Imputation Performance

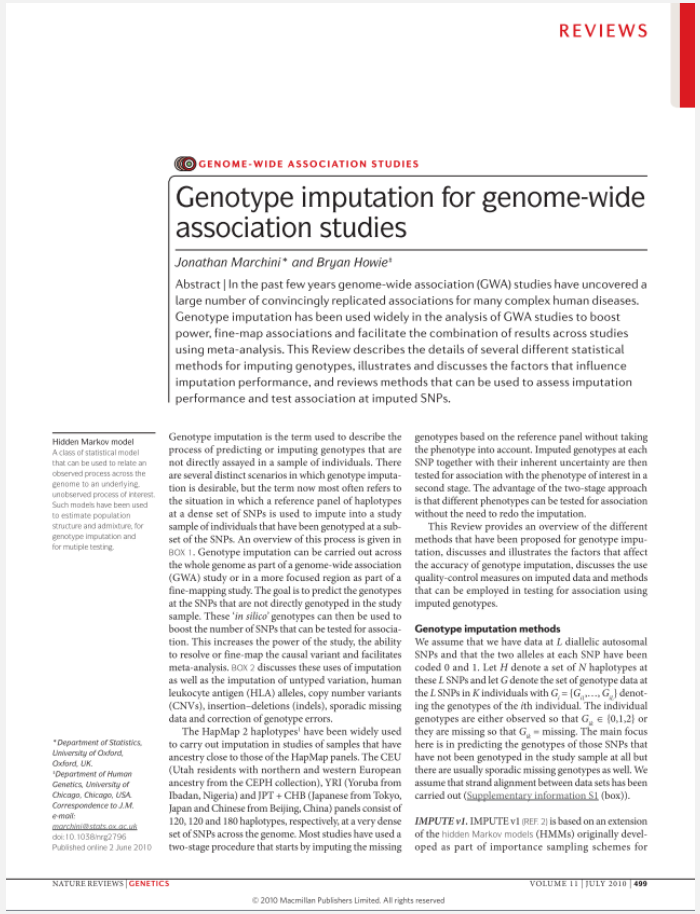
GA-Workshop 28-Sep-2020

Yu Wang (yu.wang@lic.co.nz)

genomics aotearoa

Topics today

- How to evaluate the imputation accuracy?
- What are the parameters influence accuracy?
- Imputation to the sequence level



REVIEWS

GENOME-WIDE ASSOCIATION STUDIES

Genotype imputation for genome-wide association studies

Jonathan Marchini and Bryan Howie‡*

Abstract | In the past few years genome-wide association (GWA) studies have uncovered a large number of convincingly replicated associations for many complex human diseases. Genotype imputation has been used widely in the analysis of GWA studies to boost power, fine-map associations and facilitate the combination of results across studies using meta-analysis. This Review describes the details of several different statistical methods for imputing genotypes, illustrates and discusses the factors that influence imputation performance, and reviews methods that can be used to assess imputation performance and test association at imputed SNPs.

Hidden Markov model
A class of statistical model that can be used to relate an observed process across the genome to an underlying, unobserved process of interest. Such models have been used to estimate population structure and admixture, for genotype imputation and for multiple testing.

Genotype imputation is the term used to describe the process of predicting or imputing genotypes that are not directly assayed in a sample of individuals. There are several distinct scenarios in which genotype imputation is desirable, but the term now most often refers to the situation in which a reference panel of haplotypes at a dense set of SNPs is used to impute into a study sample of individuals that have been genotyped at a subset of the SNPs. An overview of this process is given in BOX 1. Genotype imputation can be carried out across the whole genome as part of a genome-wide association (GWA) study or in a more focused region as part of a fine-mapping study. The goal is to predict the genotypes at the SNPs that are not directly genotyped in the study sample. These '*in silico*' genotypes can then be used to boost the number of SNPs that can be tested for association. This increases the power of the study, the ability to resolve or fine-map the causal variant and facilitates meta-analysis. BOX 2 discusses these uses of imputation as well as the imputation of untyped variation, human leukocyte antigen (HLA) alleles, copy number variants (CNVs), insertion–deletions (indels), sporadic missing data and correction of genotype errors.

The HapMap 2 haplotypes[1] have been widely used to carry out imputation in studies of samples that have ancestry close to those of the HapMap panels. The CEU (Utah residents with northern and western European ancestry from the CEPH collection), YRI (Yoruba from Ibadan, Nigeria) and JPT + CHB (Japanese from Tokyo, Japan and Chinese from Beijing, China) panels consist of 120, 120 and 180 haplotypes, respectively, at a very dense set of SNPs across the genome. Most studies have used a two-stage procedure that starts by imputing the missing genotypes based on the reference panel without taking the phenotype into account. Imputed genotypes at each SNP together with their inherent uncertainty are then tested for association with the phenotype of interest in a second stage. The advantage of the two-stage approach is that different phenotypes can be tested for association without the need to redo the imputation.

This Review provides an overview of the different methods that have been proposed for genotype imputation, discusses and illustrates the factors that affect the accuracy of genotype imputation, discusses the use quality-control measures on imputed data and methods that can be employed in testing for association using imputed genotypes.

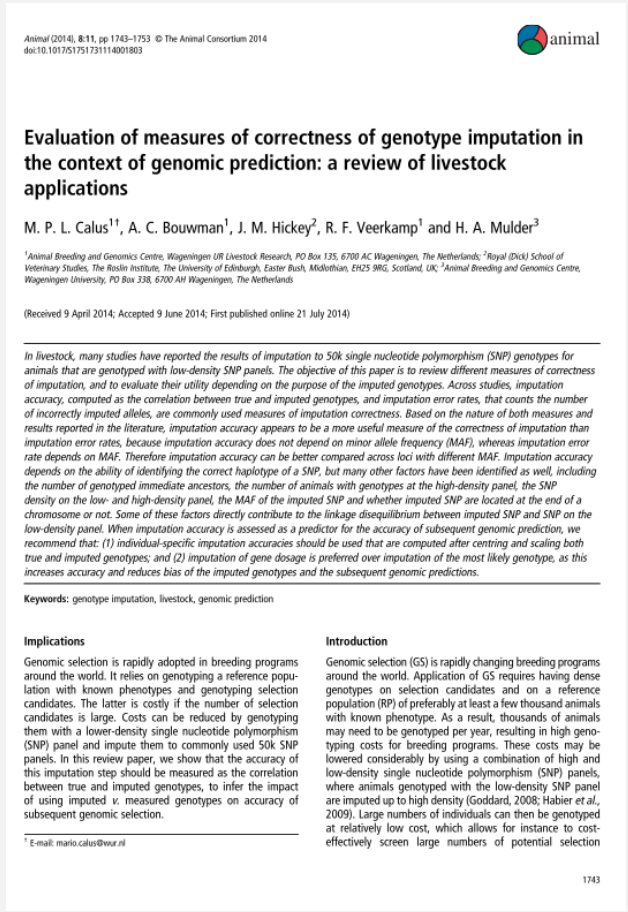
Genotype imputation methods

We assume that we have data at L diallelic autosomal SNPs and that the two alleles at each SNP have been coded 0 and 1. Let H denote a set of N haplotypes at these L SNPs and let G denote the set of genotype data at the L SNPs in K individuals with $G_i = [G_{i1}, \ldots, G_{iL}]$ denoting the genotypes of the ith individual. The individual genotypes are either observed so that $G_{ij} \in \{0,1,2\}$ or they are missing so that G_{ij} = missing. The main focus here is in predicting the genotypes of those SNPs that have not been genotyped in the study sample at all but there are usually sporadic missing genotypes as well. We assume that strand alignment between data sets has been carried out (Supplementary information S1 (box)).

