

# Genotype imputation procedures

...

Evangelina López de Maturana & Oscar González-Recio

# Topics

Background

Imputation

Imputation  
performance

Imputation strategies

Why imputation  
is important?

Definition  
Advantages

Factors affecting  
the imputation  
performance

Tools  
Strategies

# Background

Next-generation sequencing (NGS) has revolutionized human, plant and animal research by providing powerful genotyping methods

They provide a straightforward workflow to identify, validate, and screen genetic variants in a short time with a low cost

One limitation of SNP genotyping arrays is that they assay only a small fraction of human genetic variation

The variants assayed on SNP arrays are chosen based on the linkage disequilibrium structure of the human or other species genome

# Background

Without imputation, GWASs that test variants on a commercial genotyping array must rely on pairwise linkage disequilibrium between an assayed SNP and a causal variant to detect association between the assayed SNP and trait

However, rare variants, which are more often associated with dramatic functional consequences, tend to have low levels of pairwise linkage disequilibrium with common variants on SNP genotyping arrays

Although NGS have significantly reduced the cost of sequencing a genome, it remains prohibitively expensive to whole-genome sequence millions of samples

# Imputation

Genotype imputation is a process of estimating missing genotypes from the haplotype or genotype reference panel

Imputation works by copying haplotype segments from a densely genotyped reference panel into individuals typed at a subset of the reference variants

# Typical imputation scenario

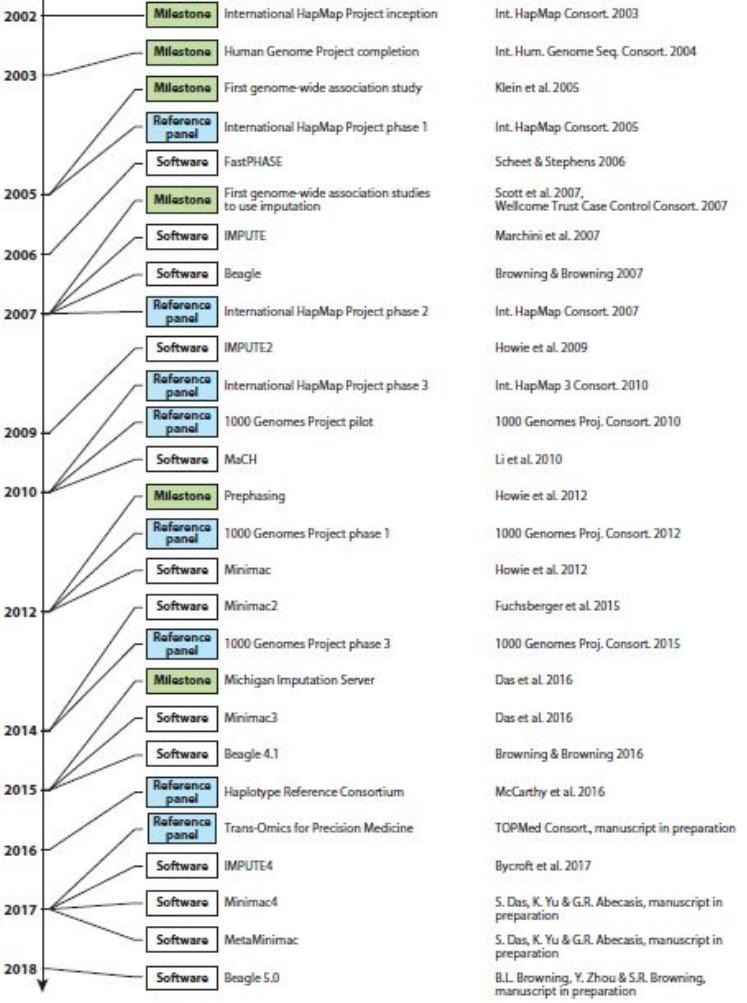
The diagram illustrates a typical imputation scenario. At the top, a block of 12 reference haplotypes is shown, with a bracket on the right labeled "Reference haplotypes". Below this is a horizontal line with tick marks, representing the genomic position of the SNPs. Below the line, a block of 8 study genotypes is shown, with a bracket on the right labeled "Study genotypes". The reference haplotypes and study genotypes are aligned by position. The study genotypes contain many missing data points, represented by question marks.

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

<https://jmarchini.org/impute2/>

Figure 1

A brief time line summarizing the major developments in genotype imputation. Each major development has been categorized as a milestone (green), reference panel (blue), or software (white).



Das et al, 2018

# Imputation

Advantages:

- Reconstruction of rare-variant genotypes
- Boost the power of detecting SNPs in GWAS
- Integrate multiple studies for meta-analysis
- Building PRS
- Guide fine mapping studies
- Reduces cost
- To estimate other types of genetic variations, such as CNV or classical HLA alleles

# Tools

Phasing methods: since genotype imputation is a highly computationally intensive process, prephasing may reduce the computational burden

- MACH: it uses an HMM model (<http://csg.sph.umich.edu/~yli/mach/download/>)
- SHAPEIT2: haplotype estimation method using HMM on a graph structure of haplotypes ([https://mathgen.stats.ox.ac.uk/genetics\\_software/shapeit/shapeit.html](https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html))

Imputation tools:

- Beagle (<https://faculty.washington.edu/browning/beagle/beagle.html>)
- IMPUTE2 ([http://mathgen.stats.ox.ac.uk/impute/impute\\_v2.html](http://mathgen.stats.ox.ac.uk/impute/impute_v2.html))
- Minimac (<https://genome.sph.umich.edu/wiki/Minimac#Download>)

# Imputation strategies

- MACH + Minimac
- SHAPEIT2 + IMPUTE2
- IMPUTE2
- Beagle

# Imputation strategies

Most of the available tools rely on a general framework that uses a hidden Markov model (HMM) to describe the data (Li and Stephens model)

- The observed genotypes of unknown phase in a study sample represent the observed data of the HMM
- An underlying and unobserved set of phased genotypes represent the hidden states of the HMM

# Li and Stephens model state space



Figure 2

An illustration of genotype imputation, showing the process of imputation for a study haplotype ( $S_G$ ) genotyped at 6 markers using a reference panel of sequenced haplotypes at 21 markers. The alleles in  $S_G$  are used to match short segments from the reference panel. For example, in the first genomic segment, the alleles T and G imply that the corresponding segment might have been copied from haplotype  $X_3$ . In the second segment, the alleles A and T imply that haplotype  $X_5$  might have been copied. Proceeding similarly, the study haplotype can be represented as a mosaic of DNA segments from haplotypes  $X_3$ ,  $X_5$ , and  $X_6$ . Consequently, the missing sites can be imputed to obtain the final imputed haplotype,  $S_I$ .

- Each allele on each reference haplotype corresponds to an HMM state
- Each study sample haplotype is assumed to trace an unobserved path through the grid, proceeding left to right from the first reference marker to the last reference marker

Das et al 2018

# Li and Stephens model state space

Determined by the HMM transition probabilities (related to population recomb rate)



Figure 2

An illustration of genotype imputation, showing the process of imputation for a study haplotype ( $S_G$ ) genotyped at 6 markers using a reference panel of sequenced haplotypes at 21 markers. The alleles in  $S_G$  are used to match short segments from the reference panel. For example, in the first genomic segment, the alleles T and G imply that the corresponding segment might have been copied from haplotype  $X_3$ . In the second segment, the alleles A and T imply that haplotype  $X_5$  might have been copied. Proceeding similarly, the study haplotype can be represented as a mosaic of DNA segments from haplotypes  $X_3$ ,  $X_5$ , and  $X_6$ . Consequently, the missing sites can be imputed to obtain the final imputed haplotype,  $S_I$ .

