Genomic selection in farm animals: accuracy of prediction and applications with imputed whole-genome sequencing data in chicken

Dissertation
to obtain the Ph. D. degree
in the International Ph. D. Program for Agricultural Sciences
in Goettingen (IPAG)
at the Faculty of Agricultural Sciences,
Georg-August-University Göttingen, Germany

presented by

Guiyan Ni

born in Shandong, P. R. China

Göttingen, February 2016

D7

1st Referee: Prof. Dr. Henner Simianer
Animal Breeding and Genetics Group
Department of Animal Sciences
Georg-August-University Göttingen

2nd Referee: Prof. Dr Jörn Bennewitz
Farm Animal Genetics and Breeding
Institute of Animal Science
University of Hohenheim

Date of dissertation: 18th of February, 2016

Table of contents

Summary

Genomic prediction has been successfully applied in many livestock breeding schemes, based on different densities of single nucleotide polymorphism (SNP) array data. With the availability of whole-genome sequencing (WGS) data, which may contain the causal mutations, there are a growing number of studies to conducting genomic prediction with WGS data.

The main objective of this thesis was to investigate the possibility of imputing SNP array data up to the whole genome sequence level (**Chapter 2**) and then perform genomic prediction based on the imputed WGS data and SNP array data with different genomic relationship matrices to account for genetic architecture (**Chapter 3**). To further understand the accuracy of genomic prediction, a simulation study was performed to determine the degree of overestimation of the accuracy of genomic prediction, in order to propose a new method (**Chapter 4**).

The technical progress in the last decade has made it possible to sequence millions of DNA reads in a relatively short time frame. Several variant callers based on different algorithms have emerged and have made it possible to extract SNPs out of the whole-genome sequence. Often, only a few individuals of a population are sequenced completely and imputation is used to obtain genotypes for all sequence-based SNP loci for other individuals that have been genotyped for a subset of SNPs using a genotyping array. Thus, in **Chapter 2** we first compared the sets of variants detected with different variant callers, namely GATK, freebayes and SAMtools, and checked the quality of genotypes of the called variants in a set of 50 fully sequenced white and brown layers. There were 1,741,573 SNPs detected by all three callers on the studied chromosomes 3, 6, and 28, which was 71.6% (81.6%, 88.0%) of SNPs detected by GATK (SAMtools, freebayes) in total. Genotype concordance (GC), defined as the proportion of individuals whose array-derived genotypes are the same as the sequence-derived genotypes over all non-missing SNPs on the array, was 0.98 with GATK, 0.98 with SAMtools, and 0.97 with freebayes averaged over all SNPs on the studied chromosomes, respectively. Furthermore, for GATK (SAMtools, freebayes) 90 (88, 75) percent of variants had high values (>0.9) for other quality measures (non-reference sensitivity, non-reference genotype concordance and precision). Performance of all variant callers studied was very good in general, particularly for GATK and SAMtools. Second, we assessed the imputation accuracy (measured as the correlation between imputed and true genotype per SNP and per individual and genotype conflict between father-progeny pairs) when imputing from high density

SNP array data to whole-genome sequence using data from approximately 1000 individuals from six generations. Three different imputation programs (Minimac, FImpute and IMPUTE2) were checked in different validation scenarios. Across all imputation programs, correlation between true and imputed genotypes was >0.95 on average with randomly masked 1000 SNPs from the SNP array and >0.85 for a leave-one-out cross-validation within sequenced individuals. FImpute performed slightly worse than Minimac and IMPUTE2 in terms of genotype correlation, especially for SNPs with low minor allele frequency, however, it did have the lowest numbers in Mendelian conflicts in available father-progeny pairs. Correlations of real and imputed genotypes remained constantly high even if individuals to be imputed were several generations away from the sequenced individuals. In conclusion, among three variant callers tested GATK proved the relatively better performance; Minimac proved the relatively better performance comparing to the other two imputation programs tested.

Based on the conclusions in **Chapter 2**, we applied a genomic prediction with imputed WGS in **Chapter 3**. A commercial brown layer line comprising of 892 chickens from 6 generations was used in the study. These chickens were genotyped with a high density array data. Using the WGS data of 25 individuals, those array data were imputed up to the sequence level. The imputation was done with Minimac3, which needs pre-phased data generated with Beagle4. Accuracy of genomic prediction was measured as the correlation between de-regressed proofs and direct genomic breeding values of eggshell strength, feed intake and laying rate. In this study, besides the accuracy of genomic prediction based on array data and WGS data, accuracy based on different genomic relationship matrices to account for genetic architecture was investigated. The alternative weighting factors used were uniform, $-(log_{10}P)$ from a t-test of genome wide association study, and the square of estimated SNP effects from random regression BLUP. Best linear unbiased prediction given genetic architecture (BLUP|GA) was investigated as well. Prediction with uniform weights (the original GBLUP) was implemented with all SNPs or with only genic SNPs, both based on array and imputed whole sequence data. Averaging over the studied traits, predictive ability with only genic SNPs in WGS data was 0.366 ± 0.075, which was the highest predictive ability observed in the current study. Genomic prediction with genic SNPs in high density array data provided the second highest accuracy (0.361 ± 0.072). The prediction with $-(log_{10}P)$ or squares of SNP effects as weighting factors for building a genomic relationship matrix or BLUP|GA did not lead to higher accuracy, compared to that with uniform weights, regardless of the SNP set used. The results from this study showed that little or no benefit was gained when using all imputed WGS data to perform genomic prediction compared to using HD array data, regardless of the different SNP weightings tested. However, higher predictive ability was observed when using only genic SNPs extracted from the WGS data for genomic prediction.

Decisions of genomic selection schemes are made based on the genomic breeding values (GBV) of selection candidates. Thus, the accuracy of GBV is a relevant parameter, as it reflects the stability of the prediction and the possibility that the GBV might change when more information becomes available. It is also one of the key factors in expected response to selection, which is also known as breeders' equation. Accuracy of genomic prediction, however, is difficult to assess, considering true breeding values (TBV) of the candidates are not available in reality. In previous studies, several methods are proposed to assess the accuracy of GBV by using population and trait parameters (e.g. the effective population size, the reliability of quasi-phenotypes used, the number of independent chromosome segments) or parameters inferred from the mixed model equations. In practice, most approaches were found to overestimate the accuracy of genomic prediction. Thus, in **Chapter 4** we tested several approaches used in previous studies based on simulated data under a variety of parameters mimicking different livestock breeding programs (i.e. a cattle-like and a pig-like as well as a basic scenario) and measured the magnitude of overestimation. Then we proposed a novel and computationally feasible method. Based on the comparison in **Chapter 4**, the new method provided a better prediction for the accuracy of GBV. The method still had one unknown parameter, for which we suggested an approach to approximate its value from a suitable data set reflecting two separate time points. In conclusion, the new approach provided a better assessment of the accuracy of GBVs in many cases.

Zusammenfassung

Methoden zur genomischen Vorhersage basierend auf Genotypinformationen von Single Nucleotide Polymorphism (SNP)-Arrays mit unterschiedlicher Markeranzahl sind mittlerweile in vielen Zuchtprogrammen für Nutztiere fest implementiert. Mit der zunehmenden Verfügbarkeit von vollständigen Genomsequenzdaten, die auch kausale Mutationen enthalten, werden mehr und mehr Studien veröffentlicht, bei denen genomische Vorhersagen beruhend auf Sequenzdaten durchgeführt werden.

Das Hauptziel dieser Arbeit war zu untersuchen, inwieweit SNP-Array-Daten mit statistischen Verfahren bis zum Sequenzlevel ergänzt werden können (sogenanntes „Imputing“) (Kapitel 2) und ob die genomische Vorhersage mit imputeten Sequenzdaten und zusätzlicher Information über die genetische Architektur eines Merkmals verbessert werden kann (Kapitel 3). Um die Genauigkeit der genomischen Vorhersage besser verstehen und eine neue Methode zur Approximation dieser Genauigkeit ableiten zu können, wurde außerdem eine Simulationsstudie durchgeführt, die den Grad der Überschätzung der Genauigkeit der genomischen Vorhersage verschiedener bereits bekannter Ansätze überprüfte (Kapitel 4).

Der technische Fortschritt im letzten Jahrzehnt hat es ermöglicht, in relativ kurzer Zeit Millionen von DNA-Abschnitten zu sequenzieren. Mehrere auf unterschiedlichen Algorithmen basierende Software-Programme zur Auffindung von Sequenzvarianten (sogenanntes „Variant Calling“) haben sich etabliert und es möglich gemacht, SNPs in den vollständigen Genomsequenzdaten zu detektieren detektieren. Oft werden nur wenige Individuen einer Population vollständig sequenziert und die Genotypen der anderen Individuen, die mit einem SNP-Array an einer Teilmenge dieser SNPs typisiert wurden, imputet.

In Kapitel 2 wurden deshalb anhand von 50 vollständig sequenzierten Weiß- und Braunleger-Individuen die mit drei unterschiedlichen Variant-Calling-Programmen (GATK, freebayes and SAMtools) detektierten Genomvarianten verglichen und die Qualität der Genotypen überprüft. Auf den untersuchten Chromosomen 3,6 und 26 wurden 1.741.573 SNPs von allen drei Variant Callers detektiert was 71,6% (81,6%, 88,0%) der Anzahl der von GATK (SAMtools, freebayes) detektierten Varianten entspricht. Die Kenngröße der Konkordanz der Genotypen („genotype concordance“), die durch den Anteil der Individuen definiert ist, deren Array-basierte Genotypen mit den Sequenz-basierten Genotypen an allen auch auf dem Array vorhandenen SNPs

übereinstimmt, betrug 0,98 mit GATK, 0,98 mit SAMtools und 0,97 mit freebayes (Werte gemittelt über SNPs auf den untersuchten Chromosomen). Des Weiteren wiesen bei Nutzung von GATK (SAMtools, freebayes) 90% (88 %, 75%) der Varianten hohe Werte (>0.9) anderer Qualitätsmaße (non-reference sensitivity, non-reference genotype concordance und precision) auf.

Die Leistung aller untersuchten Variant-Calling-Programme war im Allgemeinen sehr gut, besonders die von GATK und SAMtools. In dieser Studie wurde außerdem in einem Datensatz von ungefähr 1000 Individuen aus 6 Generationen die Güte des Imputings von einem hochdichten SNP-Array zum Sequenzlevel untersucht. Die Güte des Imputings wurde mit Hilfe der Korrelationen zwischen imputeten und wahren Genotypen pro SNP oder pro Individuum und der Anzahl an Mendelschen Konflikten bei Vater-Nachkommen-Paaren beschrieben. Drei unterschiedliche Imputing-Programme (Minimac, FImpute und IMPUTE2) wurden in unterschiedlichen Szenarien validiert.

Bei allen Imputing-Programmen betrug die Korrelation zwischen wahren und imputeten Genotypen bei 1000 Array-SNPs, die zufällig ausgewählt und deren Genotypen im Imputing-Prozess als unbekannt angenommen wurden, durchschnittlich mehr als 0.95 sowie mehr als 0.85 bei einer Leave-One-Out-Kreuzvalidierung, die mit den sequenzierten Individuen durchgeführt wurde. Hinsichtlich der Genotypenkorrelation zeigten Minimac und IMPUTE2 etwas bessere Ergebnisse als FImpute. Dies galt besonders für SNPs mit niedriger Frequenz des selteneren Allels. FImpute wies jedoch die kleinste Anzahl von Mendelschen Konflikten in verfügbaren Vater-Nachkommen-Paaren auf. Die Korrelation zwischen wahren und imputeten Genotypen blieb auf hohem Niveau, auch wenn die Individuen, deren Genotypen imputet wurden, einige Generationen jünger waren als die sequenzierten Individuen. Zusammenfassend zeigte in dieser Studie GATK die beste Leistung unter den getesteten Variant-Calling-Programmen, während Minimac sich unter den untersuchten Imputing-Programmen als das beste erwies.

Aufbauend auf den Ergebnissen aus Kapitel 2 wurden in Kapitel 3 Studien zur genomischen Vorhersage mit imputeten Sequenzdaten durchgeführt. Daten von 892 Individuen aus 6 Generationen einer kommerziellen Braunlegerlinie standen hierfür zur Verfügung. Diese Tiere waren alle mit einem hochdichten SNP-Array genotypisiert. Unter der Nutzung der Daten von 25 vollständig sequenzierten Individuen wurden jene Tiere ausgehend von den Array-Genotypen bis zum Sequenzlevel hin imputet. Das

Imputing wurde mit Minimac3 durchgeführt, das bereits haplotypisierte Daten (in dieser Studie mit Beagle4 erzeugt) als Input benötigt.

Die Genauigkeit der genomischen Vorhersage wurde durch die Korrelation zwischen de-regressierten konventionellen Zuchtwerten und direkt genomischen Zuchtwerten für die Merkmale Bruchfestigkeit, Futteraufnahme und Legerate gemessen. Neben dem Vergleich der Genauigkeit der auf SNP-Array-Daten und Sequenzdaten basierenden genomischen Vorhersage wurde in dieser Studie auch untersucht, wie sich die Verwendung verschiedener genomischer Verwandtschaftsmatrizen, die die genetische Architektur berücksichtigen, auf die Vorhersagegenauigkeit auswirkt. Hierbei wurden neben dem Basisszenario mit gleichgewichteten SNPs auch Szenarien mit Gewichtungsfaktoren, nämlich den $-(log_{10}P)$-Werten eines t-Tests basierend auf einer genomweiten Assoziationsstudie und den quadrierten geschätzten SNP-Effekten aus einem Random Regression-BLUP-Modell, sowie die Methode BLUP|GA („best linear unbiased prediction given genetic architecture“) überprüft. Das Szenario GBLUP mit gleichgewichteten SNPs wurde sowohl mit einer Verwandtschaftsmatrix aus allen verfügbaren SNPs oder nur derer in Genregionen, jeweils ausgehend von der Grundmenge aller imputeten SNPs in der Sequenz oder der Array-SNPs, getestet.

Gemittelt über alle untersuchten Merkmale war die Vorhersagegenauigkeit mit SNPs aus Genregionen, die aus den imputeten Sequenzdaten extrahiert wurden, mit 0,366 ± 0,075 am höchsten. Den zweithöchsten Wert erreichte die genomische Vorhersage mit SNPs aus Genregionen, die im SNP-Array erhalten sind (0,361 ± 0,072). Weder die Verwendung gewichteter genomischer Verwandtschaftsmatrizen noch die Anwendung von BLUP|GA führten im Vergleich zum normalen GBLUP-Ansatz zu höheren Vorhersagegenauigkeiten. Diese Beobachtung war unabhängig davon, ob SNP-Array- oder imputete Sequenzdaten verwendet wurden. Die Ergebnisse dieser Studie zeigten, dass kaum oder kein Zusatznutzen durch die Verwendung von imputeten Sequenzdaten generiert werden kann. Eine Erhöhung der Vorhersagegenauigkeit konnte jedoch erreicht werden, wenn die Verwandschaftsmatrix nur aus den SNPs in Genregionen gebildet wurde, die aus den Sequenzdaten extrahiert wurden.

Die Auswahl der Selektionskandidaten erfolgt in genomischen Selektionsprogrammen mit Hilfe der geschätzten genomischen Zuchtwerte (GBVs). Die Genauigkeit des GBV ist hierbei ein relevanter Parameter, weil sie die Stabilität der geschätzten Zuchtwerte beschreibt und zeigen kann, wie sich der GBV verändern kann, wenn mehr Informationen verfügbar werden. Des Weiteren ist sie einer der entscheidenden Faktoren beim erwarteten Zuchtfortschritt (auch als so genannte „Züchtergleichung“ beschrieben). Diese Genauigkeit der genomischen Vorhersage ist jedoch in realen Daten schwer zu quantifizieren, da die wahren Zuchtwerte (TBV) nicht verfügbar sind. In früheren Studien wurden mehrere Methoden vorgeschlagen, die es ermöglichen, die Genauigkeit von GBV durch Populations- und Merkmalsparameter (z.B. effektive Populationsgröße, Sicherheit der verwendeten Quasi-Phänotypen, Anzahl der unabhängigen Chromosomen-Segmente) zu approximieren. Weiterhin kann die Genauigkeit bei Verwendung von gemischten Modellen mit Hilfe der Varianz des Vorhersagefehlers abgeleitet werden.

In der Praxis wiesen die meisten dieser Ansätze eine Überschätzung der Genauigkeit der Vorhersage auf. Deshalb wurden in Kapitel 4 mehrere methodische Ansätze aus früheren Arbeiten in simulierten Daten mit unterschiedlichen Parametern, mit Hilfe derer verschiedene Tierzuchtprogramme (neben einem Basisszenario ein Rinder- und ein Schweinezuchtschema) abgebildet wurden, überprüft und die Höhe der Überschätzung gemessen. Außerdem wurde in diesem Kapitel eine neue und leicht rechenbare Methode zur Approximation der Genauigkeit vorgestellt Die Ergebnisse des Vergleichs der methodischen Ansätze in Kapitel 4 zeigten, dass die Genauigkeit der GBV durch den neuen Ansatz besser vorhergesagt werden kann. Der vorgestellte Ansatz besitzt immer noch einen unbekannten Parameter, für den jedoch eine Approximation möglich ist, wenn in einem geeigneten Datensatz Ergebnisse von Zuchtwertschätzungen zu zwei verschiedenen Zeitpunkten vorliegen. Zusammenfassend kann gesagt werden, dass diese neue Methode die Approximation der Genauigkeit des GBV in vielen Fällen verbessert.

Chapter 1 General introduction

The main focus of this thesis is to investigate the prediction accuracy with imputed whole-genome sequencing data and high density array data. A short overview of different relevant theories and the availability of data will be described in this chapter.

Genomic prediction

The process of animal breeding includes selecting the best individuals from the current generation (selection) as parents for the next generation (mating). One of the methodological cornerstones was the introduction of the best linear unbiased prediction (BLUP) (Henderson, 1975), which opened up an era of comprehensive selection. The theory of BLUP uses the phenotypic records from candidates themselves or records from relatives and relationships among individuals building the mixed model equations to estimate fixed effects (mostly environment effects) and random effects (including breeding values) simultaneously to predict estimated breeding values (EBVs). Breeders can select the candidates with the largest EBVs as parents for the next generation. In addition to the basic mixed model, a number of extensions have been introduced to handle data with different structure, e.g. a sire model, a reduced animal model, and an animal model with groups (Mrode, 1996).

With the availability of the first genetic markers, combining molecular genetic information into selection has come into the scope of animal breeding and different methods have been proposed, summarized under the term marker-assisted selection (MAS). MAS refers to a method of selecting candidate individuals in a breeding scheme based on DNA molecular marker patterns in addition to their trait phenotype. MAS includes linkage disequilibrium-based MAS (LD-MAS) and linkage equilibrium-based MAS (LE-MAS) (Meuwissen and Goddard, 2010). The hypothesis for LD-MAS is that a maker is in linkage disequilibrium with a causal mutation or quantitative trait locus (QTL); thus this marker can be used as a proxy for that QTL. The genetic variance explained by QTL can be captured by the nearby marker with a factor of r^2, which is commonly a measure of linkage disequilibrium. However, in reality, the positions of QTLs are normally unknown, not to mention the linkage disequilibrium between QTLs and markers. The basic idea of LE-MAS is to scan the genome to detect the inheritance of markers in order to build the identical-by-descent probability which then is used in a statistical model. Even though a few studies have identified the casual mutation for traits controlled by a single or limited number of QTLs (Dekkers, 2004), most economically important traits are of complex nature, which means that these traits are controlled by many genes, are influenced by the environment and the observed phenotypes are mostly on a continuous scale. Consequently, the application of MAS was generally of limited success.

Another major breakthrough after BLUP and MAS was genomic selection (Meuwissen et al., 2001), which came along with available dense marker information. Genomic selection is in fact a form of LD-MAS, since it also relies on the linkage disequilibrium between the markers and the QTLs. However, different from MAS, the idea behind genomic selection is to use (ten) thousands of markers covering the whole genome, so that some of the markers inevitably are in linkage disequilibrium with the QTLs. Thus, the markers can potentially explain a major part of the genetic variance (Meuwissen et al., 2001). The basic step of genomic prediction is estimating the effect of thousands of markers in a population for individuals with both phenotypic data and genotypic data (training set) simultaneously, then summing all the marker effects for candidates that only have genotypic, but no phenotypic data to obtain their genomic estimated breeding values (Goddard and Hayes, 2007).

Compared to the previous selection methods, genomic selection holds at least two advantages. First, since the phenotypic data of selection candidates are not required, genomic selection can improve the selection process for traits that can only be measured in one sex, appear late in life (even after death), or that are too expensive to measure (Meuwissen et al., 2001). Second, in conventional BLUP the relationships between individuals are determined based on the pedigree information as an expected relationship, whilst in genomic selection the realized relationship between individuals can be estimated by using genomic information, which means that it can measure the Mendelian sampling effect which potentially increases accuracy of selection (Hayes et al., 2009b). Consequently, genomic selection has become one of the most powerful tools in animal breeding schemes.

Models for breeding value estimation

1) Conventional BLUP and genomic BLUP

Conventional BLUP in this thesis refers to the best linear unbiased prediction of Henderson (Henderson, 1975). The basic animal model of conventional BLUP is the following:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is a vector of phenotypic records with dimension number of individuals with phenotype times one $(n \times 1)$, $\boldsymbol{\beta}$ is a matrix with p fixed effects $(n \times p)$, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices relating phenotypic records to the fixed effects and random additive effects $\boldsymbol{u}$, $\boldsymbol{e}$ is a vector of random residual effects. $\boldsymbol{u}$ is assumed to be normally distributed with $\boldsymbol{u} \sim N(0, \boldsymbol{A}\sigma_u^2)$ and $\boldsymbol{e}$ is assumed to be normally distributed with $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$. $\boldsymbol{A}$ is the pedigree-based numerator relationship matrix.

In this animal-based model, replacing the pedigree-based numerator relationship matrix $\boldsymbol{A}$ by a genomic relationship matrix $\boldsymbol{G}$ will lead to genomic BLUP (GBLUP) (Goddard, 2009; Hayes et al., 2009b). The construction of the genomic relationship matrix will be presented in the following.

2) Ridge regression BLUP

$$\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{W}\boldsymbol{g} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is a vector of observations of individuals in the training set for a specific trait (quasi-phenotype); $\boldsymbol{1}$ is a vector of 1s, relating the effect of the mean to each record; μ is the overall mean; $\boldsymbol{W}$ is a design matrix relating quasi-phenotypes to the genotypes of m markers with dimension number of individuals in the training set × number of markers m; $\boldsymbol{g}$ is the vector of the random effects of the m markers and $\boldsymbol{g} \sim N(0, \boldsymbol{I}\sigma_g^2)$; $\boldsymbol{e}$ is the vector of residual terms and $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$.

In this model only the individuals with observations are used to estimate the marker effects. However, the direct genomic breeding values (DGV) for an individual i with or without observations can be assessed as the summation of estimated SNP effects times its genotypes:

$$DGV_i = \sum_{k=1}^{m} W_{ik}\widehat{g_k}$$

In fact, GBLUP and RRBLUP are two fully equivalent models if the genomic relationship matrix is specified appropriately. This equivalence has been proven in several studies, e.g. Habier et. al. (2007), Goddard (2009), VanRaden (2008). Consequently, the DGVs from RRBLUP and those from GBLUP are identical.

In conventional BLUP, the independent variables are normally the phenotypic observations of the selection candidates or observations of their relatives, while in GBLUP the independent variables are normally quasi-phenotypes, e.g. EBVs estimated from conventional BLUP, de-regressed proofs (DRPs) (Garrick et al., 2009), or daughter yield deviations (DYD).

There are more models that can be employed in animal breeding, e.g. Bayesian models. The 'Bayesian alphabet' refers to various Bayesian linear regression models used in genomic selection that differ in the priors (Gianola, 2013), for instance, BayesA and BayesB (Meuwissen et al., 2001), BayesC (Habier et al., 2011), BayesCπ (Habier et al., 2011), and BayesR (Erbe et al., 2012). More specifically, in BayesA, each SNP is assumed to have a different variance, and these variances follow an inverse-chi-squared distribution. Model BayesB assumes that a certain proportion of SNPs have no effects, while the rest of the SNPs have a SNP-specific variance. Bayesian models are quite often found to outperform GBLUP with simulated data, while yielding similar predictive ability in real data (Habier et al., 2011; Wang et al., 2015). Genomic prediction with Bayesian alphabet methods is beyond the scope of this thesis; however, more details can be found in several review papers e.g. Gianola (2013), de los Campos et al. (2013).

Conventional BLUP is used to estimate breeding values in **Chapter 4**. GBLUP for estimating direct genomic breeding values is used in **Chapter 3** and **4** of this thesis. SNP effects are estimated by RRBLUP in **Chapter 3**.

Establishment of the genomic relationship matrix G

The matrix of relationship between individuals is crucial since it can be used to estimate breeding values, to assess the covariance structure and to manage inbreeding. There are several ways to construct a relationship matrix. In conventional BLUP, relationships between individuals are estimated based on pedigree information only which can reflect the expected relationship between individuals; thus, pedigree-based relationship cannot distinguish full sibs from each other, since they share the same pedigree information. In genomic selection, a genomic relationship matrix can be constructed using the genomic markers covering the whole genome. In this genomic relationship matrix, a realized relationship between individuals can be measured, because different individuals may inherit different alleles from the last generation. In other words, Mendelian sampling effects are taken into account in a genomic relationship matrix, which is one of the sources of increased accuracy of genomic selection compared to the selection based on breeding values estimated via conventional BLUP.

A genomic relationship matrix can be built in different ways. One of the first and one of the most widely used approaches to build the genomic relationship matrix was proposed by VanRaden (2008), i.e.

$$\boldsymbol{G} = \frac{\boldsymbol{MDM^T}}{2\sum_{k=1}^{m} p_k(1-p_k)}$$

where $\boldsymbol{M}$ contains the centered SNP genotypes with individuals in rows and SNPs in columns. The elements of column k of $\boldsymbol{M}$ are $0-2p_k$ for homozygotes of the first allele, $1-2p_k$ for heterozygotes, and $2-2p_k$ for homozygotes of the second allele, where p_k is the frequency of the second allele at locus k from the current data set. $\boldsymbol{D}$ is a diagonal matrix with weights on different loci. An identity matrix is used to construct matrix $\boldsymbol{D}$ in VanRaden (2008), which implies that all loci equally contribute to the variance-covariance structure. The i^{th} diagonal value of the $\boldsymbol{G}$ matrix minus one is the genomic inbreeding coefficient of an individual i.

Considering that different traits may be affected by different SNPs and different numbers of SNPs, the assumption that all SNPs have equal contribution to all traits may be not suitable. Thus, different approaches to build the trait-specific $\boldsymbol{G}$ matrix are proposed in order to account for genetic architecture. Most of the approaches replace the identity matrix of $\boldsymbol{D}$ with a more informative diagonal matrix, for example, the squares of SNP effects estimated from the training set with RRBLUP (Su et al., 2014), or $-(log_{10}P)$ from a t-test in a GWAS (de los Campos et al., 2013b). In addition, Zhang et al. (2015) proposed an approach named best linear unbiased prediction given genomic architecture (BLUP|GA), which can also account for different genomic architecture based on variable selection.

More details regarding the construction of genomic relationship matrices based on different weighting factors is reported in **Chapter 3**. Genomic prediction results based on different genomic relationship matrices with different weighting factors as presented above are reported in **Chapter 3**.

Accuracy of genomic prediction

Empirical accuracy of prediction can be measured as the correlation between the true breeding values and breeding values estimated from different models. Accuracy of prediction is relevant because it is an important component of response to selection per year, also known as the breeder's equation (Falconer and Mackay, 1996) and shown in the following:

$$\Delta G = i\rho\sigma_A/L$$

where ΔG is the response to selection per year, i is the intensity of selection, ρ is the correlation between true and estimated breeding values (accuracy of selection), σ_A is the additive genetic standard deviation and L is the generation interval. This equation can predict how the mean value of a trait under selection changes from one year to the next.

Since the true breeding values are unknown in reality, the correlation between quasi-phenotypes and direct genomic breeding values, often called 'predictive ability', is often used as a proxy for the accuracy of prediction. Beyond this, there are several approaches to estimate the accuracy of selection at the individual level or in a population. First, accuracy of selection for each individual can be obtained from the framework of BLUP (Henderson, 1975), i.e.

$$r_{\boldsymbol{BV}_i} = \sqrt{1 - \frac{PEV_i}{var(\boldsymbol{A}_i)}}$$

where PEV_i is the prediction error variance for individual i, and can be obtained from the framework of BLUP, and $var(\boldsymbol{A}_i)$ is the genetic variance. Second, the expected accuracy of selection across all individuals can also be obtained from the definition of correlation between genomic and true breeding values, r_{GT}, (Amer and Banos, 2010), i.e.

$$r_{GT} = \frac{r_{EG}}{r_{ET}\left(1 + \frac{cov(\boldsymbol{\varepsilon_E},\boldsymbol{\varepsilon_G})}{var(\boldsymbol{T})}\right)}$$

where r_{EG} is the correlation between EBVs and GBVs; r_{ET} is the theoretical accuracy of EBVs which can be obtained from the framework of BLUP; $cov(\boldsymbol{\varepsilon_E}, \boldsymbol{\varepsilon_G})$ is the covariance between errors of EBV ($\boldsymbol{\varepsilon_E}$) and errors of GBV ($\boldsymbol{\varepsilon_G}$); $var(\boldsymbol{T})$ is the variance of true breeding values (TBVs). In this approach, the covariance between TBVs and errors of EBVs $[cov(TBV, \varepsilon_E)]$ and covariance between TBVs and errors of GBVs $[cov(TBV, \varepsilon_G)]$ are both assumed to be zero. If we further assume the covariance between errors of EBVs and errors of GBVs $[cov(\boldsymbol{\varepsilon_E}, \boldsymbol{\varepsilon_G})]$ to be zero as well, then accuracy of selection can be measured as

$$r_{GT} = \frac{r_{EG}}{r_{ET}}$$

which is the formula proposed by Hayes et al. (2009b).

Complementarily, cross-validation is often used to assess the accuracy of prediction. In general, the population is split into two groups: training set and validation set. The analysis is performed in the training set and validated in the validation set. The commonly used strategies to split the population are leave-one-out cross-validation and k-fold cross-validation. In leave-one-out cross-validation, each training set is created by taking all the individuals except one. Consequently, the validation set is the one individual left out. Thus, for a dataset with n individuals, n replicates have to be run. In k-fold cross-validation, all individuals are divided randomly into k groups with equal size. The learning process is performed using individuals in $k - 1$ folds. The validation set is the fold left out. This process is repeated k times with each of the k groups acting as validation set once. As a measure of accuracy, the correlation between estimated genomic breeding values and (available, but not used) quasi phenotypes of the individuals in the left-out fold is calculated. This is then averaged across all folds – so that each individual is predicted once – and eventually across replicates of the validation set characterized by different random assignment of individuals to the k groups. In practice, 5-fold and 10 fold cross-validation are commonly used. If k=n, the k-fold cross-validation is the same as the leave-one-out cross-validation.

In the cross-validation strategies mentioned above, individuals are randomly assigned to the training set or validation set. In reality, this may not be the most relevant case because the selection candidates are normally younger than the individuals in the training set, which can be mimicked by masking the phenotypic data of younger animal as unknown. Then predictive ability can be obtained for the young selection candidates after genomic prediction, which is called forward prediction.

The different approaches to estimate accuracy of selection are investigated in **Chapter 4** based on a simulation study. Cross-validation strategies to assess the accuracy of selection are carried out in **Chapter 3**. In **Chapter 2**, we employ cross-validation to assess imputation accuracy.

Factors affecting accuracy of genomic prediction

There are many factors which can affect the accuracy of selection, such as the number of individuals in the training population (Goddard and Hayes, 2009), relationship between training population and validation population (Erbe et al., 2012; Fangmann et al., 2015), relationship among individuals in the training population (Pszczola et al., 2012), the heritability of the trait of interest (Calus et al., 2008), availability of information (e.g. number of progenies), and marker density (Erbe et al., 2013).

Marker density can determine the strength of linkage disequilibrium between SNPs and QTLs. With higher density of genotypic data, the linkage disequilibrium between SNPs and QTLs can be potentially increased, which makes genotypic data a better proxy for the causal QTLs. Furthermore, with the availability of whole-genome sequence data, the potential causal mutations can be included in the data.

Given that the marker density is sufficient, the number of individuals in the training population (i.e. the number of individuals with both phenotypic records and genotypic data) is one of the crucial factors affecting accuracy of genomic prediction. This is due to the fact that with an increasing size of the training population, SNP effects can be estimated more accurately.

The relationship between training population and validation population is an important factor affecting accuracy of selection, especially in across breeds or across populations selection. For breeds or populations with only small training populations available, previous studies found that combining training populations from relatively close breeds may be helpful for the accuracy of selection (de Roos et al., 2009). However, genomic selection relies on the phase of linkage disequilibrium between SNPs and QTLs, which might be different from one breed to another, thus the genetic distance between training population and validation population can influence accuracy of genomic prediction.

Genomic prediction with high density commercial arrays and whole-genome sequencing data is carried out in **Chapter 3**. The effect of the heritability of traits of interest and availability of information on accuracy of genomic prediction is investigated in **Chapter 4** with a simulation study.

Implementation of genomic selection

Genomic selection was first and has been widely performed in breeding schemes of dairy cattle. The increase in accuracy of genomic selection has been reported in many studies for a range of traits and countries (Hayes et al., 2009a; Hayes et al., 2009b). For instance, based on data in year 2003, VanRaden et al. (2009) reported that realized reliability based on genomic selection was 0.50 compared to the reliability achieved with traditional selection of 0.27, averaging over more than 20 traits.

For pigs, the implementation of genomic selection so far is not as wide as in dairy cattle breeding. However, genomic prediction has started in this sector as well (e.g. the PigGS project in Germany, among others). Gierlaug-Enger et al. (2013) evaluated the accuracy of genomic selection for the trait intramuscular fat based on 4,576 Norsvin Landrace and 3,408 Norsvin Duroc pigs, and reported that accuracy of breeding values increased from 0.36 (accuracy with traditional selection) to 0.63 with genomic selection, averaged over the two breeds.

For chickens, the literature and research on genomic selection is growing as well (Wang et al., 2013; Van Eenennaam et al., 2014). Wolc et al. (2015) performed a three-year experiment by splitting a brown egg laying hen population into two sub-lines: conventional selection was performed in one sub-line, and genomic selection was performed in the other sub-line. Based on the results of 16 traits, they reported that chickens selected by genomic selection outperformed those selected by conventional selection, with a doubled response to selection for traits egg weight and yolk weight. In addition, they also found that the inbreeding per year in the genomically selected sub-line was lower than that in the conventionally selected sub-line.

Availability of SNP array data and whole-genome sequencing data in chicken

In 2004, the International Chicken Genome Sequencing Consortium released the first draft of the sequence of the chicken genome, based on DNA from an inbred Red Jungle Fowl (International Chicken Genome Sequencing Consortium, 2004). To date, the fourth version of the reference genome (Gallus_gallus-4.0) is available within public databases, and includes 28 of 38 heterosomes, two linkage groups and two sex chromosomes. A large amount of genomic variation including SNPs, copy number variants (CNVs), short insertion and deletion (INDELs) are available and are used for the design of commercial array chips. Groenen et al. (2011) designed a moderate density SNP array chip of chicken (Illumina SNP BeadChip, short for 60K chip), which in total consists of 60,800 SNPs known to be segregating in broilers and laying hens at high to medium allele frequencies. This 60K chip has been commercially available since 2011 and is designed based on the second build of the chicken genome (Gallus_gallus-2.1). In 2013, the Affymetrix Axiom® Chicken Genotyping Array with 600K SNPs (HD array) became commercially available (Kranis et al., 2013). 580,954 SNPs, including 21,534 coding variants, were selected from 243 chickens in 24 lines (including experimental, commercial broiler and layer lines). The selected SNPs are evenly distributed in the genetic map, resulting in a higher physical density (SNPs per kilo-basepairs) on micro- compared to macro-chromosomes. This HD array was designed based on the fourth build of the chicken genome (Gallus_gallus-4.0). Note that the individuals used in **Chapter 2** and **3** were genotyped with this HD array offered by Affymetrix.

The technical progress in the last decade has made it possible to sequence millions of DNA reads in a relatively short time frame at reasonable costs. Thus, whole-genome sequencing has become available which allows us to gather more information on genetic variation, genes, gene function and other characterizations of genomes (Bentley, 2006; Mardis, 2008), leading to the emergence of research projects dealing with whole-genome sequencing data in humans (Morozova and Marra, 2008; Lam et al., 2012; Goldstein et al., 2013; Sims et al., 2014), domestic animals (Rubin et al., 2010; Baes et al., 2014; Daetwyler et al., 2014) and other species (Hickey et al., 2012a). Large consortia, e.g. the 1000 bull genomes project (Grant et al., 2011; Daetwyler et al., 2014) and the human genome project (International Human Genome Sequencing Consortium, 2001; International Human Genome Sequencing Consortium, 2004) have been established to accumulate available resources, detect new variants in genomes, better understand genetic architecture of different traits and to find or narrow down positions of potential causal loci.

The basic workflow to obtain whole-genome sequencing data

The raw output of next generation sequence technology for whole-genome sequencing data normally is in a FASTQ format, which is text-based sequence information with nucleotides being represented by a single letter. The basic workflow to convert this format into the final set of SNP calls is illustrated in Figure 1.1.

In the first step, we align the raw reads to a reference genome. There are several tools available for alignment, e.g. Burrows-Wheeler Alignment tool (BWA) (Li and Durbin, 2009), Mapping and Assembly with Qualities (MAQ) (Li et al., 2008), and Bowtie (Langmead et al., 2009). BWA is one of the most popular alignment tools. BWA is based on backward search with Burrows–Wheeler Transform, which is an efficient algorithm to align short sequencing reads to a reference genome. It can also cope with sequencing errors and mismatches and supports both single-end and pair-end alignment. For each alignment, BWA also generates a mapping quality score reflecting the probability that the alignment is correct. By default, it generates a SAM/BAM format of alignment.

In the second step, we mark or remove duplicates. This step can be done with MarkDuplicates utility of Picard (http://picard.sourceforge.net) or with rmdup utility of

SAMtools. Whether duplicates are marked or removed depends on the study design, since it is not easy to distinguish PCR artifacts and real DNA duplicates.

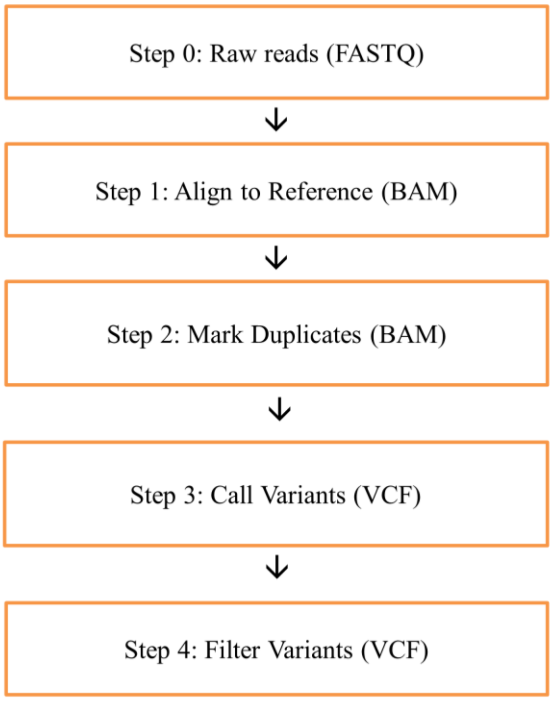

Step 0: Raw reads (FASTQ)
Step 1: Align to Reference (BAM)
Step 2: Mark Duplicates (BAM)
Step 3: Call Variants (VCF)
Step 4: Filter Variants (VCF)

Figure 1.1: Basic workflow to convert the raw output files of next generation sequence technology into the final set of SNP calls.

The next step is calling the variant from the BAM file. Several variant callers based on different algorithms have emerged using single or multiple samples simultaneously, e.g. the utilities of SAMtools (Li et al., 2009), UnifiedGenotyper or HaplotypeCaller of GATK (McKenna et al., 2010), and freebayes (Garrison and Marth, 2012). Although the priors are different, GATK and SAMtools use rather similar Bayesian methods for estimation of the posterior probability of the genotype and detection of variants relying on alignment. Freebayes also uses Bayesian methods to detect variants, but is haplotyped-based, in the sense that it calls variants based on the literal sequences of reads aligned to a particular target, not their precise alignment (https://github.com/ekg/freebayes).

To minimize the artefacts that may affect the downstream analyses and to enhance the quality of variants, it is crucial to perform a filtering for the called variants per individual and per position. Different strategies to select the so-called high-quality variants have been suggested (Qanbari et al., 2012; O'Rawe et al., 2013; Cheng et al., 2014). In our study, we only filtered depth of coverage and mapping quality as shown in **Chapter 2** and **3**. Furthermore, this workflow actually only illustrates the basic guideline. There are some custom-modifications depending on the design of a study. For instance, there are several studies in which filtering for the raw reads was performed followed by the alignment of the reads which passed the filter to the reference genome (see e.g. Patel and Jain, 2012).

There are several options for programs of variant calling; thus, we performed a comparison among three variant callers (i.e. GATK, SAMtools, and freebayes) in **Chapter 2**.