IMPUTE v1. IMPUTE v1 (REF. 2) is based on an extension of the hidden Markov models (HMMs) originally developed as part of importance sampling schemes for

*Department of Statistics, University of Oxford, Oxford, UK.
‡Department of Human Genetics, University of Chicago, Chicago, USA.
Correspondence to J.M.
e-mail: marchini@stats.ox.ac.uk
doi:10.1038/nrg2796
Published online 2 June 2010

Marchini, Jonathan, and Bryan Howie. "Genotype imputation for genome-wide association studies." *Nature Reviews Genetics* 11.7 (2010): 499-511.



Animal (2014), **8:11**, pp 1743–1753 © The Animal Consortium 2014
doi:10.1017/S1751731114001803

animal

Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications

M. P. L. Calus[1†], A. C. Bouwman[1], J. M. Hickey[2], R. F. Veerkamp[1] and H. A. Mulder[3]

[1]*Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, PO Box 135, 6700 AC Wageningen, The Netherlands;* [2]*Royal (Dick) School of Veterinary Studies, The Roslin Institute, The University of Edinburgh, Easter Bush, Midlothian, EH25 9RG, Scotland, UK;* [3]*Animal Breeding and Genomics Centre, Wageningen University, PO Box 338, 6700 AH Wageningen, The Netherlands*

(Received 9 April 2014; Accepted 9 June 2014; First published online 21 July 2014)

In livestock, many studies have reported the results of imputation to 50k single nucleotide polymorphism (SNP) genotypes for animals that are genotyped with low-density SNP panels. The objective of this paper is to review different measures of correctness of imputation, and to evaluate their utility depending on the purpose of the imputed genotypes. Across studies, imputation accuracy, computed as the correlation between true and imputed genotypes, and imputation error rates, that counts the number of incorrectly imputed alleles, are commonly used measures of imputation correctness. Based on the nature of both measures and results reported in the literature, imputation accuracy appears to be a more useful measure of the correctness of imputation than imputation error rates, because imputation accuracy does not depend on minor allele frequency (MAF), whereas imputation error rate depends on MAF. Therefore imputation accuracy can be better compared across loci with different MAF. Imputation accuracy depends on the ability of identifying the correct haplotype of a SNP, but many other factors have been identified as well, including the number of genotyped immediate ancestors, the number of animals with genotypes at the high-density panel, the SNP density on the low- and high-density panel, the MAF of the imputed SNP and whether imputed SNP are located at the end of a chromosome or not. Some of these factors directly contribute to the linkage disequilibrium between imputed SNP and SNP on the low-density panel. When imputation accuracy is assessed as a predictor for the accuracy of subsequent genomic prediction, we recommend that: (1) individual-specific imputation accuracies should be used that are computed after centring and scaling both true and imputed genotypes; and (2) imputation of gene dosage is preferred over imputation of the most likely genotype, as this increases accuracy and reduces bias of the imputed genotypes and the subsequent genomic predictions.

Keywords: genotype imputation, livestock, genomic prediction

Implications

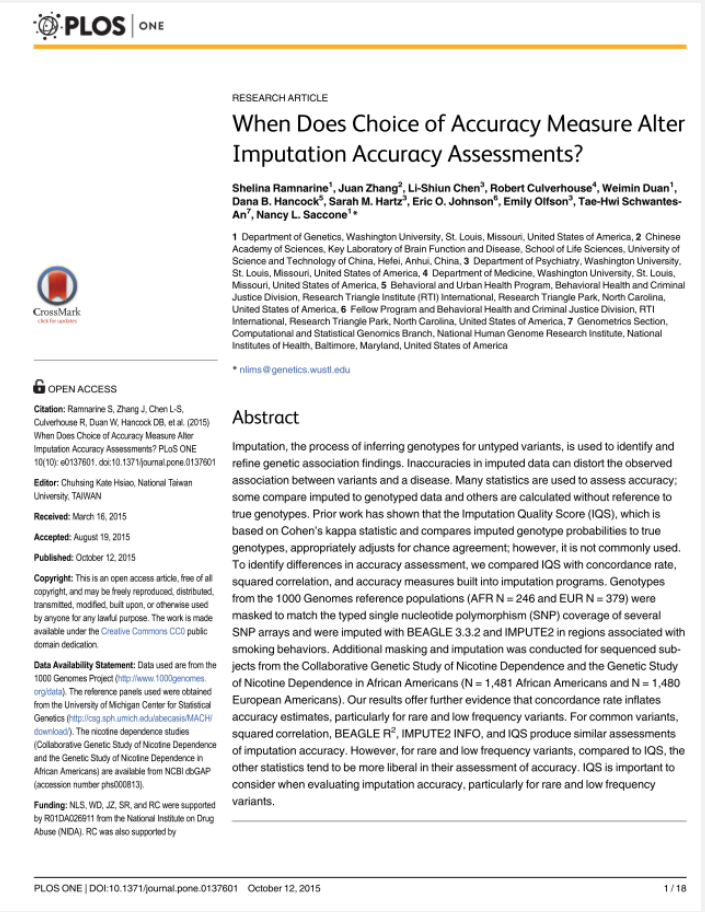
Genomic selection is rapidly adopted in breeding programs around the world. It relies on genotyping a reference population with known phenotypes and genotyping selection candidates. The latter is costly if the number of selection candidates is large. Costs can be reduced by genotyping them with a lower-density single nucleotide polymorphism (SNP) panel and impute them to commonly used 50k SNP panels. In this review paper, we show that the accuracy of this imputation step should be measured as the correlation between true and imputed genotypes, to infer the impact of using imputed *v.* measured genotypes on accuracy of subsequent genomic selection.

† E-mail: mario.calus@wur.nl

Introduction

Genomic selection (GS) is rapidly changing breeding programs around the world. Application of GS requires having dense genotypes on selection candidates and on a reference population (RP) of preferably at least a few thousand animals with known phenotype. As a result, thousands of animals may need to be genotyped per year, resulting in high genotyping costs for breeding programs. These costs may be lowered considerably by using a combination of high and low-density single nucleotide polymorphism (SNP) panels, where animals genotyped with the low-density SNP panel are imputed up to high density (Goddard, 2008; Habier *et al.*, 2009). Large numbers of individuals can then be genotyped at relatively low cost, which allows for instance to cost-effectively screen large numbers of potential selection

Calus, M. P. L., *et al.* "Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications." *Animal* 8.11 (2014): 1743-1753.



PLOS ONE

RESEARCH ARTICLE

When Does Choice of Accuracy Measure Alter Imputation Accuracy Assessments?

Shelina Ramnarine[1], Juan Zhang[2], Li-Shiun Chen[3], Robert Culverhouse[4], Weimin Duan[1], Dana B. Hancock[5], Sarah M. Hartz[3], Eric O. Johnson[6], Emily Olfson[3], Tae-Hwi Schwantes-An[7], Nancy L. Saccone[1]*

1 Department of Genetics, Washington University, St. Louis, Missouri, United States of America, **2** Chinese Academy of Sciences, Key Laboratory of Brain Function and Disease, School of Life Sciences, University of Science and Technology of China, Hefei, Anhui, China, **3** Department of Psychiatry, Washington University, St. Louis, Missouri, United States of America, **4** Department of Medicine, Washington University, St. Louis, Missouri, United States of America, **5** Behavioral and Urban Health Program, Behavioral Health and Criminal Justice Division, Research Triangle Institute (RTI) International, Research Triangle Park, North Carolina, United States of America, **6** Fellow Program and Behavioral Health and Criminal Justice Division, RTI International, Research Triangle Park, North Carolina, United States of America, **7** Genometrics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, United States of America

* nlims@genetics.wustl.edu

OPEN ACCESS

Citation: Ramnarine S, Zhang J, Chen L-S, Culverhouse R, Duan W, Hancock DB, et al. (2015) When Does Choice of Accuracy Measure Alter Imputation Accuracy Assessments? PLoS ONE 10(10): e0137601. doi:10.1371/journal.pone.0137601

Editor: Chuhsing Kate Hsiao, National Taiwan University, TAIWAN

Received: March 16, 2015

Accepted: August 19, 2015

Published: October 12, 2015

Data Availability Statement: Data used are from the 1000 Genomes Project (http://www.1000genomes.org/data). The reference panels used were obtained from the University of Michigan Center for Statistical Genetics (http://csg.sph.umich.edu/abecasis/MACH/download/). The nicotine dependence studies (Collaborative Genetic Study of Nicotine Dependence and the Genetic Study of Nicotine Dependence in African Americans) are available from NCBI dbGAP (accession number phs000813).

