Imputation and it's importance in GWAS

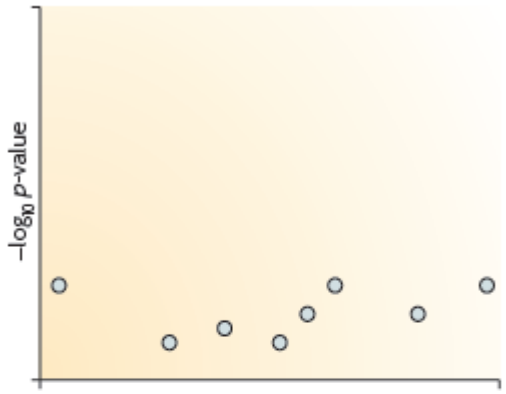
Dhriti

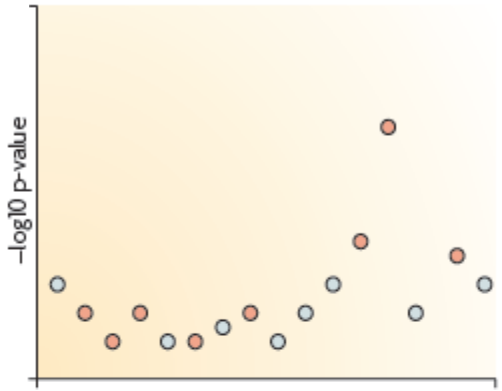
5th September 2018

Lecture 6

H3ABioNet 2018 Genotyping Chip Data Analysis and GWAS lecture series

The method of estimating genotypes or genotype probabilities at markers that have not been directly genotyped in a genetic study is known as ‘genotype imputation.





Imputation

Genotype data with missing data at untyped SNPs (grey question marks)

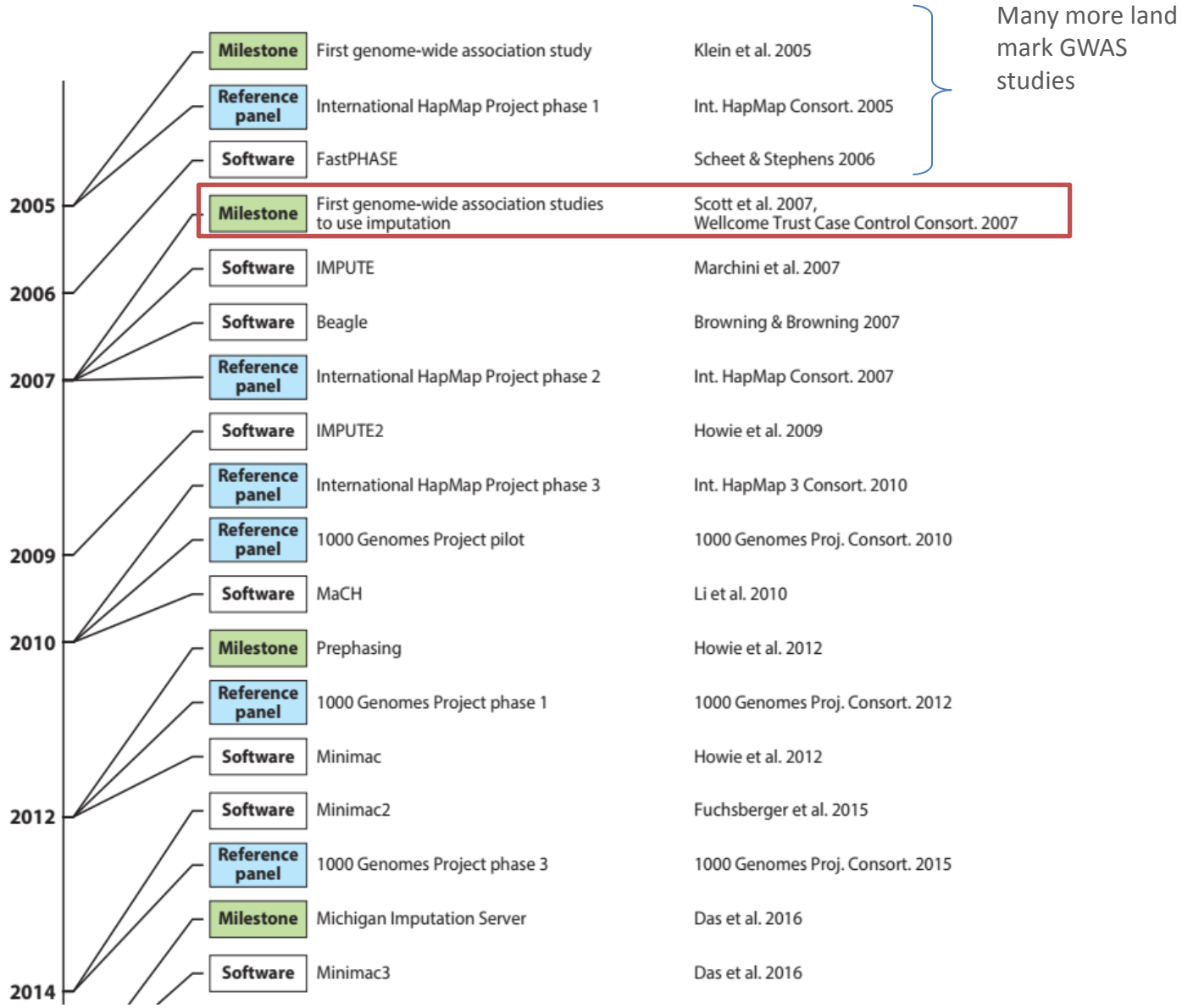
e The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

Reference Panel

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

1	1	1	1	1	2	1	0	0	2	2	0	2	2	2	0
0	0	1	0	2	2	2	0	0	2	2	2	2	2	2	0
1	1	1	1	2	2	2	0	0	2	1	1	2	2	2	0
1	1	2	0	2	2	1	0	1	2	2	1	2	2	2	0
2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	0
1	1	1	0	1	2	1	0	1	2	2	1	2	2	2	0
1	1	2	1	2	1	2	0	0	2	1	1	1	2	1	1
2	2	2	1	1	1	1	0	1	2	1	0	1	2	1	1
1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	0

Major Milestones..



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

First Imputation papers

A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants

Laura J. Scott,[1] Karen L. Mohlke,[2] Lori L. Bonnycastle,[3] Cristen J. Willer,[1] Yun Li,[1]

SCIENCE VOL 316 1 JUNE 2007

TECHNICAL REPORTS

nature genetics

A new multipoint method for genome-wide association studies by imputation of genotypes

Jonathan Marchini[1,2], Bryan Howie[1,2], Simon Myers[1], Gil McVean[1] & Peter Donnelly[1]

VOLUME 39 | NUMBER 7 | JULY 2007 NATURE GENETICS

Why do we perform Imputation ?

Fine-mapping

Imputation provides a higher resolution view of a genetic region by adding more variants, increasing the chances of identifying a causal variant.

