Imputation

Sarah Medland

The 2023 International Statistical Genetics Workshop

Session Objectives

By the end of this session, if you stay awake, you might be able to:

Explain how you can use imputation

Gain familiarity with the tasks addressed within each step involved in imputation

Identify the software commonly used for imputation

Summarize considerations when using in-house vs. online imputation

Evaluate imputation output and common metrics of imputation

cp /faculty/sarah/Imputation/qualtrics.txt to your working directory

Open up the Qualtrics and start the imputation

Why do we impute?

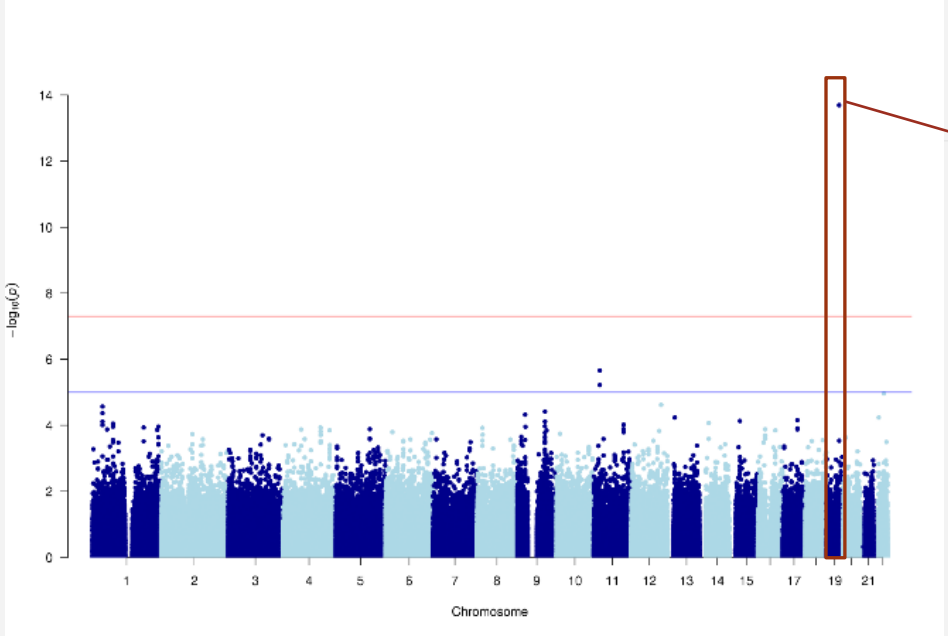
3 main reasons for imputation

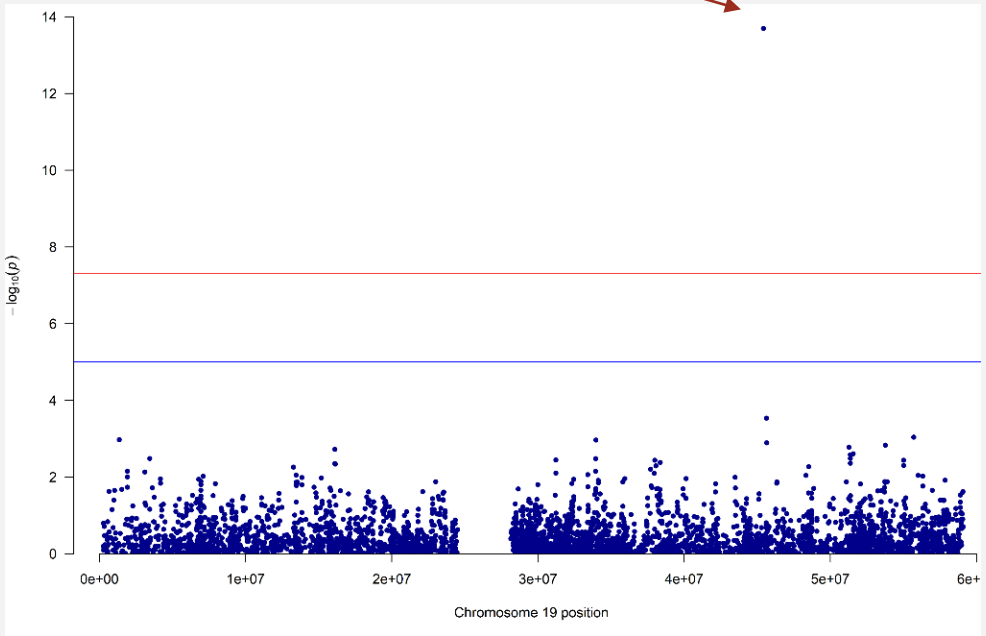
- Meta-analysis
- Fine Mapping
- Combining data from different chips

Other less common uses

- Sporadic missing data imputation
- Correction of genotyping errors
- Imputation of non-SNP variation

