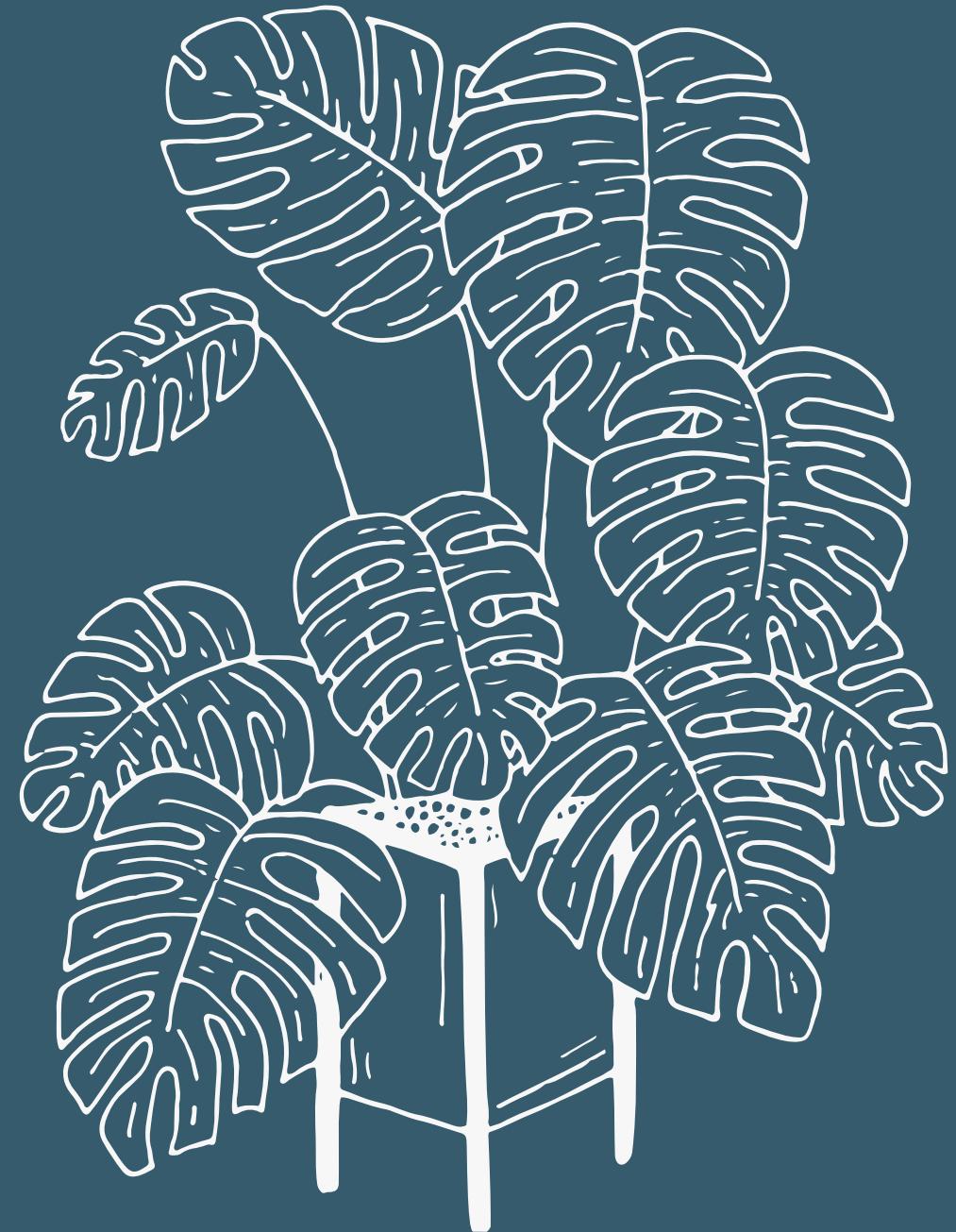




Imputation workshop

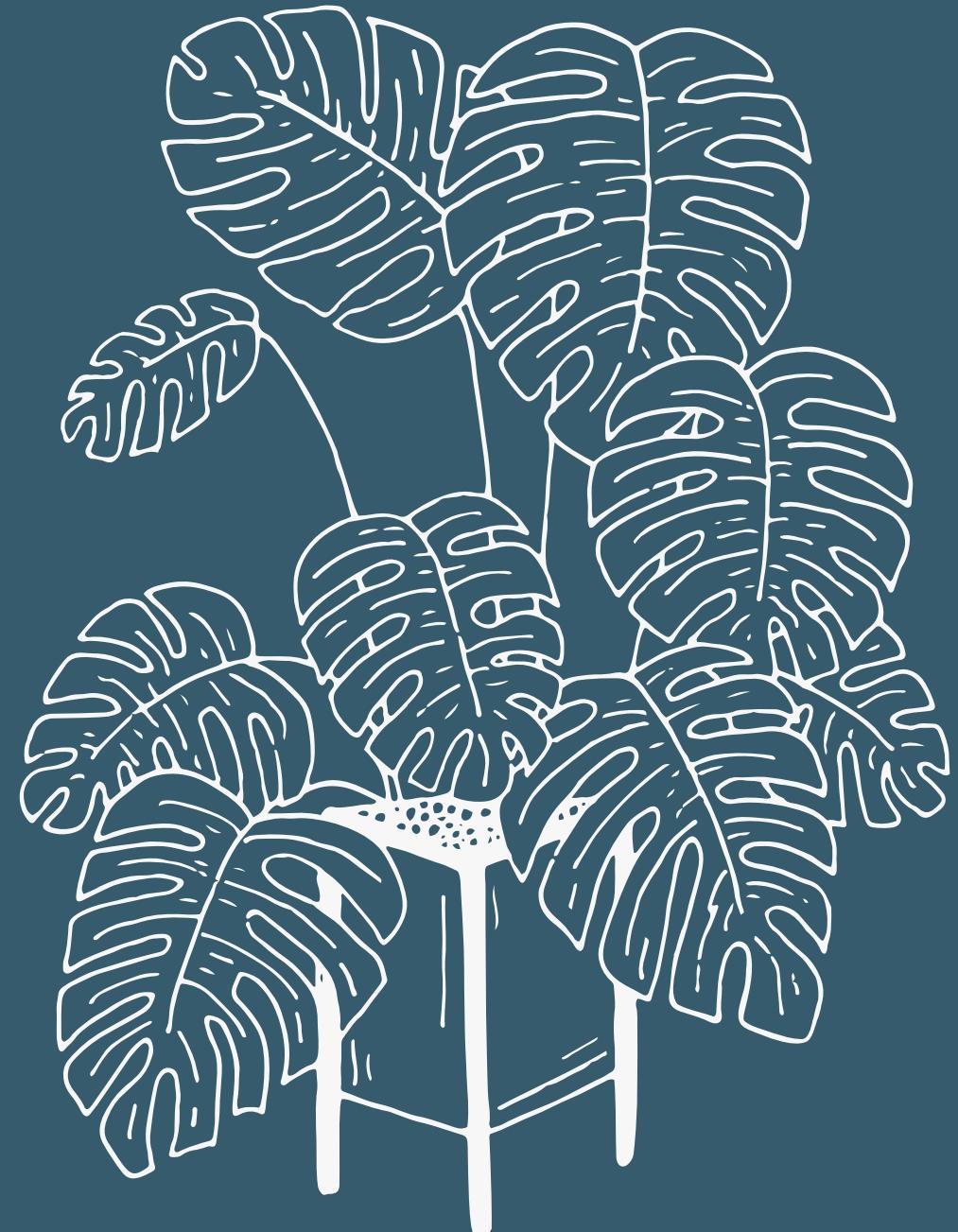
Yu Wang (Scientist in Quantitative Genetics)

yu.wang@lic.co.nz



Time table

- 9:00-9:15 Greeting and housekeeping info
- 9:15-10:15 Lecture—Yu
- 10:15-10:30 Q&A
- 10:30-10:45 Coffee break
- 10:45-11:45 Lecture—Martina
- 11:45-12:00 Q&A
- 12:00-13:00 Lunch
- 13:00-16:00 Tutorial



Content

- What is imputation
- Why we need imputation
- Why phasing is important
- Imputation accuracy
- Post-QC

Genotype Imputation from
Large Reference PanelsSayantan Das,¹ Gonçalo R. Abecasis,¹
and Brian L. Browning²¹Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, USA; email: sayantan@umich.edu, goncalo@umich.edu²Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, Washington 98195-7720, USA; email: browning@uw.eduAnnu. Rev. Genom. Hum. Genet. 2018.19:73-96. Downloaded from www.annualreviews.org
Access provided by 124.248.141.150 on 07/04/23. For personal use only.