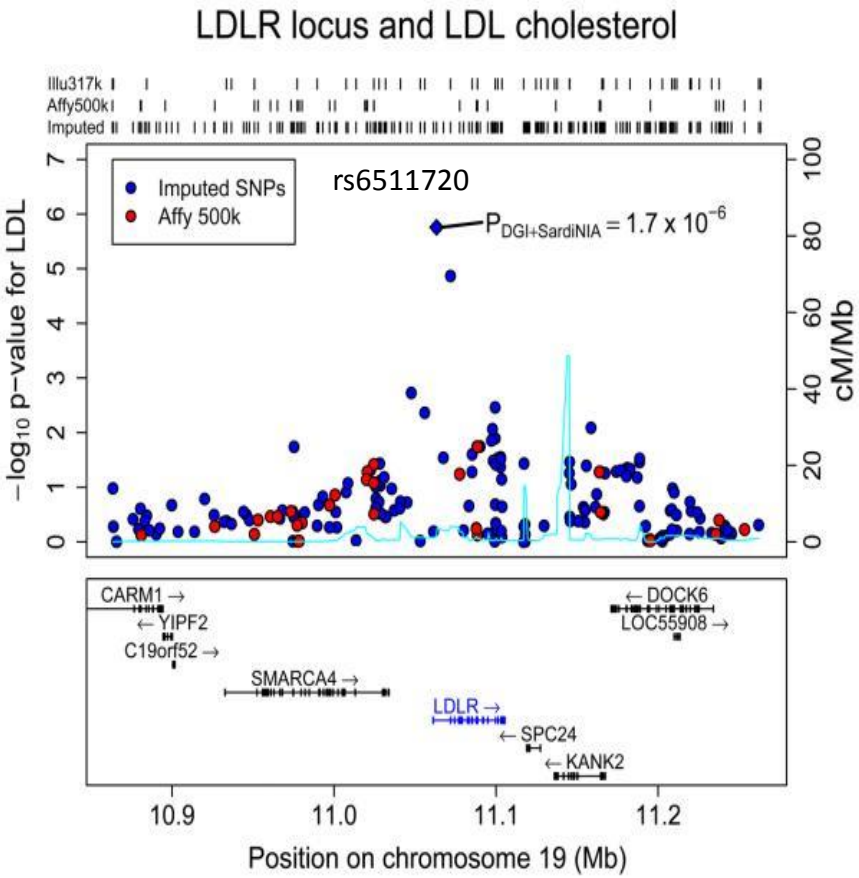
Large scale Meta-analysis

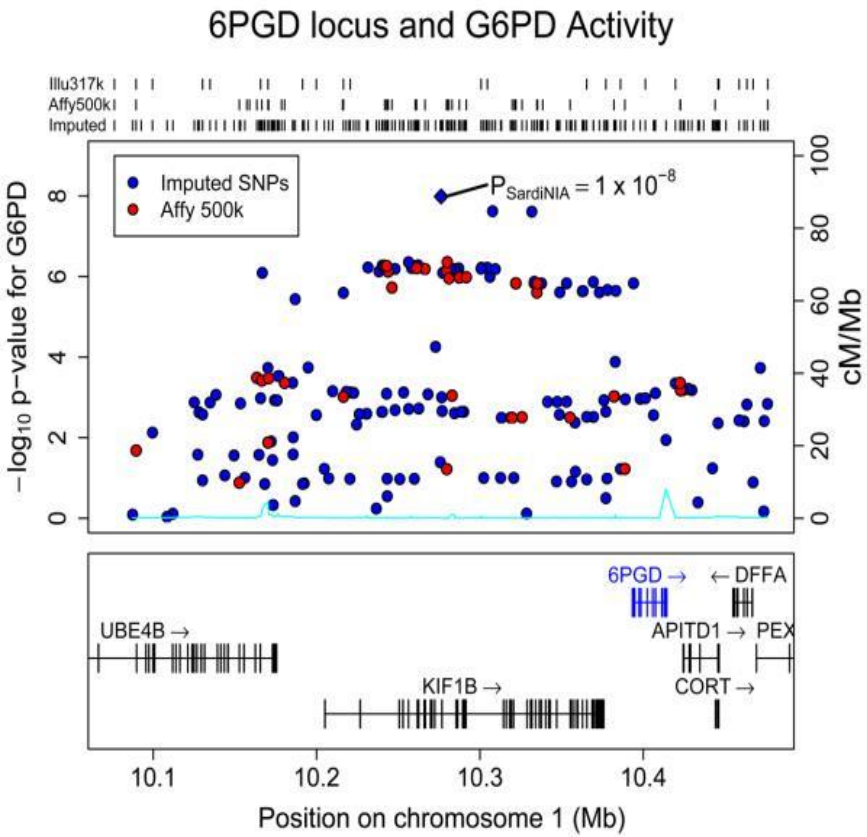
Imputation allows the combination of results across studies, generating a common set of variants which can then be analysed across all the studies to boost power.

Increased power of association

The reference panel is more likely to contain the causal variant than a GWAS array.

Success stories





In a study on triglycerides and cholesterol, where a common variant in a known risk gene (*LDLR*) was missed when only the genotyped SNPs were analysed but was then identified following imputation (Willer et al. 2008).

Although there is evidence for association in the region prior to imputation, the signal increases substantially, to reach genome wide significance, after imputation

Annu Rev Genomics Hum Genet. 2009; 10: 387–406.

Toolkit for imputation

Reference set of haplotypes, for example, HapMap

0	0	0	0	1	1	1	0	0	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	0

Genotype array data

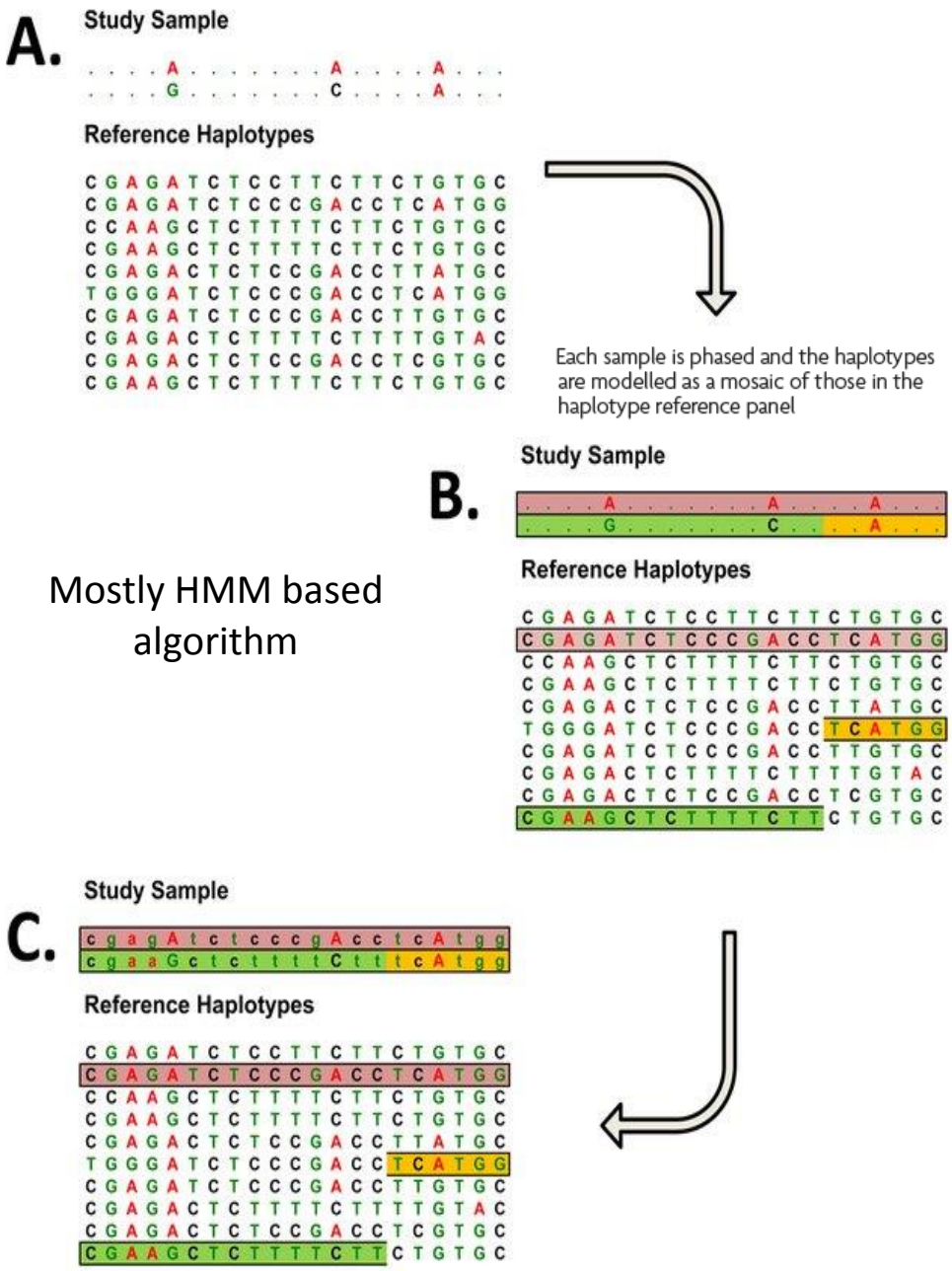
1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

0	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
1	?	?	?	1	?	1	?	0	1	1	?	?	1	?	0
⋮															
1	?	?	?	1	?	1	?	0	1	0	?	?	1	?	0
1	?	?	?	1	?	1	?	1	1	1	?	?	1	?	0
⋮															
1	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0
0	?	?	?	0	?	0	?	1	1	1	?	?	1	?	0

The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

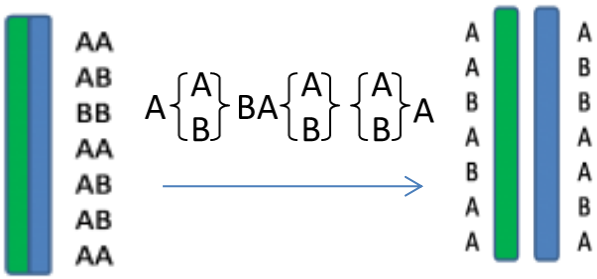
1	1	1	1	1	2	1	0	0	2	2	0	2	2	2	0
0	0	1	0	2	2	2	0	0	2	2	2	2	2	2	0
1	1	1	1	2	2	2	0	0	2	1	1	2	2	2	0
1	1	2	0	2	2	1	0	1	2	2	1	2	2	2	0
2	2	2	2	2	1	2	0	1	2	1	1	2	2	2	0
1	1	1	0	1	2	1	0	1	2	2	1	2	2	2	0
1	1	2	1	2	1	2	0	0	2	1	1	1	2	1	1
2	2	2	1	1	1	1	0	1	2	1	0	1	2	1	1
1	2	2	0	0	2	0	0	2	2	2	1	2	2	2	0

(Marchini & Howie 2010)



Mostly HMM based algorithm

- **Pre-phasing** (Haplotype estimation) of the genotypes in the study sample



Unphased data (Genotype)

Phased data (Haplotype)

Eagle2/Shapit2

- **Imputation** Study sample haplotypes are modelled as a mosaic of those on the haplotype reference panel .

