Imputation

Sarah Medland

Boulder 2019

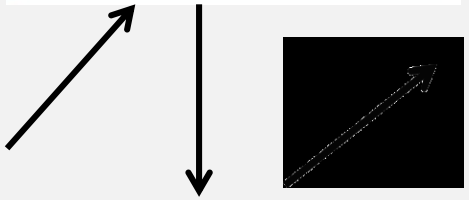
What is imputation? (Marchini & Howie 2010)

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

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

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	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	1	0



Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel

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

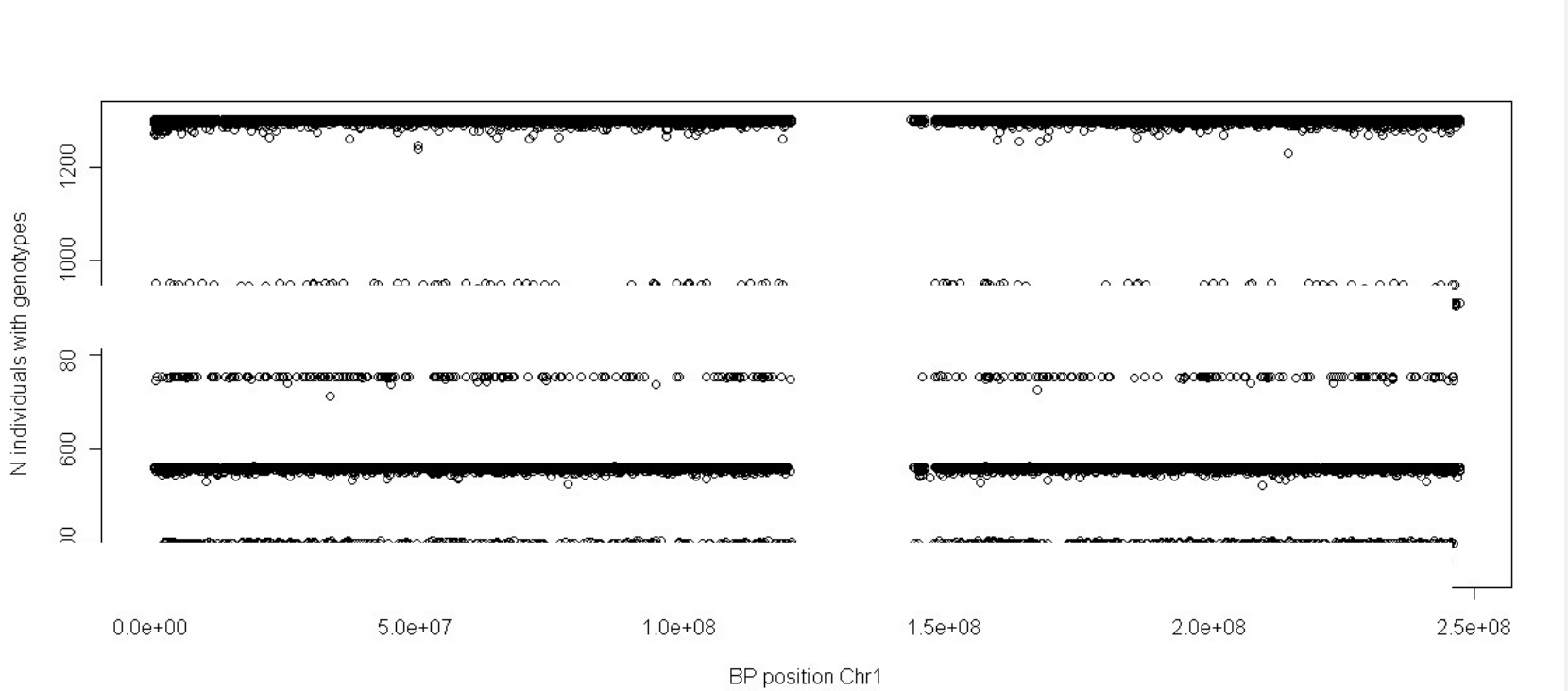
- 3 main reasons for imputation
  - Meta-analysis
  - Fine Mapping
  - Combining data from different chips

- Other less common uses
  - sporadic missing data imputation
  - correction of genotyping errors
  - imputation of non-SNP variation

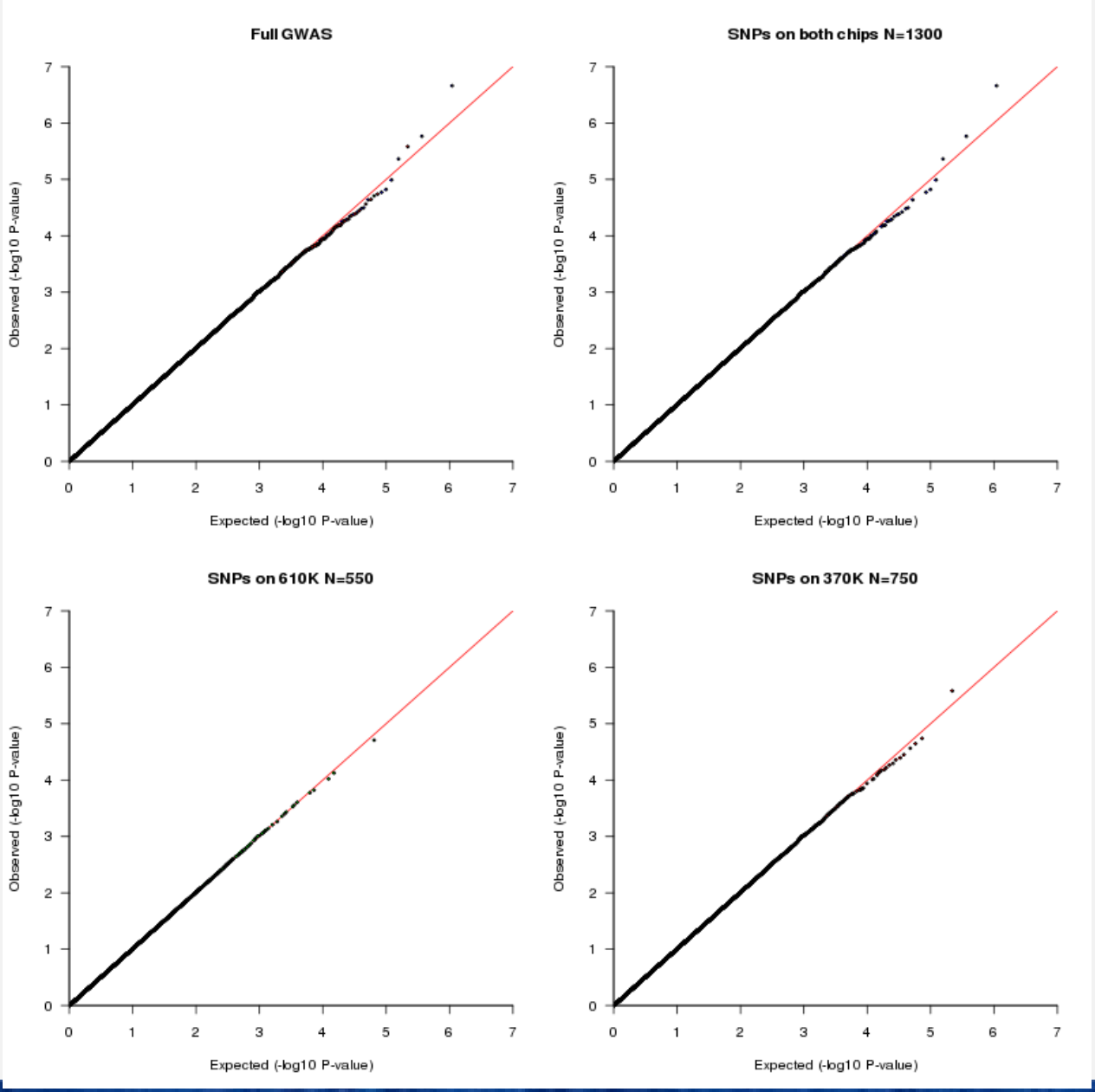
Combining data from different chips

- Example
  - 750 individuals typed on the 370K
  - 550 individuals typed on the 610K
- Power
  - MAF .2
  - SNP explaining 1% total variance
  - alpha 5e-08
    - N=1300, NCP 13.07, power .0331
    - N=750, NCP 7.54 , power .0034
    - N=550, NCP 5.53 , power .0009

Another way of looking at this



QQ-plot


Full GWAS
SNPs on both chips N=1300
SNPs on 610K N=550
SNPs on 370K N=750
Observed (-log10 P-value)
Expected (-log10 P-value)
0
1
2
3
4
5
6
7

Solution

- Impute all individuals to a single reference based on the SNPs that overlap between the chips
- Single distribution of NCP and power across all SNP
- qq plot and manhattan describes the full sample with the same degree of accuracy

Imputation programs

- minimac3
- Impute2

- Beagle – not frequently used

- never use plink for imputation!

How do they compare

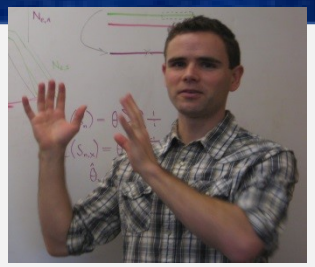
- Similar accuracy
- Similar features
- Different data formats
  - minimac3 –> custom vcf format
    - individual=row snp=column
  - Impute2 –> snp=row individual=column
- Different philosophies
  - Frequentist vs Bayesian

minimac3



- http://genome.sph.umich.edu/wiki/Minimac3
- Built by Gonçalo Abecasis, Yun Li, Christian Fuchsberger and colleagues
- Analysis options
  - SAIGE
  - BoltLMM
  - plink2

Impute2



- https://mathgen.stats.ox.ac.uk/impute/impute_v2.html
- http://genome.sph.umich.edu/wiki/IMPUTE2:_1000_Genomes_Imputation_Cookbook
- Built by Jonathan Marchini, Bryan Howie and colleges
- Downstream analysis options
  - SNPtest
  - Quicktest

Timing & Memory

from

Das et al 2016

Prior to EAGLE2

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

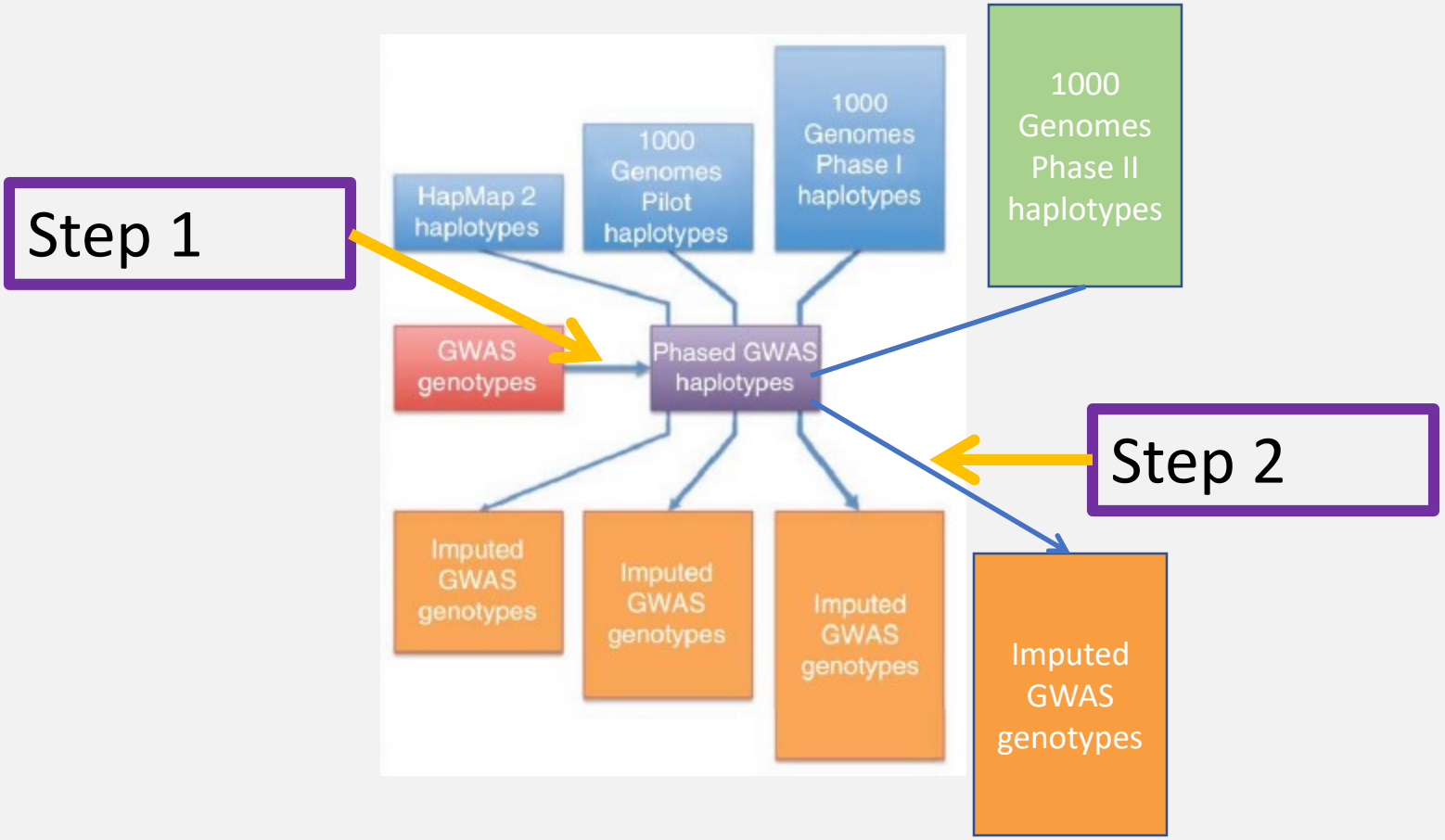
Options for imputation

- DIY – Use a cookbook!