Funding: NLS, WD, JZ, SR, and RC were supported by R01DA026911 from the National Institute on Drug Abuse (NIDA). RC was also supported by

Abstract

Imputation, the process of inferring genotypes for untyped variants, is used to identify and refine genetic association findings. Inaccuracies in imputed data can distort the observed association between variants and a disease. Many statistics are used to assess accuracy; some compare imputed to genotyped data and others are calculated without reference to true genotypes. Prior work has shown that the Imputation Quality Score (IQS), which is based on Cohen's kappa statistic and compares imputed genotype probabilities to true genotypes, appropriately adjusts for chance agreement; however, it is not commonly used. To identify differences in accuracy assessment, we compared IQS with concordance rate, squared correlation, and accuracy measures built into imputation programs. Genotypes from the 1000 Genomes reference populations (AFR N = 246 and EUR N = 379) were masked to match the typed single nucleotide polymorphism (SNP) coverage of several SNP arrays and were imputed with BEAGLE 3.3.2 and IMPUTE2 in regions associated with smoking behaviors. Additional masking and imputation was conducted for sequenced subjects from the Collaborative Genetic Study of Nicotine Dependence and the Genetic Study of Nicotine Dependence in African Americans (N = 1,481 African Americans and N = 1,480 European Americans). Our results offer further evidence that concordance rate inflates accuracy estimates, particularly for rare and low frequency variants. For common variants, squared correlation, BEAGLE R^2, IMPUTE2 INFO, and IQS produce similar assessments of imputation accuracy. However, for rare and low frequency variants, compared to IQS, the other statistics tend to be more liberal in their assessment of accuracy. IQS is important to consider when evaluating imputation accuracy, particularly for rare and low frequency variants.

Ramnarine, Shelina, *et al.* "When does choice of accuracy measure alter imputation accuracy assessments?." *PloS one* 10.10 (2015): e0137601.

Why we need accurate imputation

- Fill in missing genotypes from the lab
- Merge data sets with genotypes on different arrays
    - Eg. Affy and Illumina data
- Impute from low density to high density
    - 7K-> 50K (save $$$)
    - 50K->800K
    - capture power of higher density?
- Sequence expensive, can we impute to full sequence data?

What are the parameters to evaluate imputation accuracy

Genotype concordance (1-error rate)

Genotype concordance is computed per locus as the percentage (or proportion) of alleles or genotypes that is imputed incorrectly. A closely related measure is the percentage of correctly imputed alleles or genotypes, which can simply be calculated as 100% minus the imputation error rate. Need to know the true status.

Genotype correlation

Pearson correlation coefficient between true and imputed genotypes. Need to know the true status.

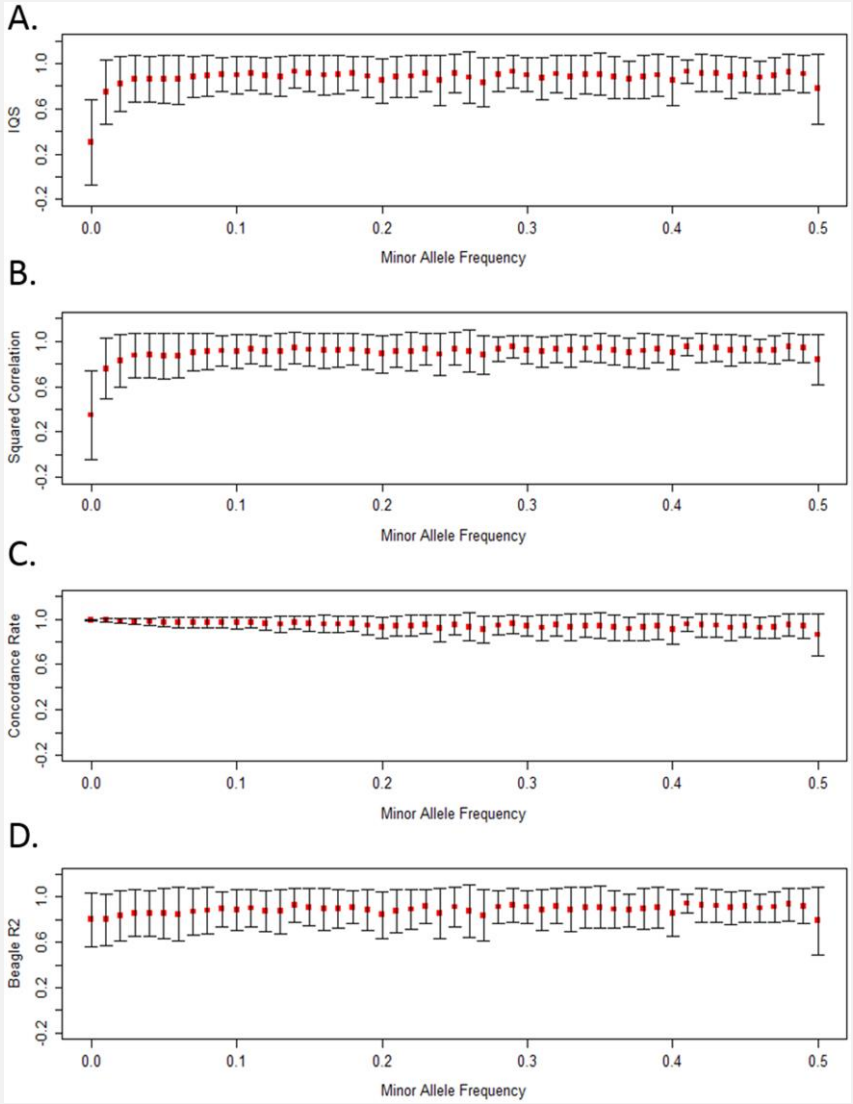
Dosage/Allelic R-square

BEAGLE R2 approximates the squared correlation between the most likely genotype and the true unobserved allele dosage. IMPUTE2/Minimac3 INFO considers allele frequency as well as the observed and expected allele dosage. Neither of these makes use of true genotypes.

Imputation Quality Score (IQS) (Lin *et al* (2010))

It adjusts the concordance between imputed and genotyped SNPs for chance, however is not widely used in accuracy assessment.