Annu Rev Genomics Hum Genet. 2009; 10: 387–406

Reference Panels

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

Larger reference panels =>Detailed catalogue of genetic variants => better imputation accuracy => Improves the power of downstream association analyses, especially for rare variants.

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

Software

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

A practical guide to Imputing a chip-based data

Step 1: Data Preparation

- The GWA data should be converted to VCF or PLINK format

- Remove samples with:
  - Excessive missingness (>5%)
  - Reported vs. genotyped sex-mismatch
  - Unusual high/low heterozygosity
  - Check for ancestry outliers (PCA/MDS) or duplicate samples

- Exclude SNPs with:
  - Excessive missingness (>5%) and low MAF (<1%)
  - HWE violations (~$P<10^{-4}$)
  - Duplicate chromosomal positions

Data Preparation continued

- SNP positions should be aligned to GRCh37 (http://genome.ucsc.edu/cgi-bin/hgLiftOver)
- REF allele should be matching to GRCh37 (plink commands like **--a2-allele** to set reference alleles)
- Careful about the PLINK major minor allele swap (plink command –keep-allele-order prevents that)
- Align the genotypes to the same strand as the reference panel (generally the forward strand)). Check allele frequency of the strand ambiguous SNPs or drop these SNPs and re-impute them

Resource for existing chip

Strand Home | Source Strand Files | Ilmn Strand Files | AB to TOP | Ref/Alt | Other Tools

Genotyping chips strand and build files

This page contains files giving the strand orientation and position of variants of the most common genotyping chips on genome builds 36 to 38 inclusive. New files will be added as new genome builds are released

Note: Genome build 35 links have now been removed from all the pages. The files are though still available and are still created by default for all new chips, in the event you require one of these files please email me

If you have a chip and/or build that is not listed here please contact me, Will Rayner (**wrayner** at **well** dot **ox** dot **ac** dot **uk**) and I can create and post the file to this page.

http://www.well.ox.ac.uk/~wrayner/strand/index.html

Illumina data files

The data for each chip and genome build combination are freely downloadable from the links below, each zip file contains three files, these are:
.strand file
.miss file
.multiple file

.strand file This contains all the variants where the match to the relevant genomic sequence >90%. The strand file contains six columns, SNP id, chromosome, position, %match to genome, strand and TOP alleles. The SNP ids used are those from the Illumina annotation file and so are not necessarily the latest ones for that position from dbSNP. The alleles listed are the Illumina TOP alleles, if you are in any doubt whether your data file can be used with these strand files a check of your non A/T G/C SNPs alleles vs the strand file should confirm this for you. If there are differences then it is likely your genotype file has been created by exporting on a different strand, such as ILMN or SOURCE in this case please check the links in the menu above to those pages. If that is still unsuccessful then if you can provide a list of the SNP ids and their alleles on the chip (a plink .bim file is ideal for this) it is likely a strand file can

Step 2: Pre-phasing your data

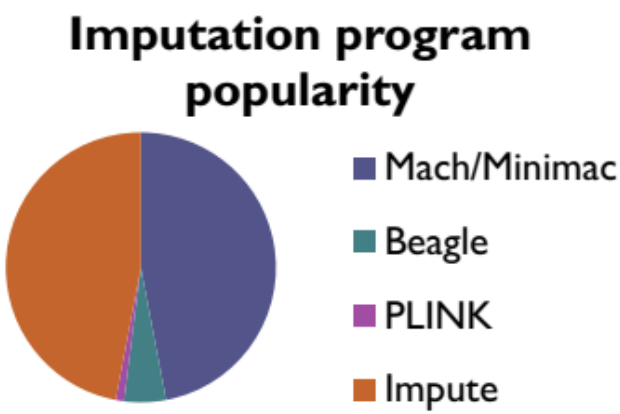
- Most commonly used tools for phasing are : Eagle2 and Shapit2

- Phasing can be done with or without a Reference panel. If the dataset to be imputed is small, it is recommended to phase using Reference Panel

- Command

```
./eagle  --vcfRef 1000GP.vcf.gz
--vcfTarget  gwas.vcf.gz
--geneticMapFile genetic_map_b37.txt
--chrom 20
--outPrefix gwas_chr2.phased
```

Step 3: Impute your data



- Reference panels
  - Hapmap
  - KGP-phase 3
  - HRC
  - CAPPA

File formats

Software	**Imputing**	*Minimac*	*Minimac3*	*IMPUTE2*
File format	Input	Mach/ HAPS	VCF	HAPS
	Output	DOSE	VCF/ DOSE	GEN

Minimac3

Download your Reference Panel

Reference Panels for Download

Some commonly used reference panels are available for download here:

Chr X Haplotypes for 1000 Genomes Phase 3 have been updated on Oct 20 to include multi-allelic variants as well (split as bi-allelic variants) !!!

Reference Panel	Number of Samples	File Format	Parameter Estimates Available	Chromosomes	Link
1000 Genomes Phase 3 (version 5)	2,504	VCF	-	1-22,X	Download
		M3VCF	YES	1-22,X	Download
			NO	1-22,X	Download
		VCF,M3VCF	YES	X	Download
1000 Genomes Phase 1 (version 3)	1,092	VCF	-	1-22,X	Download
			YES	1-22,X	Download
			NO	1-22,X	Download

MiniMac3 provides Reference panels in a custom format (m3vcf) that can handle very large references with lower memory

IMPUTE2

- Home
- Getting Started
- What's New?
- Download IMPUTE2
- Reference Data
- Examples
- Program Options
- Best Practices
- Pre-phasing
- Whole Chromosomes
- Merging Panels
- Concordance Tables
- Scripts
- FAQ
- References
- Contributors
- Mail List

Download Reference Data

IMPUTE2 can use publicly available reference datasets, such as haplotypes from major sequencing projects, a well as customized reference panels, such as SNP genotypes from a fine-mapping study. If you would like to download a public dataset, just click the relevant link below, which will take you to a page with background information and download options for that dataset.

Link to download page	NCBI build	Haplotype release date	Release status
1000 Genomes Phase 3	b37	October 2014	
1000 Genomes Phase I integrated haplotypes (produced using SHAPEIT2)	b37	June 2014	
1000 Genomes Phase I integrated haplotypes (produced using SHAPEIT2)	b37	Dec 2013	
1000 Genomes Phase I integrated haplotypes (produced using SHAPEIT2)	b37	Sep 2013	
1000 Genomes Phase I integrated variant set	b37	Mar 2012	Includes chrX; updated 24 Aug 2012

Impute 2 provides its own scripts to convert a phased VCF file into reference panel format: one legend file and one haplotypes file

Input formats

VCF Formats

Comments

```
##fileformat=VCFv4.2
##fileDate=20160525
##source=PLINKv1.90
##contig=<ID=21,length=48100156>
##INFO=<ID=PR,Number=0,Type=Flag,Description="Provisional reference allele, may not be based on real reference genome">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	sample1	sample2	sample3	sample4	sample5	sample6
21	10699687		SNP21-9721558	T	C	.	.	PR	GT	0/0	0/0	0/0	0/0	
21	10823704		SNP21-9845575	G	A	.	.	PR	GT	0/0	0/0	0/0	0/0	
21	10824667		SNP21-9846538	G	T	.	.	PR	GT	0/0	0/0	0/0	0/1	
21	10832325		SNP21-9854196	C	A	.	.	PR	GT	1/0	0/0	0/0	0/0	
21	10849744		SNP21-9871615	G	T	.	.	PR	GT	1/1	0/0	0/0	0/0	
21	10863251		SNP21-9885122	G	T	.	.	PR	GT	0/0	0/0	0/0	0/0	

Info

Genoytpes

.Gen format (unphased)

```
SNP 1 : AA AA
SNP 2 : GG GT
SNP 3 : CC CT
SNP 4 : CT CT
SNP 5 : AG GG
```

Ind1 Ind2

```
SNP1 rs1 1000 A C 1 0 0 1 0 0
SNP2 rs2 2000 G T 1 0 0 0 1 0
SNP3 rs3 3000 C T 1 0 0 0 1 0
SNP4 rs4 4000 C T 0 1 0 0 1 0
SNP5 rs5 5000 A G 0 1 0 0 0 1
```

SNP RSID BP a0 a1 Ind1 Ind2

.Hap format (phased)