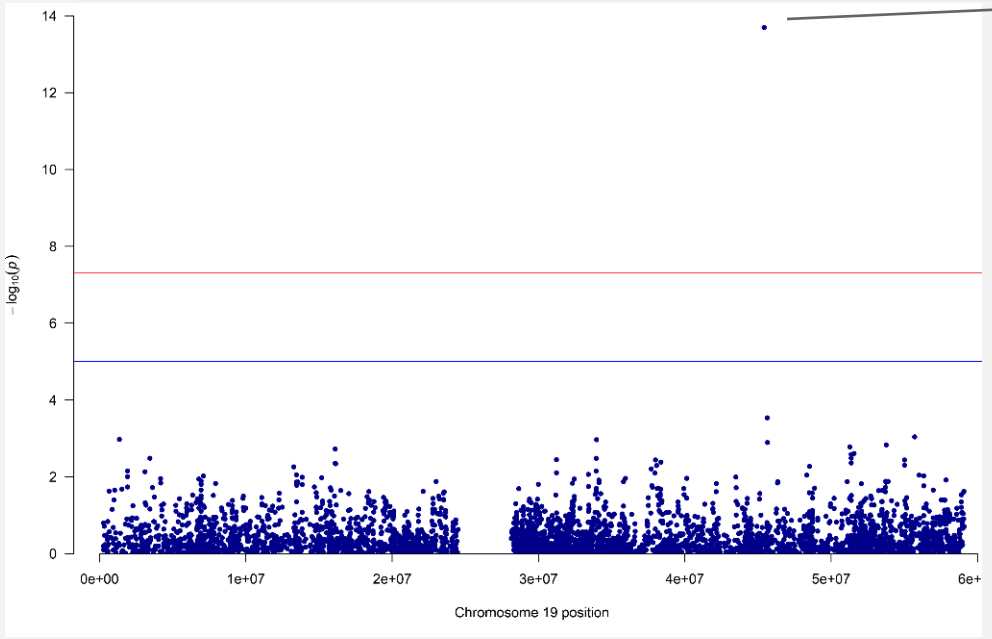
Fine Mapping



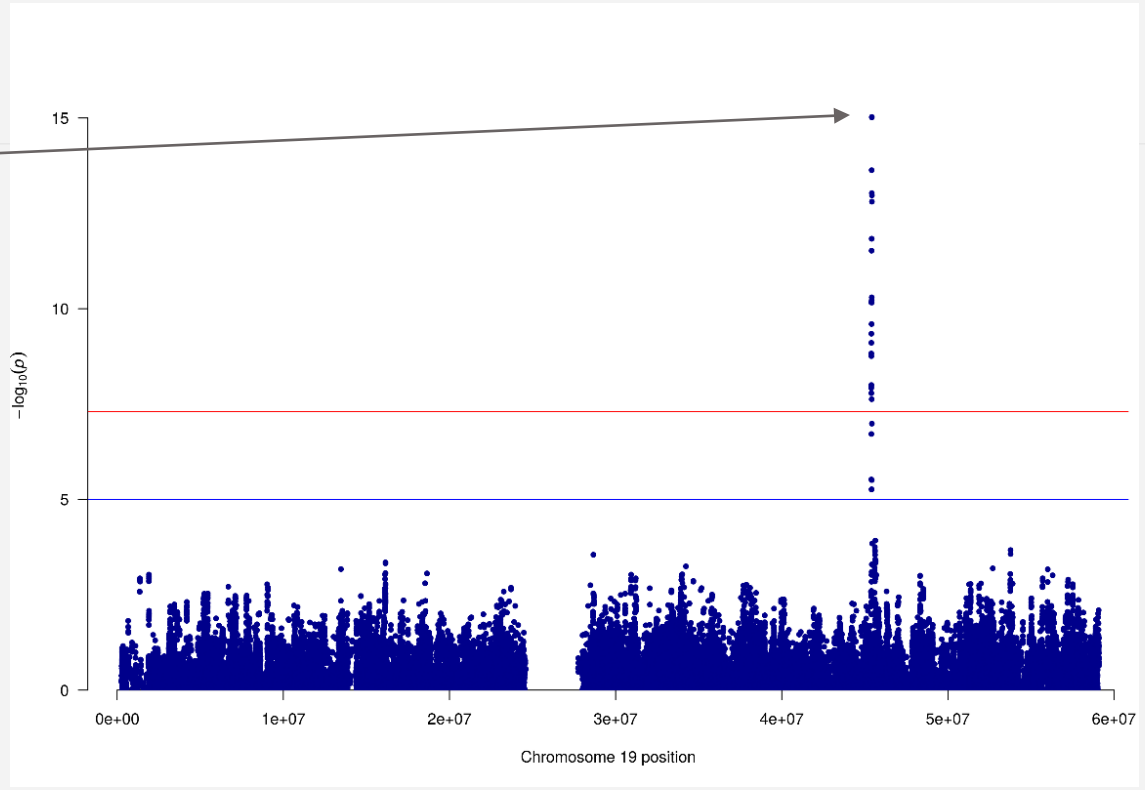


GWAS using only genotyped SNPs

Fine Mapping

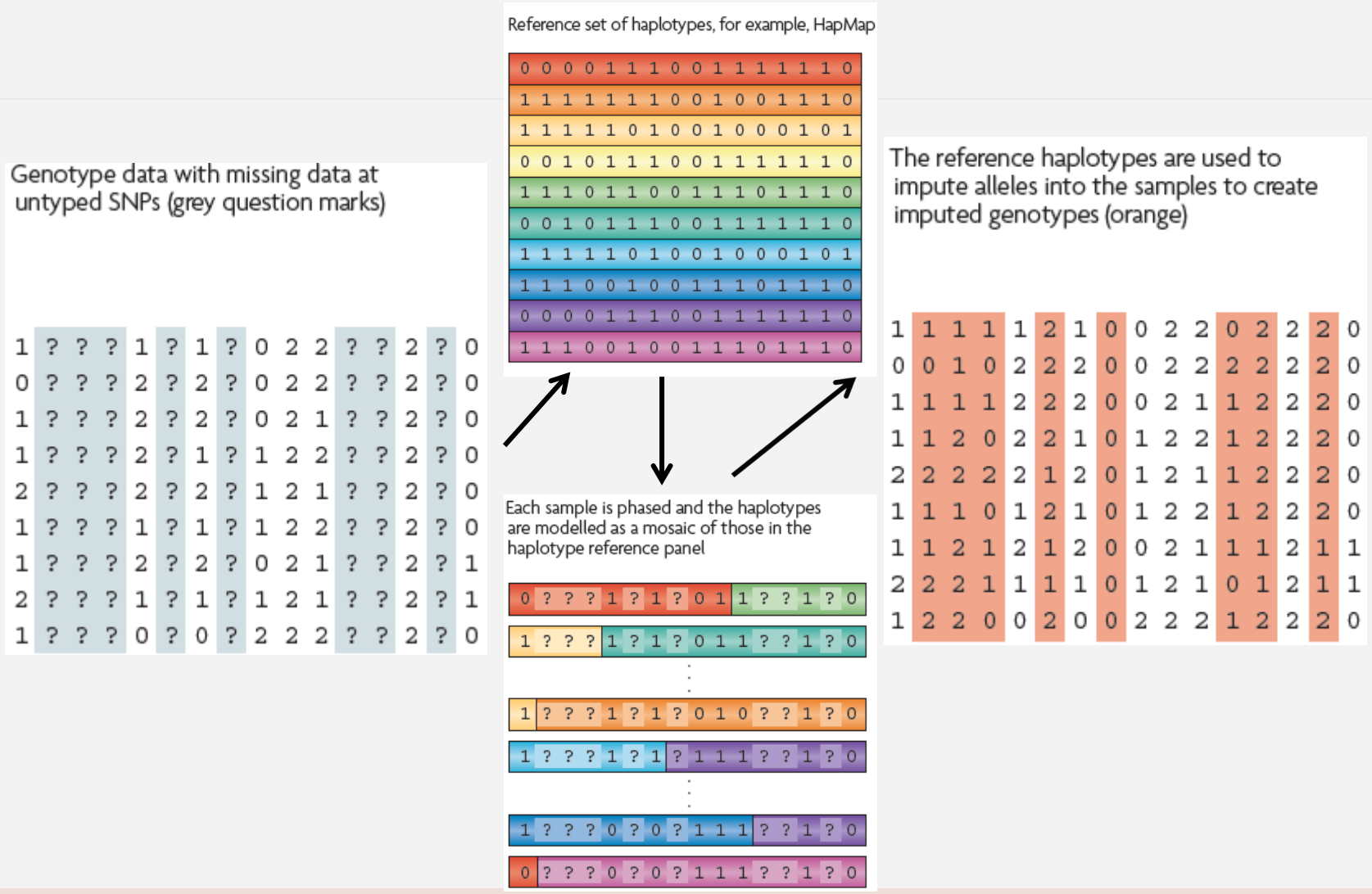


Genotyped only



Post-Imputation

What is imputation? (Marchini & Howie 2010)



1. Starting Data

Genotyped sample

. . **C** . . **G** . **C** .

Reference haplotypes

A G A T C T C C T

A G C T C T C A T

A G A T C G C C T

A G A T C T A C T

2. Identify shared regions of chromosome

Genotyped sample

.	.	**C**	.	.	**G**	.	**C**	.

Reference haplotypes

A	G	A	T	C	T	C	C	T
A	G	**C**	T	C	T	C	A	T
A	G	A	T	C	**G**	C	**C**	T
A	G	A	T	C	T	A	C	T

3. Fill in missing genotypes

Genotyped sample

A	G	**C**	T	C	**G**	C	**C**	T

Reference haplotypes

A	G	A	T	C	T	C	C	T
A	G	**C**	T	C	T	C	A	T
A	G	A	T	C	**G**	C	**C**	T
A	G	A	T	C	T	A	C	T

Step 1 – QC & references

Current Publically Available References

- HapMapII (no phased X data officially released) 2.4M SNPS
- HapMapIII 1.3M SNPS
- 1KGP – phase 3 version v5 (X imputation references released after the autosomes)

Step 1 – QC & references

References only available via custom imputation servers

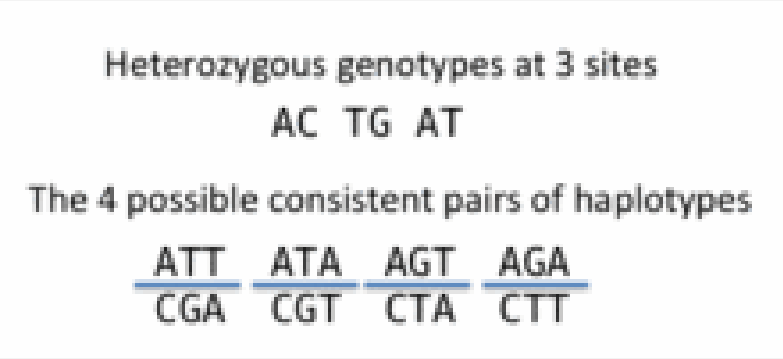
- HRC - 64,976 haplotypes 39,235,157 SNPs
- CAPPA – African American/Carabbean
- Multi-ethnic HLA
- Genome Asia Pilot - GAsP
- TopMed - 97,256 haplotypes 308,107,085 SNPs (b38)

Step 2 – Phase your data

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



Phasing in Eagle2 or Shapeit2

Hidden Markov Models are used to reconstruct haplotypes in the genotyped data

These are used to provide scaffolds for the imputation

Reference-based phasing using the Haplotype Reference Consortium panel

Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, Richard Durbin & Alkes L Price

Nature Genetics **48**, 1443–1448 (2016) | Cite this article

Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel

Olivier Delaneau, Jonathan Marchini & The 1000 Genomes Project Consortium

Nature Communications **5**, Article number: 3934 (2014) | Cite this article

Step 3: Impute

Minimac4

Impute5

Positional Burrows Wheeler Transform (PBWT)

Beagle

never use plink for imputation!