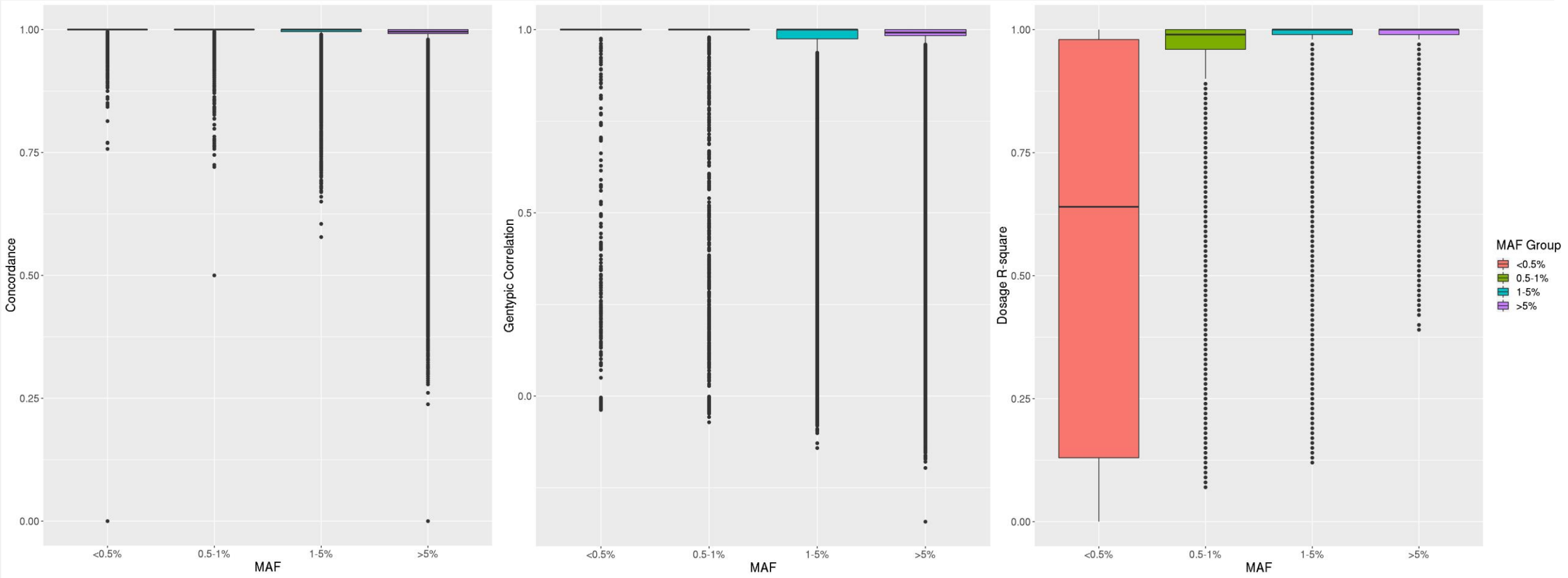
Why we don't recommend using error rate



Our results provide further evidence that concordance rate inflates accuracy estimates particularly for rare and low frequency variants. These observations highlight a need to account for chance agreement not only when assessing imputation accuracy, but also more broadly in other situations for which concordance is traditionally used to assess accuracy, such as checking genotype agreement across duplicate samples. Concordance rate will always produce a value greater than or equal to IQS due to their mathematical relationship (see Methods for proof).

Ramnarine, Shelina, *et al.* "When does choice of accuracy measure alter imputation accuracy assessments?." *PloS one* 10.10 (2015): e0137601.

Why we don't recommend using error rate



Factors affect imputation accuracy

- Number of ancestors genotyped in the reference (Hickey et al., 2011; Huang et al., 2012a)
- SNP density on the low and high panel (Mulder et al., 2012)
- MAF of the imputed SNP (van Binsbergen etal., 2014)
- Whether imputed SNP are located at the end of a chromosome or not (Badke et al., 2013; Cleveland and Hickey, 2013; Wellmann et al., 2013)
- The number of individuals with genotypes at the imputed density (Zhang and Druet, 2010)
- The relationship between imputed individuals and individuals genotyped at high density (Hickey et al., 2012)

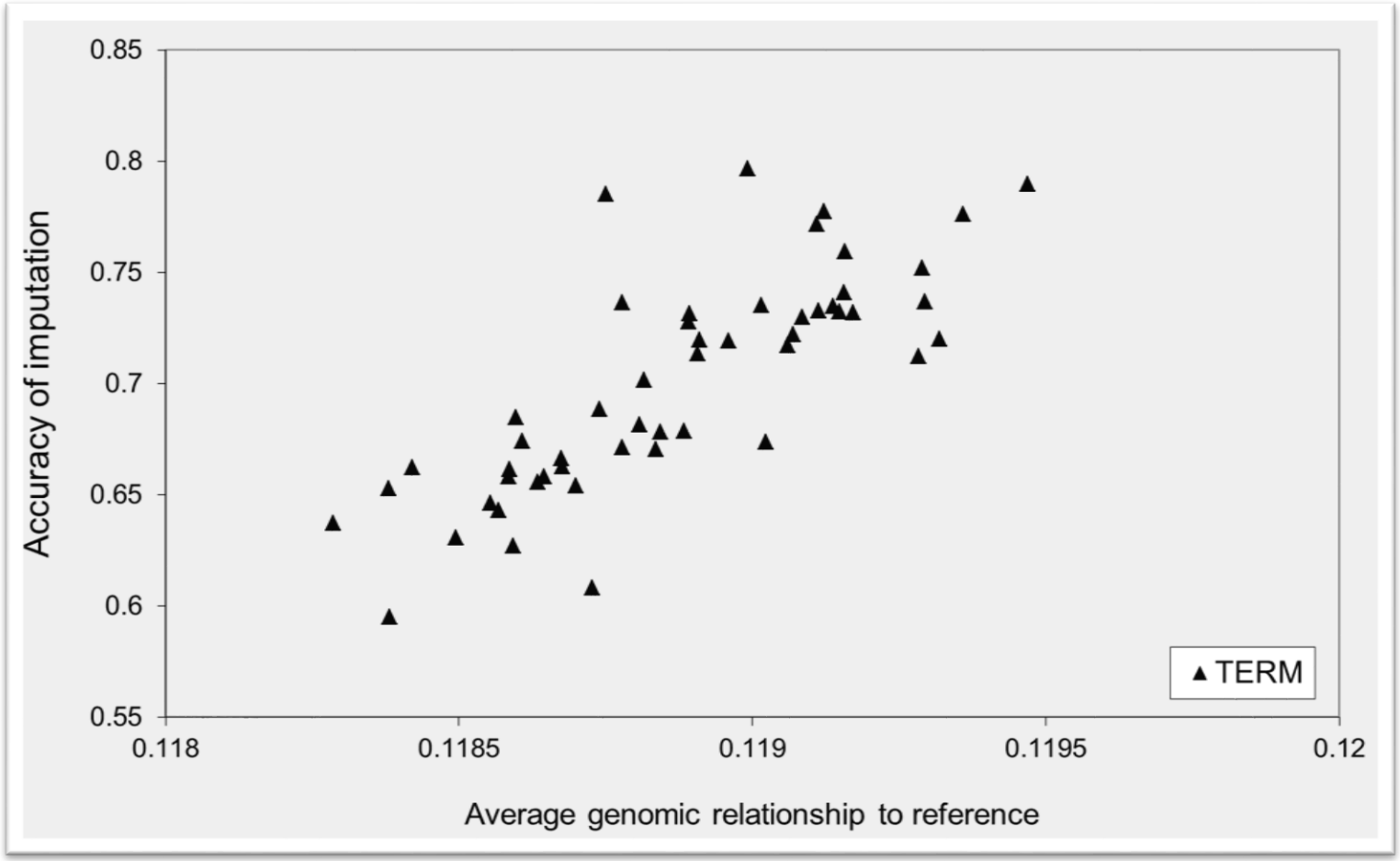
Table 6. Accuracy of imputation from BovineLD genotypes to BovineSNP50 genotypes for Australian, French, and North American breeds.