Annu. Rev. Genom. Hum. Genet. 2018. 19:73–96

First published as a Review in Advance on
May 23, 2018The *Annual Review of Genomics and Human Genetics*
is online at genom.annualreviews.org<https://doi.org/10.1146/annurev-genom-083117-021602>Copyright © 2018 by Annual Reviews.
All rights reserved

ANNUAL REVIEWS CONNECT

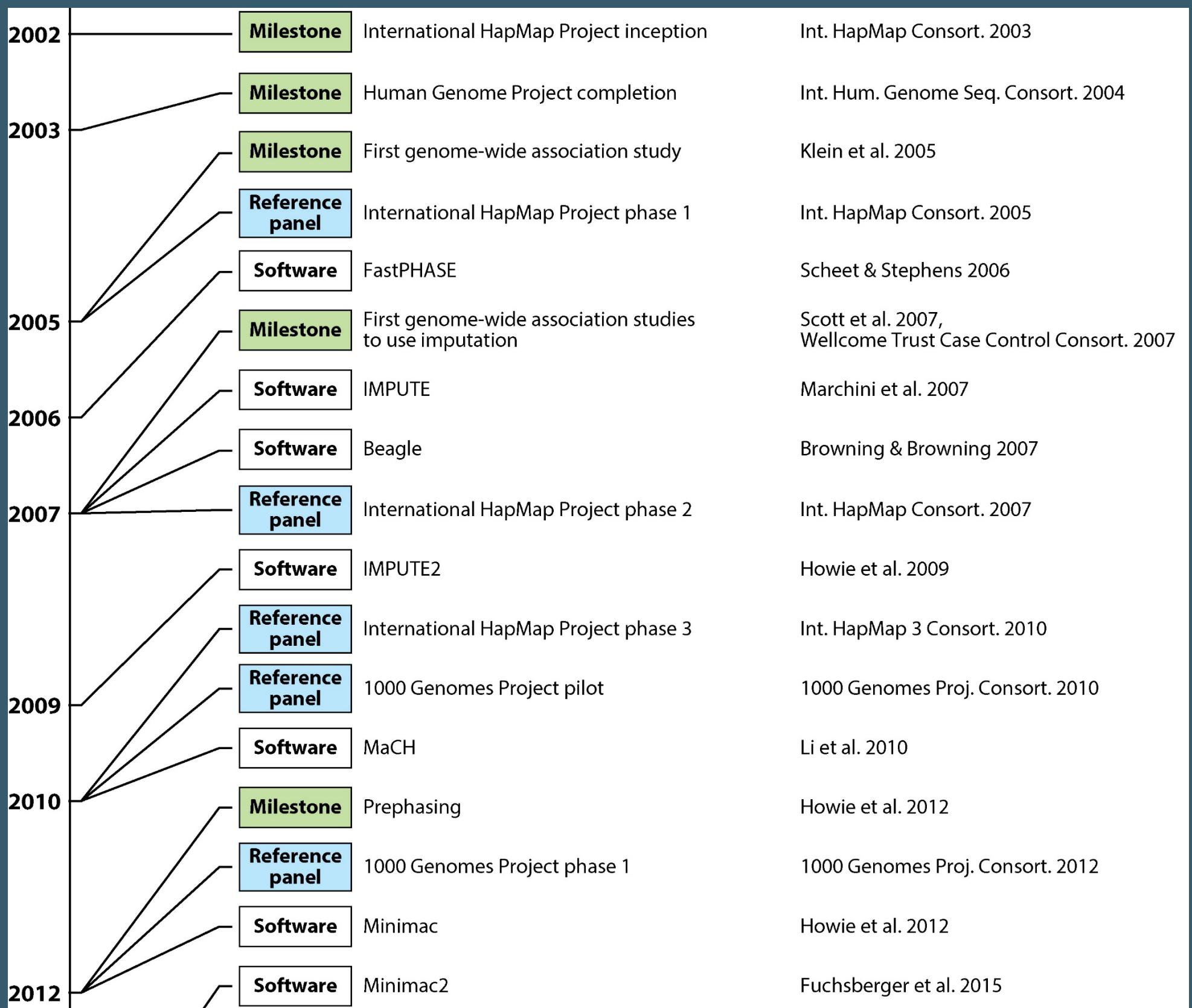
- www.annualreviews.org
- Download figures
- Navigate cited references
- Keyword search
- Explore related articles
- Share via email or social media

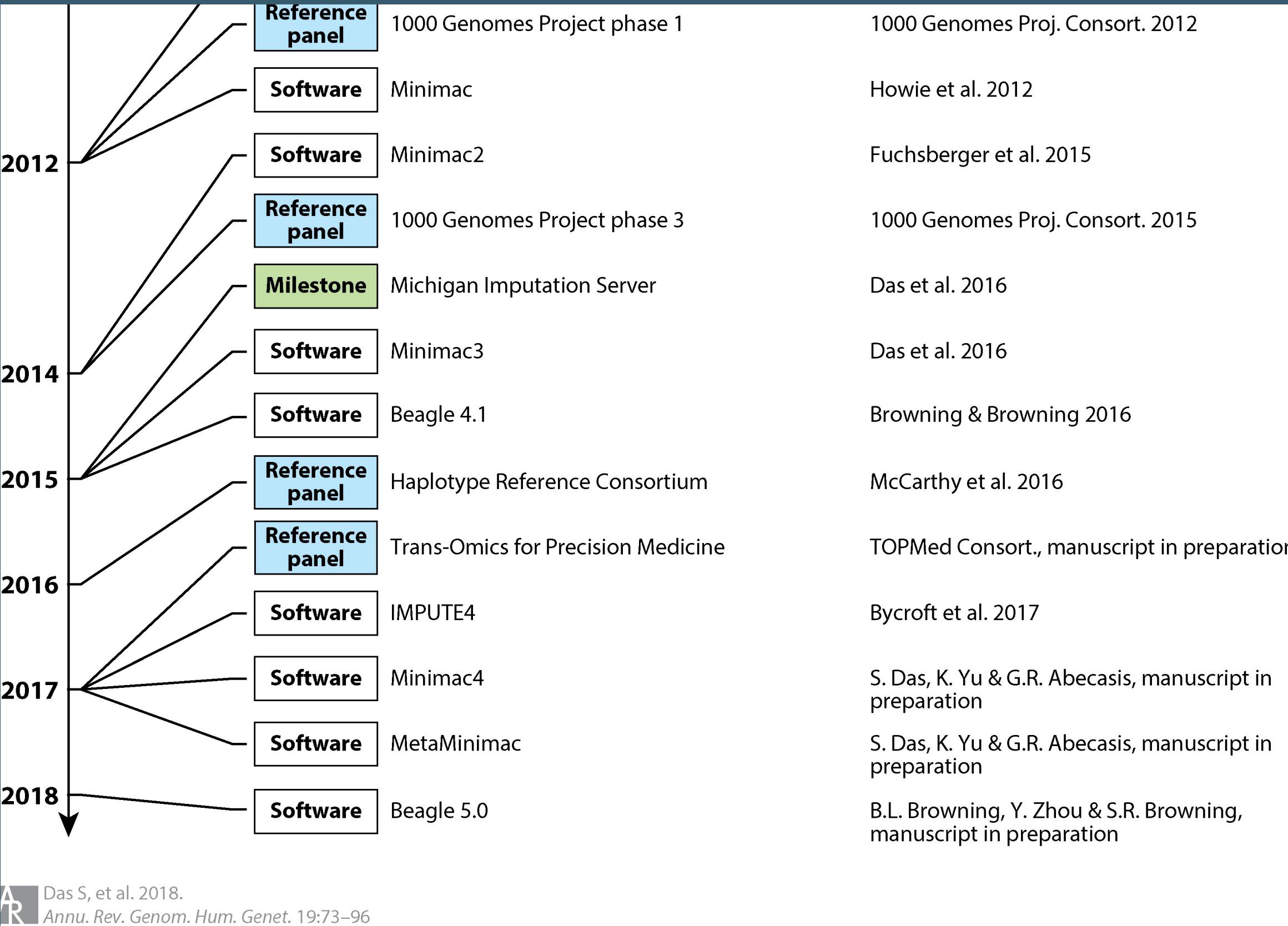
Keywords

imputation, genotype imputation, genome-wide association study, GWAS

Abstract

Genotype imputation has become a standard tool in genome-wide association studies because it enables researchers to inexpensively approximate whole-genome sequence data from genome-wide single-nucleotide polymorphism array data. Genotype imputation increases statistical power, facilitates fine mapping of causal variants, and plays a key role in meta-analyses of genome-wide association studies. Only variants that were previously observed in a reference panel of sequenced individuals can be imputed. However, the rapid increase in the number of deeply sequenced individuals will soon make it possible to assemble enormous reference panels that greatly increase the number of imputable variants. In this review, we present an overview of genotype imputation and describe the computational techniques that make it possible to impute genotypes from reference panels with millions of individuals.