```
1 0 0 1 0 0 1 1 1 1
1 0 0 1 0 0 1 1 1 1
0 0 0 0 0 0 1 0 0 0
0 1 0 1 0 0 1 1 0 0
0 1 1 1 0 1 1 1 0 1
```

Basic commands for imputation

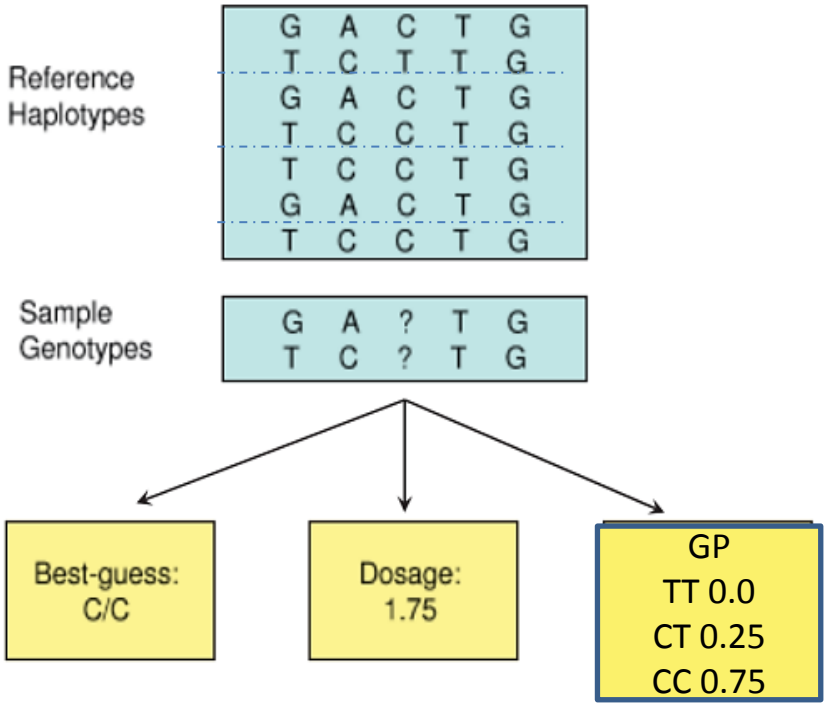
Imputing in Minimac3

```
./Minimac3  --refHaps HRC.r1-1.GRCh37.chr20.m3vcf.gz \
--haps Chr20.Phased.phased.vcf --prefix  Chr20.imputed.output \
--format GT,DS,GP –allTypedSites
```

Imputing in Impute2

```
./impute2  -m chr22.map  -h chr22.1kG.haps   -l chr22.1kG.legend   \
-g chr22.study.gens   -strand_g chr22.study.strand   -int 20.4e6 20.5e6 \
 -Ne 20000   -o chr22.one.phased.impute2
```

Terms you will come across again and again..



Recode the three genotype probabilities from any imputation tool into a single allelic dosage value with this basic equation:

[0 * p(AA)] + [1 * p(AB)] + [2 * p(BB)]
which simplifies to: p(AB) + [2 * p(BB)]

The imputed allelic dosage for SNP3 is

0*TT + 1*CT + 2*CC = 0.25 + 2*0.75 =1.75

Assessing imputation quality
Gold standard is to compare with true genotype. In absence of that, a parameter $\mathbf{r^2}$ can be estimated on the basis of posterior probabilities.

Minimac3 outputs

Info file

SNP	REF(0)	ALT(1)	ALT_Frq	MAF	AvgCall	Rsq	Genotyped	LooRsq	EmpR	EmpRsq	Dose0	Dose1
1:1005723	C	T	0.00024	0.00024	0.99976	0.00509	Imputed	-	-	-	-	-
1:1005741	G	A	0.00002	0.00002	0.99998	0.00012	Imputed	-	-	-	-	-
1:1005806	C	T	0.14489	0.14489	0.99973	0.99784	Genotyped	0.568	0.847	0.71745	0.80011	0.08737
1:1006223	G	A	0.58207	0.41793	0.94394	0.80402	Imputed	-	-	-	-	-
1:1007222	G	T	0.14226	0.14226	0.99074	0.93284	Imputed	-	-	-	-	-
1:1018598	A	G	0.054	0.054	0.97272	0.61048	Imputed	-	-	-	-	-

Rsq

This is the estimated value of the squared correlation between imputed genotypes and true, unobserved genotypes. An measure of the confidence in the imputed dosages

$$\hat{r}^2 = \frac{\frac{1}{2n} \times \sum_{i=1}^{2n} (D_i - \hat{p})^2}{\hat{p}(1-\hat{p})}$$

LooRsq This statistic can only be provided for genotyped sites. This is similar to the estimated Rsq above, but the imputed dosages value used to compare are calculated by hiding all known genotypes for the given SNP.

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
```

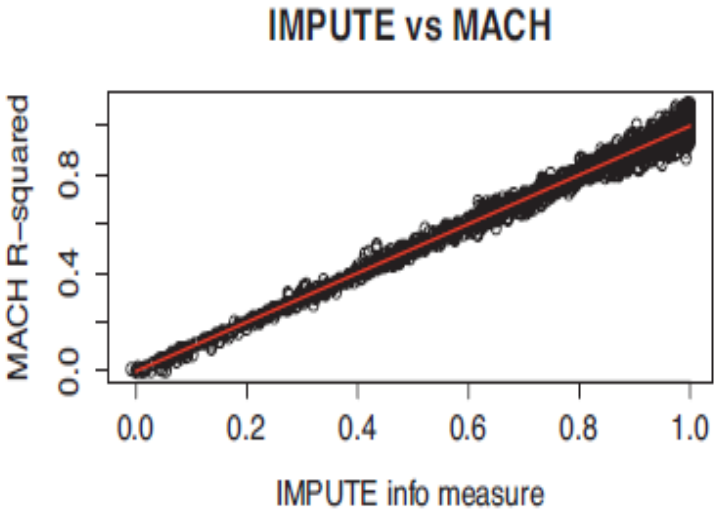
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	1005723	1:1005723	C	T	.	PASS	AF=0.00024;MAF=0.00024;R2=0.00509
1	1005741	1:1005741	G	A	.	PASS	AF=2e-05;MAF=2e-05;R2=0.00012
1	1005806	1:1005806	C	T	.	PASS;GENOTYPED	AF=0.14489;MAF=0.14489;R2=0.99784;ER2=0.71745
1	1006223	1:1006223	G	A	.	PASS	AF=0.58207;MAF=0.41793;R2=0.80402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

r^2 and info score

- In general fairly close correlation
  - rsq/ Info/ allelic Rsq
    - 1 = no uncertainty
    - 0 = complete uncertainty
    - 0.8 on 1000 individuals = amount of data at the SNP is equivalent to a set of perfectly observed genotype data in a sample size of 800 individuals
      - Note Mach uses an empirical Rsq (observed var/exp var) and can go above 1