  http://genome.sph.umich.edu/wiki/Minimac3_Imputation_Cookbook OR http://genome.sph.umich.edu/wiki/IMPUTE2:_1000_Genomes_Imputation_Cookbook

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

Today – discuss the 2 step approach



What is phasing

- In this context it is really Haplotype Estimation
- We take genotype data and try to reconstruct the haplotypes
  - Can use reference data to improve this estimation

Heterozygous genotypes at 3 sites

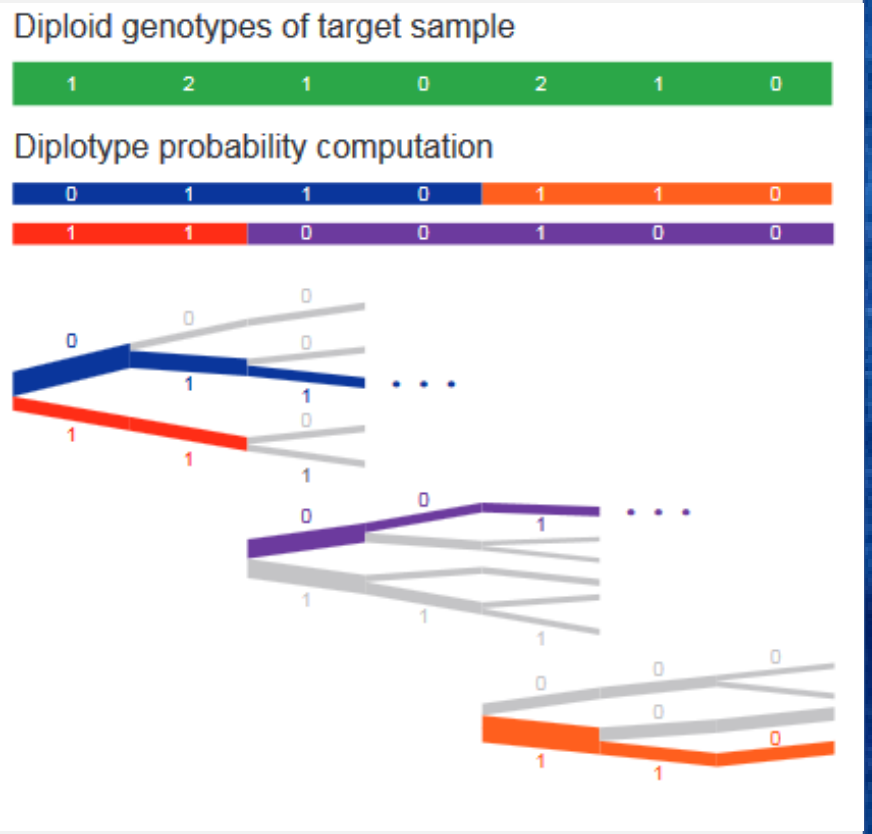
AC TG AT

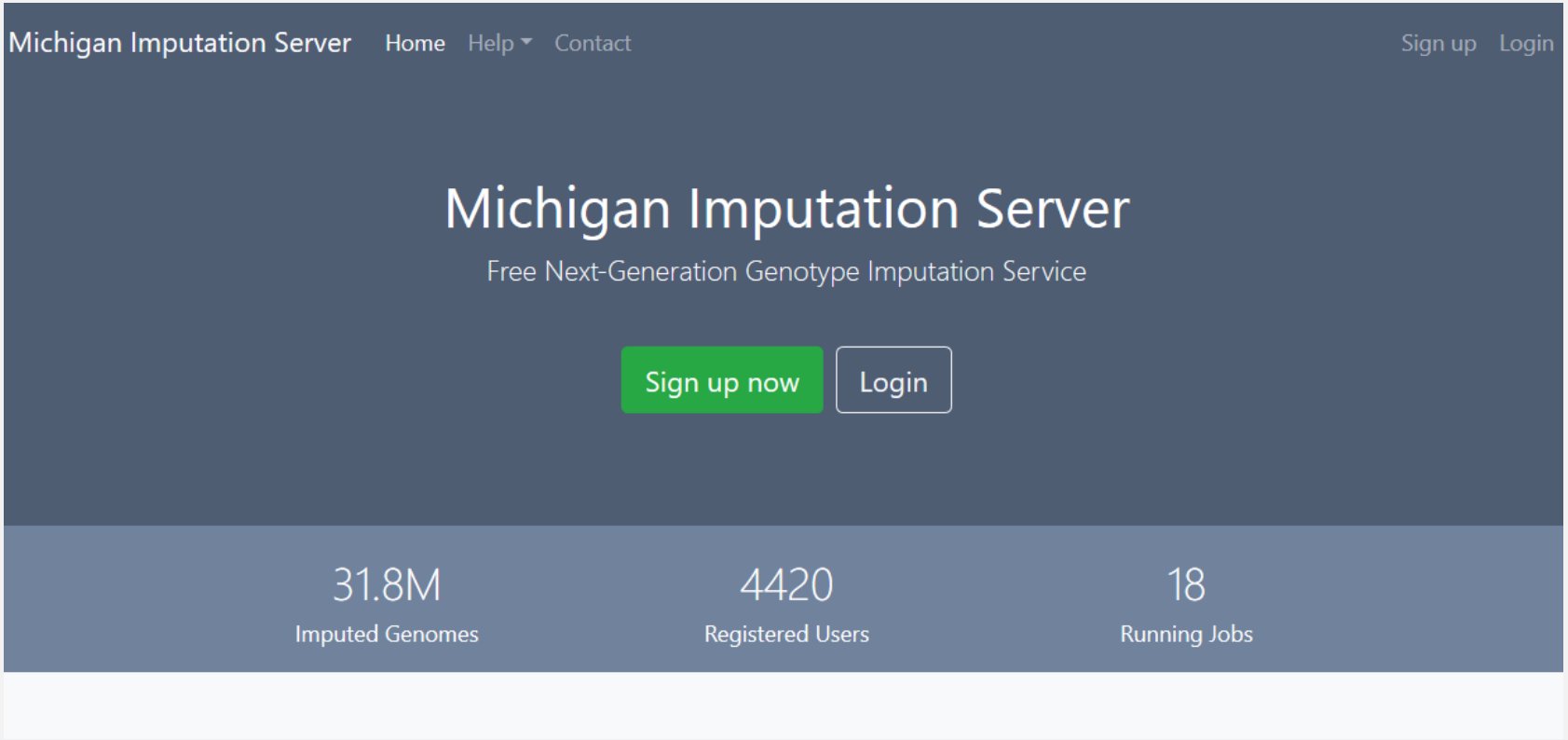
The 4 possible consistent pairs of haplotypes

ATT	ATA	AGT	AGA
CGA	CGT	CTA	CTT

Phasing in Eagle

- Input a target sample and a library of reference haplotypes
- *Selection of conditioning haplotypes.*
- *Generation of HapHedge data structure.*
- *Exploration of the diplotype space.*




Michigan Imputation Server
Home
Help
Contact
Sign up
Login
Michigan Imputation Server
Free Next-Generation Genotype Imputation Service
Sign up now
Login
31.8M
Imputed Genomes
4420
Registered Users
18
Running Jobs

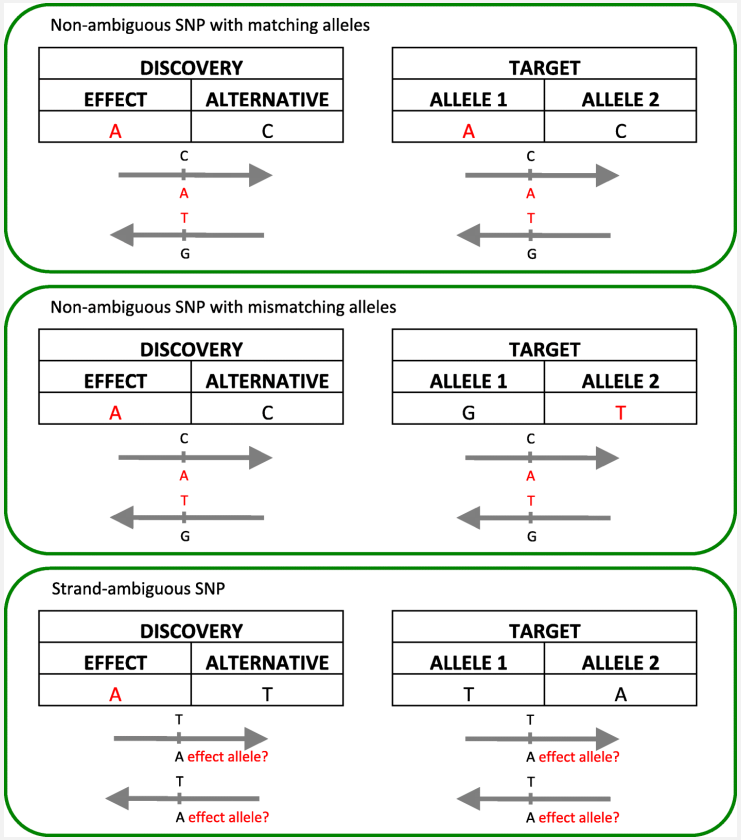
Check your data

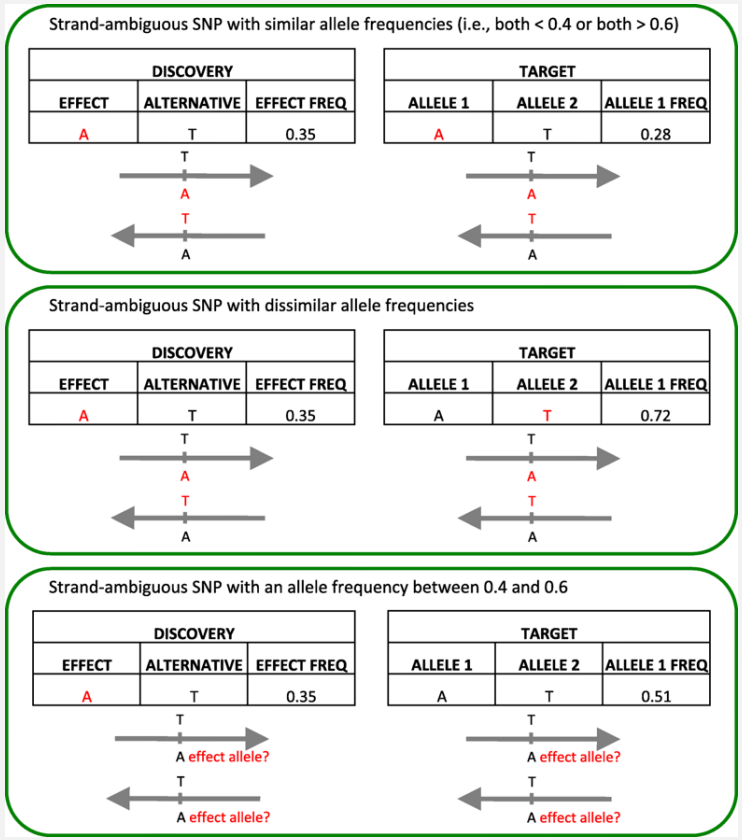
i. Exclude snps with excessive missingness (>5%), low MAF (<1%), HWE violations (~$P<10^{-6}$), Mendelian errors
ii. Drop strand ambiguous (palindromic) SNPs – ie A/T or C/G snps
iii. Update build and alignment (b37)
iv. Output your data in the expected format for the phasing program you will use

    Check the naming convention for the program and reference you want to use

    rs278405739 OR 22:395704

Ambiguous SNPs (A/T and C/G)





Getting Started

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. Our server offers imputation from HapMap, 1000 Genomes Phase 1 and 3, CAAPA and the new HRC reference panel. For all uploaded data sets an extensive QC is performed.

0. Prepare your data

Accepted Input: VCF files compressed by bgzip (*.vcf.gz).

1. To convert your ped/map file into a VCF file, please use either plink2, VCFtools or VcfCooker.
   ```
   plink --ped mystudy_chr1.ped --map mystudy_chr1.map --recode vcf --out mystudy_chr1
   ```

2. Create a sorted *.vcf.gz file using VCFtools and tabix (including bgzip):

   ```
   vcf-sort mystudy_chr1.vcf | bgzip -c > mystudy_chr1.vcf.gz
   ```

Important:

- Create a seperate vcf.gz file for each chromosome.
- Variations must be sorted by genomic position (see above).
- GRCh37 coordinates are required.
- Several *.vcf.gz files can be uploaded at once.