				Imputation accuracy	
Country/region[a]	**Breed**	**Reference**	**Target**	**Genotypes correctly imputed (%)[b]**	**Known genotypes without error (%)[c]**
Australia	Angus	200	82	92.3	93.1
	Holstein	1,831	360	97.5	97.8
	Jersey	454	86	94.9	95.7
France	Blonde d'Aquitaine	753	237	95.2	95.8
	Holstein	3,505	966	98.5	98.7
	Montbéliarde	1,170	222	98.1	98.4
	Normande	1,176	248	98.4	98.6
North America	Brown Swiss	1,994	168	97.4	97.9
	Holstein	63,288	19,506	98.8	98.9
	Jersey	8,687	1,140	98.0	98.3

[a]Beagle software (http://faculty.washington.edu/browning/beagle/beagle.html) was used for Australian and French imputations and findhap.f90 (http://aipl.arsusda.gov/software/findhap/) for North American imputations.
[b]The 6,909 SNPs on the BovineLD chip were excluded from the calculation of imputation accuracy.
[c]All SNPs included, i.e. the 6,909 SNPs on the BovineLD chip.
doi:10.1371/journal.pone.0034130.t006

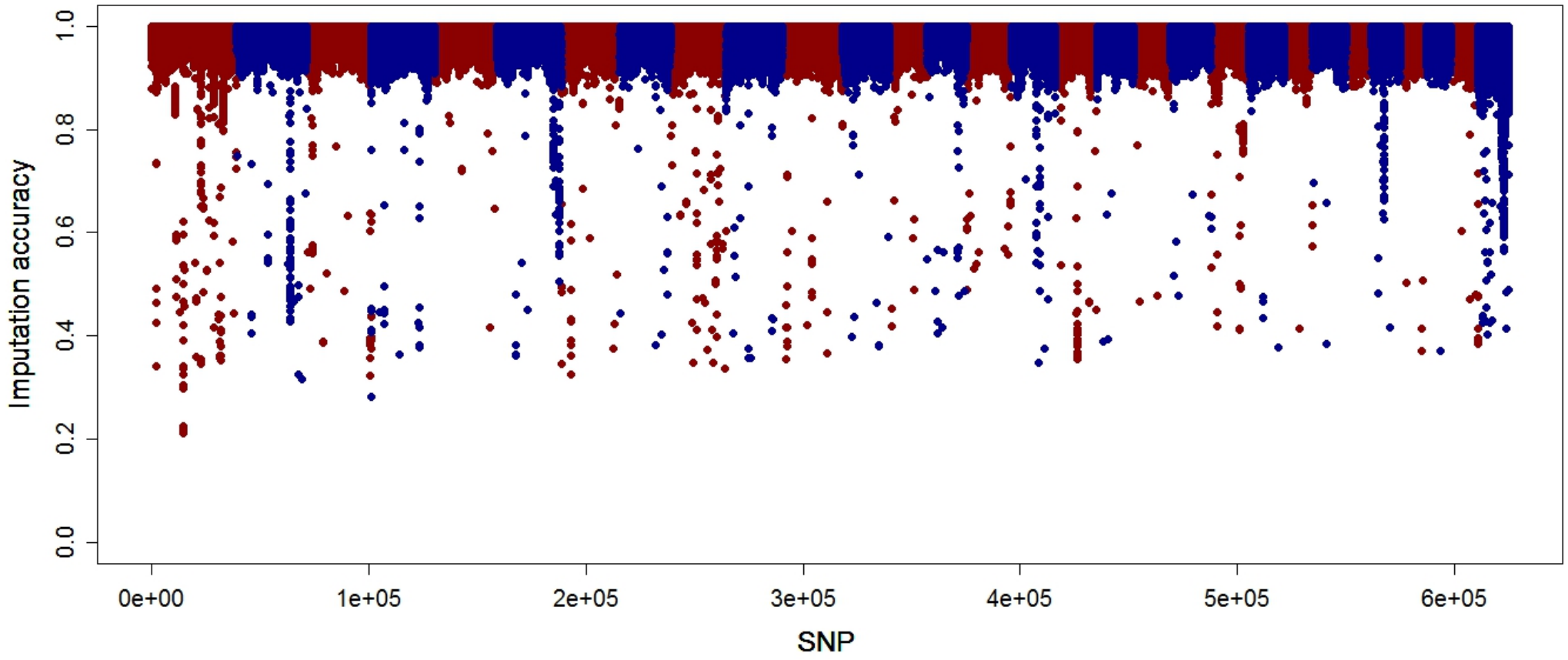
Imputation accuracy

- Relationship to reference?



Imputation accuracy

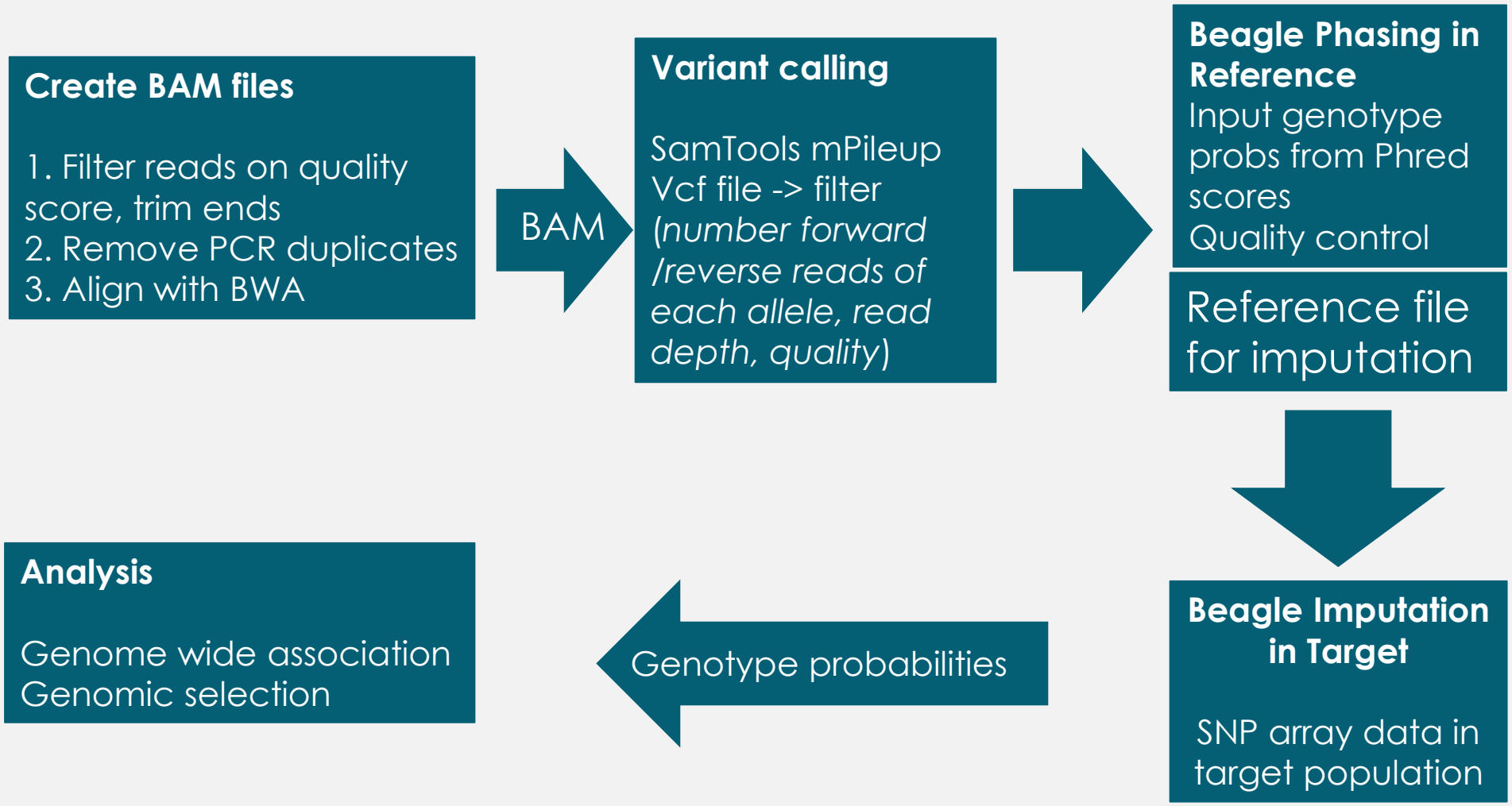
- Effect of map errors?