Imputation

Imputation refers to the methods predicting the genotypes of SNPs which were not typed (Marchini and Howie, 2008). Since the idea of imputation was first proposed by Burdick (2006), it has become an essential tool in the analysis based on genotypic data in order to maximize the number of samples and SNPs used in genomic prediction (Su et al., 2012; Wellmann et al., 2013; Badke et al., 2014; Morota and Abdollahi-Arpanahi, 2014), to improve the power of GWAS (Scuteri et al., 2007; Willer and Mohlke, 2012), and to facilitate meta-analysis (Kathiresan et al., 2008; Sanna et al., 2008).

Imputation algorithms

There are several algorithms and programs available to perform the imputation procedure, which can be classified into two main categories: rule-based approaches and model-based approaches (Browning and Browning, 2011).

One of the well-known rule-based approaches is used in the program FImpute (Sargolzaei et al., 2014). It is based on the rules that the closer relatives share longer haplotypes and the relatives distanced further apart share shorter haplotypes due to mutation and recombination over generations. It identifies haplotypes using overlapping sliding windows, which is faster than most model-based programs. In addition, FImpute is one of the programs that can combine pedigree and linkage disequilibrium information in one imputation process.

Among model-based algorithms, there are several models commonly used. Several well-known programs are based on the coalescent model such as fastPHASE (Scheet and Stephens, 2006), HaploRec (Eronen et al., 2006), and PHASE (Stephens et al., 2001). Hidden Markov models are underlying programs like Beagle (Browning and Browning, 2007), Impute2 (Howie et al., 2009), and Minimac3 (Howie et al., 2012). The general idea behind this model is estimating haplotypes that are present in both reference and study samples and then imputing the missing genotype for the study samples. Beagle is one of the most popular programs for phasing and imputation in animal breeding, however, it is not convenient to use for whole-genome sequencing data in terms of imputation speed. In some relevant cases genotypic data from low density and high density SNP arrays along with WGS data for some individuals may be available in the same population. Multiple step imputation (i.e. imputing low density array data into high density data, and then to a sequence level) could increase the imputation errors (Khatkar et al., 2012). Impute2 can address this challenge by using multiple reference panels that contain different SNP sets. Minimac3 claims to be a low memory, computationally efficient imputation program, but requires pre-phased data as an input.

Imputation accuracy

Imputation accuracy can be measured in different ways, e.g. as the correlation between true and imputed genotypes, imputation error rates, and genotype conflicts between parents and progeny.

When measuring imputation accuracy as the correlation between the true and imputed genotypes, accuracy tends to increase with the increase of minor allele frequency (MAF). Because SNPs with lower MAF tend to have weaker linkage disequilibrium (LD) with their surrounding SNPs, and it is difficult to build haplotypes as the SNPs are in weak LD (Browning and Browning, 2011; Hickey et al., 2012b). Nonetheless, correlation between true and imputed genotypes is a relatively good and commonly used measure for imputation accuracy.

Imputation error rate counts the incorrectly imputed alleles. It is difficult to compare imputation accuracy cross loci, since it depends on MAF of the respective SNP (Hickey et al., 2012a; Calus et al., 2014), in that imputation error rates tend to decrease with increasing MAF.

Measuring genotype conflict means to search for alternative homozygotes being present in parent and progeny pairs, which violates Mendelian rules. Genotype conflict cannot assess the imputation accuracy of heterozygous SNPs. Furthermore, genotype conflicts can only be verified in adjacent generations and can be easily affected by pedigree errors.

Nevertheless, the measurements mentioned above can evaluate the imputation accuracy from different aspects. In addition, some imputation programs can also provide imputation accuracy according to their methods. For instance, Minimac (Howie et al., 2012) and its later versions offer a criterion called Rsq to evaluate the quality of imputation per SNP. Rsq is the estimated squared correlation between true and imputed genotypes. Beagle and several other programs offer similar measurements and have suggested threshold values. FImpute, however, does not offer such an evaluation.

Different strategies are performed to assess the imputation accuracy in terms of correlation and genotype conflicts in **Chapter 2**.

Factors affecting imputation accuracy

A number of factors can influence the imputation accuracy, e.g. size of reference population, relationship between or within reference and study populations, marker density, allele frequency, and genotype errors.

Size of reference population has been reported as one of the most important factors affecting imputation accuracy (Druet et al., 2010; Browning and Browning, 2011; Hickey et al., 2011). With a large reference population, more haplotypes in the population can be discovered. Thus the possibility of matching the haplotypes showed in the study population is increased. The relationship between or within reference and study populations is an important factor as well (Marchini et al., 2006), because closer relatives tend to share longer haplotypes and relatives further apart tend to share shorter haplotypes due to mutation and recombination over generations. Thus, the relationship can affect the correctness of haplotypes and further affect imputation accuracy. MAF is another factor affecting imputation accuracy. As mentioned before, SNPs with low MAF normally are in low LD with surrounding SNPs. Consequently, it is difficult to build haplotypes for those SNPs, which can cause low imputation accuracy.

Objectives of this thesis

Genomic selection has brought a breakthrough to modern animal breeding since it was proposed in 2001. With the feasibility of next-generation sequencing technologies, genomic selection has become possible using whole-genome sequencing data in animal breeding schemes. This is expected to lead to higher predictive ability, since whole-genome sequencing data may contain all genomic variants including causal mutations. However, sequencing all individuals in a population is not realistic due to the lack of DNA resources and funding. Thus, the first objective of this thesis is to investigate the optimal strategy to impute array data up to the sequencing level. Further, we compare the advantage of using whole-genome sequencing data in genomic selection over high density array data. The second objective is to measure the over-estimation of accuracy of genomic prediction with several available methods and further propose a new method to access the accuracy of genomic prediction.

Chapter 2 first compares the sets of variants detected with different variant callers, namely GATK, freebayes and SAMtools, and checks the quality of genotypes of the called variants in a set of 25 white layer and 25 brown layer individuals. Second, an assessment is presented regarding the imputation accuracy from SNP array data to whole-genome sequencing with three different imputation programs, namely Minimac, FImpute and IMPUTE2, in a brown layer line.

Chapter 3 compares genomic predictions using both high density array data and imputed whole-genome sequencing data in a commercial brown layer chicken line, based on the promising results in Chapter 2. In addition, GBLUP models with a variety of weighting factors for specific SNPs are studied.

Chapter 4 investigates several available approaches that are often used as measures for accuracy of genomic prediction to assess the magnitude of over-estimation based on a simulation study. In addition, a novel and computationally feasible method is proposed. The quality of the new approximation is evaluated with both simulated and real data.

Chapter 5 includes a general discussion about several critical issues regarding genomic prediction with imputed whole-genome sequencing data.

Reference

Amer, P. R., and G. Banos. 2010. Implications of avoiding overlap between training and testing data sets when evaluating genomic predictions of genetic merit. J. Dairy Sci. 93:3320–3330.

Badke, Y. M., R. O. Bates, C. W. Ernst, J. Fix, and J. P. Steibel. 2014. Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. G3 (Bethesda). 4:623–31.

Baes, C. F., M. a Dolezal, J. E. Koltes, B. Bapst, E. Fritz-Waters, S. Jansen, C. Flury, H. Signer-Hasler, C. Stricker, R. Fernando, R. Fries, J. Moll, D. J. Garrick, J. M. Reecy, and B. Gredler. 2014. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. BMC Genomics 15:948.

Bentley, D. R. 2006. Whole-genome re-sequencing. Curr. Opin. Genet. Dev. 16:545–52.

Browning, S. R., and B. L. Browning. 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. 81:1084–97.

Browning, S. R., and B. L. Browning. 2011. Haplotype phasing: existing methods and new developments. Nat. Rev. Genet. 12:703–14.

Burdick, J. T., W.-M. Chen, G. R. Abecasis, and V. G. Cheung. 2006. In silico method for inferring genotypes in pedigrees. Nat. Genet. 38:1002–4.

Calus, M. P. L., A. C. Bouwman, J. M. Hickey, R. F. Veerkamp, and H. A. Mulder. 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. Animal 8:1743–53.

Calus, M. P. L., T. H. E. Meuwissen, a P. W. de Roos, and R. F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553–61.

Cheng, A. Y., Y.-Y. Teo, and R. T.-H. Ong. 2014. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics 30:1707–13.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M.-N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858–865.

Dekkers, J. C. M. 2004. Commercial application of marker- and gene-assisted selection in livestock : Strategies and lessons. J. Anim. Sci.:E313–E328.

Druet, T., C. Schrooten, and a P. W. de Roos. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93:5443–54.

Van Eenennaam, A. L., K. a Weigel, A. E. Young, M. a Cleveland, and J. C. M. Dekkers. 2014. Applied animal genomics: results from the field. Annu. Rev. Anim. Biosci. 2:105–39.

Erbe, M., B. Gredler, F. R. Seefried, B. Bapst, and H. Simianer. 2013. A function accounting for training set size and marker density to model the average accuracy of genomic prediction. PLoS One 8:e81046.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. a Mason, and M. E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95:4114–29.

Eronen, L., F. Geerts, and H. Toivonen. 2006. HaploRec: efficient and accurate large-scale reconstruction of haplotypes. BMC Bioinformatics 7:542.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Longmans Green, Harlow, Essex, UK.

Fangmann, A., E. Tholen, H. Simianer, and M. Erbe. 2015. Can multi-subpopulation reference sets improve the genomic predictive ability for pigs ? J. Anim. Sci. 93:5618–5630.

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55.

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing. arXiv Prepr. arXiv1207.3907:1–9.

Gianola, D. 2013. Priors in whole-genome regression: the bayesian alphabet returns. Genetics 194:573–96.

Gjerlaug-Enger, E., O. Nordbo, and E. Grindflek. 2013. Genomic selection in pig breeding for improved meat quality. :1–3.

Goddard, M. E., and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10:381–91.

Goddard, M., and B. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet.124:323–330.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245–57.

Goldstein, D. B., A. Allen, J. Keebler, E. H. Margulies, S. Petrou, S. Petrovski, and S. Sunyaev. 2013. Sequencing studies in human genetics: design and interpretation. Nat. Rev. Genet. 14:460–70.

Grant, J. R., A. S. Arantes, X. Liao, and P. Stothard. 2011. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. Bioinformatics 27:2300–1.

Groenen, M. a M., H.-J. Megens, Y. Zare, W. C. Warren, L. W. Hillier, R. P. M. a Crooijmans, A. Vereijken, R. Okimoto, W. M. Muir, and H. H. Cheng. 2011. The development and characterization of a 60K SNP chip for chicken. BMC Genomics 12:274.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–97.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186.

Hayes, B. J., P. J. Bowman, a J. Chamberlain, and M. E. Goddard. 2009a. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92:433–43.

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. (Camb). 91:47–60.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012a. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Sci. 52:654.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012b. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Sci. 52:654.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. F. Wilson, N. Dunstan, and J. H. J. van der Werf. 2011. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet. Sel. Evol. 43:12.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44:955–9.

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

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–777.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature:931–945.

Kathiresan, S., O. Melander, C. Guiducci, A. Surti, N. P. Burtt, M. J. Rieder, G. M. Cooper, C. Roos, B. F. Voight, A. S. Havulinna, B. Wahlstrand, T. Hedner, D. Corella, E. S. Tai, J. M. Ordovas, G. Berglund, E. Vartiainen, P. Jousilahti, B. Hedblad, M.-R. Taskinen, C. Newton-Cheh, V. Salomaa, L. Peltonen, L. Groop, D. M. Altshuler, and M. Orho-Melander. 2008. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat. Genet. 40:189–97.

Khatkar, M. S., G. Moser, B. J. Hayes, and H. W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13:538.

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

Lam, H. Y. K., M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. E. Dewey, L. Habegger, E. a Ashley, M. B. Gerstein, A. J. Butte, H. P. Ji, and M. Snyder. 2012. Performance comparison of whole-genome sequencing platforms. Nat. Biotechnol. 30:78–82.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–60.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9.

Li, H., J. Ruan, and R. Durbin. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18:1851–8.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus. 2013a. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–45.

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen. 2013b. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 9:e1003608.

Marchini, J., D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. S. Qin, H. M. Munro, R. Abecasis, P. Donnelly, and I. Hapmap. 2006. A Comparison of Phasing Algorithms for Trios and Unrelated Individuals. :437–450.

Marchini, J., and B. Howie. 2008. Comparing Algorithms for Genotype Imputation. Am. J. Hum. Genet. 83:535–539.

Mardis, E. R. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. 24:133–41.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–303.

Meuwissen, T., and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics 185:623–31.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

Morota, G., and R. Abdollahi-Arpanahi. 2014. Genome-enabled prediction of quantitative traits in chickens using genomic annotation. BMC Genomics 15.

Morozova, O., and M. a Marra. 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92:255–64.

Mrode, R. A. 1996. Linear Models for the Prediction of Animal Breeding Values: 3rd Edition. CABI pubilishing.

O'Rawe, J., T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang, and G. J. Lyon. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 5:28.

Patel, R. K., and M. Jain. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7:e30619.

Pszczola, M., T. Strabel, H. a. Mulder, and M. P. L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J. Dairy Sci. 95:389–400.

Qanbari, S., T. M. Strom, G. Haberer, S. Weigend, A. a Gheyas, F. Turner, D. W. Burt, R. Preisinger, D. Gianola, and H. Simianer. 2012. A high resolution genome-wide scan for significant selective sweeps: an application to pooled sequence data in laying chickens. PLoS One 7:e49525.

De Roos, a P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of genomic predictions across multiple populations. Genetics 183:1545–53.

Rubin, C.-J., M. C. Zody, J. Eriksson, J. R. S. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallböök, F. Besnier, O. Carlborg, B. Bed'hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, and L. Andersson. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464:587–91.

Sanna, S., A. U. Jackson, R. Nagaraja, C. J. Willer, W.-M. Chen, L. L. Bonnycastle, H. Shen, N. Timpson, G. Lettre, G. Usala, P. S. Chines, H. M. Stringham, L. J. Scott, M. Dei, S. Lai, G. Albai, L. Crisponi, S. Naitza, K. F. Doheny, E. W. Pugh, Y. Ben-Shlomo, S. Ebrahim, D. a Lawlor, R. N. Bergman, R. M. Watanabe, M. Uda, J. Tuomilehto, J. Coresh, J. N. Hirschhorn, A. R. Shuldiner, D. Schlessinger, F. S. Collins, G. Davey Smith, E. Boerwinkle, A. Cao, M. Boehnke, G. R. Abecasis, and K. L. Mohlke. 2008. Common variants in the GDF5-UQCC region are associated with variation in human height. Nat. Genet. 40:198–203.

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

Scheet, P., and M. Stephens. 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.

Scuteri, A., S. Sanna, W.-M. Chen, M. Uda, G. Albai, J. Strait, S. Najjar, R. Nagaraja, M. Orrú, G. Usala, M. Dei, S. Lai, A. Maschio, F. Busonero, A. Mulas, G. B. Ehret, A. a Fink, A. B. Weder, R. S. Cooper, P. Galan, A. Chakravarti, D. Schlessinger, A. Cao, E. Lakatta, and G. R. Abecasis. 2007. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS Genet. 3:e115.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15:121–32.

Stephens, M., N. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. Am. J. Hum. …:978–989.

Su, G., R. F. Brøndum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (~54,000) and high-density (~777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J. Dairy Sci. 95:4657–65.

Su, G., O. F. Christensen, L. Janss, and M. S. Lund. 2014. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J. Dairy Sci. 97:6547–59.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–23.

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

Wang, T., Y.-P. P. Chen, M. E. Goddard, T. H. E. Meuwissen, K. E. Kemper, and B. J. Hayes. 2015. A computationally efficient algorithm for genomic prediction using a Bayesian model. Genet. Sel. Evol. 47:34.

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

Willer, C. J., and K. L. Mohlke. 2012. Finding genes and variants for lipid levels after genome-wide association analysis. Curr. Opin. Lipidol. 23:98–103.

Wolc, A., H. H. Zhao, J. Arango, P. Settar, J. E. Fulton, N. P. O'Sullivan, R. Preisinger, C. Stricker, D. Habier, R. L. Fernando, D. J. Garrick, S. J. Lamont, and J. C. M. Dekkers. 2015. Response and inbreeding from a genomic selection experiment in layer chickens. Genet. Sel. Evol. 47:59.

Zhang, Z., M. Erbe, J. He, U. Ober, N. Gao, H. Zhang, H. Simianer, and J. Li. 2015. Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. G3 (Bethesda). 5:615–27.

Chapter 2 Comparison among three variant callers and assessment of the accuracy of imputation from SNP array data to whole-genome sequence level in chicken

Guiyan Ni[1], Tim M. Strom[2], Hubert Pausch[3], Christian Reimer[1], Rudi Preisinger[4], Henner Simianer[1], Malena Erbe[1,5]

[1] Animal Breeding and Genetics Group, Georg-August-Universität, Göttingen, Germany

[2] Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany

[3] Chair of Animal Breeding, Technische Universität München, Freising, Germany

[4] Lohmann Tierzucht GmbH, Cuxhaven, Germany

[5] Institute for Animal Breeding, Bavarian State Research Centre for Agriculture, Grub, Germany

Published in BMC genomics

http://www.biomedcentral.com/1471-2164/16/824

Abstract

Background: The technical progress in the last decade has made it possible to sequence millions of DNA reads in a relatively short time frame. Several variant callers based on different algorithms have emerged and have made it possible to extract single nucleotide polymorphisms (SNPs) out of the whole-genome sequence. Often, only a few individuals of a population are sequenced completely and imputation is used to obtain genotypes for all sequence-based SNP loci for other individuals, which have been genotyped for a subset of SNPs using a genotyping array. This study focused on two objectives. First, we compared the sets of variants detected with different variant callers, namely GATK, freebayes and SAMtools, and checked the quality of genotypes of the called variants in a set of 50 fully sequenced white and brown layers. Second, we assessed the imputation accuracy (measured as the correlation between imputed and true genotype per SNP and per individual, and genotype conflict between father-progeny pairs) when imputing from high density SNP array data to whole-genome sequence using data from around 1000 individuals from six different generations. Three different imputation programs (Minimac, FImpute and IMPUTE2) were checked in different validation scenarios. **Results:** There were 1,741,573 SNPs detected by all three callers on the studied chromosomes 3, 6, and 28, which was 71.6% (81.6%, 88.0%) of SNPs detected by GATK (SAMtools, freebayes) in total. Genotype concordance (GC) defined as the proportion of individuals whose array-derived genotypes are the same as the sequence-derived genotypes over all non-missing SNPs on the array were 0.98 (GATK), 0.97 (freebayes) and 0.98 (SAMtools). Furthermore, the percentage of variants that had high values (>0.9) for another three measures (non-reference sensitivity, non-reference genotype concordance and precision) were 90 (88, 75) for GATK (SAMtools, freebayes). With all imputation programs, correlation between original and imputed genotypes was >0.95 on average with randomly masked 1000 SNPs from the SNP array and >0.85 for a leave-one-out cross-validation within sequenced individuals. **Conclusions:** Performance of all variant callers studied was very good in general, particularly for GATK and SAMtools. FImpute performed slightly worse than Minimac and IMPUTE2 in terms of genotype correlation, especially for SNPs with low minor allele frequency, while it had lowest numbers in Mendelian conflicts in available father-progeny pairs. Correlations of real and imputed genotypes remained constantly high even if individuals to be imputed were several generations away from the sequenced individuals.

Background

The technical progress in the last decade has made it possible to sequence millions of DNA reads in a relatively short time frame for reasonable costs. Thus, whole-genome sequencing has become available that allows us to gather more information on genetic variation, genes, gene function and other characterizations of genomes (Bentley, 2006; Mardis, 2008) and the number of research projects dealing with whole-genome sequencing data has been emerging in humans (Morozova and Marra, 2008; Lam et al., 2012; Goldstein et al., 2013; Sims et al., 2014), domestic animals (Rubin et al., 2010; Baes et al., 2014; Daetwyler et al., 2014) and other species (Hickey et al., 2012) in the last years. Large consortia (e.g. 1000 bull genomes project (Grant et al., 2011; Daetwyler et al., 2014) or the human genome project (International Human Genome Sequencing Consortium, 2001; International Human Genome Sequencing Consortium, 2004) have been established to accumulate available resources, detect new variants in genomes, better understand genetic architecture of different traits and find or narrow down positions of potential causal loci. In dairy cattle, for example, 28.3 million variants in the whole genome were identified from 234 bulls sequenced with an average coverage of 8.3X in the first phase of the 1000 bull genomes project, and loci associated with milk production and curly coat were detected by genome-wide association studies (Daetwyler et al., 2014). In chicken, research projects using whole-genome sequencing data have been rare so far. Rubin et al. (2010) generated pooled whole-genome sequencing data representing eight populations of domestic chickens in order to identify how genetics adapt to new environments. In the study of Qanbari et al. (2012) genome regions with strong evidence of selection were identified from pooled whole-genome sequencing data of 15 laying chickens. Within the framework of the project Synbreed (http://www.synbreed.tum.de/) whole-genome sequencing data of 50 individuals from commercial layer lines were generated which built the basis for this study.

Several variant callers based on different algorithms have emerged using single or multiple samples simultaneously, e.g. SAMtools (Li et al., 2009) or GATK (McKenna et al., 2010). Recently, some studies have shown that there is significant difference in the set of variants called by different variant callers (Rosenfeld et al., 2012; O’Rawe et al., 2013; Baes et al., 2014). Baes et al. (2014) found that the number of variants varied between variant callers (i.e. Platypus, SAMtools and two difference GATK utilities: UnifiedGenotyper and HaplotypeCaller) in whole-genome sequencing data of dairy cattle. O’Rawe et al. (2013) carried out a study to examine the concordance among different variant calling pipelines with default parameters, but their analyses mainly focused on exome sequencing and did not assess multiple sample variant calling algorithms. Thus, it is still important to evaluate genotype concordance and precision obtained with different variant callers in whole-genome sequencing data in chicken.

Although over the past several years the cost of DNA sequencing has decreased by several orders of magnitude due to the rapid development of sequencing technology, it is still comparatively expensive (Meynert et al., 2014). There are two main strategies to reduce costs: One is to only sequence coding exons which has been commonly used in human clinical applications (Linderman et al., 2014), but actually none of the available kits can cover all the coding exons (Sulonen et al., 2011). Besides, it was shown that both natural and positive selection eventually occurred in the non-coding DNA blocks and some QTL have been mapped in such blocks, so that important parts of the genome may be missed by just using exome sequencing (Bird et al., 2006; Drake et al., 2006). The other major strategy to reduce costs when being interested in sequence information of a whole population is to generate whole-genome sequencing data for a small set of individuals highly related to the population and then impute SNP array data of other individuals of the same population up to sequence level based on the whole-genome sequencing data of the sequenced individuals and array based SNP array data of the remaining individuals. Before whole-genome sequencing data had been available, the technique of imputation has already been used for imputing from low to high density SNP array data with high accuracy and thus has proven to be a successful line of action, e.g. in cattle (Pausch et al., 2013) to obtain higher marker densities for a large number of individuals.

Heidaritabar et al. (2014) showed the possibility to impute SNP array data into whole-genome sequencing data based on a small reference population of 22 sequenced individuals in simulated data. Druet et al. (2014) investigated the accuracy of imputation that can be achieved with Beagle (Browning and Browning, 2007) and found that the highest imputation accuracy was 0.86 when the simulated whole-genome sequencing data for 50 bulls with a 12X coverage was used as reference dataset. Van Binsbergen et al. (2014) and Pausch et al. (2014) showed that a reasonable accuracy of imputation (e.g. correlation between observed and imputed genotypes as high as 0.83) could be achieved when imputing from SNP array data to whole-genome sequencing data in dairy cattle breeds. Nevertheless, there has been no attempt so far to evaluate the accuracy of imputation from high density SNP array data (580k) up to sequence level with real chicken data.

In this study, we first compared the sets of variants detected with different variant callers, namely GATK (McKenna et al., 2010), freebayes (Garrison and Marth, 2012) and

SAMtools (Li et al., 2009), and checked the quality of genotypes of the called variants in a set of 25 white layer and 25 brown layer individuals. Second, we assessed the imputation accuracy from SNP array data to whole-genome sequencing with three different imputation programs, namely Minimac (Howie et al., 2012), FImpute (Sargolzaei et al., 2014) and IMPUTE2 (Howie et al., 2009), in a brown layer line.

Methods

Ethics statement

Samples were collected by veterinarians in the Lohmann Company in the course of a routine health check for diagnostic reasons and a partition of retained samples was used to extract DNA. The authors collected no samples themselves.

Data

Blood samples and pedigree data of more than 5 generations backwards (2260 individuals in total) were available for purebred individuals from different generations of a brown layer line. Number of individuals per generation is shown in Additional file 2.1. Furthermore, genotypes from the Affymetrix Axiom® Chicken Genotyping Array (580k array) were available for 1081 brown layer chickens (including 24 of the 25 sequenced brown layers) from 5 different generations which were later imputed to whole sequence level. Genotyped SNPs with minor allele frequency (MAF) smaller than 0.5% and genotyping call rate smaller than 97% were removed so that 350,602 SNPs remained. Individuals with a call rate smaller than 95% in the remaining SNPs were then excluded leaving a set of 1075 genotyped brown layer individuals.

Whole-genome sequencing and alignment

Fifty individuals (25 brown layers and 25 white layers) chosen to be from one of the older generations and highly related to the set of already genotyped individuals were sequenced with the Illumina HiSeq2000 technology with a target coverage of 8X. Sequence reads were aligned to Build 4 of the chicken reference genome (galGal4) using BWA (version 0.7.9a-r786) (Li and Durbin, 2009) with default parameters for paired-end alignment. In this step SAM files were generated, which were converted to BAM files using SAMtools (Li et al., 2009) in the following step. Reads were then further processed with the MarkDuplicates utility of Picard (http://picard.sourceforge.net) to remove potential PCR duplicates.

Variant detection

Variants including SNPs and short insertion and deletion (INDELs) were called using different software programs: GATK (version 3.1-1-g07a4bf8, UnifiedGenotyper) (McKenna et al., 2010), freebayes (version v0.9.15-1-g076a2a2) (Garrison and Marth, 2012) and the mpileup utility of SAMtools (version 0.1.19-96b5f2294a) (Li et al., 2009) with default parameters, respectively. With all programs mentioned, variant calling was performed with multi-sample approaches using all 50 sequenced chickens simultaneously. Sets of variants obtained with the three different callers were processed in equal manner, but independent from each other in the following. Different versions of the same variant callers may result in a different set of variants for the same underlying sequencing data. Thus, two versions of freebayes (version v0.9.15-1-g076a2a2 and version v9.9.2-22-gc283d6d) were compared regarding the overlap of called variants.

Filtering and genotype quality enhancement

To reduce the proportion of the false positive variants, different strategies to select the so-called high-quality variants have been suggested (Qanbari et al., 2012; O'Rawe et al., 2013; Yu and Sun, 2013; Cheng et al., 2014). We applied thresholds for depth of coverage (DP) and mapping quality (MQ) according to the following protocol: Extraction of SNPs from the set of all called variants was done using the SelectVariants command of GATK. Filtering for the SNPs called on all chromosomes of the whole-genome included the following criteria: First, outlier SNPs (top 0.5% of DP) were removed. Then, mean and standard deviation of DP of the remaining SNPs were calculated and SNPs with a DP above and below 3 times the standard deviation from the mean were removed as well. For mapping quality, SNPs with a MQ score smaller than 30 within SNP sets obtained with the variant callers GATK and SAMtools and SNPs with both mean mapping quality of observed alternate alleles (MQM) and mean mapping quality of observed reference alleles (MQMR) smaller than 30 within the SNP set obtained with freebayes were excluded from further analyses. Separate SNP sets were built for brown and white layers, respectively, in which SNPs that were monomorphic in the respective set of individuals were removed. Finally, we used Beagle 3.3.2 (Browning and Browning, 2007) (see Additional file 2.2 for the pipeline) in order to enhance the original genotype quality of the remaining SNPs following the proposal of Jansen et al. (2013). For all subsequent analyses regarding imputation, only data from brown layers and variants called by GATK were used. Furthermore, considering the computational efforts especially in the imputation process, all further analyses were not performed for the entire genome, but three chromosomes (chromosomes 3, 6 and 28) of different length were selected for the following analyses.

Validation of different variant callers

Genotype concordance (GC), non-reference sensitivity (NRS), non-reference genotype concordance (NRC), and precision were calculated based on array genotypes and corresponding sequence-based genotypes obtained with different variant callers (GATK (McKenna et al., 2010), freebayes (Garrison and Marth, 2012), SAMtools (Li et al., 2009). For each SNP, GC is the proportion of individuals whose array-derived genotypes are the same as the sequence-derived genotypes over all non-missing SNPs on the array. NRS is the number of individuals who have at least one non-reference allele in both whole-genome sequencing data and SNP array data divided by total number of individuals who have at least one non-reference allele in the array data. NRC is the number of animals whose array-derived genotypes are the same as the sequence-derived genotypes and are not homozygous for the reference allele divided by total number of individuals who have at least one non-reference allele in the SNP array data. Precision is the number of animals whose array-derived genotypes are the same as the sequence-derived genotypes and are not homozygous for the reference allele divided by total number of individuals who have at least one non-reference allele in the sequencing data. The detailed calculations are shown in Additional file 2.3, which were based on the definitions in DePristo et al. (2011) and Linderman et al. (2014). Validation of variant callers was done for all positions at which SNPs from the array were available on chromosomes 3, 6 and 28 (34,311, 13,627 and 2,730 SNPs), respectively, for the 24 brown layer individuals that were both genotyped and sequenced.

Imputation

Imputation was done with three software packages: Minimac (Howie et al., 2012), FImpute (Sargolzaei et al., 2014) and IMPUTE2 (Howie et al., 2009), among which Minimac and IMPUTE2 are based on pedigree-free algorithms, while FImpute can combine linkage disequilibrium (LD) information and pedigree information in the imputation process. FImpute uses an overlapping sliding window method to detect the relationship between study and reference set, while IMPUTE2 apples a hidden Markov model. Minimac implements the MaCH (Li et al., 2010) algorithm for genotype imputation. For all software, a default number of iteration was used. As Minimac and IMPUTE2 need phased input data, pre-phasing for whole-genome sequencing and SNP array data was performed using Beagle 4 (Browning and Browning, 2007).

Assessment of imputation quality

To evaluate the accuracy of imputation from SNP array data to whole-genome sequencing data, three strategies (described below) were used.

Leave-one-out cross-validation

Since 24 out of the 25 sequenced brown layer chickens were also genotyped with the 580k array, each of these individuals was excluded from the imputation reference data set once and imputation from SNP array data to whole-genome sequencing data was performed with the respective individual being one of the validation individuals in the resulting dataset. Genotype concordance and correlation between the imputed and sequenced genotypes from these run for all non-monomorphic SNPs being not on the array was calculated afterwards per the respective individual.

Father-progeny pair conflicts

Among the genotyped brown layer individuals there were 134 individuals that were progeny of one of the sequenced individuals. Thus, genotypes on imputed SNPs in the progeny could be compared to the father's genotypic information at these SNP positions and genotype conflicts (alternative homozygotes in father and progeny) were counted. Proportion of genotype conflicts were calculated per father-progeny-pair over all SNPs excluding the ones which were also genotyped using the 580k array on chromosomes 3, 6 and 28, respectively.

Accuracy of randomly masked 1000 SNPs

As imputation accuracy depends (amongst others) on the degree of relationship between sequenced individuals and individuals to be imputed, we also checked how imputing accuracy changes when different numbers of generations are between sequenced individuals and individuals to be imputed. For this analysis, we randomly masked (i.e. setting them to missing) 1000 SNPs (680 out of total 34'311 SNPs on chromosome 3, 270 out of 13'627 on chromosome 6 and 50 out of 2736 on chromosome 28) from the SNP array data in all genotyped individuals and imputed those SNPs as if they were SNPs from the sequence data. Afterwards, imputed genotypes on these 1000 SNPs and real array genotypes were compared and genotype correlation was calculated for each SNP and also for each individual. As Calus et al. (2014) and Mulder et al. (2012) demonstrated that it is better to center and scale true and impute genotype when calculating the individual-specific imputation accuracy, we investigated the individual-specific imputation accuracy based on original genotypes and standardized genotypes. The random masking was replicated five times with different random sets of 1000 SNPs and means of these five replicates are reported in the results.

Results and discussion

Alignment and coverage

For brown layer chickens, on average 88 million paired-end reads were obtained per individual. Among these reads, 1.72% (ranging from 0.76% -1.98%) on average were marked as duplications and excluded and on average 96.7% (ranging from 96.1% - 96.9%) were mapped against the reference genome (galGal4). Coverage per sample ranged from 5.0 to 16.6, with an average of 7.6. For white layer chickens, on average 94 million paired-end reads were obtained per individual. Among these reads, 1.69% (range 1.45% - 1.92%) were marked as duplications and excluded; and 96.7% (range 96.3% - 96.9%) were mapped against reference genome. Coverage per sample ranged from 7.9 to 15.6, with an average of 10.8. Based on this data set, the number of raw paired-end reads obtained was higher in white layer chickens than in brown layer chickens. However there was no difference in percentage of duplications and percentage of mapping. Details can be seen in Additional file 2.4.

Variant detection

Depending on the variant caller, totally 14,757,670 (GATK), 13,442,923 (freebayes), and 13,642,483 (SAMtools) variants (i.e. SNPs and INDELs) were detected with multi-sample calling on the 50 available brown and white layer chicken genomes (Additional file 2.5). In the study of Cheng et al. (2014) GATK identified almost the same amount of variants as SAMtools, while Pattnaik et al. (2012) identified more SNPs with freebayes than with SAMtools or GATK in the human genome. Unlike their results, we found that GATK identified more variants than SAMtools and freebayes, thus showing the same tendency as in studies of Liu et. al (2013), O’Rawe et al. (2013) and Baes et al. (2014). On the three chromosomes 3, 6, and 28 selected for imputation, GATK identified 2,297,603 variants per animal of which 2,054,930 were SNPs. After excluding low-quality SNPs that did not match the filtering criteria (as defined in the method section), there were 2,021,911 SNPs remaining. Compared to GATK, both SAMtools (2,125,837) and freebayes (2,055,976) detected less variants, and after excluding INDELs and filtering, there were 1,759,887 (1,652,870) SNPs remaining with SAMtools (freebayes).

Figure 2.1 illustrates the number of overlapping SNPs detected by the three variant callers on the three chromosomes 3, 6 and 28. Many SNPs were detected only by one variant caller (236,322 for GATK, 29,187 for SAMtools, and 138,860 for freebayes). However, 1,471,573 SNPs were detected by all three callers, which is 71.6% (81.6%, 88.0%) of SNPs detected by GATK (SAMtools, freebayes) in total. When focusing on GATK and SAMtools only, 1,763,383 SNPs were detected by both of them, which were the 86% of total SNPs detected by GATK, and 98% of total SNPs detected by SAMtools. Baes et al. (2014) found that around 18.3 million SNPs were both detected by SAMtools and GATK in the whole-genome of 65 individuals of the Swiss dairy cattle population, which was 83% of the total number of SNPs detected by GATK and 98% of SAMtools which are very similar proportions to the ones we observed in our study. Based on data from exomes of 20 humans, Liu et al. (2013) found in an exome sequencing study that 23,824 SNPs were both detected by SAMtools and GATK, which is 95.5% of the total number of SNPs detected by GATK and 89.8% of SAMtools. A high agreement of different callers (i.e. a high percentage of SNPs detected by different callers simultaneously) with each other in terms of called variants, is an advantage if whole-genome sequencing data is handled in a way like O'Rawe (2013) suggested, namely using only the variants discovered by multiple variants callers or pipelines for further analyses.

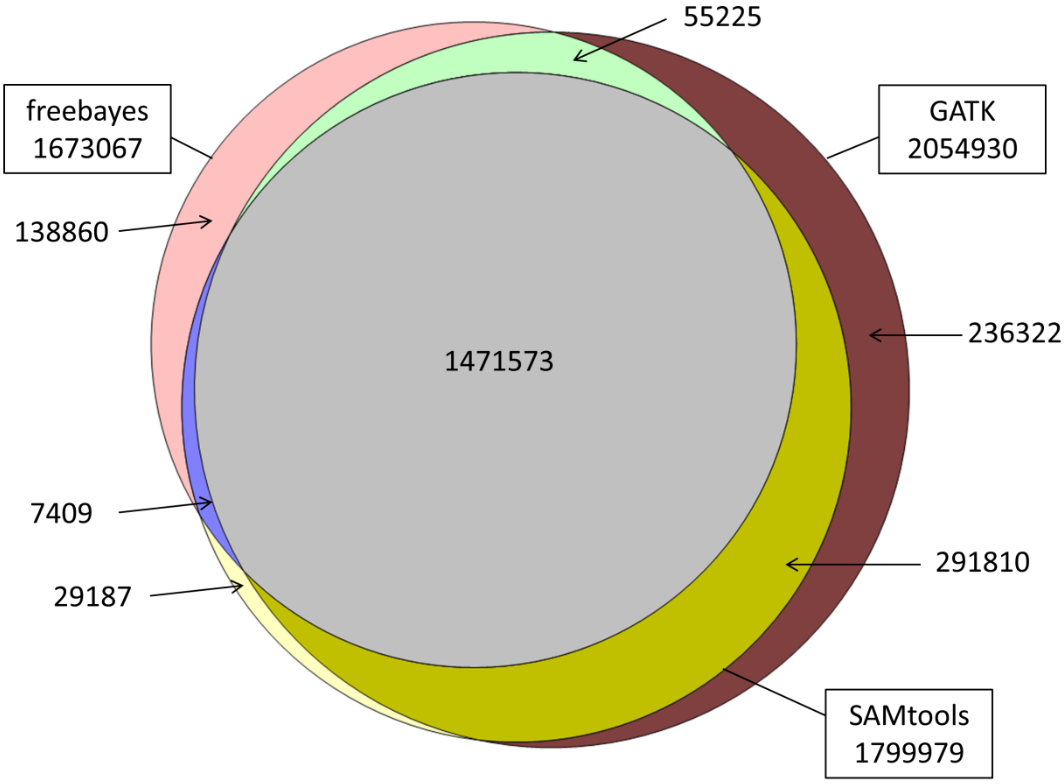


Figure 2.1: The overlap of single nucleotide polymorphisms detected by different variant callers.

Except the number of SNPs shared by different variant callers, we also compared GC, NRS, NRC, and precision of different callers as suggested in DePristo et al. (2011) and Linderman et al. (2014). Different quality measures are shown in Table 2.1 and Additional file 2.6. In general terms, we obtained very high values (> 0.9) for all metrics and all different callers, particularly for GATK and SAMtools. For 90% of variants called by GATK all four metrics were simultaneously larger than 0.9, while this was the case for 88% (75%) of variants called with SAMtools (freebayes), which was mainly due to lower values in NRC and precision. The four different metrics which were binned into 100 groups according to their array-derived MAF plotted against array-derived MAF are shown in Figure 2.2.In general, the similarity of metrics based on the SNPs called by GATK and SAMtools in different MAF bins was extremely high when compared to metrics based on SNPs called by freebayes. Results obtained with GATK and SAMtools were rather insensitive to MAF with the exception of GC, which showed a slight increase when MAF (<0.05) was low. Metrics for freebayes were in general lower (with the exception of NRS) and showed a slight increase with increasing MAF. The different properties of results of freebayes compared to GATK and SAMtools is likely due to (dis)similarities of the algorithms underlying the three programs. Although the priors are different, GATK (McKenna et al., 2010) and SAMtools (Li et al., 2009) use rather similar Bayesian methods for estimation of the posterior probability of the genotype and detection of variants relying on alignment. freebayes (Garrison and Marth, 2012) also uses Bayesian methods to detect variants, but is haplotyped-based, in the sense that it calls variants based on the literal sequences of reads aligned to a particular target, not their precise alignment (https://github.com/ekg/freebayes). Although GATK and SAMtools have equivalent performances and both perform better than freebayes, the metrics used here relied on the accuracy of array-derived genotypes which were assumed to be the 'true' genotypes. Eventually existing genotyping errors may thus bias the results. Besides, SNPs on the array were selected to be almost evenly distributed across the genome and were preselected to match a certain MAF spectrum which differs from the MAF distribution present in the sequence (Kranis et al., 2013), which also could bias the relative performance of variant callers if they differ in sensitivity to such patterns. Furthermore, the coverage of sequencing as a potential influence factor was not under consideration here. Linderman et al. (2014) discovered that insufficient coverage could bias the GC metrics, particularly the NRS. Thus, freebayes might have more similar results to GATK or SAMtools if individuals were sequenced with a higher coverage.

