Genotyping, the Usefulness of Imputation to Increase SNP Density, and Imputation Methods and Tools

Florence Phocas

Chapter 4 - Genotyping, the usefulness of imputation to increase SNP density; imputation methods and tools

Florence Phocas

Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350, Jouy-en-Josas, France

florence.phocas@inrae.fr

Running Head: Genotype imputation to increase genomic prediction accuracy

Abstract

Imputation has become a standard practice in modern genetic research to increase genome coverage and improve accuracy of genomic selection and genome-wide association study as a large number of samples can be genotyped at lower density (and lower cost) and, imputed up to denser marker panels or to sequence level, using information from a limited reference population. Most genotype imputation algorithms use information from relatives and population linkage disequilibrium. A number of softwaresfor imputation have been developed originally for human genetics and, more recently, for animal and plant genetics considering pedigree information and very sparse SNP arrays or Genotyping-By-Sequencing data. In comparison to human populations, the population structures in farmed species and their limited effective sizes allow to accurately impute high-density genotypes or sequences from very low-density SNP panels and a limited set of reference individuals. Whatever the imputation method, the imputation accuracy, measured by the correct imputation rate or the correlation between true and imputed genotypes, increased with the increasing relatedness of the individual to be imputed with its denser genotyped ancestors and as its own genotype density increased. Increasing the imputation accuracy pushes up the genomic selection accuracy whatever the genomic evaluation method. Given the marker densities, the most important factors affecting imputation accuracy are clearly the size of the reference population and the relationship between individuals in the reference and target populations.

Key Words: imputation accuracy, imputation error rate, phasing, haplotype, low density, high density, SNP array, genotyping-by-sequencing, sequence

1. Introduction

A major challenge in genome-wide association studies (GWAS) and genomic selection (GS) programs in animal and plant species is the cost of genotyping. Indeed, large numbers of densely genotyped individuals are required to get accurate results thanks to a high SNP density along the genome that constructs strong linkage disequilibrium between SNP and causative mutations ***(1, 2)***. An appealing strategy is to use a cheaper and reduced-density SNP chip with markers optimized for imputation. Imputation is a term that denotes a statistical procedure that replaces the missing values in a data set by some plausible values. Genotype imputation describes the process of predicting genotypes that are not directly assayed in a sample of individuals. While it traditionally refers to the procedure of inferring the sporadic missing genotypes in an assay, it now commonly refers to the process of predicting untyped loci in a study sample genotyped for a marker low density panel (LDP) using observed genotypes in a reference population that has been genotyped for a greater number of loci with a high density panel (HDP) ***(3, 4)***. Genotype imputation is a crucial step in many genomic studies as all existing genotyping methods result in some missing data. Missing genotypes can be imputed in order to reach a 100% genotype call rate in a single assay. Imputation is also applied to combine sample sets genotyped with different marker panels, provided enough overlap exists between panels, to allow simple integration of data and/or meta-analysis of various study results by standardizing the set of targeted markers. Imputation has become a standard practice in modern genetic research to increase genome coverage and improve GS accuracy and GWAS resolution as a large number of samples can be genotyped at lower density (and lower cost) and, imputed up to denser marker panels or to sequence level, using information from a limited reference population.

These low-cost genotyping strategies enable increased intensity of selection through the genotyping of large numbers of selection candidates or increased accuracy of estimated breeding values by expanding the training population ***(5)***. Current applications of GS are typically based on genotypes called from high and low-density SNP array data. However a lot of plant and animal species cannot afford a high development of genomic tools and genotyping-by-sequencing (GBS) has been proposed as an attractive and low-cost alternative to SNP arrays (***6, 7)***, where restriction enzymes are used to focus sequencing resources on a limited number of cut sites. Because GBS makes possible the coverage of large portions of the genome, it may have some potential advantages for GS and GWAS in animal and plant breeding ***(2, 8, 9)***. GBS also helps to avoid ascertainment bias that happens with SNP data array when marker data are not obtained from a random sample of the polymorphisms in the population of interest. Low-coverage GBS followed by imputation has also been proposed as a cost-effective genotyping approach for human disease and population genetics studies. The theoretical sequencing coverage (or depth) is the average number of times (for instance 10-fold referred as 10x) that each nucleotide is expected to be sequenced given a certain number of reads of a given length and the assumption that reads are randomly distributed across the reconstructed genome (Sims et al., 2014). In a proof-of-concept study, ***(10)*** demonstrated that very low coverage in DNA-sequencing (at 0.1–1x), followed by imputation using genotypic data from a reference population (the map of human genome variation established in the framework of the 1000 genomes project), captures almost as much of the common and low-frequency (minor allele frequency in-between 1 and 5%) variation as SNP arrays, and argued that this paradigm could become cost-effective for GWAS as sample preparation and sequencing costs would continue to fall. However, GBS data suffer from a large proportion of missing or incorrect genotype calls, in particular for low-coverage data. With GBS data, genotypes must be called from observed sequence reads that vary between loci and individuals. It is then challenging to accurately call an individual’s genotype when (almost) no reads are generated at a particular locus. Genotype calling accuracy can be increased by imputation, considering the haplotypes of other individuals in the population and detecting shared haplotype segments between individuals ***(11, 12)***.

Several methods and efficient softwares for genotype imputation have been developed over the last decade. Most imputation methods are using a reference population (RP) that is distinct from the target population (TP) although it is preferable that the two populations have similar genetic background. In this case, two categories of methods are used to predict untyped loci, depending whether haplotypes are inferred only from linkage disequilibrium (LD) information between SNP (known as “population-based” imputation), or they are inferred using both LD and pedigree information (known as “family-based” imputation). A third category of imputation methods (known as “free reference panel-based” imputation) does not imply the use of a reference population and is useful for animal and plant species that have less genomic data and tools than the main farmed species and rely on GBS strategies.

Imputation from lower density towards higher density genotype (or sequence) may be thought as a cost-effective strategy to get accurate GS and GWAS, but the accuracy of SNP imputation needs to be assessed by comparing imputed genotypes with true genotypes. Imputation accuracy is measured at the population level as the genotype correct rate (also called concordance rate) or the Pearson correlation between true and imputed genotypes in the target population. Several factors affect imputation accuracy, including the choice of the imputation method, the size of the reference population, the degree of relatedness between the reference and the target populations, the minor allele frequency (MAF) of the SNP being imputed. All these factors as well as the choice of the genomic evaluation model in relation to the number and importance of the quantitative trait loci (QTL) affect the GS accuracy.

The first objective of this review is to give an overview of the imputation methods and the advantages and drawbacks of the associated tools. The second objective is to shed light on how and under which circumstances marker density affects the imputation accuracy and thereby the genomic prediction quality.

2. Imputation methods and tools: advantages and drawbacks

Imputation requires haplotype reconstruction (known as phasing) from genotype data. Haplotype phasing is the result of a statistical inference procedure exploiting patterns of LD between SNPs by modeling haplotype frequencies and local haplotype sharing between individuals to estimate haplotype phases for a number of samples together, often augmented by a reference panel of previously estimated haplotypes ***(3, 13, 14, 15)***.

Haplotypes are needed for both individuals in TP and RP for imputation methods that require a reference population. In that case, the dense genotypes of the RP members is used to build a reference panel of haplotypes that exhibit high LD over a region of tightly linked markers, and use these haplotypes to fill untyped SNP for target individuals genotyped at LDP (Figure 1). The tag SNP that are common to both RP and TP serve as anchors for guiding genotype imputation of unobserved haplotypes within the LD block. Pre-phasing of genotypes in TP has been suggested to speed up the imputation process ***(16)***. To this end, haplotypes are constructed once and stored so they can be used for subsequent imputations. The quality of the phasing in RP is the most important factor for the accuracy of TP haplotypes ***(17)***. Some accurate phasing tools can be used such as SHAPEIT2 for common variants ***(17, 18)*** or its extension SHAPEITR for achieving greater accuracy for rare variants ***(19)***. Most of the widely used phasing methods iteratively update each individual's haplotype estimates conditional upon the current haplotype estimates of all other individuals. When a new reference set with larger numbers of variants and haplotypes is made available, TP need to be reimputed and the computational cost of this can be considerably reduced if target individuals can be 'pre-phased'. Indeed imputation to give the resulting haplotypes is considerably faster without appreciable loss of downstream accuracy when RP and TP are unrelated as it is often the case for human genomic studies ***(17)***. Because pre-phasing can only be effectively implemented in situations where individuals newly genotyped with the high density panel are not closely related to the target individuals, it is not well suited for animal and plant applications where the numbers of markers in the LDP are sparse and the genotypes of parents of young individuals are continually added to RP. In such a case, the use of pre-phased haplotypes will not lead to optimal imputation accuracy for the target individuals ***(20)***.