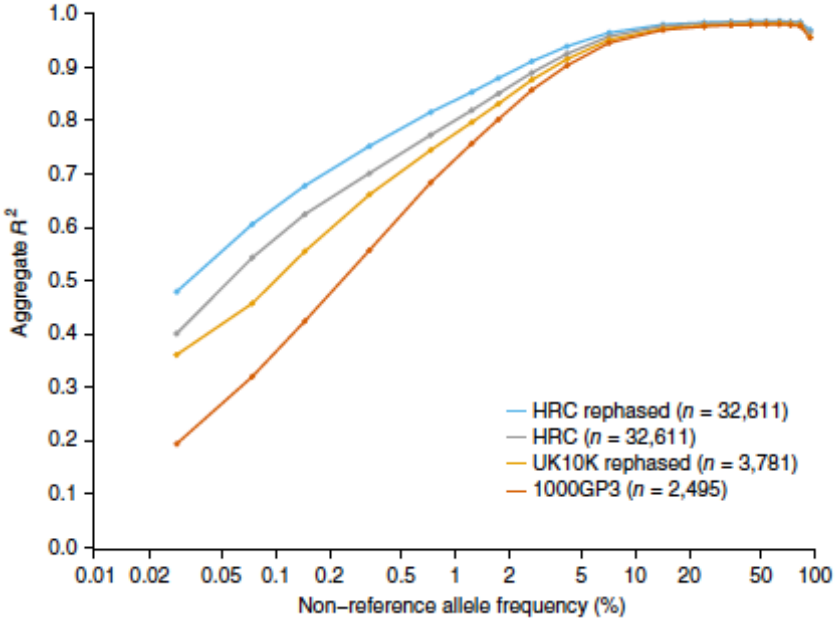


Imputation evaluation

Imputation performance

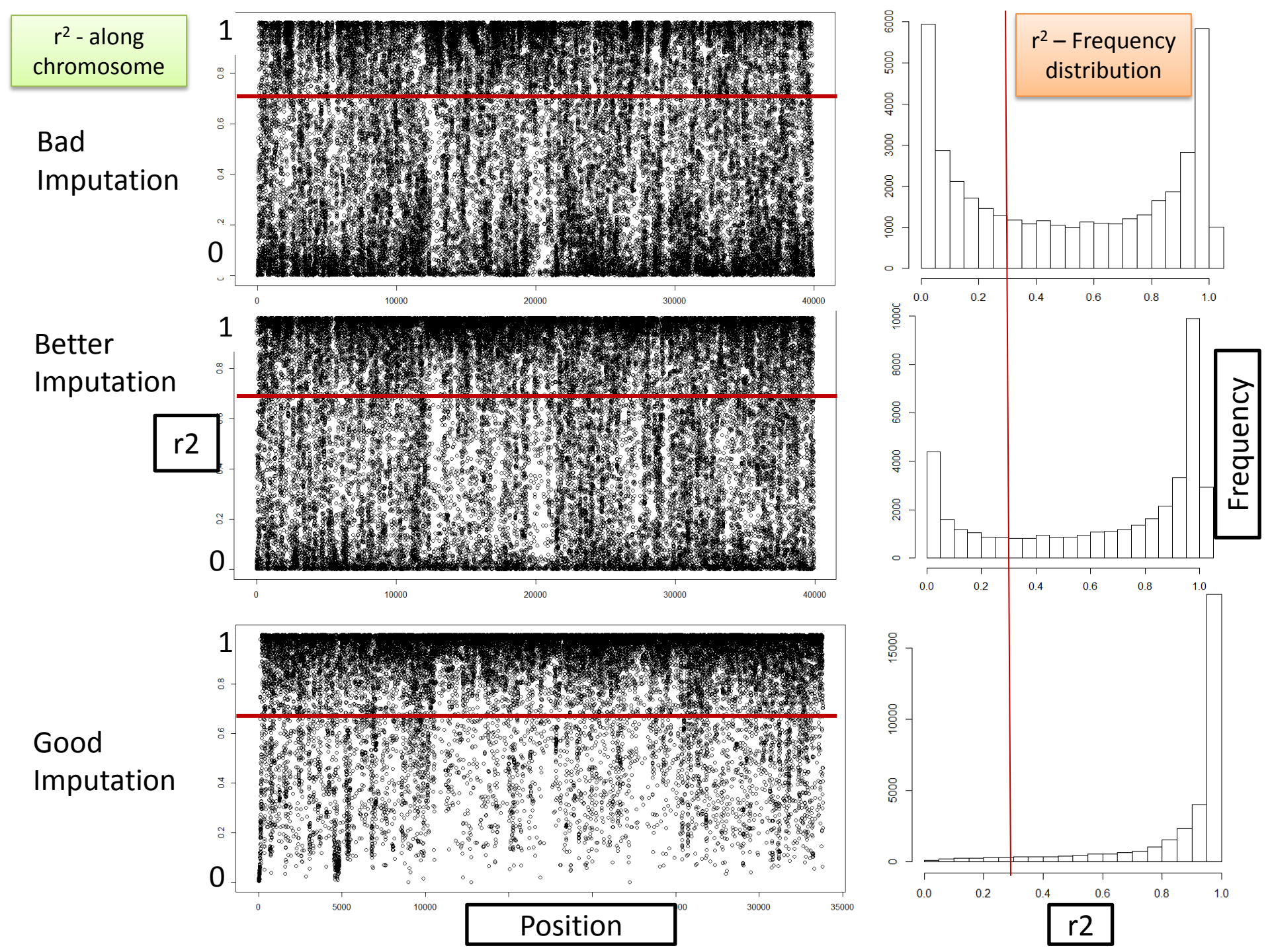
1. Number of imputed SNPs
2. Number of imputed SNPs in MAF bins
3. Number of imputed SNPs with good imputation score (~r^2 >0.8)

4. Aggregate R^2 per allele frequency bins

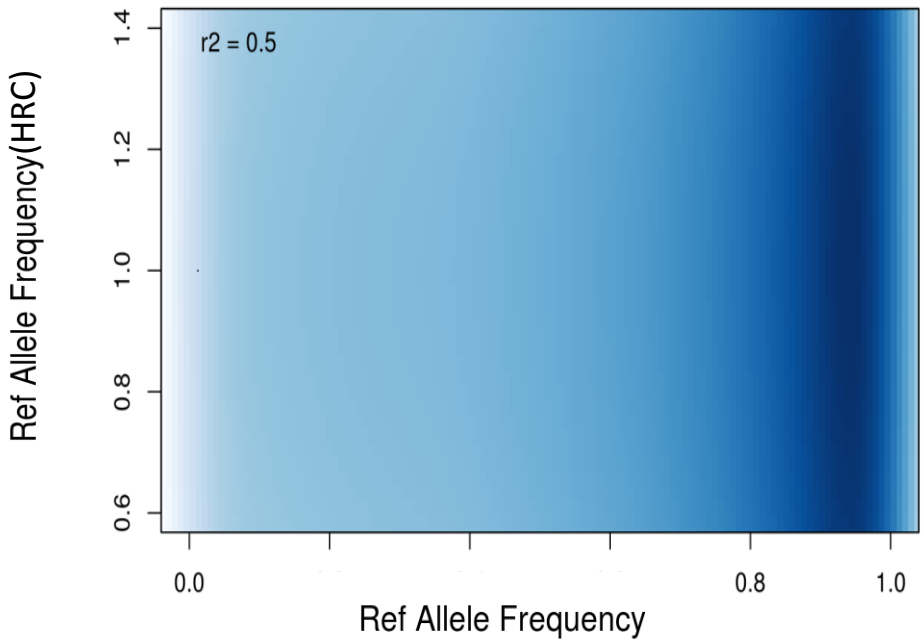


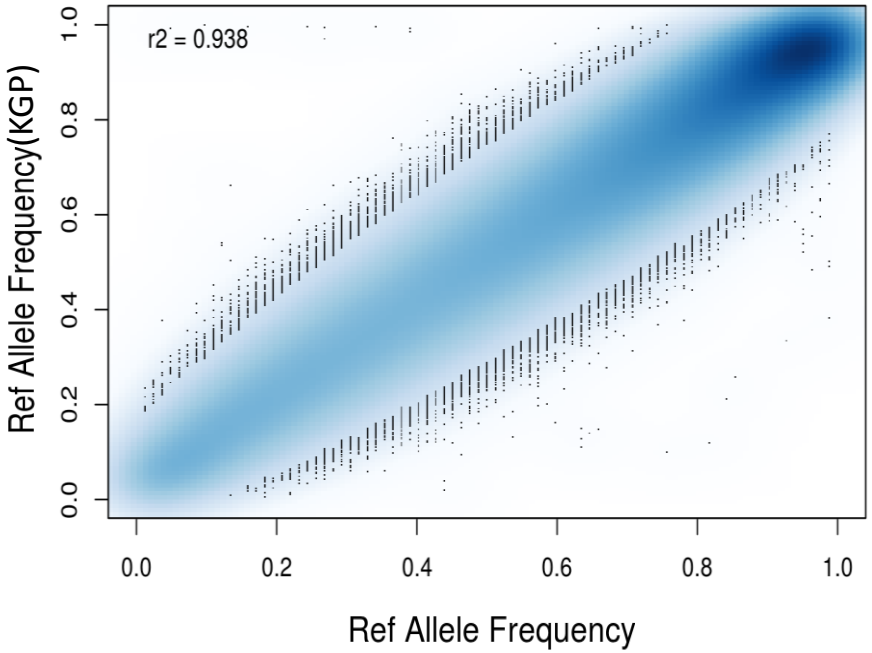
Filter SNPs with low R^2

```
bcftools view -i 'R2>0.6 & MAF>.05' -Oz
chr1.dose.vcf.gz > chr1.filtered.vcf.gz
```


r2 - along chromosome
Bad Imputation
Better Imputation
Good Imputation
r2
1
0
0
10000
20000
30000
40000
5000
35000
Position
r2 – Frequency distribution
Frequency
r2
0.0
0.2
0.4
0.6
0.8
1.0

Plot MAF-reference MAF





Bad imputation

Good imputation

Imputation Concordance Tables -IMPUTE 2

In the current analysis, IMPUTE2 masked, imputed, and evaluated 173430 genotypes that were called with high confidence (maximum probability >= 0.90) in the Panel 2 input file (-g or -known_haps_g).

When the masked study genotypes were imputed with reference data from Panel 0, the concordance between original and imputed genotypes was as follows:

Interval	#Genotypes	%Concordance
[0.0-0.1]	0	0.0
[0.1-0.2]	0	0.0
[0.2-0.3]	0	0.0
[0.3-0.4]	0	0.0
[0.4-0.5]	13	23.1
[0.5-0.6]	262	61.5
[0.6-0.7]	294	67.3
[0.7-0.8]	408	82.4
[0.8-0.9]	742	86.7
[0.9-1.0]	171711	99.8

Interval	%Called	%Concordance
[>= 0.0]	100.0	99.5
[>= 0.1]	100.0	99.5
[>= 0.2]	100.0	99.5
[>= 0.3]	100.0	99.5
[>= 0.4]	100.0	99.5
[>= 0.5]	100.0	99.5
[>= 0.6]	99.8	99.6
[>= 0.7]	99.7	99.7
[>= 0.8]	99.4	99.7
[>= 0.9]	99.0	99.8