Table 2.1: Genotype concordance metrics. The calculation based on array genotypes and corresponding sequence-based genotypes obtained with different variant callers at positions where SNPs from the array were available on chromosomes 3, 6 and 28

	Genotype concordance						Non-reference sensitivity					
	GATK		freebayes		SAMtools		GATK		freebayes		SAMtools	
	No.SNPs	Mean±SD	No.SNPs	Mean±SD	No.SNPs	Mean±SD	No.SNPs	Mean±SD	No.SNPs	Mean±SD	No.SNPs	Mean±SD
≤0.1	23	0.03±0.03	9	0.03±0.03	3	0.08±0.00	3	0.01±0.02	9	0.04±0.04	5	0.04±0.04
≤0.2	18	0.14±0.03	6	0.15±0.03	6	0.15±0.02	13	0.15±0.03	10	0.14±0.03	15	0.14±0.03
≤0.3	12	0.25±0.02	9	0.27±0.02	9	0.25±0.03	25	0.25±0.03	7	0.24±0.03	20	0.24±0.03
≤0.4	13	0.33±0.02	5	0.33±0.01	18	0.35±0.03	20	0.34±0.02	11	0.34±0.03	19	0.34±0.03
≤0.5	22	0.45±0.03	18	0.45±0.03	24	0.45±0.03	20	0.45±0.02	14	0.45±0.03	24	0.44±0.03
≤0.6	42	0.55±0.03	39	0.55±0.04	43	0.55±0.03	59	0.52±0.03	89	0.51±0.02	82	0.51±0.03
≤0.7	73	0.65±0.03	64	0.66±0.03	62	0.66±0.03	96	0.66±0.02	105	0.66±0.02	158	0.66±0.03
≤0.8	110	0.76±0.03	183	0.76±0.03	123	0.76±0.03	152	0.75±0.02	150	0.75±0.02	215	0.75±0.02
≤0.9	679	0.87±0.03	3828	0.88±0.02	738	0.87±0.03	1206	0.86±0.03	1171	0.86±0.03	1561	0.86±0.03
≤1	50441	0.98±0.02	46725	0.97±0.02	50334	0.98±0.02	49732	0.99±0.02	49206	0.99±0.02	49148	0.99±0.02

Table 2.1: Genotype concordance metrics. The calculation based on array genotypes and corresponding sequence-based genotypes obtained with different variant callers at positions where SNPs from the array were available on chromosomes 3, 6 and 28

	Non-reference genotype concordance						Precision					
	GATK	freebayes	SAMtools	GATK	freebayes	SAMtools						
	No.SNPs	Mean±SD	No.SNPs	Mean±SD	No.SNPs	Mean±SD	No.SNPs	Mean±SD	No.SNPs	Mean±SD	No.SNPs	Mean±SD
≤0.1	55	0.02±0.03	34	0.02±0.03	35	0.02±0.03	46	0.02±0.04	28	0.01±0.03	28	0.01±0.03
≤0.2	18	0.13±0.03	10	0.14±0.03	23	0.14±0.03	11	0.14±0.03	12	0.15±0.03	11	0.16±0.03
≤0.3	24	0.26±0.03	20	0.25±0.03	24	0.25±0.03	22	0.25±0.03	22	0.25±0.03	17	0.25±0.03
≤0.4	32	0.34±0.02	35	0.34±0.02	43	0.35±0.02	26	0.35±0.03	21	0.34±0.02	20	0.35±0.03
≤0.5	32	0.43±0.03	41	0.43±0.03	38	0.43±0.03	15	0.42±0.02	36	0.43±0.03	21	0.44±0.03
≤0.6	144	0.52±0.03	249	0.52±0.03	168	0.52±0.03	99	0.52±0.03	161	0.52±0.03	99	0.53±0.03
≤0.7	245	0.66±0.03	540	0.66±0.03	297	0.65±0.03	157	0.66±0.03	499	0.66±0.03	156	0.66±0.03
≤0.8	445	0.75±0.02	1725	0.76±0.03	584	0.76±0.02	270	0.75±0.02	2246	0.76±0.02	240	0.75±0.02
≤0.9	3019	0.86±0.03	6607	0.86±0.03	3720	0.86±0.03	1644	0.86±0.03	6808	0.85±0.03	1964	0.86±0.03
≤1	47312	0.98±0.03	41511	0.97±0.03	46315	0.98±0.03	49038	0.99±0.02	40941	0.97±0.03	48692	0.98±0.03

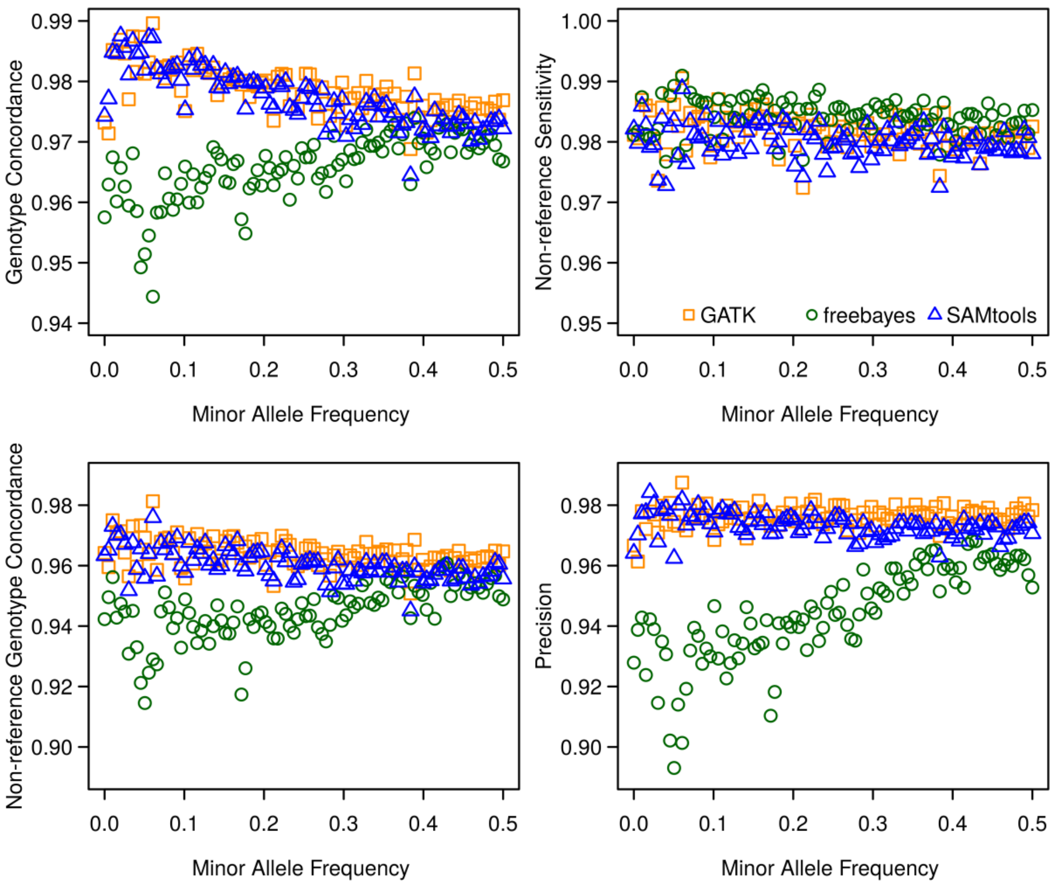


Figure 2.2: Comparison of the genotype concordance, non-reference sensitivity, non-reference genotype concordance and precision of GATK, freebayes and SAMtools over various minor allele frequency bins. SNPs were binned into 100 groups according to their array-derived MAF. The mean of each metric was calculated within each minor allele frequency bin. The statistics of different genotype concordance metrics were measured according to Linderman et.al (2014). The orange squares represent variant caller GATK. The green circles stand for variant caller freebayes. The blue triangles stand for variant caller SAMtools.

In this study, the analyses were mainly focused on the comparison of specific versions (namely the newest at the time point of performing the analyses) of different variant callers. Different versions of the same variant callers may result in a different set of variants for the same underlying sequencing data. We thus compared two different versions of freebayes. The older version of freebayes (version v9.9.2-22-gc283d6d) called 524,938 SNPs (29%) more than the newer version (version v0.9.15-1-g076a2a2) based on the same data material on chromosomes 3, 6, and 28. Thus, our ranking of callers is only valid for the specific versions used here.

It needs to be mentioned that the only alignment tool used in this study was BWA with different variant callers to make sure that variants are called based on the same basic data sets, to ensure a fair comparison of the callers. However, each step listed in the pipeline (Additional file 2.7) affects the quality of the final SNP calls, including the alignment step. Besides, it is possible that different callers combined with different alignment tools may have a different performance, an aspect which was not investigated in this study. However, BWA is one of the widely used alignment tools, which is known to have a good performance as well (Li and Durbin, 2009). Based on the results on the quality measures and regarding computation time (GATK which supports multiple threads was faster than SAMtools and freebayes, as shown in Additional file 2.8) and usefulness (i.e. quality, completeness and availability) of the software documentation, we decided to use the variants called by GATK as basis for the imputation part of this study. It turned out that there were 1,652,105 SNPs remaining on chromosomes 3, 6 and 28 totally in the brown layer chicken dataset for the imputation study. As mentioned before, there were different pipelines to deal with the whole-genome sequencing data with or without strictly filtering on genotype quality of each SNP and each individual. In this study, we did not filter genotype quality while enhancing the original genotype quality with Beagle 3.3.2 as was done in the study of Jansen et al. (2013) in order to use more SNPs in the analysis.

Imputation accuracy

Leave-one-out cross-validation

A leave-one-out cross-validation was performed for each individual that was both sequenced and genotyped (i.e. 24 out of 25 sequenced individuals) to assess imputation accuracy from SNP array data to whole-genome sequence. For chromosomes 3, 6 and 28, the genotype correlation and concordance achieved by three different imputation packages (Minimac, FImpute and IMPUTE2) are shown in Figure 2.3. Generally speaking, imputation accuracy assessed as correlation and concordance between imputed and sequence-derived genotypes within sequenced individuals was high with all imputation packages, with the performance of FImpute being slightly worse than the one of Minimac and IMPUTE2.

Over all three chromosomes, the average genotype correlation ± standard deviation between imputed and sequence-derived genotypes was 0.91±0.028 for Minimac, 0.89 ±0.028 for FImpute and 0.90 ±0.027 for IMPUTE2. These results implied that also pedigree-free imputation software (Minimac and IMPUTE2) yielded accurate genotypes for whole sequence variants in this data set. Most of the sequenced individuals in this study were contemporaries in a commercial breeding program which controls for the level of relationship and inbreeding, thus pedigree relationship among these individuals was relatively low. This may explain why imputation programs based on pedigree algorithms, such as FImpute, have no advantage in this leave-one-out cross-validation strategy.

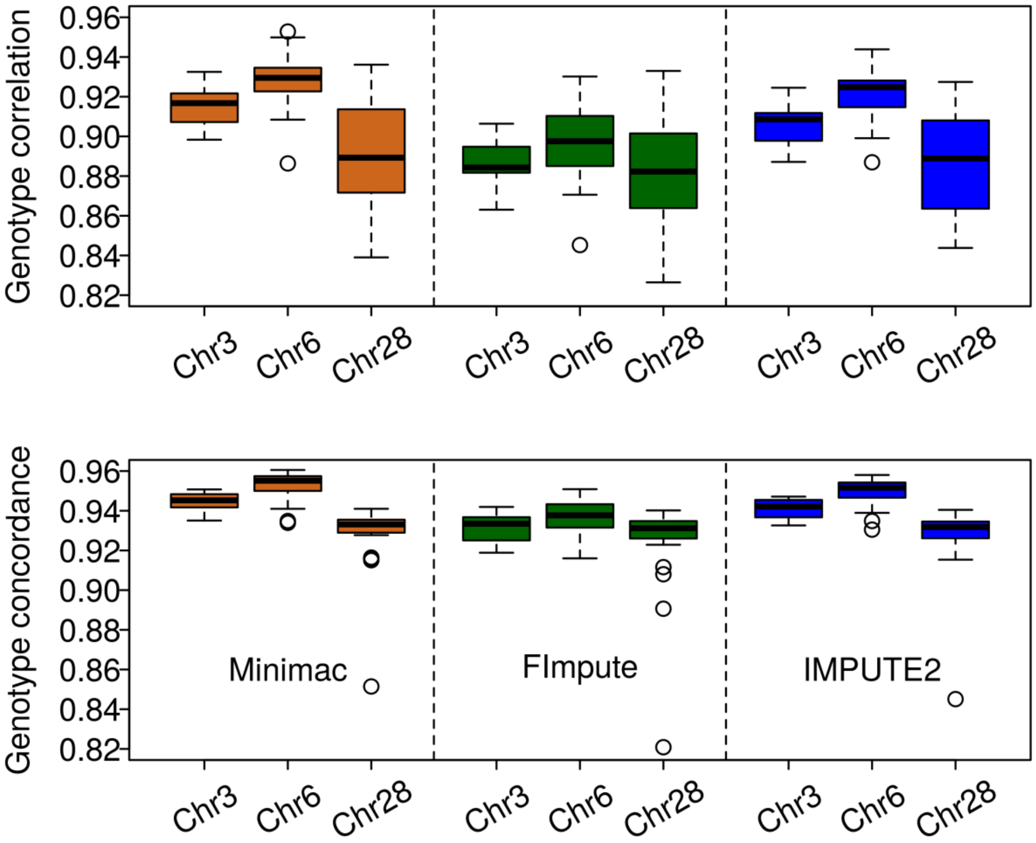


Figure 2.3: Imputation accuracy assessed by leave-one-out cross-validation. Genotype correlation (top panel) and genotype concordance (bottom panel) between the sequenced and imputed genotypes for 24 sequenced individuals with different imputing programs.

Over all three imputation programs, the average genotype correlations for SNPs on chromosomes 3 and 6 were quite similar. However, for chromosome 28, which is much smaller than the other two chromosomes studied, the average genotype correlation was slightly lower and the standard deviation was larger compared to chromosomes 3 and 6. In the study of Hancock et al. (2012), it was also found that imputation accuracy tended to be better on larger chromosomes (i.e. chromosome 1) than on smaller chromosomes (i.e. chromosome 22) in the human genome, even when there was no significant difference between these two chromosomes in typical characteristics (e.g. SNP density). The results of genotype concordance had a similar tendency as genotype correlation, but with a smaller standard deviation, particularly for chromosome 28. Overall, the imputation accuracies of different programs were largely similar in this scheme, although FImpute again performed slightly worse than the other two on chromosomes 3 and 6.

Leave-one-out cross-validation is the only strategy of assessment of imputation quality that allows taking all SNPs from sequenced data into account when calculating measures like genotype correlation. However, it should be mentioned that in most cases in practice individuals to be imputed are descendants of the sequenced individuals, which was not the case in this leave-one-out setup. Assessing whether the sequence of offspring is correctly imputed can only be done if a sample of such offspring is actually sequenced, and such data are presently not available in sufficient quantities. Thus these results of this analysis should only be extrapolated with caution to the practically most relevant case of imputing sequence of current selection candidates based on sequenced founder animals.

Genotype conflicts in father-progeny pairs

Based on the available pedigree, it is possible to apply Mendelian rules to estimate the percentage of genotype conflicts for all SNPs in father-progeny pairs (i.e. progeny’s genotype is alternatively homozygous to father’s homozygous genotype) for an imputed progeny compared to each sequenced individual. There were 134 father-progeny pairs in the available pedigree for which the father was sequenced and the progeny was imputed. The number of progenies per father varied from 1 to 44. Comparisons of the imputation performance based on the father-progeny pair conflict (Figure 2.4) show that FImpute (on average 0.01%) outperformed Minimac and IMPUTE2 clearly, which should be expected, since pedigree information is used in FImpute while both other programs are pedigree-free algorithms. Furthermore, Minimac was still much better (on average 0.11%) than IMPUTE2 which produced conflicts with 2.5% of the imputed SNPs on average. When focusing on the performance of Minimac and FImpute on each chromosome, Minimac showed better performance on the larger chromosome 3 than on smaller chromosomes, likely due to the fact that there is more recombination on the micro-chromosomes which can result in less LD (Megens et al., 2009). The results of FImpute for the three chromosomes were similar, in spite of the fact that the percentage of conflicts was slightly higher on chromosome 28 than on the other two.

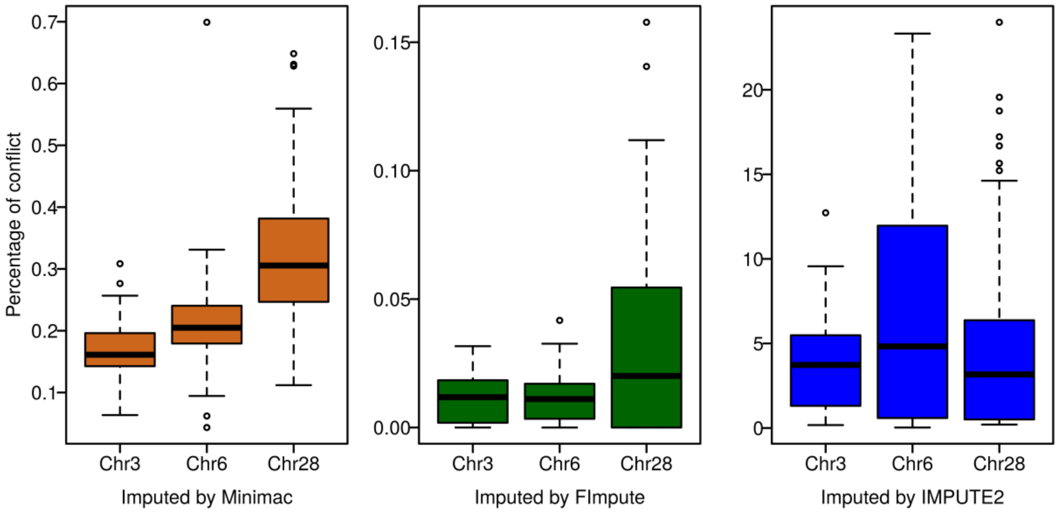


Figure 2.4: The percentage of genotype conflicts in father-progeny pairs. The conflicts were calculated within 952,826 (365,802, 37,556) imputed SNPs on chromosome 3 (6, 28) for 134 pairs of sequenced fathers and imputed progeny. Imputation was performed using minimac (left), FImpute (middle) or IMPUTE2 (right).

Imputation accuracy for randomly masked 1000 SNPs in genotyped individuals

In this scenario, the quality of imputation was assessed by the correlation between imputed and masked true genotypes per individual and/or per SNP. The average imputation accuracy of different software programs plotted for MAF bins is shown in Figure 2.5. Correlation between imputed and true genotypes per SNP over all individuals was calculated. For this, all SNPs randomly masked in all 5 replicates were binned based by their sequence-derived MAF, and the average correlation in each bin was assessed. In general, the imputation accuracy (± S.D.) of FImpute was lower (0.90±0.11) compared to Minimac (0.97±0.033) and IMPUTE2 (0.97±0.036). FImpute performed particularly poor for SNPs with a MAF smaller than 0.2. Imputation accuracies from Minimac and IMPUTE2 were comparatively stable with different MAFs, with a small reduction when MAF was low (< 0.1). Our results are in general agreement with several previous studies (e.g. in cattle (Ma et al., 2013; Bouwman and Veerkamp, 2014; Daetwyler et al., 2014) or human (Howie et al., 2011; Deelen et al., 2014) which also found that imputation accuracy decreased rapidly when MAF is low with different imputation software packages. Hence, the ability to accurately impute SNPs with low MAF is one of the most important criteria to assess the imputation programs. Usage of a diverse reference population may increase the imputation accuracy of rare variants (Deelen et al., 2014; Liu et al., 2014; Zheng et al., 2015), however, the computational burden also increases with the increase of the size of the reference population (Howie et al., 2011; Howie et al., 2012). Thus, the trade-off between the imputation accuracy and imputation efficiency needs to be considered.

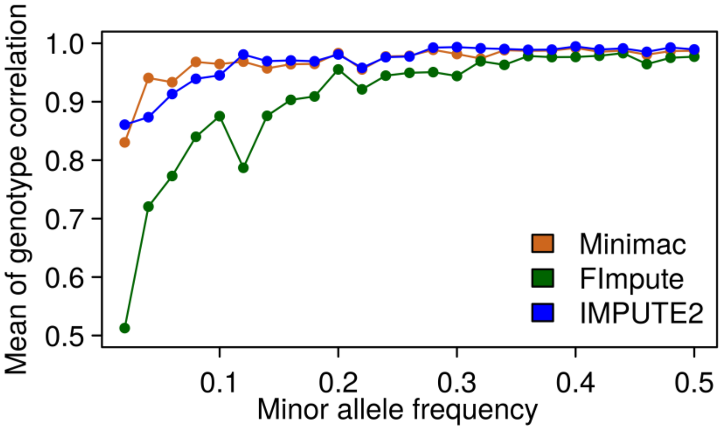


Figure 2.5: Mean of imputation accuracy of different software against minor allele frequency among 5 replications. SNP were binned by their sequence-derived MAF.

As imputation accuracy depends (amongst others) on the relationship between reference individuals and individuals to be imputed, the trend of imputation accuracy when there was a different number of generations between reference and validation individuals was investigated. The relationship between sequenced individuals and genotyped individuals, which was estimated as the percentage of genotyped individuals having a high relationship ≥ 0.25 (or 0.5) with at least one of sequenced individual is shown in Additional file 2.9. Imputation accuracy with 95% confidence interval obtained for individuals from different generations with different imputation programs is shown in Figure 2.6. Imputation accuracy measured as the correlation between original imputed and original true genotype per individual is shown in Figure 2.6A, while imputation accuracy measured as the correlation between standardized imputed and standardized true genotype per individual is shown in Figure 2.6B. Generally, imputation accuracies for all three programs based on standardized genotype were lower than based on original genotype with larger standard deviation; however, the tendency of imputation accuracy along generations was the same for both measures. In the scenario with original genotype, comparing the three imputation software studied here, IMPUTE2 showed the highest genotype correlation for individuals from all generations, while FImpute showed the lowest genotype correlation. From generations 1 to 3, the average genotype correlation increased slightly, while from generations 4 to 6 hardly any trend was observed. However, there was no significant difference between adjacent generations while there was significant increase when comparing generation 1 to generations 4, 5 and 6 respectively for Minimac and IMPUTE2, and there was a significant increase between generation 1 and 4 for FImpute. In the scenario with standardized genotype, there was a significant increase between generation 1 and generation 2. These results suggested that imputing SNP array data up to sequence level is possible with high accuracy even across several generations. Our results thus confirm results from a previous study (Heidaritabar et al., 2014) which suggested that imputation quality did not deteriorate when the imputed population was three generations away from the sequencing population. It should be mentioned that the data we used here were from a closed line (Qanbari et al. (2010) estimated the effect population size (N_e) for individuals from a commercial brown layer line cross and found a recent N_e of 70) and the results may differ in more open populations with higher effective population size, migration or variability in mating schemes.

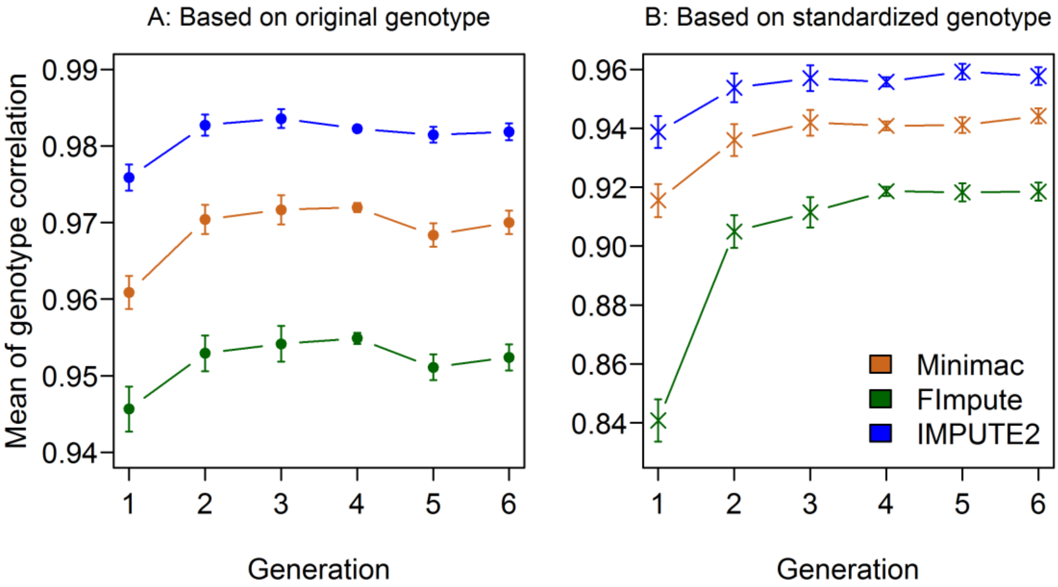


Figure 2.6: Imputation accuracy with 95% CI of masked SNPs in different generations obtained with different imputation software package. The imputation accuracy is the correlation between the sequenced and imputed genotypes which were masked as dummy genotypes on 3 chromosomes (3, 6 and 28) with 5 replications. Imputation accuracy measured as the correlation between original imputed and original true genotype per individual is shown in Figure 2.6A, while imputation accuracy measured as the correlation between standardized imputed and standardized true genotype per individual is shown in Figure 2.6B.

Conclusions

Based on data from 50 sequenced individuals from two layer lines, we compared the performance of three variant callers for a subset of SNPs (~50K) that were available from whole-genome sequencing and SNP array in 24 out of 1081 individuals that were both fully sequenced and genotyped with the 580k array. Results showed that a high proportion of SNP calls had high values in different measures of quality (amongst others genotype concordance and non-reference sensitivity) with all variant callers. GATK showed a slightly better performance than SAMtools and freebayes. We further demonstrated that three commonly used imputation programs were capable of imputing from SNP array data up to whole-genome level in a brown layer line based on a small number of sequenced individuals with substantial imputation accuracy, even across several generations. FImpute performed slightly worse than Minimac and IMPUTE2 in terms of genotype correlation, especially for SNPs with low minor allele frequency, while it yielded the lowest numbers of Mendelian conflicts in available father-progeny pairs. Imputation accuracy was lower for rare SNPs than for common SNPs, which confirmed previous results in other species. Overall, sequence imputation from a very limited number of sequenced individuals appears to yield reasonably accurate results in closed breeding populations as available in many nucleus breeding programs.

Availability of supporting data

The reference genome used for alignment was taken from a public database and is available for download from UCSC genome browser (http://hgdownload.soe.ucsc.edu/downloads.html#chicken). PLINK binary files containing genotype and map information of all variants on chromosomes 3, 6 and 28 detected by GATK in the 50 sequenced individuals are available at doi: 10.6070/H47H1GKK.

Additional file

Additional files are available online

http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-015-2059-2#Sec22

Abbreviations

SNPs: Single nucleotide polymorphisms; MAF: Minor allele frequency; INDELs: Insertion and deletion; DP: Read depth; MQ: Mapping quality; MQM: mean mapping quality of observed alternate alleles; MQMR: mean mapping quality of observed reference alleles; GC: Genotype concordance; NRS: non-reference sensitivity; NRC: Non-reference genotype concordance; LD: Linkage disequilibrium; CI: Confidence interval; SAM: Sequence Alignment/Map format; BAM file: binary version of SAM file

Acknowledgements

We thank Annett and Steffen Weigend (Friedrich-Loeffler-Institut Mariensee) for the management and DNA extraction of the samples. We are indebted to those researchers who programmed the software used in this study. This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr "Synbreed – Synergistic plant and animal breeding" (Grant ID 0315528C).

Reference

Baes, C. F., M. a Dolezal, J. E. Koltes, B. Bapst, E. Fritz-Waters, S. Jansen, C. Flury, H. Signer-Hasler, C. Stricker, R. Fernando, R. Fries, J. Moll, D. J. Garrick, J. M. Reecy, and B. Gredler. 2014. Evaluation of variant identification methods for whole genome sequencing data in dairy cattle. BMC Genomics 15:948.

Bentley, D. R. 2006. Whole-genome re-sequencing. Curr. Opin. Genet. Dev. 16:545–52.

Van Binsbergen, R., M. C. Bink, M. P. Calus, F. a van Eeuwijk, B. J. Hayes, I. Hulsegge, and R. F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 46:41.

Bird, C. P., B. E. Stranger, and E. T. Dermitzakis. 2006. Functional variation and evolution of non-coding DNA. Curr. Opin. Genet. Dev. 16:559–64.

Bouwman, A. C., and R. F. Veerkamp. 2014. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genet. 15:105.

Browning, S. R., and B. L. Browning. 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. 81:1084–97.

Calus, M. P. L., A. C. Bouwman, J. M. Hickey, R. F. Veerkamp, and H. A. Mulder. 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. Animal 8:1743–53.

Cheng, A. Y., Y.-Y. Teo, and R. T.-H. Ong. 2014. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. Bioinformatics 30:1707–13.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M.-N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858–865.

Deelen, P., A. Menelaou, E. M. van Leeuwen, A. Kanterakis, F. van Dijk, C. Medina-Gomez, L. C. Francioli, J. J. Hottenga, L. C. Karssen, K. Estrada, E. Kreiner-Møller, F. Rivadeneira, J. van Setten, J. Gutierrez-Achury, H.-J. Westra, L. Franke, D. van

Enckevort, M. Dijkstra, H. Byelas, C. M. van Duijn, P. I. W. de Bakker, C. Wijmenga, and M. A. Swertz. 2014. Improved imputation quality of low-frequency and rare variants in European samples using the “Genome of The Netherlands”. Eur. J. Hum. Genet. 22:1321–6.

DePristo, M. a, E. Banks, R. Poplin, K. V Garimella, J. R. Maguire, C. Hartl, A. a Philippakis, G. del Angel, M. a Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–8.

Drake, J. a, C. Bird, J. Nemesh, D. J. Thomas, C. Newton-Cheh, A. Reymond, L. Excoffier, H. Attar, S. E. Antonarakis, E. T. Dermitzakis, and J. N. Hirschhorn. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat. Genet. 38:223–7.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity (Edinb). 112:39–47.

Garrison, E., and G. Marth. 2012. Haplotype-based variant detection from short-read sequencing. arXiv Prepr. arXiv1207.3907:1–9.

Goldstein, D. B., A. Allen, J. Keebler, E. H. Margulies, S. Petrou, S. Petrovski, and S. Sunyaev. 2013. Sequencing studies in human genetics: design and interpretation. Nat. Rev. Genet. 14:460–70.

Grant, J. R., A. S. Arantes, X. Liao, and P. Stothard. 2011. In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. Bioinformatics 27:2300–1.

Hancock, D. B., J. L. Levy, N. C. Gaddis, L. J. Bierut, N. L. Saccone, G. P. Page, and E. O. Johnson. 2012. Assessment of genotype imputation performance using 1000 Genomes in African American studies. PLoS One 7:e50610.

Heidaritabar, M., M. P. L. Calus, A. Vereijken, M. a M. Groenen, and J. W. M. Bastiaansen. 2014. High Imputation Accuracy in Layer Chicken from Sequence Data on a Few Key Ancestors. 10th World Congr. Genet. Appl. to Livest. Prod.:2009–2011.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Sci. 52:654.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44:955–9.

Howie, B., J. Marchini, and M. Stephens. 2011. Genotype imputation with thousands of genomes. G3 (Bethesda). 1:457–70.

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

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature:931–945.

Jansen, S., B. Aigner, H. Pausch, M. Wysocki, S. Eck, A. Benet-Pagès, E. Graf, T. Wieland, T. M. Strom, T. Meitinger, and R. Fries. 2013. Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics 14:446.

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

Lam, H. Y. K., M. J. Clark, R. Chen, R. Chen, G. Natsoulis, M. O'Huallachain, F. E. Dewey, L. Habegger, E. a Ashley, M. B. Gerstein, A. J. Butte, H. P. Ji, and M. Snyder. 2012. Performance comparison of whole-genome sequencing platforms. Nat. Biotechnol. 30:78–82.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–60.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9.

Li, Y., C. J. Willer, J. Ding, P. Scheet, and G. R. Abecasis. 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34:816–34.

Linderman, M. D., T. Brandt, L. Edelmann, O. Jabado, Y. Kasai, R. Kornreich, M. Mahajan, H. Shah, A. Kasarskis, and E. E. Schadt. 2014. Analytical validation of whole exome and whole genome sequencing for clinical applications. BMC Med. Genomics 7:20.

Liu, Q., E. T. Cirulli, Y. Han, S. Yao, S. Liu, and Q. Zhu. 2014. Systematic assessment of imputation performance using the 1000 Genomes reference panels. Brief. Bioinform.

Liu, X., S. Han, Z. Wang, J. Gelernter, and B.-Z. Yang. 2013. Variant callers for next-generation sequencing data: a comparison study. PLoS One 8:e75619.

Ma, P., R. F. Brøndum, Q. Zhang, M. S. Lund, and G. Su. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. J. Dairy Sci. 96:4666–77.

Mardis, E. R. 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. 24:133–41.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome

Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–303.

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

Meynert, A. M., M. Ansari, D. R. Fitzpatrick, and M. S. Taylor. 2014. Variant detection sensitivity and biases in whole genome and exome sequencing. BMC Bioinformatics 15:247.

Morozova, O., and M. a Marra. 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics 92:255–64.

Mulder, H. A., M. P. L. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95:876–89.

O’Rawe, J., T. Jiang, G. Sun, Y. Wu, W. Wang, J. Hu, P. Bodily, L. Tian, H. Hakonarson, W. E. Johnson, Z. Wei, K. Wang, and G. J. Lyon. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. Genome Med. 5:28.

Pattnaik, S., S. Vaidyanathan, D. G. Pooja, S. Deepak, and B. Panda. 2012. Customisation of the exome data analysis pipeline using a combinatorial approach. PLoS One 7:e30080.

Pausch, H., B. Aigner, and R. Emmerling. 2013. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet Sel … 45:3.

Pausch, H., C. Wurmser, and C. Edel. 2014. Exploiting Whole Genome Sequence Data for the Identification of Causal Trait Variants in Cattle. 10th World Congr. ….

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

Qanbari, S., T. M. Strom, G. Haberer, S. Weigend, A. a Gheyas, F. Turner, D. W. Burt, R. Preisinger, D. Gianola, and H. Simianer. 2012. A high resolution genome-wide scan for significant selective sweeps: an application to pooled sequence data in laying chickens. PLoS One 7:e49525.

Rosenfeld, J. a, C. E. Mason, and T. M. Smith. 2012. Limitations of the human reference genome for personalized genomics. PLoS One 7:e40294.

Rubin, C.-J., M. C. Zody, J. Eriksson, J. R. S. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallböök, F. Besnier, O. Carlborg, B. Bed’hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, and L. Andersson. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464:587–91.

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

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15:121–32.

Sulonen, A.-M., P. Ellonen, H. Almusa, M. Lepistö, S. Eldfors, S. Hannula, T. Miettinen, H. Tyynismaa, P. Salo, C. Heckman, H. Joensuu, T. Raivio, A. Suomalainen, and J. Saarela. 2011. Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol. 12:R94.

Yu, X., and S. Sun. 2013. Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics 14:274.

Zheng, H.-F., J.-J. Rong, M. Liu, F. Han, X.-W. Zhang, J. B. Richards, and L. Wang. 2015. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. PLoS One 10:e0116487.

Chapter 3 Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture

Guiyan Ni[1], David Cavero[2], Anna Fangmann[1], Malena Erbe[1, 3], Henner Simianer[1]

[1] Animal Breeding and Genetics Group, Georg-August-Universität, Göttingen, Germany

[2] Lohmann Tierzucht GmbH, Cuxhaven, Germany

[3] Institute for Animal Breeding, Bavarian State Research Centre for Agriculture, Grub, Germany

Submitted to Genetics Selection Evolution

Abstract

Background: Applying genomic prediction with whole-genome sequencing data in animal breeding schemes has become possible with the feasibility of next-generation sequencing technologies. This was expected to lead to higher predictive ability, since these data may contain all genomic variants including causal mutations. Our objective was to compare prediction ability with high density array data and whole-genome sequencing data in a commercial brown layer line with GBLUP models using a variety of SNP weightings. **Methods:** A total of 892 chickens from a commercial brown layer line were genotyped with 336K segregating SNPs (array data) including 157K characterized as genic SNPs (i.e. SNPs in or around a gene). For these individuals genome-wide sequence information was imputed based on data from re-sequencing runs of 25 individuals, leading to 5.2M imputed SNPs (whole-genome sequencing data), including 2.6M SNPs characterized as genic. De-regressed proofs for the traits eggshell strength, feed intake and laying rate were used as quasi-phenotypic data in genomic prediction analyses. Four different weighting factors for building a trait-specific genomic relationship matrix were investigated: identical weights, $-(log_{10}P)$ from genome-wide association study results, squares of SNP effects from random regression BLUP, and variable selection based weights (known as BLUP|GA). Predictive ability was measured as correlation between DRPs and direct genomic breeding values in five replications of a 5-fold cross-validation. **Results:** Averaging over the studied traits, predictive ability with only genic SNPs in whole-genome sequencing data was 0.366 ± 0.075, which was the highest predictive ability in the current study. Genomic prediction with genic SNPs in high density array data provided the second highest accuracy (0.361 ± 0.072). The prediction with $-(log_{10}P)$ or squares of SNP effects as weighting factors for building a genomic relationship matrix or BLUP|GA did not lead to higher accuracy, compared to that with identical weights, regardless of the SNP set used. **Conclusions:** The results from this study showed that little or no benefit was gained when using all imputed whole-genome sequencing data to perform genomic prediction compared to using HD array data regardless of the different SNP weightings tested. However, higher predictive ability was observed when using only genic SNPs extracted from the whole-genome sequencing data for genomic prediction.

Background

Genomic prediction (GP) refers to a method using genomic information to obtain estimated breeding values which are subsequently used to select candidate individuals(Meuwissen et al., 2001). It has been widely implemented in livestock (Hayes et al., 2009; VanRaden et al., 2009; Daetwyler and Hickey, 2010) and plant (Daetwyler et al., 2013) breeding schemes. The feasibility of next-generation sequencing technologies has made it possible to apply GP with whole-genome sequencing (WGS) data. GP with WGS was expected to lead to higher predictive ability, since WGS data may contain all genomic variants including causal mutations. Thus, prediction is no longer depending on linkage disequilibrium (LD) between single-nucleotide polymorphisms (SNPs) and causal mutations. Furthermore, Georges (2014) claimed that WGS data can measure the SNP segregation properly, which is the drawback that commercial chips suffer from, particularly for rare SNPs. Based on a simulated study, Pérez-Enciso et al. (2015) stated that using WGS data did not increase the prediction accuracy over high density array data. In a first study using sequenced inbred lines of *Drosophila melanogaster* prediction based on whole sequence data using ~2.5 mio SNPs did not increase accuracy compared to an approach using only ~5 per cent of the segregating SNPs (Ober et al., 2012). In cattle data, Hayes et al. (2014) found that accuracy of GP was improved by only 2% with WGS data compared to the 800K array data when using BayesRC and imputed 1000 Bull Genomes data. In addition, Van Binsbergen et al. (2015) reported that GP with imputed WGS data did not lead to a higher prediction accuracy, compared to the HD array data from more than 5000 Holstein Friesian bulls. Brøndum et al. (2015) showed that the reliability of GP can be improved by adding several significant quantitative trait loci (QTLs), detected by genome-wide association studies (GWAS) of WGS data, into the regular 54K array data of cattle, especially for production traits. Thus, GP with whole-genome sequence data could be attractive, although the expectations for higher accuracies have not been realized so far with real cattle data.

In chicken, most previous studies regarding GP were based on commercial array data. For instance, Morota et al. (2014) reported that GP accuracy was higher when using all available SNPs than only using validated SNPs from partial genome (e.g. coding regions), based on the 600K SNP array data of 1,351 commercial broiler chicken. Abdollahi-Arpanahi et al. (2015) studied 1,331 chicken which were genotyped with a 600K Affymetrix platform and phenotyped for the trait body weight, and reported that predictive ability could be increased by adding the top 20 SNPs with largest effects from GWAS as fixed effects in the genomic best linear unbiased prediction (GBLUP) model. So far, studies to evaluate the predictive ability with WGS data in chicken are lacking.

Regardless the genotyping source (i.e. WGS data or array data) used, GBLUP has been widely used in genomic prediction studies. Besides GBLUP in its classical form, in which each SNP is assumed to have the same contribution to the genetic variance, several weighting factors for SNPs or parts of the SNP set were proposed to account for the genetic architecture (de los Campos et al., 2013; Zhou et al., 2014; Zhang et al., 2015). de los Campos et al. (2013) proposed a method using the $-(log_{10}P)$ from GWAS as a weighting factor for each SNP to build a genomic relationship matrix (G matrix). They observed that prediction accuracy of height was improved compared to the original GBLUP, based on around 6,000 records drawn from a public human type-2 diabetes case-control data set with a 500K SNP platform. Zhou et al. (2014) used LD phase consistency, the estimated SNP effects or both as weighting factors to build a weighted G matrix, and reported that GBLUP with those weighted G matrices did not lead to higher genomic prediction accuracy, based on 5,215 Nordic Holstein bulls and 4,361 Nordic Red bulls. With a German Holstein dataset, Zhang et al. (2015) reported that best linear unbiased prediction given genomic architecture (BLUP|GA), which puts an optimal weight on a subset of SNPs with highest effects in the training set, had similar performance as GBLUP for trait somatic cell score (SCS), but outperformed GBLUP for the traits fat percentage and milk yield. Advantages of BLUP|GA were larger when the datasets were relatively small.

The objective of this study was to compare results from genomic prediction analyses using both HD array data and WGS data performed with GBLUP models with a variety of weighting factors for specific SNPs in a purebred commercial brown layer chicken line.

Material and Methods

Data

High density array data

Individuals from a purebred commercial brown layer line were used in this study. The sample comprised 892 female and male chickens from 6 generations. These chickens were genotyped with the Affymetrix Axiom® Chicken Genotyping Array (hereinafter denoted as HD array), which initially had 580K SNPs. This HD array data were pruned by discarding sex chromosomes, unmapped linkage groups, and SNPs with minor allele frequency (MAF) lower than 0.5% or genotyping call rate smaller than 97%. Individuals with call rates smaller than 95% were also discarded. After filtering, there were 336,224 segregating SNPs for 892 individuals available in the HD array data set.