For the last decade, the increase in the size of RP and in the density of marker panels, on one hand, and the development of GBS technology, on the other hand, have motivated the development of many new computational methods and the optimization of the oldest ones (Table 1). Current imputation methods are making use of a rich palette of computational techniques, including the use of pre-phasing to reduce computational complexity ***(16)***, the use of identity-by-descent (IBD) ***(21, 22)***, haplotype clustering ***(23, 24)*** and linear interpolation ***(25)*** to reduce the state space in haplotype models, and the use of specific reference file formats to reduce size and memory needs ***(23, 25, 26).*** For instance, it is now possible to provide imputation using RP with tens of thousands of individuals as a free web service ***(23)***. Due to this recent and tremendous development of computational strategies, the different imputation algorithms may strongly differ in accuracy (especially for rare variants), computing speed and memory requirement ***(20, 22, 26, 27)***.

When using a reference panel, imputation methods can be broadly divided into population-based methods, which use population LD information ***(28)*** and pedigree-based methods, which use linkage information from close relatives.

2.1. Population-based methods requiring a reference panel

Population-based methods assume that individuals are unrelated. They do not make use of close relationships directly. However, they can still capture close relationships between individuals by finding long shared haplotypes ***(29, 30)***. Long haplotype blocks of individuals in the target population can be phased and imputed using a group of surrogate parents (individuals sharing

IBD regions with the target individuals) instead of true parents ***(29)***. Population-based methods are highly accurate if both number of markers and number of reference individuals are high enough, but they are computationally intensive. In general, population-based imputation methods use a hidden Markov model (HMM) of the full set of typed and untyped loci for each target sample to infer missing genotypes by maximum likelihood optimization, considering that each reference haplotype represents a hidden state path of the HMM ***(4).*** Additionally, SNP tagging-based imputation approaches such as the one proposed in PLINK ***(31)*** carry out genotype imputation using LD information on tag SNP. Specifically, for each SNP to be imputed, the reference haplotypes are used to search for a small set of tag SNPs in the flanking region that forms a local haplotype background in high LD with the target SNP to be imputed.

The most popular imputation algorithms, Beagle ***(32)***, IMPUTE2 ***(22)*** and MaCH ***(3)*** were initially developed for applications in human genetics. Beagle's first two versions (released in 2006-2007) were only dedicated to haplotype phasing and sporadic missing data inference in unrelated individuals ***(32)***. Late 2008, the major release of version 3.0 added phasing of parent-offspring trios and imputation of ungenotyped markers that have been genotyped in a reference panel ***(24)***. The Beagle imputation method constructs a tree of haplotypes and summarizes it in a direct acyclic graph model by joining nodes of the tree based on haplotype similarity in order to cluster haplotypes at each marker. Then Beagle uses a HMM to find the most likely haplotype pairs based on the individual's known genotypes. It works iteratively by fitting the model to the current set of estimated haplotypes and then resampling new estimated haplotypes for each individual using the fitted model. Beagle predicts the most likely genotype at missing SNP from the model that is fitted at the final iteration.

The three popular imputation algorithms, Beagle, IMPUTE and MaCH are currently in their fifth major version (Table 1). Methods are all based on a HMM based pedigree-free imputation approach and have been compared to each other in several studies ***(4, 23, 26, 27)***. Generally speaking they give similar results in terms of accuracy, but computation times and memory requirements vary strongly depending on the versions of the algorithms. In general, the RP in human includes a sample of representative individuals that are unrelated to the target individuals. Genotype imputation must be performed using the largest available RP because the number of accurately imputed variants increases with the RP size. However, one impediment to using larger RP is the increased computational cost of imputation. Therefore, the latest versions of the imputation algorithms are less memory-intensive and more computationally efficient implementations of the original ones with comparable imputation accuracy.

For instance, Minimac4 is the latest version in the series of genotype imputation software - preceded by Minimac3 ***(23)***, Minimac2 ***(33)***, Minimac ***(16)*** and MaCH ***(3)***. Das et al. ***(23)*** showed that Minimac3 was twice as fast that Beagle 4.1 and about 30 times faster than IMPUTE2 or Minimac2 when considering 100 individuals in the target sample and about 30,000 sequenced individuals in the reference panel. In addition, increasing panel size of sequenced indviduals about 30 fold (from ~1,000 to 30,000) increased memory requirement only sixfold while Beagle 4.1, Minimac2 and Impute2 memory requirement increased almost linearly with panel size.

Browning et al. ***(27)*** showed that the Beagle 5.0 computational cost of imputation from large reference panels is drastically reduced compared to Beagle 4.1, IMPUTE4 and Minimac4 when considering 1000 phased individuals in the target sample and 10k, 100k, 1M, and 10M individuals in reference panels, although all methods produce nearly identical accuracy. In addition, Beagle 5.0 has the best scaling of computation time with increasing reference panel size: its computation time is 33 (10k), 123 (100k), 433 (1M), and 5333 (10M) faster than the fastest alternative method.

Recently, a new version IMPUTE5 ***(26)*** has been developped from the initial IMPUTE2 algorithm ***(22)*** that can also scale to RP with millions of samples and appears to be even faster than Beagle 5.1 for such large RP sizes. IMPUTE5 assumes that both the reference and target samples are phased and contain no missing alleles at any site. This method continues to refine the observation made in the IMPUTE2 method, that imputation accuracy is optimized via the use of a custom subset of haplotypes when imputing each individual. It achieves fast, accurate, and memory-efficient imputation by selecting best matching haplotypes using the Positional Burrows Wheeler Transform. The method then uses the selected haplotypes as conditioning states within the IMPUTE HHM. Using a reference panel with 65,000 sequenced haplotypes, ***(26)*** showed that IMPUTE5 was up to 30x faster than Minimac4 and up to 3x faster than BEAGLE5.1, and used less memory than both these methods. They also showed that IMPUTE5 scales sub-linearly with reference panel size: less than twice the initial computation time is required for an increase of 10,000 to 1 million reference haplotypes, because IMPUTE5 is able to utilize a smaller number of reference haplotypes. Therefore at the end of 2020, IMPUTE5 appeared to be the most computationnaly efficient software for population-based imputation handling large reference panels with millions of haplotypes, including ones with unphased and incomplete genotypes.

Finallly, we mention in this section two other programs, GeneImp ***(34)*** and GLIMPSE ***(35)*** that perform genotype imputation to a dense reference panel given genotype likelihoods computed from low coverage (< 1X) sequencing as inputs. Compared to SNP genotyping, low-coverage sequencing data present a different challenge for imputation because we are not certain about any genotypes. It requires a probabilistic representation of the genotypes in the form of genotype probabilities or genotype likelihoods, rather than fixed genotype calling. Imputation is used to refine the genotype likelihoods and to fill in the gaps between the sparsely mapped reads by leveraging information from a large reference panel of thousands of haplotypes, assuming that these haplotypes adequately represent the target haplotypes over short unaltered regions. Most recent versions of the popular imputation algorithms are not well suited for this situation, as they rely on prephasing for computational efficiency, and, without definite genotype calls, the prephasing task becomes computationally expensive. It should be noticed that genotype likelihood input is not supported by the latest versions of Beagle (after Beagle 4.1 which does not scale to RP larger than a few tens of thousand genomes) ***(25)***. GeneImp was shown to achieve imputation accuracy very close to that of Beagle 4.1, but needed one to two orders of magnitude less time for similar memory requirements ***(34)***. GLIMPSE achieved higher imputation accuracy than GeneImp and, in a lesser extent, than Beagle 4.1 for common variants, but it outperformed the two methods with an increased accuracy of more than 20% for rare variants.

2.2. Pedigree-based imputation methods

Pedigree-based imputation consists in the use of HDP genotypes for a subset of individuals in a pedigree to infer genotypes for the remaining relatives genotyped with a LDP. It uses the correlation of genotypes among relatives derived from sharing of IBD genomic segments within pedigree.

While all population-based imputation methods are based upon HMM to model haplotype frequencies and are computationally intensive due to an intensive sampling process under such probabilistic approaches, most of the pedigree-based methods are mainly deterministic, rule-based methods ***(29)*** and thus are less-time consuming. They are reasonably accurate in comparison to population-based methods, especially if the target individuals are genotyped at very low density. In human, two main software were developped using pedigree-based imputation methods: Merlin and GIGI (Table 1). Merlin ***(36, 37)*** relys on a deterministic approach, uses pedigree structure to identify inheritance vectors within a family, then propagates genotypes at high-density markers observed in a subset of individuals to others individuals in the pedigree genotyped at LDP. GIGI (Genotype Imputation Given Inheritance) uses a two-stage procedure to infer inheritance vectors at sparse markers, then uses Markov chain Monte Carlo sampling method to estimate genotypes of a dense marker set ***(38)***.