Das S, et al. 2018.
Annu. Rev. Genom. Hum. Genet. 19:73–96



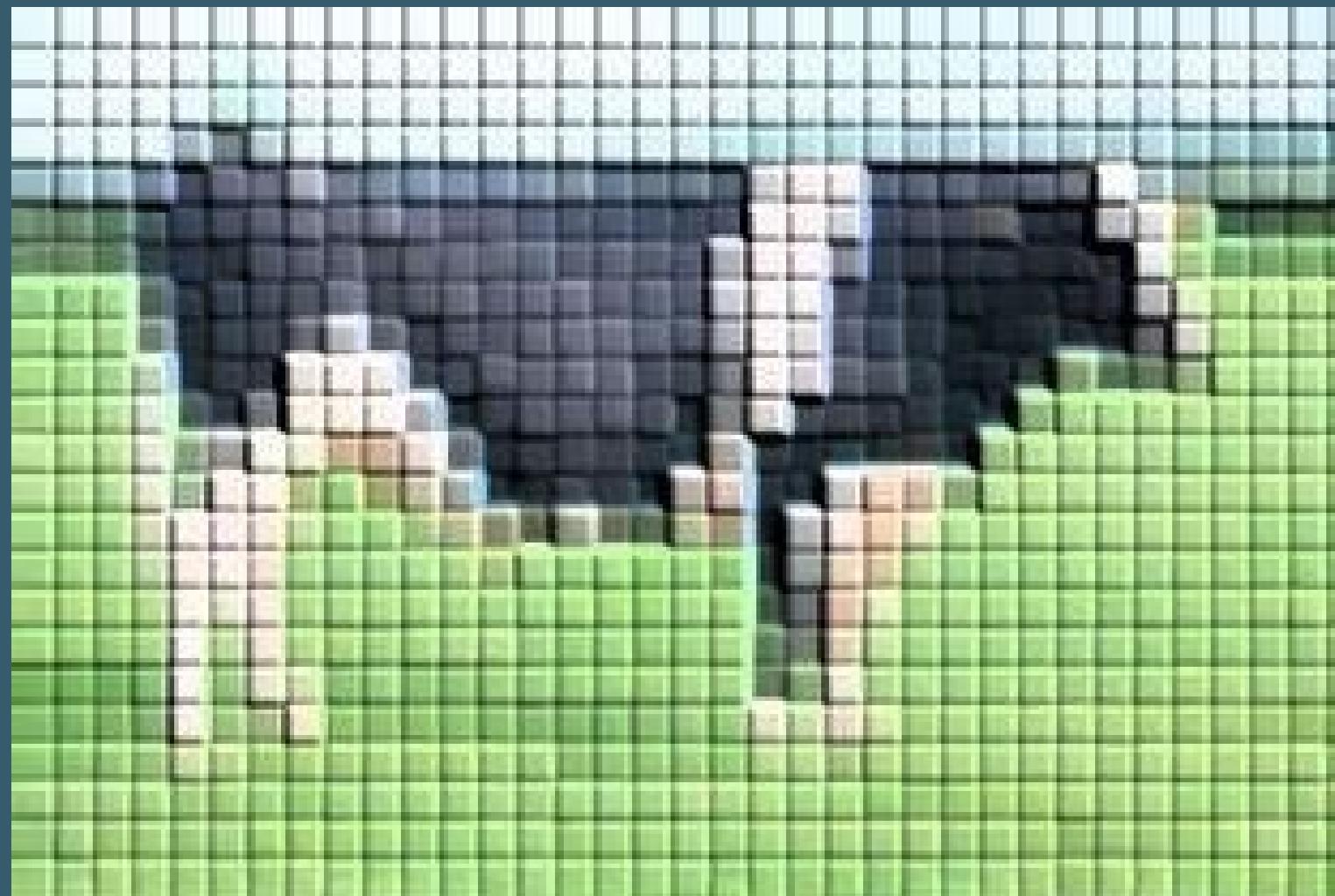
What is imputation



Imputation describes the process of predicting genotypes that have not been directly typed in a sample of individuals:

- missing genotypes at typed variants (**study**)
- genotypes at un-typed variants that are present in an external high-density "reference panel" of **phased haplotypes**

[https://en.wikipedia.org/wiki/Imputation_\(genetics\)](https://en.wikipedia.org/wiki/Imputation_(genetics))

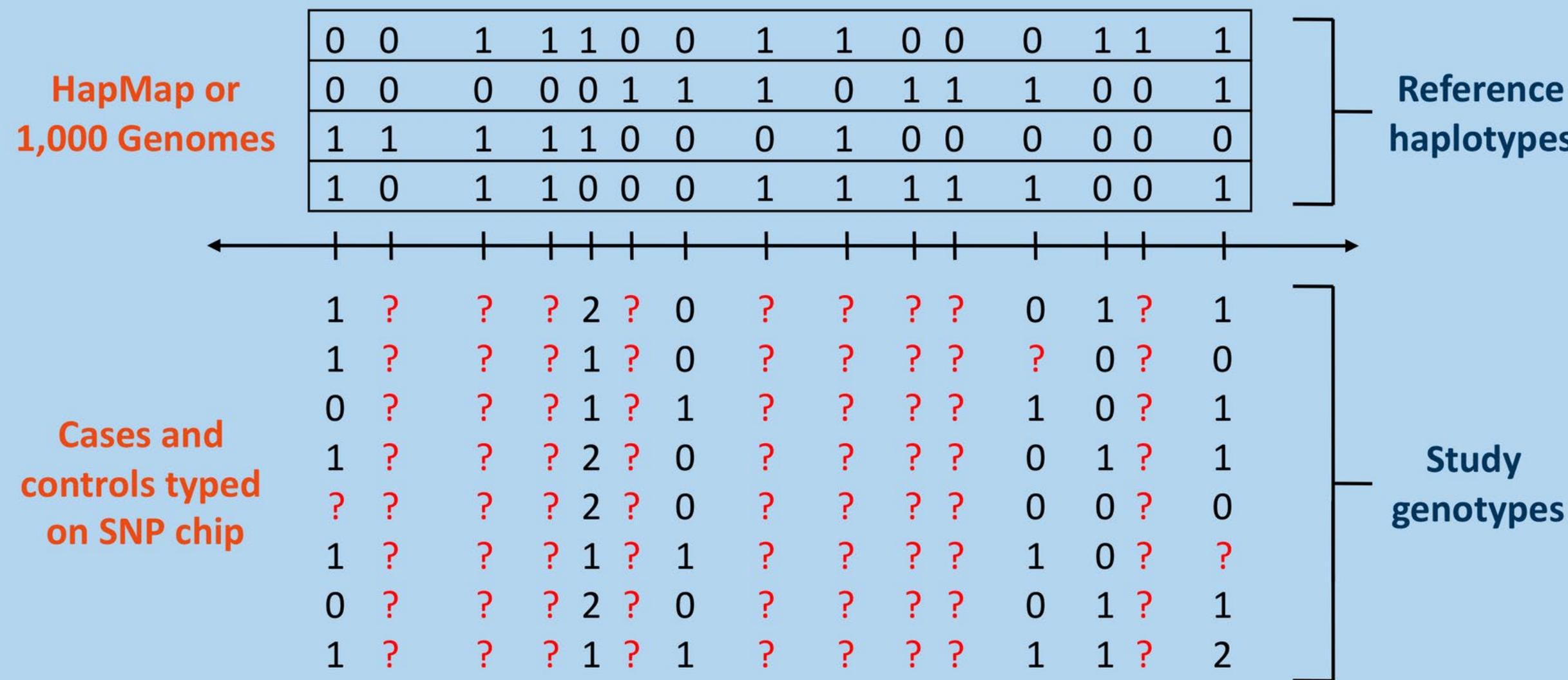


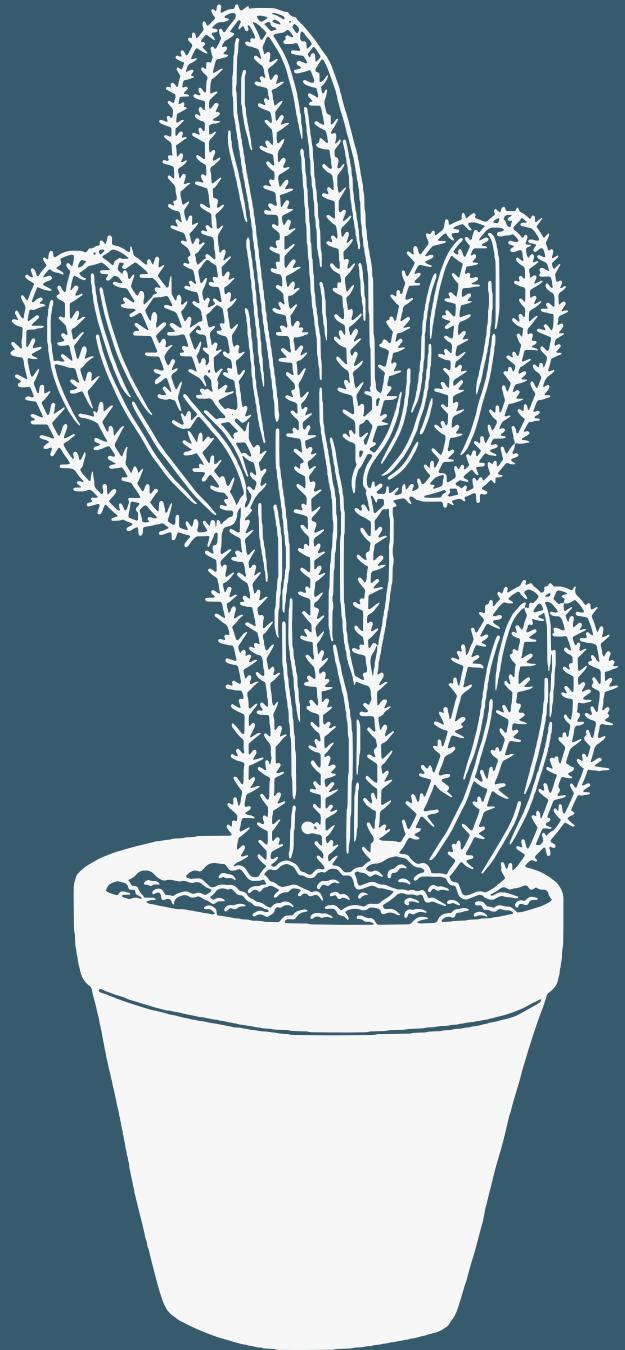
low density panel
Study population



high density panel
Reference population

Typical imputation scenario





What is phasing



What is Phasing?