Imputation of full sequence data

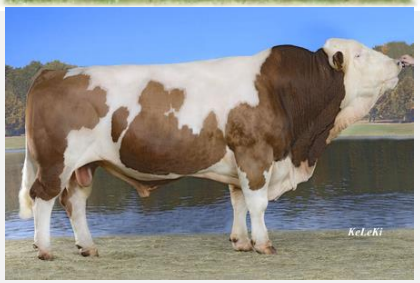
- Two groups of individuals
  - Sequenced individuals: reference population
  - Individuals genotyped on SNP array: target individuals
- Steps:
  - Step 1. Find polymorphisms in sequence data
  - Step 2. Phase genotypes (eg Beagle) in sequenced individuals, create reference file
  - Step 3. Genotype all study animals for polymorphisms (SNP, Indels)
  - Step 4. Impute all polymorphisms into individuals genotyped with SNP array

Imputation of full sequence data



Run4.0 1000 bull genomes Run 4.0

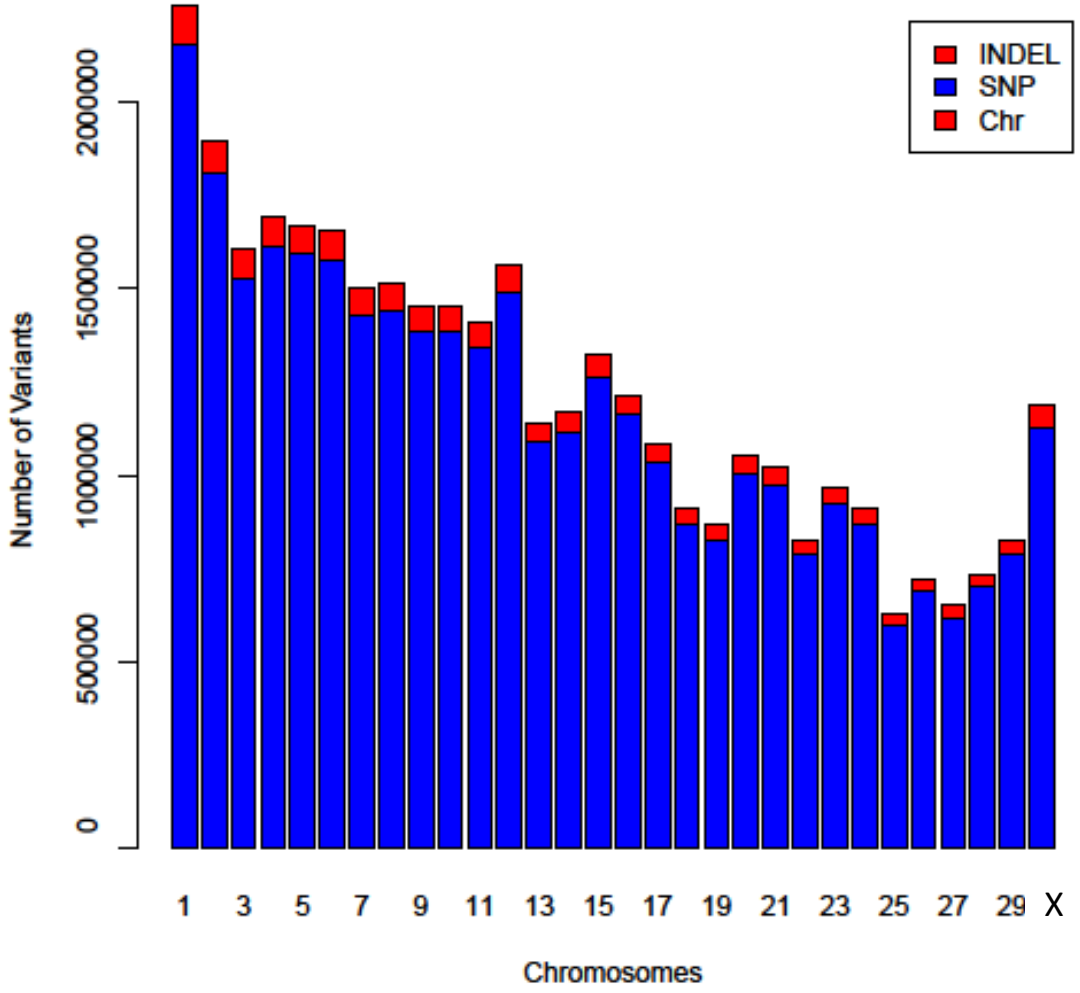
- 1147 animals sequenced
- 27 breeds
- 20 Partners
- Average 11X



Breed/Cross	Number
Holstein (Black and White)	**288**
Simmental (Dual and Beef)	216
Angus (Black and Red)	138
Jersey	61
Brown Swiss	59
Gelbvieh	34
Charolais	33
Hereford	31
Limousin	31
Guelph Composite	30
Beef Booster	29
Alberta Composite	28
Montbeliarde	28
AyrshireFinnish	25
Normande	24
Holstein (Red and White)	23
Swedish Red	16
Danish Red	15
Other Crosses	11
Belgian Blue	10
Piedmontese	5
Eringer	2
Galloway	2
Unknown	2
Scottish Highland	2
Pezzata Rossa Italiana	1
Romagnola	1
Salers	1
Tyrolean Grey	1
Total	**1147**

1000 bull genomes Run 4.0

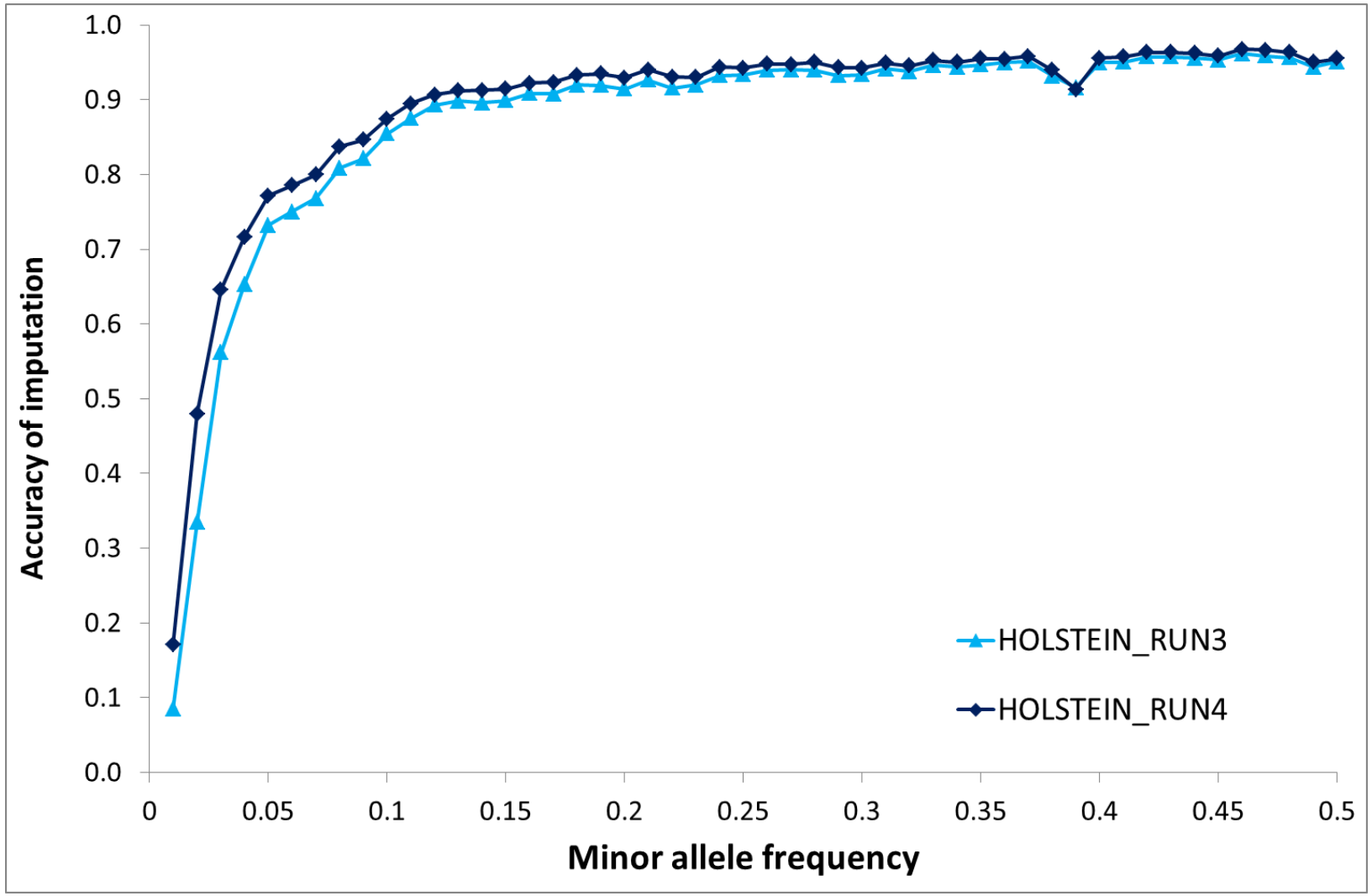
- 36.9 million filtered variants
- 35.2 million SNP
- 1.7 million INDEL