Comparison of minimac3, minimac2, IMPUte2, and Beagle 4.1

Reference panel	Number of samples	minimac3	minimac2	IMPUTE2	Beagle 4.1
			Time (in CPU-hours)		
1000G Phase 1	1,092	**4**	27	34	5
AMD	2,074	**9**	59	73.5	9
1000G Phase 3	2,504	**6**	61	78	9
SardiNIA	3,489	**7**	85	108	11
COMBINED	9,341	**17**	236	288	31
Mega	11,845	**21**	304	364	40
HRC v1.1	32,390	**31**	925	951	128
			Memory (in CPU-GB)		
1000G Phase 1	1,092	**0.09**	0.34	0.91	0.51
AMD	2,074	**0.14**	0.62	1.58	0.39
1000G Phase 3	2,504	**0.13**	0.75	1.88	0.56
SardiNIA	3,489	**0.13**	1.03	2.55	0.46
COMBINED	9,341	**0.28**	2.73	6.57	0.41
Mega	11,845	**0.33**	3.51	8.28	0.43
HRC v1.1	32,390	**0.55**	9.31	22.08	1.98

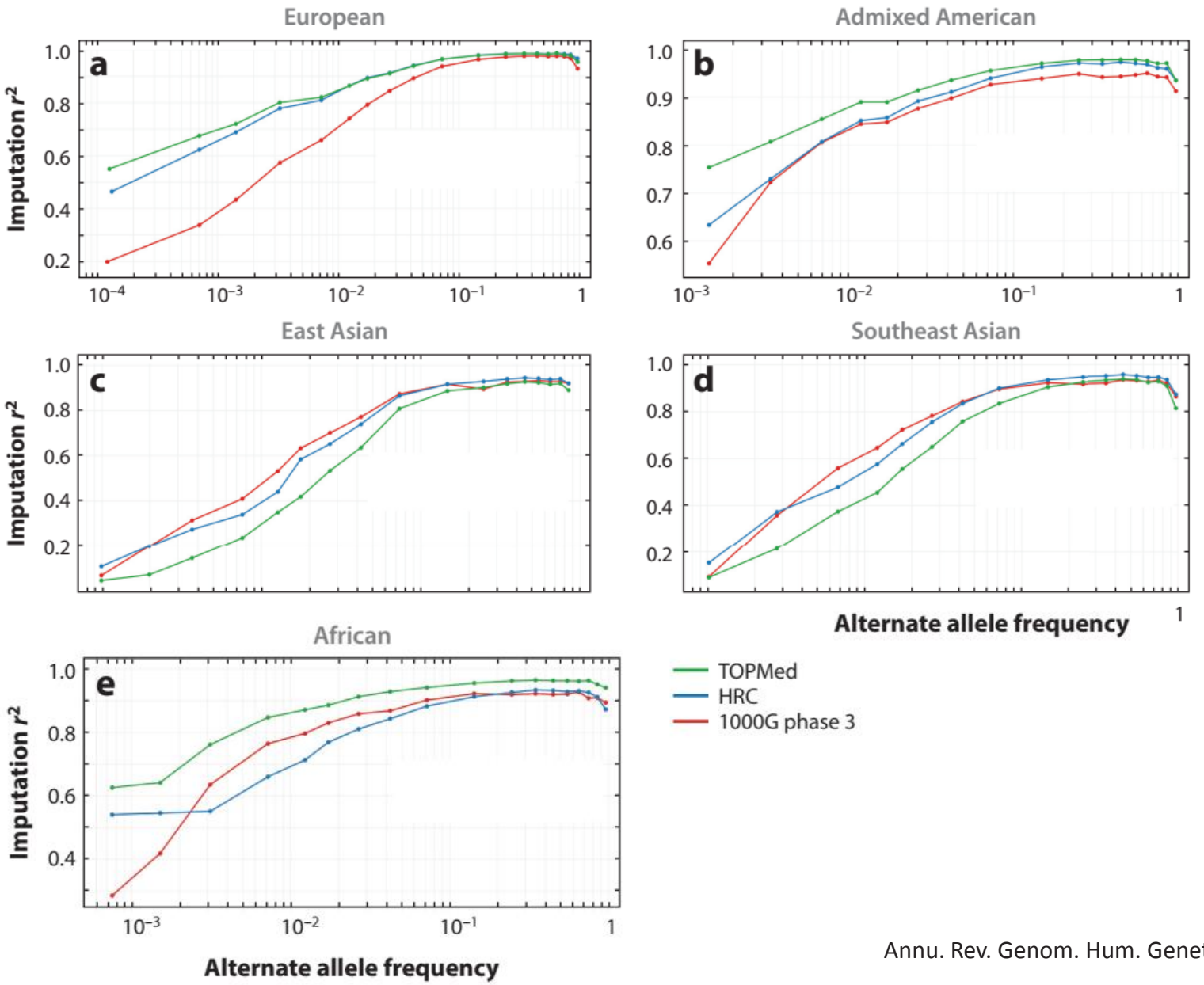
VOLUME 48 | NUMBER 10 | OCTOBER 2016 NATURE GENETICS

Reference panel	Number of samples	minimac3	minimac2	IMPUTE2	Beagle 4.1
		Imputation accuracy (mean r^2), MAF = 0.0001–0.5%			
1000G Phase 1	1,092	**0.45**	**0.45**	0.43	0.42
AMD	2,074	**0.54**	**0.54**	0.51	0.52
1000G Phase 3	2,504	**0.52**	**0.52**	0.49	**0.52**
SardiNIA	3,489	**0.55**	**0.55**	0.53	0.54
COMBINED	9,341	**0.76**	**0.76**	0.74	**0.76**
Mega	11,845	**0.76**	**0.76**	0.74	**0.76**
HRC v1.1	32,390	**0.77**	**0.77**	0.75	**0.77**
		Imputation accuracy (mean r^2), MAF = 0.5–5%			
1000G Phase 1	1,092	**0.77**	**0.77**	0.76	0.73
AMD	2,074	**0.82**	**0.82**	0.80	0.80
1000G Phase 3	2,504	**0.79**	**0.79**	0.78	**0.79**
SardiNIA	3,489	0.79	0.79	0.78	**0.80**
COMBINED	9,341	**0.89**	**0.89**	0.88	**0.89**
Mega	11,845	**0.89**	**0.89**	0.88	**0.89**
HRC v1.1	32,390	**0.90**	**0.90**	0.89	**0.90**
		Imputation accuracy (mean r^2), MAF = 5–50%			
1000G Phase 1	1,092	**0.96**	**0.96**	0.95	0.95
AMD	2,074	**0.96**	**0.96**	**0.96**	**0.96**
1000G Phase 3	2,504	**0.96**	**0.96**	**0.96**	**0.96**
SardiNIA	3,489	**0.96**	**0.96**	**0.96**	**0.96**
COMBINED	9,341	**0.97**	**0.97**	**0.97**	**0.97**
Mega	11,845	**0.97**	**0.97**	**0.97**	**0.97**
HRC v1.1	32,390	**0.98**	**0.98**	**0.98**	**0.98**

Factors affecting imputation

- Size and Sequencing coverage of reference panel
- Genetic similarity between the reference panel and study samples
- Minor allele frequency of variant being imputed (in the reference panel)
- Density of genotyping array
- Demographic history of the population

Why do we need so many reference panaels..



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

SCIENTIFIC REPORTS

OPEN

A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples

Wen-Chi Chou[1,2,*], Hou-Feng Zheng[3,*], Chia-Ho Cheng[1], Han Yan[4], Li Wang[3], Fang Han[4], J. Brent Richards[5,6], David Karasik[1], Douglas P. Kiel[1,2] & Yi-Hsiang Hsu[1,2,7]

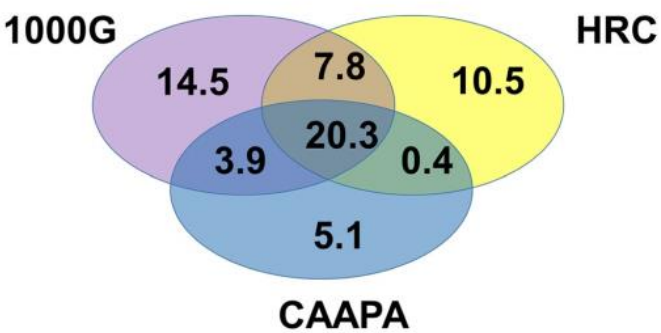
Received: 01 December 2015
Accepted: 16 November 2016
Published: 22 December 2016

Human Genetics (2018) 137:281–292
https://doi.org/10.1007/s00439-018-1881-4

ORIGINAL INVESTIGATION

Genotype imputation performance of three reference panels using African ancestry individuals

Candelaria Vergara[1] · Margaret M. Parker[2] · Liliana Franco[3,4] · Michael H. Cho[2,5] · Ana V. Valencia-Duarte[4] · Terri H. Beaty[6] · Priya Duggal[6]



Overlap between reference panels

Human Genetics (2018) 137:431–436
https://doi.org/10.1007/s00439-018-1894-z

ORIGINAL INVESTIGATION

Genotype imputation for Han Chinese population using Haplotype Reference Consortium as reference

Yuan Lin[1] · Lu Liu[2,3] · Sen Yang[2,3] · Yun Li[4] · Dongxin Lin[5] · Xuejun Zhang[2,3] · Xianyong Yin[2,3,6]

SCIENTIFIC REPORTS

OPEN

A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples

Received: 01 December 2015
Accepted: 16 November 2016
Published: 22 December 2016