Imputed whole-genome sequence data

Data of re-sequencing runs with the Illumina HiSeq2000 technology with a target coverage of 8X were available for 25 brown layer chickens from the same population (18 of them were also genotyped with the HD array), together with another 25 white layer chickens. Those data from re-sequencing runs (brown and white layer together) were aligned to Build 4 of the chicken reference genome (galGal4) with BWA (version 0.7.9a-r786) (Li and Durbin, 2009) using default parameters for paired-end alignment and SNP variants were called using GATK (version 3.1-1-g07a4bf8, UnifiedGenotyper) (McKenna et al., 2010). Called variants (25 brown layers only) were edited for depth of coverage (DP) and mapping quality (MQ) based on the following protocols: for DP, outlier SNPs (top 0.5% of DP) were removed. Then, mean and standard deviation of DP of the remaining SNPs were calculated and SNPs with a DP above and below 3 times the standard deviation from the mean were removed as well. For MQ, SNPs with MQ lower than 30 were removed. After filtering, within the set of 25 re-sequenced brown layers there were 10,420,560 SNPs left, which were used as a reference data set to impute HD array data up to sequence level. Imputation of all genotyped individuals was then performed using Minimac3 (Howie et al., 2009) which needs pre-phased data as input. The pre-phasing procedure was done with the BEAGLE 4 package (Browning and Browning, 2007). Default numbers of iteration were used in pre-phasing and imputation. According to our previous study (Ni et al., 2015), in which we used the same population, phasing genotype data with BEAGLE 4 and further imputing with Minimac3 provided the highest imputation accuracy under different validation strategies. After imputation, post-imputation filtering criteria were applied per SNP, namely SNPs with MAF smaller than 0.5% or SNPs with imputation accuracy smaller than 0.8 were removed. The imputation accuracy used here was the Rsq measurement from Minimac3, which was the estimated value of the squared correlation between true and imputed genotype. After this step, there were 5,243,860 imputed SNPs for 892 individuals available, hereinafter denoted as WGS data.

In addition, SNPs regardless the data set were classified with gene-based annotation of ANNOVAR (Wang et al., 2010) with default parameters and galGal4 as reference genome. SNPs were categorized into a total of 9 classes (Curwen et al., 2004). Our set of genic SNPs (SNP_genic) included all SNPs from categories exonic, splicing, ncRNA, UTR5', UTR3', intronic, upstream, and downstream. There were 2,593,054 SNPs characterized as genic SNPs in WGS data (hereinafter denoted as WGS_genic data) and 157,393 SNPs characterized as genic SNPs in HD array data (hereinafter denoted as HD_genic data).

Phenotypic observations

The quasi-phenotypic data were de-regressed proofs (DRPs) for eggshell strength (ES), feed intake (FI), and arcsine transformed laying rate in the last third of the laying period (LR). The arcsine transformation of the latter trait was performed to achieve an approximate normalization. To obtain the de-regressed proofs, a single trait BLUP animal model was performed for each trait using raw phenotypic and pedigree data, respectively. Estimated breeding values from these models were then de-regressed following Garrick et al.(Garrick et al., 2009). The de-regression process included removing of the parent average information.

Genomic prediction

Genomic prediction was performed using the following GBLUP model with difference genomic relationship matrices which will be described later:

$$\boldsymbol{y} = \boldsymbol{X}\mu + \boldsymbol{Z}\boldsymbol{g} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is the vector of DRPs of individuals in the training set for a specific trait; μ is the overall mean; $\boldsymbol{g}$ is the vector of additive genetic values (i.e. genomic breeding values) for all genotyped chickens; $\boldsymbol{e}$ is the vector of residual terms; $\boldsymbol{X}$ and $\boldsymbol{Z}$ are design matrices assigning DRPs to the overall mean and additive genetic values with dimension number of individuals in the training set times number of all genotyped individuals.

The distribution of residual term $\boldsymbol{e}$ is assumed to be $\boldsymbol{e} \sim N(0, \boldsymbol{R}\sigma_e^2)$, where $\boldsymbol{R}$ is a diagonal matrix, and its diagonal elements $R_{ii} = (1 - r_{DRPi}^2)/r_{DRPi}^2$ (Su et al., 2014) if individual i in the training set, where r_{DRPi}^2 is the reliability of DRP for individual i, and σ_e^2 is the residual variance. The distribution of additive genetic values is assumed to be $\boldsymbol{g} \sim N(0, \boldsymbol{G}\sigma_g^2)$, where σ_g^2 is the additive genetic variance and $\boldsymbol{G}$ is a realized genomic relationship matrix including all genotyped individuals, which can be calculated with difference approaches resulting in difference GBLUP models.

The general approach to build a $\boldsymbol{G}$ matrix is:

$$\boldsymbol{G} \quad = \frac{\boldsymbol{MDM^T}}{2\sum_{i=1}^{m} p_i(1-p_i)},$$

where $\boldsymbol{M}$ contains the corrected SNP genotypes with individuals in rows and SNPs in columns. The elements of column i of $\boldsymbol{M}$ are $0-2p_i$ (for homozygotes of the first allele), $1-2p_i$ (for heterozygotes), and $2-2p_i$ (for homozygotes of the second allele), where p_i is the frequency of the second allele at locus i from the current data set. $\boldsymbol{D}$ is a diagonal matrix giving weights to different loci in various alternatives. An identity matrix was used ($\boldsymbol{D} = \boldsymbol{I}$) in the original GBLUP (VanRaden, 2008) which implies that all loci contribute to the variance-covariance structure equally. The resulting G matrix is denoted as $\boldsymbol{G_I}$ in the following. de los Campos et al. (2013) suggested using the corresponding $-(log_{10}P)$ from a t-test of a GWAS as weighting factors to consider the relative importance of different SNPs on a specific trait. The genomic relationship matrix including a $\boldsymbol{D}$ matrix based on this weighing factor will be denoted as $\boldsymbol{G_P}$. The corresponding P-values were derived from different GWAS models each performed for each trait of interest separately in the respective training set. In order to correct for population stratification and relationship between individuals, a principal component analysis (PCA) was performed and significance among principal components (PCs) was tested in advance with a Tracy Widom test as implemented in the program EIGENSTRAT (Price et al., 2006). Then, those PCs with P-values ≤ 10^{-100} (or ≤ 0.05) were used as fixed covariates in single SNP GWAS runs. The resulting genomic relationship matrix was denoted as $\boldsymbol{G_{P100}}$ (or $\boldsymbol{G_{P005}}$). Genomic relationship matrices with weightings based on results from single SNP GWAS may not adequately represent or may overweight regions because different SNPs can capture the effect from the same QTL due to long-range LD. However, SNP effects are not corrected for each other in a single SNP GWAS. We also investigated the usefulness of weighting the G matrix with results from a random-regression BLUP (RRBLUP) in which random SNP effects are fitted simultaneously. For matrix $\boldsymbol{G_S}$**,** we thus used the squares of the estimated SNP effects of the respective trait as weighting factors to build matrix $\boldsymbol{D}$ (as also done in (Su et al., 2014)). BLUP|GA (Zhang et al., 2015) was investigated in this study as well. To account for genetic architecture, the trait-specific genomic relationship matrix $\boldsymbol{G_Z}$ was constructed as a weighted sum of a genetic architecture matrix $\boldsymbol{S}$ and a realized relationship matrix $\boldsymbol{G_I}$ (i.e. $\boldsymbol{G_Z} = \omega\boldsymbol{S} + (1-\omega)\boldsymbol{G_I}$). The construction of the $\boldsymbol{S}$ matrix was similar to the construction of $\boldsymbol{G_S}$**,** but only based on selected SNPs according to the size of their absolute SNP effects (top%) from RRBLUP. The optimal choices for top% and ω were identified with a grid search strategy applied in the training population. The combinations for searching for optimal parameters were the same as in the original study of Zhang et al. (2015) (top% within a range of [0.05, 10] and ω within a range of [0.1, 0.99]). To make sure that the weighted $\boldsymbol{G}$ matrices were in the same scale as $\boldsymbol{G_I}$, all weighting factors were divided by their mean. To mimic the real situation best and avoid overfitting, all weighing factors in all models were derived exclusively with individuals in the respective training set. To assess whether focusing on functional information improves prediction accuracy, the original GBLUP was applied to the functional subset of the WGS data (HD array data) by building a genomic relationship matrix $\boldsymbol{G_G}$ based on WGS_genic data (HD_genic data) with weights in $\boldsymbol{D}$ being 1.

Each approach mentioned above was investigated using 5-fold random cross validation with five replications, and was applied to both WGS and HD array data. Predictive ability was measured as the correlation between the obtained direct genomic values (DGVs) and DRPs for each trait of interest. DGVs and corresponding variance components were estimated using ASReml 3.0 (Gilmour et al., 2009).

In layer breeding, genomic breeding values are especially interesting for selecting the best individuals out of full-sib families. Thus we performed the Spearman's rank correlation to evaluate the ranking full-sibs according to DRPs and DGVs in a full-sib family with 12 individuals. Results presented here were from the validation sets of the first replicate of a 5-fold cross validation.

Results and discussion

Data summary

Numbers of SNPs in different MAF bins for different data sets are shown in Figure 3.1. The difference in terms of total number of SNPs between HD array data and data from re-sequencing runs can be clearly seen in the top panel. The last bin ($0.48 < \text{MAF} \leq 0.5$) contains only half the number of SNPs since in this bin only one allele frequency class (25 out of 50 alleles) is represented while in all other bins two frequency classes (e.g. 24 and 26 out of 50 alleles in the adjacent class) are reflected. For the HD data set, the number of SNPs per MAF bin was uniformly distributed over all MAF classes, reflecting the ascertainment process in the construction of the array (Kranis et al., 2013). In contrast, for data from re-sequencing runs of the 25 sequenced individuals, the number of SNPs per bin decreased with increasing MAF. SNPs with very small MAF are not that extremely overrepresented in the re-sequenced set as in other studies with sequenced data (Fujimoto et al., 2010; Eynard et al., 2015), which may be due to the following two reasons. First, the size of reference data set was relatively small (25 chickens). Thus, some extremely rare variants might not be captured. Second, the commercial broilers and layers have been subject to an intensive within-line selection, which could reduce the genetic diversity dramatically, further resulting in the lacking of rare SNPs (Muir et al., 2008). The number of SNPs in different MAF bins in the WGS data set before and after post-imputation filtering is shown in the bottom panel of Figure 3.1. Unlike Van Binsbergen et al. (2015), in which 429 sequenced individuals from several cattle breeds were used as a reference set for imputation process, we did not observe a clear U-shaped distribution of MAFs in the imputed WGS data. This means that some of the rare variants in the re-sequenced individuals were either not present in all other individuals of the population or got lost during the imputation process, partly caused by the poor imputation accuracy for SNPs with low MAF (Hickey et al., 2012; Calus et al., 2014). Starting from more than 9 million SNPs after imputation (monomorphic SNPs excluded), 200,679 SNPs were filtered due to low MAF and 85% of those filtered SNPs had low imputation accuracy (Rsq of minimac3 < 0.8) as well, which makes SNPs with low MAF even less represented in the SNP set. Further, 1.3 million SNPs among the imputed SNP set which have passed the MAF criteria were filtered due to low imputation accuracy alone, which were evenly distributed over all MAF bins. In total, more than 50% of SNPs were filtered due to low imputation accuracy in the most left three MAF bins ($0 < \text{MAF} \leq 0.06$). The fact that we found high rates of low Rsq values within the set of SNPs with low MAF could be due to low linkage disequilibrium of these SNPs with adjacent SNPs. This can lead to lower imputation accuracy (Additional file 3.1 showing imputation accuracy in different MAF bins) (Ma et al., 2013; Daetwyler et al., 2014; Deelen et al., 2014; Liu et al., 2014; Zheng et al., 2015). Filtering a large amount of SNPs with low MAF (in many cases, because imputation accuracy is too low) could weaken the advantage of imputed WGS data which contains a large number of rare SNPs (Georges, 2014), although GP with all imputed SNPs without quality-based filtering did not improve the prediction accuracy in our case (results not shown).

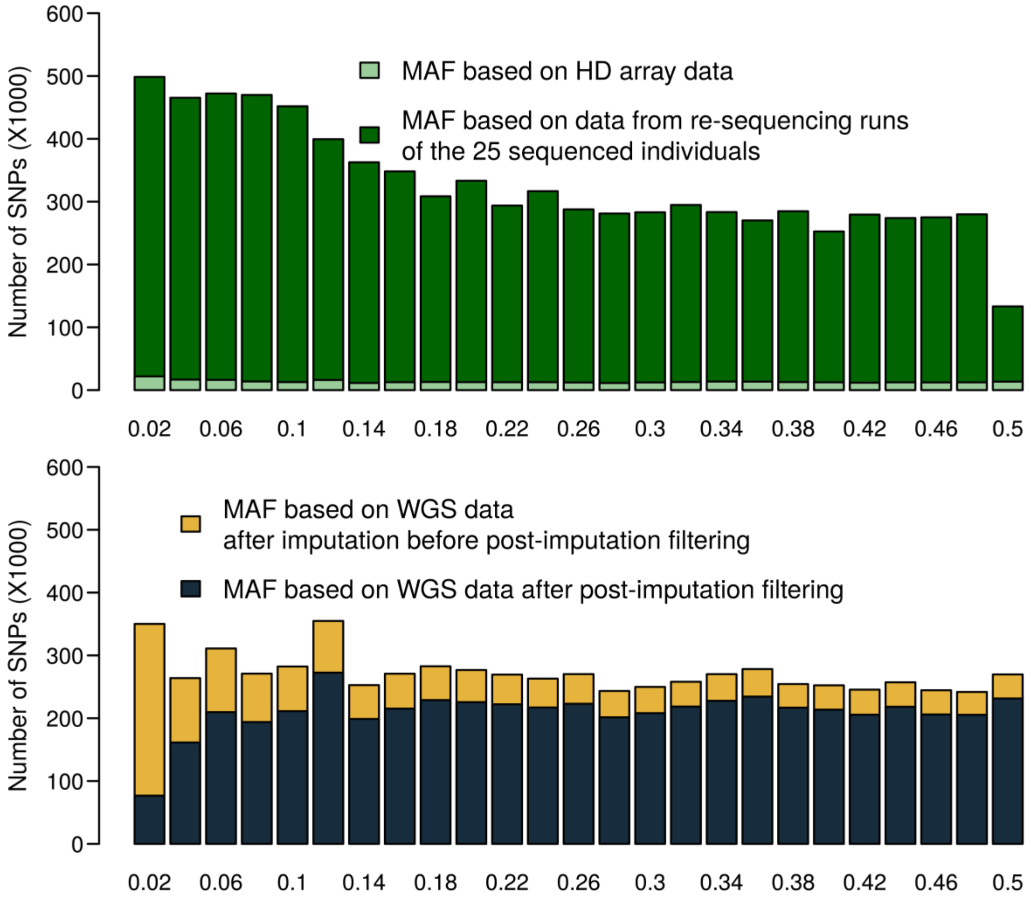


Figure 3.1: Number of SNPs (x1000) in each MAF bin for high density (HD) array data and data from re-sequencing runs of the 25 sequenced individuals (top), and for imputed whole-genome sequence (WGS) data after imputation and after post-imputation filtering (bottom). The values on the x-axis are the upper limit of the respective bin.

Comparison between HD array data and WGS data using different weighting factors

Predictive ability with GBLUP under different weighting factors based on HD array data and WGS data is shown in Figure 3.2 for the traits ES, FI, and LR, respectively. The predictive ability was defined as the correlation between DGVs and DRPs of individuals in the validation set. Generally speaking, predictive ability could not be enhanced clearly when using WGS data compared to using HD array data regardless of the different weighting factors studied. However, genomic prediction with genic SNPs annotated in WGS data gave the highest predictive ability averaged over the three studied traits.

Averaging over three traits, the predictive ability ± standard deviation for the original GBLUP was 0.353 ± 0.074 based on HD array data and 0.358 ± 0.076 based on WGS data. In the case that employing $-(log_{10}P)$ (with P-values from GWAS with different covariates in the model) as weighting factors, predictive abilities were 0.352 ± 0.062

(0.347 ± 0.072) for $\boldsymbol{G_{P100}}$ ($\boldsymbol{G_{005}}$) based on HD array data and were 0.356 ± 0.062 (0.354 ± 0.073) based on WGS data. Unlike SNP effects estimated from RRBLUP, in which effects are assessed simultaneously, SNP effects were estimated independently in GWAS. Thus, effects of a group of SNPs which present the same QTL could not be fitted simultaneously and thus the overall weighting of a region might depend on the marker density. de los Campos et al. (2013) studied a public human type-2 diabetes case-control data set containing genotype data from a 500K SNP platform and around 6,000 phenotype records. They reported that the predictive reliability (square of predictive ability) with a prediction model weighted by $-(log_{10}P)$ was increased by a factor of 110% compared to that with the original GBLUP. Similarly, Su et al. (2014) reported that predictive ability using $-(log_{10}P)$ as weighting factors was higher than the original GBLUP, based on more than 5,000 Nordic Holstein bulls which were genotyped with the Illumina Bovine SNP50 BeadChip. However, the improvement of predictive ability by employing $-(log_{10}P)$ as weighting factors in GP was not observed in our dataset.

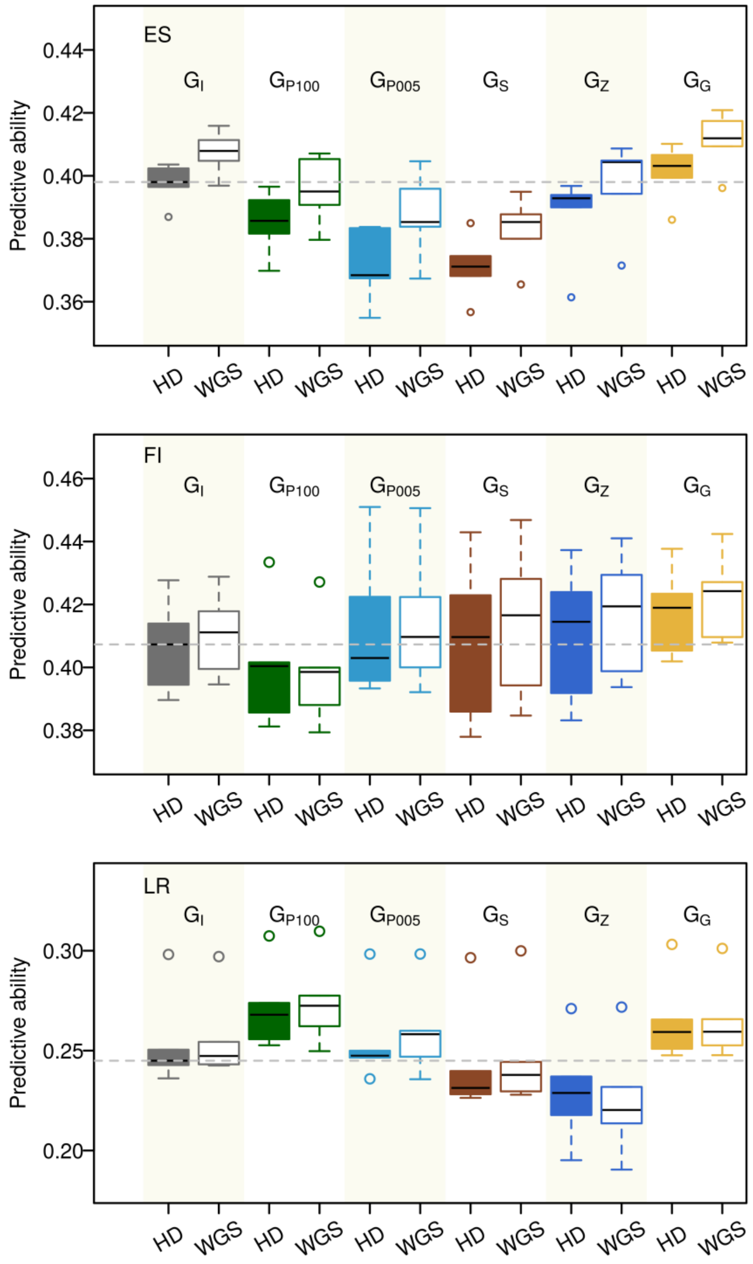


Figure 3.2: Predictive ability with GBLUP using different weighting factors based on high density array data and whole-genome sequencing data. The predictive ability was measured as correlation between direct genomic breeding values (DGVs) and de-regressed proofs (DRPs) in the validation set. Results are for the traits eggshell strength (ES), feed intake (FI) and arcsine transformed laying rate in the last third of the laying period (LR). HD stands for high density data and WGS stands for whole-genome sequencing data. $\boldsymbol{G_I}$ stands for the original GBLUP, $\boldsymbol{G_P}$ represents the model with $-(log_{10}P)$ of GWAS as weighting factors where principal components being significant with P-values ≤ 1E-100 (≤ 0.05) were used as covariates in the GWAS model, denoted as $\boldsymbol{P}_{100}$ ($\boldsymbol{P}_{005}$). $\boldsymbol{G_S}$ stands for the GBLUP model with the squares of estimated SNPs effect as weighing factors. $\boldsymbol{G_z}$ stands for results from BLUP|GA as described in the methods section. $\boldsymbol{G_G}$ stands for the original GBLUP but only based on genic SNPs. The dashed horizontal line denotes the median predictive ability of GBLUP with HD data as a reference. Note that all the outliers for trait LR were from the same replicate.

Furthermore, using the squares of SNP effects as weighting factors in GBLUP ($\boldsymbol{G_S}$) gave slightly lower predictive ability compared to the original GBLUP, in both analyses based on HD array data and on WGS data, respectively, as shown in Figure 3.2. For $\boldsymbol{G_S}$, averaging over the 3 traits, predictive ability was 0.341 ± 0.076 based on HD data and 0.348 ± 0.078 based on WGS array data, compared to 0.353 ± 0.074 (for HD array data) and 0.358 ± 0.076 (for WGS data) with the original GBLUP. These results are in agreement with Su et al. (2014), who described that GBLUP with the squares of SNP effects as weighting factors did not improve the predictive ability compared to the original GBLUP or compared to the model with $-(log_{10}P)$ as weighting factors. The lack of improvement of predictive ability when using the squares of SNP effects as weighting factors might be due to the following reasons: First, this could be due to sequencing or imputation errors. In this study, the most probable genotypes imputed from Minimac3 instead of genotype probabilities were used as WGS data, which does not account for the uncertainty of imputation. Second, the noise and uncertainty of estimated SNP effects could also bias the predictive ability (Su et al., 2014). DGVs of the training population were assigned to millions of SNPs (Figure 3.3 and Additional file 3.2). Thus the effect of each SNP was tiny. The prediction error of a SNP effect, however, might be even larger than the SNP effect itself. In addition, the size of the training set was relatively small which could further enhance the uncertainty of SNP effects. The accumulation of the above two reasons could lead to lower predictive ability, since the DGV of individual i is the summation of estimated SNP effects times its genotypes (i.e. $DGV_i = \sum_{k=1}^{m} X_{ik}\beta_k$).

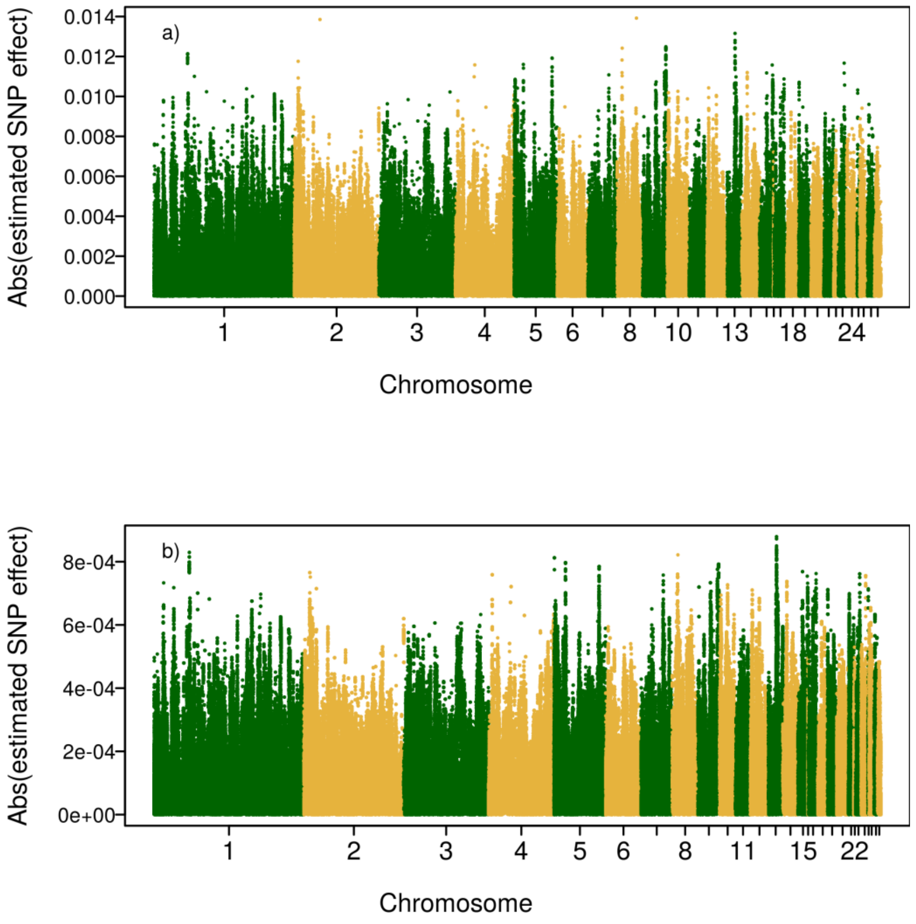


Figure 3.3: Manhattan plot of absolute estimated SNP effects for trait eggshell strength based on high density (HD) array data (a) and whole-genome sequence (WGS) data (b), respectively. SNP effects were obtained from RRBLUP in the training set of the first replicate.

With BLUP|GA, the predictive ability was 0.342 (± 0.085) based on HD array data and 0.346 (± 0.091) based on WGS data averaged over three traits of interest (Figure 3.2). Generally speaking, BLUP|GA did not improve predictive ability with WGS or HD data, compared to the original GBLUP. Zhang et al. (2015) reported that BLUP|GA outperformed the original GBLUP for production traits (i.e. fat percentage, milk yield) in a German Holstein cattle population, while it had a similar performance as GBLUP for the trait SCS. A well-known candidate gene DGAT1 has a big influence on fat percentage (Grisart et al., 2002; Winter et al., 2003), while for SCS no major genes are known. This suggests that BLUP|GA is especially useful when there are QTL regions in the genome that heavily influence the trait. The genetic architecture of ES, FI, and LR seems to be more similar to SCS than to fat percentage which means no large candidate genes have been found yet and also no large SNP effects have been found in GWAS runs performed in this study (Additional file 3.3). The SNP effects estimated from RRBLUP based on HD array data and WGS data are in Figure 3.3, which further illustrates that ES, FI, and LR are controlled by numerous SNPs with tiny effects.

When focusing on the training stage of BLUP|GA, the burden of calculation was huge to identify the optimal combination for parameters top% and ω with a grid strategy. Prediction abilities of BLUP|GA in the training stage are displayed in Figure 3.4 for each parameter combination exemplarily for the first fold of the first replicate. The combination of large ω and small top% tended to give lower predictive ability. With the increasing of top% and decreasing of ω, predictive ability tended to increase. In most of cases, the optimal option for ω based on HD data and WGS data were 0.1 in our study, which is the minimal ω we analyzed. The optimal option for top% were 10%, which is the maximal top% we analyzed, and is different from the findings of Zhang et al. (2015). They tended to select a smaller top% while there was no obvious pattern in the selection of ω. The optimal combination in each 5-fold cross-validation of each replicate for each trait is shown in Additional file 3.4 and 3.5. It should be noted that, as described in Zhang et al. (2015), accuracy of GP based on the optimal parameters obtained from training stage by cross validation may not lead to the highest accuracy in the application stage.

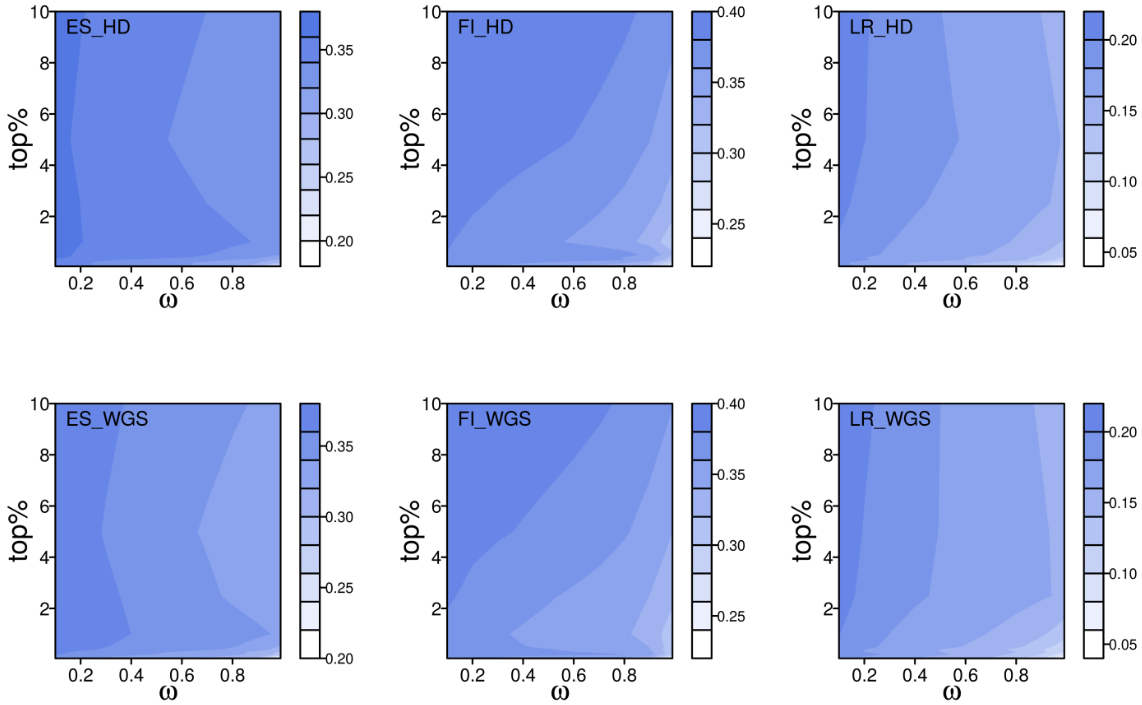


Figure 3.4: Predictive ability of the best linear unbiased prediction given genetic architecture (BLUP|GA) in the training stage to select the optimal parameter combination for the application stage. Predictive ability in this figure is the mean correlation between direct genomic breeding values (DGV) and de-regressed proofs (DRP). The first row is for high density (HD) array data, while the second row is for whole-genome sequence (WGS) data. The x-axis stands for the overall weighting factor *ω*; y-axis stands for the percentage of SNPs selected based on the SNP effects (top%); different colors stand for different levels of predictive ability.

Averaging over three traits, the predictive ability ± standard deviation was 0.366 ± 0.075 based on the WGS_genic data and was 0.361± 0.072 based on HD_genic data, compared to 0.353 (HD array data) and 0.358 (WGS data), which means GP with WGS_genic offered the highest predictive ability in our study. Similarly, Do et al. (2015) reported that predictive ability was increased with only SNPs in genes for residual feed intake based on 1,272 Duroc pigs which were genotyped with 60K SNP chip, although the increase was not significantly different from that with 1,000 randomly SNPs. In chicken, Morota et al. (2014) studied predictive ability with 1,351 commercial broiler chickens which were genotyped with Affymetrix 600K chip and found that prediction based on SNPs in or around genes did not lead to a higher accuracy using kernel-based Bayesian ridge regression. In our data set, predictive ability with HD_genic data was slightly higher than that with all HD data. Furthermore, the benefit was observed when using WGS_genic. This could be due to that using only genic SNPs reduces the noise in WGS data and might increase the chance to identify the potential causal mutations. Koufariotis et al. (2014) found that significant SNPs in the GWAS were enriched in coding region based on 17,425 bulls and cows of Holstein or Jersey which were genotyped with 777K Illumina Bovine HD array. The enrichment of significant SNPs could further imply that using genic SNPs can help us to achieve higher predictive ability.

Comparison within a full-sib family

To get an insight into the ranking of 12 full-sibs within a family according to DRPs and DGVs, DGVs predicted in validation sets with different G matrices in the first of the five replicates of the cross-validation runs is shown in Figure 3.5 for ES (Additional files 6-7 for trait FI and LR). The higher the rank correlation is, the higher the possibility is to select the same candidates. Based on HD array data, DGVs from different weighting models had a relative high rank correlation with that from $\boldsymbol{G_I}$ (from 0.88 to 0.97 for ES). This suggested that the same candidate tended to be selected in different models. Likewise, the rank correlations based on WGS data were relatively high as well, with minimal values 0.91 between $\boldsymbol{G_G}$ and $\boldsymbol{G_{P005}}$. In addition, Spearman's rank correlation between $\boldsymbol{G_I}$ based on HD array data and that based on WGS data was 0.98. Spearman's rank correlation between $\boldsymbol{G_G}$ with WGS_genic data and $\boldsymbol{G_I}$ with WGS data was 0.99, indicating that there is hardly any difference in selecting candidates based on HD array data, or WGS data, or WGS_genic data with GBLUP. Generally, the same set of candidates tended to be selected regardless of the data set (HD array data or WGS data) and weighting factors (identity weights, the squares of SNPs effect, or P-values from GWAS) used in the model. When comparing the DGVs from different models with DPRs, the Spearman's rank correlations were modest (from 0.38 to 0.54 with HD data, from 0.31 to 0.50 with WGS data) and within the expected range considering the overall predictive ability obtained in the cross-validation study (see Figure 3.2). Even though DGVs from different models were highly correlated, Spearman's rank correlation of the respective DGVs to DRPs varied clearly. This fact, however, should not be overvalued regarding the small sample size we used here (n=12).

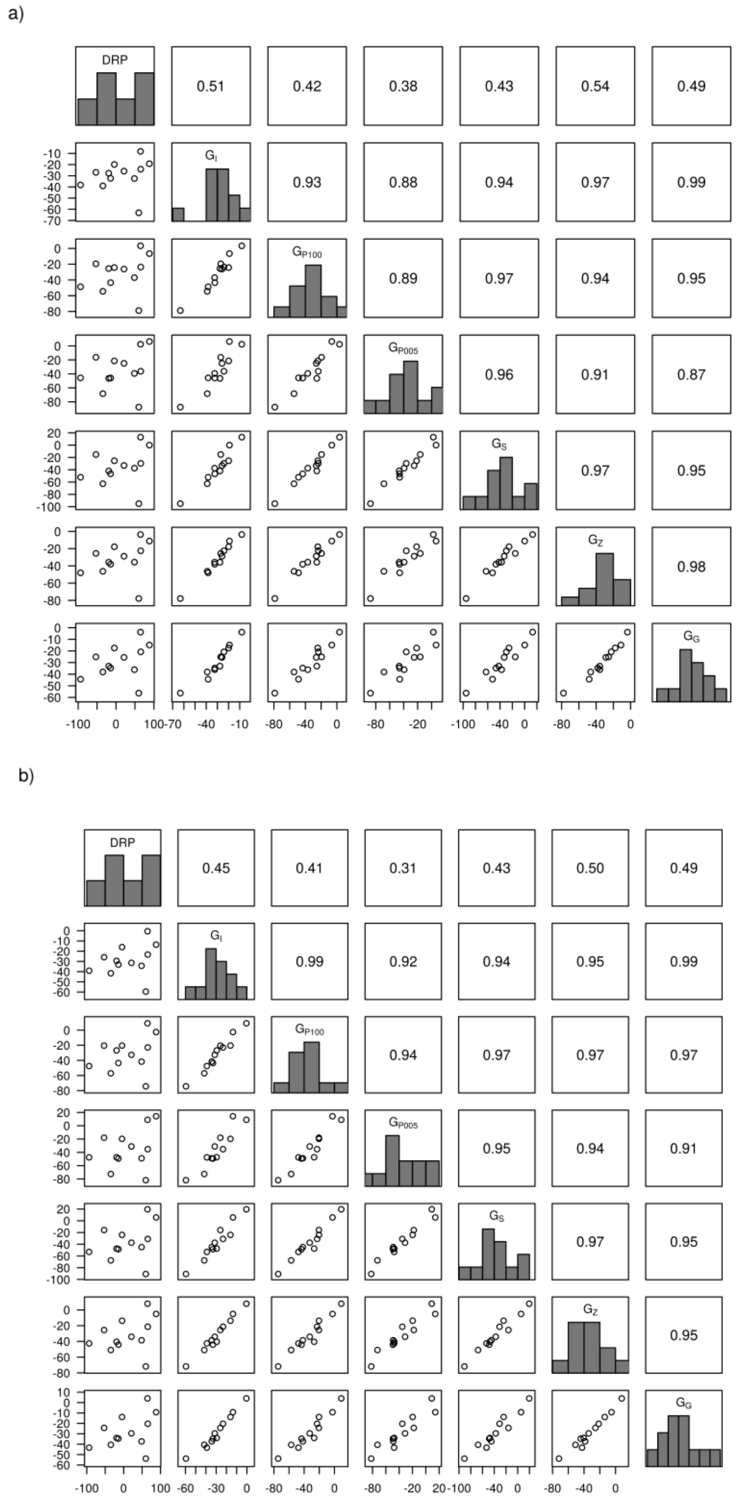


Figure 3.5: Predictive ability in a full-sib family with 12 individuals for eggshell strength based on high density (HD) array data (a) and whole-genome sequence (WGS) data (b) of one replicate. In each plot matrix, the diagonal shows the histograms of DRPs and DGVs obtained with various G matrices. The upper triangle shows the Spearman’s rank correlation between DGVs with different G matrices and with DRPs. The lower triangle shows the scatter plot of DGVs with different G matrices and DRPs.

Perspectives and implication

Using WGS data in GP was expected to lead to higher predictive ability, since WGS data should include the causal mutations that influence the trait and prediction is no longer limited by LD between SNPs and causal mutations. On the contrary to this expectation, little gain was found in our study. One possible reason could be that QTL effects were not estimated properly, due to the relatively small dataset (892 chickens) with imputed WGS data (Druet et al., 2014). Imputation has been employed widely in many livestock (Mulder et al., 2012; Segelke et al., 2012; Ma et al., 2013; Chen et al., 2014), however, the magnitude of the potential imputation errors was difficult to detect. In fact, Van Binsbergen et al. (2015) reported based on data of more than 5000 Holstein Friesian bulls that predictive ability was lower with imputed HD array data than that with the actual genotyped HD array data, which confirms our assumption that imputation could lead to lower predictive ability. In addition, discrete genotype data were used as imputed WGS data in this study, instead of genotype probabilities which can account for the uncertainty of imputation and may be more informative (Kutalik et al., 2011). At present, sequencing all individuals in a population is not realistic. In practice, there is a trade-off between predictive ability and cost efficiency. When focusing on the post-imputation filtering criteria, the threshold for imputation accuracy was 0.8 in our study to guarantee the high quality of the imputed WGS data. Numerous rare SNPs, however, were filtered due to the low imputation accuracy as shown in Figure 3.1 and Additional file 3.1. This could increase the risk of excluding rare causal mutations. Ober et al. (2012), however, did not observe the increase of predictive ability for starvation resistance with rare SNPs included in the GBLUP based on ~2.5 million SNPs determined in *Drosophila melanogaster*. Further investigation needs to be done in chicken.

A further reason why we do not observe any increase in predictive ability when using WGS data could be that we did not apply variable selection. The density of WGS data was around 15 times higher than of HD array data, which increased the linkage disequilibrium between SNPs. Thus, QTL effects were assigned to more SNPs in WGS data than in HD array data, which could be overcome by variable selection. Su et al. (2014) reported that reliability of GP increased by more than 5% when grouping 30 adjacent SNPs. In each group, a common weight was assigned reflecting the mean over the SNP effects in the same group. In addition, Brøndum et al. (2015) reported that the reliability of GP can be improved by adding several significant QTLs into the regular 54K array data of cattle. In our study, twenty top SNPs were selected according to their estimated effects from

RRBLUP or $-(log_{10}P)$ of GWAS and used as fixed effect in GBLUP. However, in the present study, this did not lead to an improvement of predictive ability (results not shown). GP with genic SNPs in WGS (the WGS_genic data) offered the highest predictive ability, comparing to that with all SNPs in WGS data. This implies that selecting the proper variables could help us to reduce noise and increase predictive ability. Thus, with increasing knowledge about gene networks and pathways, blending biological knowledge based on gene annotations and complex interactions may provide insights to guide GP (Snelling and Cushman, 2013).

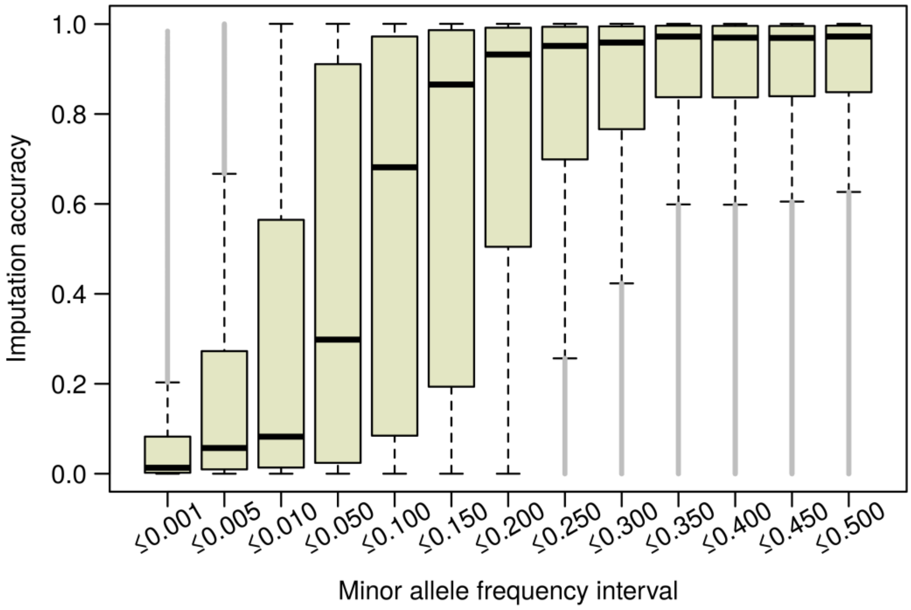
Our fourth possible explanation for the small improvement with WGS data refers to the population structure. Commercial chickens have been subject to an intensive within-line selection, which has a strong effect on the population structure. Macleod et al. (2014) studied the accuracy of genomic prediction based on WGS data for two simulation populations with different demographic history. They found that in a highly selected population with small effective population size there was almost no gain in prediction accuracy when using WGS data compared to HD data, which is in agreement with our findings.