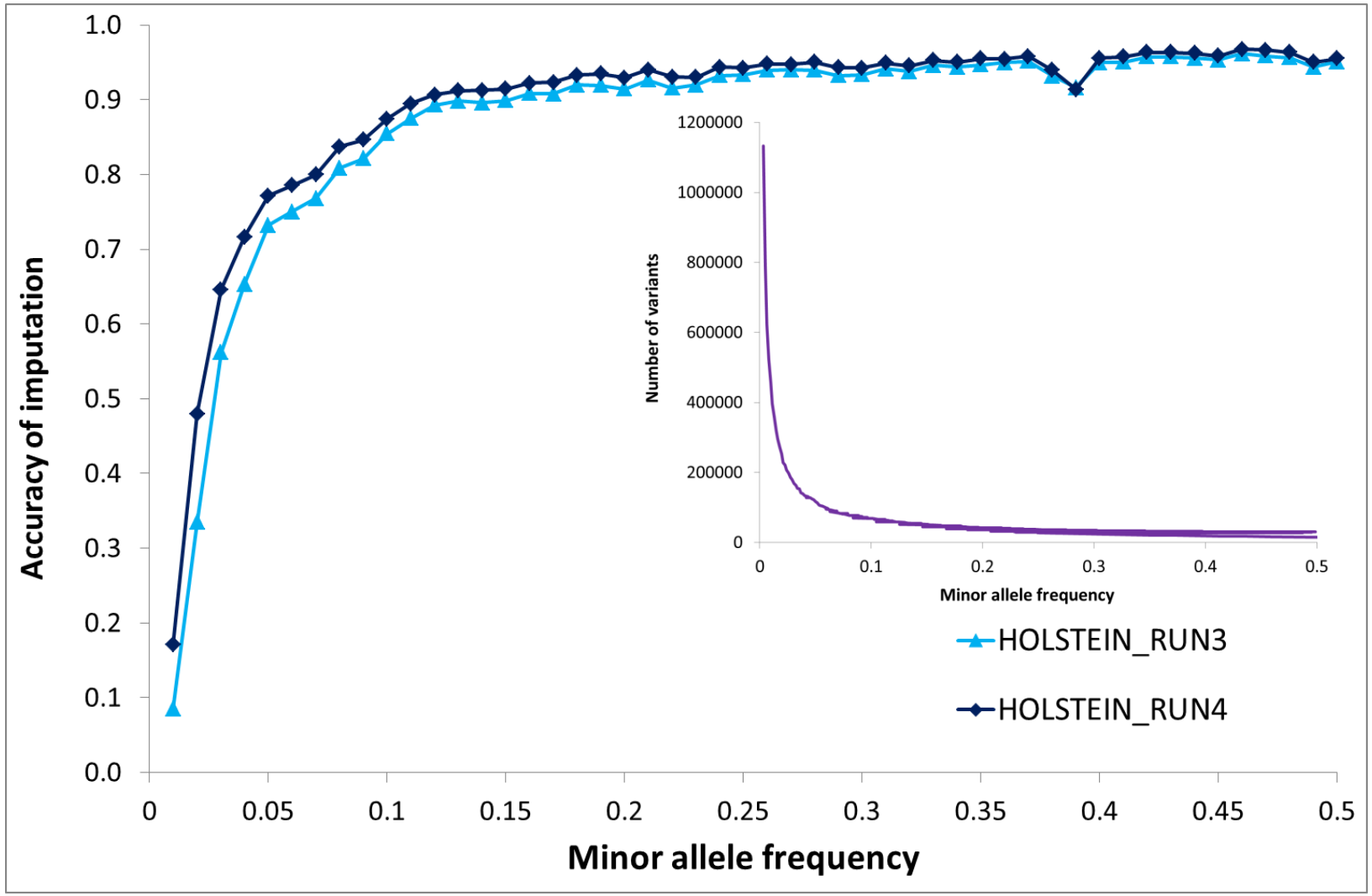
Imputation of full sequence data

- Accuracy?
  - Chromosome 14
  - Remove 50 Holsteins, 20 Jerseys from data set
  - Reduce genotypes to 800K for these animals
  - Impute full sequence using rest of animals as reference