UMich imputation Server

4. Start the Imputation!

After specifying the data location, the imputation process can be started immediately. The default values are:

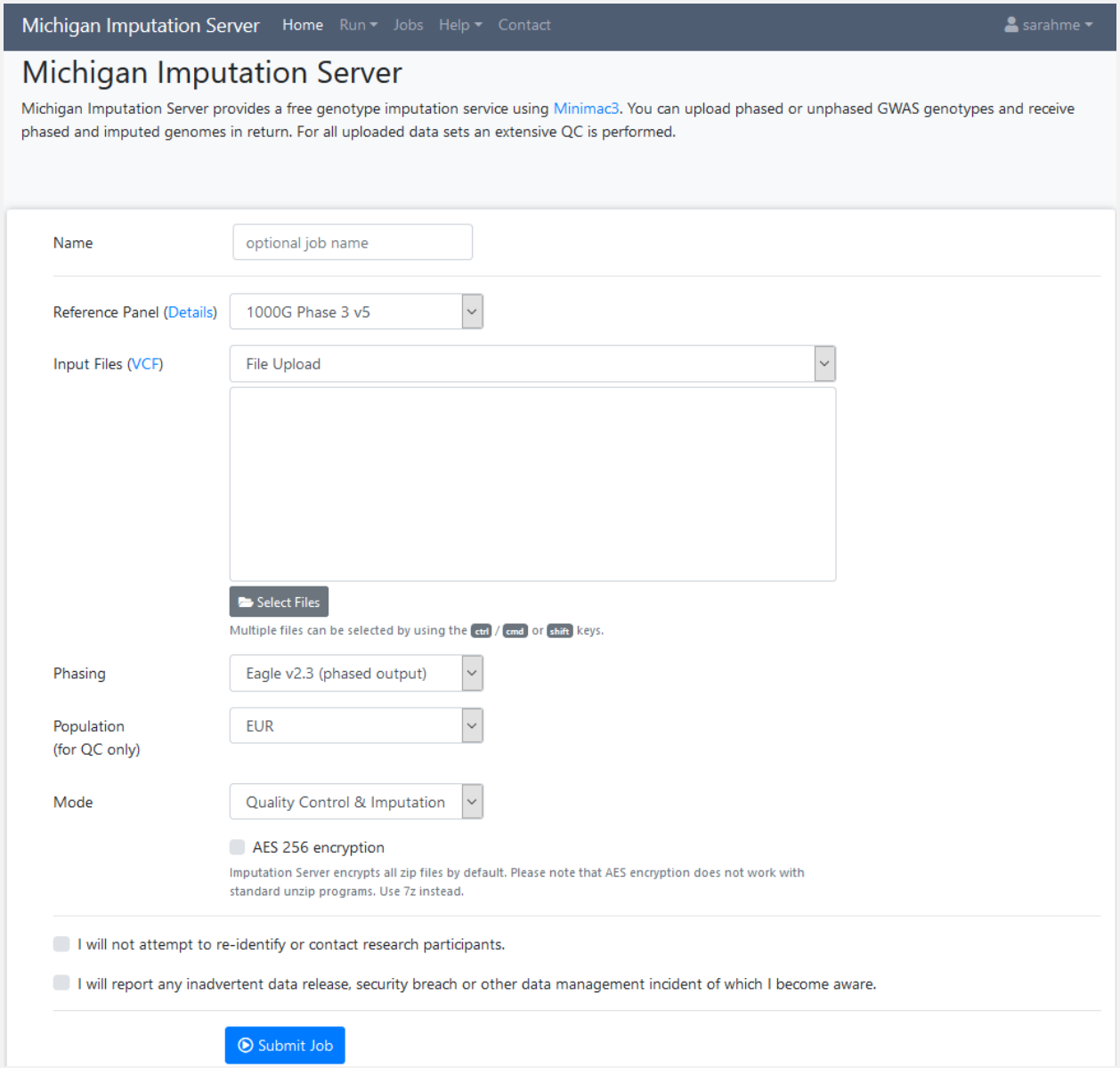
- **Reference panel: HRC**. All available reference panels can be found here.
- **Phasing:** Eagle2. This can be toggled to ShapeIT or HAPI-UR.
- **Population: EUR**. The selected population is used for the allele frequency check in quality control only. We compare the frequencies of your uploaded data with the reference panel.

In the background several steps are executed:

- **VCF check:** validity + statistics such as #samples, chromosomes, SNPs, chunks, phased / unphased, reference build.
- **Quality control statistics:** duplicate sites, SNPs removed, NonSNP sites, monomorphic sites, MAF check.
- **Imputation**: Imputation is achieved with minimac3. An overview of running / waiting / completed steps per user can be displayed.

Chose your reference

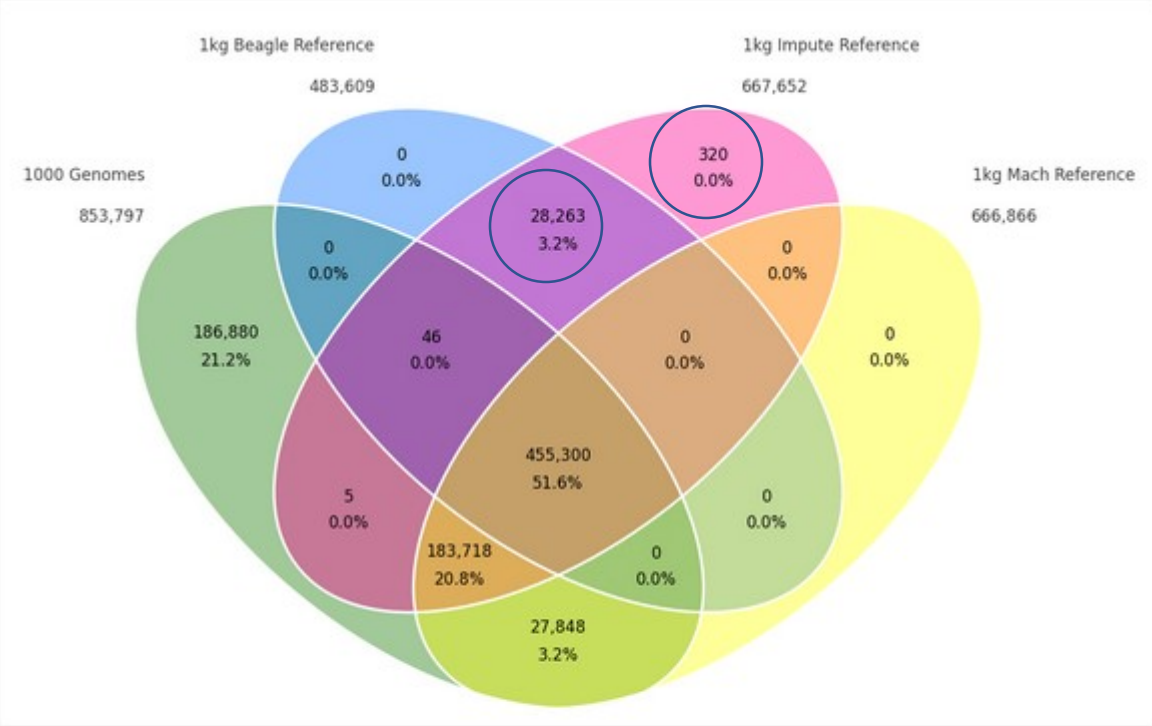
- Current Publically Available References
  - HapMapII (no phased X data officially released)
  - 1KGP – phase 1 version v3
  - 1KGP – phase 3 version v5
- Future non-public references only available via custom imputation servers
  - HRC - 64,976 haplotypes 39,235,157 SNPs
  - CAPPA – African American/Carabbean
  - TopMed


Michigan Imputation Server
Home
Run
Jobs
Help
Contact
sarahme
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)
File Upload
Select Files
Multiple files can be selected by using the ctrl / cmd or shift keys.
Phasing
Eagle v2.3 (phased output)
Population
(for QC only)
EUR
Mode
Quality Control & Imputation
AES 256 encryption
Imputation Server encrypts all zip files by default. Please note that AES encryption does not work with standard unzip programs. Use 7z instead.
I will not attempt to re-identify or contact research participants.
I will report any inadvertent data release, security breach or other data management incident of which I become aware.
Submit Job

References are in vcf format

```
##fileformat=VCFv4.1
##INFO=<ID=LDAF,Number=1,Type=Float,Description="MLE Allele Frequency Accounting for LD">
##INFO=<ID=AVGPOST,Number=1,Type=Float,Description="Average posterior probability from MaCH/Thunder">
##INFO=<ID=RSQ,Number=1,Type=Float,Description="Genotype imputation quality from MaCH/Thunder">
##INFO=<ID=ERATE,Number=1,Type=Float,Description="Per-marker Mutation rate from MaCH/Thunder">
##INFO=<ID=THETA,Number=1,Type=Float,Description="Per-marker Transition rate from MaCH/Thunder">
##INFO=<ID=CIEND,Number=2,Type=Integer,Description="Confidence interval around END for imprecise variants">
##INFO=<ID=CIPOS,Number=2,Type=Integer,Description="Confidence interval around POS for imprecise variants">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant described in this record">
##INFO=<ID=HOMLEN,Number=.,Type=Integer,Description="Length of base pair identical micro-homology at event breakpoints">
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology at event breakpoints">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Difference in length between REF and ALT alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate Allele Count">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total Allele Count">
##ALT=<ID=DEL,Description="Deletion">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Genotype dosage from MaCH/Thunder">
##FORMAT=<ID=GL,Number=.,Type=Float,Description="Genotype Likelihoods">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/ancestral_alignments/README">
##INFO=<ID=AF,Number=1,Type=Float,Description="Global Allele Frequency based on AC/AN">
##INFO=<ID=AMR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from AMR based on AC/AN">
##INFO=<ID=ASN_AF,Number=1,Type=Float,Description="Allele Frequency for samples from ASN based on AC/AN">
##INFO=<ID=AFR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from AFR based on AC/AN">
##INFO=<ID=EUR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from EUR based on AC/AN">
##INFO=<ID=VT,Number=1,Type=String,Description="indicates what type of variant the line represents">
##INFO=<ID=SNPSOURCE,Number=.,Type=String,Description="indicates if a snp was called when analysing the low coverage or exome alignment data">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096 HG00097 HG00099 HG00100 HG00101 HG00102 HG00103 HG00104 HG00106 HG00108 HG00109 HG00110 HG00111
10      60523   rs148087467     T       G       100     PASS    AN=2184;NS=1092;AC=32   GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      60969   rs187110906     C       A       100     PASS    AN=2184;NS=1092;AC=155  GT      0|1     0|0     0|0     0|0     0|0     0|0     0|0     0|1     0|0     0|0
10      61005   rs192025213     A       G       100     PASS    AN=2184;NS=1092;AC=15   GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      61020   rs115033199     G       C       100     PASS    AN=2184;NS=1092;AC=8    GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      61334   rs183305313     G       A       100     PASS    AN=2184;NS=1092;AC=5    GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      66326   rs12260013      A       G       100     PASS    AN=2184;NS=1092;AC=113  GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      66627   .       TAAAC   T       378     PASS    AN=2184;NS=1092;AC=953  GT      1|1     0|0     0|1     1|1     0|0     0|0     0|1     0|0     0|1     0|0
10      67193   rs182646175     C       T       100     PASS    AN=2184;NS=1092;AC=34   GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      68258   .       GA      G       0       PASS    AN=2184;NS=1092;AC=47   GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
10      68523   rs186971761     A       C       100     PASS    AN=2184;NS=1092;AC=4    GT      0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0     0|0
```

Not all references are equal!!

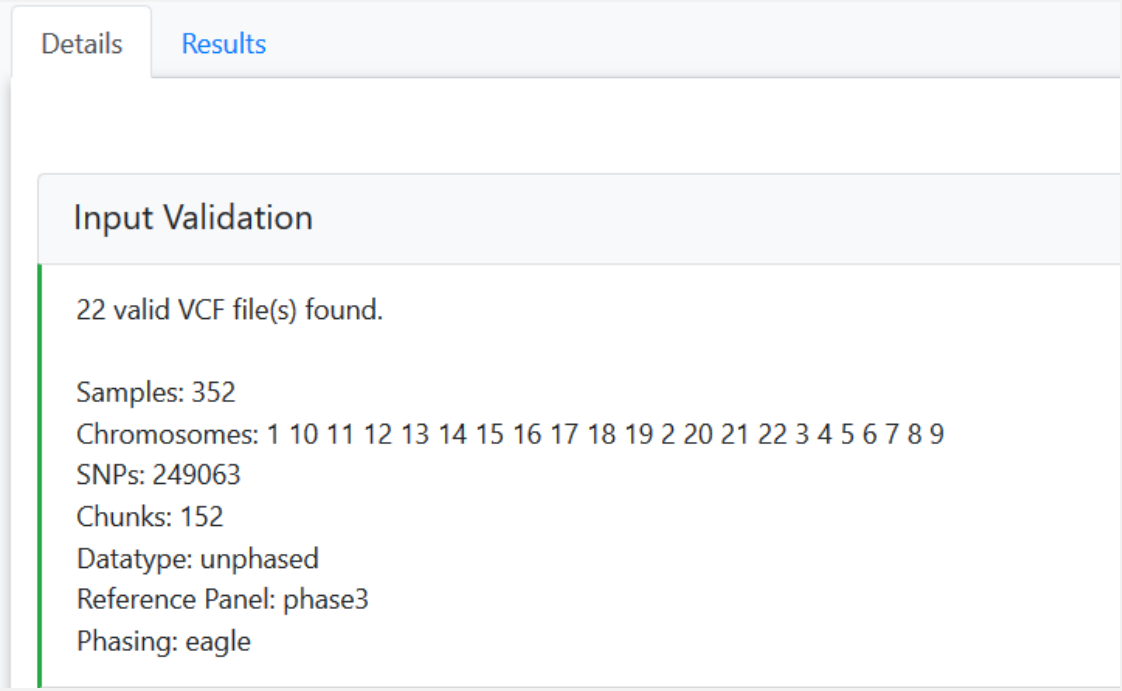


This Venn diagram displays how markers in the reference panels for each imputation program and the original 1000 Genomes data overlap on chromosome 20. Unique lists of genomic position were compared across datasets.
The original dataset and the MaCH reference panels came with genomic position in the format of VCF files. For the BEAGLE reference panel, genomic position was determined with the .markers files and with the legend file for IMPUTE2.

The Beagle and IMPUTE versions of the references contain variants that do not appear in the publically available 1KGP data!

The 1KGP references still contain multiple locations with more than 1 variant
&
Multiple variants in more than one place!

QC on the imputation server



Quality Control

Execution successful.

Statistics:
Alternative allele frequency > 0.5 sites: 0
Reference Overlap: 100.00%
Match: 108,529
Allele switch: 39,315
Strand flip: 85,444
Strand flip and allele switch: 39,207
A/T, C/G genotypes: 23,771
Filtered sites:
Filter flag set: 0
Invalid alleles: 0
Duplicated sites: 0
NonSNP sites: 0
Monomorphic sites: 0
Allele mismatch: 971
SNPs call rate < 90%: 0

Excluded sites in total: 125,622
Remaining sites in total: 171,615

Error: More than 100 obvious strand flips have been detected. Please check strand. Imputation cannot be started!

Quality Control

Execution successful.

