Imputation workshop

Yu Wang (Postdoctoral Scientist Quantitative Genetics & Genomics)
yu.wang@lic.co.nz

Today's Agenda

Welcome to the workshop!

9:00-9:30 Test NeSI & Greeting

9:30-10:30 Imputation introduction (Dr. Yu Wang)

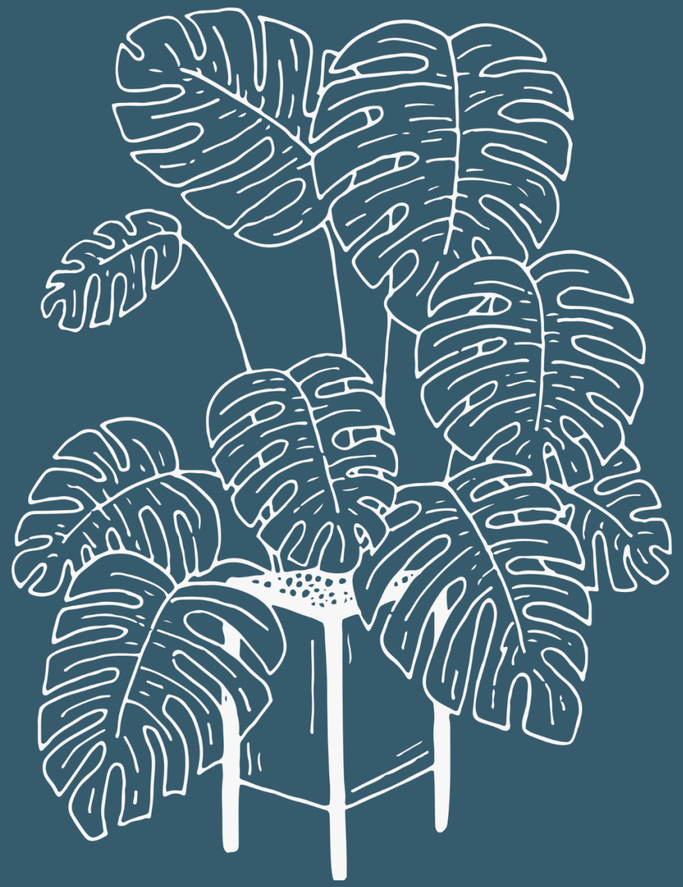
10:30-11:00 Coffee break

11:00-12:00 Imputation application (Dr. Andrew Wallace)