Minimac4


https://github.com/statgen/Minimac4

Building on the work from Gonçalo Abecasis, Christian Fuchsberger and colleagues

Analysis options

- SAIGE
- BoltLMM
- plink2

Next-generation genotype imputation service and methods

Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, David Schlessinger, Dwight Stambolian, Po-Ru Loh, William G Iacono, Anand Swaroop, Laura J Scott, Francesco Cucca, Florian Kronenberg, Michael Boehnke, Gonçalo R Abecasis & Christian Fuchsberger

Nature Genetics **48**, 1284–1287 (2016) | Cite this article

5242 Accesses | **724** Citations | **80** Altmetric | Metrics

Impute5

https://jmarchini.org/software/#impute-5

Built by Jonathan Marchini and colleges

Incorporating Positional Burrows Wheeler Transform (PBWT)

Downstream analysis options

- BGENIE
- SNPtest
- Quicktest

PLOS GENETICS

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Genotype imputation using the Positional Burrows Wheeler Transform

Simone Rubinacci, Olivier Delaneau, Jonathan Marchini

Version 2 Published: November 16, 2020 • https://doi.org/10.1371/journal.pgen.1009049

Issue – in house vs online imputation

Legal restrictions

- eg General Data Protection Regulation (GDRP)

IRB/Ethics restrictions

- eg UK Biobank

Data sovereignty restrictions

- Study specific

In house Options for imputation

Use a cookbook!

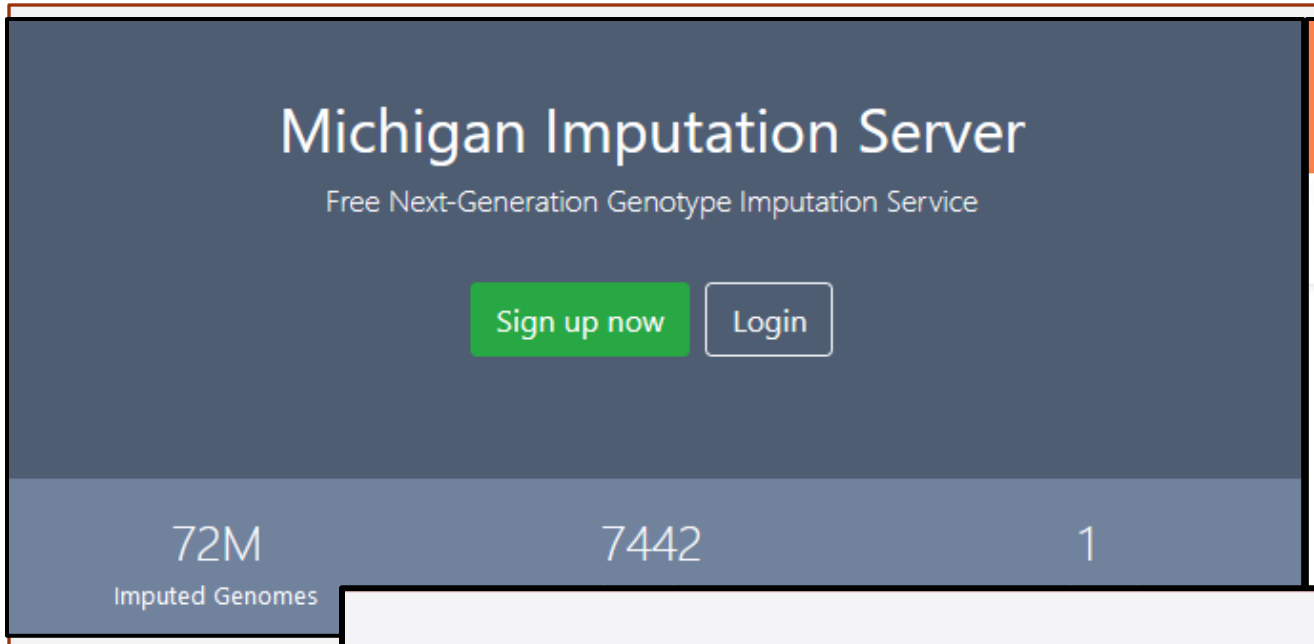
http://genome.sph.umich.edu/wiki/Minimac3_Imputation_Cookbook OR http://genome.sph.umich.edu/wiki/IMPUTE2:_1000_Genomes_Imputation_Cookbook