- A new segment in the mosaic begins when the path switches reference haplotypes (rows) between one marker and the next

Das et al 2018

prediction

# Li and Stephens model state space

Determined by the HMM transition probabilities (related to population recomb rate)



Figure 2

An illustration of genotype imputation, showing the process of imputation for a study haplotype ( $S_G$ ) genotyped at 6 markers using a reference panel of sequenced haplotypes at 21 markers. The alleles in  $S_G$  are used to match short segments from the reference panel. For example, in the first genomic segment, the alleles T and G imply that the corresponding segment might have been copied from haplotype  $X_3$ . In the second segment, the alleles A and T imply that haplotype  $X_5$  might have been copied. Proceeding similarly, the study haplotype can be represented as a mosaic of DNA segments from haplotypes  $X_3$ ,  $X_5$ , and  $X_6$ . Consequently, the missing sites can be imputed to obtain the final imputed haplotype,  $S_I$ .

- The probability of a template switch between markers is determined by the **HMM transition probabilities** (related to population **recomb rate**)

- The probability that an observed allele differs from the template allele is determined by the HMM emission probabilities

Das et al 2018

# Li and Stephens model state space

Determined by the HMM transition probabilities (related to population recomb rate)



Figure 2

An illustration of genotype imputation, showing the process of imputation for a study haplotype ( $S_G$ ) genotyped at 6 markers using a reference panel of sequenced haplotypes at 21 markers. The alleles in  $S_G$  are used to match short segments from the reference panel. For example, in the first genomic segment, the alleles T and G imply that the corresponding segment might have been copied from haplotype  $X_3$ . In the second segment, the alleles A and T imply that haplotype  $X_5$  might have been copied. Proceeding similarly, the study haplotype can be represented as a mosaic of DNA segments from haplotypes  $X_3$ ,  $X_5$ , and  $X_6$ . Consequently, the missing sites can be imputed to obtain the final imputed haplotype,  $S_I$ .

- Given an observed haplotype with missing alleles, the **probability of each possible path through the HMM states can be calculated**
- This path is **penalized** when the path **switches reference haplotypes** and when the **reference allele differs from the observed allele**

Das et al 2018

# Li and Stephens model state space

Determined by the HMM transition probabilities (related to population recomb rate)



Figure 2

An illustration of genotype imputation, showing the process of imputation for a study haplotype ( $S_G$ ) genotyped at 6 markers using a reference panel of sequenced haplotypes at 21 markers. The alleles in  $S_G$  are used to match short segments from the reference panel. For example, in the first genomic segment, the alleles T and G imply that the corresponding segment might have been copied from haplotype  $X_3$ . In the second segment, the alleles A and T imply that haplotype  $X_5$  might have been copied. Proceeding similarly, the study haplotype can be represented as a mosaic of DNA segments from haplotypes  $X_3$ ,  $X_5$ , and  $X_6$ . Consequently, the missing sites can be imputed to obtain the final imputed haplotype,  $S_I$ .

- The probability that the target haplotype carries a particular allele is the sum of the state probabilities corresponding to reference haplotypes that carry the allele

Das et al 2018

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 <sup>a</sup>	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

This table describes the typical state space and parameter functions used to model the Li and Stephens framework. Minimac and IMPUTE2 were the first tools to use the prephasing approach. Minimac3 and Beagle 4.1 exploit local haplotype redundancy to reduce the size of the state space and hence the computational burden.

<sup>a</sup>IMPUTE4 uses the same HMM as IMPUTE2; however, to reduce memory usage and increase speed, it uses compact binary data structures and takes advantage of high correlations between inferred copying states in the HMM to reduce computation.

Das et al 2018



# Beagle <http://faculty.washington.edu/browning/beagle/beagle.html>

Software for phasing genotypes and for imputing ungenotyped markers

It is a hidden Markov model, which uses a clustering graphical model on haplotypes

Different versions have different characteristics:

- BEAGLE 5.0 and 5.1: default settings show better performance than BEAGLE 4.0 and 4.1, especially in less diverse populations
- BEAGLE 5.0 and 5.1: reduced computing times and memory requirements
- BEAGLE 4.0: can incorporate pedigree data and genotype likelihoods
- Up to BEAGLE 4.0 all markers are assumed to be equidistant, whereas in BEAGLE 4.1, 5.0 and 5.1 the genetic distance between markers can be provided

## IMPUTE2 ([http://mathgen.stats.ox.ac.uk/impute/impute\\_v2.html](http://mathgen.stats.ox.ac.uk/impute/impute_v2.html))

It uses a McMC algorithm in which each iteration includes two-steps, phasing and imputation, to maximize the posterior probabilities of the missing alleles for imputation

First step: it infers haplotype conditioning from information of the study sample, reference, and recombination rate using a Markov Chain Monte Carlo approach

Second step: it uses a hidden markov model to impute the missing genotypes on the haplotypes inferred in the first step

This MCMC algorithm is run for a number of iterations (typically 30, including 10 burn-in iterations), then the probabilities from Step 2 are averaged across iterations to produce marginal posterior genotype probabilities at each untyped SNP

# Minimac (<https://github.com/statgen/Minimac4>)

It relies on a two step approach:

1. Samples that are to be analyzed must be phased into a series of estimated haplotypes.
2. Imputation is carried out directly into these phased haplotypes

It uses state space reduction HMM to reduce the running time

[https://genome.sph.umich.edu/wiki/Minimac:\\_1000\\_Genomes\\_Imputation\\_Cookbook#Minimac\\_Imputation](https://genome.sph.umich.edu/wiki/Minimac:_1000_Genomes_Imputation_Cookbook#Minimac_Imputation)

Howie B, Fuchsberger C, Stephens M, Marchini J, and Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics* 2012

# Online imputation servers

Phasing and imputation servers provide an alternative way to perform phasing and imputation from large reference panels that have restrictive sample consents,

For example, the HRC reference panel may be used for genotype phasing and imputation but not for any other purpose

(<https://www.ebi.ac.uk/ega/studies/EGAS00001001710>)