Animal and plant breeding populations present some interesting advantages for rapid, pedigree-based, imputation. Firstly, they are populations of small effective sizes in comparison to human populations. This limits the number of haplotypes and conserved haplotypes are long within population, which makes haplotype inference easier; all individuals are related and, therefore, share haplotypes which differ in length and frequency based on their relationships. Secondly, there is a large contribution of recent ancestors to the gene pool of each breeding population. Genotyping these ancestors to constitute the reference panel greatly help imputation, as conserved haplotypes from ancestors to present individuals are then very long. So, despite the existence of softwares dedicated to pedigree-based imputation in human, specific methods and softwares were developed for pedigreed animal and plant populations (Table 1). Indeed computing time of algorithms dedicated to human genetics is considered to be incompatible with the very large candidate populations and with the frequent routine genetic evaluation runs in farmed species. Fast and deterministic approaches which make use of family information have been developped for animal and plant breeding, the two most popular algorithms being AlphaImpute ***(39)*** and FImpute ***(20)***.

AlphaImpute involves simple phasing and imputation rules, long-range phasing and haplotype library imputation ***(29)*** as implemented in AlphaPhase1.1 ***(40)***. It uses information from close and distant relatives and from close and distant SNP loci to impute genotypes for individuals for which genotype information may or may not be available, and for individuals which have close or distant relatives densely genotyped. According to ***(39)***, imputation accuracy is greater with AlphaImpute than with IMPUTE2 ***(22)***, the higher accuracy of AlphaImpute over IMPUTE2 increasing with reducing marker density of the LDP. As the marker density of the panel increases, the importance of pedigree information decreases because the likelihood of finding truely shared haplotypes increases, especially for short segments, and increases crossover resolution ***(41)***.

FImpute ***(20)*** was mainly developed for large scale genotype imputation in livestock where hundreds of thousands of individuals are genotyped with different marker panels. Imputation and phasing are more accurate when using information from close relatives (i.e. long haplotypes with usually low frequency) than when using information from distant relatives (i.e. shorter haplotypes with usually higher frequency). Therefore, the key idea of FImpute algorithm ***(20)*** is to exploit the pedigree relationships between individuals by searching for haplotypes from the longest to the shortest. It is worth mentioning that FImpute has an option to impute missing genotypes based on population and/or pedigree information. The importance of pedigree information increases with the decrease of marker density in the LDP. The method starts with family imputation if pedigree information is available, and then exploits close relationships by searching for long haplotype similarities between target and reference individuals using overlapping sliding windows. After each chromosome sweep, the window size is shrunk by a constant factor allowing for shorter haplotype similarity to be taken into account and the search continues in order to capture more distant relationships. The algorithm assums that all individuals are related to each other at different degrees. To speed up the imputation process, FImpute has the capability to use pre-constructed haplotypes. However, for livestock populations, the use of pre-phased haplotypes for imputation is not a recommended option and reducing the reference population to a group of animals that have high genomic relationships with the target individuals might be a better strategy than using pre-constructed haplotypes ***(20)***.

FImpute (version v2) computing requirements are considerably lower than those of Beagle 3.3 and IMPUTE2 ***(20)***. In addition, FImpute gives higher or similar imputation accuracy than Beagle 3.3 and IMPUTE2 in cattle data sets when all available information is used ***(20)***. However these results should be updated to most recent versions of FImpute (v3), Beagle (v5) and IMPUTE (v5). When close relatives of target individuals are present in the reference panel, FImpute results in higher accuracy compared to the other two methods even when the pedigree is not used. Rare variants (e.g. MAF < 0.05) are also imputed with higher accuracy ***(20, 42)***. FImpute imputes rare alleles with high accuracy because it is efficient at finding the long haplotype matches on which rare alleles are most likely located ***(43)***.

Accurate imputation of SNP with rare alleles is important when the imputed genotypes are to be used in GWAS. Rare alleles may contribute substantially to the genetic variance and may account for a substantial part of the so-called “missing heritability” ***(44)***. To identify those rare variants, study of unrelated individuals is not as efficient as family study. Indeed, rare population variants can be frequent in families where a founder has the variant. However, the family-based approach tends to have a lower representation of the global set of rare variants as a limited number of families will be observed at a constant RP size. Pedigree-based methods provide much higher accuracy in calling rare alleles than population-based methods, because explicitly modeling the transmission of IBD genomic segments via the pedigree structure allows rare alleles on such segments to be reliably called. It has been shown that family-based algorithms such as FImpute but also GIGI or MERLIN outperformed population-based approaches such as Beagle 3.3 or IMPUTE2 in calling rare alleles ***(20, 38, 45)***.

2.3. Imputation methods that do not require a reference panel

Most imputation algorithms rely not only on reference panels, but also on physical or genetic maps for ordering SNP and are not suitable for use in species with limited genomic resources. Such species can only rely on GBS technology to perform at the same time SNP discovery, GWAS and GS ***(46, 47, 48)***. Compared to SNP array, it is much challenging to accurately call an individual’s genotype with GBS technologies, specialy when (almost) no reads are generated at a particular locus. Genotype calling accuracy can be increased by imputation, considering the haplotypes of other individuals in the population and detecting shared haplotype segments between individuals ***(11, 12)***. However, the quality of genotypes obtained with GBS tends to be lower than with SNP array since it depends on the genome-wide sequence read depth (*x*). By increasing *x*, the proportion of correctly called genotypes increases but so do the costs. Since *x* varies along each sequenced genome, the number and the quality of genotype calls also vary along the genome of each individual. It complicates the use of GBS data, but can be partially overcome by specific imputation algorithms (Table 1) recently developped to provide powerful new ways to obtain accurate GWAS and GS at lower prices than with SNP arrays.

While methods such as Beagle in its version 4 ***(25)***, findhap ***(49)*** in its version 4 ***(50)*** or GLIMPSE ***(35)*** can be applied for genotype calling and imputation from GBS data, they are tailored to work with reference panels. The first method specifically dedicated to genotype imputation in population samples of any species sequenced at low coverage is named STITCH for Sequencing To Imputation Through Constructing Haplotypes ***(11)***. It is based on HMM, but does not require a haplotype reference panel. However STITCH needs a high-quality reference assembly for read-mapping and SNP ordering, which is still a limiting factor for a large set of animal and plant species. In addition, ***(35)*** pointed out that while STITCH can be used efficiently to capture variation at common variants, its performance drops considerably at rare variants compared to reference-based approaches such as GLIMPSE or Beagle. Recently, ***(51)*** presented a novel deep learning model called SCDA for reference-free genotype imputation based on sparse convolutional denoising autoencoders. This SCDA model seems to achieve good imputation accuracy and to be robust to high levels of missing data and heterogeneity of genotype data. However, as the SCDA is based on a deep learning architecture, training the model is a computationally very demanding process and further developments are still needed to propose more efficient training mechanisms and automatic hyperparameter learning before that kind of algorithms can be efficiently applied to solve large routine genotype imputation issues.

In plant breeding, low-coverage GBS technology has become a cost-effective tool for multiparental populations produced to increase genetic diversity and resolution in QTL mapping ***(52)***. In the last decade, several genotype imputation methods have focused on biparental populations in experimental plant crosses ***(53, 54, 55)***. More recently, ***(56)*** proposed a more general approach for genotype imputation from low-coverage GBS data, applicable to many scenarios in experimental plant crosses where the target individuals are produced by multigenerational crossing from two or more founders. This algorithm is called magicImpute and is based on HMM. It integrates with genotype calling to account for the uncertainties in identifying heterozygous genotypes due to low read numbers (< 1X) in GBS data. The founders of multiparental populations are used as the reference panel for genotype imputation. It applies to both bi- and multiparental populations, realizes parental phasing and can be used even if some founders' genotypes are not available as it particularly happens if both founders and offspring are genotyped by low-coverage sequencing ***(57)***.

Money et al. ***(58)*** introduced LinkImpute, a software package based on a k nearest neighbor genotype imputation method which was designed for unordered markers (no physical nor genetic map required) and for unphased genotype data. LinkImpute exploits the fact that markers useful for imputation often are not physically close to the missing genotype but rather distributed throughout the genome. Using GBS data from diverse and heterozygous accessions of apples, grapes, and maize, ***(58)*** showed that their algorithm has a runtime similar to Beagle 4.0 on all three datasets while achieving slightly better accuracy. However LinkImpute is applied to a table of genotypes that have been called by a genotype calling algorithm and therefore is not using genotypes likelihoods, which limit its interest for low coverage GBS. Money et al. ***(59)*** proposed a new version called LinkImputeR that exploits the read count information and makes use of all available DNA sequence information for the purposes of genotype calling and imputation. They demonstrated that LinkImputeR can significantly improve both the quantity and quality of genotype data generated from next-generation sequencing technologies.