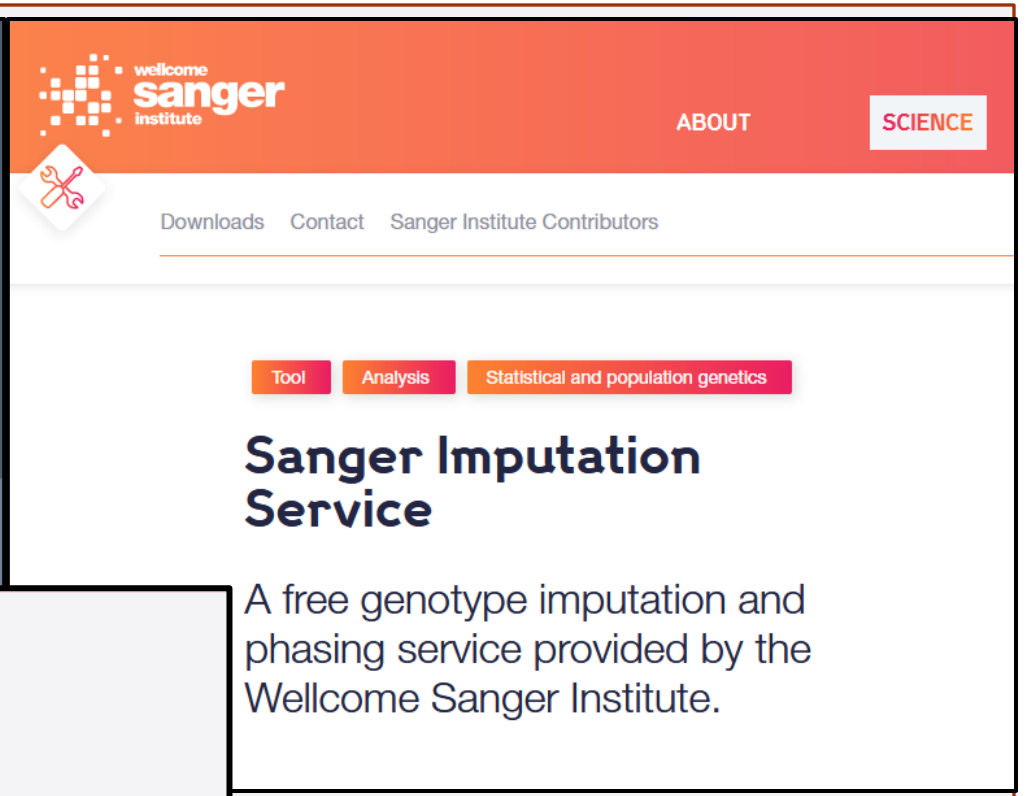
Use a container

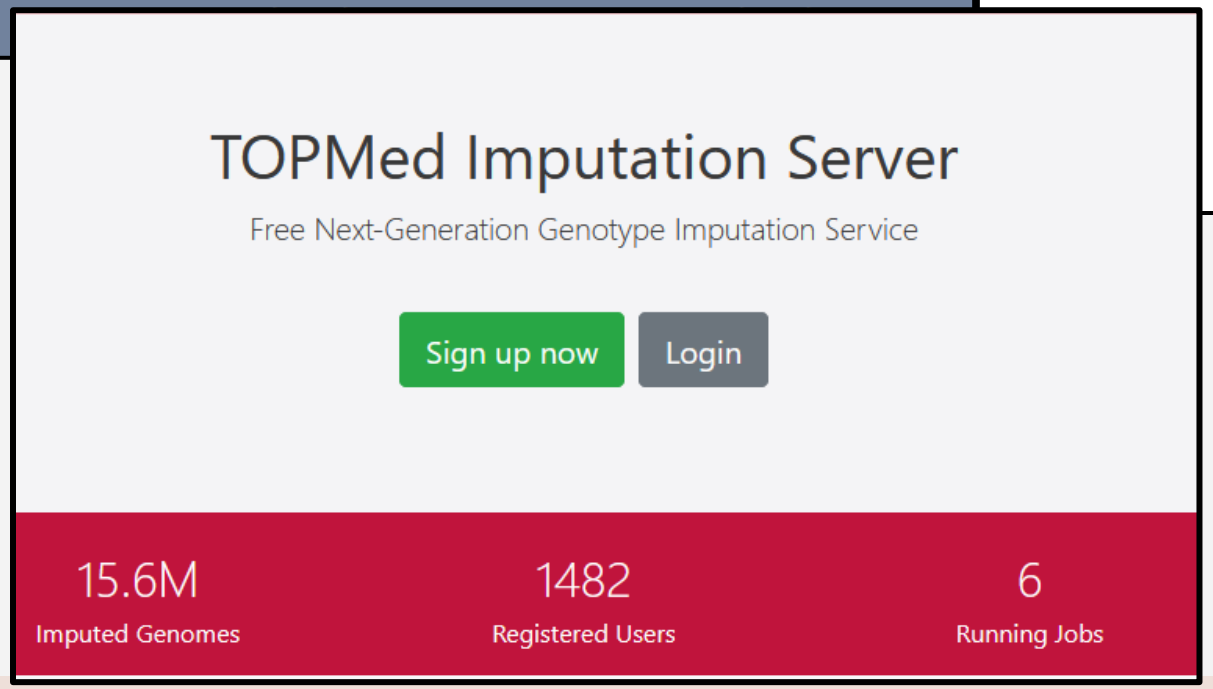
- eg https://github.com/HippocampusGirl/ImputationProtocol
- Use with singularity or docker
- Enables you to run the imputation on a stand alone computer or local server
- (You can think of this as a container that holds a pre-built environment with everything already set up for your task)

Online options for imputation

- UMich Imputation Server
  - https://imputationserver.sph.umich.edu/
- Sanger Imputation Server
  - https://imputation.sanger.ac.uk/
- TOPMed Imputation Server
  - https://imputation.biodatacatalyst.nhlbi.nih.gov/

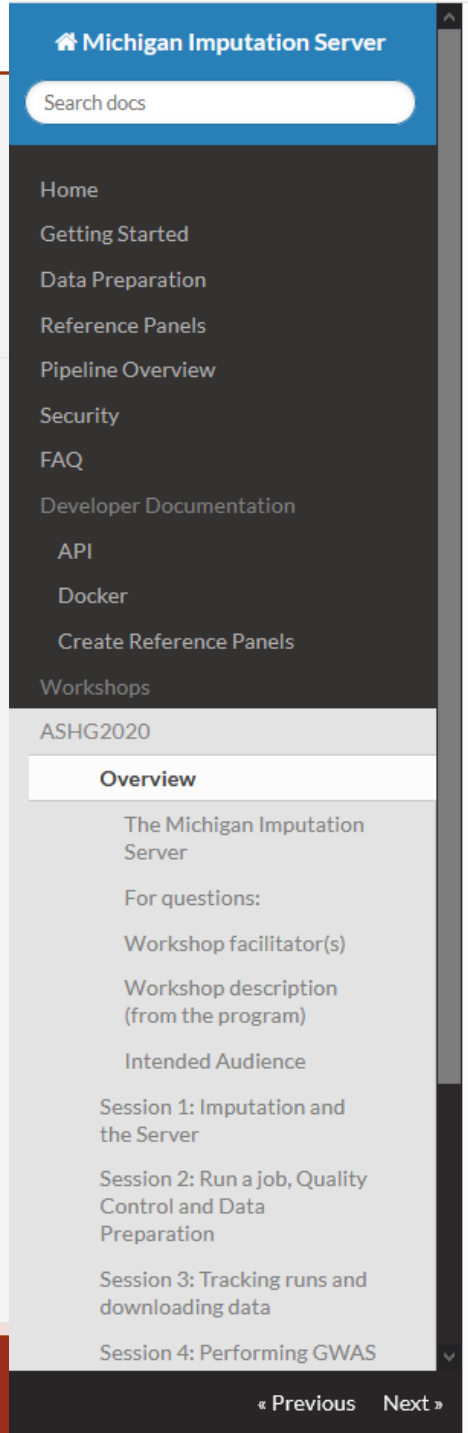

Michigan Imputation Server
Free Next-Generation Genotype Imputation Service
Sign up now
Login
72M
Imputed Genomes
7442
1


wellcome sanger institute
ABOUT
SCIENCE
Downloads
Contact
Sanger Institute Contributors
Tool
Analysis
Statistical and population genetics
Sanger Imputation Service
A free genotype imputation and phasing service provided by the Wellcome Sanger Institute.


TOPMed Imputation Server
Free Next-Generation Genotype Imputation Service
Sign up now
Login
15.6M
Imputed Genomes
1482
Registered Users
6
Running Jobs

On the Michigan Imputation Server Site - Great practical workshop sessions from ASHG 2020

https://imputationserver.readthedocs.io/en/latest/workshops/ASHG2020/



Docs » Workshops » ASHG2020 » Overview

Edit on genepi/imputationserver

Workshop ASHG2020

The Michigan Imputation Server

Data Preparation, Genotype Imputation, and Data Analysis

For questions:

- Please email us: mis-ashg2020@umich.edu
- Slack channel: Slack sign-up

Workshop facilitator(s)

- Christian Fuchsberger, christian.fuchsberger@eurac.edu (Eurac Research)
- Lukas Forer, lukas.forer@i-med.ac.at (Medical University of Innsbruck)
- Sarah Hanks, schanks@umich.edu (University of Michigan)
- Sebastian Schoenherr, sebastian.schoenherr@i-med.ac.at (Medical University of Innsbruck)
- Albert Smith, albertvs@umich.edu (University of Michigan)
- Cassie Spracklen, cspracklen@umass.edu (University of Massachusetts-Amherst)