12:00-13:00 Lunch break

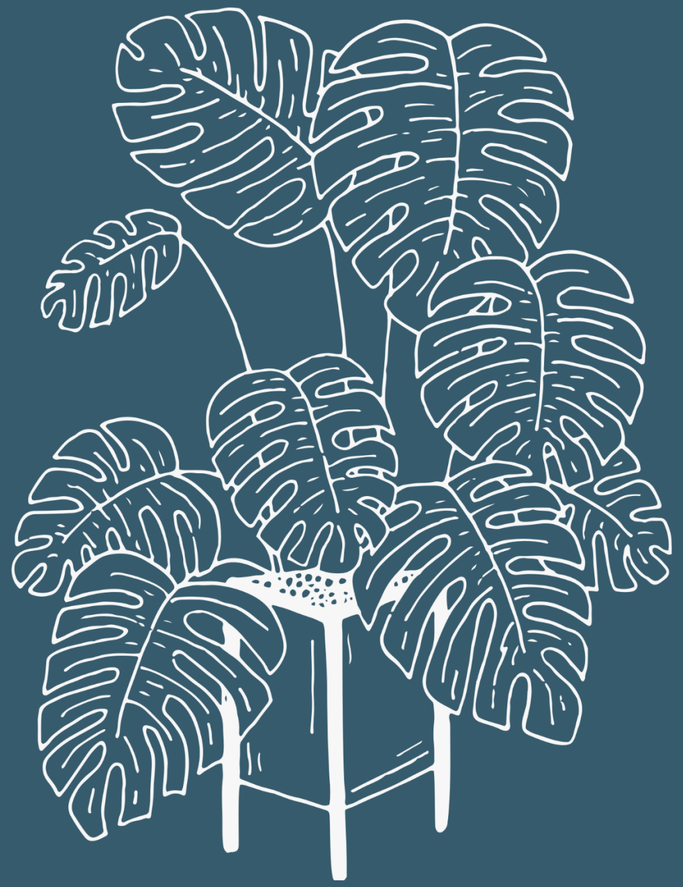
13:00-15:30 Tutorial

15:30-16:00 Wrap up



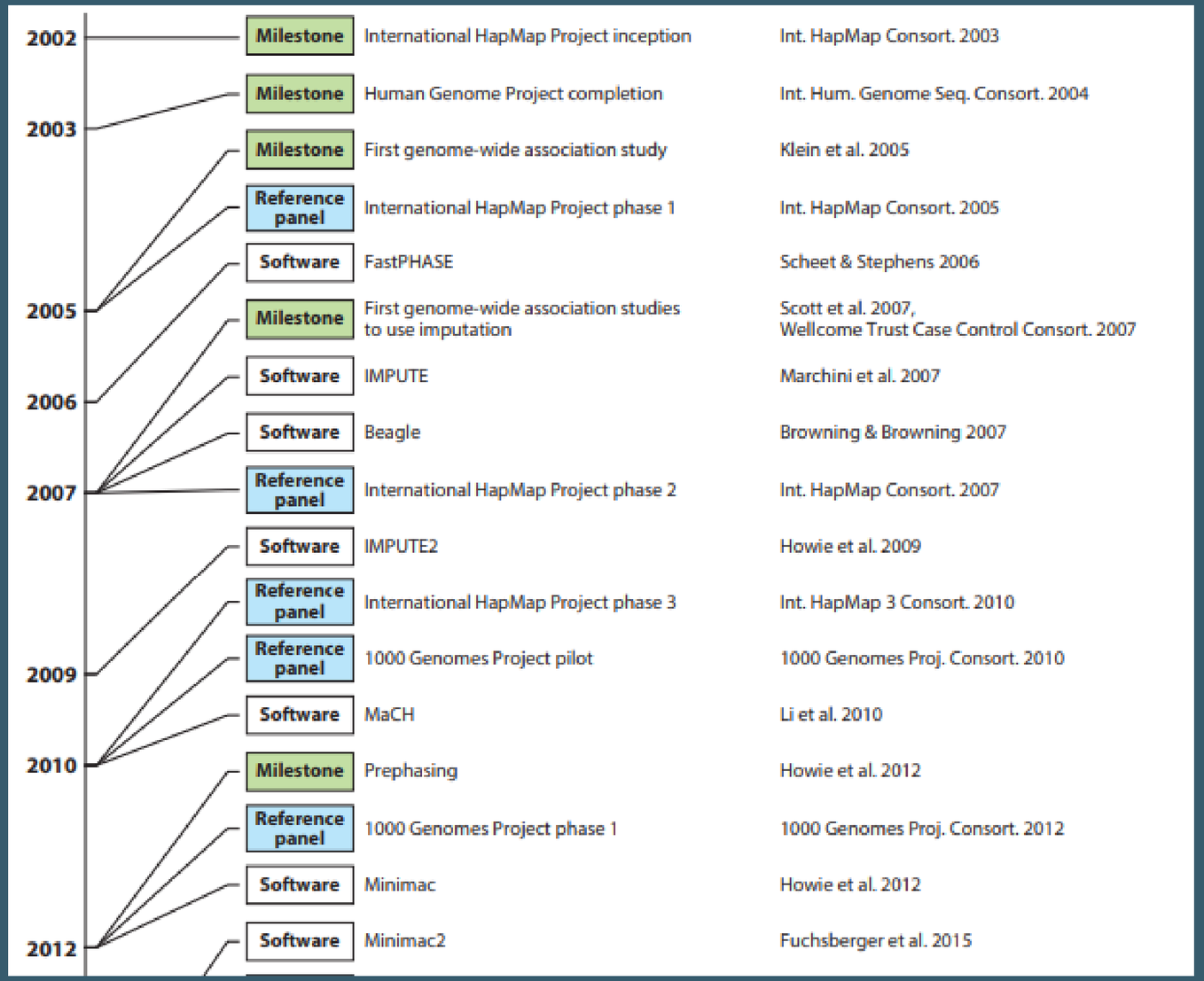
Self intro & What species you are working on