However, all these previous algorithms are not designed to exploit the specific structure of haplotype sharing observed in large outbred full-sib families which is a population structure commonly found in animal and plant breeding programs. In the context of an outbred full-sib family, imputation can be simplified by recognizing that we only need to consider the four parental haplotypes and identify which pair of haplotypes the offspring inherited at each locus. AlphaFamImpute ***(60)*** considers this particular population structure to improve the accuracy of calling, phasing and imputing genome-wide genotypes and to decrease run-time as demonstrated by comparison with Beagle 4.0 ***(25)***. AlphaFamImpute performs imputation using a two-step approach. In the first step, it phases and imputes parental genotypes based on the segregation states of their offspring (i.e. which pair of parental haplotypes the offspring inherited). In the second step, it phases and imputes the offspring genotypes by detecting which haplotype segments the offspring inherited from their parents. AlphaFamImpute achieves a higher imputation accuracy than Beagle 4.0, in both presence and absence of parental GBS data. It was possible to obtain a very high imputation accuracy (> 0.99) when sufficient sequencing resources (> 2x) were spent on the offspring, even if the parents were not sequenced. In addition, the computational costs were strongly decreased: when imputing 100 full-sib families with 100 offspring each, AlphaFamImpute took less than 1 minute for 1000 loci on one chromosome while Beagle 4.0 took 11h for similar memory needs ***(60)***.

3. Factors affecting imputation accuracy and subsequent genomic prediction quality

Empirical evidence from various animal and plant breeding populations ***(52, 61, 62, 63, 64, 65, 66, 67)*** suggest that imputation of low density to higher density genotypes can be highly accurate and that the estimated breeding values (EBV) derived from imputed genotypes can reach similar levels of accuracy to that derived from high density genotypes.

Nevertheless, accuracy of EBV increases when imputation error rate decreases ***(67, 68)***. It is therefore important to define what are the most influential factors affecting the imputation accuracy and, when possible, methods to optimize those characteristics. Both the imputation and GS accuracies depend on: (a) the imputation method; (b) the characteristics of the low-density marker panel with respect to the MAF of the SNP, their number, localization, spacing and linkage between adjacent SNP; (c) the characteristics of the reference population including its size and its relationship and proportion of common genotyped SNPs with the target population; (d) the genomic evaluation method linked to the genetic architecture of the evaluated trait.

3.1. Choice of the imputation method

An optimal imputation strategy for application in animal and plant breeding programs must : (a) allow both ungenotyped and low-density genotyped individuals to be imputed ; (b) functions well in small and large datasets of moderately related individuals; (c) use information from close and distant relatives and from close and distant SNP loci; (d) accurately impute genotypes for all individuals in the pedigree for all SNP (including rare variants) and whatever the position of high-density genotyped individuals in the pedigree; (e) have efficient computing time and memory usage when routine genomic evaluations are required.

Imputation accuracy can be measured as the allele correct rate, the genotype correct rate (also called concordance rate) or the Pearson correlation between true and imputed genotypes in the target population. Genotype error (i.e 1 – concordance rate) is the proportion of genotypes called incorrectly and allele error is the proportion of alleles called incorrectly. Those two rates give similar results although allele error is approximately half of genotype error, because all methods that are likely to impute one allele correctly are unlikely to impute both alleles incorrectly. Those statistics of sample imputation quality can also be derived at the SNP level.

The allele/genotype correct rates are allele-frequency dependent. With a naive imputation procedure based on the most frequent genotypes, the proportion of genotypes correctly imputed approaches 100% as allele frequencies approach zero or one ***(39)***. When considering rare alleles, it is therefore recommended to look at the correlation between imputed and true genotypes rather than to the rate of correct allele/genotype as the latter will always be high when the MAF are low despite the fact that the rare alleles will not be well-predicted ***(39, 42)***.

Browning and Browing ***(24)*** also proposed the squared correlation between the allele dosage (number of minor alleles) of the most likely imputed genotype and the allele dosage of the true genotype as a metrics of imputation accuracy at the marker level. They called this quantity, the allelic R^2. Its interpretation does not depend on allele frequency. Allelic R^2 measures the loss of power when the most likely imputed genotypes are used in place of the true genotypes for a marker. Browning and Browing ***(24)*** showed that allelic R^2 can be estimated from the imputed posterior genotype probabilities without knowledge of the true genotypes, which is an important feature because the true genotype is generally unknown. This internal quality metrics of imputation is given by softwares such as Beagle or Minimac. Another internal quality metrics, the INFO score, is proposed in IMPUTE2. Both imputation quality scores were shown to give highly correlated results ***(25)***. Their values range from 0 to 1, where a higher value indicates increased quality of an imputed SNP. The allelic R^2 and the INFO score can be used for identifying or excluding markers with poor imputation accuracy prior to downstream analysis.

In a study that was independent of any of the co-authors of imputation algorithms, ***(42)*** compared the five most popular imputation algorithms in animal and plant breeding, using SNP array (Beagle 3.3, IMPUTE2, findhap, AlphaImpute and FImpute). Two dairy cattle datasets with low (3K), medium (54K) and high (777K) density SNP panels were used to investigate imputation accuracy, considering about 30% of individuals in the reference panels and relatedness between target and reference individuals. Results demonstrated that the accuracy was always high (allele correct rate > 93%), but lower when imputing from 3K to 54K (93 – 97%) than from 54K to 777K (97 – 99%). IMPUTE2 and Beagle 3.3 resulted in higher accuracies and were more robust under various conditions than the other 3 methods when imputing from 3K to 54K. The accuracy of imputation using FImpute was similar to the ones of Beagle and IMPUTE2 when imputing from 54K to 777K, and higher than findhap and AlphaImpute. Considering computing time and memory usage, FImpute was proposed as a relevant alternative tool to IMPUTE2 and Beagle 3.3. ***(69)*** also investigated the imputation accuracies for dairy cattle when the reference population, genotyped with 50K SNP panel, contained sires, halfsibs, or both sires and halfsibs of the individuals in the target population genotyped with a low density panel using three imputation softwares (FImpute, findhap and Beagle 3.3). They showed that FImpute performed the best in all cases, with correlations between true and imputed genotypes from 0.92 to 0.98 when imputing from sires to their daughters or between halfsibs. Recently a study compared Beagle 4.1 and FImpute for phasing quality ***(70)***. Although similar phasing quality was observed when at least one parent was genotyped and pedigree information was considered for FImpute, ***(70)*** concluded that, since in most actual breeding programs there will be a certain amount of individuals without genotyped parents and progeny, Beagle 4.1 was the most robust and recommendable option for phasing quality, despite a 29 times longer computing time compared to FImpute for their poultry dataset.

Currently, efficient algorithms for imputation of missing genotypes in GBS data are still in their earliest steps of development, especially with regard to very low sequencing read depth (< 1x). Therefore there are yet not enough independent studies from the co-authors of imputation algorithms that can help to define the best algorithms for GBS data imputation. In a recent study, ***(71)*** compared Beagle, IMPUTE2 and FImpute softwares based on simulated GBS data of livestock population. Sequencing read depth varyed between 2 and 10 and different MAF editing criteria (from no lower limit to MAF > 0.03) were investigated. The results showed that imputation accuracies were all low (r < 0.90) for GBS at 2x, but FImpute had a slightly lower imputation accuracy than Beagle and IMPUTE2 at this depth. The three algorithms had similar imputation accuracy of r > 0.95, when the depth of sequencing read depth was ≥ 4x. As the depth increased to 10x, the prediction accuracies approached those using true genotypes in the GBS loci. The authors also analysed the reliability of genomic prediction with the different imputation hypotheses. They concluded that, retaining more SNPs with no MAF limit resulted in higher reliability of genomic prediction.

To sum up, there are nowadays a rich palette of imputation methods and algorithms useful for either low density SNP array or low coverage GBS data, although none of them appears to be efficient for all situation in terms of both genomic ressources (reference assembly genome, density of SNP panels, RP size) and target population structure. In most cases, Beagle and FImpute performed better than other methods. An obvious advantage of FImpute over Beagle is that it uses much less computing time. However comparisons have only been performed with early versions of Beagle. Due to the computational efforts made in the latest version of Beagle (v5.1) and the recent development of specific softwares for GBS data in plant and animal breeding, new comparison studies of imputation quality and computational costs are needeed to help users in choosing the relevant imputation software according to the characteristics of their genotyping datasets.