Statistics:
Alternative allele frequency > 0.5 sites: 0
Reference Overlap: 100.00%
Match: 108,529
Allele switch: 39,315
Strand flip: 85,444
Strand flip and allele switch: 39,207
A/T, C/G genotypes: 23,771
Filtered sites:
Filter flag set: 0
Invalid alleles: 0
Duplicated sites: 0
NonSNP sites: 0
Monomorphic sites: 0
Allele mismatch: 971
SNPs call rate < 90%: 0

Excluded sites in total: 125,622
Remaining sites in total: 171,615

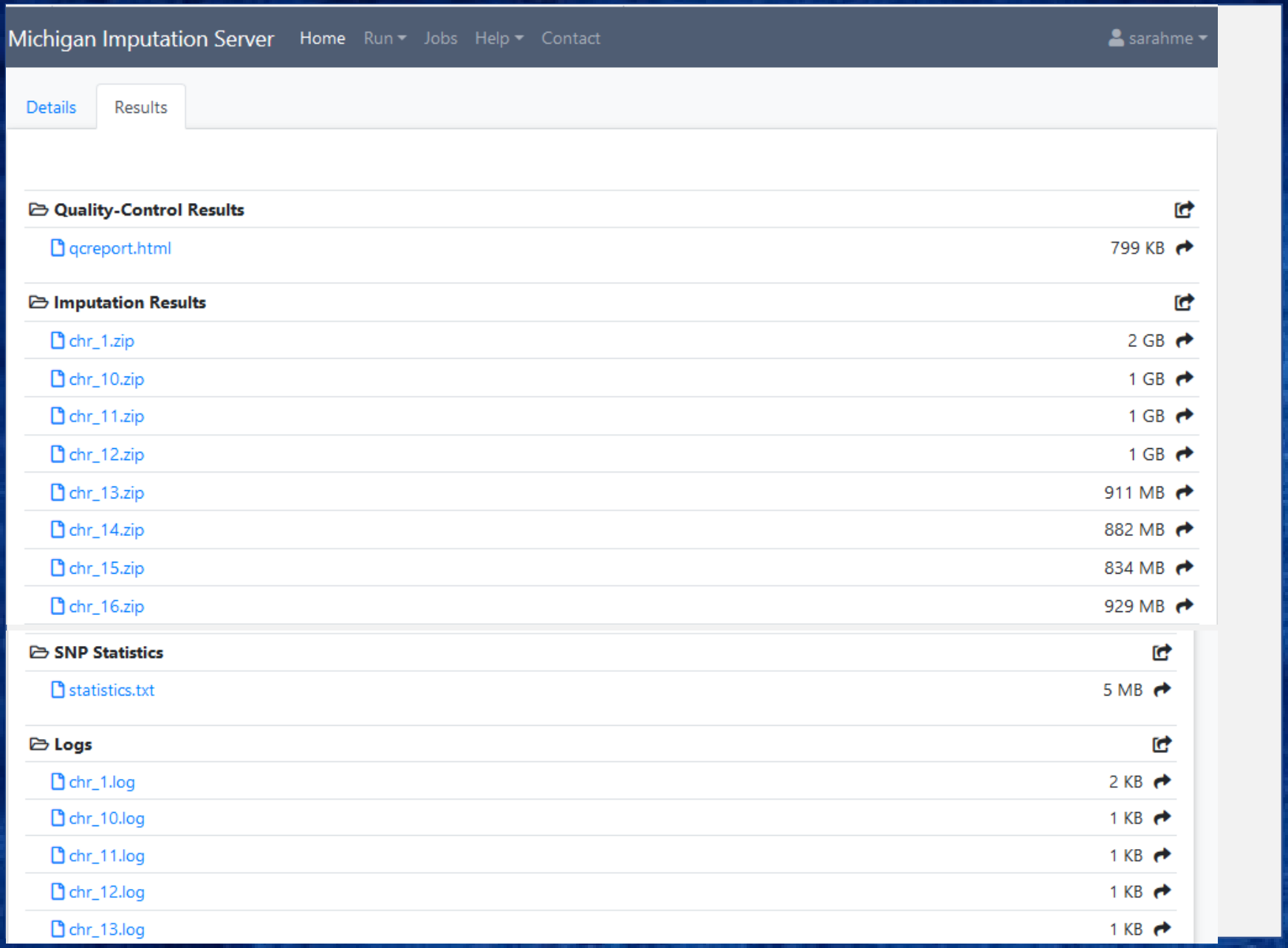
Error: More than 100 obvious strand flips have been detected. Please check strand imputation cannot be started.

Quality Control

Execution successful.

Statistics:
Alternative allele frequency > 0.5 sites: 0
Reference Overlap: 100.00%
Match: 170,541
Allele switch: 78,522
Strand flip: 0
Strand flip and allele switch: 0
A/T, C/G genotypes: 0
Filtered sites:
Filter flag set: 0
Invalid alleles: 0
Duplicated sites: 0
NonSNP sites: 0
Monomorphic sites: 0
Allele mismatch: 0
SNPs call rate < 90%: 0

Excluded sites in total: 0
Remaining sites in total: 249,063

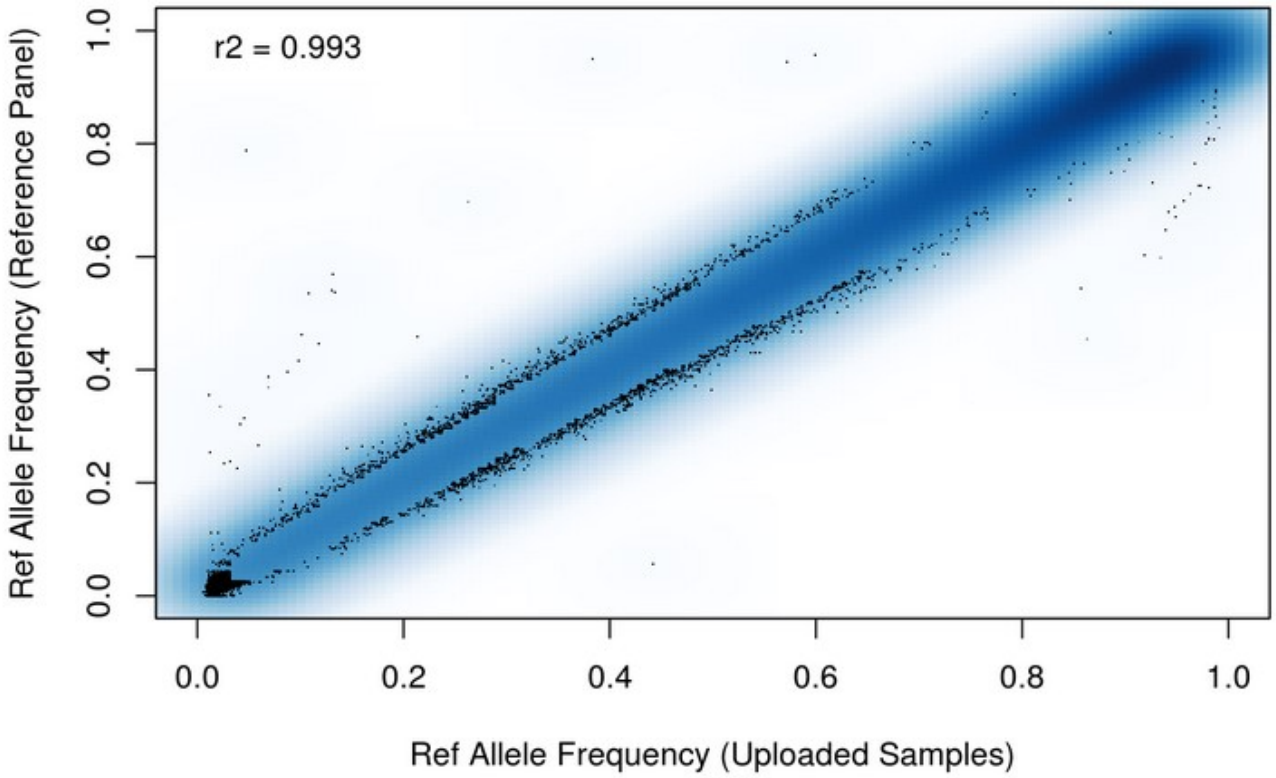

Michigan Imputation Server
Home
Run
Jobs
Help
Contact
sarahme
Details
Results
Quality-Control Results
qcreport.html
799 KB
Imputation Results
chr_1.zip
2 GB
chr_10.zip
1 GB
chr_11.zip
1 GB
chr_12.zip
1 GB
chr_13.zip
911 MB
chr_14.zip
882 MB
chr_15.zip
834 MB
chr_16.zip
929 MB
SNP Statistics
statistics.txt
5 MB
Logs
chr_1.log
2 KB
chr_10.log
1 KB
chr_11.log
1 KB
chr_12.log
1 KB
chr_13.log
1 KB

QC-Report

Allele-Frequency Correlation

Uploaded Samples vs. Reference Panel

The plot shows the densities of frequencies falling into each part (excluding chromosome X). The first 5000 points from areas of lowest regional densities will be plotted.

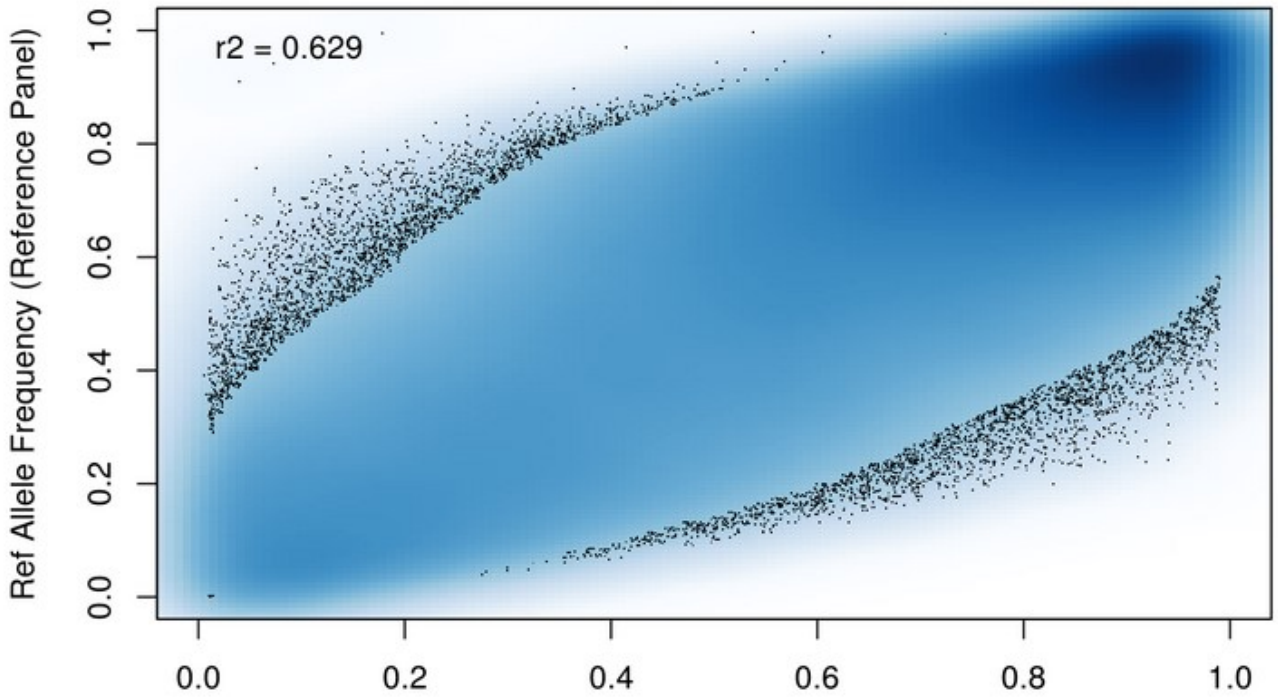


QC-Report

Allele-Frequency Correlation

Uploaded Samples vs. Reference Panel

The plot shows the densities of frequencies falling into each part (excluding chromosome X). The first 5000 points from areas of lowest regional densities will be plotted.