The use of incomplete whole-genome sequence information could also weaken the predictive ability of WGS data. First, in most studies, sexual chromosomes were disregarded in the GP scheme, considering that the inheritance of sexual chromosomes differs from autosomes and the density of SNPs and LD structure on sexual chromosomes is lower compared to autosomes in commercial SNP chips. However, recent studies have discovered an increasing number of genes on heterosomes affecting economic traits: For example, Su et al. (2014) found that including sexual chromosomes in the GP scheme could increase the predictive ability averaging over 15 traits which were included in the Nordic Total Merit index (e.g. milk yield, fat yield). Second, WGS data, technically, includes all the DNA variations (e.g. CNV, INDELs), but, the studies in GP of livestock so far are generally focusing on SNPs. However, according to previous studies (McCarroll et al., 2005; Redon et al., 2006), CNV and other types of structural variation play an important role in gene expression and phenotypic variation. Third, the chicken karyotype consists of 39 chromosomes, however, data from re-sequencing represent only 30 chromosomes and two linkage groups since the reference genome was not available for some of the microchromosomes which are also supposed to be gene-rich (Mcqueen et al., 1998; International Chicken Genome Sequencing Consortium, 2004). Beyond that, chromosome 16, which hosts the chicken major histocompatibility complex, is included in the reference sequence but has a low marker density (Kranis et al., 2013) and the quality of the reference sequence is expected to be inferior due to the high genetic variability. Further, non-nuclear DNA hosted in the mitochondria is also not accounted for. In general, further work is necessary to assess the importance of the entire DNA variation on the predictive ability in chicken.

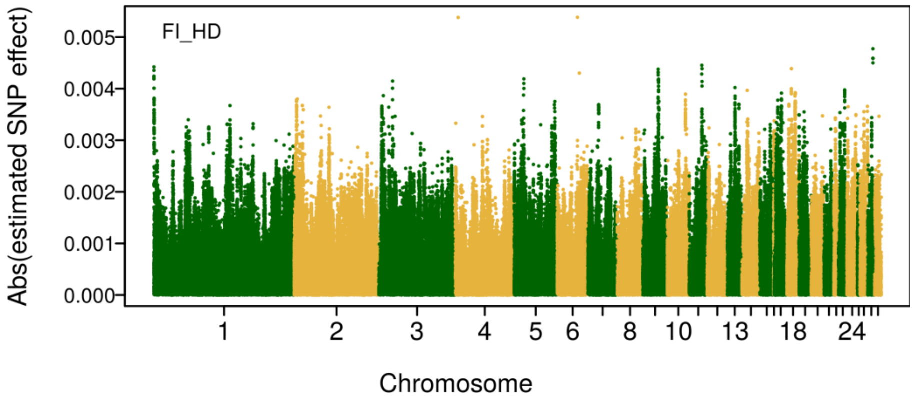
Conclusion

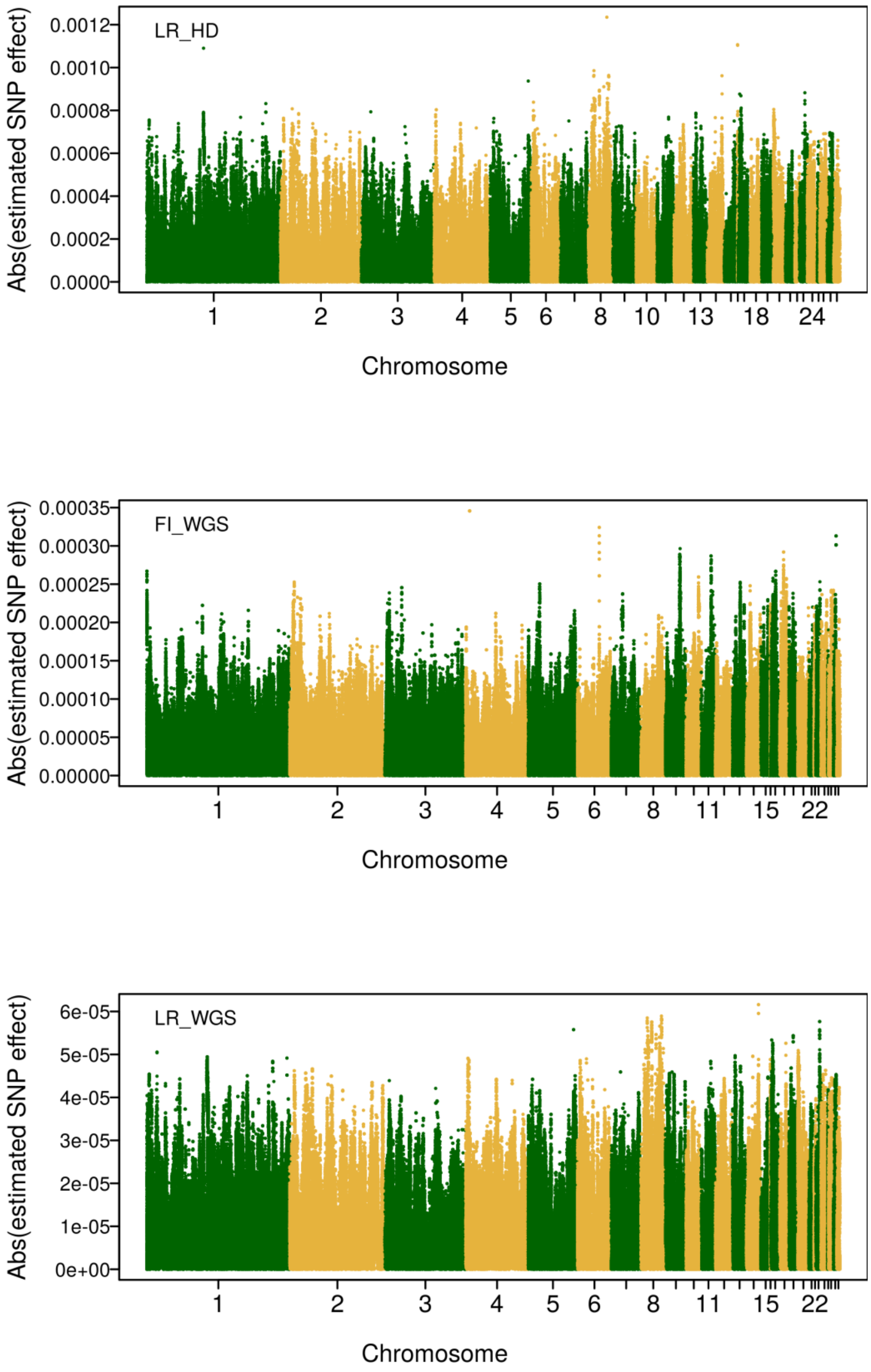
In this study, we compared ability of genomic prediction using both high density array data and imputed whole-genome sequencing data. Further comparisons were performed based on GBLUP with different genomic relationship matrices to account for genetic architecture of the three traits eggshell strength, feed intake, and laying rate. The results show that little or no benefit was gained when using all imputed WGS data compared to using HD array data with different weighting approaches in the GBLUP model. However, our results provide evidence that using genic SNPs for genomic prediction has the potential to improve the predictive ability both with HD and WGS data. Overall the same candidates tend to be selected from a full-sib family of interest regardless of the genotype data and weighting factors used.

Additional file

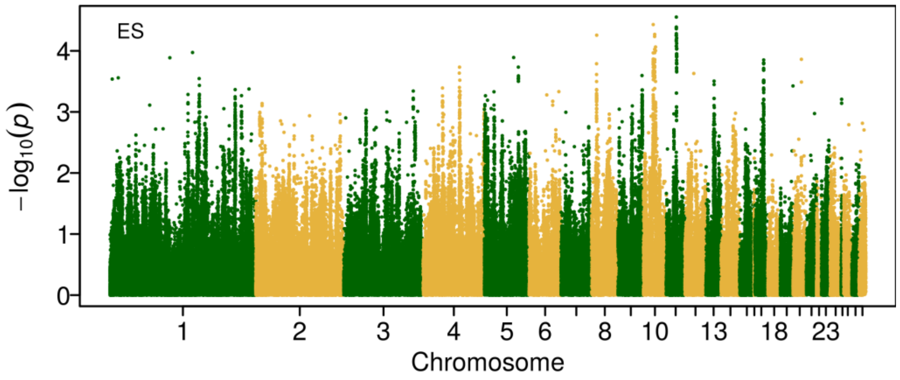


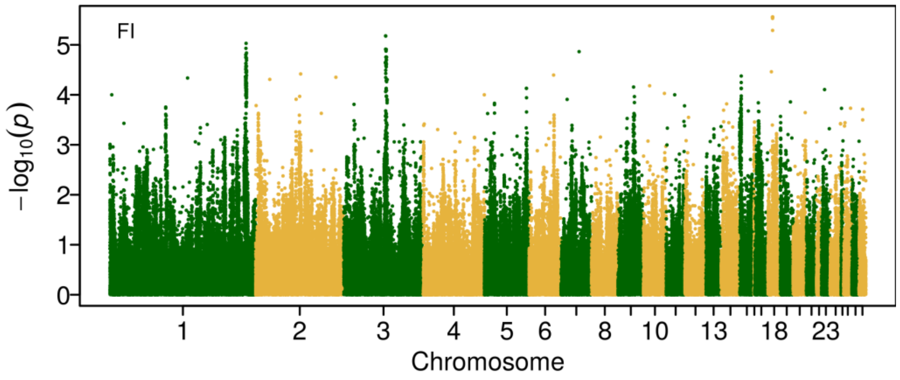
Additional file 3.1: Imputation accuracy (Rsq of Minimac3) in each minor allele frequency (MAF) interval.

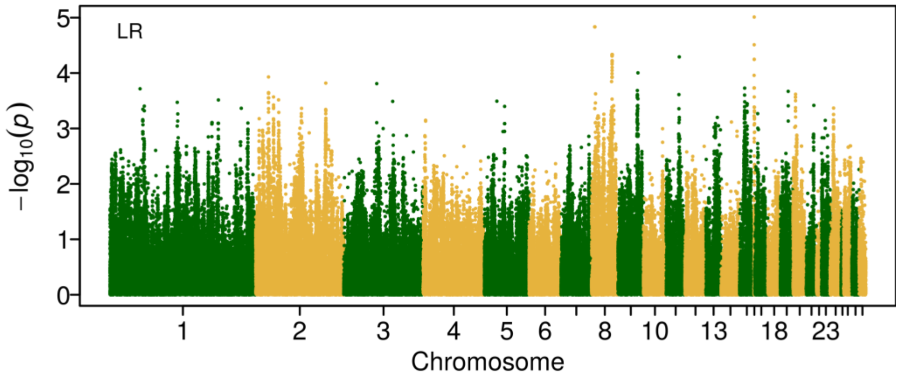


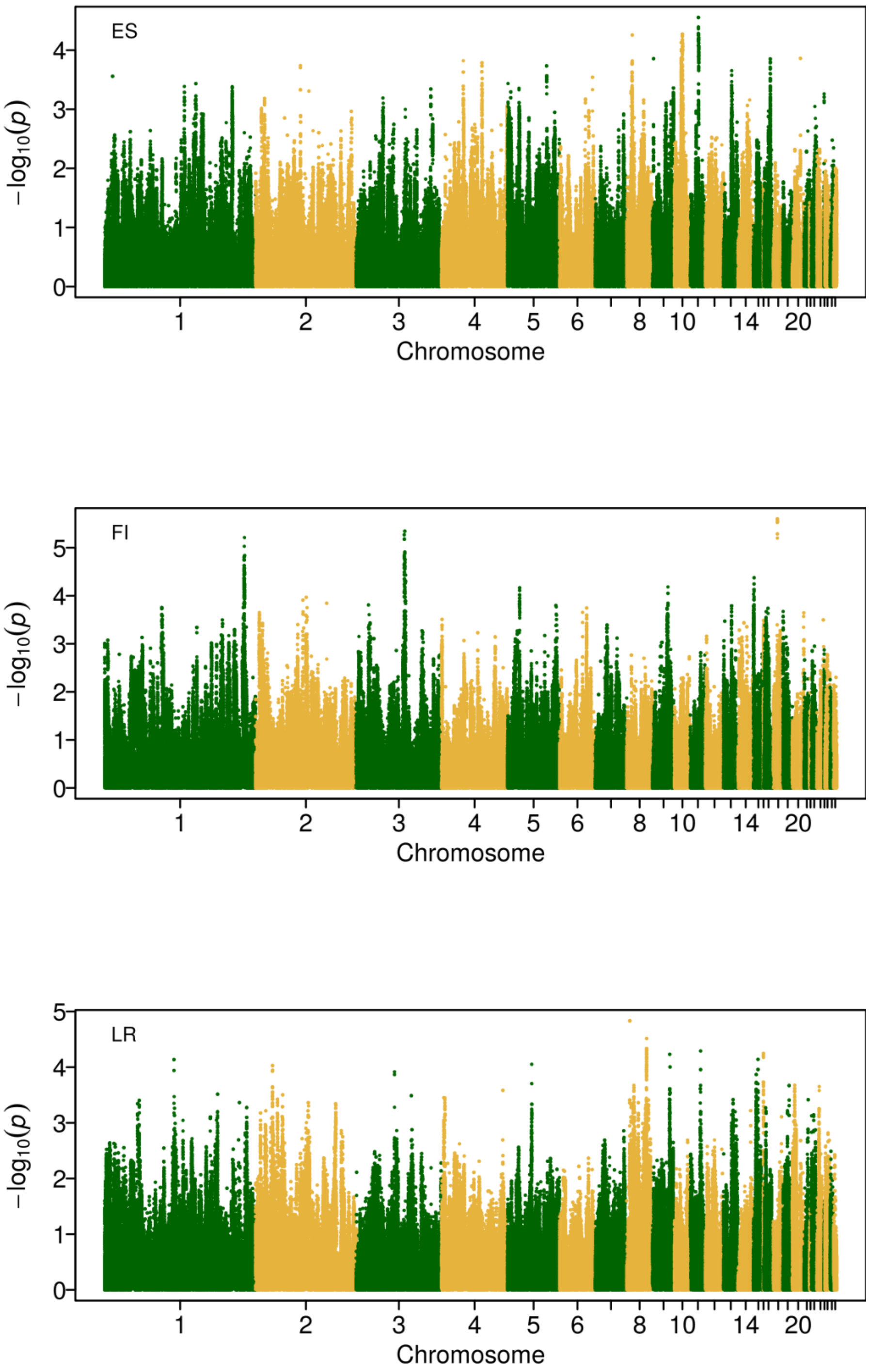


Additional file 3.2: Manhattan plot of absolute estimated SNP effects for traits FI and LR based on high density (HD) array data and whole-genome sequence (WGS) data respectively.


ES
$-\log_{10}(p)$
0
1
2
3
4
1
2
3
4
5
6
8
10
13
18
23
Chromosome


FI
$-\log_{10}(p)$
0
1
2
3
4
5
1
2
3
4
5
6
8
10
13
18
23
Chromosome


LR
$-\log_{10}(p)$
0
1
2
3
4
5
1
2
3
4
5
6
8
10
13
18
23
Chromosome



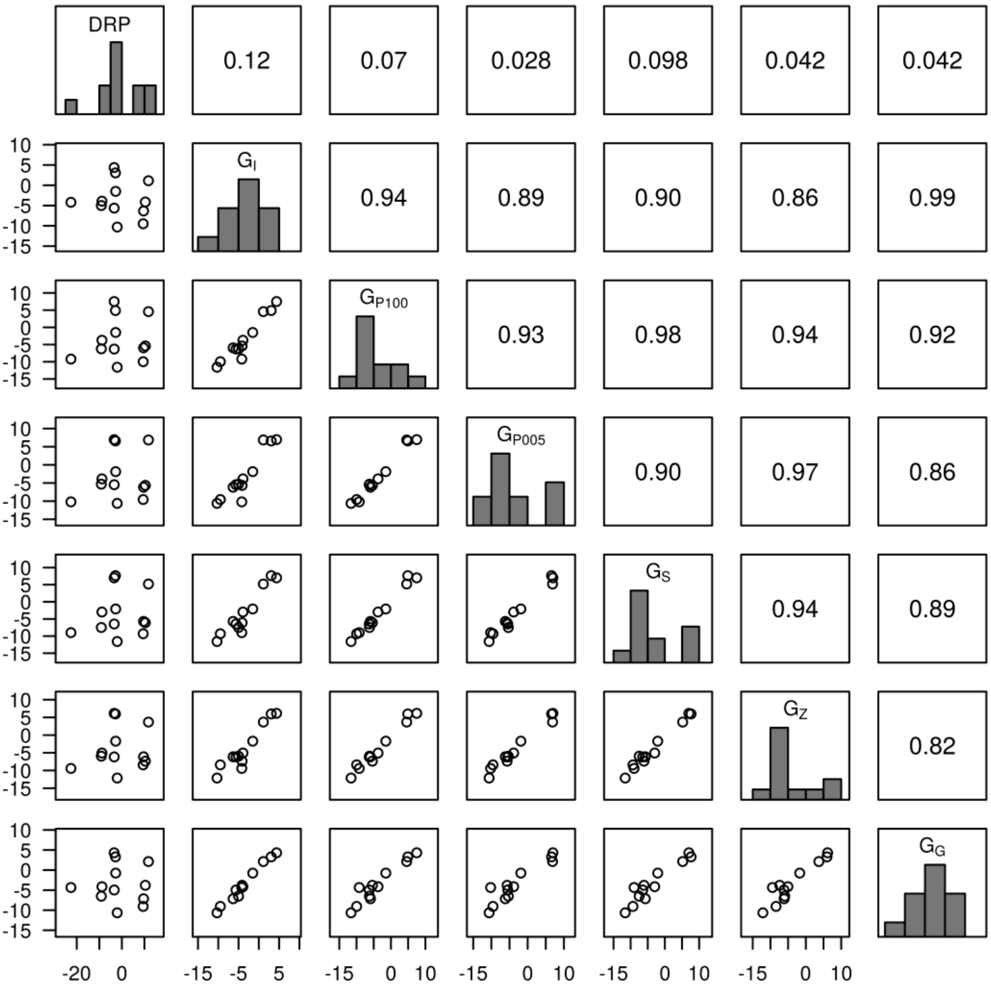
Additional file 3.3: Manhattan plots of $-(log_{10}P)$ for the three traits based on high density array data (panels1-3) and the whole-genome sequence (WGS) data (panels 4-6). Significance among principal components (PCs) was tested in advance with a Tracy Widom test and PCs with P-values ≤ 0.05 were used as fixed covariates in single SNP GWAS runs.

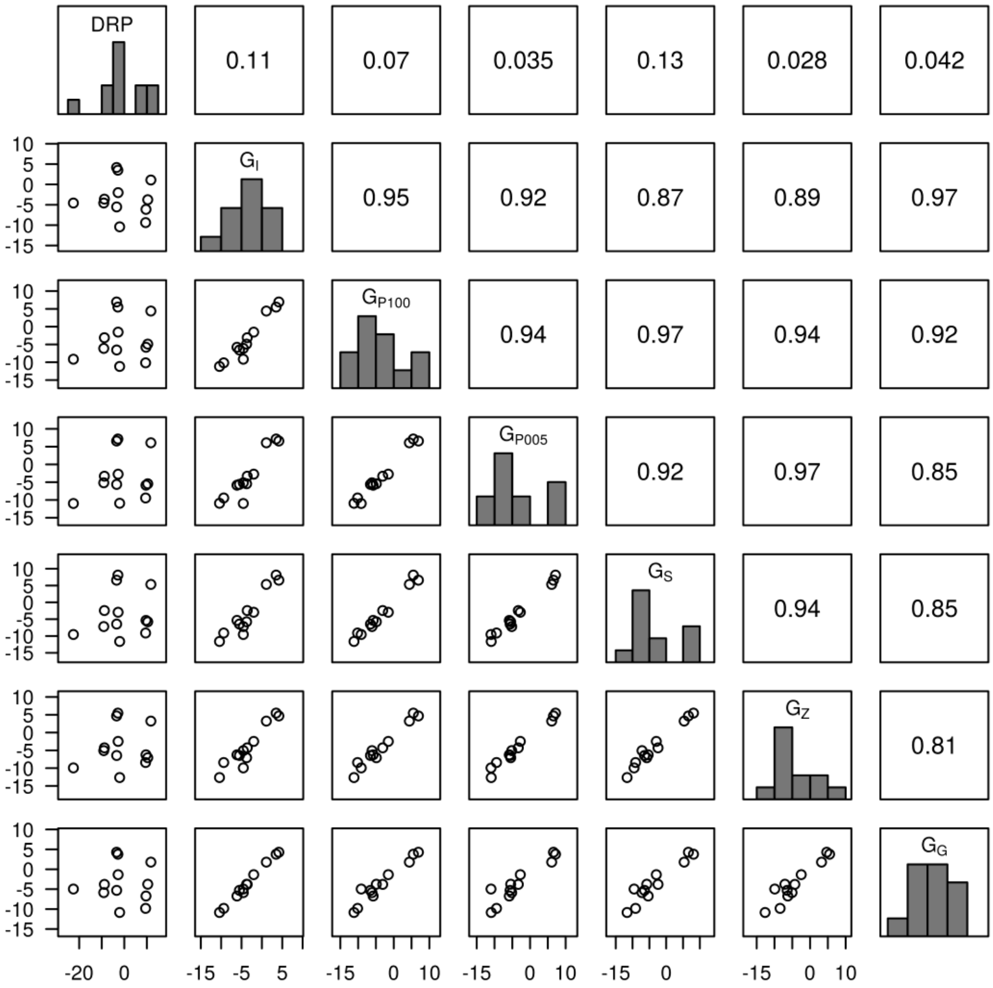
Additional file 3.4: The optimal parameter in the training stage of BLUP|GA based on HD array data for each fold 5-fold cross-validation of each replicate

		Fold1			**Fold2**			**Fold3**			**Fold4**			**Fold5**		
Trait	rep	top%	ω	acc	top%	ω	acc	top%	ω	acc	top%	ω	acc	top%	ω	acc
Eggshell strength	1	10	0.1	0.368	10	0.1	0.333	10	0.1	0.358	10	0.1	0.373	10	0.1	0.413
	2	10	0.1	0.326	10	0.1	0.284	10	0.1	0.360	2.5	0.1	0.373	0.2	0.1	0.413
	3	1	0.1	0.395	10	0.1	0.369	10	0.1	0.357	10	0.1	0.361	0.1	0.1	0.367
	4	1	0.1	0.393	10	0.1	0.305	0.3	0.1	0.342	10	0.1	0.391	10	0.1	0.378
	5	10	0.1	0.379	10	0.1	0.336	10	0.1	0.400	10	0.1	0.329	10	0.1	0.399
Feed intake	1	10	0.2	0.399	5	0.1	0.388	10	0.1	0.419	5	0.2	0.394	0.4	0.1	0.375
	2	10	0.1	0.427	5	0.2	0.362	10	0.1	0.379	10	0.7	0.413	10	0.1	0.366
	3	5	0.1	0.392	10	0.7	0.385	5	0.4	0.352	10	0.6	0.371	10	0.1	0.443
	4	10	0.2	0.439	10	0.1	0.369	10	0.4	0.406	10	0.1	0.392	5	0.2	0.450
	5	10	0.5	0.379	10	0.1	0.426	10	0.2	0.409	2.5	0.1	0.394	0.2	0.2	0.395
Laying rate	1	10	0.1	0.214	10	0.1	0.239	10	0.1	0.212	10	0.1	0.225	10	0.1	0.280
	2	10	0.1	0.244	10	0.1	0.239	10	0.1	0.238	10	0.1	0.204	10	0.1	0.196
	3	10	0.1	0.181	10	0.1	0.283	10	0.1	0.260	10	0.1	0.230	0.05	0.1	0.184
	4	10	0.1	0.183	10	0.1	0.209	1	0.7	0.255	10	0.1	0.235	10	0.1	0.212
	5	10	0.1	0.248	0.5	0.7	0.260	0.05	0.1	0.237	10	0.1	0.222	5	0.1	0.292

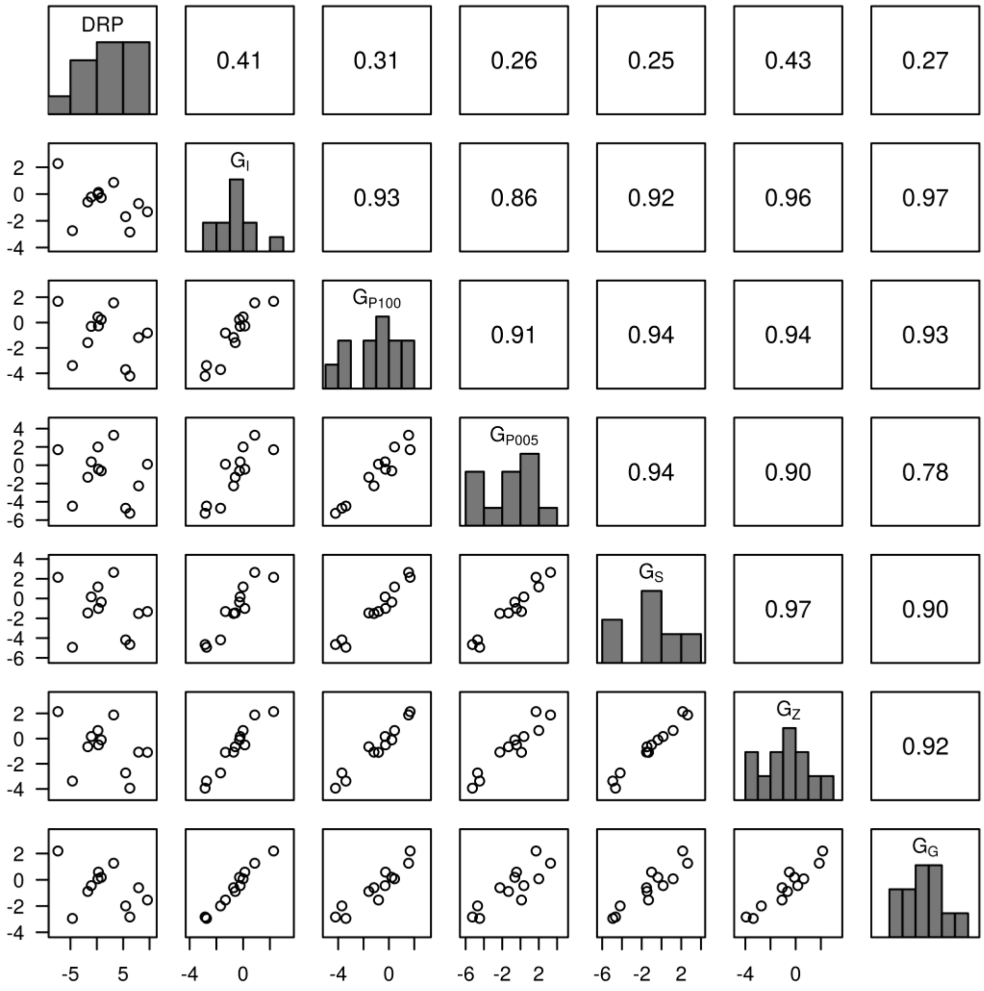
Additional file 3.5: The optimal parameter in the training stage of BLUP|GA based on WGS data for each fold 5-fold cross-validation of each replicate

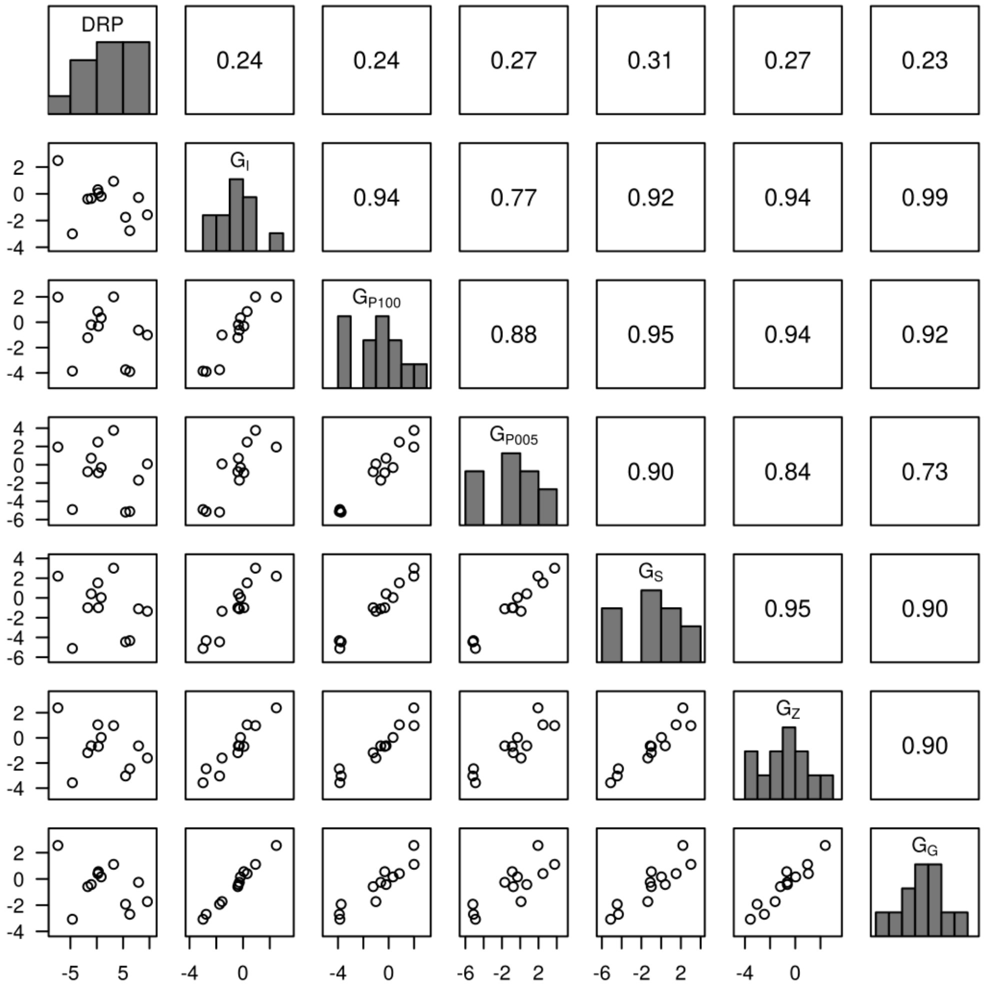
		Fold1			**Fold2**			**Fold3**			**Fold4**			**Fold5**		
Trait	rep	top%	*ω*	acc	top%	*ω*	acc	top%	*ω*	acc	top%	*ω*	acc	top%	*ω*	acc
Eggshell strength	1	10	0.1	0.378	10	0.1	0.346	0.1	0.1	0.356	10	0.1	0.382	10	0.1	0.427
	2	10	0.1	0.329	10	0.1	0.302	10	0.1	0.367	2.5	0.1	0.386	1	0.1	0.422
	3	5	0.1	0.415	10	0.1	0.373	10	0.1	0.367	10	0.1	0.364	0.1	0.1	0.375
	4	1	0.1	0.408	1	0.1	0.308	10	0.1	0.350	10	0.1	0.405	10	0.1	0.384
	5	10	0.1	0.389	10	0.1	0.340	10	0.1	0.412	10	0.1	0.336	10	0.1	0.407
Feed intake	1	10	0.2	0.394	10	0.1	0.394	10	0.1	0.418	10	0.3	0.391	0.2	0.1	0.368
	2	10	0.1	0.420	10	0.3	0.356	10	0.1	0.385	10	0.6	0.413	10	0.1	0.372
	3	10	0.1	0.395	10	0.7	0.390	10	0.2	0.353	10	0.6	0.375	10	0.1	0.447
	4	10	0.2	0.441	10	0.1	0.371	10	0.3	0.409	10	0.1	0.388	5	0.2	0.447
	5	10	0.5	0.395	10	0.1	0.434	10	0.3	0.411	5	0.1	0.400	0.4	0.1	0.378
Laying rate	1	10	0.1	0.218	0.5	0.1	0.232	0.05	0.8	0.218	10	0.1	0.227	10	0.1	0.285
	2	0.5	0.1	0.254	10	0.1	0.240	10	0.1	0.241	10	0.1	0.204	10	0.1	0.204
	3	10	0.1	0.178	10	0.1	0.285	10	0.1	0.261	10	0.1	0.229	0.1	0.99	0.189
	4	10	0.1	0.180	10	0.1	0.216	2.5	0.7	0.253	10	0.1	0.240	2.5	0.1	0.217
	5	10	0.1	0.247	0.3	0.2	0.278	10	0.1	0.236	10	0.1	0.228	2.5	0.1	0.285


DRP
0.12
0.07
0.028
0.098
0.042
0.042
G_I
0.94
0.89
0.90
0.86
0.99
G_P100
0.93
0.98
0.94
0.92
G_P005
0.90
0.97
0.86
G_S
0.94
0.89
G_Z
0.82
G_G
10
5
0
-5
-10
-15
-20
0
-15
-5
5
-15
0
10



Additional file 3.6: Predictive ability in a full-sib family with 12 individuals for feed intake based on high density (HD) array data (top) and whole-genome sequence (WGS) data (bottom) of one replicate. In each plot matrix, the diagonal shows the histograms of DRPs and DGVs obtained with various G matrices. The upper triangle shows the Spearman's rank correlation between DGVs with different G matrices and DRPs. The lower triangle shows the scatter plot of DGVs with different G matrices and DRPs.


DRP
0.41
0.31
0.26
0.25
0.43
0.27
G$_I$
0.93
0.86
0.92
0.96
0.97
G$_{P100}$
0.91
0.94
0.94
0.93
G$_{P005}$
0.94
0.90
0.78
G$_S$
0.97
0.90
G$_Z$
0.92
G$_G$
2
0
-2
-4
4
-6
-5
5
-4
0
-6
-2
2



Additional file 7: Predictive ability in a full-sib family with 12 individuals for laying rate based on high density (HD) array data (left) and whole-genome sequence (WGS) data (right) of one replicate. In each plot matrix, the diagonal shows the histograms of DRPs and DGVs obtained with various G matrices. The upper triangle shows the Spearman's rank correlation between DGVs with different G matrices and DRPs. The lower triangle shows the scatter plot of DGVs with different G matrices and DRPs.

Abbreviations

GP: Genomic prediction; HD: High density; WGS: Whole-genome sequence; GWAS: Genome-wide association studies; QTL: Quantitative trait loci; GBLUP: Genomic best linear unbiased prediction; LD: Linkage disequilibrium; BLUP|GA: Best linear unbiased prediction given genetic architecture ; MAF: Minor allele frequency; MQ: Mapping quality; DP: Depth of coverage; DRP: De-regressed proofs; ES: Eggshell strength; FI: Feed intake; LR: Arcsine transformed laying rate in the last third of the laying period ; PC: Principal components; DGV: Direct genomic breeding values; RRBLUP: Ridge regression best linear unbiased prediction; CNV: Copy number variation; INDELs: Insertion and deletions

Acknowledgements

We thank Annett and Steffen Weigend (Friedrich-Loeffler-Institute Mariensee) for the management and DNA extraction of the samples. We also thank Tim M. Strom (Helmholtz Zentrum München) for sequencing and Hubert Pausch (Technische Universität München) for genotyping. We are indebted to those researchers who programmed the software used in this study. This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr "Synbreed – Synergistic plant and animal breeding" (Grant ID 0315528C). GN personally thanks China scholarship council for the financial support.

Reference

Abdollahi-Arpanahi, R., G. Morota, B. D. Valente, a Kranis, G. J. M. Rosa, and D. Gianola. 2015. Assessment of bagging GBLUP for whole-genome prediction of broiler chicken traits. J. Anim. Breed. Genet. 132:218–28.

Van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 47:71.

Brøndum, R. F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, and M. S. Lund. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J. Dairy Sci. 98:4107–16.

Browning, S. R., and B. L. Browning. 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. 81:1084–97.

Calus, M. P. L., A. C. Bouwman, J. M. Hickey, R. F. Veerkamp, and H. A. Mulder. 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. Animal 8:1743–53.

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

Curwen, V., E. Eyras, T. D. Andrews, L. Clarke, E. Mongin, S. M. J. Searle, M. Clamp, T. Wellcome, T. Genome, and T. Broad. 2004. The Ensembl Automatic Gene Annotation System. :942–950.

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–65.

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M.-N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858–865.

Daetwyler, H., and J. Hickey. 2010. Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. Anim. Prod. Sci.:1004–1010.

Deelen, P., A. Menelaou, E. M. van Leeuwen, A. Kanterakis, F. van Dijk, C. Medina-Gomez, L. C. Francioli, J. J. Hottenga, L. C. Karssen, K. Estrada, E. Kreiner-Møller, F. Rivadeneira, J. van Setten, J. Gutierrez-Achury, H.-J. Westra, L. Franke, D. van Enckevort, M. Dijkstra, H. Byelas, C. M. van Duijn, P. I. W. de Bakker, C. Wijmenga, and M. A. Swertz. 2014. Improved imputation quality of low-frequency and rare variants in European samples using the "Genome of The Netherlands". Eur. J. Hum. Genet. 22:1321–6.

Do, D. N., L. L. G. Janss, J. Jensen, and H. N. Kadarmideen. 2015. SNP annotation-based whole genomic prediction and selection : An application to feed efficiency and its component traits in pigs. j:2056–2063.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity (Edinb). 112:39–47.

Eynard, S. E., J. J. Windig, G. Leroy, R. van Binsbergen, and M. P. Calus. 2015. The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genet. 16:24.

Fujimoto, A., H. Nakagawa, N. Hosono, K. Nakano, T. Abe, K. A. Boroevich, M. Nagasaki, R. Yamaguchi, T. Shibuya, M. Kubo, S. Miyano, Y. Nakamura, and T. Tsunoda. 2010. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nat. Genet. 42:931–6.

Garrick, D. J., J. F. Taylor, and R. L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55.

Georges, M. 2014. Towards sequence-based genomic selection of cattle. Nat. Genet. 46:807–9.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2009. ASReml User Guide 3.0. VSN International Ltd, Hemel Hempstead, UK.

Grisart, B., W. Coppieters, and F. Farnir. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res.:222–231.

Hayes, B. J., P. J. Bowman, a J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92:433–43.

Hayes, B., I. M. MacLeod, H. D. Daetwyler, P. J. Bowman, A. Chamberlian, C. Vander Jagt, A. Capitan, H. Pausch, P. Stothard, X. Liao, C. Schrooten, E. Mullaart, R. Fries, B. Guldbrandtsen, M. S. Lund, D. A. Boichard, R. F. Veerkamp, C. P. VanTassell, B. Gredler, T. Druet, A. Bagnato, J. Vilkki, D. J. DeKoning, E. Santus, and M. E. Goddard. 2014. Genomic prediction from whole genome sequence in livestock: the 1000 bull genomes project. 10th WCGALP.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Sci. 52:654.

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

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–777.

Koufariotis, L., Y.-P. P. Chen, S. Bolormaa, and B. J. Hayes. 2014. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. BMC Genomics 15:436.

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

Kutalik, Z., T. Johnson, M. Bochud, V. Mooser, P. Vollenweider, G. Waeber, D. Waterworth, J. S. Beckmann, and S. Bergmann. 2011. Methods for testing association between uncertain genotypes and quantitative traits. Biostatistics 12:1–17.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–60.

Liu, Q., E. T. Cirulli, Y. Han, S. Yao, S. Liu, and Q. Zhu. 2014. Systematic assessment of imputation performance using the 1000 Genomes reference panels. Brief. Bioinform.

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen. 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 9:e1003608.

Ma, P., R. F. Brøndum, Q. Zhang, M. S. Lund, and G. Su. 2013. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. J. Dairy Sci. 96:4666–77.

MacLeod, I. M., B. J. Hayes, and M. E. Goddard. 2014. The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. Genetics 198:1671–84.

McCarroll, S. A., T. N. Hadnott, G. H. Perry, P. C. Sabeti, M. C. Zody, J. C. Barrett, S. Dallaire, S. B. Gabriel, C. Lee, M. J. Daly, and D. M. Altshuler. 2005. Common deletion polymorphisms in the human genome. Nat. Genet. 38:86–92.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–303.

Mcqueen, H. A., G. Siriaco, and A. P. Bird. 1998. Chicken Microchromosomes Are Hyperacetylated , Early Replicating , and gene rich. :621–630.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

Morota, G., and R. Abdollahi-Arpanahi. 2014. Genome-enabled prediction of quantitative traits in chickens using genomic annotation. BMC Genomics 15.

Muir, W. M., G. K. Wong, Y. Zhang, J. Wang, M. A. M. Groenen, R. P. M. A. Crooijmans, H. Megens, H. Zhang, R. Okimoto, A. Vereijken, A. Jungerius, G. A. A. Albers, C. Taylor, M. E. Delany, S. Maceachern, and H. H. Cheng. 2008. Genome-wide assessment of worldwide chicken SNP genetic diversity indicates significant absence of rare alleles in commercial breeds. 105.