3.2. Characteristics of the low-density panel and its optimized choice

3.2.1. Characteristics of LDP influencing the imputation accuracy

For all species and study populations, a limit exists upon which increasing the number of SNP in the array used for GS will not induce higher prediction accuracy ***(72, 73, 74).*** The upper bound of GS accuracy is the proportion of the genetic variance which is captured by the array and is determined by the LD between the markers and the causative mutations affecting the trait. Thus this upper limit depends on the genetic architecure of the traits. In wheat, ***(52)*** hypothesized that the limit will be reach at a lower density level for monogenic traits than for polygenic traits for which imputed SNP increased the chances of capturing most of the QTL linked to these traits.

If the major factor affecting the imputation quality of a low-density panel is its number of SNP it is composed of, in relation with the existing LD between adjacent SNP ***(5, 62, 64, 66)***, imputation quality and GS accuracy are also dependent of the MAF and location of tag SNP in the low density panel. The individual SNP imputation accuracy is strongly dependent of the

MAF as reported in maize ***(39)***, sheep ***(76)***, cattle ***(42)***, pig ***(65)*** or salmon ***(77)*.** This is specialy the case for SNP with MAF below 10% that are difficult to correctly impute unless the tag SNP density is sufficient and the size of reference panel is large ***(78)***. Regarding localization along the chromosomes, lower accuracy are generally observed for SNP located at the two end of the chromosomes, in centromeres and more generally in regions with high similarity or high recombination rates (such as HLA/MHC in humans). The telomeres have very long patterns of repeats which generate problems in reads mapping and imputation. Another explanation for the low imputation accuracy is that SNP imputation relies on surrounding markers, but for SNP at the very end of the telomere, surrounding information is only on one side of the chromosome ***(79)***. An additionnal explanation is the fact that recombination is higher around the telomeres, which may decrease the precision of haplotype reconstruction and imputation accuracy ***(61, 65, 79)***. Therefore it is often recommended to increase the number of SNP at the chromosome extremities ***(80)***. ***(81)*** observed that imputation accuracy was positively associated with chromosome size due to the fact that longer chromosomes harbour more markers, and hence provide more information for inferring unknown haplotypes. In longer chromosomes, the problem of low imputation accuracy at the beginning and end of the chromosomes are relatively less important than in shorter chromosomes. Low imputation accuracies have also been observed in some centromere regions ***(61)*** that might be attributed to incorrect order of markers on the reference genome in regions difficult to assemble ***(82)***. By contrast, in other studies the imputation accuracy of SNP in centromere regions was close to 1 ***(65, 79)***.

3.2.2. Optimization of the low-density panel

Several avenues are possible to optimize the design of the low-density chips. In animal and plant breeding, the choice of SNP for low-density arrays is often based on the selection of markers that are uniformly distributed along the genome (equidistant spacing based on physical position along the genome) and that have high MAF to ensure segregation ***(80, 83)***. This strategy was shown to be more relevant than choosing at random the SNP ***(74)***, especially for traits with large-effect QTL for which prediction accuracy crucially depends on capturing specific regions that explain a high proportion of the phenotypic variance. If the optimal choice of SNP in a LDP chip is crucial for the accuracy of genomic prediction only based on low-density genotypes, it also significantly impacts the accuracy of genomic prediction based on high-density imputed genotypes as SNP in the LDP are the only ones that are not subject to imputation errors.

However, it has also been shown that a LD-based strategy could allow more accurate imputation ***(84, 85)*** and that densification of markers at recombination hot spots and telomeres improves accuracy ***(64, 86)***. A mixed strategy combining LD and physical distance has also been proposed to design low-density chips. It consists in LD based marker pruning in user-defined sliding windows.

An alternative strategy is to choose the markers for their effects on the important traits to be improved ***(88, 89, 90)***. Results suggest that a low density panel comprising SNP with the largest effects has the potential to preserve the accuracy of genomic prediction from higher density panels ***(91)***. However, this strategy limits the interest of the genotyping tool to a single population and a limited number of traits with similar genetic architectures ***(83)***.

While arrays with at least 3000 SNP must be used in dairy cattle to obtain mean allelic imputation error rates below 5% ***(66, 89, 92)***, very low density SNP (< 900 SNP) panels and associated cost-effective genotyping tools can be used in populations with higher LD at long distance and close relationship between reference and target populations. This kind of "light" genomic selection was initially proposed in pig and poultry ***(5, 63, 86, 93)*** using panels of ~ 400 SNPs to reduce GS costs with less than 5% loss in prediction accuracy compared to GS using only high density genotyping. Considering the parents of previous generations as reference population reduces the cost of high density genotyping per generation to a few hundred breeding individuals. More recently the interest of this approach has been shown in Atlantic salmon ***(77, 94)*** with extremely low density panels (~200 SNP). When considering 600 SNP in the low density panel, imputation makes it possible to obtain similar accuracy than with the high density panels. The loss of accuracy was small when considering only 200 SNPs and the genotyping cost of the breeding program was reduced by 62% ***(94)***. However, it is not obvious that the same very low density chip allows precise imputation for genetically diverse populations, because the accuracy of imputation depends on the existence of a sufficiently strong linkage between adjacent markers. If as many low density chips have to be developed as there are different populations to be evaluated within a species, then the chip orders cannot be pooled to reduce costs and the economic interest of such technical optimization may vanish.

A last strategy is to exploit GBS data for developing genomic selection in farmed species because it makes it possible to cover large fractions of the genome and to vary the sequence read depth per individual. Gorjanc et al. ***(8)*** quantified by simulation the value of GBS to increase genetic gain, considering three parameters (i) using SNP array genotyping or GBS with sequence read depth (*x*) varying per individual from 0.01*x* to 20*x*; (ii) number of genotyped markers from 3000 to 300 000; and (iii) size of training and validation sets from 500 to 50 000 individuals. The latter was achieved by distributing the total available *x* of 1000*x*, 5000*x*, or 10 000*x* per genotyped locus among the varying number of individuals. Gorjanc et al. ***(8)*** found that accuracies of genomic predictions using GBS data or SNP array data were comparable when large numbers of markers were used and *x* per individual was ~1*x* or higher. The bias of genomic predictions was very high at a very low *x*. When the total available *x* was distributed among the training individuals, the GS accuracy was maximized with the large number of individuals genotyped with low *x* for a large number of loci. Similarly, response to selection was maximized under the same conditions due to increased both GS accuracy and selection intensity.

3.3. Characteristics of the reference population and its optimized choice

3.2.1. Characteristics of RP influencing the imputation accuracy

A crucial component of most genotype-imputation methods is to correctly infer the local haplotypes from reference populations ***(3, 22)***. If a pedigree-free imputation method is used, the most important characteristics of the RP affecting the accuracy of imputation appear to be its size and its ability to capture the genetic diversity of the target population ***(25, 66, 82)***. Whenever a significantly larger reference population becomes available, it is useful to re-impute the target population for subsequent analysis. The size of the RP is less important when pedigree-based imputation is used and the initial RP already includes parents from the TP ***(79)***.

The effect of the size of the RP depends also on the structure of the TP. For a TP of low genetic diversity, few RP individuals are required to achieve a given imputation accuracy because LD is high and individuals derive from a small set of ancestors. The accuracy of imputation for any variant depends on how well individuals of RP match individuals of TP in terms of ancestral haplotypes to be imputed ***(22, 25)***. Therefore, smaller number of animals in RP generally results in lower imputation accuracy, with the difference all the more evident that fewer ancestors are present in the reference population ***(82, 61)***. Reference sets composed of diverse lines very distantly related, as is often the case in plant breeding programs, do not provide highly accurate imputation because, in such cases, individuals share only short chromosome segments and this makes imputation of missing genotypes difficult, especially when TP is genotyped with a very low density panel ***(39)***. Indeed the importance of the size of RP is also strongly dependent on the number of common markers between the HDP and LDP arrays. The benefit of having less missing genotypes in the target panel is higher with fewer individuals in the reference population ***(79).***

3.3.2. Optimization of the reference population

One of the most important factor to optimize the accuracy of genotype imputation in farmed species is the degree of relationship between the individuals in the RP and in the TP. The importance of these genetic relationships has been well documented in various animal species such as cattle ***(42, 66, 82, 92)***, sheep ***(95)***, pig ***(5, 39, 86, 96)***, poultry ***(63, 85, 97)*** and fish ***(61)***. In particular, imputation accuracy strongly increase when parents of the TP are present in the RP ***(61, 63, 66, 92)***. Simulation studies ***(62)*** and ***(98)*** quantified the impact of successive generations of genotype imputation on genomic predictions. Results showed that GS accuracy decays substantially in one or two generations without updating, by a small proportion, the RP to reflect the genetic change in the TP at each generation. ***(62)*** argued that this decay was mainly due to the impact on the genomic estimated breeding values of the increase in genetic distance between TP and RP rather than due to a strong increase in imputation error rate. Indeed, concordance rates only decay by about 0.5% per generation in their study. When the RP was updated by either 1% or 5% of the top animals in the previous generations, decay of GS accuracy was substantially reduced ***(62)***. In addition, ***(98)*** showed that GS accuracy for a trait of moderate heritability was higher using a small reference population of true genotypes than using a larger population of imputed genotypes. But, when the heritability was low (0.03), the accuracy of genomic predictions benefited from a larger RP, even if SNP were imputed. To reduce the accumulation of imputation errors over generations, it is then recommended to routinely generate dense genotypes on influential ancestors.