- Human
- Plants
- Animals
- Microbioms
- Others

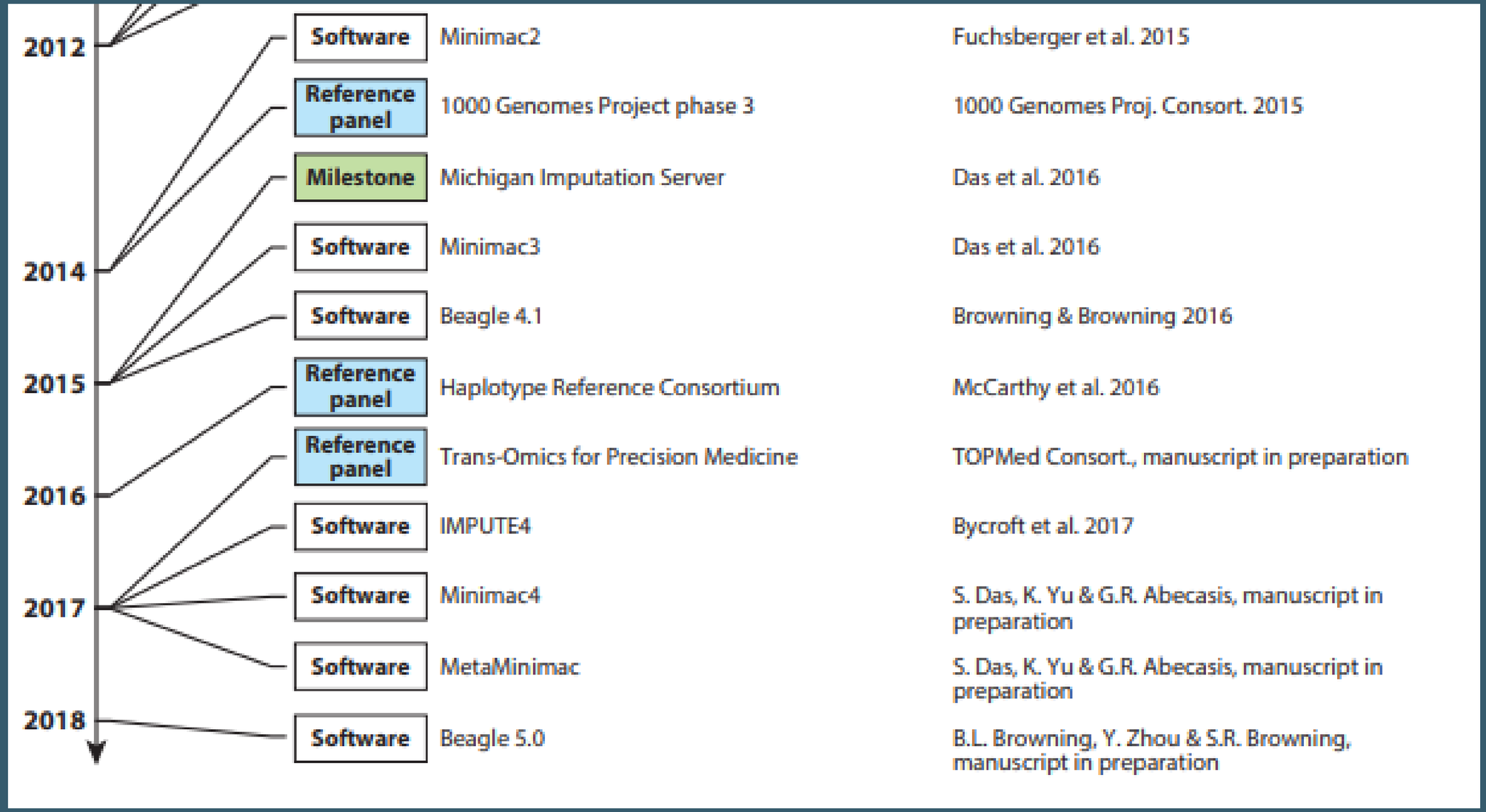


Content

- What is imputation
- Why we need imputation
- Phasing
- How to evaluate the imputation accuracy



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



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

What is imputation

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

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

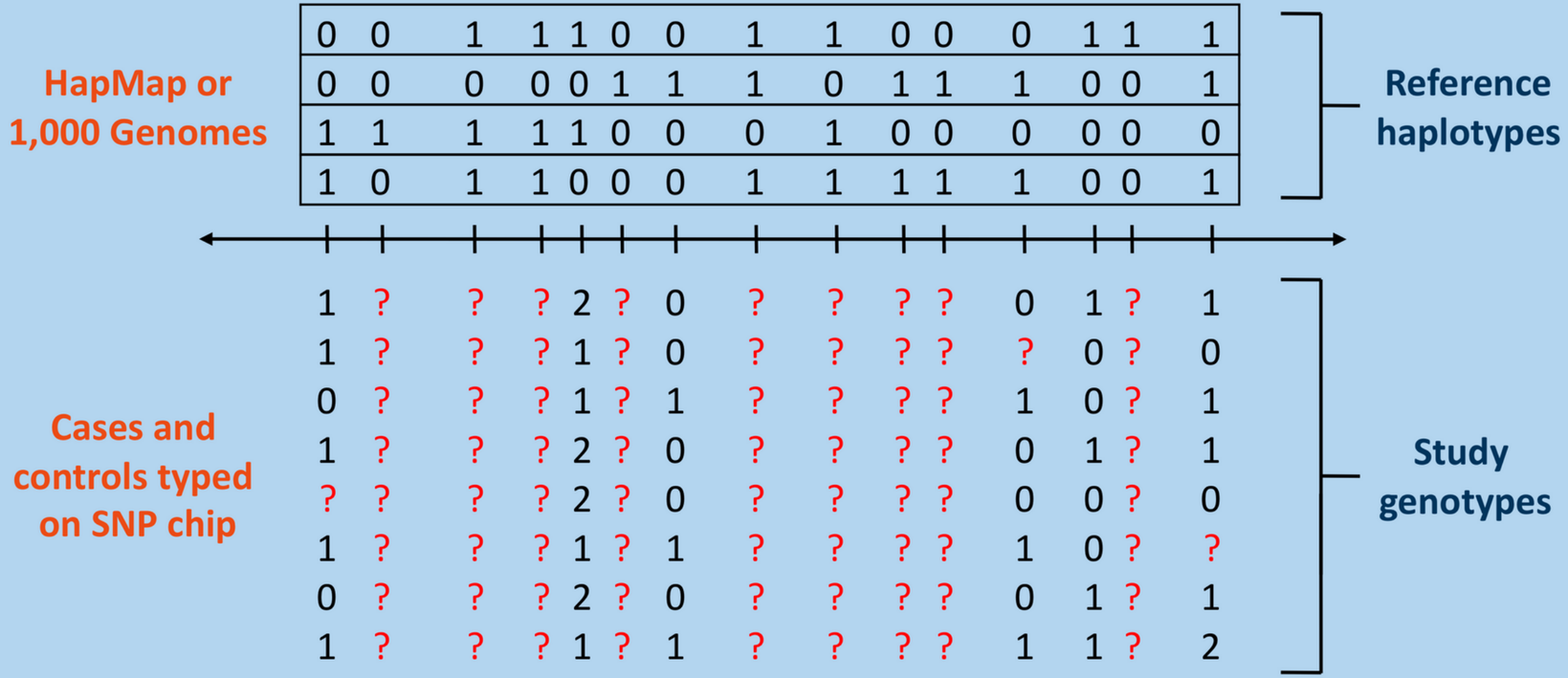
- missing genotypes at typed variants
- genotypes at un-typed variants that are present in an external high-density "reference panel" of phased haplotypes