# Online imputation servers

<https://imputationserver.sph.umich.edu/index.html#!>

# Michigan Imputation server

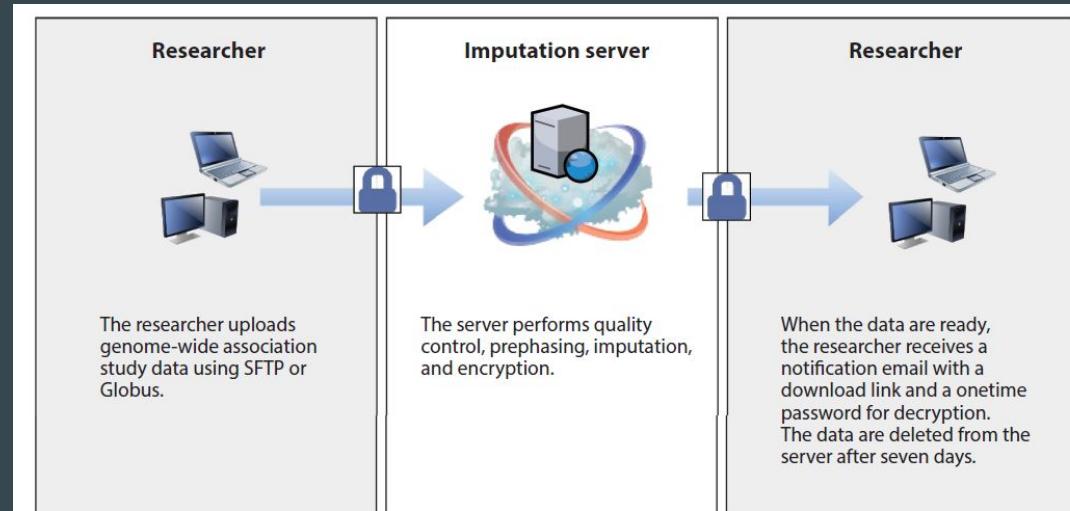
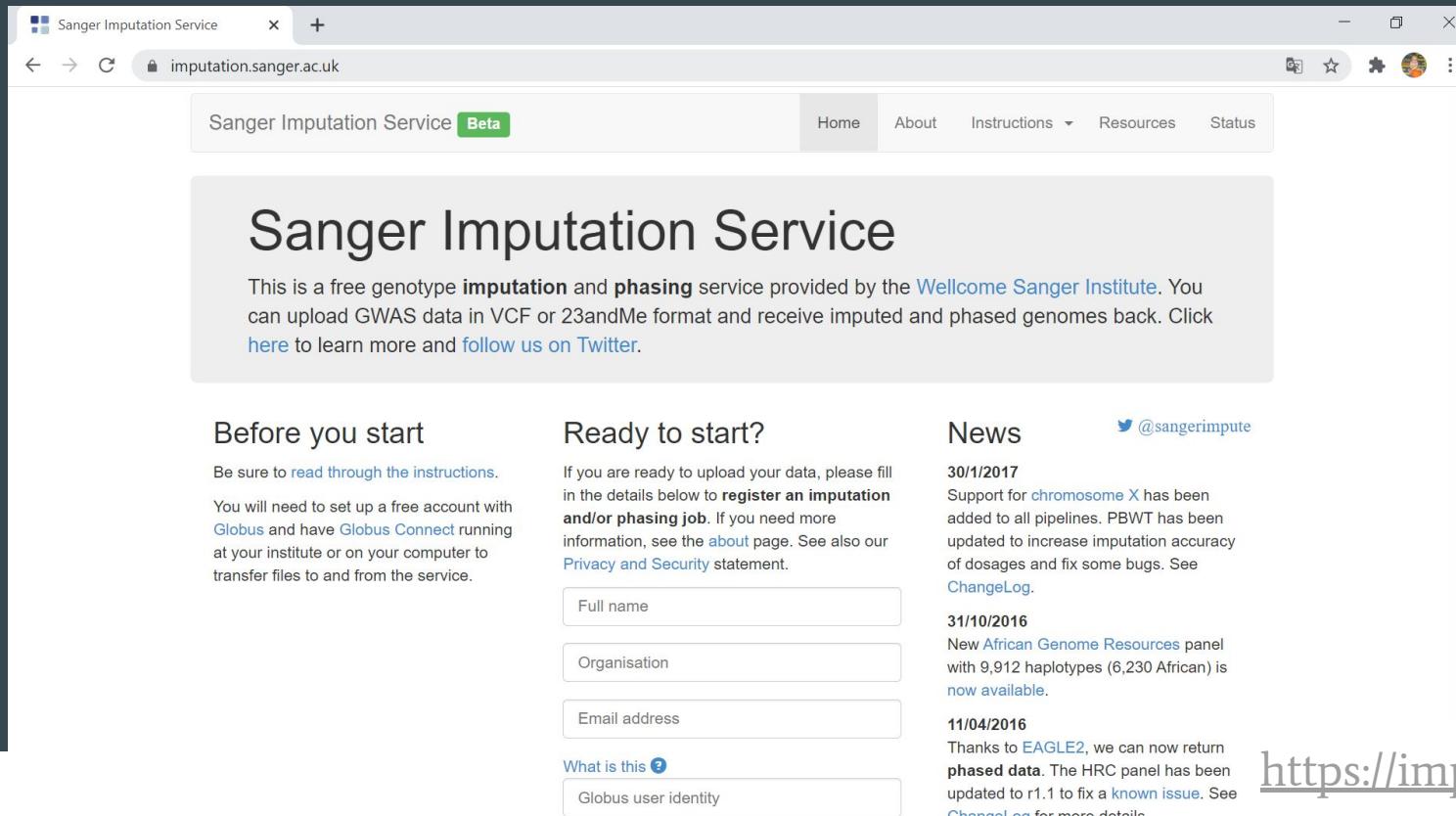


Figure 3

An outline of the pipeline in the Michigan imputation server.

# Online imputation servers



The screenshot shows a web browser window for the Sanger Imputation Service. The address bar shows the URL [imputation.sanger.ac.uk](https://imputation.sanger.ac.uk). The page title is "Sanger Imputation Service" with a "Beta" badge. The navigation menu includes Home, About, Instructions (with a dropdown arrow), Resources, and Status. The main content area features a large heading "Sanger Imputation Service" and a descriptive paragraph about the service, mentioning genotype imputation and phasing, and providing links to the Wellcome Sanger Institute and Twitter. Below this, there are three main sections: "Before you start", "Ready to start?", and "News".

**Before you start**

Be sure to [read through the instructions](#). You will need to set up a free account with [Globus](#) and have [Globus Connect](#) running at your institute or on your computer to transfer files to and from the service.

**Ready to start?**

If you are ready to upload your data, please fill in the details below to [register an imputation and/or phasing job](#). If you need more information, see the [about](#) page. See also our [Privacy and Security](#) statement.

Full name  
Organisation  
Email address  
What is this ?  
Globus user identity

**News**

**30/1/2017**  
Support for [chromosome X](#) has been added to all pipelines. PBWT has been updated to increase imputation accuracy of dosages and fix some bugs. See [ChangeLog](#).

**31/10/2016**  
New [African Genome Resources](#) panel with 9,912 haplotypes (6,230 African) is now available.

**11/04/2016**  
Thanks to [EAGLE2](#), we can now return [phased data](#). The HRC panel has been updated to r1.1 to fix a [known issue](#). See [ChangeLog](#) for more details.

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

# Important considerations

- It is advisable to remove low-quality variants and individuals (standard GWAS quality control filterings)
- Convert the genotype data to the build of the reference panel
- All panels have their allele codings aligned to a fixed reference:
  - In human genetics, the genome of reference
  - Softwares as IMPUTE2 will align the strand between panels (flipping A/C to G/T in the reference) except for ambiguous alleles (A/T; G/C)
  - For ambiguous alleles, the MAF can be used to get the alignment except for those with MAF near 50%
  - It is important to check if the id of the SNP in the study sample corresponds to that in the reference population (SHAPEIT2 and IMPUTE2)
- Chromosome X: specific options