Mulder, H. A., M. P. L. Calus, T. Druet, and C. Schrooten. 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95:876–89.

Ni, G., T. M. Strom, H. Pausch, C. Reimer, R. Preisinger, H. Simianer, and M. Erbe. 2015. Comparison among three variant callers and assessment of the accuracy of imputation from SNP array data to whole-genome sequence level in chicken. BMC Genomics:1–12.

Ober, U., J. F. Ayroles, E. A. Stone, S. Richards, D. Zhu, R. a Gibbs, C. Stricker, D. Gianola, M. Schlather, T. F. C. Mackay, and H. Simianer. 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet. 8:e1002685.

Pérez-Enciso, M., J. C. Rincón, and A. Legarra. 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet. Sel. Evol. 47:43.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–9.

Redon, R., S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. 2006. Global variation in copy number in the human genome. Nature 444:444–454.

Segelke, D., J. Chen, Z. Liu, F. Reinhardt, G. Thaller, and R. Reents. 2012. Reliability of genomic prediction for German Holsteins using imputed genotypes from low-density chips. J. Dairy Sci. 95:5403–11.

Snelling, W., and R. Cushman. 2013. Breeding and Genetics Symposium: networks and pathways to guide genomic selection. J. Anim. Sci.91:537-52.

Su, G., O. F. Christensen, L. Janss, and M. S. Lund. 2014. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. J. Dairy Sci. 97:6547–59.

Su, G., B. Guldbrandtsen, G. P. Aamand, I. Strandén, and M. S. Lund. 2014. Genomic relationships based on X chromosome markers and accuracy of genomic predictions with and without X chromosome markers. Genet. Sel. Evol. 46:47.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–23.

Wang, K., M. Li, and H. Hakonarson. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38:e164.

Winter, A., G. Ewald, O. Bellmann, J. Wegner, and H. Zu. 2003. DGAT1, a new positional and functional candidate gene for intramuscular fat deposition in cattle. Anim. Genet. 34:354–357.

Zhang, Z., M. Erbe, J. He, U. Ober, N. Gao, H. Zhang, H. Simianer, and J. Li. 2015. Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. G3 (Bethesda). 5:615–27.

Zheng, H.-F., J.-J. Rong, M. Liu, F. Han, X.-W. Zhang, J. B. Richards, and L. Wang. 2015. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. PLoS One 10:e0116487.

Zhou, L., M. S. Lund, Y. Wang, and G. Su. 2014. Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. J. Anim. Breed. Genet. 131:249–57.

Chapter 4 Comparison between approaches to estimate the accuracy of genomic breeding values

Guiyan Ni[1], Sandra Kipp[2], Henner Simianer[1], Malena Erbe[1,3]

[1]Animal Breeding and Genetics Group, Georg-August-Universität, Göttingen, Germany

[2]Vereinigte Informationssysteme Tierhaltung w.V. (vit), Verden, Germany

[3]Institute for Animal Breeding, Bavarian State Research Centre for Agriculture, Grub, Germany

Publication in preparation

Abstract

Decisions of genomic selection schemes are made based on the genomic breeding values (GBV) of selection candidates. Thus, the accuracy of GBV is a relevant parameter, as it reflects the stability of the prediction and the possibility that the GBV might change when more information becomes available. Accuracy of genomic prediction, however, is difficult to assess, considering true breeding values of the candidates are not available in reality. In previous studies, several methods were proposed to assess the accuracy of GBV by using population and trait parameters or parameters inferred from the mixed model equations. In practice, most approaches were found to overestimate the accuracy of genomic prediction. Thus, we tested several approaches used in previous studies based on simulated data under a variety of parameters mimicking different livestock breeding programs in order to measure the magnitude of overestimation. Further we proposed a novel and computationally feasible method and tested in a real Holstein data set. Based on the comparisons with simulated data, the new method provided a better prediction for the accuracy of GBV. The new method still has one unknown parameter, for which we suggest an approach to approximate its value from a suitable data set reflecting two separate time points. In conclusion, the new approach has the potential to provide a better assessment of the accuracy of GBVs in many cases.

Introduction

With the widespread availability of high throughput single-nucleotide polymorphism (SNP) genotyping, genomic selection (GS) has been widely used in livestock (Hayes et al., 2009a; Meuwissen et al., 2013) and plant (Jannink et al., 2010; Rincent et al., 2012) breeding, and displays dramatic advantages in genetic progress in both simulated (Habier et al., 2013) and real (Hayes et al., 2009a) selection scenarios, especially for sex limited traits or traits that can only be measured late in life (Meuwissen et al., 2013). Decisions of genomic selection schemes are made based on the genomic breeding values (GBV) of selection candidates. A GBV as used in this study is the prediction of an individual’s true breeding value (TBV) derived from its SNP genotype and marker effects estimated based on a set of genotyped and phenotyped animals of the same population. Thus, the accuracy of GBV, defined as the correlation between TBV and GBV, is a relevant parameter, since it reflects the stability of the prediction and the possibility that the GBV might change when more information becomes available (Bijma, 2012). Furthermore, it is also one of the key factors in expected response to selection which is also known as breeders' equation (Falconer and Mackay, 1996; Nirea et al., 2012).

In two-step genomic prediction, conventional estimated breeding values (EBV) are first estimated with pedigree based best linear unbiased prediction (BLUP), and then GBV are estimated in genomic BLUP with EBV or their derivates, e.g. de-regressed proofs (DRP), daughter yield deviations (DYD), used as quasi-phenotypes. To assess the predictive ability of GBV, it is possible to easily calculate the correlation between GBV and EBV in genomic prediction schemes. The actually interesting correlation between GBV and TBV, namely the accuracy of genomic prediction, however, is difficult to assess, considering TBV of the candidates are not available in reality.

In previous studies, several branches have been suggested to assess the accuracy of GBV: one of the branches is using population and trait parameters, such as the effective population size, the size of a chromosome, the number of independent chromosome segments, and the heritability of the considered trait to approximate the accuracy. The advantage of this branch is that these approaches can be used before data of selection candidates are collected (Wientjes et al., 2013). These approaches give an overall assessment of the expected accuracy, which treat the studied samples as a whole but are independent on the set of information available for a specific animal in question. Different suggested equations predicting accuracy from known population parameters, however, were shown to provide different results not always matching with the real values, especially when extrapolating parameters beyond the actually observed space (Erbe et al., 2013).

The second branch of estimating the accuracy of GBV is using parameters inferred from the mixed model equations (MMEs). Following Henderson (1975), accuracy of estimated breeding values for a given animal i ($r_{\boldsymbol{BV}_i}$) can be calculated from the prediction error variance for individual i, $\boldsymbol{PEV}_i$, which can be obtained from the inverse of the coefficient matrix of the MMEs, and the genetic variance $var(\boldsymbol{A_i})$, as

$$r_{\boldsymbol{BV}_i} = \sqrt{1 - \frac{\boldsymbol{PEV}_i}{var(\boldsymbol{A_i})}}.$$

In principle, this type of assessment can also be used in genomic breeding value estimation implemented with GBLUP (VanRaden, 2008), where the pedigree-based numerator relationship matrix in BLUP is replaced by the genomic relationship matrix. However, this accuracy only holds under the absence of selection (Dekkers, 1992) and is biased when ignoring the changes in the (co)variance structure in selected populations (Henderson, 1975; Dekkers, 1992). A basic advantage of this approach is that a specific accuracy can be obtained for each individual for which a breeding value is estimated.

A third branch to assess the average accuracy of GBV is based on observed correlations between different quantities obtained or derived in the course of breeding value estimation. Assuming no covariance between TBV and errors of GBV ($\boldsymbol{\varepsilon_G}$) and errors of EBV ($\boldsymbol{\varepsilon_E}$), Amer and Banos (2010) suggested the accuracy of GBV to be

$$r_{GT} = \frac{r_{EG}}{r_{ET}\left(1 + \frac{cov(\boldsymbol{\varepsilon_G},\boldsymbol{\varepsilon_E})}{var(\boldsymbol{T})}\right)} \tag{1}$$

where r_{EG} is the empirical correlation between EBV and GBV, r_{ET} is the accuracy of EBV obtained from conventional BLUP MMEs, $\varepsilon_G(\varepsilon_E)$ is the error of GBV (EBV) which is defined as the deviation of predicted values divided by its own reliability from TBV:

$$\boldsymbol{\varepsilon_G} = \boldsymbol{T} - \frac{\boldsymbol{G}}{r_G^2}$$

However, ε_G and ε_E usually are not available in real data so that their covariance is not known. Hence, ignoring the covariance between ε_G and ε_E (or more precisely, assuming this covariance to be zero), equation (1) simplifies to

$$r_{GT} = \frac{r_{EG}}{r_{ET}}.$$

This last formula or formulas with derivatives of EBV (e.g. DYD, or DRP) have been widely used in real data analysis (Hayes et al., 2009a; Luan et al., 2009; Saatchi et al., 2011) and are easy to implement, since r_{ET} exactly or approximately is available from MMEs of conventional BLUP and r_{EG} can be empirically calculated as the correlation between EBV and GBV.

In practical applications, most approaches were found to overestimate the accuracy of GBV (Goddard, 2009; Hayes et al., 2009b; Goddard et al., 2011). In addition, the magnitude of overestimation is unknown in real data sets, and little attention has been given to the quantification of how much these approaches overestimate the accuracy of GBV.

The first objective of this study was thus to test several approaches mentioned above with simulated data under a variety of parameters mimicking different livestock breeding programs (i.e. a cattle-like and a pig-like as well as a basic scenario) and to measure the magnitude of overestimation. The second objective of this study was to suggest a novel and computationally feasible method that can provide a better prediction for the accuracy of GBV in real data sets and to assess the quality of the new approximation with both simulated and real data.

Material and Methods

Different approaches of estimating accuracy of genomic breeding values

Previous approaches of estimating accuracy of GBV

According to the definition of a correlation, the correlation between GBV and EBV is defined as:

$$r_{GE} = \frac{cov(\boldsymbol{G}, \boldsymbol{E})}{\sqrt{var(\boldsymbol{G})var(\boldsymbol{E})}} \tag{2}$$

GBV or EBV of individuals can be written as $\boldsymbol{G} = r_{GT}^2(\boldsymbol{T} + \boldsymbol{\varepsilon_G})$ or $\boldsymbol{E} = r_{ET}^2(\boldsymbol{T} + \boldsymbol{\varepsilon_E})$, and $var(\boldsymbol{G}) = cov(\boldsymbol{G}, \boldsymbol{T})$ and $var(\boldsymbol{E}) = cov(\boldsymbol{E}, \boldsymbol{T})$ based on the assumption of BLUP. Thus equation (2) can be expressed as follows:

$$r_{GE} = \frac{cov(r_{GT}^2(\boldsymbol{T} + \boldsymbol{\varepsilon_G}), r_{ET}^2(\boldsymbol{T} + \boldsymbol{\varepsilon_E}))}{\sqrt{cov(r_{GT}^2(\boldsymbol{T} + \boldsymbol{\varepsilon_G}), \boldsymbol{T})cov(r_{ET}^2(\boldsymbol{T} + \boldsymbol{\varepsilon_E}), \boldsymbol{T})}}$$

$$= \frac{r_{GT}^2 r_{ET}^2(var(\boldsymbol{T}) + cov(\boldsymbol{\varepsilon_G}, \boldsymbol{\varepsilon_E}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_G}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_E}))}{\sqrt{r_{GT}^2(var(\boldsymbol{T}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_G}))r_{ET}^2(var(\boldsymbol{T}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_E}))}}$$

$$= \frac{r_{GT}^2 r_{ET}^2(var(\boldsymbol{T}) + cov(\boldsymbol{\varepsilon_G}, \boldsymbol{\varepsilon_E}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_G}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_E}))}{r_{GT} r_{ET}\sqrt{(var(\boldsymbol{T}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_G}))(var(\boldsymbol{T}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_E}))}} \tag{3}$$

By rearranging formula (3), we get

$$r_{GT} = \frac{r_{GE}}{r_{ET}\frac{(var(\boldsymbol{T}) + cov(\boldsymbol{\varepsilon_G}, \boldsymbol{\varepsilon_E}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_G}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_E}))}{\sqrt{(var(\boldsymbol{T}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_G}))(var(\boldsymbol{T}) + cov(\boldsymbol{T}, \boldsymbol{\varepsilon_E}))}}} \tag{4}$$

Assuming the covariance between TBV and ε_E, and the covariance between TBV and ε_G are both equal to 0 (i.e. $cov(T, \varepsilon_G) = cov(T, \varepsilon_E) = 0$), formula (4) can be simplified as:

$$r_{GT} = \frac{r_{GE}}{r_{ET}\frac{(var(\boldsymbol{T})+cov(\boldsymbol{\varepsilon_G},\boldsymbol{\varepsilon_E}))}{\sqrt{var(\boldsymbol{T})var(\boldsymbol{T})}}}$$

$$= \frac{r_{GE}}{r_{ET}(1 + \frac{cov(\boldsymbol{\varepsilon_G},\boldsymbol{\varepsilon_E})}{var(\boldsymbol{T})})} \qquad (5)$$

which is the approximation for predicting the accuracy of GBV suggested by Amer and Banos (2010) and will be denoted as ‘Acc_AB’ in the following. By further assuming that $\boldsymbol{\varepsilon_G}$ and $\boldsymbol{\varepsilon_E}$ are independent ($cov(\boldsymbol{\varepsilon_G}, \boldsymbol{\varepsilon_E}) = 0$), we get

$$r_{GT} = \frac{r_{GE}}{r_{ET}} \qquad (6)$$

which is the formula used in Hayes et al. (2009b) and will be referred as ‘Acc_H’ in the following.

Since the covariance between two errors is not available in reality, Acc_AB is not immediately applicable in reality. Acc_H was found to overestimate the accuracy of GBV (Goddard, 2009; Hayes et al., 2009b; Goddard et al., 2011). Thus it is necessary to suggest a novel and computationally feasible method that provides a better prediction for the accuracy of GBV in real data sets.

A novel approach for approximating the accuracy of GBV

Let $\boldsymbol{t}$, $\boldsymbol{e}$ and $\boldsymbol{g}$ denote the standardized transformed vectors TBV, EBV and GBV for individuals in a population, i.e.

$$\boldsymbol{t} = \frac{\boldsymbol{T}-1_n\overline{T}}{\sqrt{var(\boldsymbol{T})}},\ \boldsymbol{e} = \frac{\boldsymbol{E}-1_n\bar{E}}{\sqrt{var(\boldsymbol{E})}},\ \text{and}\ \boldsymbol{g} = \frac{\boldsymbol{G}-1_n\bar{G}}{\sqrt{var(\boldsymbol{G})}};$$

where $\overline{T}, \overline{E}, \overline{G}$ is the mean of TBV ($\boldsymbol{T}$), EBV ($\boldsymbol{E}$) and GBV ($\boldsymbol{G}$), respectively. n is the length of vector TBV, 1_n is a column vector of 1s of length n.

Then, the sums of the elements of $\boldsymbol{t}$, $\boldsymbol{e}$ and $\boldsymbol{g}$ are equal to 0, the variances of $\boldsymbol{t}$, $\boldsymbol{e}$ and $\boldsymbol{g}$ are equal to 1, and the correlation coefficients equal the regression coefficients. Since such a scaling does not affect correlations,

$$r_{GT} \equiv r_{gt},\, r_{ET} \equiv r_{et},\text{ and } r_{EG} \equiv r_{eg}.$$

Now**,** $\boldsymbol{t}$ can be rewritten as

$$\boldsymbol{t} = r_{et}\ \boldsymbol{e} + \boldsymbol{m}_{et}$$

with r_{et} being the regression coefficient of $\boldsymbol{t}$ on $\boldsymbol{e}$ (as all variances are 1). Thus, the vector of true breeding values $\boldsymbol{t}$ is expressed as its expectation given the estimated breeding values $\boldsymbol{e}$ and an error term, denoted as $\boldsymbol{m}_{et}$. Note that $\boldsymbol{e}$ and $\boldsymbol{m}_{et}$ are uncorrelated, i.e. $cor(\boldsymbol{e}, \boldsymbol{m_{et}}) =\ 0$. In the same way, we can write $\boldsymbol{g} = r_{eg}\ \boldsymbol{e} + \boldsymbol{m}_{eg}$ with $cor(\boldsymbol{e}, \boldsymbol{m_{eg}}) = 0$.

Then, the accuracy of GBV can be written as

$$r_{GT} \equiv r_{gt} \equiv \frac{cov(r_{et}\ \boldsymbol{e} + \boldsymbol{m_{et}}, r_{eg}\ \boldsymbol{e} + \boldsymbol{m_{eg}})}{\sqrt{var(\boldsymbol{g})var(\boldsymbol{t})}}$$

Since $var(\boldsymbol{g}) = var(\boldsymbol{t}) = 1$, it follows

$$r_{GT} \equiv cov(r_{et}\ \boldsymbol{e} + \boldsymbol{m_{et}}, r_{eg}\ \boldsymbol{e} + \boldsymbol{m_{eg}})$$

$$\equiv r_{eg}r_{et}cov(\boldsymbol{e}, \boldsymbol{e}) + r_{et}cov(\boldsymbol{e}, \boldsymbol{m_{eg}}) + r_{eg}cov(\boldsymbol{e}, \boldsymbol{m_{et}}) + cov(\boldsymbol{m_{et}}, \boldsymbol{m_{eg}})$$

Because $\boldsymbol{e}$ and $\boldsymbol{m}_{et}$ are uncorrelated, therefore, $cov(\boldsymbol{e}, \boldsymbol{m_{eg}}) = cov(\boldsymbol{e}, \boldsymbol{m_{et}}) = 0$, thus,

$$r_{GT} \equiv r_{eg}r_{et} + cov(\boldsymbol{m_{et}}, \boldsymbol{m_{eg}})$$

$$\equiv r_{eg}r_{et} + r_{mm}\sqrt{var(\boldsymbol{m_{et}})var(\boldsymbol{m_{eg}})}$$

with r_{mm} being the correlation between $\boldsymbol{m_{et}}$ and $\boldsymbol{m_{eg}}$.

$var(\boldsymbol{m_{et}})$ is

$$var(\boldsymbol{m_{et}}) = var(\boldsymbol{t} - r_{et}\ \boldsymbol{e})$$

$$= var(\boldsymbol{t}) + r_{et}^2 var(\boldsymbol{e}) - 2r_{et}\, cov(\boldsymbol{t}, \boldsymbol{e}),$$

As $var(\boldsymbol{t}) = var(\boldsymbol{e}) = 1$,

$$var(\boldsymbol{t}) + r_{et}^2 var(\boldsymbol{e}) - 2r_{et}\, cov(\boldsymbol{t}, \boldsymbol{e}) = 1 + r_{et}^2 - 2r_{et}\, r_{et}$$

$$= 1 - r_{et}^2$$

and analogously

$$var(\boldsymbol{m_{eg}}) = 1 - r_{eg}^2.$$

This results in

$$r_{GT} \equiv r_{eg} r_{et} + r_{mm} \sqrt{(1 - r_{et}^2)\left(1 - r_{eg}^2\right)}$$

$$\equiv r_{EG} r_{ET} + r_{mm} \sqrt{(1 - r_{ET}^2)\,(1 - r_{EG}^2)} \qquad (7)$$

which we will hereinafter denote as Acc_N. Since r_{EG} (as the empirical correlation of EBV and GBV) and r_{ET} (as the average theoretical value based on the prediction error variance) are available from breeding value estimation runs in real data, r_{mm}, which we will call a weighting factor, is the only unknown parameter which needs to be determined.

We will study the range of the optimal weighting factor r_{mm} by minimizing the squared difference between the true accuracy of GBV, which is available in simulated data, and the approximation with the new formula over replicates of the simulation data described below. We will further assess the usefulness of the new approximation in real data by using highly accurate progeny-based breeding values as a proxy of TBV, thus demonstrating that at least a good approximation of r_{mm} can be obtained with real data. In addition, we will discuss the influence of using the average theoretical values based on the prediction error variance instead of empirical correlation between EBV and TBV, $cor(E, T)$ which is not available in real data, on all the approaches listed above.

Simulation of data

Accuracy of genomic prediction was estimated for three alternative main simulated data sets with different population parameters, called cattle-like, pig-like, and basic scenario, for which the details are presented in the following. The simulation was performed by using the software QMSim (Sargolzaei and Schenkel, 2009). The whole simulation process was repeated 20 times.

Genome

The simulated genome for all scenarios was the same: The genome consisted of 10 chromosomes with 100 centiMorgan each. Initially, there were 3,000 polymorphic markers and 50 quantitative trait loci (QTLs) randomly distributed on each chromosome. Markers and QTLs with a minor allele frequency (MAF) ≥ 0.01 in the last historical population were selected and used in the simulation of the recent population. The additive allelic effects of QTLs were drawn from a gamma distribution with shape parameter 0.2. The positions of markers and QTLs across the genome were randomized in each of the 20 replications.

A quantitative trait with heritability of 0.2 or 0.5 was simulated. TBV were simulated by summing up all true additive QTL allelic effects. The phenotypes were obtained by adding random residual effects to TBV. The simulation included random selection of parents (abbreviated as ‘noSel’) or selection of parents based on EBV with predefined accuracy (abbreviated as ‘Sel’) in each sex in each scenario, in which the predefined accuracy were calculated based on the available information in each scenario e.g. the heritability of the trait, and the number of progeny, as suggested in Falconer and Mackay (1996).

Cattle-like scenario (Additional file 4.1a):

A historical random mating population with a constant size of 1,000 in the first 900 generations and with a continuous increase in size to 100,000 for the last 100 generations was simulated. 500 founder males and 10,000 founder females were randomly chosen from the last generation of the historical population.

To mimic a real cattle breeding scheme, a recent population with 12 generations was simulated. 500 sires were mated to 10,000 dams per generation. Each dam produced 2 progenies with a probability of 50% for male progenies. Therefore, the number of simulated individuals in each generation was 20,000. Each sire had 20 female offspring and 20 male offspring. In the selection scenario, the male parents were selected based on an EBV with an accuracy ≥0.7 (0.85) for heritability 0.2 (0.5); the female parents were selected based on EBV with an accuracy ≥ 0.45 for both heritabilities. A sex limited trait was simulated; consequently, phenotypes were assigned only to females in generation 6 to 12.

Pig-like scenario (Additional file 4.1b):

The simulation of the first 900 generations of the historical random mating population was performed with a fixed population size of 1,000 followed by 100 generations with a gradual increase in size to 10,000. The founder population was built up by 500 sires and 500 dams randomly sampled form generation 1,000 of the historical population.

The parameters used in the recent population mimicked a 12-generation pig breeding scheme. In each generation, each litter consisted of 4 pigs (2 males, 2 females). It needs to be mentioned that we only simulated the individuals used in the breeding scheme. In the selection scenario, both parents were selected based on EBV with an accuracy ≥0.6 (0.7) for heritability 0.2 (0.5). Phenotypes were assigned only to females in generation 6 to 12.

Basic scenario (Additional file 4.1c):

A historical random mating population was simulated over 1,000 generations with a constant population size of 10,000. From the last generation of the historical population, 500 males and 1,000 females were randomly chosen to act as founders of the recent population.

The recent population consisted of 12 generations in which each of the 500 sires mated with 2 out of 1000 dams randomly per generation. Each dam produced 2 offspring. The proportion of male offspring was 0.5. For heritability 0.2 (0.5), the selection of female parents was based on EBV with accuracy ≥0.65 (0.75); the selection of female parents was based on EBV with accuracy 0.45 for both heritabilities. Starting from generation 6, both males and females got phenotypic records.

Training and validation sets for the genomic breeding value prediction

In reality, there are two situations in genomic selection schemes in which a measure of accuracy is considered. The first one is cross-validation (abbreviated as ‘CV’). In this case, the phenotypes of individuals in the validation set have already been collected and the accuracy is validated based on the correlation between predicted and observed phenotypes in the validation set. The other situation, called ‘forward prediction’ (abbreviated as ‘FP’), is similar the real challenge in genomic selection, since the phenotypes of candidates have not been collected at the time point the genomic breeding values are predicted. Both situations were investigated for each scenario and each replication (Additional file 4.1), and the accuracy of GBV of sires in the validation set obtained with different formulas was assessed for each scenario in each replicate separately. The validation set were sires in generation 11 for both CV and FP scenarios. Progeny information from generation 12 was available for the CV scenario while it was not available for the FP scenario, when estimating EBV and GBV with the models described in the following.

Estimation of conventional and genomic breeding values in a two-step model

In the first step, the conventional EBV of individuals were estimated based on the following animal model:

$$\boldsymbol{y} = \mathbf{1}\mu_1 + \boldsymbol{Za} + \boldsymbol{e_1}$$

where $\boldsymbol{y}$ is a vector of phenotypic records, μ_1 is the overall mean, $\boldsymbol{Z}$ is the design matrix of breeding values, $\boldsymbol{a}{\sim}N\left(0,\ \boldsymbol{A}\sigma_a^2\right)$ is the vector of breeding values and $\boldsymbol{e_1}$ is a vector of random errors following a normal distribution $\boldsymbol{e_1}{\sim}N(0, \boldsymbol{I}\sigma_{e_1}^2)$. $\boldsymbol{A}$ is the pedigree-based numerator relationship matrix. Based on this model, the vector of conventional breeding values (EBV) is obtained via a BLUP estimation.

In the second step, the GBV of sires were estimated based on the following model:

$$\boldsymbol{y_2} = \mathbf{1}\mu_2 + \boldsymbol{Wg} + \boldsymbol{e_2}$$

where $\boldsymbol{y_2}$ is a vector of quasi phenotypes (EBVs in this case) of sires in the training population, μ_2 is the overall mean, $\boldsymbol{W}$ is the design matrix corresponding to $\boldsymbol{g}$, the vector of the animals’ GBV which was assumed to be distributed $\boldsymbol{g}{\sim}N(0,\ \boldsymbol{G}\sigma_g^2)$, and $\boldsymbol{e_2}$ is a vector of random errors following a normal distribution $\boldsymbol{e_2}{\sim}N(0, \boldsymbol{I}\sigma_{e_2}^2)$. $\mathbf{G}$ is the genomic relationship matrix according to VanRaden (2007).

EBV, GBV and corresponding variance components were estimated using ASReml 3.0 for each scenario and each replicate (Gilmour et al., 2009).

Real cattle data

The real Holstein cattle data used for testing the validity of the proposed approach are from two routine breeding value runs 2010 and 2014 conducted by vit Verden (http://www.vit.de/). For each individual, EBV, reliabilities of EBV, and GBV were provided for the traits milk yield (MY), fat yield (FY), protein yield (PY), and somatic cell score (SCS). Individuals were selected based on two criteria: First, individuals had to be in the candidate set in 2010 and in the training set in 2014 in the genomic prediction. Second, individuals had EBV with reliability ≤0.85 in 2010 and EBV with reliability ≥0.95 in 2014. For studying on how to estimate r_{mm} best in real data, the EBV estimated in 2014 with $r^2 \geq 0.95$ (i.e. an accuracy ≥ 0.974) was used as a proxy for the TBV of that individual for the respective trait, denoted as $\mathrm{TBV_E}$ hereinafter. Besides the empirical correlation between EBV and GBV (r_{EG}), the empirical correlation between $\mathrm{TBV_E}$ and EBV, and GBV (r_{ET_E} , r_{GT_E}) can be approximated as well. Consequently, the optimal weighting factor can be approximated as $\frac{r_{GT\_E}-r_{EG}r_{ET\_E}}{\sqrt{(1-r_{ET\_E}^2)\,(1-r_{EG}^2)}}$. Based on this, the robustness of the weighting factor in different traits can be investigated.

Results

Simulation

On average, there were 27,585 SNPs and 459 QTLs randomly distributed on the simulated chromosomes. The minor allele frequency (MAF) based on the simulated genotypes of sires from generation 6 to generation 11 in the first replicate of cattle_5 noSel scenario is shown in Additional file 4.2. The average linkage disequilibrium (r^2) between SNPs whose distance was smaller than 2 cM is shown in Additional file 4.3 for different scenarios. There was no significant difference in the level of LD between pig-like and cattle-like scenarios, which both had a higher LD level than the basic scenario.

MSE of Acc_N with different weighting factors and the optimal weighting factor

Since r_{mm} in Acc_N is defined as a correlation, r_{mm} can only take values between -1 and +1. In other words, all possible MSE can be inspected when r_{mm} is moved from -1 to +1, as shown in Figure 4.1, in which MSE is the average of the squares of the difference between the empirical accuracy and theoretical accuracy. Compared to the Sel scenarios, the curves of the noSel scenarios were more flat. The curves of FP scenarios were more convex than the curves of CV scenarios (e.g. Sel_FP vs Sel_CV), which means that FP scenarios were more sensitive to the choice of the weighting factor. In FP scenarios, the optimal weighting factor of cattle-like scenarios were larger than pig-like and basic scenarios, however, there were no differences in CV scenarios. Furthermore, the optimal weighting factor which is defined as r_{mm} giving the minimum MSE in each scenario can be located and is shown in Table 4.1. Across all the scenarios, the average (± standard deviation) of the optimal weighting factor was 0.25 (± 0.15) for Sel_FP, 0.25 (±0.15) for Sel_CV, 0.40 (±0.14) for noSel_FP, and 0.33 (±0.1) for noSel_CV. The optimal weighting factors of noSel_FP were larger than Sel_FP in each scenario, while there was no systematic pattern in the CV scenarios.

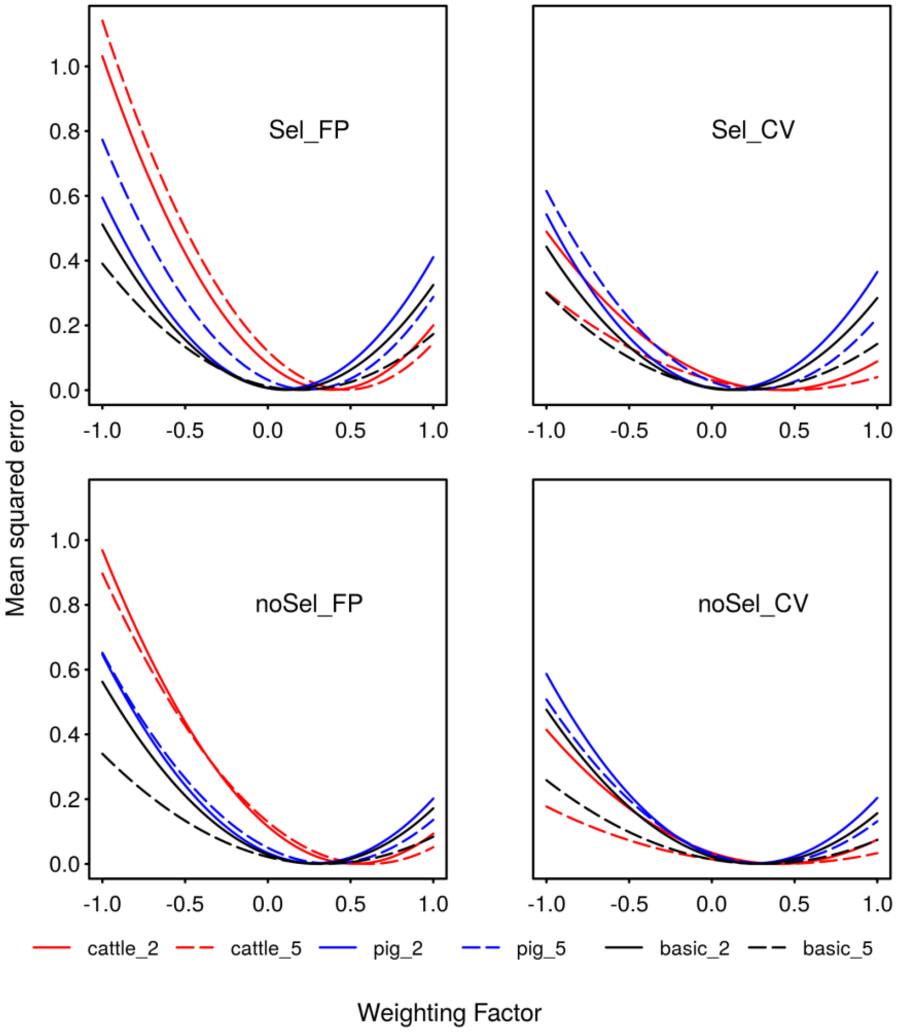


Figure 4.1: Mean squared errors (MSE) of the predicted accuracy of GBV calculated by Acc_N in different simulation scenarios plotted against all possible weighting factors r_{mm}. Red stands for cattle-like scenarios, blue represents pig-like scenarios and black denotes intermediate scenarios. Solid lines (—) stand for scenarios with heritability equal to 0.2 and dashed lines (- -) denote scenarios with heritability equal to 0.5.

Table 4.1: The optimal weighting factors in different scenarios

	cattle_2	cattle_5	pig_2	pig_5	basic_2	basic_5
Sel_FP	0.38	0.48	0.10	0.24	0.12	0.20
Sel_CV	0.40	0.46	0.10	0.24	0.12	0.18
noSel_FP	0.52	0.62	0.28	0.38	0.28	0.34
noSel_CV	0.40	0.40	0.26	0.32	0.28	0.30

Predicted accuracy with different approaches in different simulated scenarios

Predicted accuracies of GBV calculated from different approaches in validation data sets are shown in Figure 4.2 where they were plotted against the empirical accuracies of GBV defined as the correlation between GBV and TBV. It should be mentioned that the empirical (true) accuracy is not available in real applications. The grey dashed lines in the plots stand for equality of predicted accuracies of GBV and empirical accuracy, which means that the closer the values are to the grey line, the more accurate the predication is. In general, Acc_N with the optimal weighting factor showed considerably better performance than the other three approaches in most of the simulated cases. Acc_H systematically overestimated the empirical accuracy in this simulated dataset, with a substantial proportion of accuracies larger than 1 and thus outside the parameter space, except for Sel_CV in the cattle scenario. The overestimation of Acc_H was stronger in the noSel scenarios than that in Sel scenarios. Acc_AB had a relative good performance in Sel scenarios, especially in the case when more information is available (i.e. Sel_CV). However, in some cases of noSel scenarios, predicted accuracies with Acc_AB were larger than 1 as well. Compared to Acc_AB and Acc_H, r_{EG} had equal or better performance. It should be noticed that when the empirical accuracy was lower, Acc_H, Acc_AB, and r_{EG} had a stronger tendency to overestimate the empirical accuracy. Focusing on prediction with Acc_N, the results were close to even along with the grey dash lines, especially in the Sel scenarios. The prediction with Acc_N, however, was always in the same range regardless of the level of empirical accuracies of GBV in noSel scenarios. All in all, Acc_H and r_{EG} had relative good prediction in breeding schemes with more information and higher accurate EBVs available (e.g. cattle). However, overestimation still needs to be paid attention. The agreement between Acc_N and empirical accuracy was higher than that with Acc_H, Acc_AB, and r_{EG} in most simulated scenarios, especially in the breeding schemes under selection.

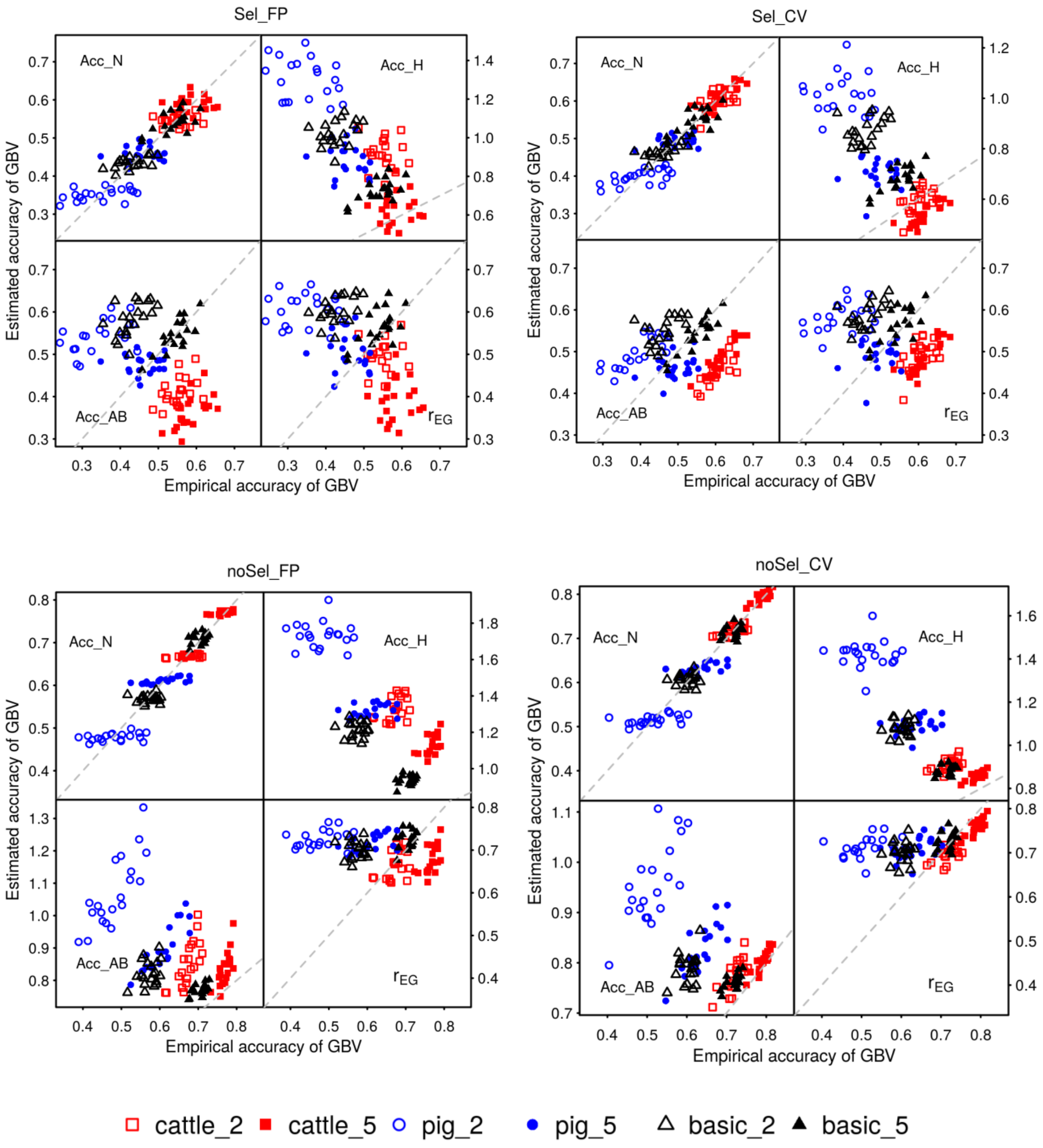


Figure 4.2: Accuracy of GBV estimated by different approaches in validation sets plotted against the empirical accuracy of GBVs calculated as the correlation of GBVs and TBVs. Values on the X axis are the empirical accuracies of GBV, $cor(G, T)$, among different simulated scenarios. Values on the Y axis are the accuracies of GBV estimated by Acc_N, Acc_H, Acc_AB and r_{EG}. Grey lines in the plots stand for where estimated accuracies equaling to empirical accuracies. Values above the grey line denote overestimation of empirical accuracies while values below the grey line mean underestimation of empirical accuracies. Number two (five) in the legend stands for the respective scenario with simulated heritability equaling to 0.2 (0.5).

Results from the real cattle data

In order to check the different approaches with real data, there were 1,271 individuals selected based on the two criteria mentioned above. Distributions of reliabilities of EBVs for MY, FY, PY, and SCS are shown in Additional file 4.4. Since TBV$_E$, EBV, and GBV are available, the correlations among them are available. Thus, the optimal r_{mm} [$=\frac{r_{GT\_E}-r_{EG}r_{ET\_E}}{\sqrt{(1-r_{ET\_E}^2)\,(1-r_{EG}^2)}}$], Acc_H and r_{EG} can be assessed in this real data set as shown in Table 4.2 for several traits of interest. Acc_AB is not available based on this dataset, because in Acc_AB, errors of GBV is defined as the deviation of predicted values divided by its own reliability from TBV, i.e.

$$\boldsymbol{\varepsilon_G} = \boldsymbol{T} - \frac{\boldsymbol{G}}{r_{GT}^2}$$

in which $r_{GT}^2$is the values that we want to estimate. The optimal weighting factors were relatively stable in all traits and ranged between 0.608 and 0.676. Empirical correlations between EBV and GBV were smaller than the accuracy of GBV. Acc_H overestimated the accuracy of GBV.