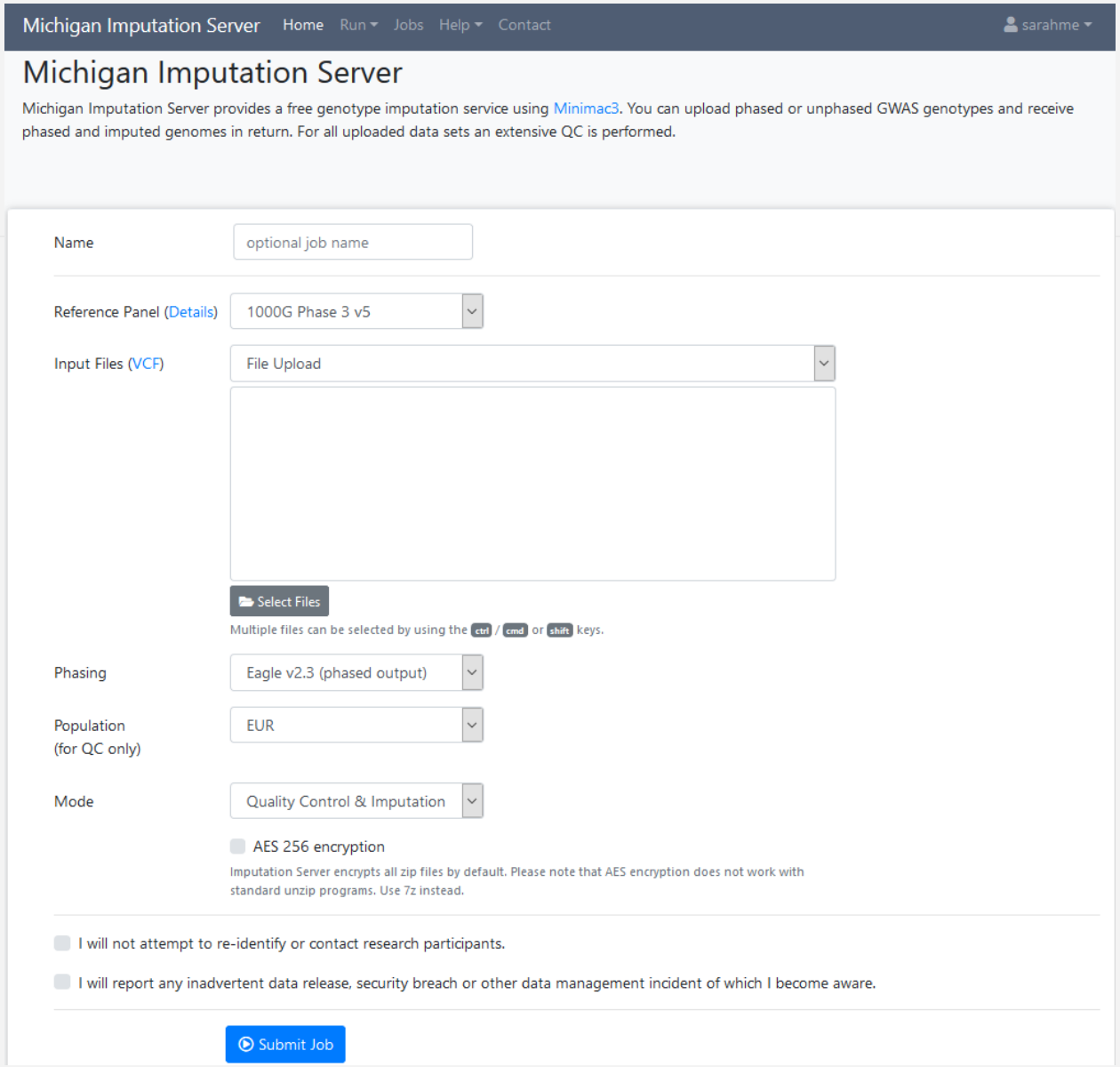
Workshop description (from the program)

Genotype imputation is a key component of modern genetic studies. This interactive workshop is intended for anyone interested in learning how to use the Michigan Imputation Server (MIS; https://imputationserver.sph.umich.edu) to impute genotypes and how to use the imputed genotypes, with a special focus on up-coming reference panels, including the multi-ancestry panel from the TOPMed program. A brief overview of imputation and the server will be followed by demonstrations and exercises, including:

1. quality control and preparation of genetic data for use on the MIS with a special focus on diverse ancestries and chromosome X
2. tracking runs and use of the application program interface for larger jobs
3. downloading data from the MIS and preparing data for genetic analysis
4. performing a GWAS using imputed data and interpreting results, taking into account imputation

Preparing 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 vs b38)
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


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

Output

3 main genotype output formats

- Hard call or best guess
- Dosage data (most common – 1 number per SNP, 1-2)
- Probs format (probability of AA AB and BB genotypes for each SNP)

```
##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
```

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.).

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

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

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

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

For this likelihood the score and information are given by

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

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

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.

$$I_A = \frac{\mathbb{E}_{G_{\cdot j}}[I(\hat{\theta})] - V_G[U(\hat{\theta})]}{\mathbb{E}_{G_{\cdot j}}[I(\hat{\theta})]} \quad (13)$$

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

$$\mathbb{E}_{G_{\cdot j}}[I(\hat{\theta})] = \frac{2N}{\hat{\theta}(1-\hat{\theta})} \quad (14)$$

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

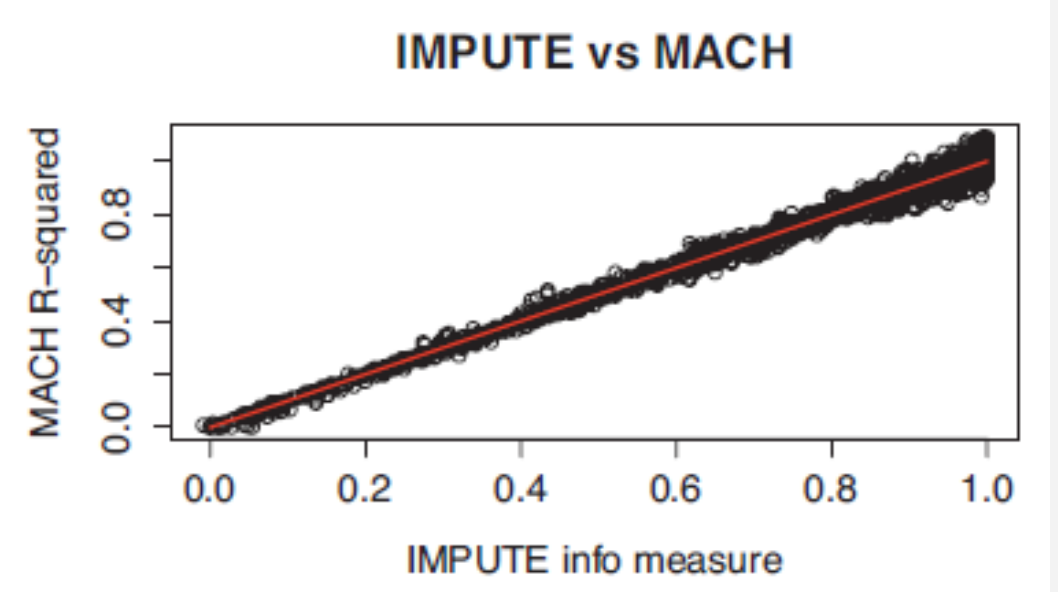
so that

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

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

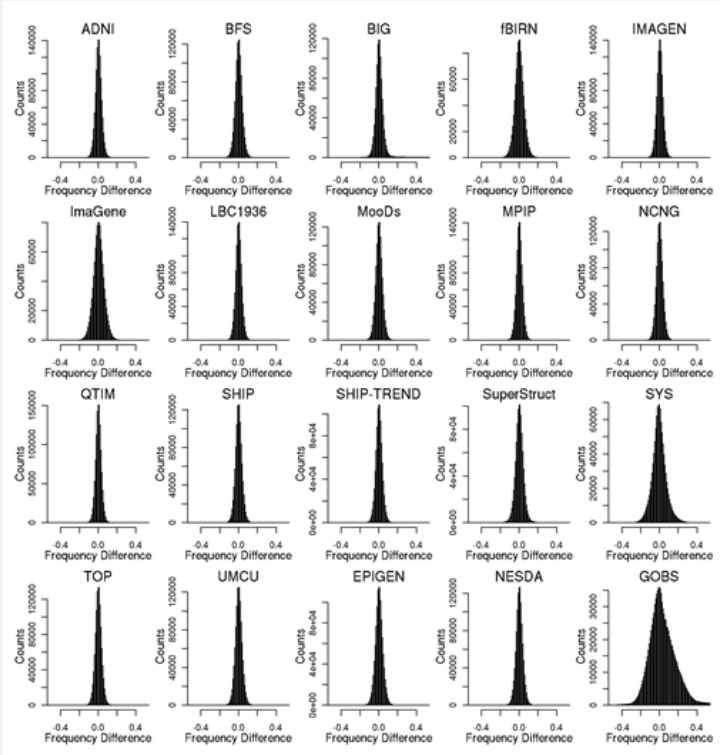


Post imputation QC

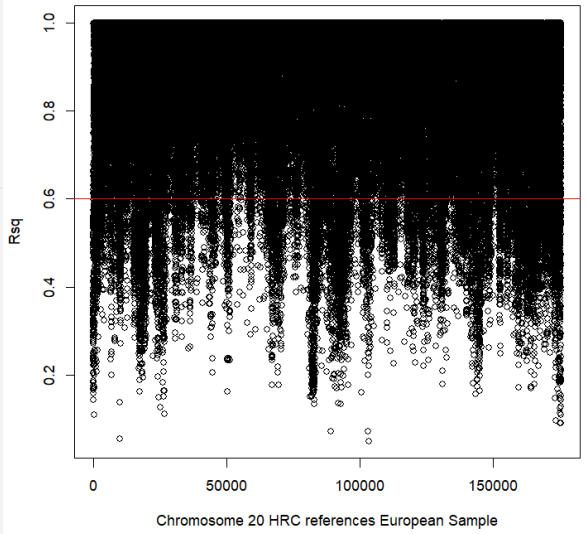
After imputation you need to check that it worked and the data look ok