Potential Frequency Mismatches

Markers where chisq is greater than 300.

```
## Total mismatches: 13Mismatched frequencies for '10:111850503' f[A,G] = [0.1301775,0.8698225] vs [0.5408,0.4592], chisq 3
62.9649
## Mismatched frequencies for '11:14181174' f[T,C] = [0.5716332,0.4283668] vs [0.9453,0.0547], chisq 652.8791
## Mismatched frequencies for '11:114432867' f[G,A] = [0.8627167,0.1372832] vs [0.4543,0.5457], chisq 365.2667
## Mismatched frequencies for '11:119648458' f[A,C] = [0.1331361,0.8668639] vs [0.5368,0.4632], chisq 351.1041
## Mismatched frequencies for '15:78269472' f[A,G] = [0.04782608,0.9521739] vs [0.7883,0.2117], chisq 1275.596
## Mismatched frequencies for '19:55357424' f[T,C] = [0.3836207,0.6163793] vs [0.9503,0.0497], chisq 1232.612
## Mismatched frequencies for '1:65872597' f[A,C] = [0.1079882,0.8920118] vs [0.5358,0.4642], chisq 395.3629
```

Potential Frequency Mismatches

Markers where chisq is greater than 300.

```
## Total mismatches: 4831Mismatched frequencies for '10:327832' f[A,C] = [0.4571429,0.5428572] vs [0.8749,0.1251], chisq 31
8.6727
## Mismatched frequencies for '10:499573' f[C,T] = [0.9755747,0.02442529] vs [0.5657,0.4343], chisq 345.4074
## Mismatched frequencies for '10:514138' f[C,T] = [0.03276353,0.9672365] vs [0.4536,0.5464], chisq 353.7223
## Mismatched frequencies for '10:526773' f[G,A] = [0.03994083,0.9600592] vs [0.4592,0.5408], chisq 335.7513
## Mismatched frequencies for '10:618409' f[C,T] = [0.8859649,0.1140351] vs [0.3267,0.6733], chisq 492.09
## Mismatched frequencies for '10:2349064' f[C,T] = [0.08045977,0.9195402] vs [0.5866,0.4134], chisq 430.9494
## Mismatched frequencies for '10:2349772' f[C,T] = [0.04505814,0.9549419] vs [0.4926,0.5074], chisq 371.247
## Mismatched frequencies for '10:2766759' f[G,A] = [0.3211144,0.6788856] vs [0.7854,0.2146], chisq 341.2071
## Mismatched frequencies for '10:2766921' f[G,A] = [0.319242,0.680758] vs [0.7854,0.2146], chisq 344.753
## Mismatched frequencies for '10:2772078' f[C,T] = [0.319242,0.680758] vs [0.8148,0.1852], chisq 394.4303
## Mismatched frequencies for '10:2805185' f[C,T] = [0.4176136,0.5823864] vs [0.8443,0.1557], chisq 315.5917
```

Output

- Info files

SNP	Al1	Al2	Freq1	MAF	AvgCall	Rsq	Genotyped	LooRsq	EmpR	EmpRsq	Dose1	Dose2
1:10583	G	A	0.79288	0.20712	0.79288	-0.00000	-	-	-	-	-	-
1:10611	C	G	0.97889	0.02111	0.97889	0.00000	-	-	-	-	-	
1:13302	C	T	0.86280	0.13720	0.86280	-0.00000	-	-	-	-	-	-
1:13327	G	C	0.96042	0.03958	0.96042	-0.00000	-	-	-	-	-	-
1:95207182	T	C	0.99547	0.00453	0.99547	0.10108	-	-	-	-	-	-
1:95207382	T	T	1.00000	0.00000	1.00000	0.00000	-	-	-	-	-	-
1:95207442	C	T	0.62754	0.37246	0.99999	1.00507	Genotyped	0.98810	0.99822	0.99645	0.99484	0.00421
1:95207524	G	A	0.78061	0.21939	1.00000	1.00511	Genotyped	1.00059	1.00000	1.00000	0.99924	0.00083
1:95207532:TG_T	R	D	0.78620	0.21380	0.99441	0.97729	-	-	-	-	-	-
1:95207558	C	T	0.99399	0.00601	0.99399	0.05165	-	-	-	-	-	-
1:95207633	A	C	0.93366	0.06634	0.99998	1.00482	Genotyped	0.94847	0.99901	0.99802	0.99621	0.00372
1:95207846	G	T	0.98937	0.01063	0.98942	0.31316	-	-	-	-	-	-

Imputation quality evaluation

Minimac hides each of the genotyped SNPs in turn and then calculates 3 statistics:

- looRSQ - this is the estimated rsq for that SNP (as if SNP weren't typed).
- empR - this is the empirical correlation between true and imputed genotypes for the SNP. If this is negative, the SNP alleles are probably flipped.
- empRSQ - this is the actual R2 value, comparing imputed and true genotypes.

These statistics can be found in the *.info file

Be aware that, unfortunately, imputation quality statistics are not directly comparable between different imputation programs (MaCH/minimac vs. Impute vs. Beagle etc.).

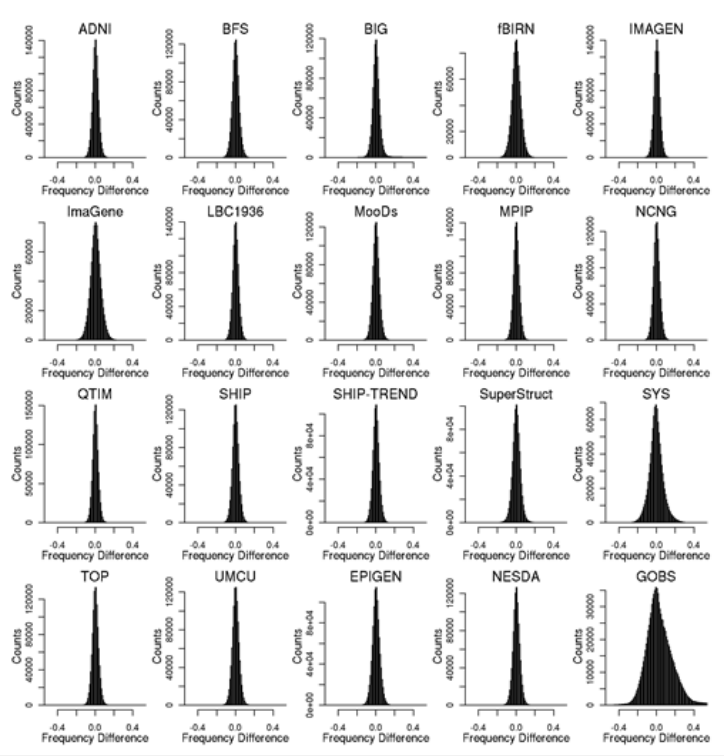
Output

- 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)

```
##fileformat=VCFv4.1
##filedate=2015.7.12
##source=Minimac3
##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=MAF,Number=1,Type=Float,Description="Estimated Alternate 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)">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  A0001_A0001     A0003_A0003     A0004_A0004     A0007_A0007     A0008_A0008     A0009_A0009     A0010_
10      27754636        10:27754636     C       G       .       PASS    MAF=0.00032;R2=0.81788  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27754678        10:27754678     G       A       .       PASS    MAF=0.00042;R2=0.77190  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27754849        10:27754849     C       G       .       PASS    MAF=0.00001;R2=0.00262  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27754857        10:27754857     T       C       .       PASS    MAF=0.00120;R2=0.72916  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27754954        10:27754954     T       C       .       PASS    MAF=0.11410;R2=0.97841  GT:DS:GP        1/1:2.000:0.000,0.000,1.000     1/1:2.000:0.000,0.000,1.000
10      27755014        10:27755014     G       T       .       PASS    MAF=0.00000;R2=0.00082  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755016        10:27755016     C       T       .       PASS    MAF=0.00003;R2=0.01909  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755047        10:27755047     T       C       .       PASS    MAF=0.02255;R2=0.87665  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755175        10:27755175     C       T       .       PASS    MAF=0.00004;R2=0.13821  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755281        10:27755281     C       T       .       PASS    MAF=0.00061;R2=0.86168  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755330        10:27755330     A       G       .       PASS    MAF=0.00273;R2=0.90295  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755439        10:27755439     A       C       .       PASS    MAF=0.00000;R2=0.00138  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
10      27755489        10:27755489     C       A       .       PASS    MAF=0.00003;R2=0.39172  GT:DS:GP        0/0:0.000:1.000,0.000,0.000     0/0:0.000:1.000,0.000,0.000
```

Post imputation QC

- After imputation you need to check that it worked and the data look ok
- Things to check
  - Plot r^2 across each chromosome look to see where it drops off
  - Plot MAF-reference MAF



Issue – the r^2 metrics differ between imputation programs

The MACH $\hat{r}^2$ measure

This is the ratio of the empirically observed variance of the allele dosage to the expected binomial variance at Hardy-Weinberg equilibrium. At the jth SNP this is defined as

$$\hat{r}_j^2 = \begin{cases} \dfrac{\frac{\sum_{i=1}^{N} e_{ij}^2}{N} - \left(\frac{\sum_{i=1}^{N} e_{ij}}{N}\right)^2}{2\hat{\theta}(1-\hat{\theta})} & \text{when } \hat{\theta} \in (0,1) \\ 1 & \text{when } \hat{\theta} = 0, \hat{\theta} = 1 \end{cases} \quad (1)$$

When all the genotypes are predicted with high certainty this ratio will be close to 1, although it can go above 1 (Figure 1). As the amount of uncertainty increases the allele dosages will tend to 2θ, the empirical variance will tend to 0 and so $\hat{r}^2$ tends to 0.

The IMPUTE info measure I_A

This is based on measuring the relative statistical information about the population allele frequency, θ_j. If the G_{ij}'s were observed then the full data likelihood is given by

$$L(\theta_j) = \prod_{i=1}^{N} \theta_j^{G_{ij}} (1-\theta_j)^{2-G_{ij}} \tag{10}$$

For this likelihood the score and information are given by

$$U(\theta_j) = \frac{d \log L(\theta_j)}{d\theta_j} = \frac{X - 2N\theta_j}{\theta_j(1-\theta_j)} \tag{11}$$

$$I(\theta_j) = \frac{-d^2 \log L(\theta_j)}{d\theta_j^2} = \frac{X}{\theta_j^2} + \frac{2N - X}{(1-\theta_j)^2} \tag{12}$$

The IMPUTE info measure is based on the same idea used to calculate the SNPTEST information measure i.e. the ratio of the observed and complete information.

$$I_A = \frac{\mathbb{E}_{G.j}[I(\hat{\theta})] - V_G[U(\hat{\theta})]}{\mathbb{E}_{G.j}[I(\hat{\theta})]} \tag{13}$$

where the expectations are taken over the imputed genotype distribution and evaluated at the allele frequency estimate, $\hat{\theta}_j$. The exact terms are given by

$$\mathbb{E}_{G.j}[I(\hat{\theta})] = \frac{2N}{\hat{\theta}(1-\hat{\theta})} \tag{14}$$

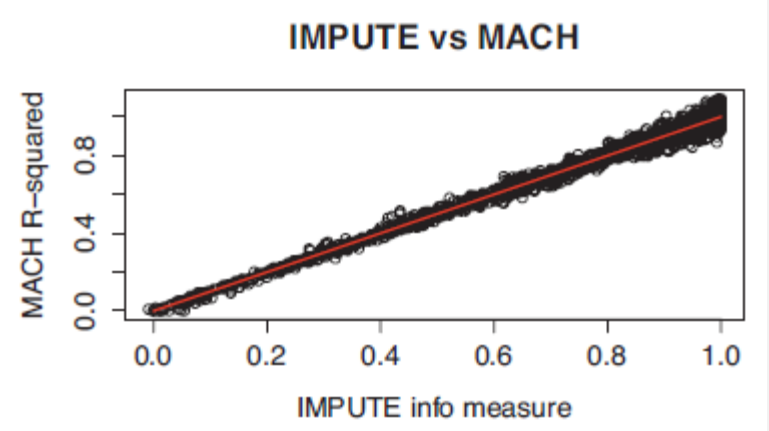
$$V_G[U(\hat{\theta})] = \frac{\sum_{i=1}^{N}(f_{ij} - e_{ij}^2)}{\hat{\theta}^2(1-\hat{\theta})^2} \tag{15}$$

so that

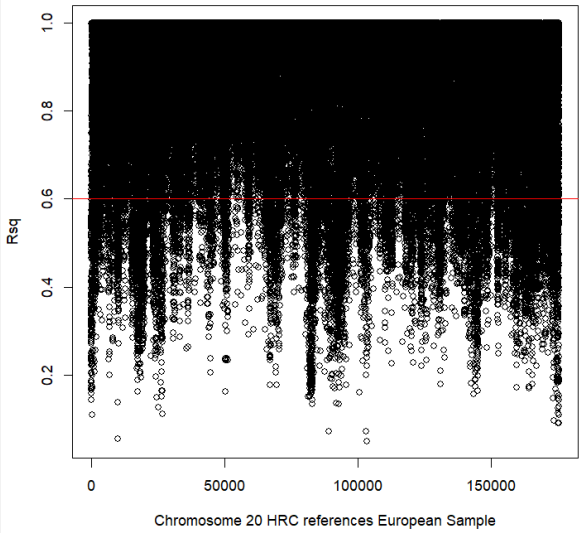
$$I_A = \begin{cases} 1 - \frac{\sum_{i=1}^{N}(f_{ij} - e_{ij}^2)}{2N\hat{\theta}(1-\hat{\theta})} & \text{when } \hat{\theta} \in (0,1) \\ 1 & \text{when } \hat{\theta} = 0, \hat{\theta} = 1. \end{cases} \tag{16}$$

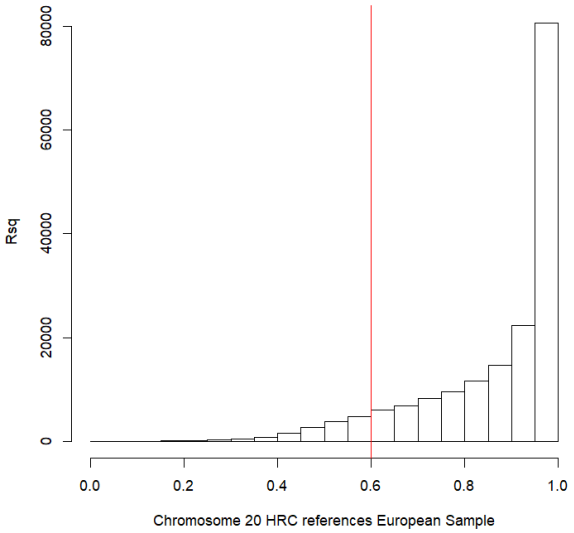
So I_A is bounded above at 1 and will equal 0 when the sample mean variance of the imputed genotypes equals the variance you would expect if alleles where sampled with frequency $\hat{\theta}$.