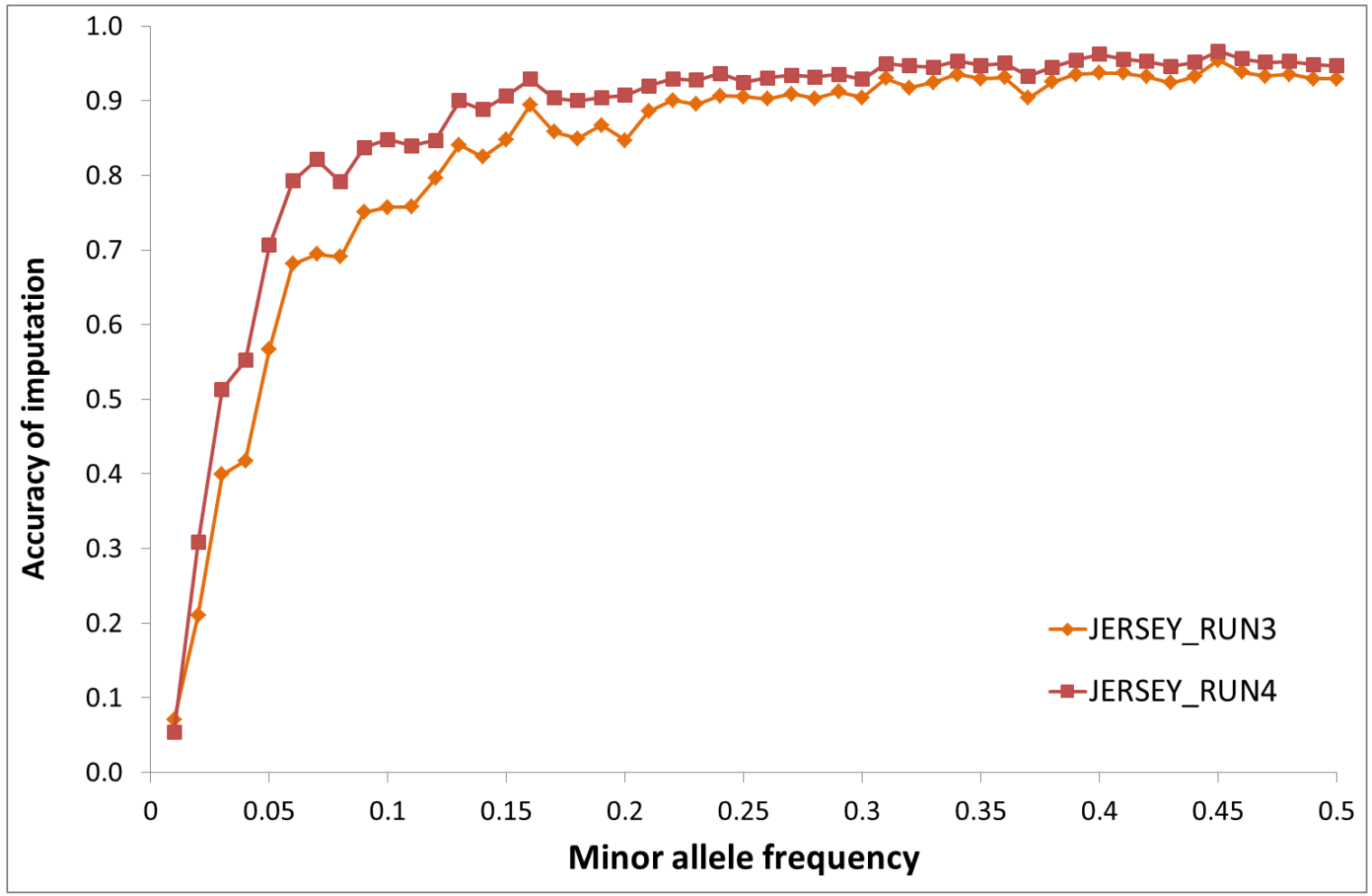
Imputation of full sequence data



Imputation of full sequence data



Imputation of full sequence data



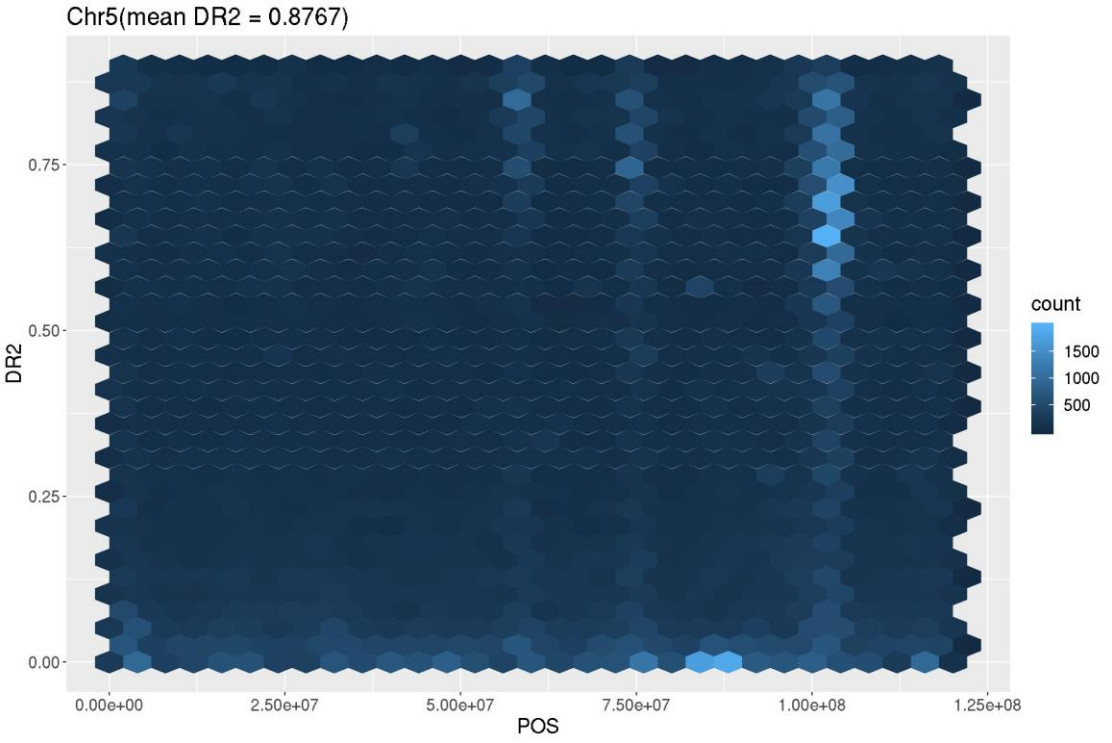
Imputation of full sequence data

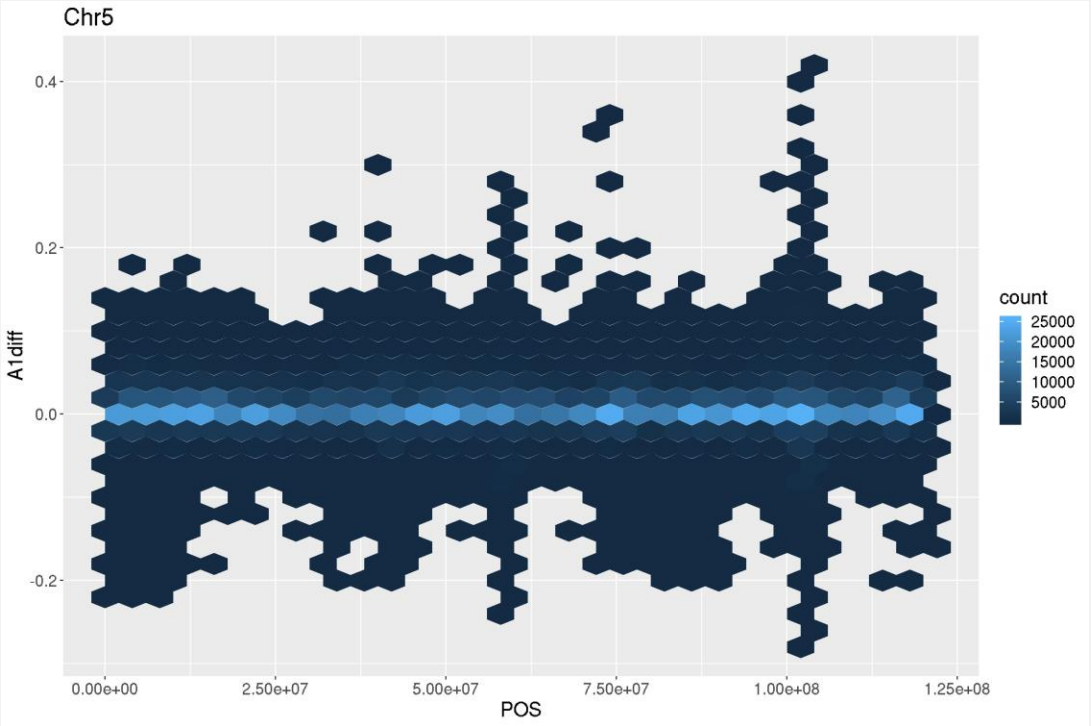
- Beagle – run imputation in chromosome segments, free, stable, new version (version 5.0 & 5.1) fast and accurate
  - Computational inefficient (large amount of memory required)
- Minimac3 – free, stable, computational efficient, slightly slower than Beagle
- FImpute – much faster than Beagle, computational efficient, be able to use pedigree information
  - Does not give probabilities, sometimes can have fatal error in the poorly assembled areas, input format is not friendly

Post imputation QC

- After imputation you need to check that it worked and the data look ok
- Things to check
  - Plot r2 across each chromosome look to see where it drops off
  - Plot MAF-reference MAF
- For each chromosome check N and % of SNPs:
  - MAF <.5%
  - With r2 0-0.3, 0.3-0.6,0.6-1
  - If you have hard calls or probs data HWE P < 10E-6
  - If you have families convert to hard calls and check for Mendelian errors (should be ~.2%)

Post imputation QC





Take Home Message

- Impute
  - to fill in missing genotypes
  - low density to high density to save $$
- Accuracy depends on size of reference, effective population size, relationship to reference, marker density
- Imputation to sequence possible, relatively low accuracies for rare alleles
- Use genotype probabilities from imputation in GWAS and genomic prediction

Practical for the afternoon

- Go through the imputation pipeline using 1000 Genome data (human)
- Understand the importance of quality control in imputation
- Have a look at how to use Beagle and Miminac3 for phasing and imputation

Choices of analysis methods