In silico genotypes can be tested for association within standard generalised linear regression framework

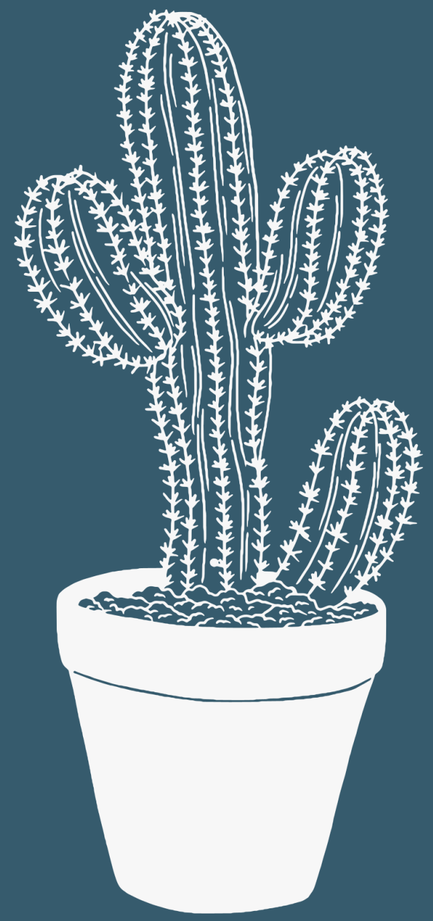
low density
panel

high density
panel

Typical imputation scenario



http://mathgen.stats.ox.ac.uk/impute/impute_v2.html



What is phasing


What is Phasing?

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
- Animal-ImputeDB (animals:)

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

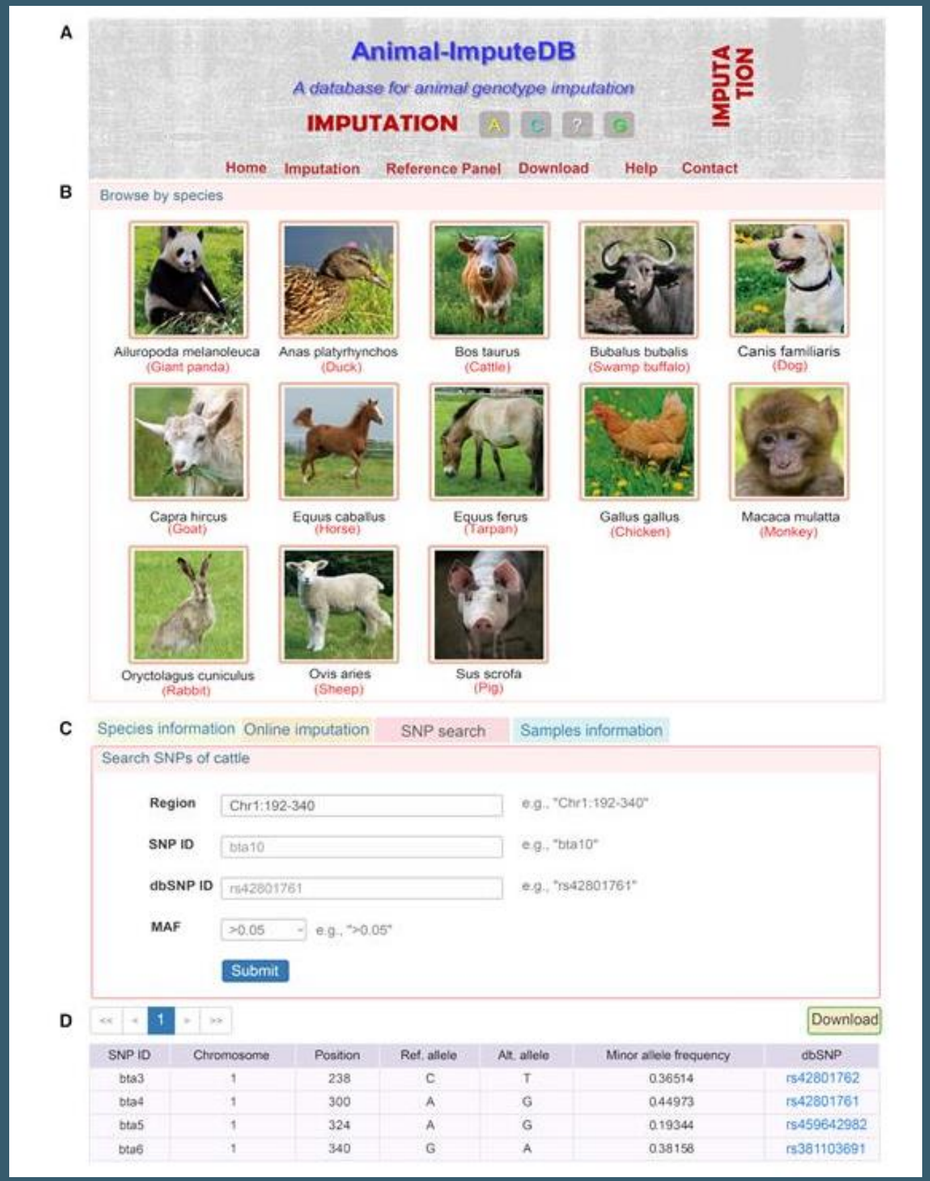


Table 1 Genotype imputation tools that employ a hidden Markov model (HMM)