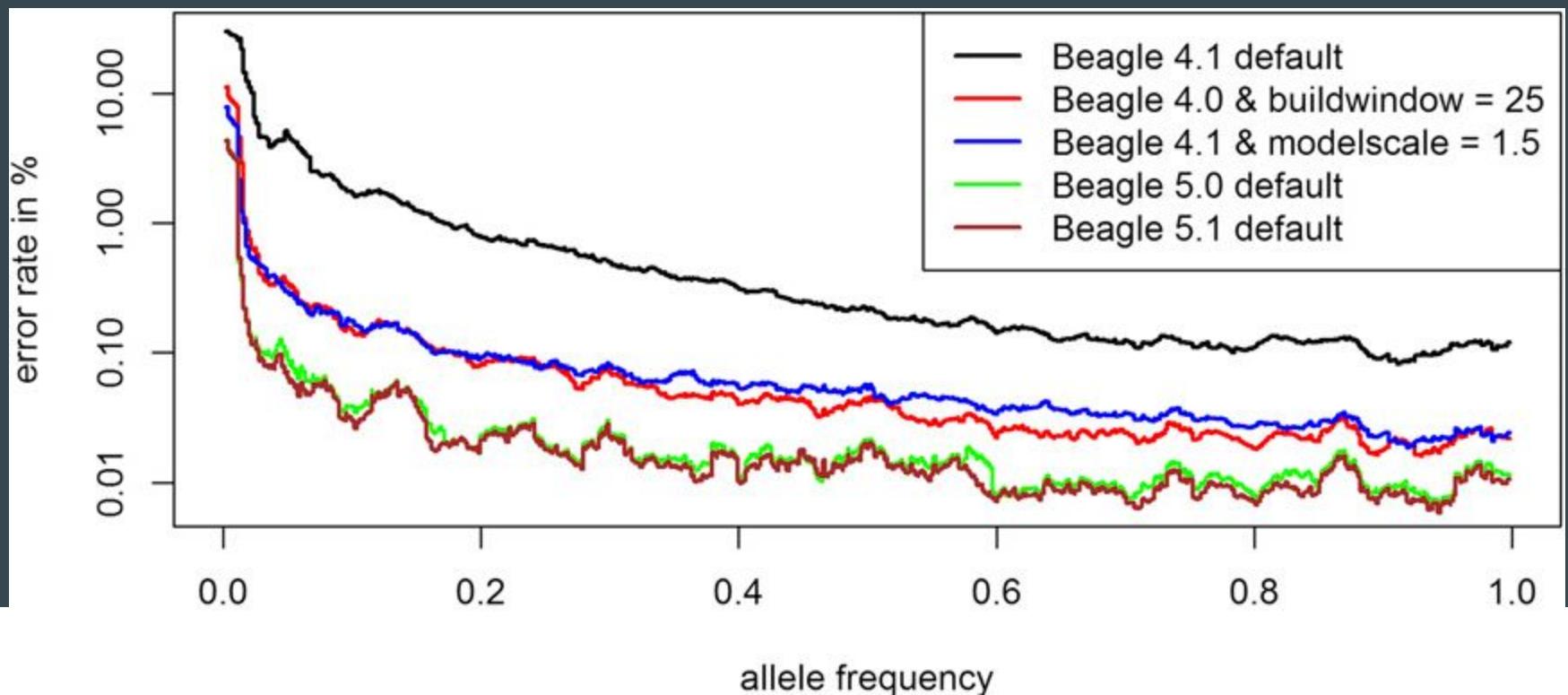
# Evaluate the imputation performance

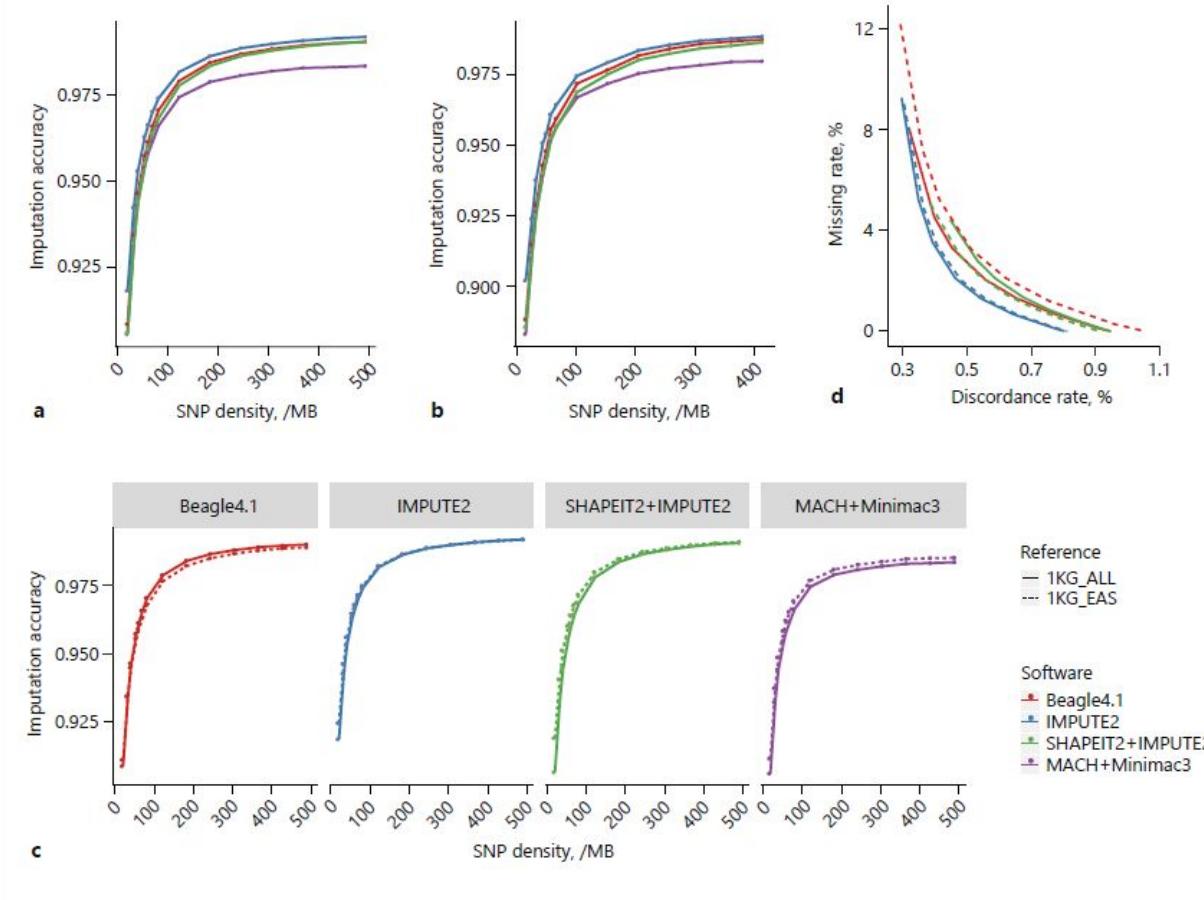
- $r^2$ : also known as the info parameter is the squared Pearson correlation coefficient between imputed genotype dosages (0-2) and masked sequence genotypes {0, 1, 2}:
  - The info metric is commonly used to remove poorly imputed SNPs from their association testing
  - There is no universal cutoff value for post-imputation SNP filtering:
    - Common cutoffs: 0.3 and 0.5
    - SNPs to impute HLA alleles → info threshold = 0.9
- Concordance rate: the percentage of correctly imputed genotypes of the test set

# Factors affecting the imputation performance

- Imputation strategy: software, phasing..

# Allele specific error rate depending on the allele frequency under different BEAGLE settings for the maize data





**Fig. 2. a, b** Imputation accuracy of the 112 kinds of strategies for chr1 and chr22. The x-axis represents SNP density. The y-axis represents the imputation accuracy, which is the rate of consistent sites between imputed genotypes and masked genotypes. **c** Imputation accuracy of Beagle4.1, IMPUTE2, MACH+Minimac3, and

SHAPEIT2+IMPUTE2 with two references (1KG\_ALL and 1KG\_EAS) for chr1. **d** Percentage discordance versus percentage missing genotypes for calling thresholds ranged from 0.33 to 0.99 for chr1.

Shi et al, 2017



# Factors affecting the imputation performance

- Imputation strategy: software, phasing..
- Reference panel: different strategies work better with different reference panels

# Reference population

It is important to consider the population similarity between the reference and to-be-imputed population