- In general fairly close correlation
  - rsq/ ProperInfo/ allelic Rsq
    - 1 = no uncertainty
    - 0 = complete uncertainty
    - .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

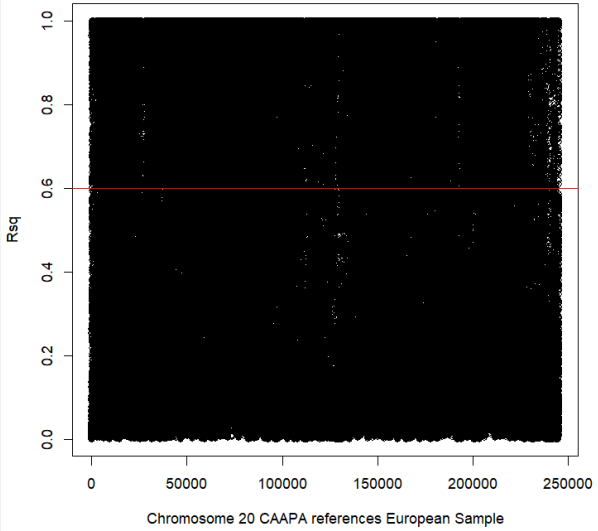


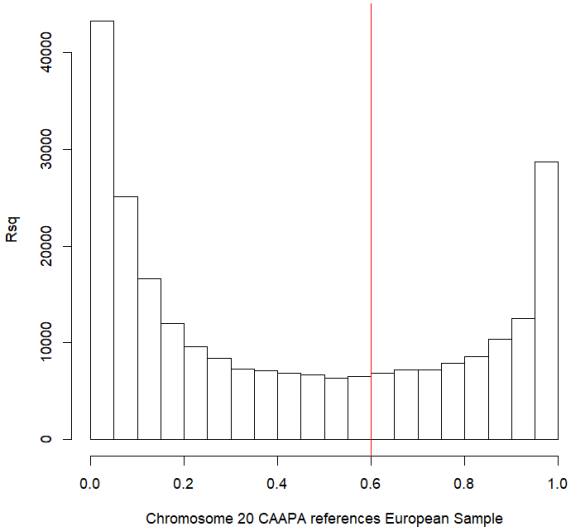
Good imputation





Bad imputation





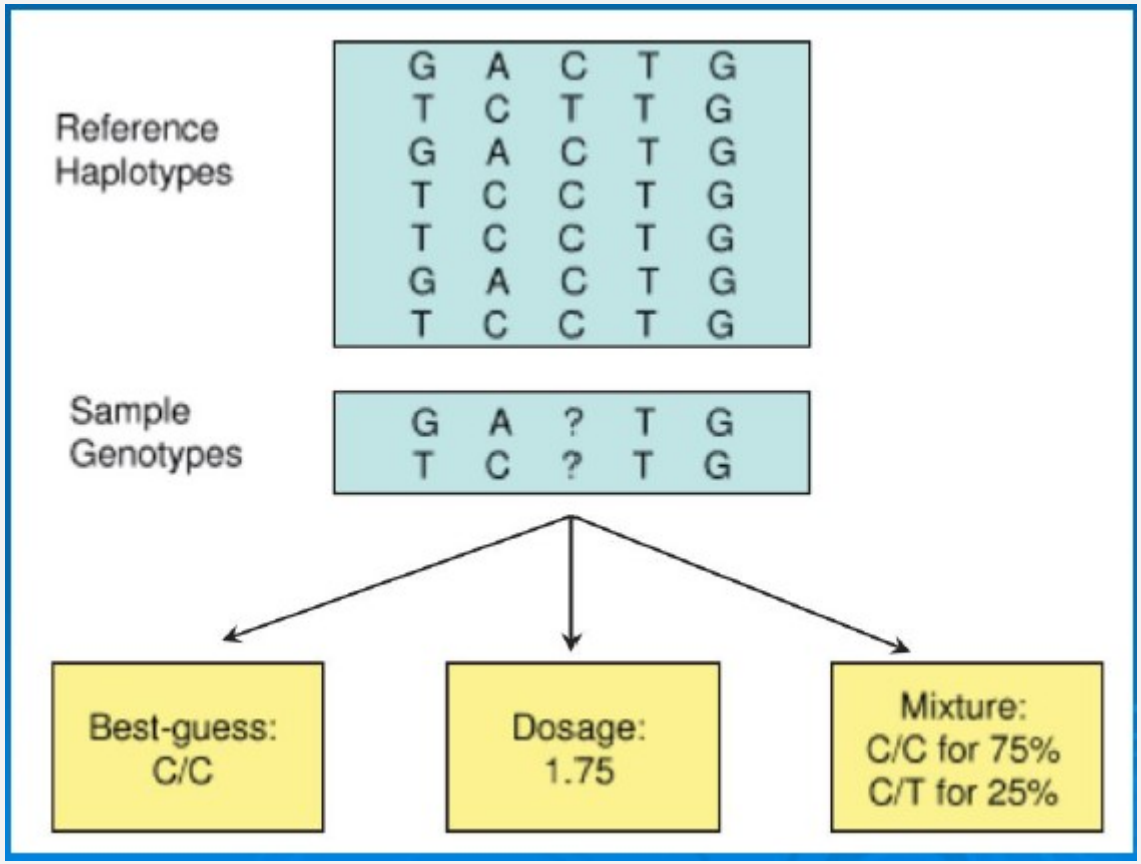
Post imputation QC

- Next run GWAS for a trait – ideally continuous, calculate lambda and plot:
  - QQ
  - Manhattan
  - SE vs N
  - P vs Z
- Run the same trait on the observed genotypes – plot imputed vs observed

If you are running analyses for a consortium they will probably ask you to analyse all variants regardless of whether they pass QC or not…

(If you are setting up a meta-analysis consider allowing cohorts to ignore variants with MAF <.5% and low r2 – it will save you a lot of time)

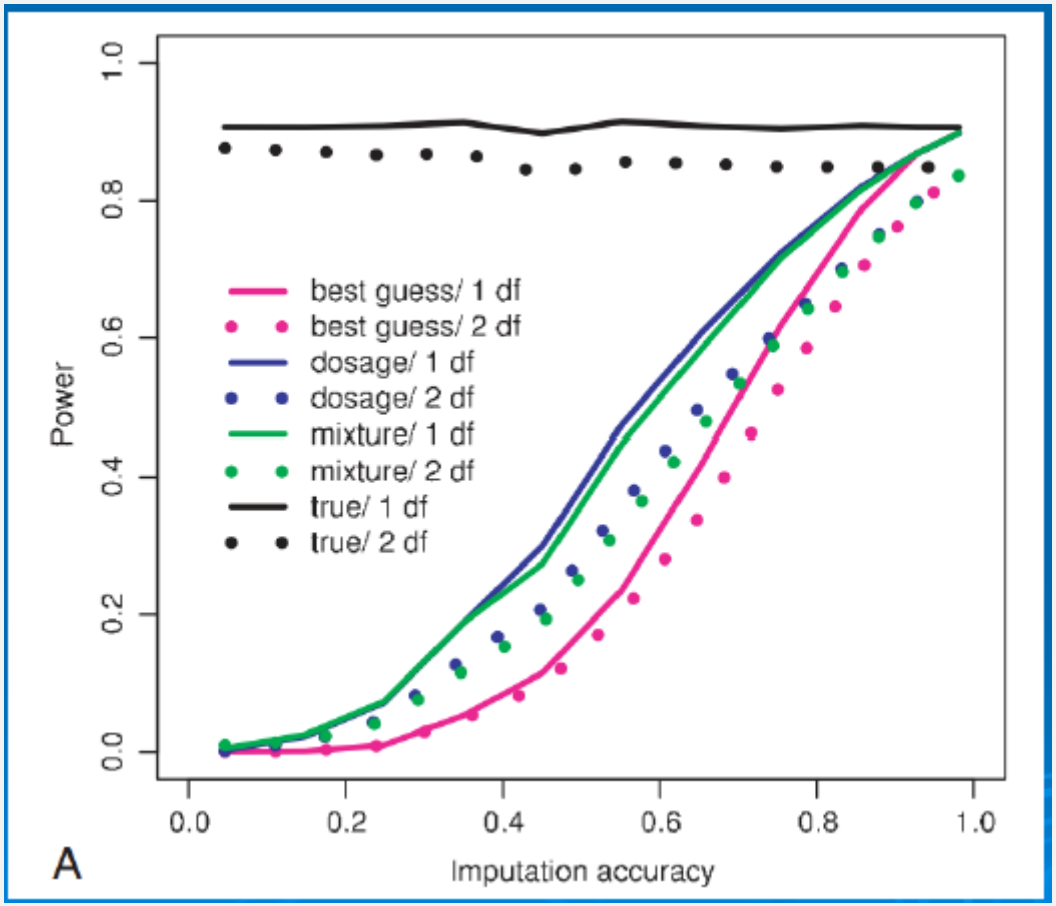
Choices of analysis methods



Choices for Analysis

Scenario	N	H^2	Power: Best Guess	Power: Dosage	Power: Mixture
Large sample, small effect					
	1000	3%	63.5%	66.0%	66.8%
Small sample, large effect					
	50	60%	70.1%	75.5%	85.0%

- When effect sizes are small, difference between dosage and mixture models becomes even smaller


Power
0.0
0.2
0.4
0.6
0.8
1.0
best guess/ 1 df
best guess/ 2 df
dosage/ 1 df
dosage/ 2 df
mixture/ 1 df
mixture/ 2 df
true/ 1 df
true/ 2 df
0.0
0.2
0.4
0.6
0.8
1.0
Imputation accuracy
A

Analysis of Imputed data using SAIGE

Why SAIGE

PROS

- Fast
- Low memory
- Can cope with UKB size data
- Correctly models zygosity and relatedness
- Continuous or Binary

CONS

- 2 stages
- 1 trait at a time
- Phenotypes and covariates all in one file

Step 1

- Runs the LMM and creates a pre-processed r data file

```
Rscript step1_fitNULLGLMM.R
--plinkFile=sparse
--phenoFile=phenoAD.txt
--phenoCol=AD
--covarColList=PC1,PC2,PC3,PC4
--sampleIDColinphenoFile=FID_IID
--traitType=binary
--outputPrefix=AD > AD.log
```

Phenotype/Covariate file

- Simple format – headers, space delim
- Binary traits are 0/1
- Need to join FID and IID with an underscore _ as this is the ID format in the imputed data

```
FID_IID AD PC1 PC2 PC3 PC4
WGAAD_10 1 0.0550949 0.0507711 0.00845787 -0.00116914
WGAAD_15 1 0.0470604 0.0474843 0.00315769 0.00810905
WGAAD_18 1 0.0564277 0.0471303 0.00803162 -0.00242266
WGAAD_20 1 0.0564051 0.0436962 0.00419304 -0.007482
WGAAD_24 1 0.0540288 0.0477145 0.00711973 -0.00223988
WGAAD_25 1 0.0475798 0.0504094 0.00207224 0.00637812
WGAAD_28 1 0.0570727 0.0493075 0.00609508 -0.00164
WGAAD_29 1 0.054579 0.0496459 0.00995307 -0.00327564
WGAAD_31 1 0.0552207 0.0516809 0.00705046 -0.00485599
```

Plink (hard call) genotypes for relatedness estimation

- Doesn’t need to be all available data – can be a sparse file
  - Todays files contains ~75,000 snps

```
plink  --bfile QCed  --indep-pairwise 10000kb 5 .2
--out prune
plink  --bfile QCed  --extract prune.prune.in
--make-bed  --out sparse
```

Step 1

- Runs the LMM and creates a pre-processed r data file

```
Rscript step1_fitNULLGLMM.R
--plinkFile=sparse
--phenoFile=phenoAD.txt
--phenoCol=AD
--covarColList=PC1,PC2,PC3,PC4
--sampleIDColinphenoFile=FID_IID
--traitType=binary
--outputPrefix=AD > AD.log
```

It will make 4 files

- Log, Rda, varianceRatio.txt, _30markers.SAIGE.results.txt

- `R`
- `load("example.rda")`
- `names(modglmm)`
- `modglmm$theta`

- #theta: a vector of length 2. The first element is the dispersion parameter estimate and the second one is the variance component parameter estimate
- #coefficients: fixed effect parameter estimates
- #linear.predictors: a vector of length N (N is the sample size) containing linear predictors
- #fitted.values: a vector of length N (N is the sample size) containing fitted mean values on the original scale
- #Y: a vector of length N (N is the sample size) containing final working vector
- #residuals: a vector of length N (N is the sample size) containing residuals on the original scale

Step 2

```
Rscript step2_SPAtests.R
--vcfFile=chr19.dose.vcf.gz
--vcfFileIndex=chr19.dose.vcf.gz.tbi
--sampleFile=sample.txt
--vcfField=DS
--chrom=19
--minMAF=0.01 --minMAC=5
--GMMATmodelFile=AD.rda
--varianceRatioFile=AD.varianceRatio.txt
--SAIGEOutputFile=AD.chr19.SAIGE.txt
--numLinesOutput=2
--IsOutputAFinCaseCtrl=TRUE &> AD.chr19.SAIGE.log
```



```
WGAAD_10
WGAAD_15
WGAAD_18
WGAAD_20
WGAAD_24
WGAAD_25
WGAAD_28
WGAAD_29
```

Output

CHR	POS	SNPID	Allele1	Allele2	AC_Allele2	AF_Allele2	N	BETA	SE	Tstat	p.value	varT	varTstar
19	89282	19:89282	C	T	268.85693359375	0.381899058818817	352	-0.243892512960461	0.211526149486516	-0.332437280097603	0.248904829858843	0.(	
19	95981	19:95981	G	A	61.9459915161133	0.0879914686083794	352	0.0343642743023006	0.484427618730728	0.0186055615054843	0.943447185535922	(	
19	105021	19:105021	G	C	61.5499954223633	0.0874289721250534	352	0.153042756015932	0.493144880976199	0.0802139137592557	0.756302148310946		
19	240554	19:240554	C	T	16.2939929962158	0.0231448765844107	352	-0.162009508677698	0.360888521999669	-0.308163077838583	0.653490500578882		
19	240963	19:240963	G	A	15.2640085220337	0.0216818302869797	352	-0.204254254224956	0.23975796353646	-0.909474922804863	0.394259305960713		
19	249357	19:249357	C	T	54.6259994506836	0.0775937512516975	352	-0.202778104360785	0.199395348215412	-0.690067090576171	0.309170046189126		
19	251430	19:251430	G	C	16.2120018005371	0.0230284109711647	352	0.122260719602352	1.16206157903867	0.0224859034252765	0.91620902509518	0.	
19	253155	19:253155	T	C	52.0609817504883	0.0739502608776093	352	-0.20922129536113	0.206403088423198	-0.68063977874865	0.31074792911264	0.	
19	253938	19:253938	A	G	422.252685546875	0.599790751934052	352	0.0212743717510442	0.0402777842365444	0.781260215526934	0.597366626898937		
19	254304	19:254304	A	G	418.611663818359	0.594618856906891	352	0.0208301822368687	0.0399736958538765	0.771660646056235	0.602299039544212		
19	254448	19:254448	T	C	422.284973144531	0.599836587905884	352	0.0212370063691577	0.0402133197303094	0.782435302299136	0.597423851974419		
19	254899	19:254899	A	G	418.603393554688	0.59460711479187	352	0.0208251345591604	0.0399037992918946	0.774167470835337	0.601751445122865	2	