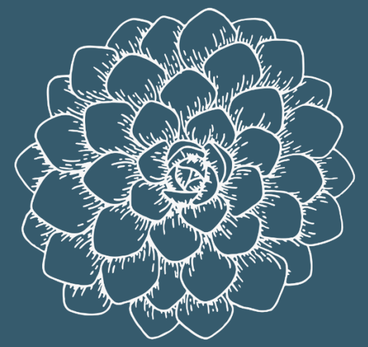
Tool	Year	Description of state space	Computational complexity	HMM parameter functions
FastPHASE	2006	All genotype configurations from a fixed number of localized haplotype clusters	Maximization-step linear in number of haplotypes, quadratic in number of clusters	Depends on recombination and mutation 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

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

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 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.

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.

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).

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

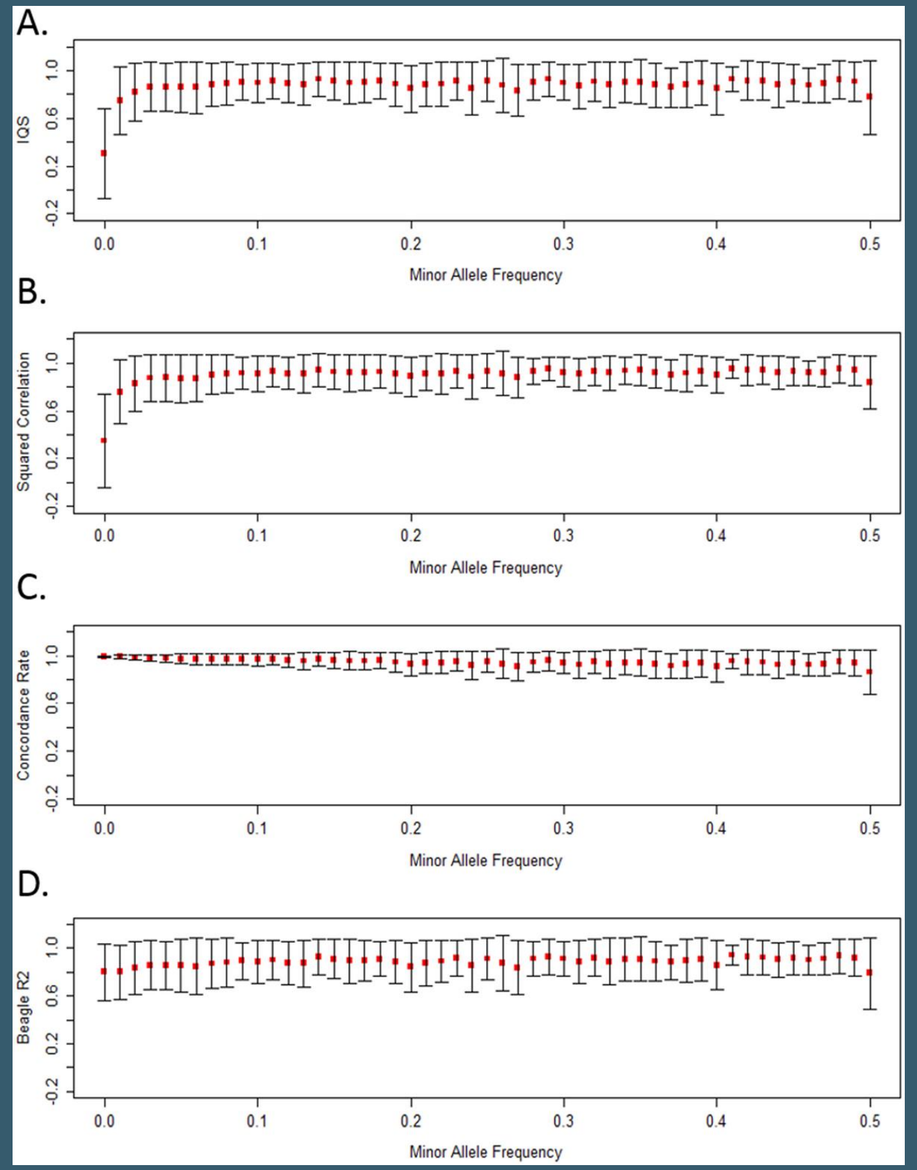


Table 3: The correlation between R^2 and the imputation quality measure

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

[a]The Spearman correlation coefficient between R^2 and the imputation quality measure.
[b]The 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): 549-562.