Medical Monitoring
Devices

Why we need
to do
imputation



- **Cheap**

Directly sequence a large population is still quite expensive

- **Increase power**

The reference panel is more likely to contain the causal variant (or a better tag) than a lower density panel

- **Fine-mapping**

Imputation provides a high-resolution overview of an association signal across a locus

- **Meta-analysis**

Imputation allows different genotype panels to be combined up to the variants in the reference panel



Imputation Server

- Michigan Imputation Server (human)

<https://imputationserver.sph.umich.edu>

- Sanger Imputation Service (human)

<https://imputation.sanger.ac.uk/>

- TOPMed Imputation Server (human) = MIS but different reference

<https://imputation.biodatacatalyst.nhlbi.nih.gov>

- SWineIMputation (SWIM) server (pig)

<https://quantgenet.msu.edu/swim/>

- Animal-ImputeDB (animals)

http://gong_lab.hzau.edu.cn/Animal_ImputeDB#/

The screenshot displays the Animal-ImputeDB interface, which is a database for animal genotype imputation. The top navigation bar includes links for Home, Imputation, Reference Panel, Download, Help, and Contact. A red 'IMPUTATION' button is prominently displayed.

Panel A: The header features the 'Animal-ImputeDB' logo and a subtext 'A database for animal genotype imputation'. To the right is a vertical 'IMPUTATION' menu with options A, C, ?, and G.

Panel B: Titled 'Browse by species', this section shows a grid of 15 images of various animals, each with its scientific name and common name. The animals include a giant panda, duck, cattle, swamp buffalo, dog, goat, horse, tarpan, chicken, monkey, rabbit, sheep, and pig.

Panel C: This panel is titled 'Species information' and contains tabs for Online imputation, SNP search, and Samples information. It includes a search bar for 'Search SNPs of cattle' and fields for Region (Chr1:192-340), SNP ID (bta10), dbSNP ID (rs42801761), and MAF (>0.05). A 'Submit' button is present.

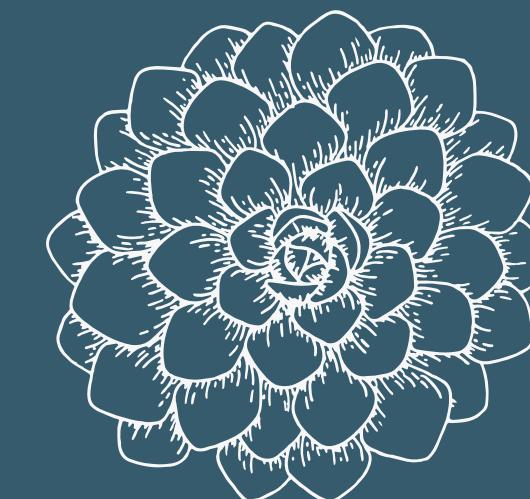
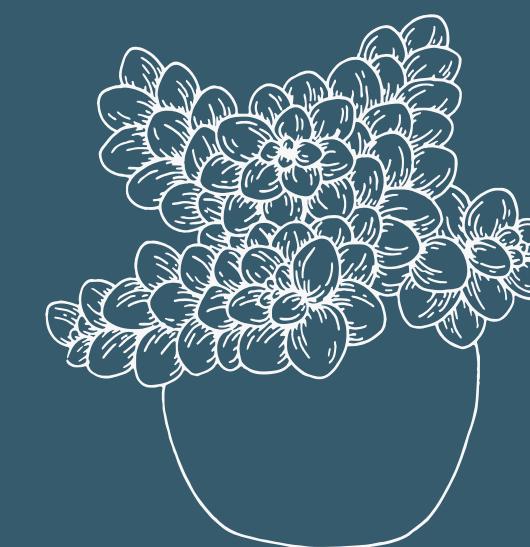
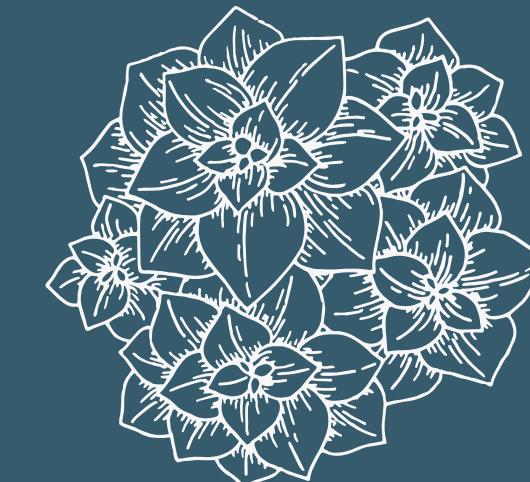
Panel D: This panel displays a table of SNP information for cattle. The columns are SNP ID, Chromosome, Position, Ref. allele, Alt. allele, Minor allele frequency, and dbSNP ID. The table shows five rows of data:

SNP ID	Chromosome	Position	Ref. allele	Alt. allele	Minor allele frequency	dbSNP
bta3	1	238	C	T	0.36514	rs42801762
bta4	1	300	A	G	0.44973	rs42801761
bta5	1	324	A	G	0.19344	rs459642982
bta6	1	340	G	A	0.38158	rs381103691

A 'Download' button is located at the bottom right of the table.

Table 1 Genotype imputation tools that employ a hidden Markov model (HMM)

Tool	Year	Description of state space	Computational complexity	HMM parameter functions
FastPHASE	2006	All genotype configurations from a fixed number of localized haplotype clusters	Maximization-step linear in number of haplotypes, quadratic in number of clusters	Depends on recombination and mutation rates; parameters are fit using an expectation–maximization algorithm
IMPUTE	2007	All genotype configurations from all reference haplotypes	Quadratic in number of haplotypes	Depends on a fine-scale recombination map that is fixed and provided internally by the program
Beagle	2007	All genotype configurations from a variable number of localized haplotype clusters	Quadratic in number of haplotypes	Empirical model with no explicit parameter functions
IMPUTE2	2009	All reference haplotypes	Phasing quadratic in number of haplotypes, imputation linear in number of haplotypes	Same as IMPUTE
MaCH	2010	All genotype configurations from all reference haplotypes	Quadratic in number of haplotypes	Depends on recombination rate, mutation rate, and genotyping error; parameters are fit using a Markov chain Monte Carlo or expectation–maximization algorithm
Minimac and Minimac2	2012	All reference haplotypes	Linear in number of haplotypes	Same as MaCH
Minimac3	2016	All unique allele sequences observed in reference data in a small genomic segment	Linear in number of haplotypes	Same as MaCH, but parameter estimates are precalculated and fixed
Beagle 4.1	2016	All reference haplotypes at genotyped markers	Linear in number of haplotypes	Depends on recombination rates and error rates, which are precalculated and fixed
Minimac4	2017	Collapsed allele sequences from reference data that match at genotyped positions in small genomic segments	Linear in number of haplotypes	Same as Minimac3
IMPUTE4 ^a	2017	All possible reference haplotypes	Linear in number of haplotypes	Same as IMPUTE2
Beagle 5.0	2018	A user-specified number of reference haplotypes	Linear in number of haplotypes	Same as Beagle 4.1

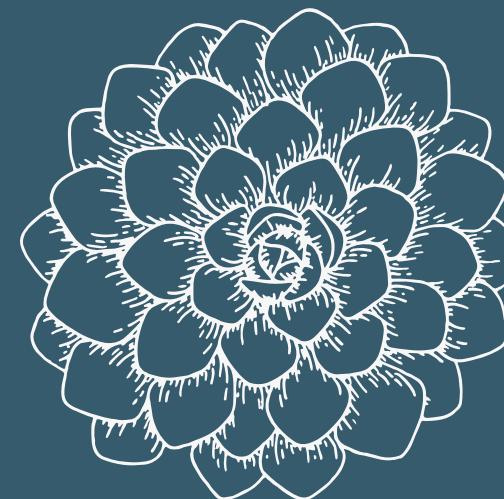
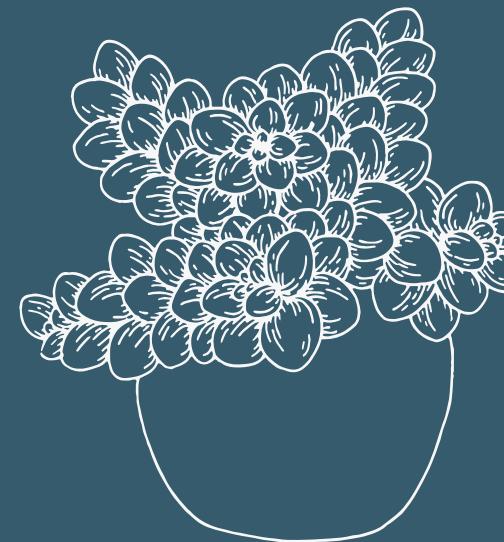
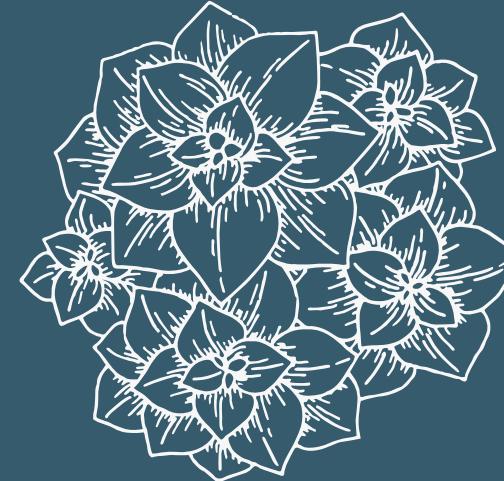


Imputation software

Pros & Cons

Pros: flexible about reference, majority are free, easy to get imputation accuracy, more flexible in all aspects

Cons: might be slow, may need large memory, software may contain bugs, license problem, performance might vary based on dataset





Do you have
questions so far?