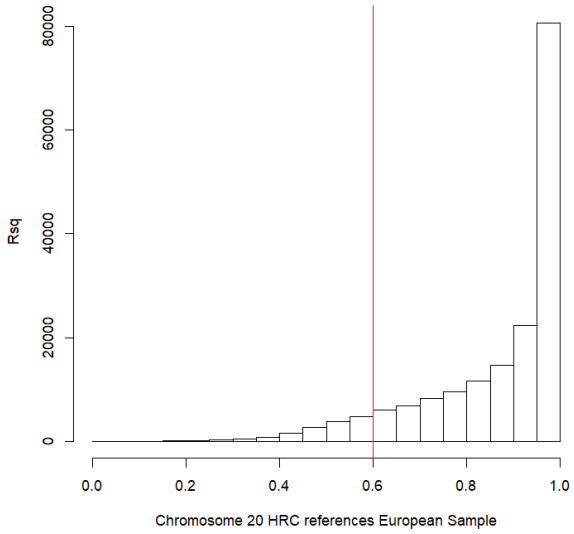
Things to check

- Plot r^2 across each chromosome look to see where it drops off
- Plot MAF-reference MAF

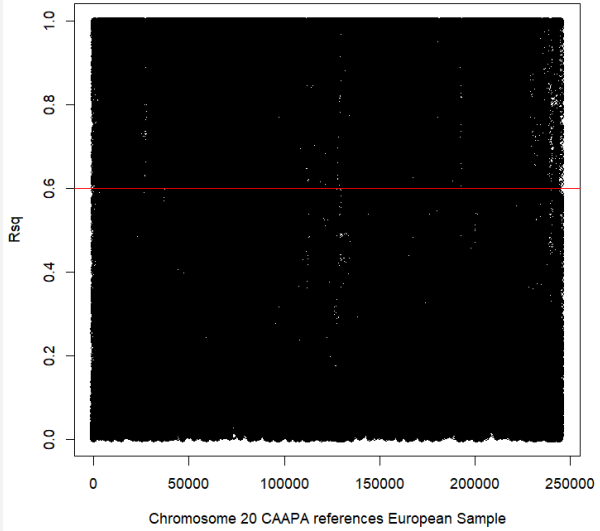


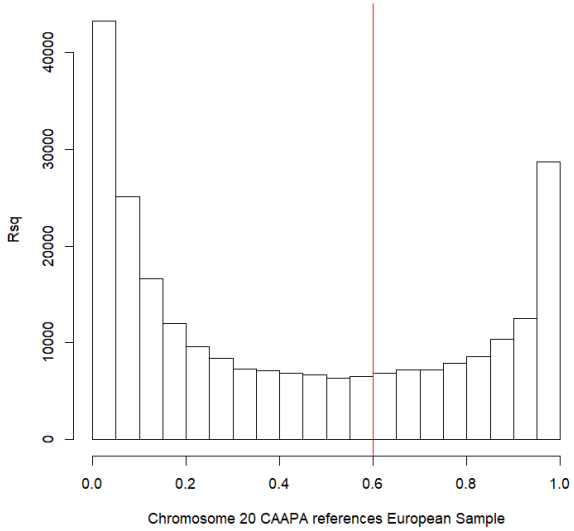
Good imputation





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

Last points

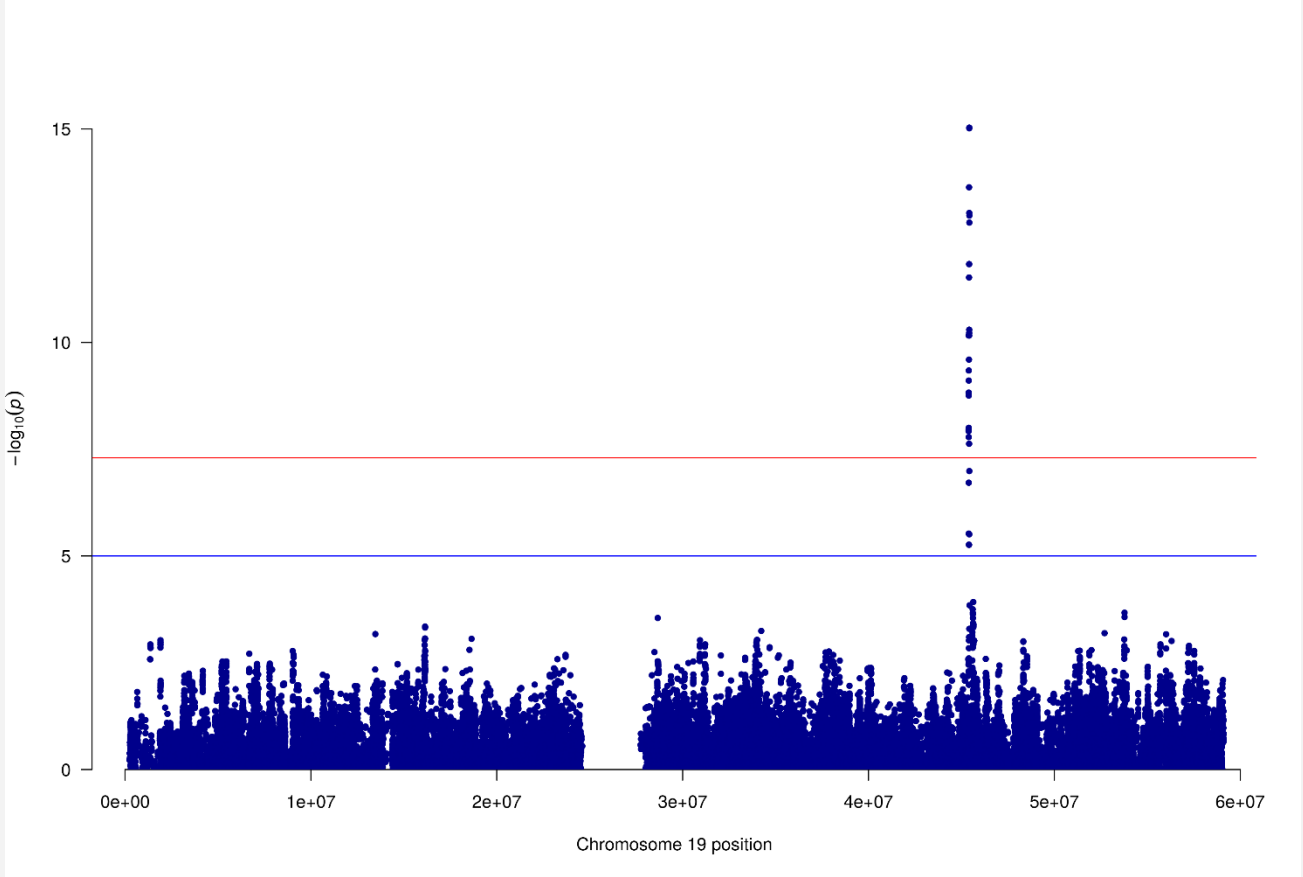
If you are running analyses for a consortium they will probably ask you to analyse all variants regardless of whether they pass QC or not…

(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)