R-squared

- Dosage R-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

[a]The Spearman correlation coefficient between R^2 and the imputation quality measure.
[b]The 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): 549-562.

R-squared

- Dosage R-squared and Allelic R-squared

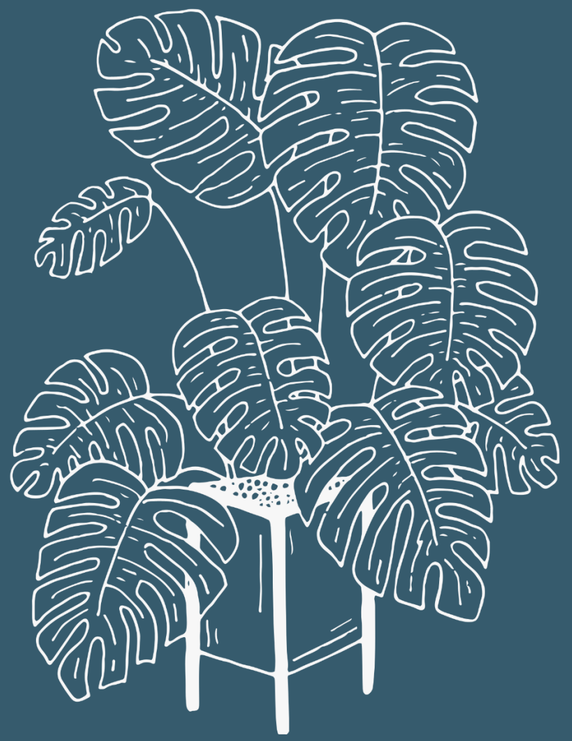
We noticed that in this setting, the quality measure of BEAGLE is comparable with (or even slightly better than) the quality measures from the other two programs in removing poorly imputed variants from their corresponding imputation results.

...

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).

Factors influence 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 in the reference population (Zhang and Druet, 2010)
- The relationship between imputed individuals and individuals genotyped at high density (Hickey et al., 2012)