CHR: chromosome

POS: genome position

SNPID: variant ID

Allele1: Ref allele

Allele2: Alt allele

AC_Allele2: allele count of Alt allele

AF_Allele2: allele freq of Alt allele

N: sample size

BETA: effect size **of A2**

SE: standard error of BETA

Tstat: score statistic

p.value: p value with SPA applied

p.value.NA: p value when SPA is not applied

Is.SPA.converge: whether SPA is converged or not

varT: estimated variance of score statistic with sample related incorporated

varTstar: variance of score statistic without sample related incorporated

Merge the GWAS results with the R2 from the imputation files and keep only those results with r2 >= 0.6 (i.e. good imputation quality)

Merge results GWAS with imputed information file and filter by R2 >= 6 (the results have been filtered by MAF before)

Extract columns of interest in GWAS results: CHR POS SNPID Allele1 Allele2 AF_Allele2 N BETA SE Tstat p.value

```
zcat AD.chr19.SAIGE.txt.gz | awk '{print $1,$2,$3,$4,$5,$7,$8,$9,$10,$11,$12}' > temp_AD.chr19.results.txt
```

Extract columns of interest in imputation information: SNP Rsq

```
zcat ../saige/chr19.info.gz | awk '{print $1,$7}' > temp_chr19.info.txt
```

Merge information from the GWAS and R2

```
awk 'NR==FNR{a[$1]=$2; next} $3 in a{print $0,a[$3]}' temp_chr19.info.txt temp_AD.chr19.results.txt > temp
```

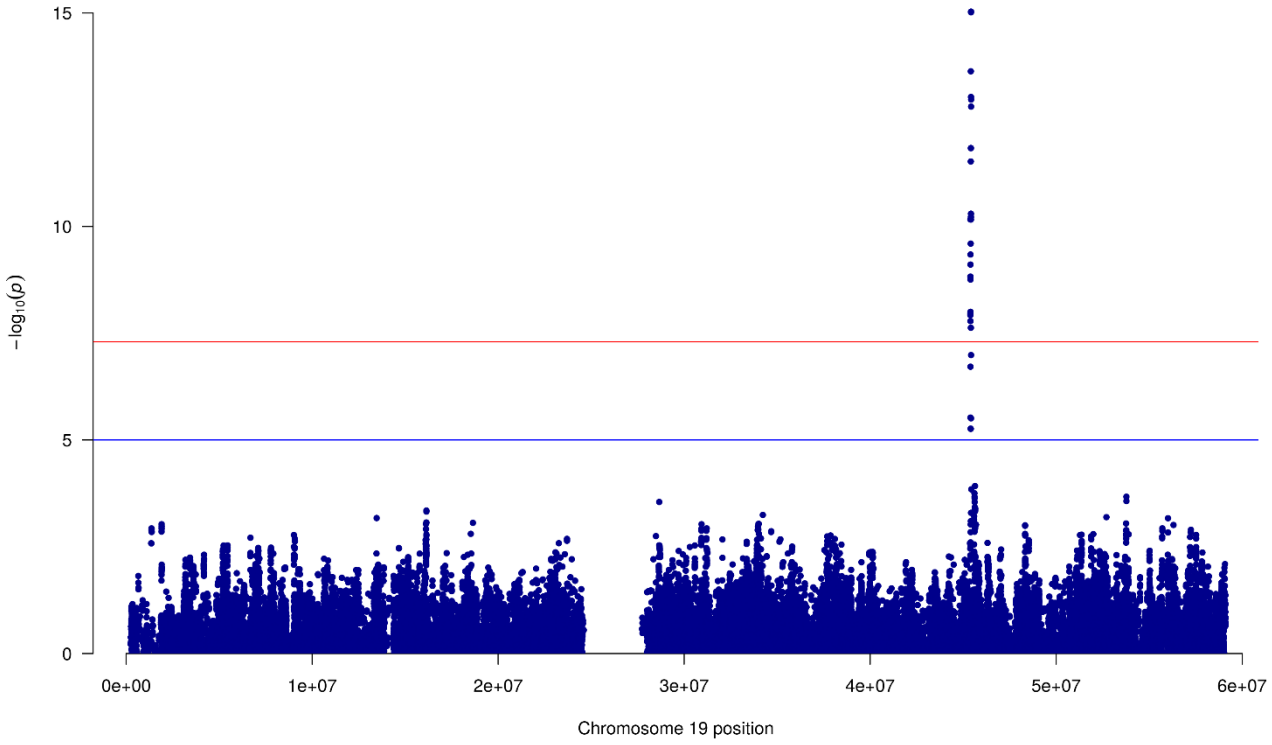
Filer by R2 >= 0.6 and put the header

```
echo 'CHR POS SNPID Allele1 Allele2 AF_Allele2 N BETA SE Tstat p.value Rsq' > AD.chr19.results.QC.txt
awk '{if ($12>=0.6) print $0}' temp >> AD.chr19.results.QC.txt
```

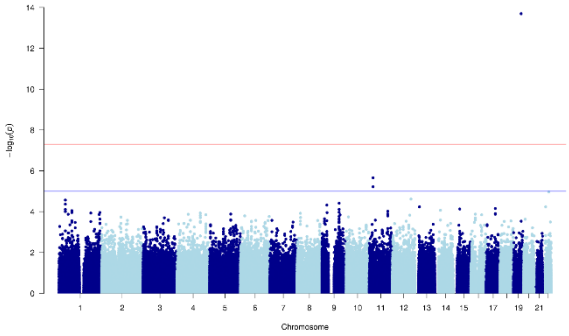
Plot the results: Manhattan plot and regional plot

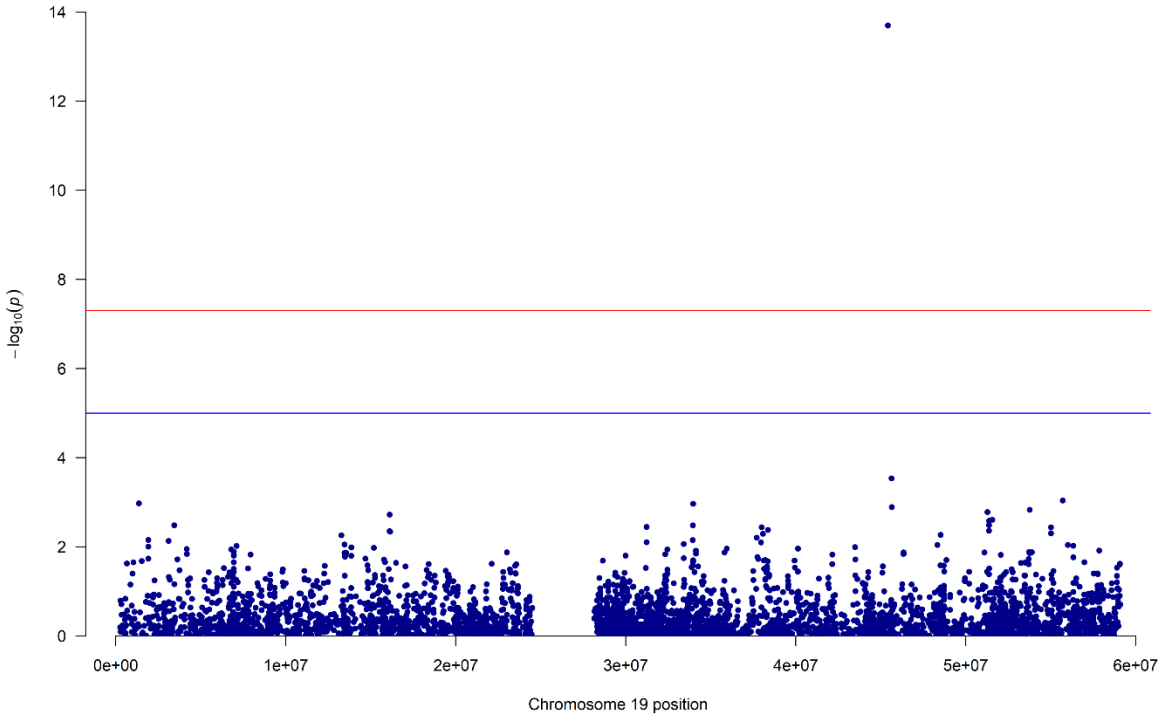
```
awk '{if (NR>1) print $1,$2,$11}' temp_AD.chr19.results.QC.txt > temp_plot.AD.chr19.txt
sort -k11 -g temp_AD.chr19.results.QC.txt | head # lowest p-value at 19:45422946 , p = 9.3515193823689e-16
awk '{if (NR==1) print $1,$2,$3,$11; else if ($2>= 44922946 && $2<= 45922946) print $1,$2,"chr"$3,$11}' AD.chr19.results.QC.txt > temp_ld.AD.chr19.txt
```

Imputed results (Manhattan plot only chr19)

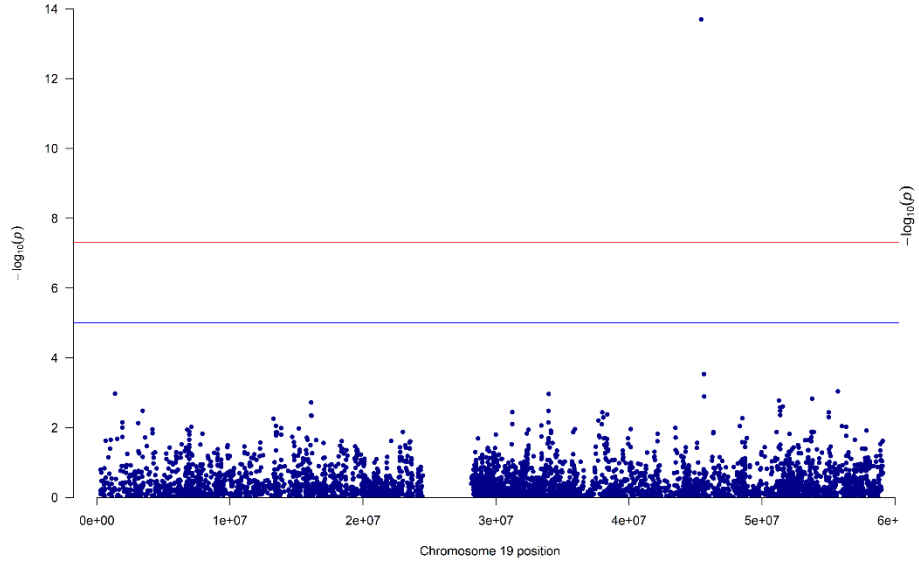


Remember the genotyped results!