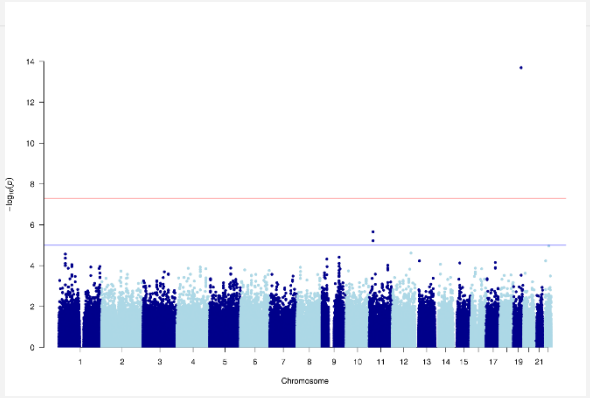
Analysis of Imputed data using SAIGE

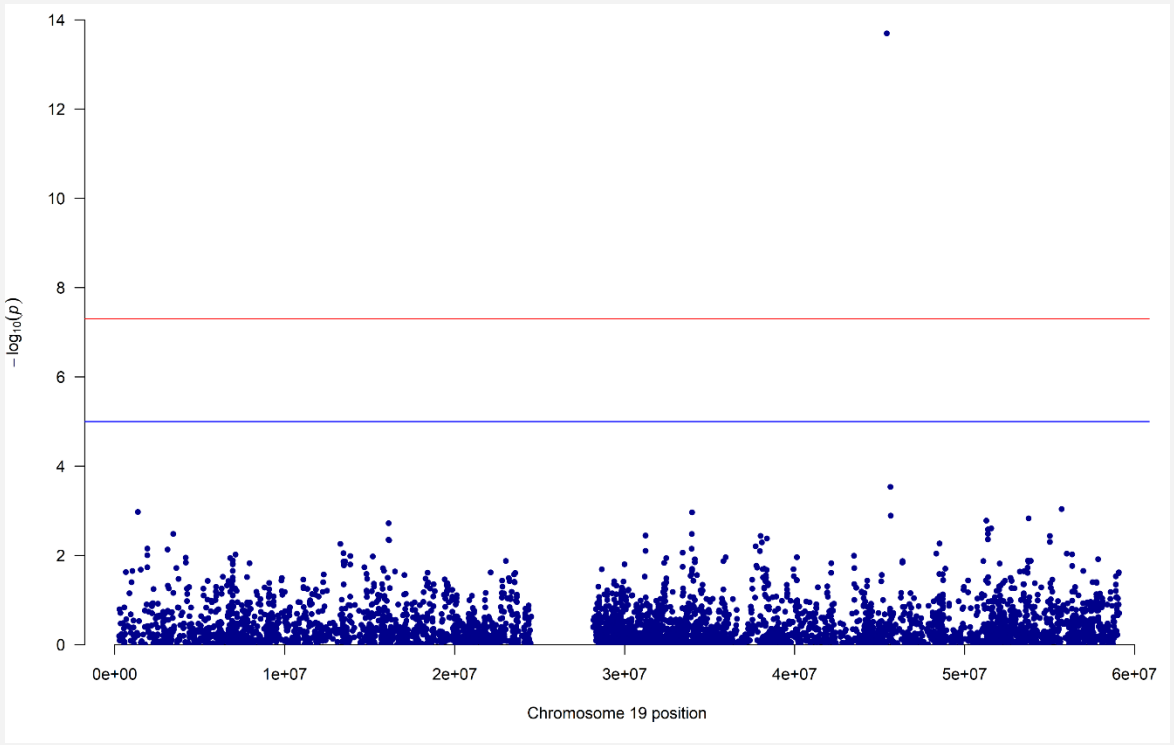
Back to qualtrics…

Imputed results (Manhattan plot only chr19)

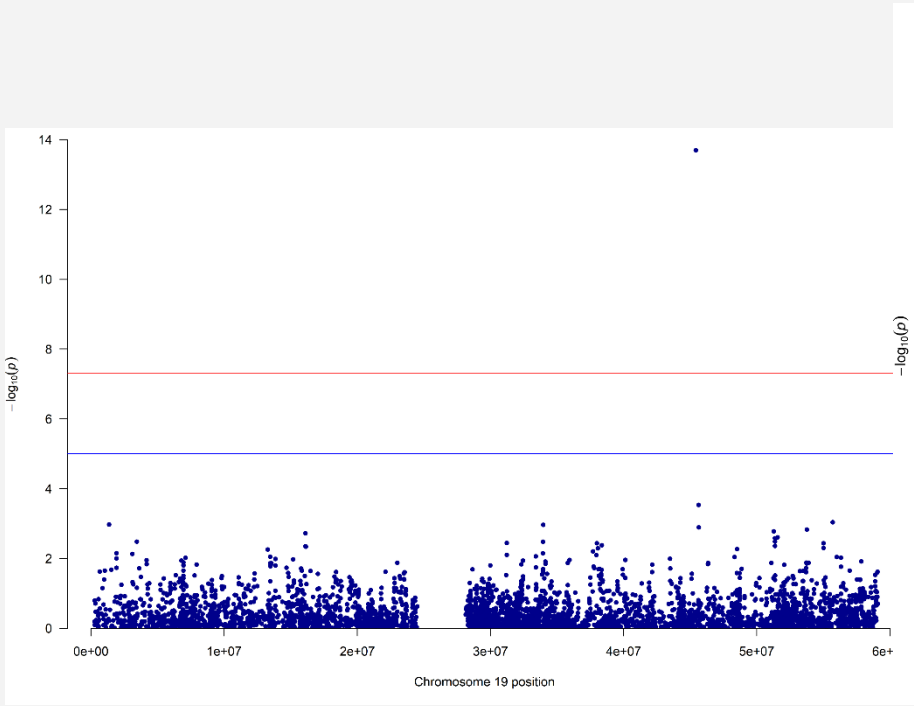


Compared to the genotyped snp results!