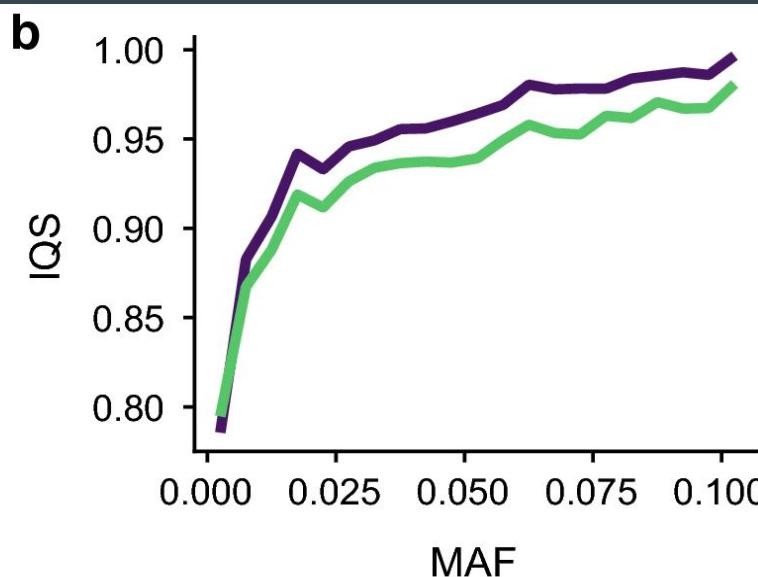
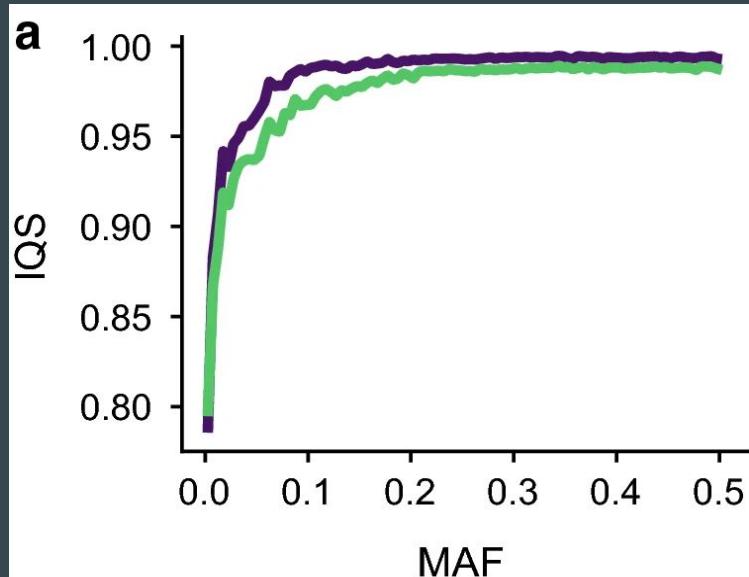
If no knowledge of the genetic structure, it is better to use all available individuals genotyped under high marker density for the reference panel (Pook et al, 2020)

In case the reference population has a lot of stratification, the design of a good reference panel is more difficult, as genetically distant individuals may introduce more noise than relevant information to the model (Pook et al, 2020)

If most of the genetic diversity of the study sample can be represented in a subset of the individuals in a reference panel, excluding genetically distant individuals improves imputation performance (Pook et al, 2020)

# Reference population

Composite reference population may increase imputation accuracy in diverse samples  
(beef cattle populations)



Rowan et al, 2020

Table 2 The most commonly used public reference panels to date

Reference panel	Number of reference samples	Number of sites (autosomes + X chromosome)	Average sequencing coverage	Ancestry distribution	Publicly available	Indels available	Reference
International HapMap Project phase 3	1,011	1.4 million	NA <sup>a</sup>	Multiethnic	Yes	No	47
1000G phase 1	1,092	28.9 million	2–6×	Multiethnic	Yes	Yes	1
1000G phase 3	2,504	81.7 million	7× genomes, 65× exomes	Multiethnic	Yes	Yes	3
UK10K Project	3,781	42.0 million	7× genomes, 80× exomes	European	Yes	Yes	89
HRC	32,470	40.4 million	4–8× <sup>b</sup>	Predominantly European <sup>c</sup>	Partially <sup>d</sup>	No	69
TOPMed	60,039	239.7 million	30×	Multiethnic	Partially <sup>e</sup>	Yes	71

Abbreviations: 1000G, 1000 Genomes Project; HRC, Haplotype Reference Consortium; indel, insertion or deletion; LOF, loss of function; NA, not applicable; TOPMed, Trans-Omics for Precision Medicine.

<sup>a</sup>The International HapMap Project phase 3 data were genotyped on the Illumina Human1M and Affymetrix 6.0 SNP arrays.

<sup>b</sup>The HRC panel was obtained by combining sequencing data across many low-coverage (4–8×) and a few high-coverage sequencing studies.

<sup>c</sup>The only non-European samples in the HRC panel are through the 1000G reference panel (which was a contributing study).

<sup>d</sup>Most of the HRC samples (~27,000) are available for download through controlled access from the European Genome-Phenome Archive.

<sup>e</sup>Some of the TOPMed samples (~18,000) are available for download through controlled access from the Database of Genotypes and Phenotypes (dbGaP).

# Reference panels - Examples

[http://gong\\_lab.hzau.edu.cn/Animal\\_ImputeDB/#/](http://gong_lab.hzau.edu.cn/Animal_ImputeDB/#/)

It is a public database with genomic reference panels of 13 animal species for online genotype imputation, genetic variant search, and free download

**Table 1.**

Data summary in Animal-ImputeDB

Species	No. of chromosome	No. of sample	Reference panel
			No. of SNPs
Ailuropoda melanoleuca (Giant panda)	28 354 scaffolds	34	4 671 936
Anas platyrhynchos (Duck)	30	106	12 682 400
Bos taurus (Cattle)	30	93	41 808 907
Bubalus bubalis (Swamp buffalo)	24	206	33 245 917
Canis familiaris (Dog)	39	658	61 065 811
Capra hircus (Goat)	30	233	29 889 815
Equus caballus (Horse)	32	53	19 257 635
Equus ferus (Tarpan)	32	19	7 809 754
Gallus gallus (Chicken)	35	103	26 864 273
Ovis aries (Sheep)	27	450	29 889 815
Sus scrofa (Pig)	19	233	40 323 709
Macaca mulatta (Monkey)	21	30	47 332 297
Oryctolagus cuniculus (Rabbit)	22	46	40 420 337