Wen-Chi Chou[1,2,*], Hou-Feng Zheng[3,*], Chia-Ho Cheng[1], Han Yan[4], Li Wang[3], Fang Han[4], J. Brent Richards[5,6], David Karasik[1], Douglas P. Kiel[1,2] & Yi-Hsiang Hsu[1,2,7]

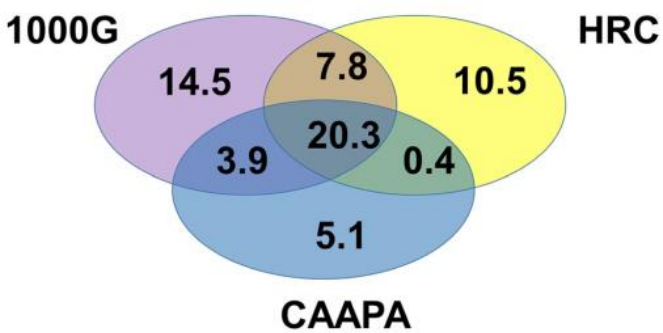
Human Genetics (2018) 137:281–292
https://doi.org/10.1007/s00439-018-1881-4

ORIGINAL INVESTIGATION

Genotype imputation performance of three reference panels using African ancestry individuals

Candelaria Vergara[1] · Margaret M. Parker[2] · Liliana Franco[3,4] · Michael H. Cho[2,5] · Ana V. Valencia-Duarte[4] · Terri H. Beaty[6] · Priya Duggal[6]

H3ABioNet is currently working on generating an African specific reference panel.



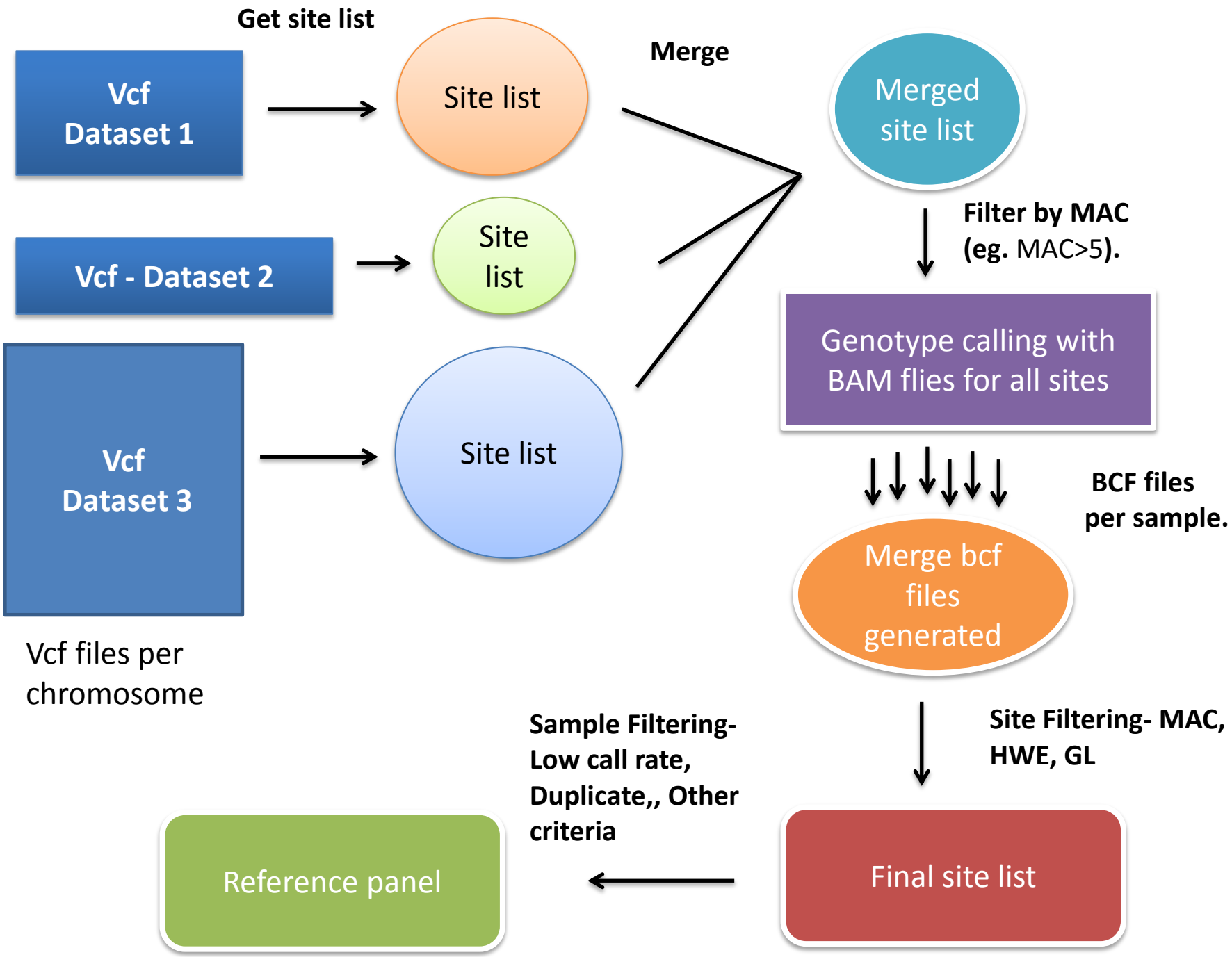
Overlap between reference panels

Human Genetics (2018) 137:431–436
https://doi.org/10.1007/s00439-018-1894-z

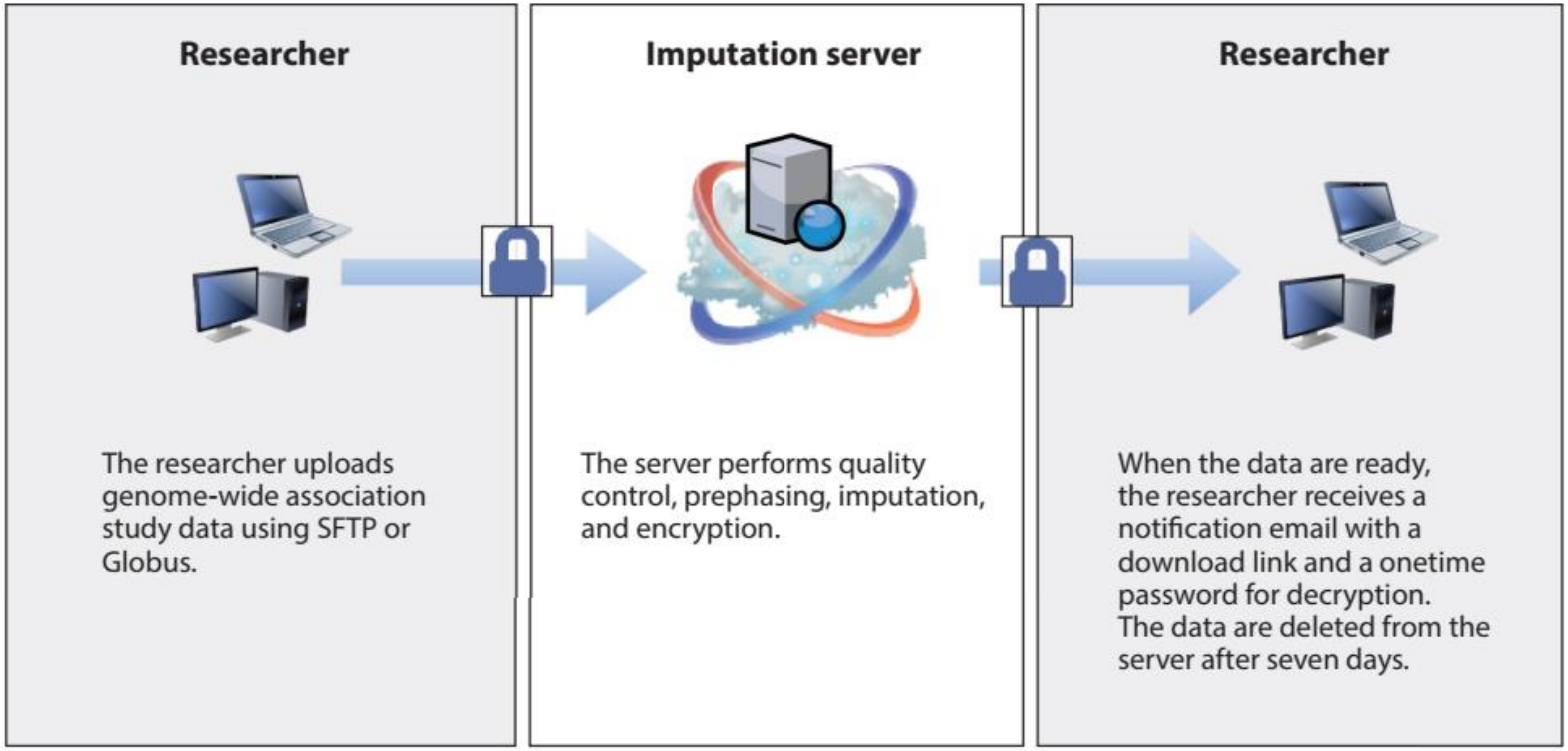
ORIGINAL INVESTIGATION

Genotype imputation for Han Chinese population using Haplotype Reference Consortium as reference