Another characteristics of RP that can be optimized is the nature of the HDP. As already mentioned in section 3.2.1, the upper bound of GS accuracy is the proportion of the genetic variance which is captured by the SNP panel and is determined by the LD between the markers and the causative mutations affecting the trait. As proposed by ***(99)***, genomic prediction from whole-genome sequence data is attractive, as the accuracy of genomic prediction is no longer bounded by extent of LD between markers and causal mutations affecting the trait as the latter are then in the HDP. Thus a cost-effective strategy can be to sequence a small number of individuals to consititute the RP ***(100)***. The idea is to choose key individuals based on either pedigree relationships or haplotype diversity that maximized the number of unique haplotypes in the RP and that are a subset from the common ancestors of the TP. Based on a Belgian Blue cattle dataset, Druet el al. ***(100)*** investigated the optimum number of individuals to sequence by fold coverage given a maximum total sequencing effort. At 600 total fold coverage (x 600), the optimum strategy was to sequence 75 individuals at eightfold coverage. At a constant sequencing cost, one interesting strategy was to sequence animals at variable fold coverage: key ancestors at x8 to ensure their alleles that are widespread in the population are called correctly, then a larger number of individuals sequenced at only x4 to capture rare alleles. Indeed, compared to dense SNP array genotypes, the use of sequence data increased GS accuracy only when many causal variants had a low MAF. The imputation accuracy of rare alleles could be also improved, by composing the RP with a set of the most common sires, instead of random animals, as it was shown in layer chickens populations ***(97)***.

3.4. Choice of the genomic prediction method

Imputation errors affect the accuracy of all genomic prediction methods. However, probably because LD between SNP and QTL is better exploited by Bayesian methods than by kinship-based methods such as GBLUP ***(101),*** Bayesian methods seem to be more impacted by imputation errors than GBLUP when traits are affected by a few large QTL. For instance, the accuracy of Bayesian prediction methods were reported to be more impacted than the accuracy of GBLUP, for milk fat percentage, a trait affected by a few large QTL in dairy cattle ***(89)***. In this case, inclusion in the LDP of SNP with largest effects substantially improved the accuracy of Bayesian genomic prediction. A similar trend was observed in a simulation study without any imputed genotypes ***(102)***, where the accuracy of genomic prediction from low density panels declined much more rapidly for traits with a smaller number of QTL.

Relative performance of Bayesian and GBLUP methods might be related to the distributions of imputation errors. If more imputation errors are distributed around the QTL, one can assume that Bayesian method may suffer more from these errors than GBLUP because, in genomic regions with a large QTL, Bayesian methods tend to select few relevant SNP surrounding the QTL while GBLUP picks all the SNP. As suggested by ***(89)***, Bayesian methods could suffer more if the few relevant SNP are imputed with error, but GBLUP would suffer from imputation errors accumulated over all SNP. Because the vast majority of economically important traits are complex traits that are controlled by hundreds or thousands of QTL with small effects, the impact of imputation errors on the GBLUP and Bayesian methods is expected to be very similar in most cases.

4 conclusion

Nowadays there is a rich palette of imputation algorithms useful for either low density SNP array or low coverage GBS data, although none of them appears to be efficient for all situation in terms of both genomic ressources and target population structure. Regardless of the imputation method, accuracies of both genotype imputation and genomic selection increase with the relatedness of the target individuals with its denser genotyped ancestors and as their own genotype density increase. At given low and high density SNP panels, the most important factors affecting imputation accuracy are clearly the size of the reference population and the relationship between individuals in the reference and target populations.

References

1. de Roos AP, Hayes BJ, Spelman R J, Goddard ME (2008) Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics 179(3): 1503–1512. doi:10.1534/genetics.107.084301
2. Meuwissen T, Goddard M (2010) Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185(2):623–631
3. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34: 816–834
4. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11:499–511
5. Huang Y, Hickey JM, Cleveland MA, Maltecca C (2012) Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet Sel Evol. 44(1):25-32. doi:10.1186/1297-9686-44-25
6. Baird NA, EtterPD, Atwood TS, CurreyMC, et al. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoSONE3: e3376
7. Davey JW, Hohenlohe.A, EtterPD, Boone JQ et al. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12(7):499-510
8. Gorjanc G, Cleveland MA, Houston RD, Hickey JM (2015) Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet. Sel. Evol. 47:12
9. Crossa J, Jarquín D, Franco J, Pérez-Rodríguez P et al. (2016) Genomic prediction of gene bank wheat landraces. G3 (Bethesda) 6:1819–1834. doi :10.1534/g3.116.029637
10. Pasaniuc B, Rohland N, McLaren PJ, Garimella K, et al. (2012) Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat Genet 44:631-635. doi: 10.1038/ng.2283
11. Davies RW, Flint J, Myers S, Mott R (2016) Rapid genotype imputation from sequence without reference panels. Nat Genet. 48(8):965-969. doi: 10.1038/ng.3594
12. Gorjanc G, Dumasy JF, Gonen S, Gaynor RS, et al. (2017) Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. Crop Sci. 57:1404–1420. doi: 10.2135/cropsci2016.08.0675
13. Browning BL, Yu Z (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am. J. Hum. Genet. 85:847–861
14. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78:629–644
15. Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165: 2213-2233
16. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44:955–959
17. Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. Nat. Methods 9:179–181
18. Delaneau O, Zagury JF, Marchini J (2013). Improved whole-chromosome phasing for disease and population genetic studies. Nat. Methods 10:5–6
19. Sharp K, Kretzschmar W, Delaneau O, Marchini J (2016) Phasing for medical sequencing using rare variants and large haplotype reference panels Bioinformatics 32(13):1974–1980. doi: 10.1093/bioinformatics/btw065

20. Sargolzaei M, Chesnais JP, Schenkel FS (2014) A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478. doi:10.1186/1471-2164-15-478

21. Liu EY, Li M., Wang W, Li Y (2013) MaCH-admix: genotype imputation for admixed populations. Genet. Epidemiol. 37:25–37

22. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5:e1000529

23. Das S, Forer L, Schonherr S, Sidore C, et al. (2016) Next-generation genotype imputation service and methods. Nature Genetics 48(10):1284–1287. doi:10.1038/ng.3656

24. Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J.Hum. Genet. 84:210–223

25. Browning, B.L., and Browning, S.R. (2016). Genotype imputation with millions of reference samples. Am. J. Hum. Genet. 98:116–126

26. Rubinacci S, Delaneau O, Marchini J (2020) Genotype imputation using the Positional Burrows Wheeler Transform PLoS Genet 16(11):e1009049. doi:10.1371/journal.pgen.1009049

27. Browning BL, Zhou Y, Browning SR (2018) A One-Penny Imputed Genome from Next-Generation Reference Panels. American Journal of Human Genetics 103(3):338–348. doi:10.1016/j.ajhg.2018.07.015

28. Li Y, Willer CJ, Sanna S, Abecasis GR (2009) Genotype imputation. Annu. Rev. Genomics Hum. Genet. 10:387–406

29. Kong A, Masson G, Frigge ML, Gylfason A, et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet 40(9):1068–1075

30. Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. Nat Rev Genet 12:703–714

31. Purcell S, Neale B, Todd-Brown K, Thomas L, et al. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575.

32. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81:1084–1097. doi:10.1086/521987

33. Fuchsberger C, Abecasis GR, Hinds DA (2014) minimac2: Faster genotype imputation. Bioinformatics 31:782–784

34. Spiliopoulou A, Colombo M, Orchard P, Agakov F, McKeigue P (2017) GeneImp: Fast Imputation to Large Reference Panels Using Genotype Likelihoods from Ultralow Coverage Sequencing. Genetics. 206(1):91-104. doi: 10.1534/genetics.117.200063

35. Rubinacci S, Ribeiro DM, Hofmeister RJ *et al.* Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat Genet* **53,** 120–126 (2021). https://doi.org/10.1038/s41588-020-00756-0

36. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

37. Burdick JT, Chen WM, Abecasis GR, Cheung VG (2006) In silico method for inferring genotypes in pedigrees. Nat Genet 38:1002–1004

38. Cheung CYK, Thompson EA, Wijsman EM (2013) GIGI: an approach to effective imputation of dense genotypes on large pedigrees. Am. J. Hum. Genet. 92:504–516. doi:10.1016/j.ajhg.2013.02.011

39. Hickey JM, Kinghorn BP, Tier B, van der Werf JH, Cleveland MA (2012) A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol 44:9. doi:10.1186/1297-9686-44-9

40. Hickey JM, Kinghorn BP, Tier B, Wilson JF, Dunstan N, van der Werf JHJ (2011) A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet Sel Evol 43:12

41. Daetwyler HD, Wiggans GR, Hayes BJ, Woolliams JA, Goddard ME (2011) Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 189: 317–327

42. Ma P, Brøndum RF, Zhang Q et al. (2013) Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish red Cattle J. Dairy Sci. 96:4666–4677. doi : 10.3168/jds.2012-6316

43. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108

44. Manolio TA, Collins FS, Cox NJ, Goldstein DB, et al. (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

45. Liu CT, Deng X, Fisher V, Heard-Costa N, et al. (2019) Revisit Population-based and Family-based Genotype Imputation. Scientific Reports 9:1800 .doi :10.1038/s41598-018-38469-4

46. Bastien M, Sonah H, Belzile F (2014) Genome wide association mapping of resistance in soybean with a genotyping-by-sequencing approach. The Plant Genome 7:1-62

47. Wang L, Liu P, Huang S, Ye B, et al. (2017) Genome-wide association study identifies loci associated with resistance to viral nervous necrosis disease in Asian seabass. Marine Biotechnology 19:255-265

48. Dong L, Han Z, Fang M, Xiao S, Wang Z (2019) Genome-wide association study identifies loci for body shape in the large yellow croaker (*Larimichthys crocea*) Aquaculture and Fisheries 4(1):3-8

49. VanRaden PM, O'Connell JR, Wiggans GR, Weigel KA (2011) Genomic evaluations with many more genotypes. Genet Sel Evol 43:10

50. VanRaden PM, Sun C, O'Connell JR (2015) Fast imputation using medium or low-coverage sequence data. BMC Genet. 16:82

51. Chen J, Shi X (2019) Sparse Convolutional Denoising Autoencoders for Genotype Imputation. Genes 10:652. doi:10.3390/genes10090652

52. Nyne M, Wang S, Kiani K et al. (2019) Genotype Imputation in Winter Wheat Using First-Generation Haplotype Map SNPs Improves Genome-Wide Association Mapping and Genomic Prediction of Traits. G3 9:125-133

53. Swarts K, Li H, Romero Navarro JA, An D, et al. (2014) Novel methods to optimize genotypic imputation for low-coverage, next-generation sequence data in crop plants. Plant Genome 7:1–12. doi :10.3835/plantgenome2014.05.0023

54. Hickey, J. M., G. Gorjanc, R. K. Varshney, and C. Nettelblad (2015) Imputation of single nucleotide polymorphism genotypes in biparental, backcross, and topcross populations with a hidden markov model. Crop Sci. 55: 1934–1946.doi : 10.2135/cropsci2014.09.0648

55. Fragoso CA, Heffelfinger C, Zhao HY, Dellaporta SL (2016) Imputing genotypes in biallelic populations from low coverage sequence data. Genetics 202:487–495. doi :10.1534/genetics.115.182071

56. Zheng C, Boer MP, van Eeuwijk FA (2018) Accurate Genotype Imputation in Multiparental Populations from Low-Coverage Sequence. Genetics 210:71–82

57. Thépot S, Restoux G, Goldringer I, Hospital F, et al. (2015) Efficiently tracking selection in a multiparental population: the case of earliness in wheat. Genetics 199:609–623. doi :10.1534/genetics.114.169995

58. Money D, Gardner K Migicovsky Z, Schwaninger H, Zhong GY, Myles S (2015) k Nearest Neighbor method : LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. G3 5:2383–2390. doi: 10.1534/g3.115.021667

59. Money D, Migicovsky Z, Gardner K, Myles S. (2017) LinkImputeR: user-guided genotype calling and imputation for non-model organisms. BMC Genomics 18(1):523. doi: 10.1186/s12864-017-3873-5

60. Whalen A, Gorjanc G, Hickey JM (2020) AlphaFamImpute: high-accuracy imputation in full-sib families from genotype-by-sequencing data. Bioinformatics 36(15): 4369–4371. doi: 10.1093/bioinformatics/btaa499

61. Yoshida GM, Carvalheiro R., Lhorente JP, Correa K, et al. (2018) Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using high-density and low-density SNP panels. Aquaculture. doi:10.1016/j.aquaculture.2018.03.004

62. Toghiani S, Aggrey SE, Rekaya R (2016) Multi-generational imputation of single nucleotide polymorphism marker genotypes and accuracy of genomic selection. Animal 10:1077–1085. doi:10.1017/S1751731115002906

63. Wolc A, Kranis A, Arango J, Settar P, et al. (2016) Implementation of genomic selection in the poultry industry. Anim Front. 6(1):23–31. doi:10.2527/af.2016-0004

64. Bolormaa S, Gore K, Van Der Werf JHJ, Hayes BJ, Daetwyler HD (2015) Design of a low density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. Anim Genet. 46(5):544–56. doi:10.1111/age.12340

65. Badke YM, Bates RO, Ernst CW, Schwab C, et al. (2013) Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC Genet. 14:8

66. Zhang Z, Druet T (2010) Marker imputation with low-density marker panels in Dutch Holstein cattle. J Dairy Sci. 93:5487–5494

67. Weigel KA, de los Campos G, Vazquez AI, Rosa GJM, Gianola D, et al. (2010) Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in jersey cattle. J Dairy Sci. 93:5423–5435

68. Mulder HA, Calus MPL, Druet T, Schrooten C (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

69. He S, Wang S, Fu W, Ding X, Zhang Q (2014) Imputation of missing genotypes from low- to high-density SNP panel in different population designs. Anim. Genet. 46:1–7

70. Frioni N, Cavero D, Simianer H, et al. (2019) Phasing quality assessment in a brown layer population through family- and population-based software. BMC Genet 20:57. doi:10.1186/s12863-019-0759-3

71. Wang X, Su G, Hao D et al. (2020). Comparisons of improved genomic predictions generated by different imputation methods for genotyping by sequencing data in livestock populations. Journal of Animal Science and Biotechnology 11:3

72. Hickey JM, Crossa J, de los Campos G, Babu R (2012) Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Sci. 52 (2): 654-663. doi :10.2135/cropsci2011.07.0358.

73. Gunia M, Saintilan R, Venot E, Hozé C, Fouilloux MN, Phocas F (2014) Genomic prediction in French Charolais beef cattle using high-density single nucleotide polymorphism markers. J. Anim. Sci. 92:3258-3269

74. Spindel J, Begum H, Akdemir D, Virk P, et al. (2015) Genomic selection and association mapping in rice (Oryza sativa): Effect of trait genetic architecture, training population

composition, marker number and statistical model on accuracy of rice genomic selection in elite tropical rice breeding lines. PLoS Genet. 11:e1004982. doi :10.1371/journal.pgen.1004982

75. Griot R, Allal F, Phocas F et al. (2021) Optimisation of genomic selection to improve disease resistance in two marine fishes, the European sea bass (Dicentrarchus labrax) and the gilthead sea bream (Sparus aurata). Frontiers in genetics (submitted)

76. Hayes BJ, Bowman PJ, Daetwyler HD, Kijas JW (2012) Accuracy of genotype imputation in sheep breeds. Anim Genet. 43:72-80

77. Tsai HY, Matika O, Edwards SMK, Antolín-Sánchez R, Hamilton A, Guy DR, et al. (2017). Genotype imputation to improve the cost-efficiency of genomic selection in farmed Atlantic salmon. G3 7(4):1377–1383. doi:10.1534/g3.117.040717

78. Howie B, Marchini J, Stephens M (2011) Genotype imputation with thousands of genomes. G3 (Bethesda) 1:457–470

79. Druet T, Schrooten C, de Roos APW (2010) Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93:5443–5454

80. Boichard D, Chung H, Dassonneville R et al. (2012) Design of a Bovine Low-Density SNP Array Optimized for Imputation. Plos One 7:e34130

81. Sun C, Wu XL, Weigel KA, Rosa G.J.M., et al. (2012) An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genet. Res. (Camb.) 94, 133–150.doi:10.1017/S001667231200033X.