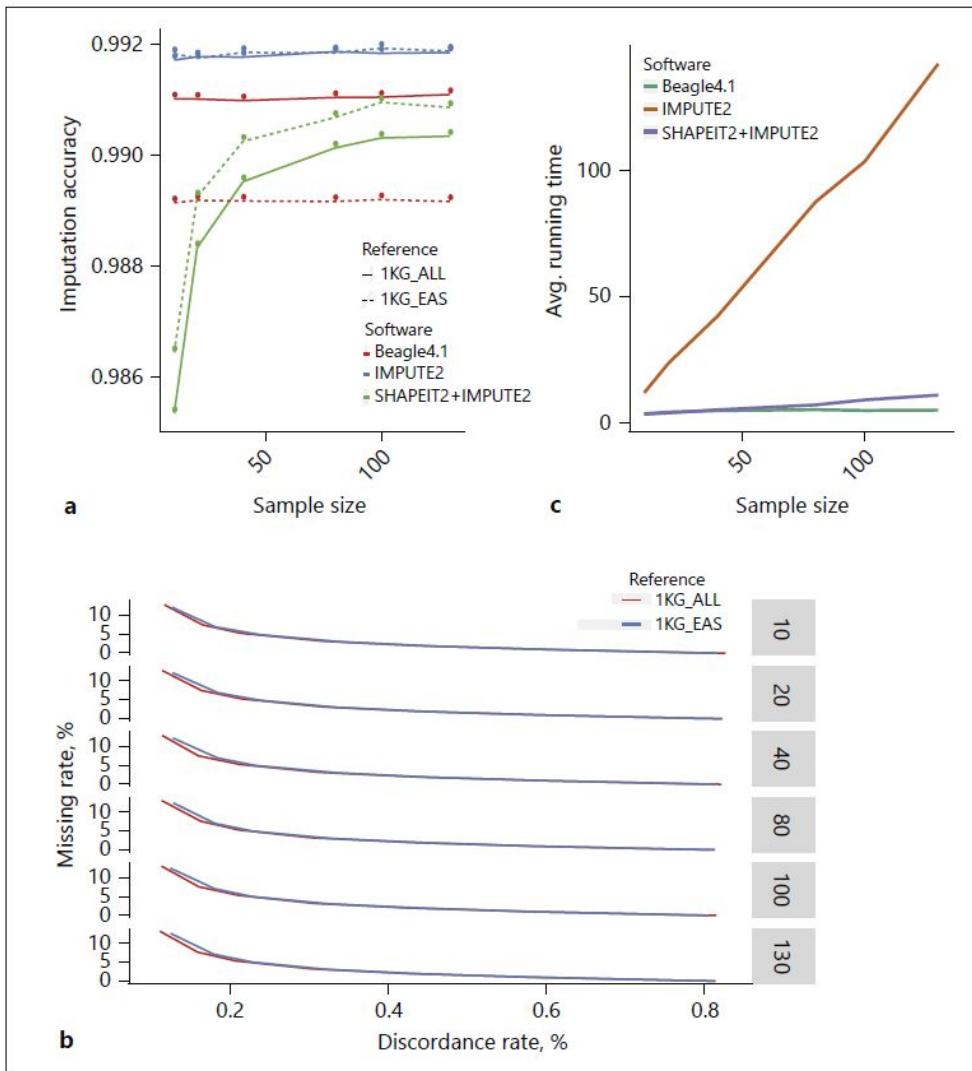
Yang et al, 2020

# Factors affecting the imputation performance

- Imputation strategy: software, phasing
- Reference panel

# Factors affecting the imputation performance

- Imputation strategy: software, phasing
- Reference panel
- Sample size

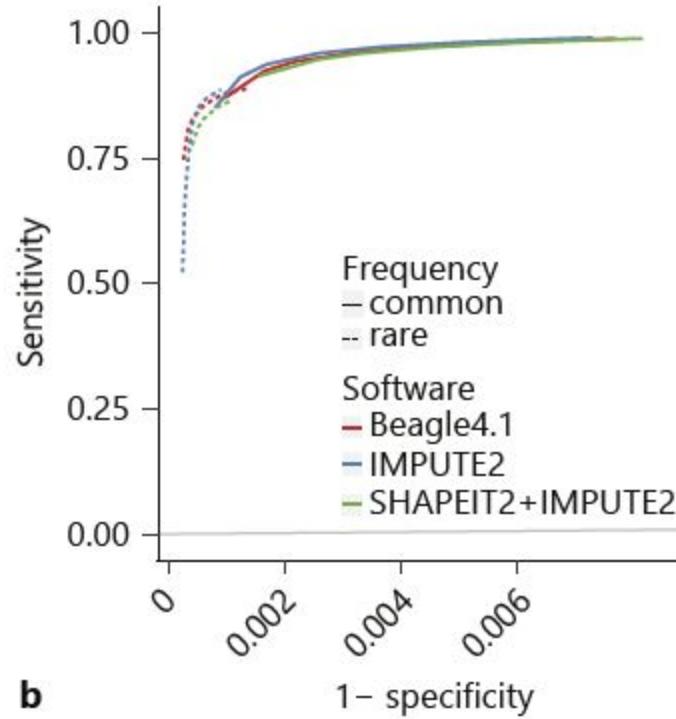
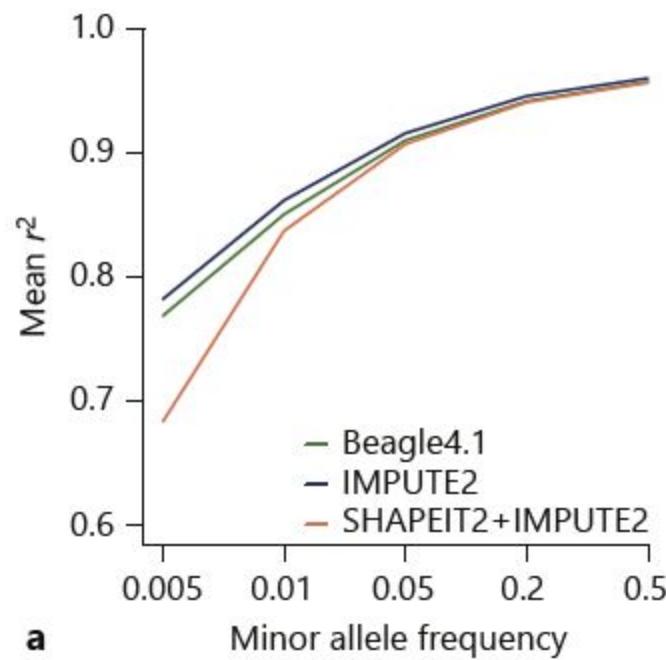


Shi et al, 2017

iction

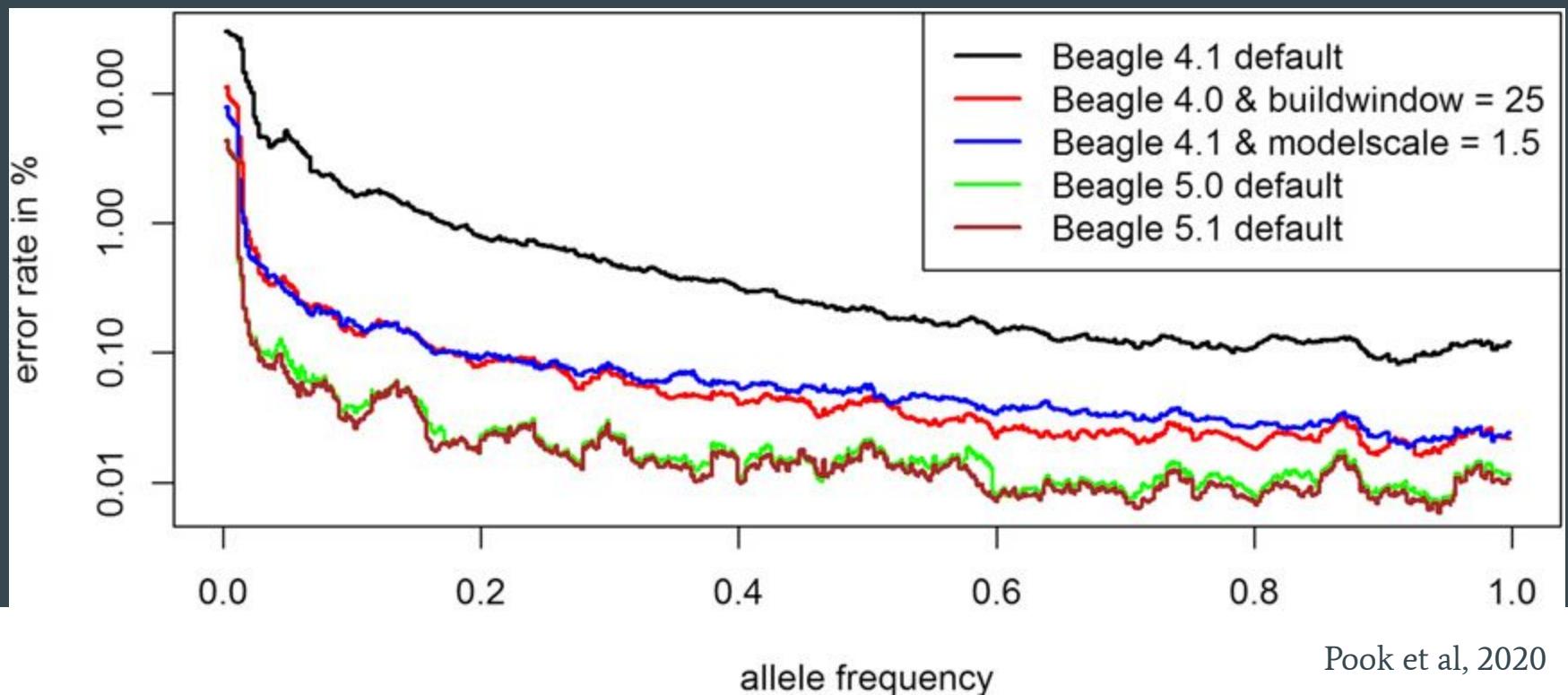
# Factors affecting the imputation performance

- Imputation strategy: software, phasing
- Reference panel
- Sample size
- Minor allele frequency of the SNPs