Table 4.2: The optimal weighting factor and correlation between EBV, GBV and TBV$_E$ for the real cattle data

Traits	$\boldsymbol{r_{mm}}$	$\boldsymbol{r_{ET\_E}}$	$\boldsymbol{r_{GT\_E}}$	$\boldsymbol{r_{EG}}$	Acc_H
Milk yield	0.670	0.525	0.768	0.571	1.088
Fat yield	0.656	0.520	0.757	0.573	1.102
Protein yield	0.608	0.518	0.723	0.571	1.101
Somatic cell score	0.676	0.466	0.752	0.515	1.104

Covariance between TBV, errors of GBV and errors of EBV

The degree of overestimation of accuracy of GBV by any method is determined by omitting $cov(\boldsymbol{T},\boldsymbol{\varepsilon_E})$, $cov(\boldsymbol{T},\boldsymbol{\varepsilon_G})$, and $cov(\boldsymbol{\varepsilon_G},\boldsymbol{\varepsilon_E})$, which are basically not known in reality. However, based on the simulated data, Figure 4.3a and 4.3b is demonstrating $cov(\boldsymbol{T},\boldsymbol{\varepsilon_E})$, $cov(\boldsymbol{T},\boldsymbol{\varepsilon_G})$, and $cov(\boldsymbol{\varepsilon_G},\boldsymbol{\varepsilon_E})$ in Sel and noSel scenarios, respectively, allowing to assess the deviation of these covariances from zero. In Sel scenarios, $cov(\boldsymbol{T},\boldsymbol{\varepsilon_E})$, $cov(\boldsymbol{T},\boldsymbol{\varepsilon_G})$, and $cov(\boldsymbol{\varepsilon_G},\boldsymbol{\varepsilon_E})$ were all significantly different from zero. $cov(\boldsymbol{T},\boldsymbol{\varepsilon_G})$ tended to be upward biased compared to ‘noSel’. The magnitudes of bias varied in different scenarios. Nowadays, most of livestock breeding scheme are under strong selection, which will lead to overestimate the accuracy of GBV if $cov(\boldsymbol{T},\boldsymbol{\varepsilon_E})$, $cov(\boldsymbol{T},\boldsymbol{\varepsilon_G})$, or $cov(\boldsymbol{\varepsilon_G},\boldsymbol{\varepsilon_E})$ assumed to zero. $cov(\boldsymbol{T},\boldsymbol{\varepsilon_E})$ approximated to zero in all noSel scenarios (Figure 4.3b), which is in agreement with the assumption of traditional BLUP. However, in GBLUP, the $cov(\boldsymbol{T}, \boldsymbol{\varepsilon_G})$ were biased from zero and tended to have negative values. $cov(\boldsymbol{\varepsilon_G}, \boldsymbol{\varepsilon_E})$ were biased from zero as well with varying positive values in different scenarios.

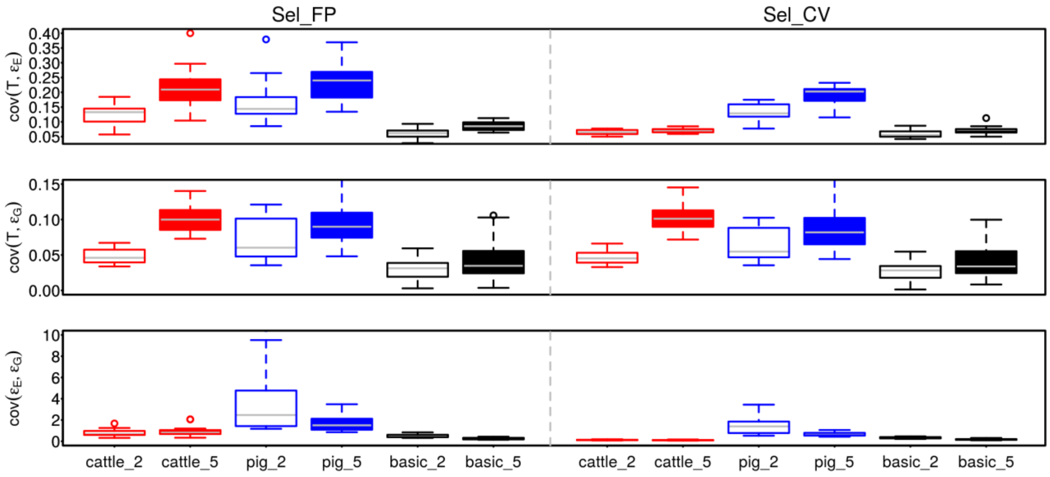


Figure 4.3a: Covariance between errors of EBV and TBV [$cov(\boldsymbol{T}, \boldsymbol{\varepsilon_E})$], covariance between errors of GBV and TBV [$cov(\boldsymbol{T}, \boldsymbol{\varepsilon_G})$], and covariance of errors of EBV and errors of GBV [$cov(\boldsymbol{\varepsilon_E}, \boldsymbol{\varepsilon_G})$] in Sel_FP and Sel_CV scenarios.

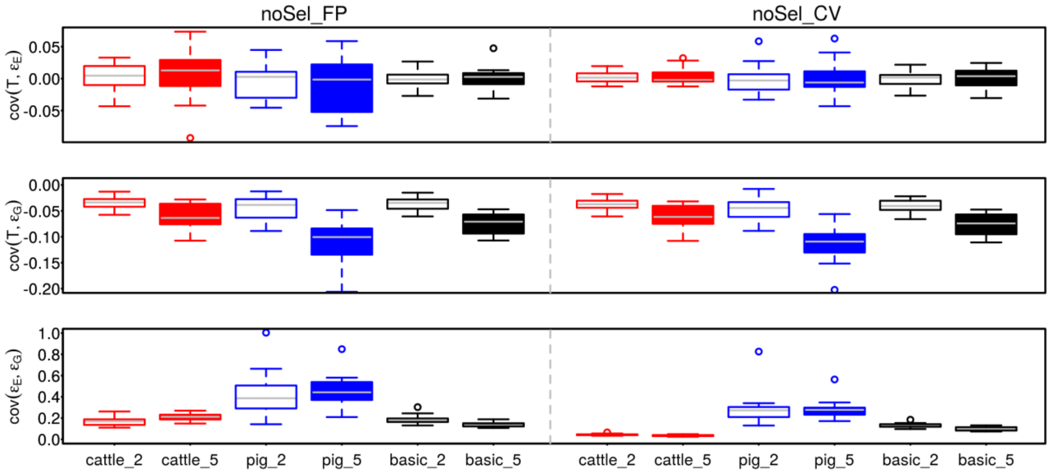


Figure 4.3b: Covariance between errors of EBV and TBV [$cov(\boldsymbol{T}, \boldsymbol{\varepsilon_E})$], covariance between errors of GBV and TBV [$cov(\boldsymbol{T}, \boldsymbol{\varepsilon_G})$], and covariance of errors of EBV and errors of GBV [$cov(\boldsymbol{\varepsilon_E}, \boldsymbol{\varepsilon_G})$] in noSel_FP and noSel_CV scenarios.

Discussion

Comparison among approaches

The results of this study suggested that the new approach proposed here held good potentials for estimating the accuracy of GBV more accurate compared to several approaches used before. In previous studies, one of the most popular approaches of estimating the accuracy of GBV was Acc_H (i.e. r_{EG}/r_{ET}), since it is easy to apply in real data sets by receiving r_{EG} as the empirical correlation between EBV and GBV, $cor(E, G)$, and receiving r_{ET} from the MMEs of traditional BLUP. In real cattle breeding schemes, particularly for bulls in the training set with information from many daughters, the accuracy of EBV is reasonably high even close to one, thus, dividing by r_{ET} does not have a strong effect on r_{EG}. Consequencely, Acc_H were in a resonable range for cattle scenarios, which also makes Acc_H works better in cattle breeding schemes than in others (Figure 4.2). However, an overestimation of accuracy of GBV was still discovered frequently (Goddard, 2009; Hayes et al., 2009b; Goddard et al., 2011). For instance, based on the DRP of more than 900 Yorkshire pigs, Badke et al. (2014) found that accuracy of trait “number of days to 250 lb” estimated by Acc_H was significantly higher than the individual accuracy estimated from MMEs of genomic evaluation. Furthermore, the magnitude of inaccurateness of accuracy of GBV can hardly be quantified in the real data, which makes this approximation of accuracy unreliable. With less favorable data structures, Acc_H appears to be substantially biased, partly resulting in accuracies larger than 1, which we also observed in some simulated scenarios (Figure 4.2) and even in the real cattle data (Table 4.2).

The correlation between EBV (or its derives e.g. DRP, DYD) and GBV (r_{EG}) is a widely used measure to assess quality of prediction especially in cross-validation scenarios (Luan et al., 2009), although it is not claimed to be an approximation of the accuracy of genomic breeding values in the strict sense. Equation (7), however, provides some insight into the link between r_{EG} and r_{GT}. Since $-1 \leq r_{mm} \leq 1$, the range of r_{GT} is between $r_{EG}r_{ET} + \sqrt{(1 - r_{EG}^2)\,(1 - r_{GT}^2)}$ and $r_{EG}r_{ET} - \sqrt{(1 - r_{EG}^2)\,(1 - r_{GT}^2)}$. This means that if you obtain estimates of e.g. r_{EG} = 0.7 and r_{ET} = 0.8, underlying r_{GT} can still be between 0.13 and 0.99. Besides, the overestimation for the accuracy of GBV is stronger when the accuracy of EBV is lower as shown in Figure 4.2, which was also found in Wellmann et al. (2013). It is even possible to underestimate the accuracy of GBV when the accuracy of

EBV is high as shown in Table 4.2 for the real cattle data. Even though, Daetwlyer et al. (2013) suggested to report r_{EG} to compare result across studies, attention has thus to be paid, especially in the case when accuracy of EBV is relatively small. Generally, it is not robust to use the correlation between EBV and GBV as the proxy for the accuracy of GBV.

Beside Acc_H and r_{EG}, Acc_AB ($\frac{r_{GE}}{r_{ET}(1+\frac{cov(\varepsilon_G,\varepsilon_E)}{var(T)})}$) conceptually is a relatively precise formula to estimate the accuracy of GBV in most simulation scenarios. Acc_AB achieved the minimal MSEs in two scenarios (Additional file 4.5), while both Acc_H and r_{EG} had relative larger MSEs in all scenarios. However, in the noSel scenarios, Acc_AB tended to overestimate the accuracy of GBV and provided estimations larger than 1 in some scenarios. In any case, the non-availability of covariance between errors of GBV and errors of EBV limits the applicability of this approach in real data analyses.

Acc_N, however, presented several potential advantages compared to the above listed methods. The items needed in Acc_N can be either obtained from the MMEs of BLUP (r_{ET}) or from the empirical correlation (r_{GE}) which is one of advantages of the new approach. The only undetermined item is r_{mm} with a boundary from -1 to 1, used as a weighting factor. Thus, the overestimation and underestimation of accuracy of GBV is limited to

$$r_{EG}r_{ET} - \sqrt{(1-r_{ET}^2)\,(1-r_{EG}^2)} \le r_{GT} \le r_{EG}r_{ET} + \sqrt{(1-r_{ET}^2)\,(1-r_{EG}^2)}.$$

Besides, we demonstrated the possibility to estimate r_{mm} from the real data set, in which the estimated r_{mm} was relative stable for production traits and SCS. Accuracies of GBV estimated from Acc_N in most scenarios (11 out of 12 in both Sel and noSel scenarios) had the minimum MSEs for the predicted accuracy of GBV (Figure 4.2 and Additional file 4.5) which proved the robust agreement between empirical and predicted accuracy. This can be seen as the second advantage.

It should be mentioned that accuracy of EBV (r_{ET}) used in all the tested approaches is the average of theoretical accuracy for MMEs, since the TBV are unknown in reality, leading to the empirical correlation between EBV and TBV, $cor(E,T)$, to be unknown. The influence of using theoretical accuracy of EBV from MMEs, r_{ET} , instead of $cor(E,T)$ is shown in Figure 4.4. The use of r_{ET} is the main reason that Acc_N tended to provide the same values regardless of the level of empirical accuracies of GBV in noSel scenarios, for example, the standard deviation of $cor(E,T)$ of pig_2 in noSel_CV scenario was 0.051, while the standard deviation of average of r_{ET} was only 0.014 over 20 replicates. By replacing r_{ET} with $cor(E,T)$, the performance of Acc_N is improved. In addition, if the theoretical accuracy of EBV and the optimal weighting factor were known, Acc_N should lead to the exact true accuracy of GBV, since it is based on the definition of a correlation and no assumption about the variance of EBV and variance of GBV in the derivation of Acc_N. When replacing the theoretical accuracy in Acc_AB with empirical correlation, the performance of Acc_AB was better, although there was still a slight overestimation in all tested scenarios. This could be due to the fact that variance of GBV was smaller than the covariance of TBV and GBV (Additional file 4.6). Acc_H with empirical correlation, however, still overestimated the empirical accuracy of GBV clearly.

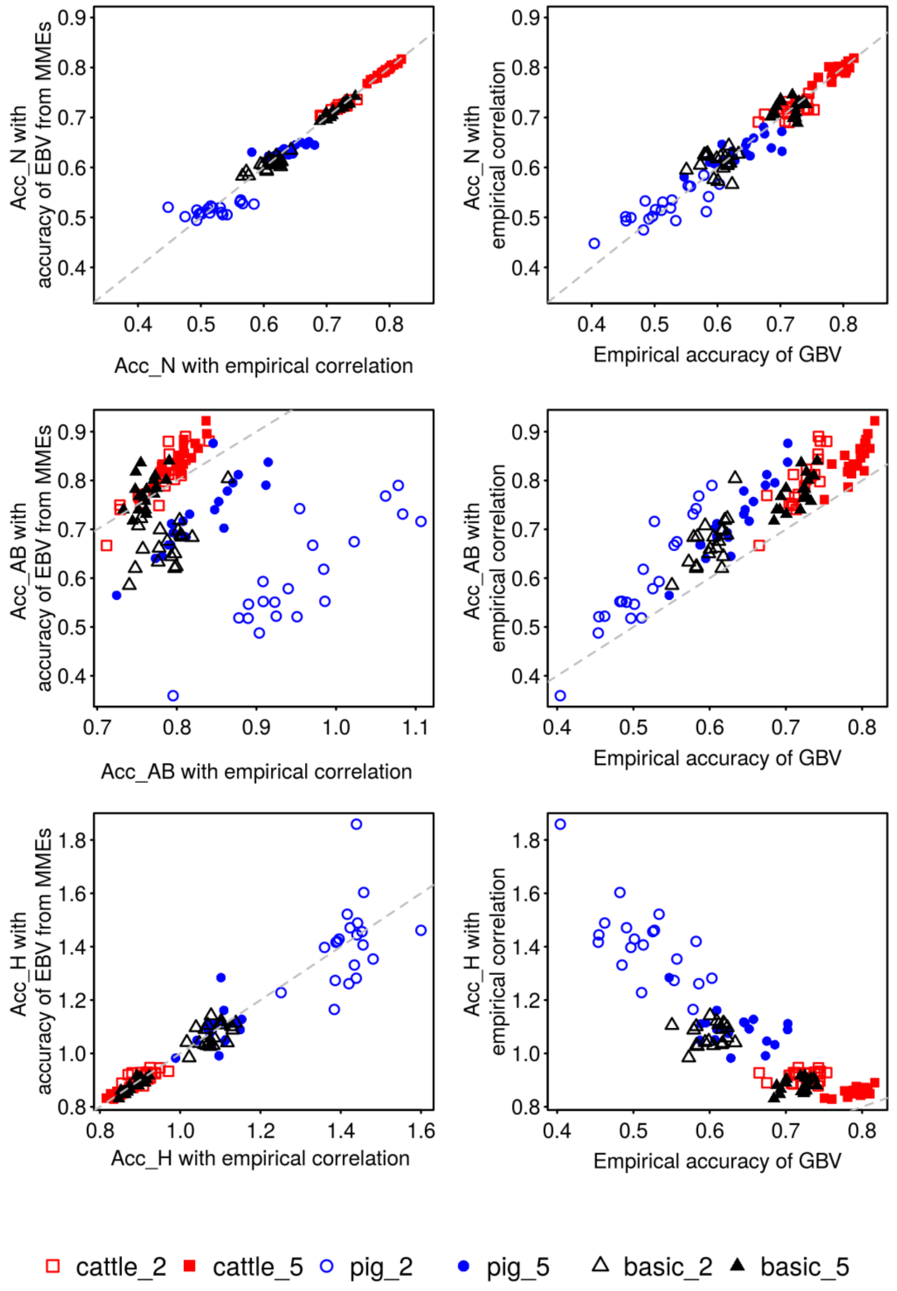


Figure 4.4: Estimated accuracy of GBV based different approaches with theoretical accuracy of EBV from MMEs against that with empirical correlation used based on the data in noSel_CV scenarios.

Weighting factor r_{mm}

As shown in Figure 4.1, with more information available, the curves of MSE for the predicted accuracy of GBV calculated by Acc_N were more flat, and less sensitive to the choice of the weighting factor. This might be because when more information is available, the accuracy of EBV (GBV) is higher and r_{EG} is increased. Therefore, $1 - r_{ET}^2$ and $1 - r_{EG}^2$ in equation (7) is reduced (tends to 0), causing the estimation being less sensitive to the choice of the weighting factors. We reported an approximation for the optimal weighting factor based on available data. Furthermore, the similarity of optimal weighting factors across traits in the real data (Table 4.3) suggested that the empirical value from the data set of a trait could be used in another trait as well, which means that for traits for which we don't have highly reliable EBVs available, it is possible to use the weighing factors derived for traits with highly reliable EBVs available.

EBV, DYD or DRP

In this study, when applying GBLUP, EBVs were used as quasi-phenotype to estimate GBV for the respective traits. However, progeny performances are often used as quasi-phenotype in reality. For example, in cattle breeding schemes, daughter yield deviations (DYD) are commonly employed, especially in the international genetic evaluation of bulls (Liu et al., 2004; Mrode and Swanson, 2004; Szyda et al., 2008), where DYD is defined as the weighted average of a bull's daughters' yields corrected for all fixed effects. Thus, we estimated prediction accuracy of GBV with Acc_N by using DYD as phenotypic data in one of simulated scenarios (cattle_2_Sel_CV) to test the variety of weighting factors. The mean (± S.D.) of the optimal weighting factor was 0.47 (± 0.04) when using DYD as input, compared to 0.45 (± 0.04) using EBV as input. Besides, the correlation between weighting factors when using DYDs as input and that using EBVs as input was 0.87. There was a small shift for the optimal weighting factors (0.45 with EBV, 0.47 with DYD) considering that correlation between EBV and DYD and correlations between GBVs based on EBV and DYD were not exactly one.

Limitation of this study

Most of the analyses were based on different simulation scenarios trying to mimic breeding schemes of cattle or pig, and some 'basic' breeding scenario. However, breeding schemes in reality are much more complex, hence the conclusions drawn from those simulation studies may be of limited value.

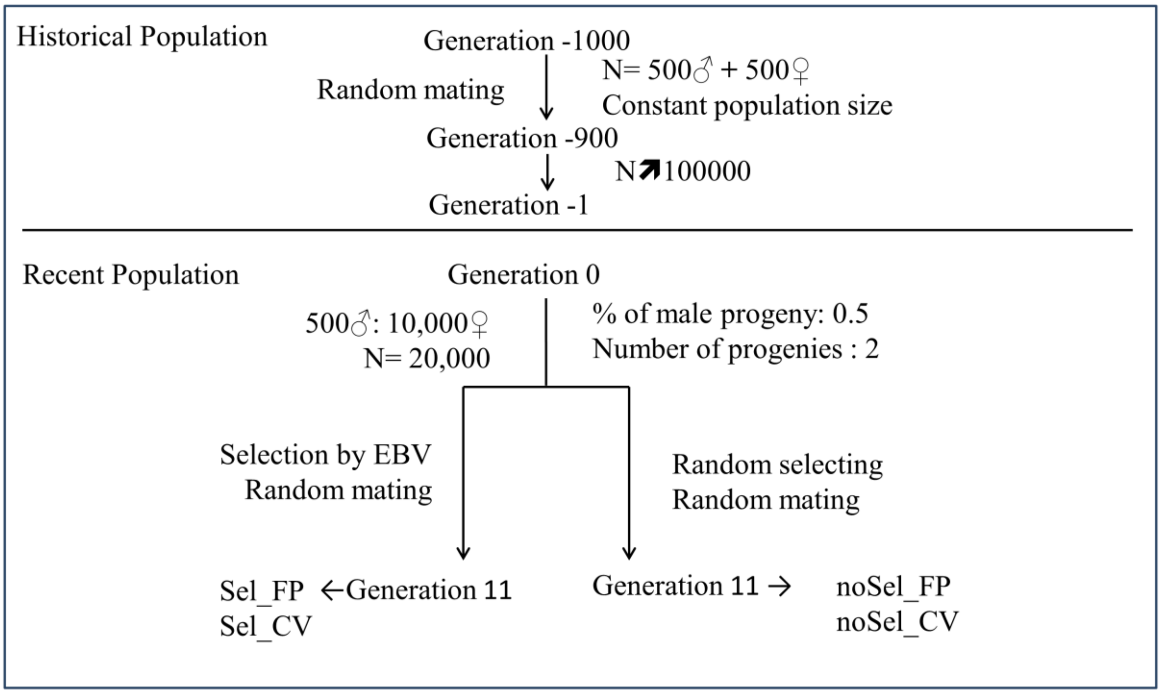
Conclusion

Accuracy of GBV is a critical parameter when performing genomic selection, because it determines the reliability of the selected individuals, and further is a critical parameter in optimization of the design of a breeding scheme. In most approximations suggested so far $cov(\boldsymbol{T}, \boldsymbol{\varepsilon_E}), cov(\boldsymbol{T}, \boldsymbol{\varepsilon_G})$, and $cov(\boldsymbol{\varepsilon_G}, \boldsymbol{\varepsilon_E})$ are assumed to be zero, which they are not, especially when the data are from a population under selection. This could be the reason that most approaches overestimate the accuracy of GBV. In this study, we compared several approaches estimating accuracy of GBV and proposed a new approach, i.e.

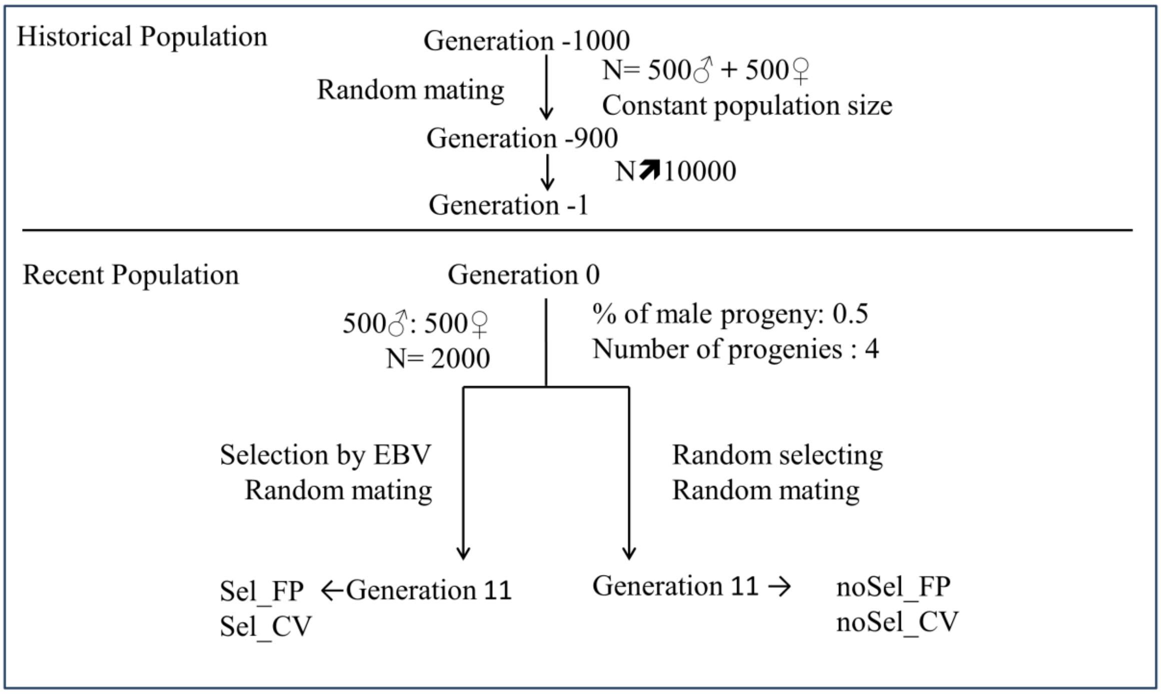
$$r_{GT} \equiv r_{EG}r_{ET} + r_{mm}\sqrt{(1 - r_{ET}^2)\,(1 - r_{EG}^2)}$$

to have a better prediction for the accuracy of GBV. This approximation provides a reliable range for the true accuracy r_{GT}. Based on simulated and real data, the new method held several advantages compared to the previously available approaches: it was computationally applicable and convenient in real data, provided smaller MSE for the predicted accuracy of GBV calculated, and appeared robust in different scenarios. The only unknown parameters here was r_{mm}, for which we suggested an approach to approximate an empirical value from a suitable data set reflecting two separate time points. In conclusion, the new approach provided a better assessment of the accuracy of GBV in many cases.

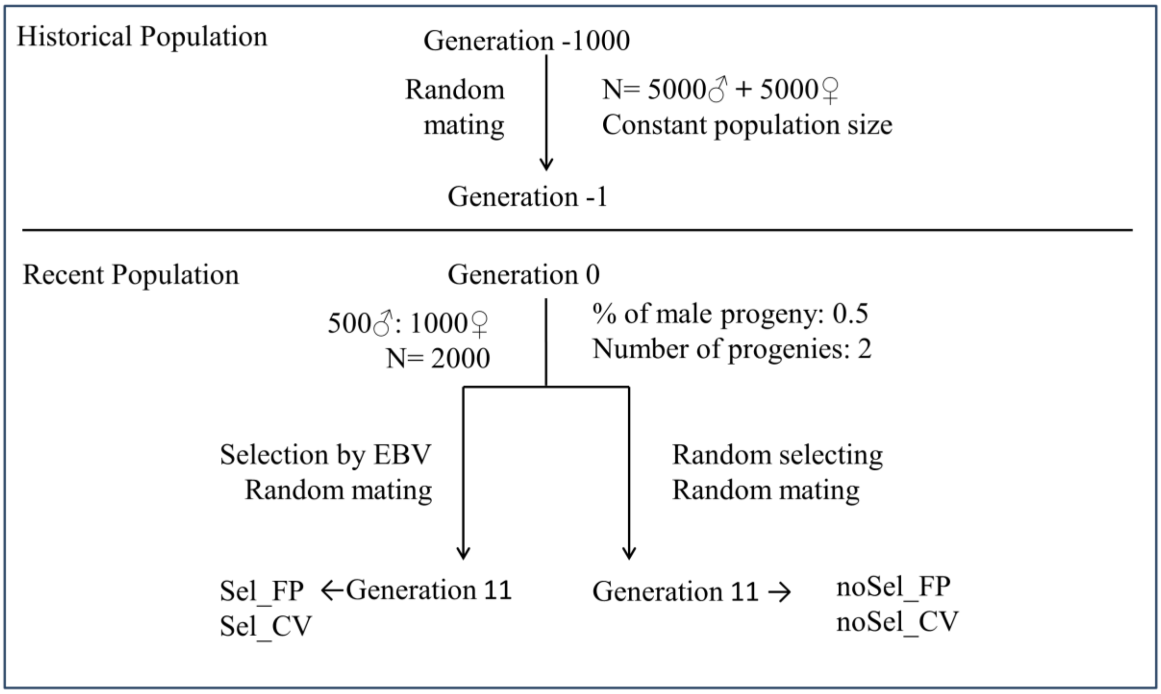
Additional files



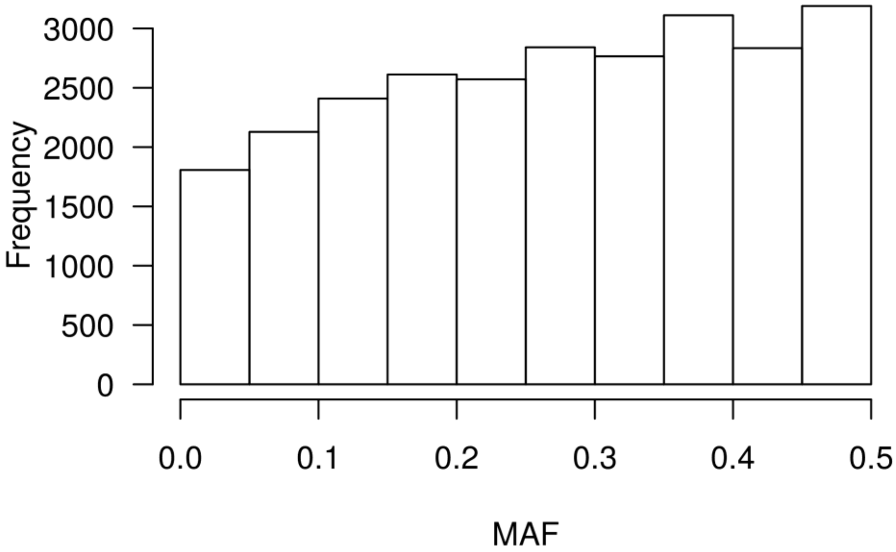
Additional file 4.1a: Schematic representation of the cattle scenario.



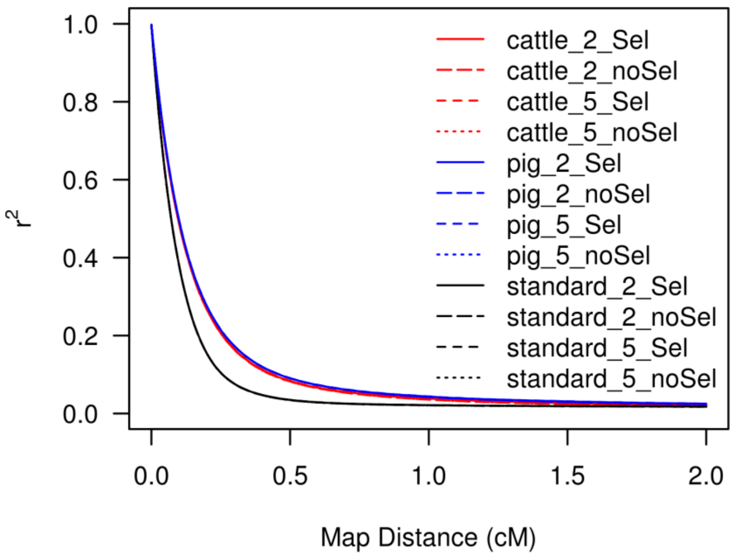
Additional file 4.1b: Schematic representation of the pig scenario.



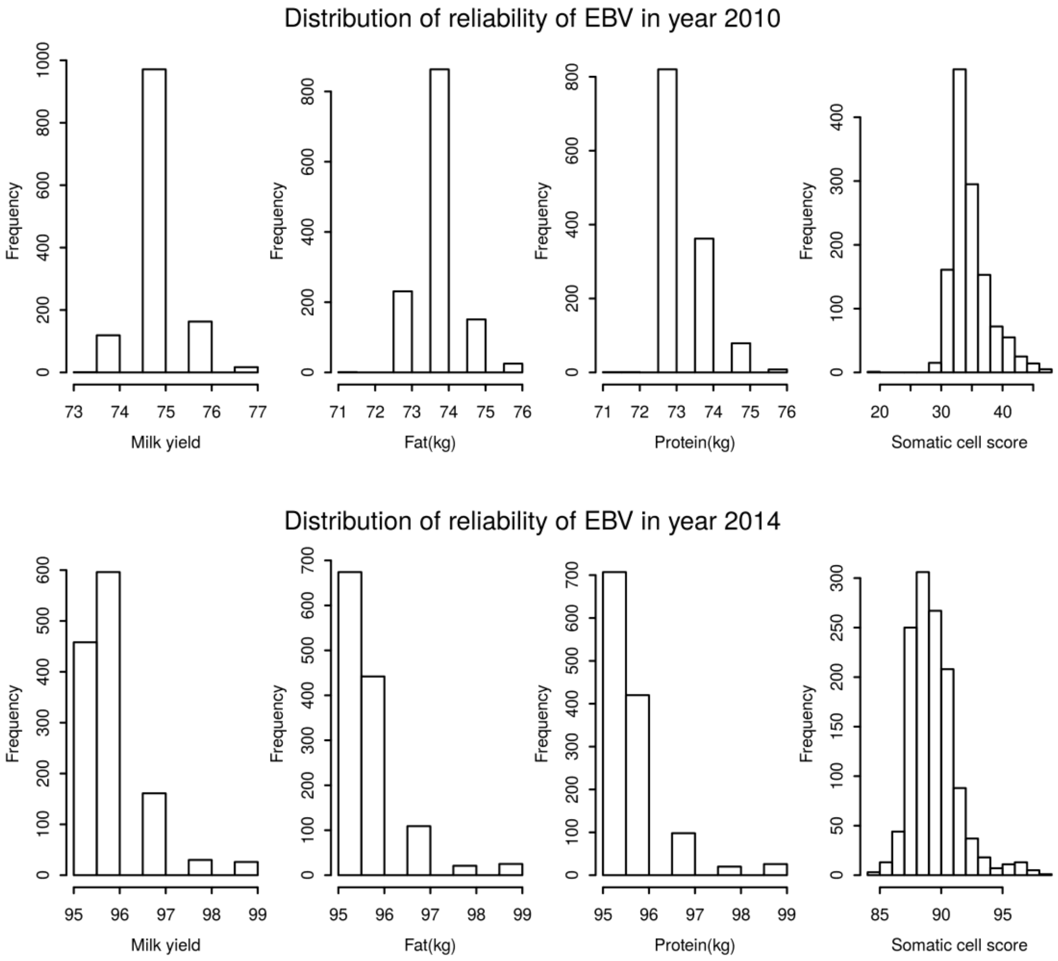
Additional file 4.1c: Schematic representation of the basic scenario.



Additional file 4.2: The minor allele frequency (MAF) based on the simulated genotypes of sires from generation 6 to generation 11 in the first replicate of cattle_5 noSel scenario.



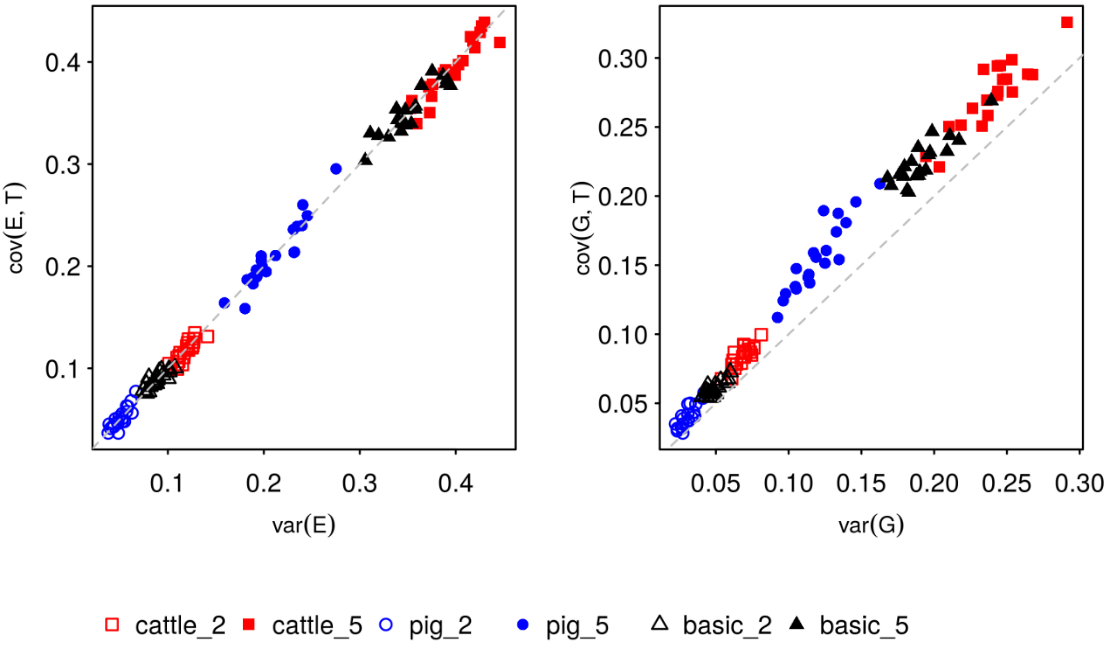
Additional file 4.3: Decline of smoothed averages of linkage disequilibrium among SNPs with distance smaller than 2 cM.



Additional file 4.4: Distribution of reliabilities of EBV for milk yield, fat yield and protein yield for the selected individuals.

Additional file 4.5: Mean squared errors of predicted accuracy of GBV calculated by different approaches in different simulated scenarios of validation sets over 20 replicates

Scenario		Acc_N	Acc_H	Acc_AB	r_{EG}
Sel_FP	cattle_2	0.0012	0.1420	0.0208	0.0057
	cattle_5	0.0017	0.0096	0.0502	0.0412
	pig_2	0.0034	0.9390	0.0421	0.0723
	pig_5	0.0022	0.1911	0.0025	0.0062
	basic_2	0.0010	0.3541	0.0243	0.0315
	basic_5	0.0010	0.0357	0.0005	0.0017
Sel_CV	cattle_2	0.0008	0.0017	0.0226	0.0138
	cattle_5	0.0005	0.0064	0.0174	0.0169
	pig_2	0.0025	0.3911	0.0114	0.0364
	pig_5	0.0017	0.0481	0.0030	0.0023
	basic_2	0.0008	0.1714	0.0090	0.0168
	basic_5	0.0008	0.0181	0.0004	0.0011
noSel_FP	cattle_2	0.0006	0.4382	0.0349	0.0009
	cattle_5	0.0002	0.1346	0.0059	0.0085
	pig_2	0.0023	1.5736	0.3527	0.0616
	pig_5	0.0015	0.5076	0.0850	0.0127
	basic_2	0.0006	0.4104	0.0633	0.0172
	basic_5	0.0003	0.0530	0.0041	0.0008
noSel_CV	cattle_2	0.0003	0.0371	0.0036	0.0009
	cattle_5	0.0001	0.0042	0.0002	0.0012
	pig_2	0.0021	0.8307	0.1991	0.0413
	pig_5	0.0012	0.2124	0.0375	0.0067
	basic_2	0.0005	0.2298	0.0357	0.0106
	basic_5	0.0003	0.0287	0.0022	0.0006



Additional file 4.6: Covariance of EBV and TBV against variance of EBV; Covariance of GBV and TBV against variance of GBV in noSel_CV scenarios.

References

Amer, P. R., and G. Banos. 2010. Implications of avoiding overlap between training and testing data sets when evaluating genomic predictions of genetic merit. J. Dairy Sci. 93:3320–3330.

Badke, Y. M., R. O. Bates, C. W. Ernst, J. Fix, and J. P. Steibel. 2014. Accuracy of estimation of genomic breeding values in pigs using low-density genotypes and imputation. G3 (Bethesda). 4:623–31.

Bijma, P. 2012. Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. J. Anim. Breed. Genet. 129:345–58.

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–65.

Dekkers, J. C. M. 1992. Asymptitic response to selection on best linear unbiased predictors of breeding values. Anim. Prod. 54:351–360.

Erbe, M., B. Gredler, F. R. Seefried, B. Bapst, and H. Simianer. 2013. A function accounting for training set size and marker density to model the average accuracy of genomic prediction. PLoS One 8:e81046.

Falconer, D. S., and T. F. C. Mackay. 1996. Introduction to Quantitative Genetics. 4th ed. Longmans Green, Harlow, Essex, UK.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson. 2009. ASReml User Guide 3.0. VSN International Ltd, Hemel Hempstead, UK.

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128:409–21.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136:245–57.

Habier, D., R. L. Fernando, and D. J. Garrick. 2013. Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194:597–607.

Hayes, B. J., P. J. Bowman, a J. Chamberlain, and M. E. Goddard. 2009a. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92:433–43.

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009b. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. (Camb). 91:47–60.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–447.

Jannink, J.-L., A. J. Lorenz, and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9:166–77.

Liu, Z., F. Reinhardt, a Bünger, and R. Reents. 2004. Derivation and calculation of approximate reliabilities and daughter yield-deviations of a random regression test-day model for genetic evaluation of dairy cattle. J. Dairy Sci. 87:1896–907.

Luan, T., J. a Woolliams, S. Lien, M. Kent, M. Svendsen, and T. H. E. Meuwissen. 2009. The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. Genetics 183:1119–26.

Meuwissen, T., B. Hayes, and M. Goddard. 2013. Accelerating Improvement of Livestock with Genomic Selection. Annu. Rev. Anim. Biosci. 1:221–237.

Mrode, R. a., and G. J. T. Swanson. 2004. Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. Livest. Prod. Sci. 86:253–260.

Nirea, K. G., A. K. Sonesson, J. a Woolliams, and T. H. E. Meuwissen. 2012. Strategies for implementing genomic selection in family-based aquaculture breeding schemes: double haploid sib test populations. Genet. Sel. Evol. 44:30.

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel, P. Revilla, V. M. Rodríguez, J. Moreno-Gonzalez, a Melchinger, E. Bauer, C.-C. Schoen, N. Meyer, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, a Charcosset, and L. Moreau. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics 192:715–28.

Saatchi, M., M. C. McClure, S. D. McKay, M. M. Rolf, J. Kim, J. E. Decker, T. M. Taxis, R. H. Chapple, H. R. Ramey, S. L. Northcutt, S. Bauck, B. Woodward, J. C. M. Dekkers,

R. L. Fernando, R. D. Schnabel, D. J. Garrick, and J. F. Taylor. 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genet. Sel. Evol. 43:40.

Sargolzaei, M., and F. S. Schenkel. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25:680–1.

Szyda, J., E. Ptak, J. Komisarek, and A. Żarnecki. 2008. Practical application of daughter yield deviations in dairy cattle breeding. J. Appl. Genet. 49:183–191.

VanRaden, P. M. 2007. Genomic Measures of Relationship and Inbreeding. Interbull 37:33–36.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–23.

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

Wientjes, Y. C. J., R. F. Veerkamp, and M. P. L. Calus. 2013. The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. Genetics 193:621–31.

Chapter 5 General discussion

The main objective of this thesis was to investigate the possibility of imputing SNP array data up to the whole-genome sequence level (**Chapter 2**), then perform genomic prediction based on the imputed whole-genome sequencing data and SNP array data with different genomic relationship matrices to account for genetic architecture (**Chapter 3**). To further understand the accuracy of genomic prediction, a simulation study was performed to determine the degree of overestimation of the accuracy of genomic prediction, in order to propose a new method measuring the accuracy of genomic breeding values (**Chapter 4**).