82. Hozé C, Fouilloux MN, Venot E, et al. (2013). High-density marker imputation accuracy in sixteen French cattle breeds. Genet Sel Evol.45:33

83. Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using low density marker panels. Genetics 182:343–353. doi: 10.1534/genetics.108.100289

84. Qin ZS, Gopalakrishnan S, Abecasis GR (2006) An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria. Bioinf22 (2): 220-225. doi:10.1093/bioinformatics/bti762.

85. Herry F, Hérault F, Picard Druet D, Varenne A, et al. (2018) Design of low density SNP chips for genotype imputation in layer chicken. BMC Genet. 19(1):1–14. doi.org;10.1186/s12863-018-0695-7

86. Wellmann R, Preuß S, Tholen E, Heinkel J, Wimmers K, Bennewitz J (2013) Genomic selection using low density marker panels with application to a sire line in pigs. Genet Sel Evol. 45(1):1–11. doi:10.1186/1297-9686-45-28

87. Porto-Neto LR, Sonstegard TS, Liu GE, Bickhart DM, et al. (2013) Genomic divergence of zebu and taurine cattle identified through high-density SNP genotyping. BMC Genomics 14:876. doi:10.1186/1471-2164-14-876

88. Moser G, Khatkar MS, Hayes BJ, Raadsma HW (2010) Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet Sel Evol.42(1):1–15. doi:10.1186/1297-9686-42-37

89. Chen L, Li C, Sargolzaei M., Schenkel F (2014) Impact of Genotype Imputation on the Performance of GBLUP and Bayesian Methods for Genomic Prediction. Plos One 9(7) : e101544 doi:10.1371/journal.pone.0101544

90. Vallejo RL, Leeds TD, Gao G, Parsons JE, et al. (2017) Genomic selection models double the accuracy of predicted breeding values for bacterial cold water disease resistance compared to a traditional pedigree-based model in rainbow trout aquaculture. Genet. Sel. Evol. 49(1):1–13. doi.org:10.1186/s12711-017-0293-6

91. Liu A, Lund MS, Boichard D et al. (2020). Improvement of genomic prediction by integrating additional single nucleotide polymorphisms selected from imputed whole genome sequencing data. Heredity 124:37–49. doi :10.1038/s41437-019-0246-7

92. Dassonneville R, Brøndum R.F, Druet T, et al. (2011) Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. Journal of Dairy Science 94(7):3679-3686. doi:10.3168/jds.2011-4299

93. Wang C, Habier D, Peiris BL, Wolc A, et al. (2013) Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide polymorphism panel in broiler chickens. Poultry Science 92(7):1712-1723. doi:10.3382/ps.2012-02941

94. Tsairidou S, Hamilton A, Robledo D, Bron JE, Houston RD (2020) Optimizing low-cost genotyping and imputation strategies for genomic selection in atlantic salmon. G3 10(2):581–590. doi.org:10.1534/g3.119.400800

95. Moghaddar N, Gore KP, Daetwyler HD, Hayes BJ, van der Werf JHJ (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

96. Cleveland MA, Hickey JM (2013) Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. Journal of Animal Science 91:3583–3592. doi.org:10.2527/jas.2013-6270

97. Heidaritabar M, Calus MPL, Vereijken A, Groenen MAM, Bastiaansen JWM (2015) Accuracy of imputation using the most common sires as reference population in layer chickens. BMC Genetics 16:101. doi:10.1186/s12863-015-0253-5

98. Judge MM, Purfield DC, Sleator RD, Berry DP (2017) The impact of multi-generational genotype imputation strategies on imputation accuracy and subsequent genomic predictions. J. Anim. Sci. 95:1489–1501. doi:10.2527/jas2016.1212

99. Meuwissen T, Goddard M (2010) The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. Genetics 185:1441–1450

100. Druet T, Macleod IM, Hayes BJ (2014) Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112(1):39–47. doi:10.1038/hdy.2013.13

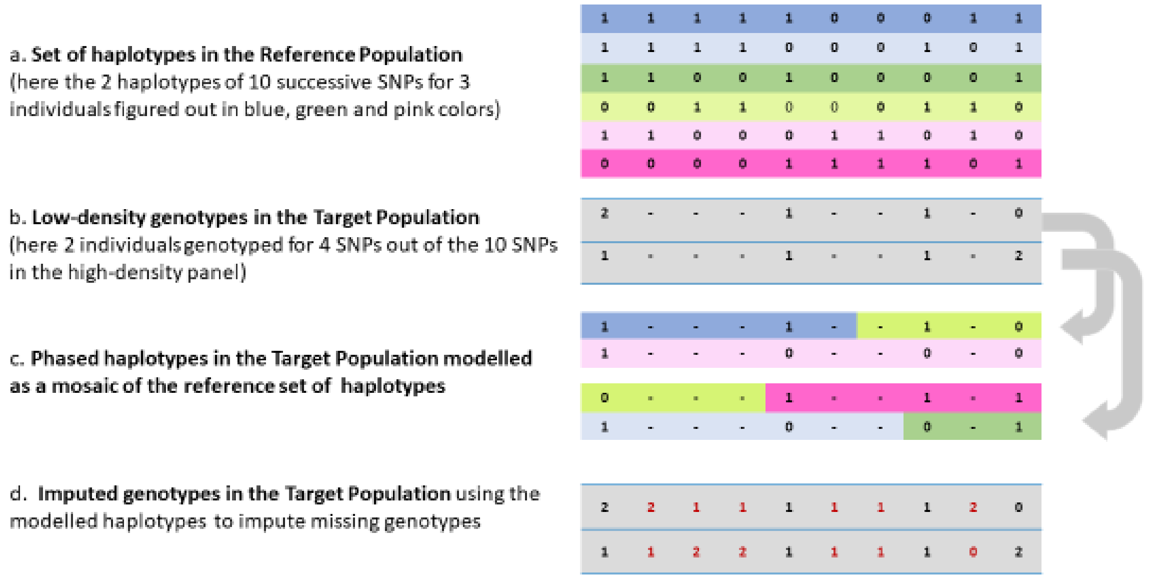
101. Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389-2397. doi: 10.1534/genetics.107.081190

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

Table 1. List of the main genotype imputation methods and their main software versions

Software Name	**Current Version**	**Referenced versions**
Population-based imputation methods requiring a reference panel		
BEAGLE	v5.1	v3.3 ***(24)*** v4.1 ***(25)*** v5.0 ***(27)***
fastPHASE	v1.4	***(14)***
GeneImp	v1.3	***(34)***
GLIMPSE	v1	***(35)***
IMPUTE	v5 named IMPUTE5	***(26)***
IMPUTE2	IMPUTE v2	***(22)***
MINIMAC	v4 named MINIMAC4	V1 named MINIMAC ***(16)*** V2 named MINIMAC2 ***(33)*** V3 named MINIMAC3 ***(23)***
PLINK	v2 named PLINK2	***(31)***
Pedigree-based imputation methods requiring a reference panel		
AlphaImpute	v1.9	***(39)***
findhap	v4	v1 ***(49)*** v4 ***(50)***
FImpute	v3	***(20)***
GIGI	v1.06	***(38)***
MERLIN	v1.1	***(36) (37)***
Free reference panel-based imputation methods		
AlphaFamImpute	v1	***(60)***
LinkImpute	v1	***(58)***
LinkImputeR	v1	***(59)***
magicImpute	v1	***(56***
SCDA	v1	***(51)***
STITCH	v1.6	***(11)***

Figure 1. Imputation process based on a set of haplotypes in a Reference Population



a. **Set of haplotypes in the Reference Population** (here the 2 haplotypes of 10 successive SNPs for 3 individuals figured out in blue, green and pink colors)

1	1	1	1	1	0	0	0	1	1
1	1	1	1	0	0	0	1	0	1
1	1	0	0	1	0	0	0	0	1
0	0	1	1	0	0	0	1	1	0
1	1	0	0	0	1	1	0	1	0
0	0	0	0	1	1	1	1	0	1

b. **Low-density genotypes in the Target Population** (here 2 individuals genotyped for 4 SNPs out of the 10 SNPs in the high-density panel)

2	-	-	-	1	-	-	1	-	0
1	-	-	-	1	-	-	1	-	2

c. **Phased haplotypes in the Target Population modelled as a mosaic of the reference set of haplotypes**

1	-	-	-	1	-	-	1	-	0
1	-	-	-	0	-	-	0	-	0
0	-	-	-	1	-	-	1	-	1
1	-	-	-	0	-	-	0	-	1

d. **Imputed genotypes in the Target Population** using the modelled haplotypes to impute missing genotypes

2	2	1	1	1	1	1	1	2	0
1	1	2	2	1	1	1	1	0	2