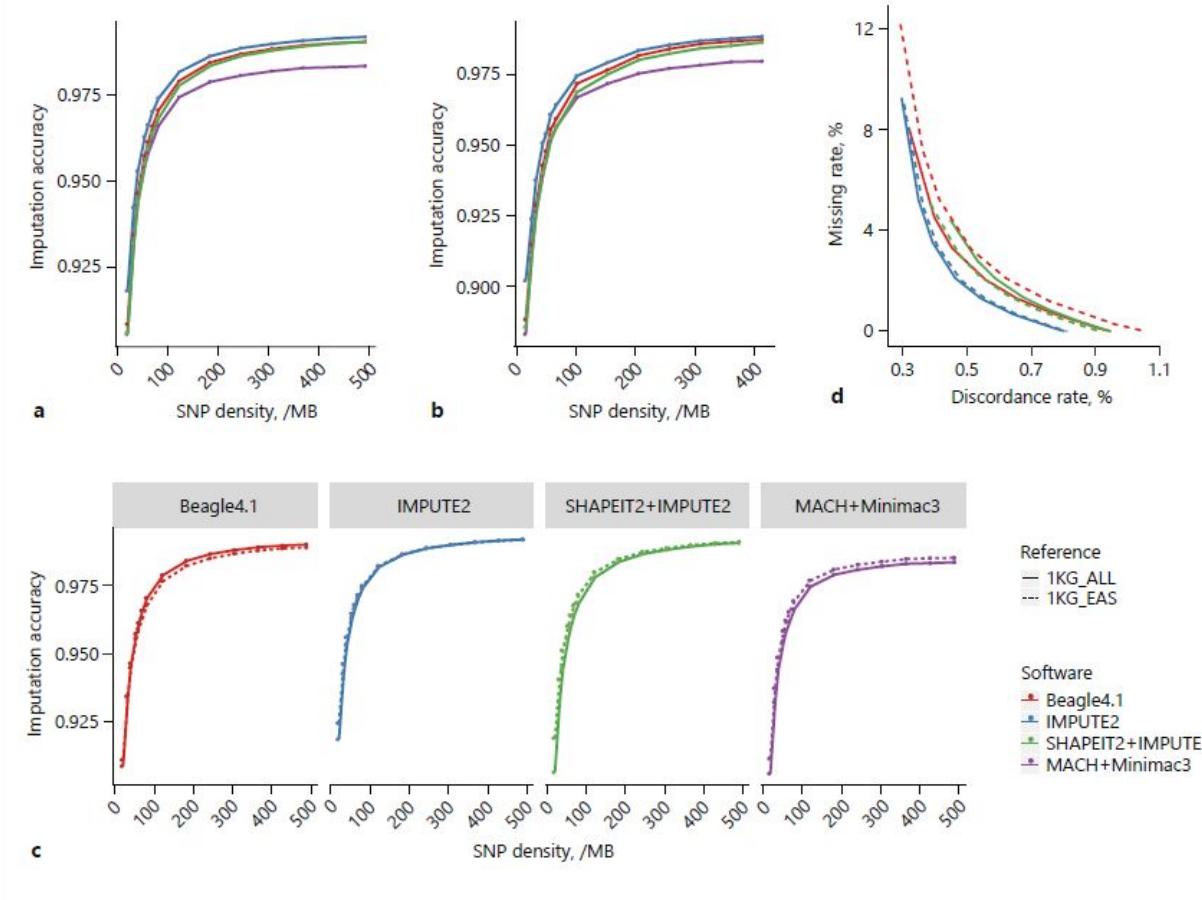
Shi et al, 2017

# Allele specific error rate depending on the allele frequency under different BEAGLE settings for the maize data



# Factors affecting the imputation performance

- Imputation strategy: software, phasing
- Reference panel
- Sample size
- Minor allele frequency of the SNPs
- SNP density/coverage

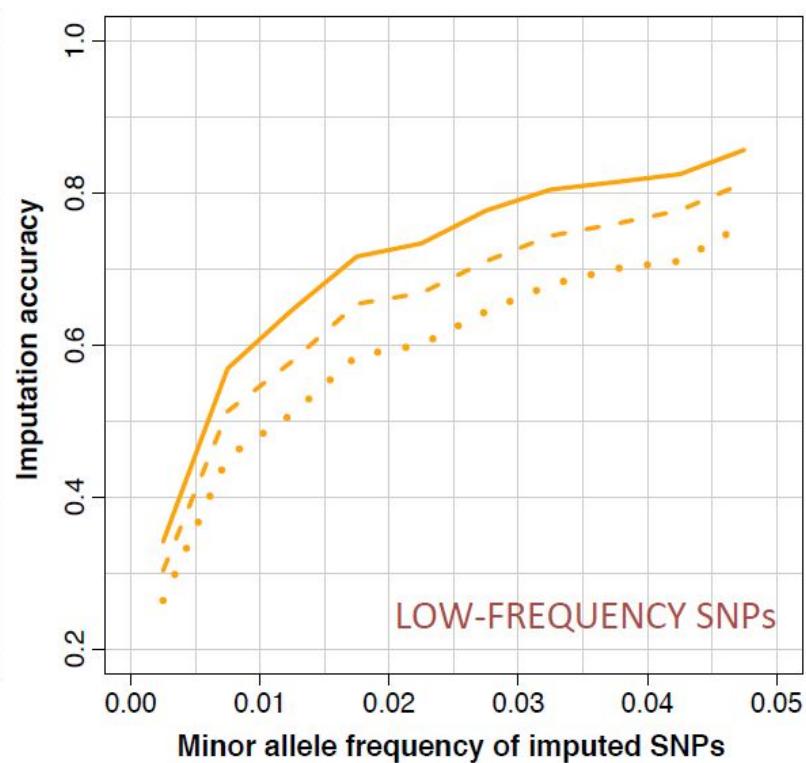
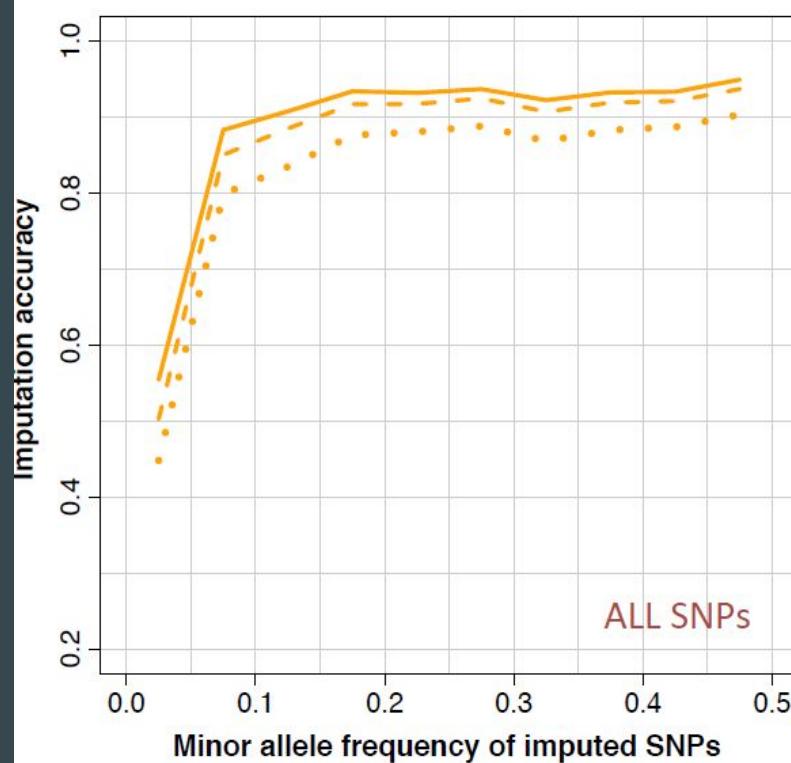


**Fig. 2. a, b** Imputation accuracy of the 112 kinds of strategies for chr1 and chr22. The x-axis represents SNP density. The y-axis represents the imputation accuracy, which is the rate of consistent sites between imputed genotypes and masked genotypes. **c** Imputation accuracy of Beagle4.1, IMPUTE2, MACH+Minimac3, and

SHAPEIT2+IMPUTE2 with two references (1KG\_ALL and 1KG\_EAS) for chr1. **d** Percentage discordance versus percentage missing genotypes for calling thresholds ranged from 0.33 to 0.99 for chr1.