In this discussion chapter, some of the important questions arising from the previous studies are discussed. In addition, needs for further investigations are presented.

Impact of imputation on genomic prediction

Genomic prediction is becoming a routine practice for a range of animals in many countries (Van Eenennaam et al., 2014). Considering cost-effectiveness, genomic prediction based on imputed genotypes is often conducted. The question we were interested in is whether conducting genomic prediction with imputed genotypic data has a negative effect on predictive ability or not.

In **Chapter 3**, genomic prediction with imputed whole-genome sequencing (WGS) data was conducted, since it is not realistic to sequence a whole population due to the cost of sequencing. Thus, it is not possible to compare the predictive ability with real WGS data and imputed WGS data so far. However, there are several studies investigating the effect of using imputed genotypic data on genomic prediction with other densities of array data. A strategy often used is: first to mask a certain amount of SNPs in the data set available to produce a lower density data set; then second, use both data sets to perform genomic prediction and then compare the results derived from the two data sets. For example, Pimentel et al. (2015) studied the direct genomic breeding values (DGVs) for 37 traits of 3,494 Brown Swiss candidates from two genotype data sets. The mimicked 6K chip was created by only keeping the SNPs that are both on Illumina BovineLD BeadChip and on Illumina Bovine SNP50 BeadChip, which is the chip that candidates were actually genotyped with. The imputation process was done with Findhap V2 (VanRaden et al., 2011) and FImpute (Sargolzaei et al., 2014) independently. Note that FImpute is one of the imputation programs tested in **Chapter 2**. 6,243 individuals were used as the reference set for imputation, which were the same individuals in the training set of genomic prediction. Pimentel et al. (2015) found that DGVs of top candidates tended to be underestimated, whilst DGVs of bottom candidates tended to be overestimated. They believe that this is because imputation programs tend to give the most frequent haplotype when a clear haplotype cannot be determined.

Unlike the study of Pimentel et al. (2015), in which how imputation error biased the DGVs of candidates was investigated, Khatkar et al. (2012) investigated how imputation errors affected the accuracy of genomic prediction. They did not observe the decrease of predictive ability with imputed 50K genotypes that were imputed from 7K compared to that with actual 50K genotypes for five production traits. Segelke et al. (2012) reported that the loss of reliability of genomic prediction was around 5.3% (1.9%) with imputed genotyped data from 3K (6K) by Findhap and around 1.9% (1%) with imputed genotyped data form 3K (6K) by Beagle (Browning and Browning, 2007), averaged over 12 traits from a EuroGenomics data set including 11,670 Holstein bulls. The predictive ability of somatic cell scores based on imputed 777K genotypes was 3.8% lower than that based on true genotype data in the study of Van Binsbergen et al. (2015) in which 5,503 Holstein Friesian bulls were employed. Chen et al. (2014) studied the de-regressed proofs of more than ten thousand Holstein bulls across several traits and found that the reduction of predictive ability with Bayesian methods was larger than that with GBLUP, especially for traits which are influenced by a few large QTLs (e.g. milk yield, fat percentage, protein percentage).

Based on previous studies, the reduction of predictive ability is affected by the studied population (e.g. number of individuals in the training set), traits of interest (e.g. heritability, genetic architecture), prediction methods (e.g. GBLUP, Bayesian) and imputation programs (e.g. Beagle, Findhap). However, the degree of reduction is almost negligible when averaged over traits and individuals. Nonetheless, the effect may differ from individual to individual, for example between individuals that do have few or many close relatives in the reference population, or individuals who are supposed to have higher direct genomic breeding values (Pimentel et al., 2015). Similar conclusions could be drawn based on WGS data. Thus, attention should be paid when conducting genomic prediction with imputed genotypic data.

Persistency of predictive ability with whole-genome sequencing data across generations

One of the assumed advantages of employing whole-genome sequencing (WGS) data in genomic prediction is that WGS may contain causal mutations for the traits of interest.

When genomic information passes through generations, the linkage disequilibrium between SNPs and QTLs might be broken down due to recombination. QTLs, however, should be passed through if the selection goals are the same as before. Thus, genomic prediction with WGS data is no longer limited by the linkage disequilibrium. Consequently, the predictive ability should be maintained across generations. Cross-validation, as employed in **Chapter 3**, is an often used strategy to compare different data sets or different methods in different fields. In animal breeding the general process of cross-validation is the following: the whole population is randomly split into two sets: the training set in which the (quasi-) phenotype of individuals are known, the validation set in which the (quasi-) phenotype of individuals are pretended to be known. Then the information derived from the training set is used to assess the performance of the validation set. In practice, we are more interested in the performance of young animals for which phenotypes are not available yet and for which the phenotype of their parents and relatives are available, which means that the population should not be split randomly but split by year or age, resulting in a so-called forward prediction. However, studies regarding forward prediction across generations with real WGS data are so far lacking.

To investigate whether predictive ability can be better maintained across generations with WGS data, we carried out the following forward prediction: the data set, which was the same as used in **Chapter 3,** was divided into a training set and validation sets by generation. The training set contained individuals from generation 1 to generation 4, whereas three validation sets were used: individuals in generations 5 and 6 (G5+G6), individuals in generation 5 (G5), and individuals in generation 6 (G6). The same prediction model as in **Chapter 3** was used. The genomic relationship matrix was built according to VanRaden (VanRaden, 2007), which was denoted as $\boldsymbol{G_I}$ in **Chapter 3**. The number of individuals in each generation is given in Table 5.1. Four different data sets were used: HD array data (including 336,224 SNPs), HD_genic data (including 157,393 SNPs), WGS data (including 5,243,860 SNPs), WGS_genic data (including 2,593,054 SNPs).

Table 5.1: Number of individuals in each generation

Generation	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Number	52	49	49	596	78	68	892

Predictive ability (i.e. the correlation between DRPs and DGVs in a respective validation set) of trait eggshell strength based on different data sets is shown in Figure 5.1. In general, a higher consistency of predictive ability with WGS data was not observed for eggshell strength. There was a clear decrease when individuals were two generations away from the training set, and even more so with WGS data, which is contrary to our expectation. In addition, predictive ability with genic SNPs (HD_genic or WGS_genic) was slightly lower than that with the full SNP set (HD and WGS), which contrasts with our results in **Chapter 3**. In **Chapter 3** a 5-fold cross validation strategy was used and the highest predictive ability was achieved by the WGS_genic SNP set for trait eggshell strength.

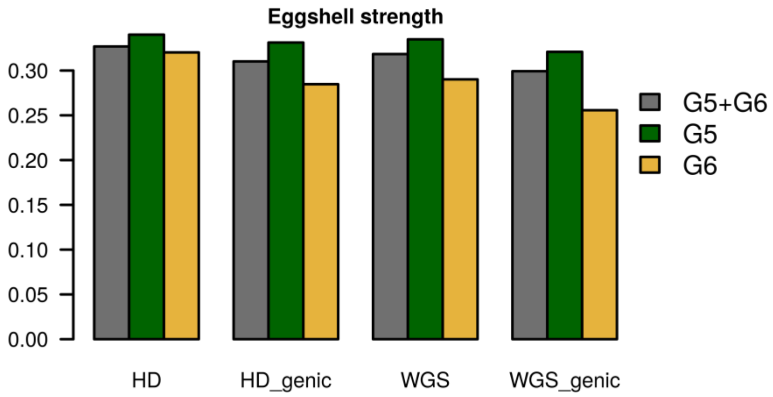


Figure 5.1: Predictive ability of forward prediction for different validation sets based on different SNP data sets.

Based on a simulated population mimicking a bovine breeding scheme, MacLeod et al. (2014) studied the persistency of accuracy with WGS and HD data by comparing the accuracy of genomic prediction with validation individuals either from the same generation as the training set or ten generations away from the training set. They found that the accuracy of genomic prediction based on GBLUP was reduced to ~75% in generation 10 compared to that in generation 0, regardless of the data set (WGS data or HD data) and number of QTLs (15 or 50) simulated. The drop of accuracy based on BayesR was smaller for traits controlled by a small number of QTLs than that by a large number of QTLs. van Binsbergen et al. (2015) carried out a forward genomic prediction with imputed WGS data based on 5,505 bulls. The imputation was done with Beagle by using 429 cattle sequences from the third run of the 1000 bull genomes project as the reference set. They split the validation set according to the presence of close relatives in the training set or not, and found that the drop of predictive reliability, which is the square of predictive ability, was substantial when the validation set was one generation away from the training set. They also did not find that the persistency of predictive reliability exists across generations, which is in agreement with our results.

So far, conducting genomic prediction with WGS did not fulfill all of our expectations. The reasons could be complex. First, imputed WGS data for genomic perdition was used in our study and van Binsbergen's study (2015). Even though it is speculative to assume that predictive ability will be increased with true WGS data, it is possible that persistency of predictive ability will be higher than that with imputed WGS data. Second, we hypothesized that persistency of predictive ability should be higher with WGS data than that with array data, because WGS data should contain the causal mutations for the traits of interest. The traits that we are interested in for animal breeding, however, are commonly quantitative traits. On one hand, quantitative traits mean that there exists numerous variants (including SNPs, CNVs among others) contributing to the variance of the traits. Thus, it might be that there is a large and complex set of true causal mutations. On another hand, the quantitative traits also indicate that they are substantially affected by the environment. Even though some environmental effects are considered as fixed effects in the model to estimate EBVs, it is possible that different individuals have different responses to the same environment, which might be controlled by epigenetics. Consequently, WGS data might not be enough, but data from other –omics layers might be required. Third, based on the previous studies, predictive ability with GBLUP methods were slightly different from that with Bayesian models. Thus, it is possible that the models we have used so far might not capture the advantages of WGS data. Nevertheless, our expectations and findings regarding WGS data should be perceived with caution.

Genomic prediction with DNA structural variations

A major focus of genomic prediction nowadays is on estimating direct genomic breeding values based on SNPs. In fact, DNA variation is far more than single base-pair changes; it also includes copy number variants (CNVs), short insertion and deletion (INDELs), segmental duplications and other motifs. Even though there is no common definition of what a CNV is, it normally refers to a segment of DNA with an arbitrarily defined minimal length of 500bp (Valsesia et al., 2013) for which different individuals vary in the number of copies they carry.

Due to the technological advances in sequencing, numerous DNA structural variations have become accessible. Among others, CNVs have been found to provide a non-negligible contribution of genetic diversity and also have a substantial effect on gene expression. There are several studies that perform association between CNVs and phenotypes by CNV-based genome-wide association studies in humans (Valsesia et al., 2012), chicken (Yi et al., 2014) and other species (Wang et al., 2015). Stranger et al. (2007) performed an association study to investigate the impact of SNPs and CNVs on gene expression. To assay transcript levels, they sequenced total RNA from lymphoblastoid cell lines of 210 individuals with two arrays: Illumina's commercial whole genome expression array and Sentrix Human-6 Expression BeadChip. After filtering, 14,925 transcripts were available for association analyses of expression levels. Stranger et al. (2007) found that SNPs contribute 83.6% of the total detected genetic variation in gene expression while CNVs contribute 17.7%. In addition, the contributions of CNVs are largely independent to the contributions of SNPs.

CNV-based genome-wide association study (GWAS) are more difficult than SNP-based GWAS, not only because there is no fixed starting and ending position for each CNV in each individual (i.e. the same CNV might have different lengths in different individuals), but also because it is difficult to determine the number of independent tests. When using CNVs in genomic prediction, effects of all CNVs are estimated simultaneously, and then multiple-testing is no longer needed. To our knowledge, there is no research so far using CNVs for genomic prediction along with SNPs data. Nonetheless, to better understand the genomic architecture of quantitative traits, CNVs should be taken into account, which might provide interesting insights in the missing heritability.

SNP annotation-based genomic prediction

In **Chapter 3**, predictive abilities with SNPs located in or around genes were higher than those with high density (HD) array data and whole-genome sequencing (WGS) data, which made us wonder which annotated classes contributed to this increase in predictive abilities, or more generally, whether genomic prediction can benefit from SNP annotation.

To investigate the performance of annotated SNP classes in our data, the original GBLUP with a 5-fold cross-validation was carried out, in which the $\boldsymbol{G}$ matrix was built according to VanRaden (2008). The classification of SNPs was the same as in **Chapter 3**. The predictive ability of eggshell strength and the number of SNPs in each class is shown in Figure 5.2 and Table 5.2. Predictive abilities with SNP classified in intron, exon, upstream, downstream, and genic (ranged from 0.409 to 0.411) were slightly higher than that with WGS data (0.407) and higher than that with all HD data (0.397). The highest predictive ability was 0.423 offered by SNPs in the UTR class (including 5’ UTR and 3’ UTR), while the lowest predictive ability was 0.400 offered by SNPs in the intergenic class.

Table 5.2: Number of SNPs in each class

Class	Intronic	Exonic	Upstream	Downstream	UTR
#	2,330,659	63,093	67,701	10,635	60,772
Class	Genic	Intergenic	WGS	HD	-
#	2,593,054	2,650,806	5,243,860	336,224	-

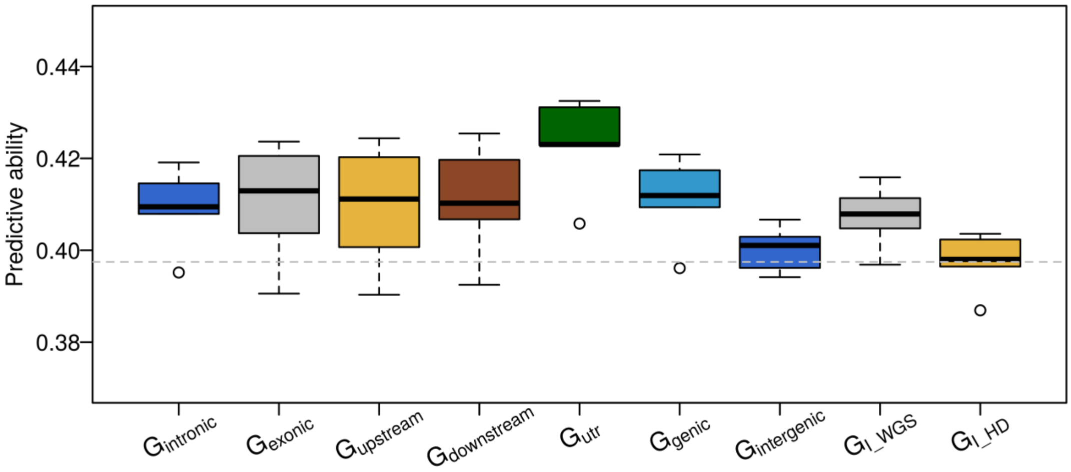


Figure 5.2: Predictive ability based on annotated classes in whole-genome sequencing data for eggshell strength versus the predictive ability based on all SNPs in whole-genome sequencing data and high density data. G_{I_WGS} stands for the predictive ability based on all SNPs in whole-genome sequencing data. G_{I_HD} stands for the predictive ability based on high density array data. The dashed horizontal line denotes the median predictive ability of GBLUP with HD data as a reference.

Do et al. (2015) carried out a SNP annotation-based genomic prediction in a Duroc population that was genotyped by the PorcineSNP60 BeadChip with 30,234 segregating SNPs. They divided a total of 1,272 Duroc pig into a training set (968 pigs) and a validation set (304 pigs) containing the youngest pigs and considered various traits (i.e. residual feed intake, back fat, and average daily gain). SNPs were characterized into 14 different classes including intergenic, upstream, and downstream. Genomic prediction was applied in each SNP class with GBLUP and several Bayesian methods. Predictive ability of residual feed intake in several annotated classes (i.e. gene, upstream, and gene ± 5kb) was higher than the mean predictive ability based on 1,000 random SNPs; however the results were not statistically significant. Predictive ability in the intergenic class was lower than that based on the 1,000 random SNPs.

Although the increased predictive ability in Do et al. (2015) was not significant, which could be due to the small number of SNP in total (30K), it shows the same tendency as our results. The tendency is that predictive ability is higher based on SNPs in or around genes than that with intergenic SNPs. Thus, based on our data set, genomic prediction clearly benefits from SNP annotation. In addition, annotation-based genomic prediction might be useful to reduce the substantial computational demand of incorporating WGS data, because SNP that are in or around genes are only 50% of the WGS data and functional annotated SNPs are even less, e.g. the most informative class of UTR SNPs made only up for 1.16 per cent of all SNPs.

Why understanding genome annotation is important for genomic prediction

Genomic prediction with GBLUP assumes that all SNPs effects are sampled form the same distribution while genomic prediction with Bayesian methods arbitrarily assigns a subset of SNPs having a relatively larger effect than others. However, different SNPs may play different roles in different pathways and gene networks which can lead to different precedence relating to the phenotype. Thus, understanding the genome annotation might help us to determine the precedence of SNPs in respective traits.

Genome annotation normally refers to two annotation processes: structural genome annotation and functional genome annotation (Yandell and Ence, 2012). Structural genome annotation means that the process of classifying SNPs and further identifying genes and their intron-exon structures. As shown in Figure 5.2, genomic prediction with classified SNPs has shown a benefit compared to genomic prediction without structural genome annotation.

Functional genome annotation refers to attaching the biological functions to the structural genome annotations (Yandell and Ence, 2012). Combining functional annotation with the results of a GWAS is an often employed pipeline to using annotation information. More specifically, genes which harbor the significant SNPs of GWAS can be identified from the database. Further the pathways or gene networks of those genes can be determined from the annotation databases. Next, overrepresentation or enrichment analysis for the pathways or gene networks can be done. SNPs related to the significant pathways can be used for the following genomic prediction. The general idea behind overrepresentation and enrichment analysis is to determine whether the representation of certain annotation categories is statically higher than expected by chance or not (Subramanian et al., 2005).

There are several databases to identify gene sets to guide functional genomic prediction. For example, gene ontology (**GO**) which provides the descriptions of gene products across databases for more than 45 species (The Gene Ontology Consortium, 2000; The

Gene Ontology Consortium, 2014). In addition, the National Center for Biotechnology Information (**NCBI**; http://www.ncbi.nlm.nih.gov/), Ensembl Genome Browser (http://www.ensembl.org/index.html), and Kyoto Encyclopedia of Genes and Genomes (**KEGG**; http://www.genome.jp/kegg/) also provide functional annotation of SNPs.

Although studies of genomic prediction based on genome annotation still are rare in livestock, there is, as discussed in **Chapter 3**, an increasing number of such studies showing the benefit of conducting genomic prediction with annotation information e.g. Do et al. (2015) and Pérez-Enciso et al. (2015). In addition, Snelling et al. (2013) performed genomic prediction of heifer pregnancy rate using SNPs detected by genome annotation and all SNPs from the BovineHD chip and found that the accuracy of genomic breeding values with annotation information was higher than that without annotation information. Thus, the value of conducting genomic prediction with functional annotation appears promising.

Annotation of livestock genomes so far is not as complete and good as that of the human genome and some lab animals, such as mouse. Fortunately, evidence exists showing that it is possible to borrow information from human and model animals to infer functions of certain genes. For instance, with the help of the myostatin knock-out mice, a deletion in bovine myostatin genes was confirmed as the causal mutation of the double-muscled phenotype (Grobet et al., 1997) in some beef cattle breeds, which illustrates that a gene has a similar function on the same trait across rather distant species. Thus, annotation information of similar traits from one species might be useful to assist the detection of gene sets or pathways contributing to the traits of other species, even for quantitative traits. One should be aware that the possibility of using functional annotation from human or model animals does not mean that it is not necessary to fill the annotation gap of the respective species based on studies of biological processes, because gene interactions might differ between species, breeds, and even between subpopulations selected for different goals (International Chicken Genome Sequencing Consortium, 2004).

Computational demands with whole-genome sequencing data

With the availability of whole-genome sequencing (WGS) data, computational demands have increased dramatically; especially for genomic prediction. WGS studies revealed up to ~28 million SNPs in cattle (Daetwyler et al., 2014) and ~7 million SNPs in chicken (Rubin et al., 2010). The processing of a GWAS analysis using single marker regression can easily be parallelized; therefore the overall demands are not as big as with genomic prediction. There are several approaches that might be useful to reduce computational demands.

The first aspect that needs to be clarified is whether or not it is necessary to conduct genomic prediction with WGS data. In other words, are the densities of the current commercial array chips sufficient for genomic prediction? Erbe et al. (2013) have suggested a general formula for the average accuracy of genomic breeding values, suggesting that accuracy of genomic prediction is inversely proportional to the log of the marker density. This means, that substantial gains in accuracy can be observed when going from low or moderate marker densities to higher ones (Su et al., 2012; Wellmann et al., 2013) while the scope of improvement is limited when the basic marker density is already high (Vazquez et al., 2010; Makowsky et al., 2011; de los Campos et al., 2013). Given the fact that accuracy of genomic prediction is operating through modelling the segregation of QTLs controlling the trait of interest, the density of array data is sufficient as long as SNPs are in strong linkage disequilibrium with potential QTLs, which means that there is no need to keep on increasing the density of array data once this objective is sufficiently achieved.

If the density of the current array data is considered sufficient, there are several strategies that might be useful to reduce the computational demands. One can remove the SNPs that are in very high or perfect linkage disequilibrium with adjacent SNPs. Harris et al. (2011) reported the predictive ability of first-lactation protein yield with imputed HD data was 0.594, and that with imputed HD data which were edited by deleting one of a pair of SNPs within an interval of 250 SNPs if they are nearly in perfect LD was 0.589. Calus et al. (2015) discarded ~ 59.7 % SNPs on whole-genome sequencing data due to high LD in a simulation study. Gonzalez-Recio et al. (2015) excluded more than 70% of SNPs due to high LD for 3,311 Holstein bulls, which were imputed from the coding regions of 122 Holstein and 26 Jersey animals. Excluding SNPs which are in perfect LD is an effective way to reduce computation demands.

Apart from reducing redundant SNPs, the use of variable selection can be a strategy to reduce the computational demands. Calus et al. (2015) conducted genomic prediction with split-and-merge Bayesian variable selection methods, and reduced the computation time from months to hours by parallelization. Genome annotation could be considered as variable selection as well. As discussed before, SNPs that are related to overrepresented or enriched genes for a trait can be determined by functional annotation, which normally returns only a subset of all available SNPs. In **Chapter 3**, a subset of SNPs which are located around genes was selected from the whole-genome sequencing data and provided the highest predictive ability when averaging over three studies traits. In addition, Figure 5.2 in **General Discussion** shows the predictive ability based on annotated classes in whole-genome sequencing data for eggshell strength and implies that conducting genomic prediction with functional annotation can reduce the computational burden (i.e. the number of SNPs) and potentially increase the predictive ability, although the increase of predictive ability was not observed in the forward prediction (Figure 5.1 of **General Discussion**).

Availability and unavailability of data

The International Chicken Genome Sequencing Consortium provided the first build of the chicken genome in 2004 using DNA from an inbred Jungle Fowl (International Chicken Genome Sequencing Consortium, 2004). The second build was released in 2006, which corrected some of the insufficiency found in the first version. Although the reference genome used in **Chapter 2** and **3** is the fourth build of the chicken genome, which was released in 2011, there are still several deficiencies.

In our data set used in **Chapter 2,** which includes 25 sequenced white layers and 25 sequenced brown layers, there were only 18,641 variants (i.e. SNPs and INDELs) called on chromosome 16 by GATK and 7,434 segregating SNPs. This is extremely few compared to Chromosome 15 and 17, which is almost the same physical length as chromosome 16 (Wallis et al., 2004), where 164,806 and 156,344 variants were called. Thus, massive numbers of presumably segregating SNPs could not be identified due to the poor quality of the reference genome for chromosome 16. In addition, the coverage of chromosome 16 (with an average of 240) was lower than with other chromosomes (with an average of 400) in our re-sequencing runs. Chromosome 16, which contains the major histocompatibility complex (MHC), is still poorly sequenced in the reference genome. The chicken MHC consists of several clusters of highly polymorphic genes with significant effects on the immune system and play an important role in disease resistance (Lamont, 1998; Taylor, 2004). When health and welfare comes into the scope of animal breeding schemes, there are more and more studies focusing on traits relating to functional disorders or diseases e.g. (Liu et al., 2014), whose genomic analysis might be hampered by the poor sequence quality of chromosome 16.

The reference sequence used was derived from the DNA of a single female Red Jungle Fowl. In chicken, it is the female who is the heterogametic sex (i.e. carrying a single copy of the Z and W chromosome, respectively), thus, chromosome Z might be poorly represented in the final assembly. In addition, chromosome W, which has a high repeat content and is different from the rest of chromosomes, is also poorly represented in the current version of the reference genome (International Chicken Genome Sequencing Consortium, 2004; Burt, 2005).

To date, the fourth version of the chicken reference genome is the newest version, which includes 28 pairs of chromosomes, two linkage groups and two sex chromosomes, in total 30 pairs. The chicken karyotype consists of 39 pairs of chromosomes, often classified into five pairs of macro-chromosomes, five pairs of intermediate chromosomes, twenty-eight pairs of micro-chromosomes and two sexual chromosomes (Masabanda et al., 2004). Several of the micro-chromosomes are not present in the reference genome, and it has been found that micro-chromosomes are gene-rich, having twice as many genes as the macro-chromosomes (Burt, 2004). Further, non-nuclear DNA hosted in the mitochondria is also not accounted for in the present assembly. In general, further work is necessary to assess the importance of the entire DNA variation on the predictive ability in chicken.

Sequencing design

Among all the factors characterizing the design of a sequencing experiment, we are only focusing on the following three factors.

1. *Whole-genome sequencing (WGS) vs. whole-genome exome sequencing (WES).* Basically, WGS generates all variations including SNPs, copy number variations, and INDELs across the entire genome, whilst WES only generates variations in protein-coding genes and other functional regions. WES is commonly used in human clinical applications as a cost-effective alternative to WGS (Linderman et al., 2014). In addition, in **General Discussion**, predictive ability with exonic SNPs of WGS (similar to WES data) was higher than that with WGS data based on 5-fold cross-validation (seen in Figure 5.2), even though there is no difference between the predictive ability with WGS data and that with WGS_genic data in the forward prediction (shown in Figure 5.1 in **General Discussion**). But WES has limitations. First, none of the kits available in human genetics can cover all the coding exons (Sulonen et al., 2011). Second, it was shown that both natural and positive selection eventually occurred in the non-coding DNA blocks and some QTLs (Bird et al., 2006; Plomin and Schalkwyk, 2007; Kellis et al., 2014) and selection signatures (Ponting and Lunter, 2006; Haddrill et al., 2008) have been mapped in such blocks, so that important parts

of the genome may be missed by just using exome sequencing (Bird et al., 2006; Drake et al., 2006). However, WGS is more expensive than WES for the same depth of coverage and certainly deserves further consideration in livestock genetics.

2. *Depth of coverage*. This quantity, which is defined as the average number of times a nucleotide is expected to be sequenced (International Human Genome Sequencing Consortium, 2001) is one of the crucial parameters determining the reliability of a SNP called from sequence reads. Theoretical coverage can be estimated by the Lander-Waterman equation as following:

   $$c = LN/G$$

   where c is the theoretical coverage, L is the read length, N is the number of reads and G is haploid genome length (Sims et al., 2014). Empirical coverage per-base is the exact number of times that a base in the reference genome is coverage by reads (Sims et al., 2014). High coverage (on average > 20× coverage) of whole-genome sequencing can provide almost complete genetic variation at an individual level, which also comes at higher costs (Nielsen et al., 2011). In animal breeding, genomic prediction or genomic wide association study are population-based studies, which means that a large number of individuals are needed with genotypic data and phenotypic data. Thus, sequencing with a medium (5-20×) or low (on average < 5× per site per individual) coverage, and genotyping by sequencing with extremely low-coverage (0.1 – 0.5×) are often employed to achieve the maximum number of individuals to be sequenced out of cost-effective concerns (Nielsen et al., 2011; Pasaniuc et al., 2012).

3. *Number and identity of individuals to be sequenced.* The price of sequencing is still relatively expensive, and it is not realistic to sequence all the breeding animals. Thus, identifying the individuals to be sequenced is crucial. In **Chapter 2** and **3**, the sequenced individuals were selected to maximize the explained genetic variation with a limited number of individuals as proposed in Druet et al. (2014). Even though the 25 selected brown layer chickens explained around 85% of total genetic variation probability of missing rare SNPs in the population is large if too few chickens are sequenced. Furthermore, a clear U-shaped distribution of MAFs was not observed in **Chapter 3**, and it is difficult to say whether or not it should exist, because the commercial chickens have been subject to an intensive within-line selection, which could reduce the genetic diversity dramatically, further resulting in the lacking of rare SNPs. Thus, increasing the number of sequencing individuals is crucial.

Main conclusions from this thesis

Based on the results of this thesis, the main conclusions can be summarized as:

- A high proportion of sequence-based SNP calls had high values in different measures of quality with all variant callers. GATK showed a slightly better performance than SAMtools and freebayes.
- When measuring imputation accuracy as the correlation between true and imputed genotypes, Minimac and IMPUTE2 performed slightly better than FImpute.
- When measuring imputation accuracy as the occurrence of Mendelian inconsistencies, FImpute performed better than Minimac and IMPUTE2.
- Imputation accuracy was clearly lower for rare than for common SNPs in all tested imputation programs.
- Imputing high density data to the sequencing level yielded reasonable accuracy, even across several generations from a very limited number of sequenced individuals.

- Little or no benefit was gained when using all imputed whole-genome sequencing data compared to using high density array data with different weighting approaches in the GBLUP model.
- Evidence was found that using genic SNPs for genomic prediction has the potential to improve the predictive ability both with high density array data and whole-genome sequencing data in a cross-validation study.
- The same candidates tended to be selected from a full-sib family of interest regardless of the genotypic data and weighting factors used.

- A newly suggested approach provided a better prediction for the average accuracy of genomic prediction.
- The single unknown parameters in the new approach could be estimated from a real data set.

Reference

Bernstein, B. E., E. Birney, I. Dunham, E. D. Green, C. Gunter, and M. Snyder. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74.

Van Binsbergen, R., M. P. L. Calus, M. C. A. M. Bink, F. A. van Eeuwijk, C. Schrooten, and R. F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 47:71.

Bird, C. P., B. E. Stranger, and E. T. Dermitzakis. 2006. Functional variation and evolution of non-coding DNA. Curr. Opin. Genet. Dev. 16:559–64.

Browning, S. R., and B. L. Browning. 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. 81:1084–97.

Burt, D. W. 2004. The chicken genome and the developmental biologist. Mech. Dev. 121:1129–35.

Burt, D. W. 2005. Chicken genome: current status and future opportunities. Genome Res. 15:1692–1698.

Calus, M. P. L., C. Schrooten, and R. F. Veerkamp. 2015. Split-and-merge Bayesian varibale selection enables efficient genomic prediction using sequence data. In: 66th EAAP Annual Meeting (Warsaw, Poland), p505. p. 505.

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

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R. F. Brøndum, X. Liao, A. Djari, S. C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M.-N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P. J. Bowman, D. Coote, A. J. Chamberlain, C. Anderson, C. P. VanTassell, I. Hulsegge, M. E. Goddard, B. Guldbrandtsen, M. S. Lund, R. F. Veerkamp, D. A. Boichard, R. Fries, and B. J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46:858–865.

Do, D. N., L. L. G. Janss, J. Jensen, and H. N. Kadarmideen. 2015. SNP annotation-based whole genomic prediction and selection : An application to feed efficiency and its component traits in pigs. J. Anim. Sci.:2056–2063.

Drake, J. a, C. Bird, J. Nemesh, D. J. Thomas, C. Newton-Cheh, A. Reymond, L. Excoffier, H. Attar, S. E. Antonarakis, E. T. Dermitzakis, and J. N. Hirschhorn. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat. Genet. 38:223–7.

Druet, T., I. M. Macleod, and B. J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity (Edinb). 112:39–47.

Van Eenennaam, A. L., K. a Weigel, A. E. Young, M. a Cleveland, and J. C. M. Dekkers. 2014. Applied animal genomics: results from the field. Annu. Rev. Anim. Biosci. 2:105–39.

Erbe, M., B. Gredler, F. R. Seefried, B. Bapst, and H. Simianer. 2013. A function accounting for training set size and marker density to model the average accuracy of genomic prediction. PLoS One 8:e81046.

Gonzalez-Recio, O., H. D. Daetwyler, I. M. MacLeod, J. E. Pryce, P. J. Bowman, B. J. Hayes, and M. E. Goddard. 2015. Rare Variants in Transcript and Potential Regulatory Regions Explain a Small Percentage of the Missing Heritability of Complex Traits in Cattle. PLoS One 10:e0143945.

Grobet, L., L. J. R. Martin, D. Poncelet, D. Pirottin, B. Brouwers, J. Riquet, A. Schoeberlein, S. Dunner, F. Menissier, J. Massabanda, R. Fries, R. Hanset, and M. Georges. 1997. A deletion in the bovine myostation gene causes the double-muscled phenotype in cattle. Nat. Genet. 17:71–74.

Haddrill, P. R., D. Bachtrog, and P. Andolfatto. 2008. Positive and negative selection on noncoding DNA in Drosophila simulans. Mol. Biol. Evol. 25:1825–34.

Harris, B. L., F. E. Creagh, A. M. Winkelman, and D. L. Johnson. 2011. Experiences with the Illumina High Density Bovine BeadChip.3–7.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–777.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature 409.

Kellis, M., B. Wold, M. P. Snyder, B. E. Bernstein, A. Kundaje, G. K. Marinov, L. D. Ward, E. Birney, G. E. Crawford, J. Dekker, I. Dunham, L. L. Elnitski, P. J. Farnham, E. a Feingold, M. Gerstein, M. C. Giddings, D. M. Gilbert, T. R. Gingeras, E. D. Green, R. Guigo, T. Hubbard, J. Kent, J. D. Lieb, R. M. Myers, M. J. Pazin, B. Ren, J. a Stamatoyannopoulos, Z. Weng, K. P. White, and R. C. Hardison. 2014. Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. U. S. A. 111:6131–8.

Khatkar, M. S., G. Moser, B. J. Hayes, and H. W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13:538.

Koufariotis, L., Y.-P. P. Chen, S. Bolormaa, and B. J. Hayes. 2014. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. BMC Genomics 15:436.

Lamont, S. J. 1998. The chicken major histocompatibility complex and disease. 7:128–142.

Linderman, M. D., T. Brandt, L. Edelmann, O. Jabado, Y. Kasai, R. Kornreich, M. Mahajan, H. Shah, A. Kasarskis, and E. E. Schadt. 2014. Analytical validation of whole exome and whole genome sequencing for clinical applications. BMC Med. Genomics 7:20.

Liu, T., H. Qu, C. Luo, X. Li, D. Shu, M. S. Lund, and G. Su. 2014. Genomic selection for the improvement of antibody response to Newcastle disease and avian influenza virus in chickens. PLoS One 9:e112685.

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis, and D. Sorensen. 2013. Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet. 9:e1003608.

MacLeod, I. M., B. J. Hayes, and M. E. Goddard. 2014. The effects of demography and long-term selection on the accuracy of genomic prediction with sequence data. Genetics 198:1671–84.

Makowsky, R., N. M. Pajewski, Y. C. Klimentidis, A. I. Vazquez, C. W. Duarte, D. B. Allison, and G. de los Campos. 2011. Beyond missing heritability: prediction of complex traits. PLoS Genet. 7:e1002051.

Masabanda, J. S., D. W. Burt, P. C. M. O. Brien, A. Vignal, V. Fillon, P. S. Walsh, H. Cox, H. G. Tempest, J. Smith, F. Habermann, M. Schmid, Y. Matsuda, M. A. Ferguson-smith, R. P. M. A. Crooijmans, M. A. M. Groenen, D. K. Griffin, U. Mu, and D.- Wu. 2004. Molecular Cytogenetic Definition of the Chicken Genome : The First Complete Avian Karyotype. 1373:1367–1373.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12:443–51.

Nobrega, M., Y. Zhu, I. Plajzer-frick, V. Afzal, and E. M. Rubin. 2004. Megabase deletions of gene deserts result in viable mice. Nature 431:988–993.

Ovcharenko, I., G. G. Loots, M. A. Nobrega, R. C. Hardison, W. Miller, and L. Stubbs. 2005. Evolution and functional classification of vertebrate gene deserts. :137–145.

Pasaniuc, B., N. Rohland, P. J. Mclaren, K. Garimella, H. Li, N. Gupta, B. Neale, M. Daly, P. Sklar, F. Sullivan, S. Bergen, J. L. Moran, C. M. Hultman, P. Lichtenstein, P. Magnusson, S. M. Purcell, D. W. Haas, L. Liang, N. Patterson, P. I. W. De Bakker, D. Reich, and A. L. Price. 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. Nat. Genet. 44:631–635.

Pérez-Enciso, M., J. C. Rincón, and A. Legarra. 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet. Sel. Evol. 47:43.

Pimentel, E. C. G., C. Edel, R. Emmerling, and K.-U. Götz. 2015. How imputation errors bias genomic predictions. J. Dairy Sci. 98:4131–8.

Plomin, R., and L. C. Schalkwyk. 2007. Microarrays. Dev. Sci. 10:19–23.

Ponting, C. P., and G. Lunter. 2006. Signatures of adaptive evolution within human non-coding sequence. Hum. Mol. Genet. 15 Spec No 2:R170–5.

Rubin, C.-J., M. C. Zody, J. Eriksson, J. R. S. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallböök, F. Besnier, O. Carlborg, B. Bed'hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, and L. Andersson. 2010. Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464:587–91.

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

Segelke, D., J. Chen, Z. Liu, F. Reinhardt, G. Thaller, and R. Reents. 2012. Reliability of genomic prediction for German Holsteins using imputed genotypes from low-density chips. J. Dairy Sci. 95:5403–11.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15:121–32.

Snelling, W. M., R. A. Cushman, J. W. Keele, C. Maltecca, M. G. Thomas, M. R. S. Fortes, and A. Reverter. 2013. BREEDING AND GENETICS SYMPOSIUM : Networks and pathways to guide genomic selection. J. Anim. Sci.:537–552.

Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. De Grassi, C. Lee, C. Tyler-smith, N. Carter, S. W. Scherer, S. Tavaré, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis. 2007. Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. Science (80-.). 315:848–854.

Su, G., R. F. Brøndum, P. Ma, B. Guldbrandtsen, G. P. Aamand, and M. S. Lund. 2012. Comparison of genomic predictions using medium-density (∼54,000) and high-density (∼777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. J. Dairy Sci. 95:4657–65.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, and B. L. Ebert. 2005. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. PNAS 102:15545–15550.

Sulonen, A.-M., P. Ellonen, H. Almusa, M. Lepistö, S. Eldfors, S. Hannula, T. Miettinen, H. Tyynismaa, P. Salo, C. Heckman, H. Joensuu, T. Raivio, A. Suomalainen, and J. Saarela. 2011. Comparison of solution-based exome capture methods for next generation sequencing. Genome Biol. 12:R94.

Taylor, R. L. 2004. Major Histocompatibility (B) Complex Control of Responses Against Rous Sarcomas Poult. Sci. 83 :638–649.

The Gene Ontology Consortium. 2000. Gene Ontology : tool for the unification of biology. Nat. Genet. 25:25–29.

The Gene Ontology Consortium. 2014. Gene Ontology Consortium: going forward. Nucleic Acids Res. 43:1049–1056.

Valsesia, A., A. Macé, S. Jacquemont, J. S. Beckmann, and Z. Kutalik. 2013. The Growing Importance of CNVs: New Insights for Detection and Clinical Interpretation. Front. Genet. 4:92.

Valsesia, A., B. J. Stevenson, D. Waterworth, V. Mooser, P. Vollenweider, G. Waeber, C. V. Jongeneel, J. S. Beckmann, Z. Kutalik, and S. Bergmann. 2012. Identification and validation of copy number variants using SNP genotyping arrays from a large clinical cohort. BMC Genomics 13:241.

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

VanRaden, P. M. 2007. Genomic Measures of Relationship and Inbreeding. Interbull 37:33–36.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–23.

Vazquez, a I., G. J. M. Rosa, K. a Weigel, G. de los Campos, D. Gianola, and D. B. Allison. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J. Dairy Sci. 93:5942–9.

Wallis, J. W., J. Aerts, and M. A. M. Groenen. 2004. A physical map of the chicken genome. Nature 432:761–764.

Wang, L., L. Xu, X. Liu, T. Zhang, N. Li, E. H. Hay, Y. Zhang, H. Yan, K. Zhao, G. E. Liu, L. Zhang, and L. Wang. 2015. Copy number variation-based genome wide association study reveals additional variants contributing to meat quality in Swine. Sci. Rep. 5:12535.

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

Yandell, M., and D. Ence. 2012. A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. 13:329–42.

Yi, G., L. Qu, J. Liu, Y. Yan, G. Xu, and N. Yang. 2014. Genome-wide patterns of copy number variation in the diversified chicken genomes using next-generation sequencing. BMC Genomics 15:962.

Acknowledgements

I would like to thank

Prof. Dr. Henner Simianer for being my main supervisor and giving me the opportunity to work on those interesting topics.

Dr. Malena Erbe for guiding me patiently through my PhD study. I have gained priceless knowledge and skills from you.

Prof. Dr Jörn Bennewitz for accepting to co-reference this thesis.

Prof. Dr. Christoph Knorr for being my third supervisor.

Dr. Xiangdong Ding for recommending me to Goettingen University.

Chinese Scholarship Council for the financial support.

All the colleagues from the Department of Animal Sciences, especially the Animal Breeding and Genetics Group, and in highlights, Mahmood, Christian, Anna, Reza, Saber and Swetlana for cultivating a great scientific and social working environment.

Ms. Döring for helping me with all the work-related and personal matters of life in Goettingen.

My Mother for her never-ending supports through my whole life.