two sets: the training set with oldest rams and the validation set with youngest rams (1,144 in LA, 430 in RFM, 80 in BFM, 140 in BB and 117 in CO). The 50K genotypes of the validation population after quality control were "pierced" in order to mimic LD genotypes. Out of 16,331 SNPs from LD panel, 9,822 SNPs were located on 50K panel. These genotypes were then imputed to the 50K panel, from the training population, using the FImpute software. Comparison of 50K imputed and true genotypes permitted to assess imputation accuracy. Concordance rates (CR: percentage of alleles properly imputed) per animal or per SNP were high, ranging from 96.6% to 99.1%, depending on breed. The squared correlations (r^2) between true and imputed genotypes ranged from 81.3% to 95%. On the 3 criteria, the Lacaune breed had the higher scores followed by RFM, BB, BFM and finally CO breeds, in accordance with training population size. For validation rams, correlations between GEBV computed with true or imputed genotypes exceeded 0.99 for production and type traits whatever the breed. Consequently, the 4 breeds in GS have decided to use LD genotypes for pre-selection of young rams.

Action Innovante GOLD (Genomic Ovine Low Density)

Intégration d'une puce basse densité en sélection génomique ovine laitière

Le programme GOLD porté par le Comité National Brebis Laitières (CNBL) est une action innovante bénéficiant d'un financement de France Génétique Elevage (FGE). Ce programme associe l'INRA et l'Institut de l'Élevage associés au sein de l'UMT GGPR et l'ensemble des OS/ES ovines laitières fédérés au sein du CNBL.

L'objectif du projet GOLD était d'intégrer des génotypages sur la puce basse densité (LD, 16K) ovine (conçue par l'International Sheep Genomics Consortium) dans l'évaluation génomique des races ovines laitières françaises : Lacaune, Manech Tête Rousse, Manech Tête Noire, et Basco-Béarnaise. La race Corse a également fait partie du projet. L'usage de puces LD nécessite, pour les évaluations génomiques, de pouvoir harmoniser les génotypages LD et ceux réalisés sur la puce 54k. Pour cela, des techniques d'imputation permettent de rendre compatible ces génotypages avec ceux de la puce 54k en prédisant les SNP [Single Nucleotide Polymorphism] manquants

La première partie de ce projet a consisté à s'assurer de la **qualité de l'imputation**, en fonction de deux critères : la précision de l'imputation et son impact sur la précision des index génomiques.

Pour cela, les génotypages disponibles dans chacune des races ont été scindés en 2 populations: une population dite d'apprentissage et une population dite de validation. Les génotypages 54K, de la population de validation, ont été « dégradés » afin de mimer des génotypages basse densité (LD), puis imputés en génotypages 54K à partir de la population d'apprentissage à l'aide du logiciel Flmpute. Ces génotypages 54K imputés ont été ensuite comparés aux génotypages 54K vrais afin d'évaluer la **précision de l'imputation** par deux critères: le taux de concordance (pourcentage d'allèles correctement imputés) par animal puis par SNP, et les corrélations entre génotypages vrais et génotypages imputés. Pour les différentes races, les taux de concordance par animal varient de 96.56% à 99.05%, et par SNP ils vont de 96.68% à 99.12%. Les corrélations entre génotypages vrais et génotypages imputés varient de 81.34% à 94.95%. Ainsi pour les différentes races les précisions d'imputation sont correctes, notamment pour les 4 races en sélection génomique actuellement ou à l'avenir avec des taux de concordance supérieurs à 98%.

Le deuxième critère d'évaluation de la qualité d'imputation est **l'impact de l'imputation sur la précision des index génomiques**. Pour cela, les valeurs génomiques des animaux de la population de validation ont été calculées en utilisant soit leurs génotypages 54k réels soit leurs génotypages 54k imputés. Les valeurs génomiques (GEBV) calculées dans les 2 cas ont été comparées en calculant leurs corrélations. Pour l'ensemble des races, les corrélations entre GEBV sont élevées : de 0.991 à 0.998 pour le lait, de 0.990 à 0.999 pour la quantité de matière grasse (MG), de 0.989 à 0.999 pour la quantité de matière protéique (MP), de 0.997 à 0.999 pour le taux butyreux (TB), de 0.997 à 0.999 pour le taux protéique (TP). En race Lacaune, ces corrélations pour les caractères de comptages de cellules somatiques du lait (CCS) et de morphologie de la mamelle vont de 0.995 à 0.996. De même, les corrélations entre coefficients de détermination des évaluations (CD : précision des évaluations) varient de 0.975 à 0.999 pour le lait, de 0.975 à 0.998 pour la MG, de 0.975 à 0.998 pour la MP, de 0.963 à 0.998 pour le TB, et de 0.939 à 0.996 pour le TP. Pour les index CCS et morphologie de la mamelle, en race Lacaune, les corrélations entre CD varient de 0.978 à 0.995. Globalement, l'évaluation génomique semble très peu impactée par l'imputation. L'ensemble des résultats est donc favorable à l'intégration des génotypages sur puce basse densité dans l'évaluation génomique.

La qualité de l'imputation étant considérée comme correcte, la deuxième partie du projet a consisté à faire évoluer la chaîne actuelle de traitement des génotypages pour l'évaluation génomique en ovins laitiers et à définir avec les ES ovines laitières les modalités d'intégration de la puce LD dans la sélection génomique ovine laitière. Il s'agissait : (i) en amont de l'évaluation, de s'appuyer sur la chaîne de traitement des génotypages intégrant l'imputation développée par l'UMT 3G en bovins, (ii) d'adapter la chaîne d'évaluation génomique ovine laitière disponible au sein de l'UMT GGPR et (iii) de tester la bonne réalisation des évaluations. Les premiers génotypages sur puce basse densité seront réalisés à partir de janvier 2017 pour le choix des jeunes béliers de millésime 2017. Le maintien de la qualité d'imputation demandera de re-génotyper sur puce 54k ceux d'entre eux qui seront gardés pour la diffusion.