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

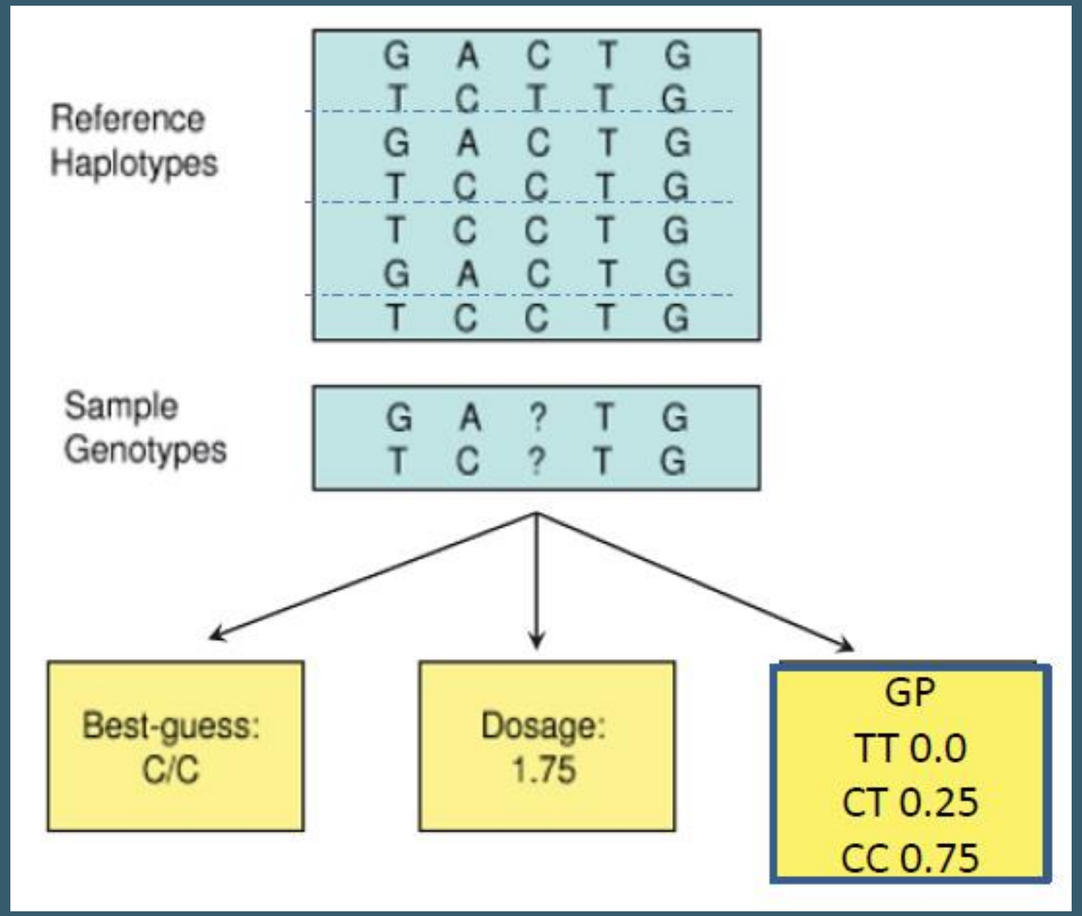
VCF file

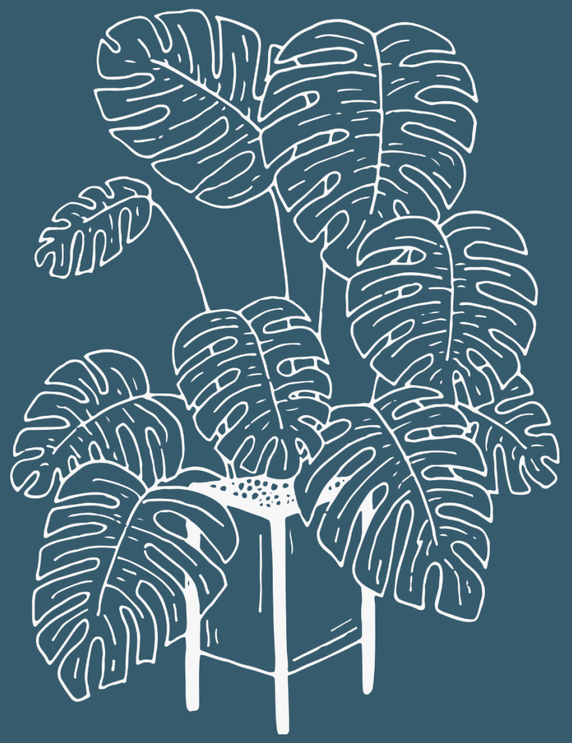
```
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="All filters passed">
##filedate=2018.7.25
##source=Minimac3
##contig=<ID=1>
##FILTER=<ID=GENOTYPED,Description="Marker was genotyped AND imputed">
##FILTER=<ID=GENOTYPED_ONLY,Description="Marker was genotyped but NOT imputed">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##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)


Reference Haplotypes
G A C T G
T C T T G
G A C T G
T C C T G
T C C T G
G A C T G
T C C T G
Sample Genotypes
G A ? T G
T C ? T G
Best-guess: C/C
Dosage: 1.75
GP
TT 0.0
CT 0.25
CC 0.75