How do I know if the
imputation was done well



Parameters to evaluate imputation accuracy



Genotype concordance

Genotype concordance is computed per locus as the percentage (or proportion) of alleles or genotypes that is imputed correctly. 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.**



Imputation Quality Score (IQS) (*Liu et al. (2010)*)

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

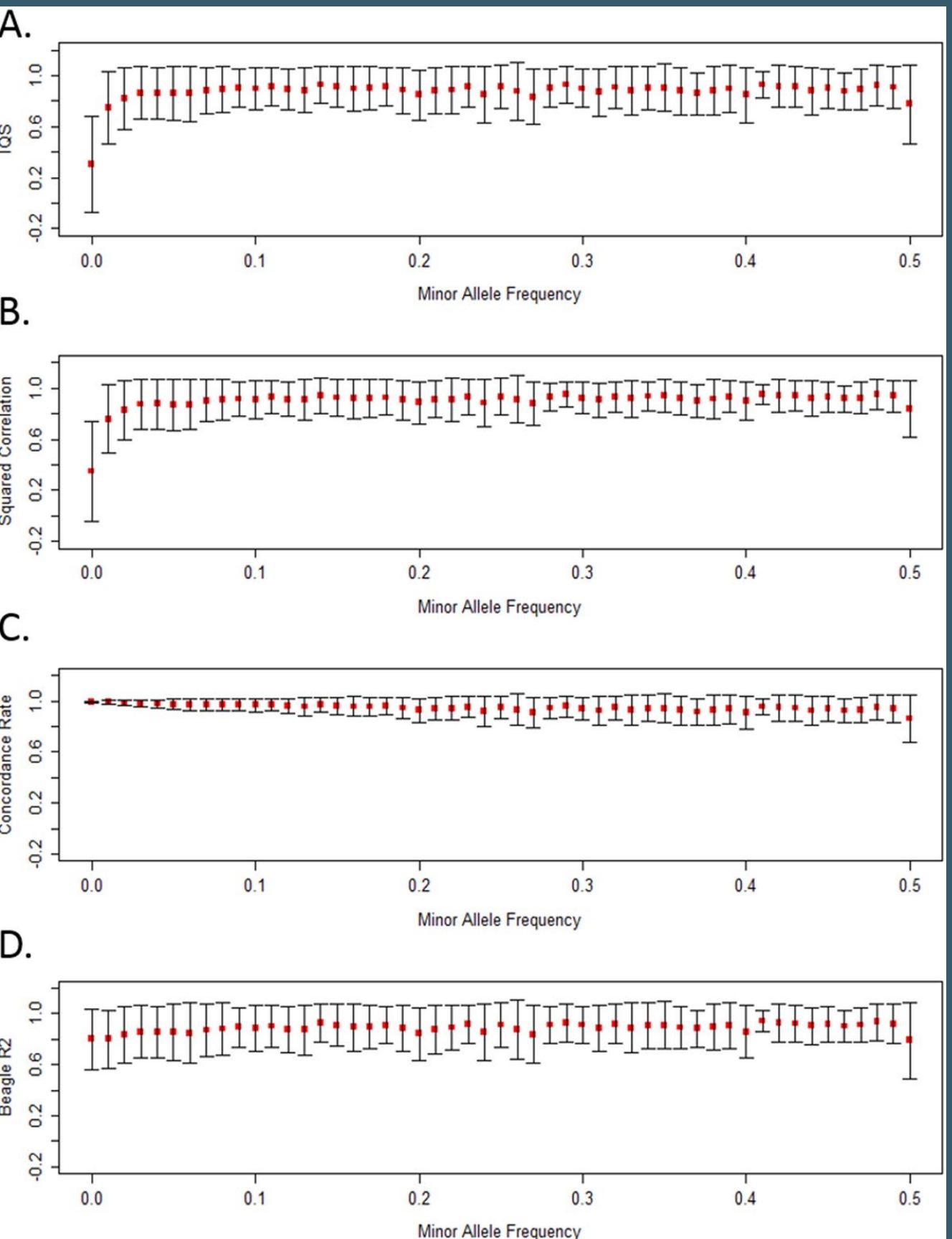


Dosage/ Allelic R-square

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 need true genotypes.

Why error rate is not recommended

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).



R-squared

- DosageR-squared and Allelic R-squared

The allelic R² from BEAGLE is the squared correlation between the best-guess genotype and the allele dosage. The information metric from IMPUTE2 measures the relative statistical information about the population allele frequency. The minimac Rsq is the ratio of the observed variance of the allele dosage to the expected binomial variance at HWE. Despite differences in the calculation of these imputation quality measures (reviewed by Marchini and Howie [2]), they all range between 0 and 1, with larger values corresponding to higher imputation quality.

Table 3: The correlation between R^2 and the imputation quality measure

Software	Reference	Correlation ^a	P (NA) ^b (%)
BEAGLE	ALL	0.776	18.546
	EUR	0.751	16.825
IMPUTE2	ALL	0.832	0.364
	EUR	0.806	0.217
minimac	ALL	0.827	0.022
	EUR	0.827	0.028

^aThe Spearman correlation coefficient between R^2 and the imputation quality measure.

^bThe percentage of imputed variants with NA R^2 values or invalid value of imputation quality measure. These imputed variants were removed in calculation of the Spearman correlation coefficient.

Liu, Qian, et al. "Systematic assessment of imputation performance using the 1000 Genomes reference panels." *Briefings in bioinformatics* 16.4 (2015): 545-562.

R-squared

- DosageR-squared and Allelic R-squared

We noticed that in this setting, the quality measure of BEAGLEs comparable with (or even slightly better than) the quality measures from the other two programs in removing poorly imputed variants from their corresponding imputation results.

...

Therefore, to achieve the highest accuracy, we recommend setting the imputation quality cutoff to small values for BEAGLE (e.g. <0.4), intermediate values for minimac (e.g. between 0.2 and 0.6) and large values for IMPUTE2 (e.g. between 0.6 and 0.9).

Table 3: The correlation between R^2 and the imputation quality measure

Software	Reference	Correlation ^a	P (NA) ^b (%)
BEAGLE	ALL	0.776	18.546
	EUR	0.751	16.825
IMPUTE2	ALL	0.832	0.364
	EUR	0.806	0.217
minimac	ALL	0.827	0.022
	EUR	0.827	0.028

^aThe Spearman correlation coefficient between R^2 and the imputation quality measure.