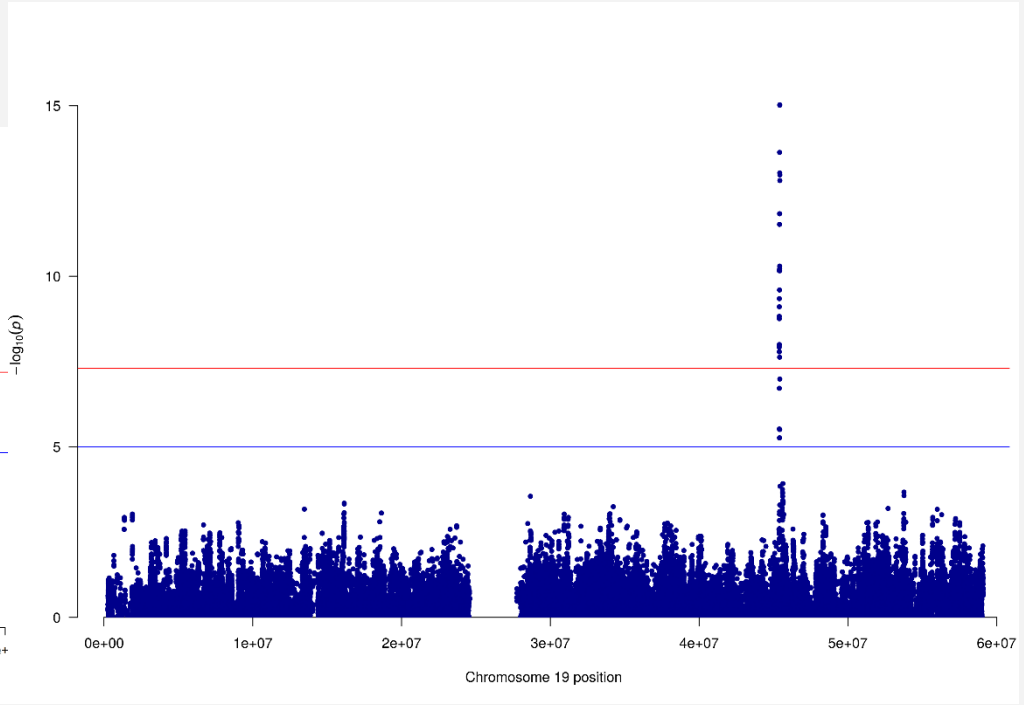




And Compare…

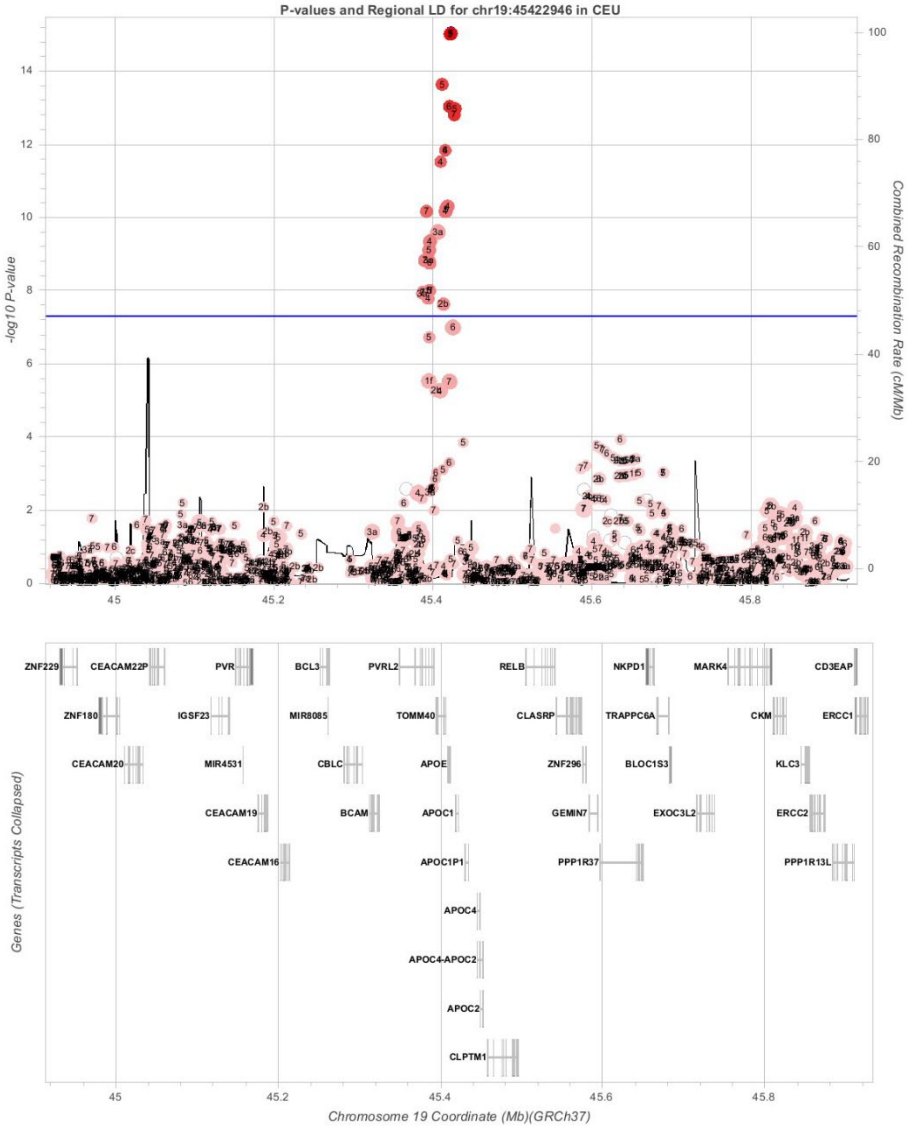


Genotyped

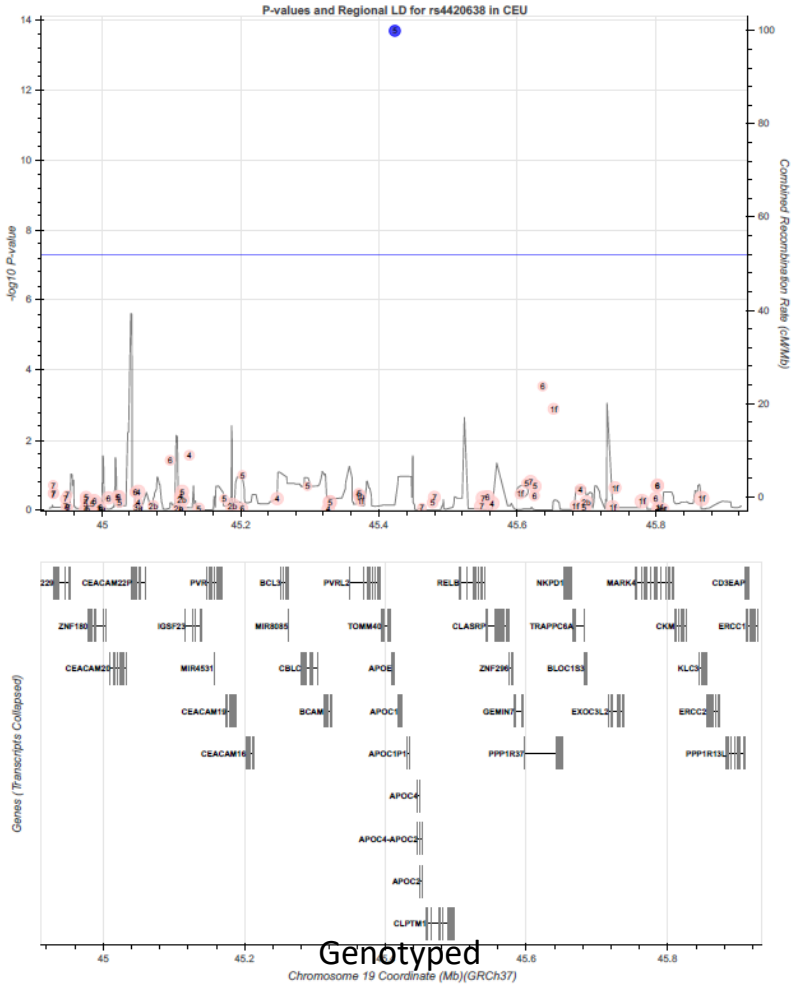


Imputed

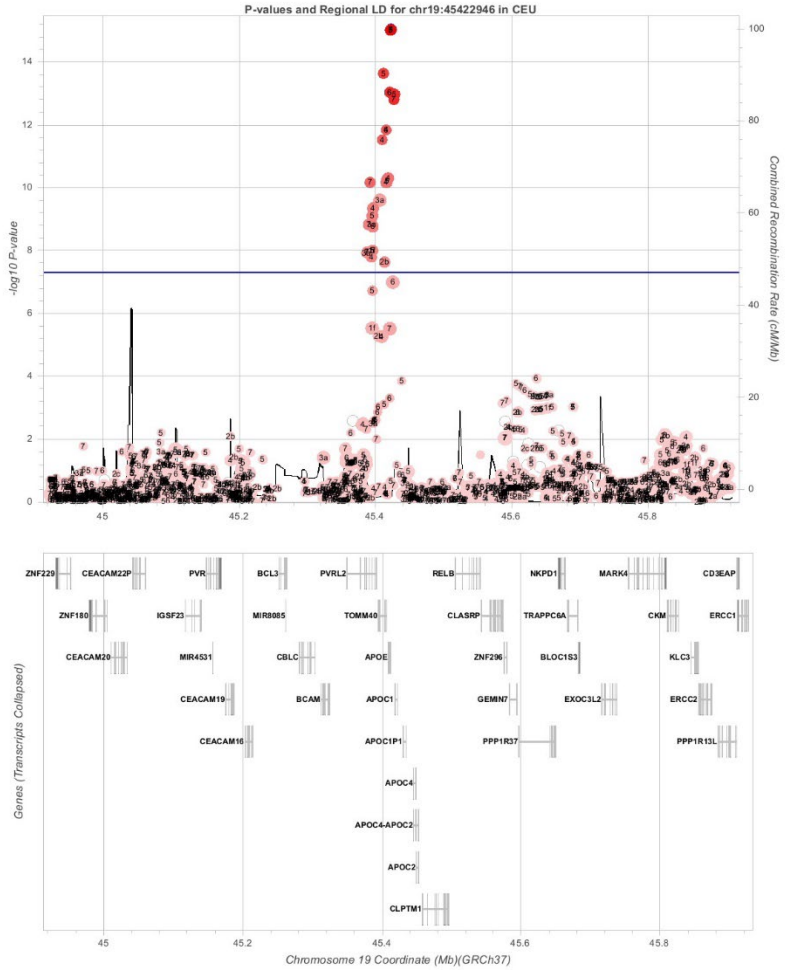
Imputed results (regional plot)



And Compare…



Genotyped



Questions