Shi et al, 2017





— Omni 2.5M

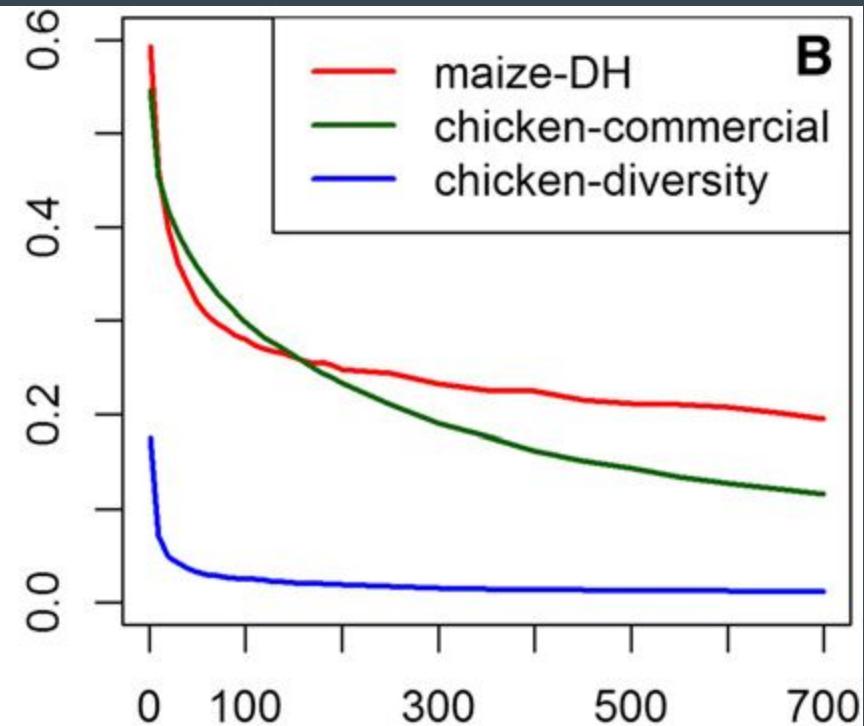
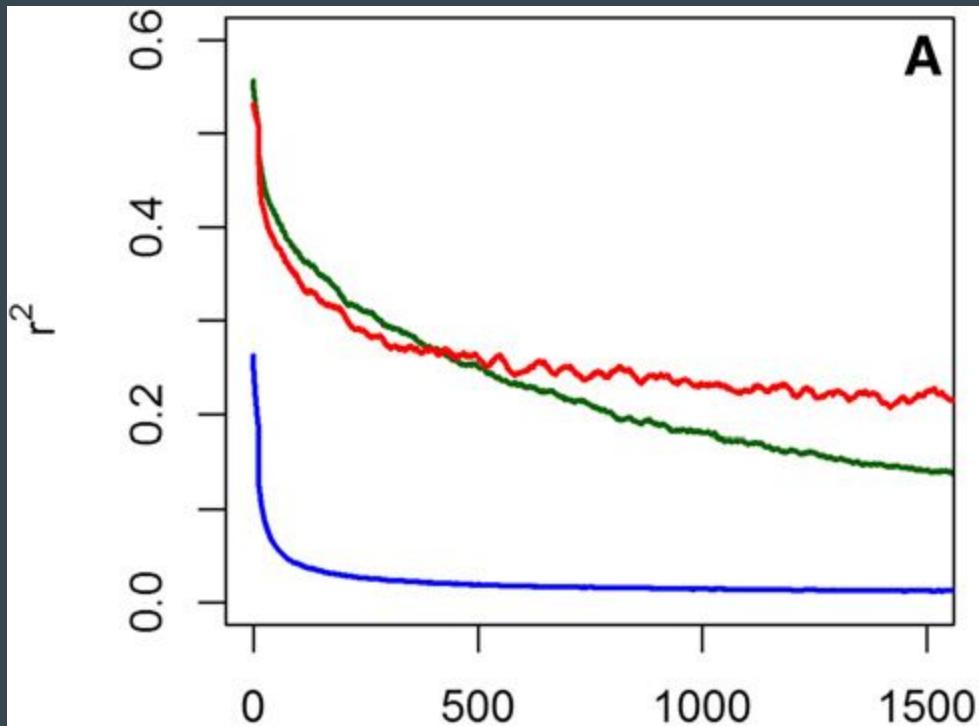
- - Illumina 550k

... Affymetrix 500k

Bryan Howie, ICHG 2011, Montreal

# Factors affecting the imputation performance

- Imputation strategy: software, phasing
- Reference panel
- Sample size
- Minor allele frequency of the SNPs
- SNP density/coverage
- LD pattern

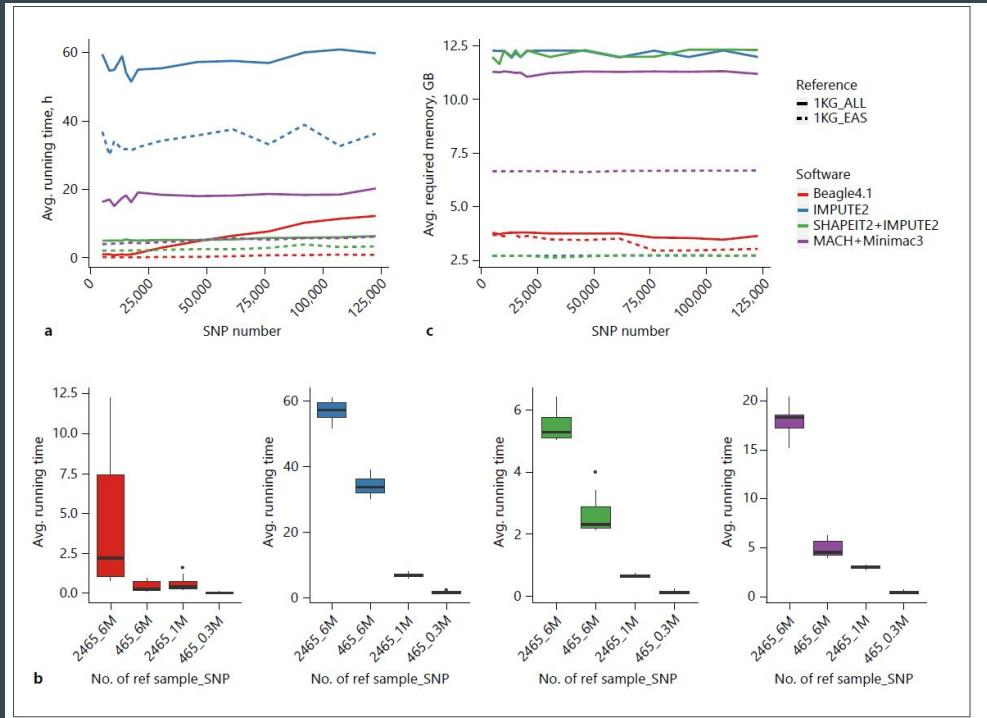


Pook et al, 2020

# Factors affecting the imputation performance

- Imputation strategy: software, phasing
- Reference panel
- Sample size
- Minor allele frequency of the SNPs
- SNP density/coverage
- LD pattern
- Computational burden

# Computational burden



**Fig. 3.** **a** Average running time of all 112 kinds of strategies for chr1. **b** Average running time of software on different references (sample number\_snp number). **c** Average computer memory of all 112 kinds of strategies for chr1.

- IMPUTE2 most time-consuming
- Except for Beagle, times affected by SNP number and size of reference (when reference size is larger than study size)

Shi et al, 2017

# Computational burden

Computational burden



Increasing the # of iterations to  
diminish the error rate

# Topics

Background

Imputation

Imputation  
performance

Imputation strategies

Why imputation  
is important?

Definition  
Advantages

Factors affecting  
the imputation  
performance

Tools  
Strategies