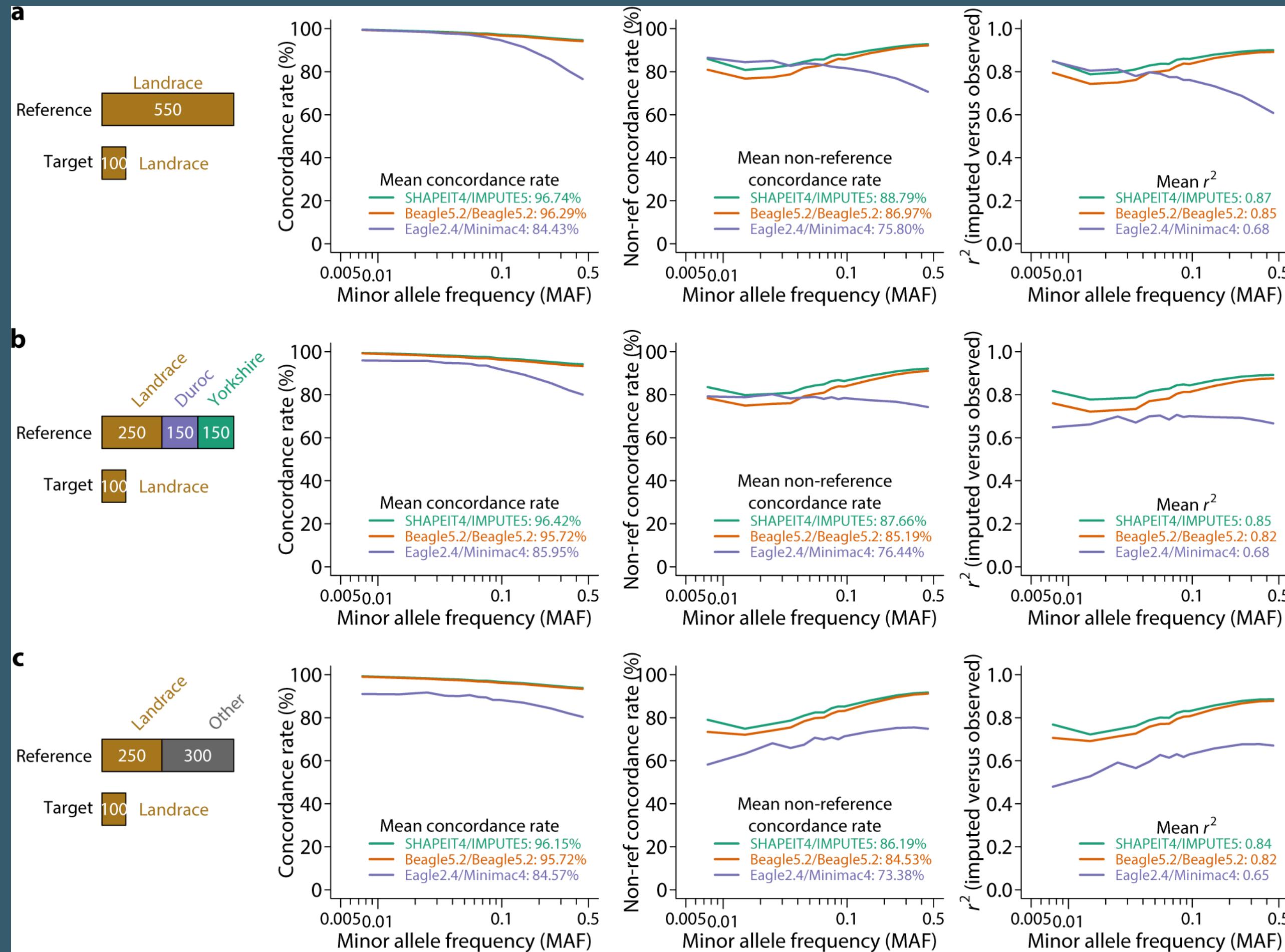
^bThe percentage of imputed variants with NA R^2 values or invalid value of imputation quality measure. These imputed variants were removed in calculation of the Spearman correlation coefficient.

Liu, Qian, et al. "Systematic assessment of imputation performance using the 1000 Genomes reference panels." *Briefings in bioinformatics* 16.4 (2015): 545-562.



Factors influence imputation accuracy

- Number of ancestors genotyped in the reference (Hickey et al., 2011 Huanget al., 2012a)
- SNPdensity on the low and high panel (Mulderet al., 2012)
- MAFof the imputed SNP(vanBinsbergenetal., 2014)
- Whether imputed SNP are located at the end of a chromosome or not (Badke et al., 2013 Cleveland and Hickey, 2013 Wellmannet al., 2013)
- The number of individuals in the reference population (Zhang and Druet, 2010)
- The relationship between imputed individuals and individuals genotyped at high density (Hickeyet al., 2012)