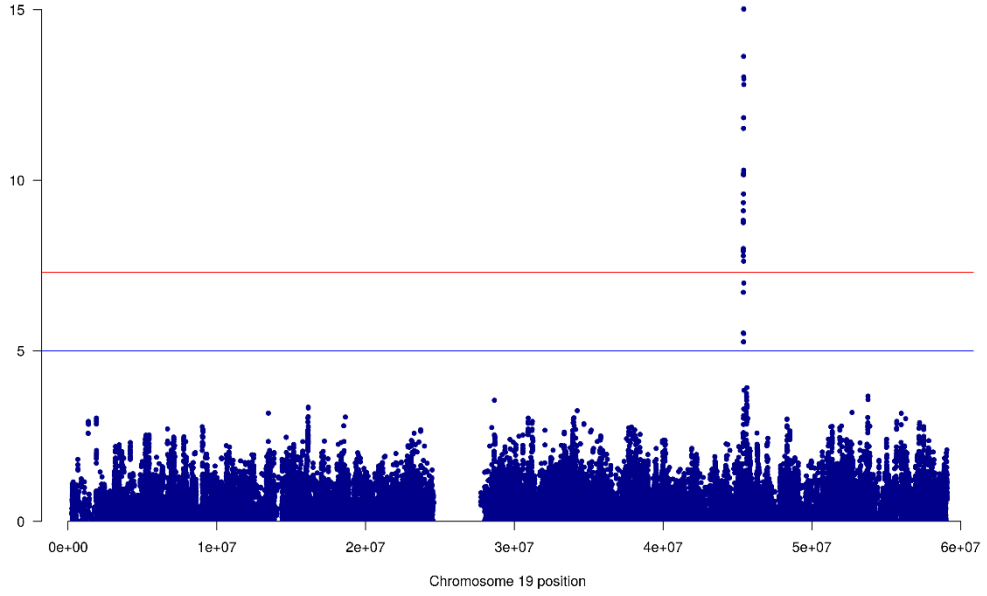




And Compare…

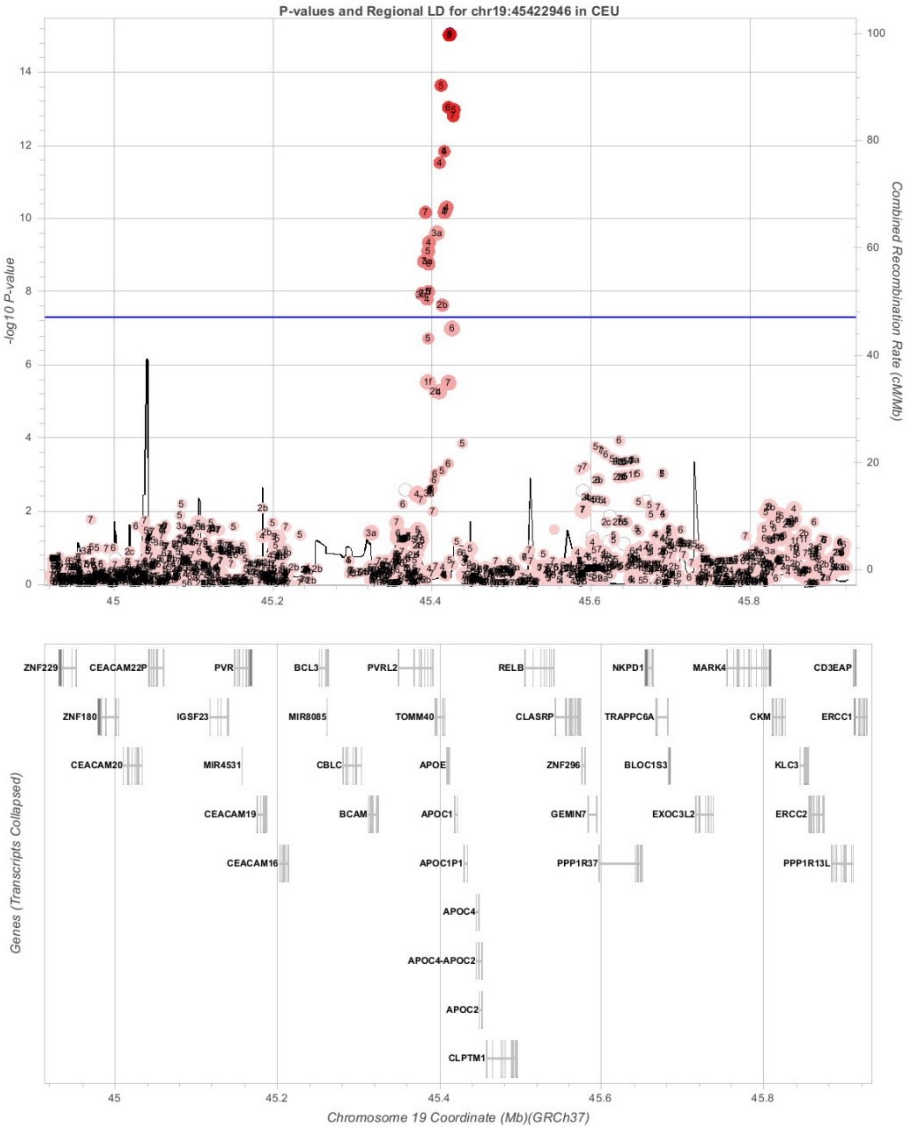


Genotyped

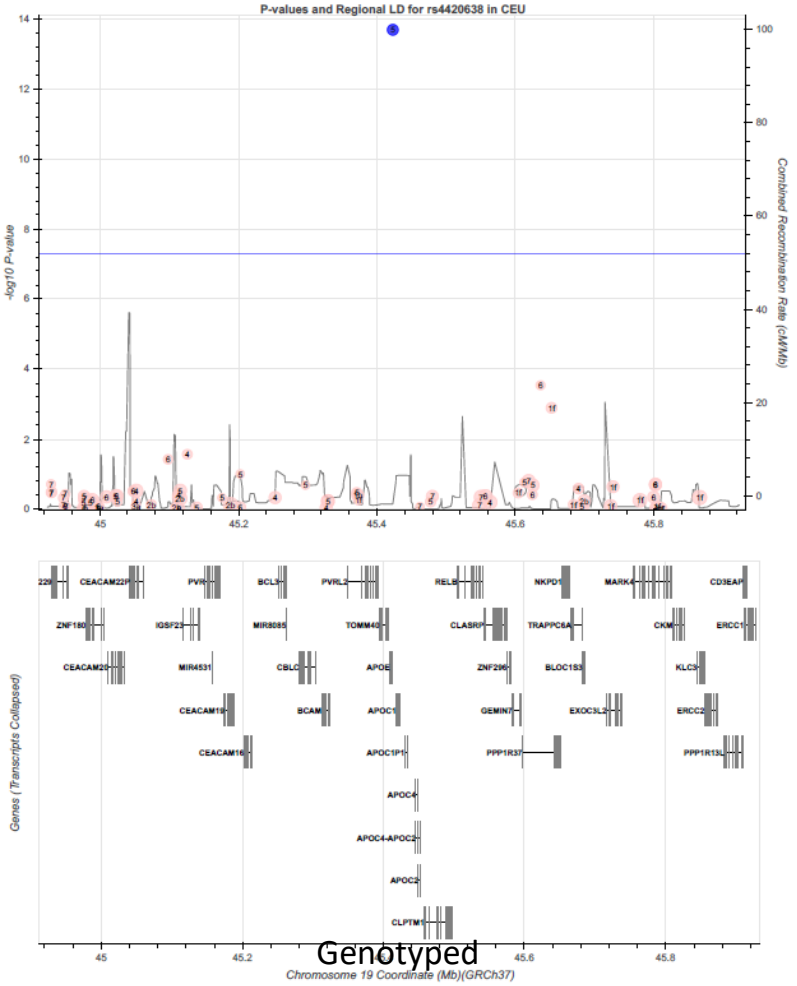


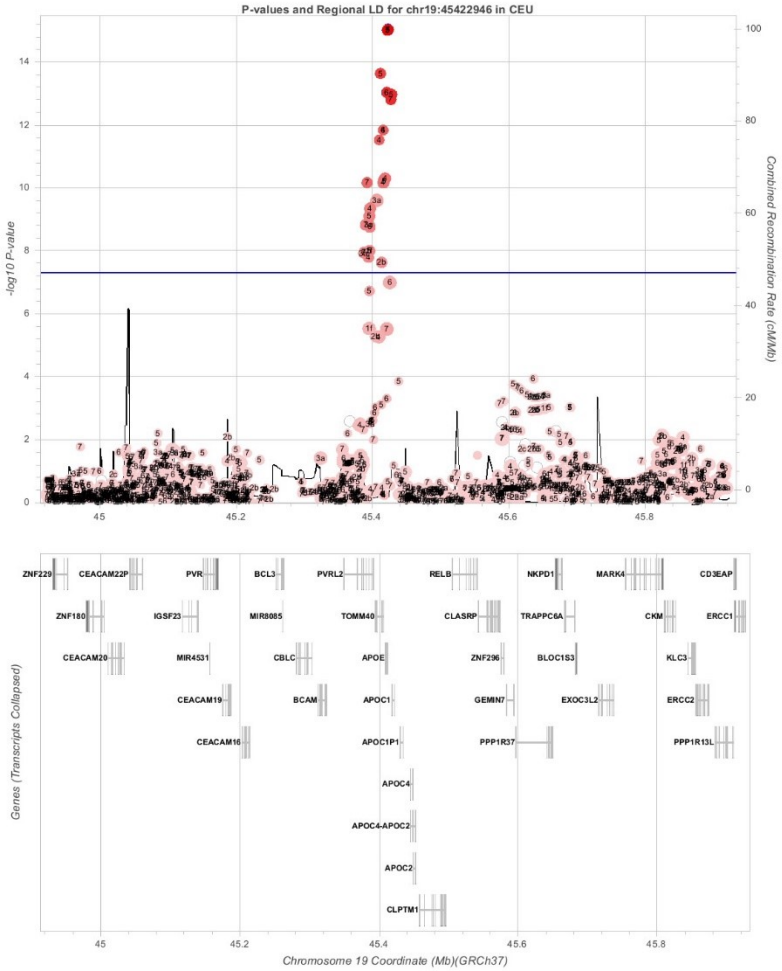
Imputed

Imputed results (regional plot)



And Compare…


P-values and Regional LD for rs4420638 in CEU
-log10 P-value
Combined Recombination Rate (cM/Mb)
Genes (Transcripts Collapsed)
Chromosome 19 Coordinate (Mb)(GRCh37)
Genotyped


P-values and Regional LD for chr19:45422946 in CEU
-log10 P-value
Combined Recombination Rate (cM/Mb)
Genes (Transcripts Collapsed)
Chromosome 19 Coordinate (Mb)(GRCh37)

Questions?

Extra slides

Step 1 – How to phase

Data is usually broken into manageable chunks
~20Mb
Each phased independently

```
./eagle
--vcfRef HRC.r1-
1.GRCh37.chr20.shapeit3.mac5.aa.genotypes.bcf
--vcfTarget
chunk_20_0000000001_0020000000.vcf.gz
--geneticMapFile
genetic_map_chr20_combined_b37.txt
--outPrefix
chunk_20_0000000001_0020000000.phased
--bpStart 1
--bpEnd 25000000
--chrom 20
--allowRefAltSwap
```

Step 2 - Imputation

- Compares the phased data to the references and infers the missing genotypes. Calculate accuracy metrics

```
./Minimac3
--refHaps HRC.r1-
1.GRCh37.chr1.shapeit3.mac5.aa.genotypes.m3vcf.gz
--haps chunk_1_0000000001_0020000000.phased.vcf
--start 1
--end 20000000
--window 500000
--prefix chunk_1_0000000001_0020000000
--chr 20
--noPhoneHome
--format GT,DS,GP
--allTypedSites
```

Phasing in Eagle

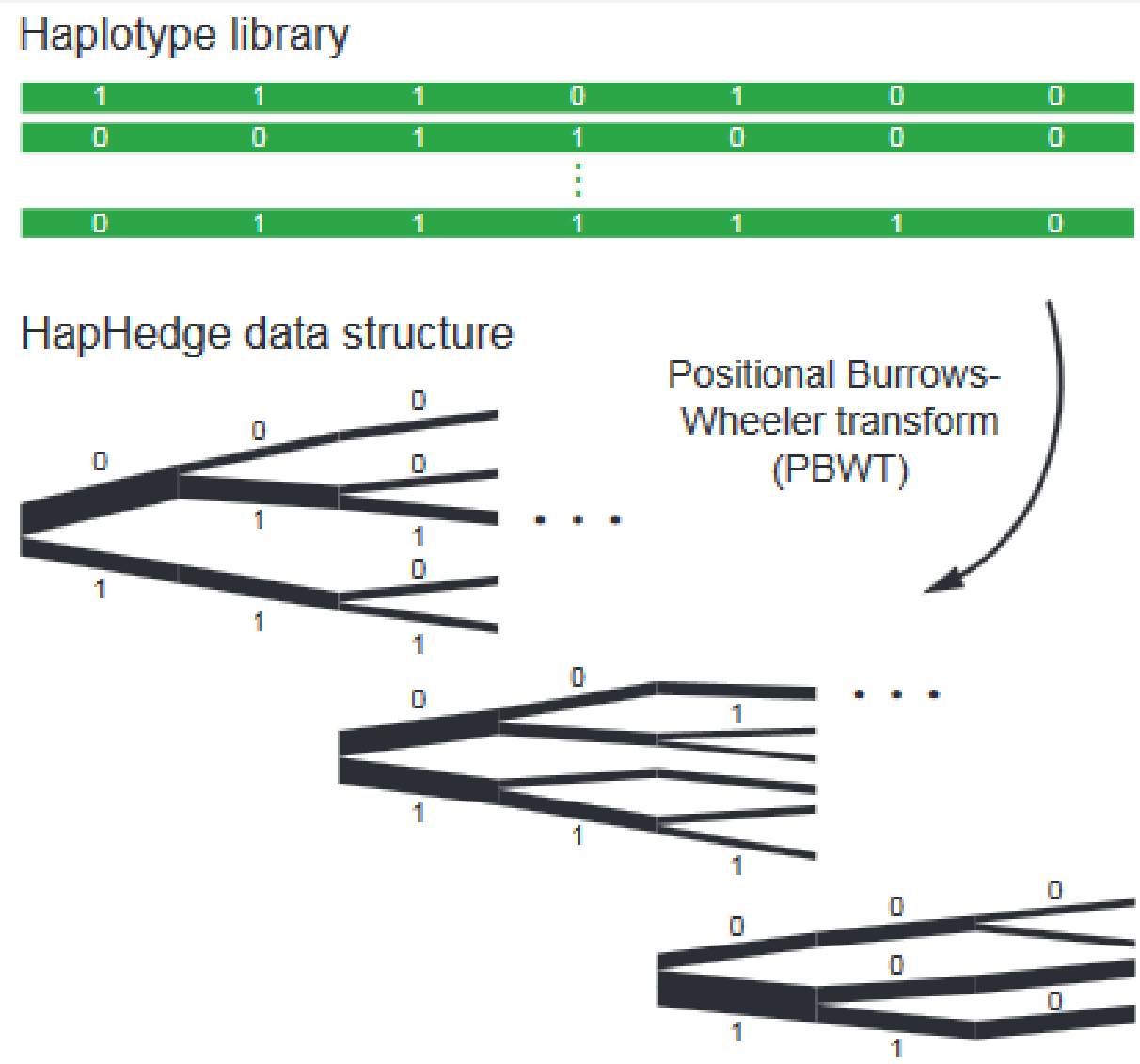
- Input a target sample and a library of reference haplotypes
- *Selection of conditioning haplotypes.*

  Eagle2 first identifies a subset of 10,000 conditioning haplotypes by ranking reference haplotypes according to the number of discrepancies between each reference haplotype and the homozygous genotypes of the target sample.

- *Generation of HapHedge data structure.*

  Eagle2 next generates a HapHedge data structure on the selected conditioning haplotypes.

  The HapHedge encodes a sequence of haplo- type prefix trees (i.e., binary trees on haplotype prefixes) rooted at a sequence of starting positions along the chromosome, thus enabling fast lookup of haplotype frequencies


Haplotype library
1 1 1 0 1 0 0
0 0 1 1 0 0 0
0 1 1 1 1 1 0
HapHedge data structure
Positional Burrows-Wheeler transform (PBWT)

- *Exploration of the diplotype space.*

  Having prepared a HapHedge of conditioning haplotypes, Eagle2 performs phasing using a hidden Markov model

  Consolidates reference haplotypes sharing common prefixes reducing computation.

Diploid genotypes of target sample

Diplotype probability computation