Yuan Lin[1] · Lu Liu[2,3] · Sen Yang[2,3] · Yun Li[4] · Dongxin Lin[5] · Xuejun Zhang[2,3] · Xianyong Yin[2,3,6]


Get site list
Vcf
Dataset 1
Site list
Merge
Merged
site list
Vcf - Dataset 2
Site
list
Filter by MAC
(eg. MAC>5).
Genotype calling with
BAM flies for all sites
Vcf
Dataset 3
Site list
BCF files
per sample.
Merge bcf
files
generated
Vcf files per
chromosome
Site Filtering- MAC,
HWE, GL
Sample Filtering-
Low call rate,
Duplicate,, Other
criteria
Reference panel
Final site list

Online Imputation Servers



- Michigan Imputation Server : https://imputationserver.sph.umich.edu/
- Sanger Imputation Server: https://imputation.sanger.ac.uk/

Michigan Imputation Server | Home | Help | Contact | Sign up | Login

Michigan Imputation Server

This server provides a free genotype imputation service. You can upload GWAS genotypes (VCF or 23andMe format) and receive phased and imputed genomes in return. Our server offers imputation from HapMap, 1000 Genomes (Phase 1 and 3), CAAPA and the updated Haplotype Reference Consortium (HRC version r1.1) panel. Learn more or follow us on Twitter.

18.4M Genomes

3,141 Users

Sign up now | Login

The easiest way to impute genotypes

Upload your genotypes to our server located in Michigan. All interactions with the server are **secured**.

Choose a reference panel. We will take care of pre-phasing and imputation.

Download the results. All results are encrypted with a one-time password. **After 3 days**, all results are deleted from our server.

*__Minimac3__ used for imputation

Michigan Imputation Server | Home | Run | Jobs | Help | Contact | dhriti

Michigan Imputation Server

Michigan Imputation Server provides a free genotype imputation service using Minimac3. You can upload phased or unphased GWAS genotypes and receive phased and imputed genomes in return. For all uploaded data sets an extensive QC is performed.

Name: optional job name

Reference Panel (Details): 1000G Phase 3 v5

Input Files (VCF & 23andMe): File Upload

Select Files

Multiple files can be selected by using the ctrl / cmd or shift keys.

Phasing: Eagle v2.3 (phased output)

- -- select an option --
- Eagle v2.3 (phased output)
- HapiUR (phased output)
- ShapeIT v2.r790 (unphased)

Population (for QC only)

Mode: Quality Control & Imputation

wellcome sanger institute

SCIENCE | Programmes | Groups | Collaborations | Tools | Data | Facilities

Home » Science » Tools and Software » Analysis

Sanger Imputation Service

Overview

Download

Contact

Authors

Overview

This is a free genotype **imputation** and **phasing** service
data in VCF or 23andMe format and receive imputed ar
SHAPEIT2 and imputation is with PBWT into a cho
and the Haplotype Reference Consortium.

Who is this for? This service is aimed at researchers w
consistent reference in a consistent manner. The PBWT
support genotype data from few samples, such as 23a
Globus is aimed at the larger files produced by imputin

Download and Installation

The service and instructions on how to use it are availa

Label your job:

test job

Choose a reference panel:

Haplotype Reference Consortium (r1.1)

Choose the pipeline:

- (•) pre-phase with EAGLE and impute
- ( ) pre-phase with SHAPEIT2 and impute
- ( ) phase with EAGLE, no imputation
- ( ) impute with PBWT, no pre-phasing

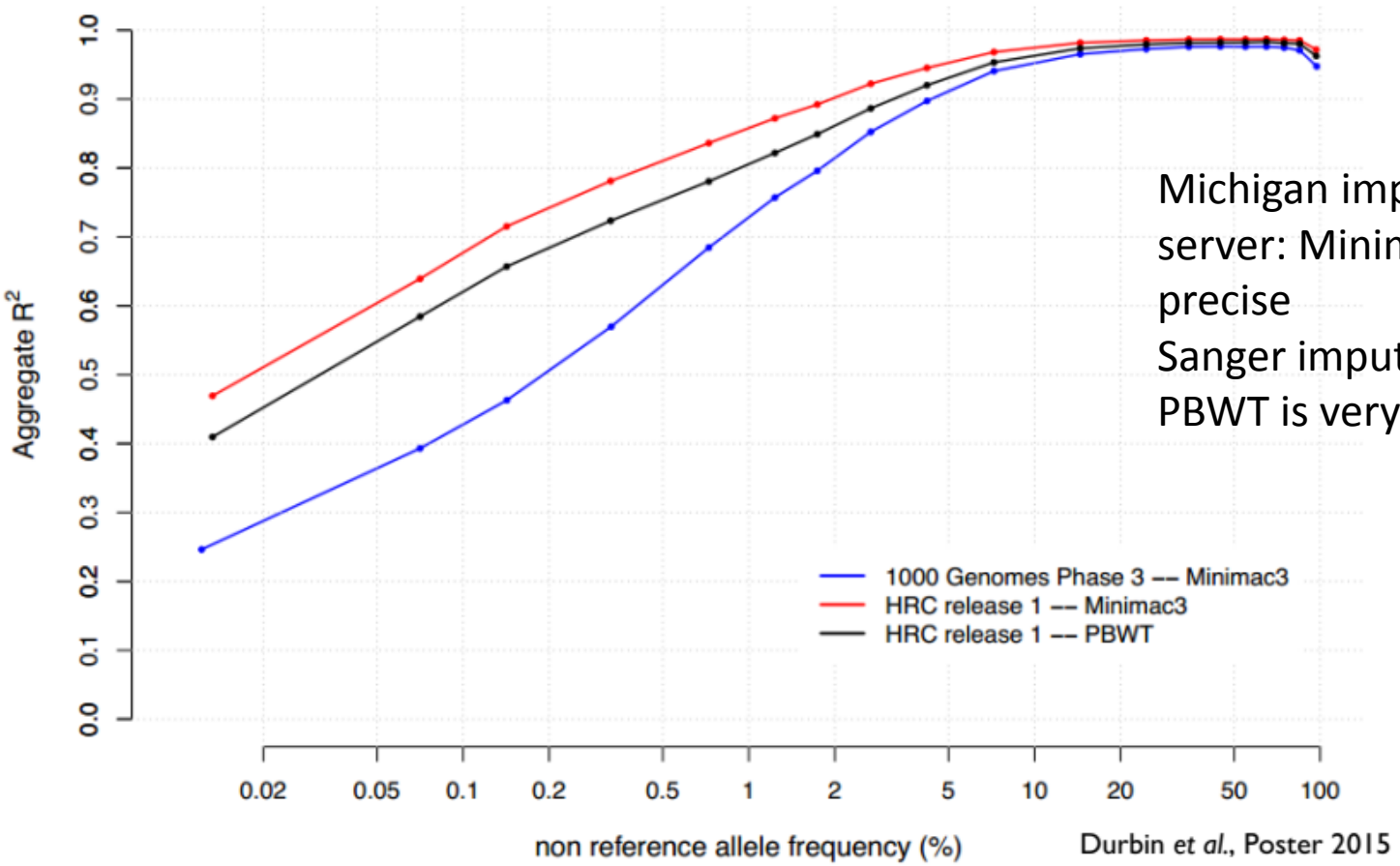
Back | Submit

* **PBWT** algorithm for imputation

Impute and Download

- QC on the GWA study data
- Data upload
  - Input Formats: VCF files for both the imputation Severs
  - Upload
- Imputation
  - Does pre-phasing and imputation
  - Time needed depends on how many samples/sites are there in input data
  - Generally within few days
- Download data
  - Notify by email
  - Download your imputed data.
  - Data deleted in few weeks
- Output file
  - The format of the returned data will be in the Variant Call Format (VCF).
  - One VCF per-chromosome

Main Difference between the Imputation servers



Michigan imputation server: Minimac3 is very precise

Sanger imputation server: PBWT is very fast

Durbin *et al.*, Poster 2015

Take home

1. Imputation is a powerful method to improve the outcome from GWAS study by making a number of additional analyses possible or simpler.
2. We have discussed parameters, such as genotype probability, allelic dosage, IQS (info scores and r2) to evaluate imputation quality at SNP level. Parameters such as number of variants imputed, number of high quality variants imputed, genomic distribution of IQS, allele frequency comparisons provide an estimate at the dataset level.
3. Factors such as size and sequencing coverage of the reference panel, genetic similarity between the reference panel and study sample, density of the genotyping array affects the quality of imputation.
4. There are various tools for imputation- we have discussed Mimimac3 and Impute2. Although Mimimac3 performed better to Impute 2, now we have Impute 4 and Minimac4, which are claimed to perform comparably. So the software choice might not be critical in the long run.
5. What seems critical at this moment is the choice of reference panels, as in a non-European population, as we all dealing with, it might make a huge difference.

THANK YOU!

IMPUTING!!!!