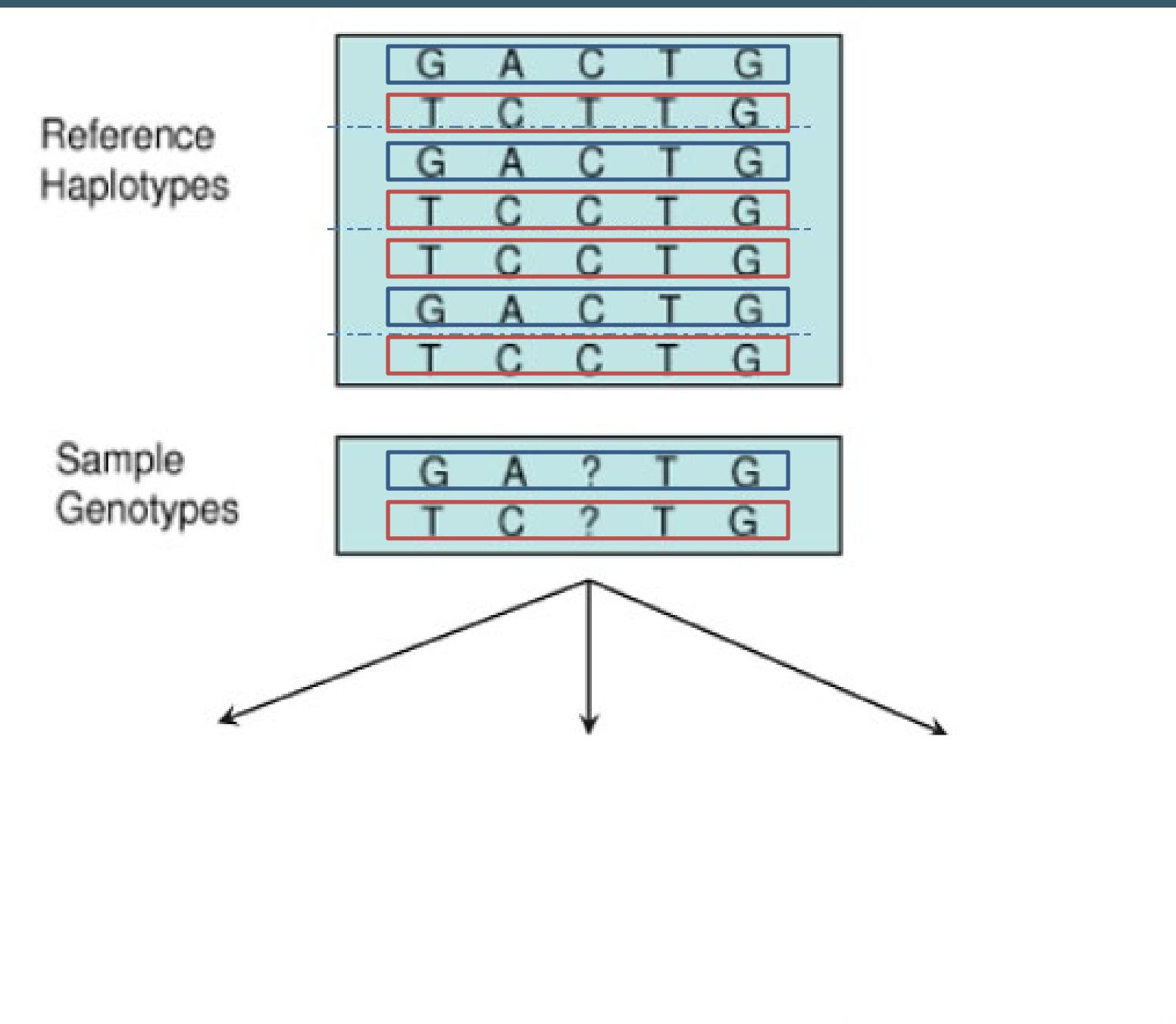
VCF file

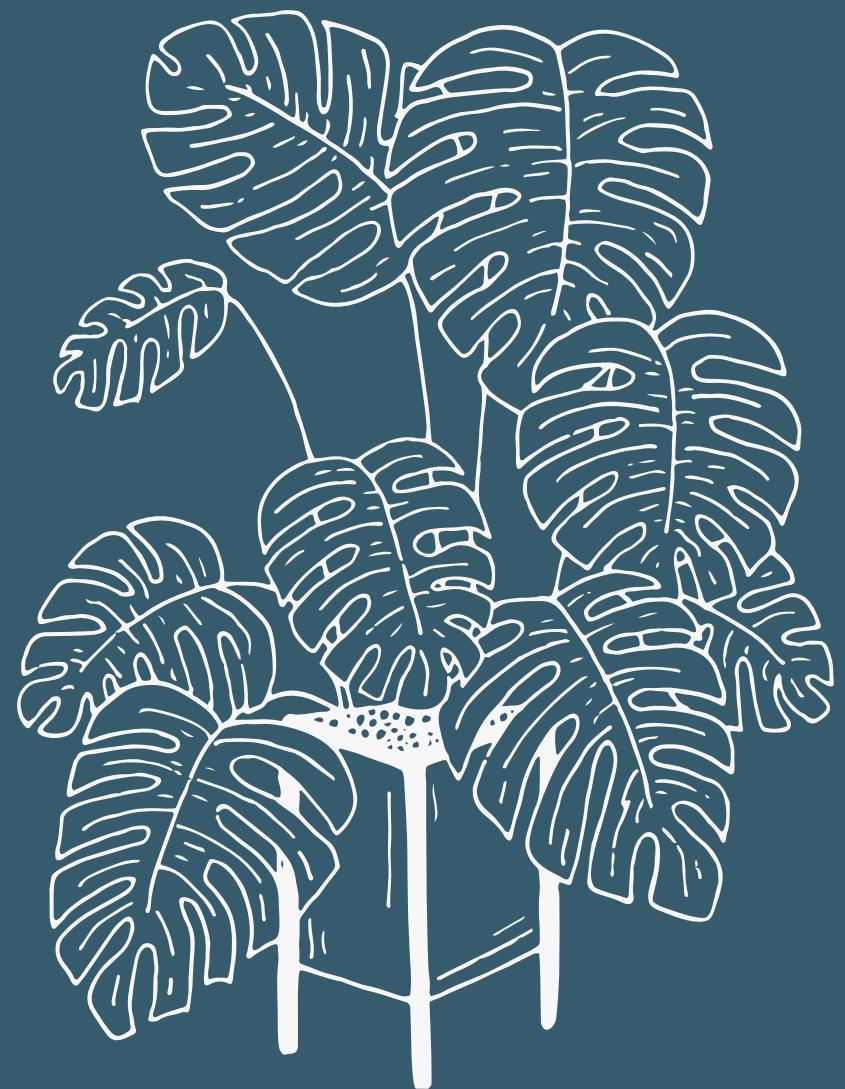
```
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="All filters passed">
##filedate=2018.7.25
##source=Minimac3
##contig=<ID=1>
##FILTER=<ID=GENOTYPED,Description="Marker was genotyped AND imputed">
##FILTER=<ID=GENOTYPED_ONLY,Description="Marker was genotyped but NOT imputed">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Estimated Alternate Allele Dosage : [P(0/1)+2*P(1/1)]">
##FORMAT=<ID=GP,Number=3,Type=Float,Description="Estimated Posterior Probabilities for Genotypes 0/0, 0/1 and 1/1">
##INFO=<ID=AF,Number=1,Type=Float,Description="Estimated Alternate Allele Frequency">
##INFO=<ID=MAF,Number=1,Type=Float,Description="Estimated Minor Allele Frequency">
##INFO=<ID=R2,Number=1,Type=Float,Description="Estimated Imputation Accuracy">
##INFO=<ID=ER2,Number=1,Type=Float,Description="Empirical (Leave-One-Out) R-square (available only for genotyped variants)">
##bcftools_viewVersion=1.3.1+htslib-1.3.1
##bcftools_viewCommand=view -h chr1.dose.vcf.gz
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	1005723	1:1005723	C	T	.	PASS	AF=0.00024;MAF=0.00024;R2=0.00509
1	1005741	1:1005741	G	A	.	PASS	AF=2e-05;MAF=2e-05;R2=0.00012
1	1005806	1:1005806	C	T	.	PASS;GENOTYPED	AF=0.14489;MAF=0.14489;R2=0.99784;ER2=0.71745
1	1006223	1:1006223	G	A	.	PASS	AF=0.58207;MAF=0.41793 R2=0.80402
1	1007222	1:1007222	G	T	.	PASS	AF=0.14226;MAF=0.14226;R2=0.93284
1	1018598	1:1018598	A	G	.	PASS	AF=0.054;MAF=0.054;R2=0.61048

FORMAT	Sample1	Sample2	Sample3
GT:DS:GP	0 0:0:1,0,0	0 0:0:1,0,0	0 0:0.012:0.988,0.012,0
GT:DS:GP	0 0:0:1,0,0	0 0:0:1,0,0	0 0:0:1,0,0