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

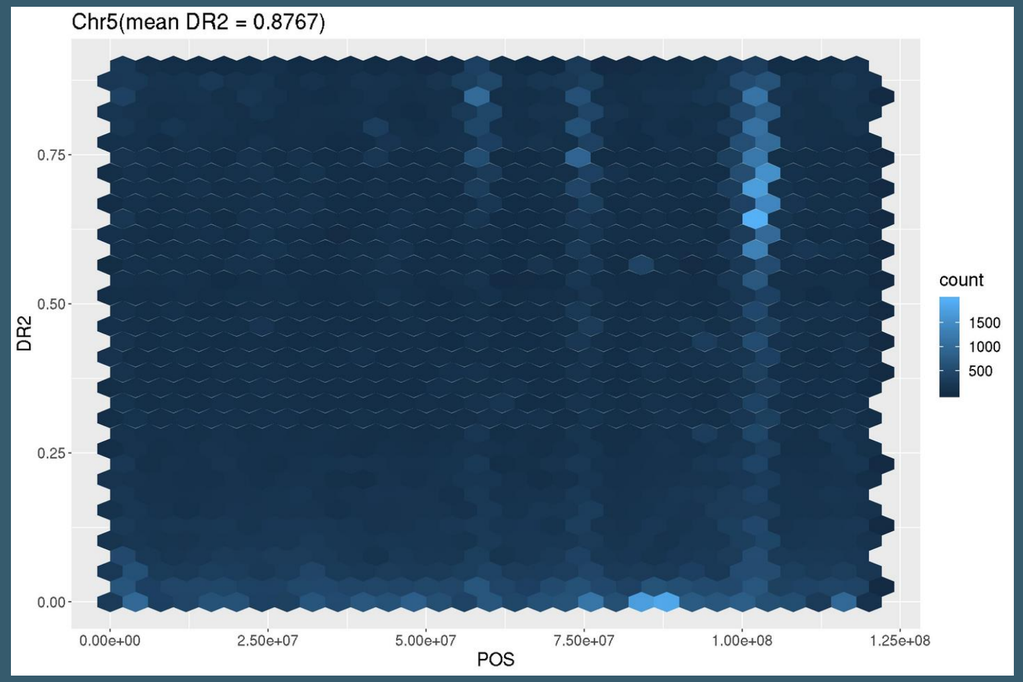
•For each chromosome check N and % of SNPs:

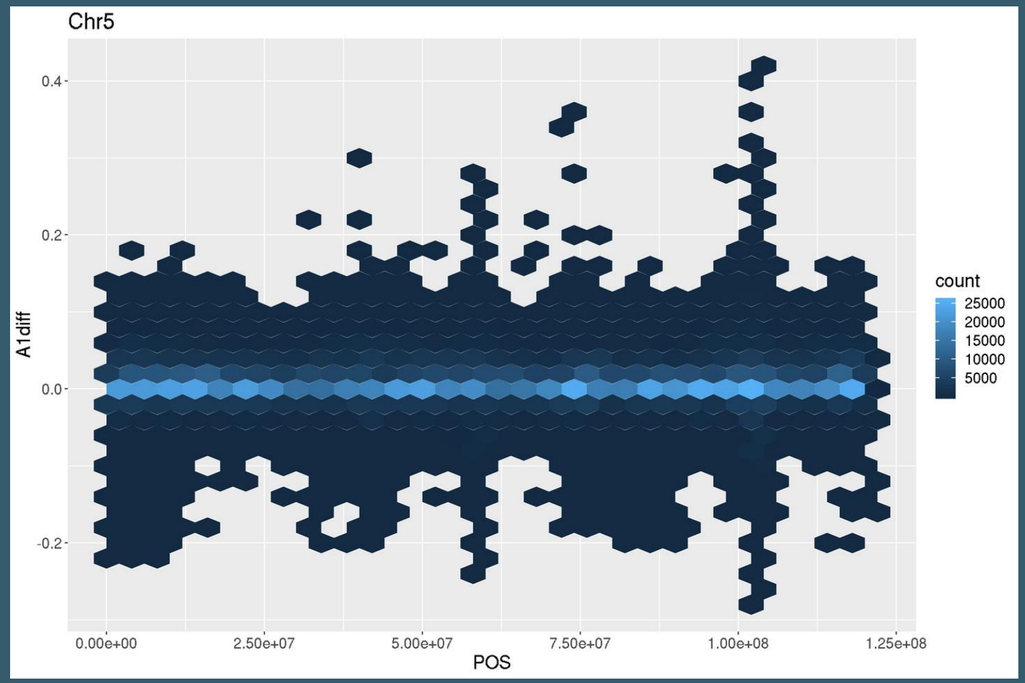
•MAF <.5%

•With r2 0-0.3, 0.3-0.6,0.6-1

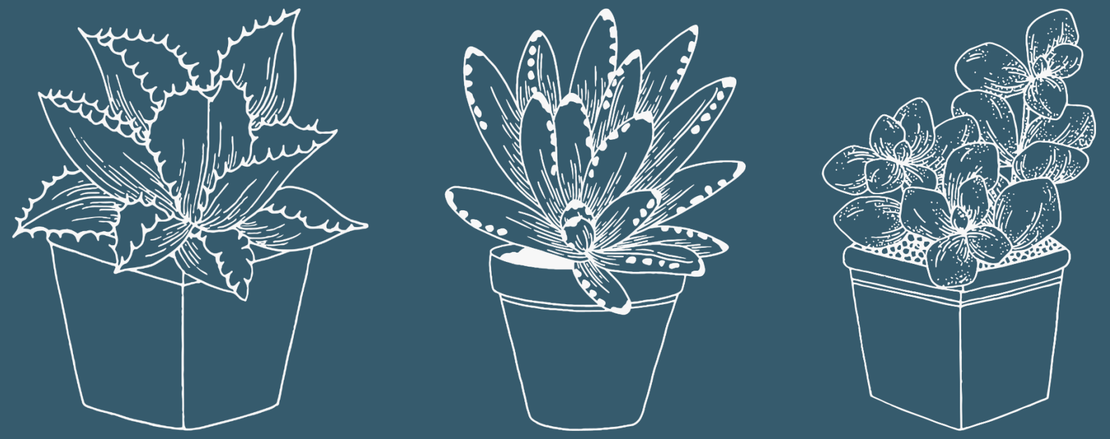
•If you have hard calls or probs data HWE P < 1E-6

•If you have families convert to hard calls and check for Mendelian errors



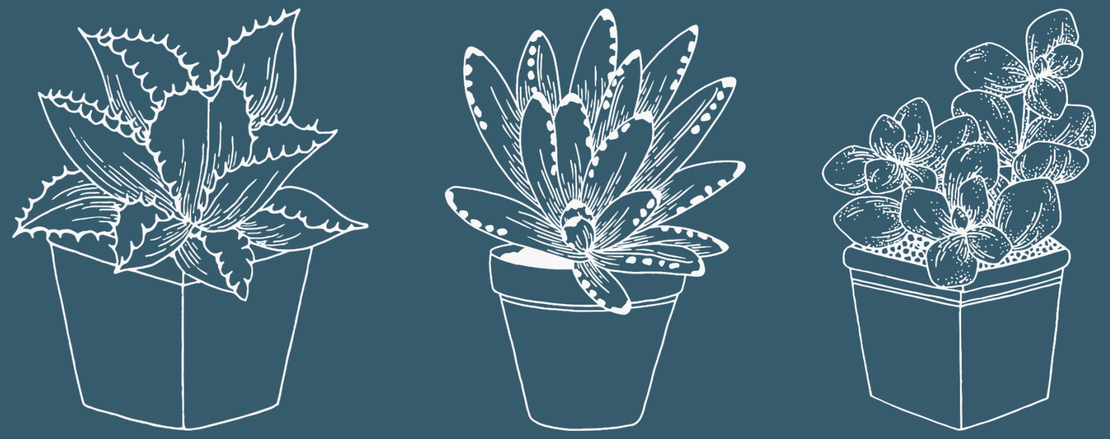


(Wang et al. 2020 ICQG6)



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 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 and how the performance of imputation is evaluated

Thank you!

Yu Wang (Postdoctoral Scientist Quantitative Genetics & Genomics)
yu.wang@lic.co.nz