GT:DS:GP	0 0:0:1,0,0	0 1:1:0,0.999,0.001	0 0:0:1,0,0
GT:DS:GP	1 1:1.912:0.002,0.085,0.913	0 0:0.366:0.635,0.365,0	0 1:1.29:0.012,0.685,0.302
GT:DS:GP	0 0:0.001:0.999,0.001,0	0 1:0.987:0.017,0.979,0.004	0 0:0.001:0.999,0.001,0
GT:DS:GP	0 0:0.002:0.998,0.002,0	0 0:0.01:0.99,0.01,0	0 0:0.493:0.507,0.493,0

3 main genotype output formats
Probs format (probability of AA AB and BB genotypes for each SNP)
Hard call or best guess (output as A C T or G allele codes)
Dosage data (most common – 1 number per SNP, 1-2)



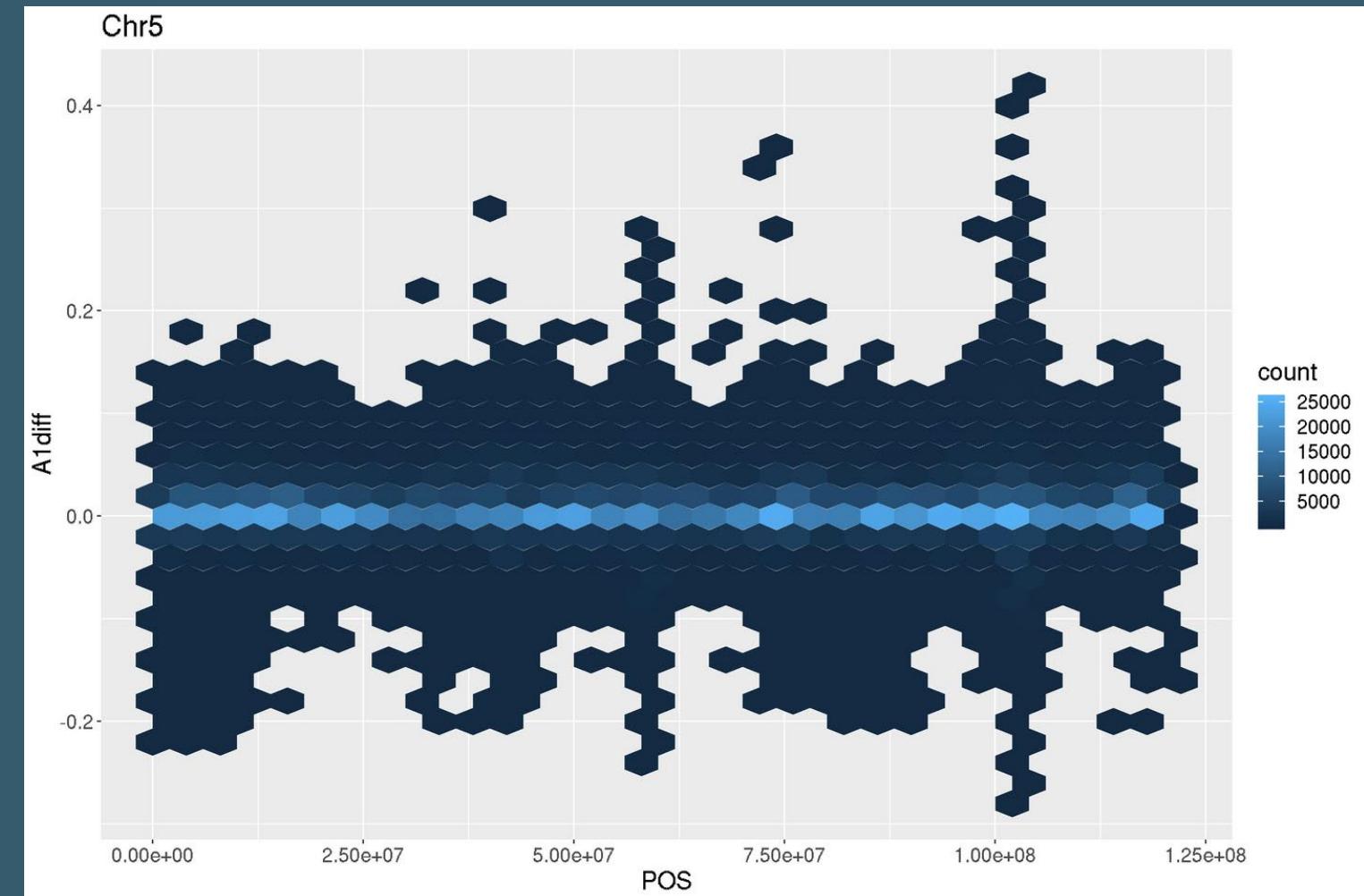
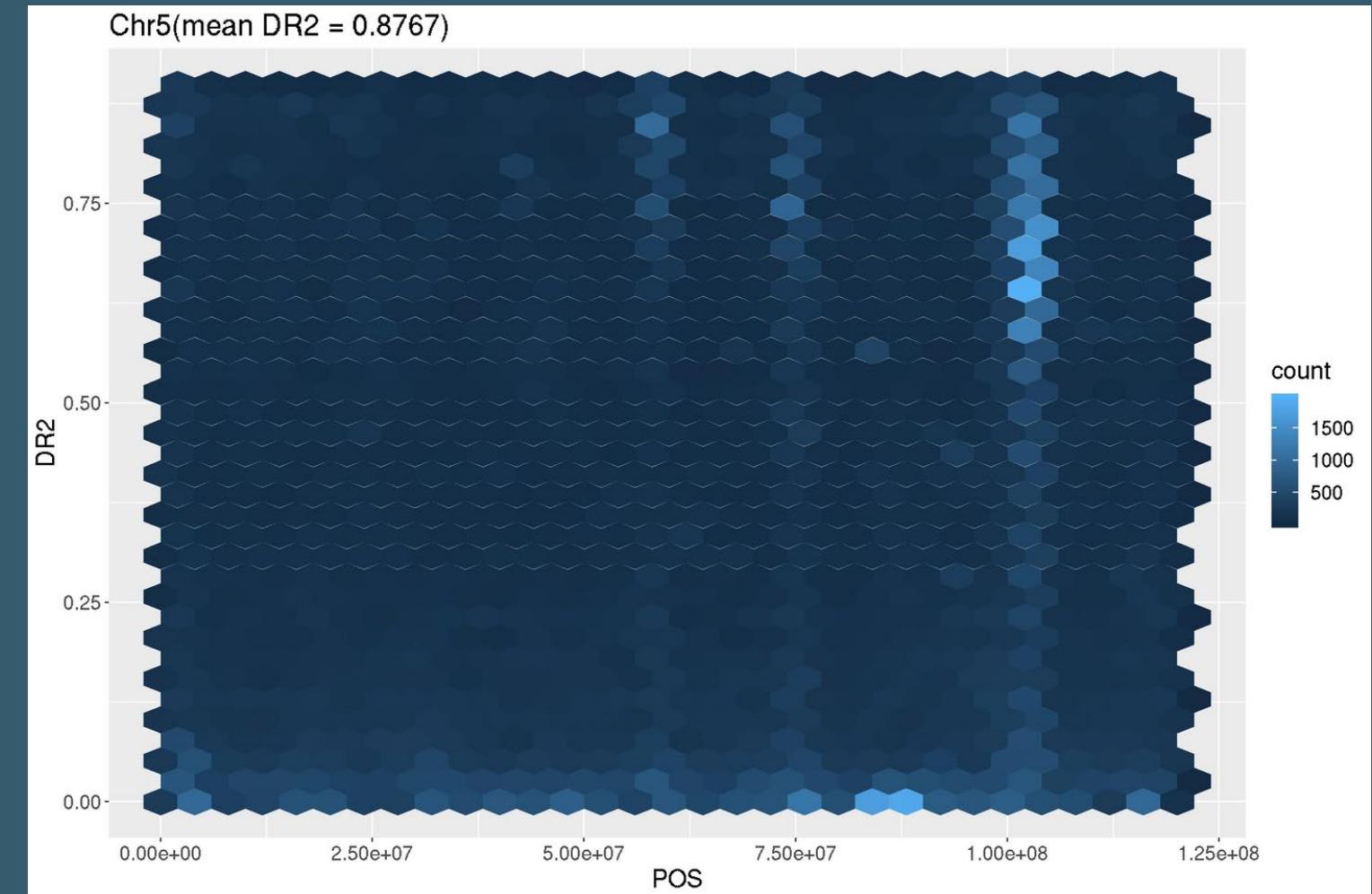


Post imputation QC

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 number and % of SNPs:
 - MAF < 0.5%
 - With r² 0-0.3, 0.3-0.6, 0.6-1
 - If you have hard calls or probs data HWE P < 1E-6
 - If you have families convert to hard calls and check for Mendelian errors



(Wang et al. 2020 ICQG6)



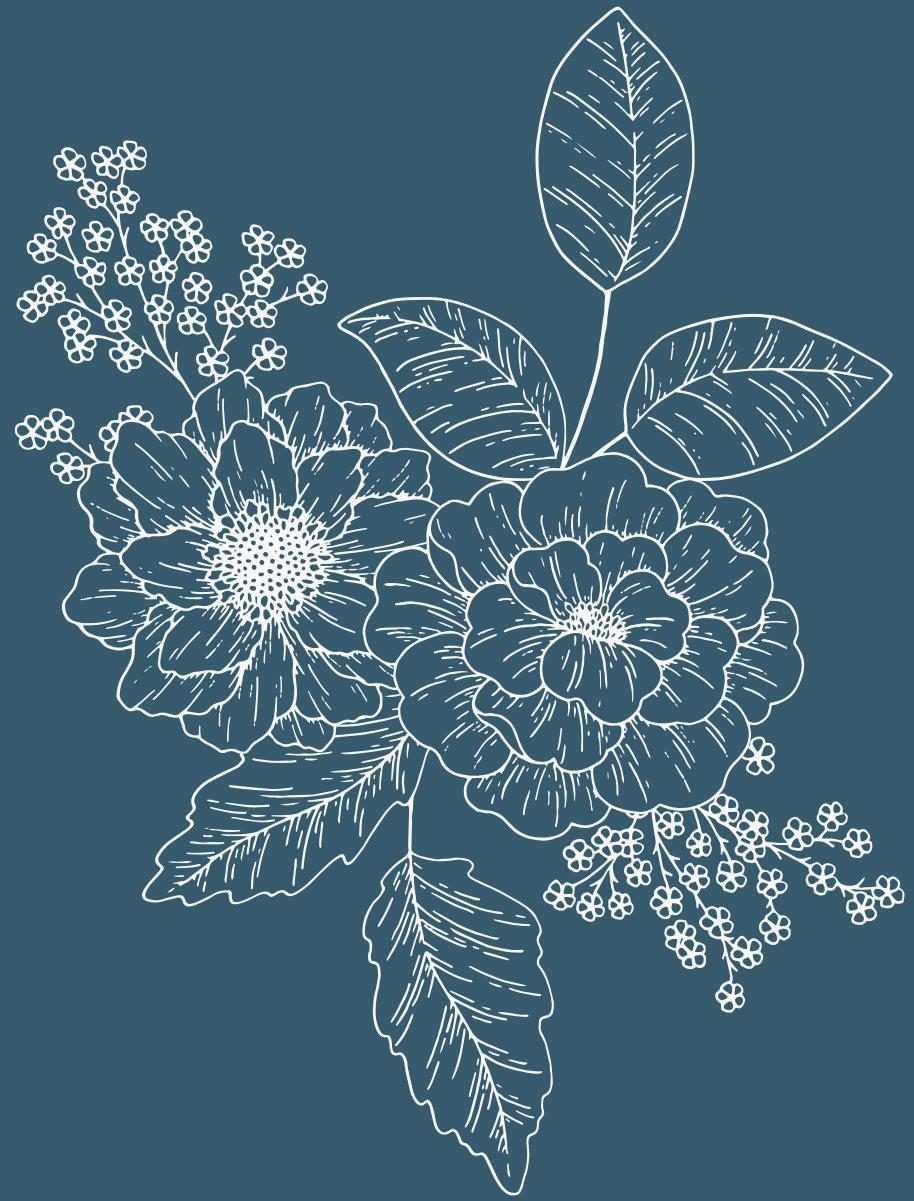
Take home message

- Genotype concordance is not an appropriate parameter for evaluating imputation performance
- Accuracy depends on size of reference, relationship to reference, marker density, map quality etc.
 - Imputation to sequence possible, relatively low accuracies for rare alleles
- Use genotype probabilities from imputation in GWAS and genomic prediction



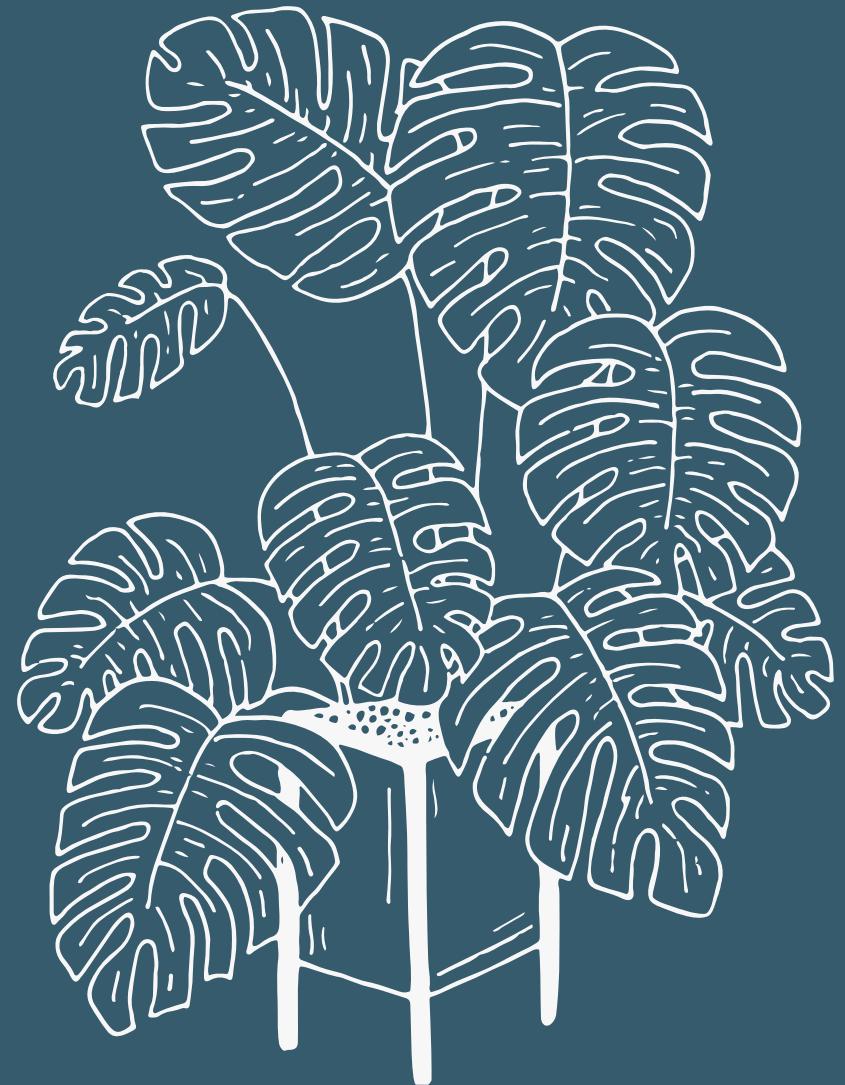
Practical after the break

-
- 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 and
 - how the performance of imputation is evaluated



Thank you!

Yu Wang (Scientist in Quantitative Genetics)
yu.wang@lic.co.nz



Imputation to sequence

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