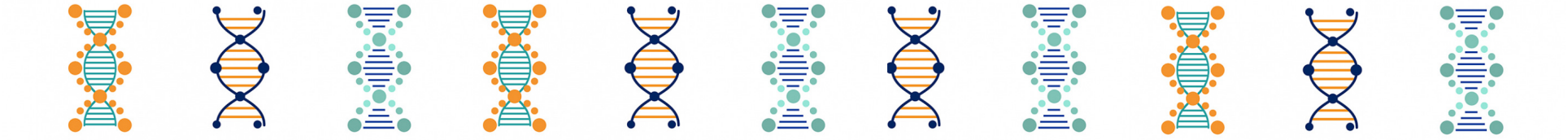
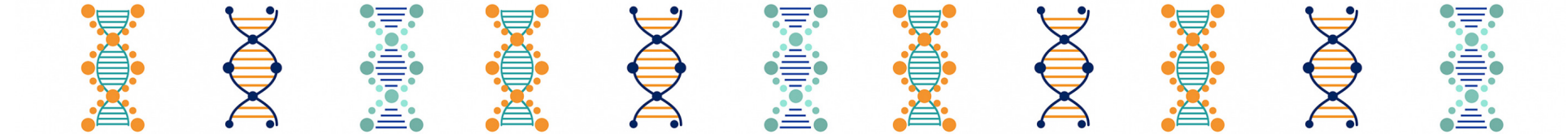


Genotype Imputation Tutorial

Anna Díez-Villanueva
17-05-2021



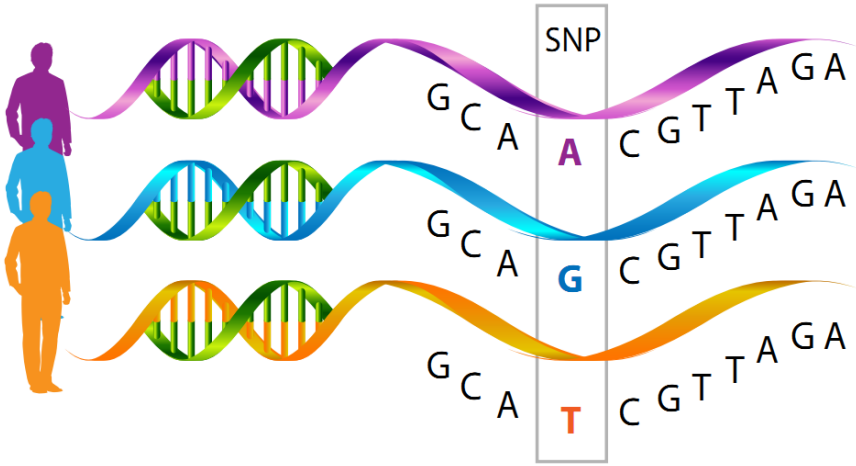
- Basic concepts
- Reference panels
- Genotype imputation background
- Quality control filters before imputation
- Data preparation before imputation
- Imputation
    a) Manually
    b) Server
        - Michigan Imputation Server
        - TOPMed Imputation Server
    c) Bot

Coffee break

- Practical

Introduction to some concepts

Basic concepts

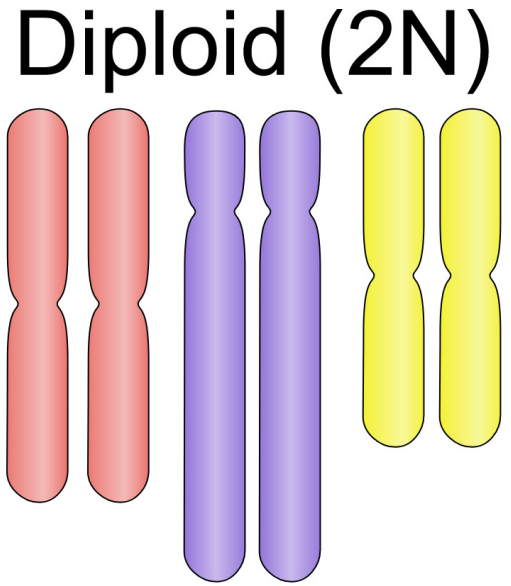


- Human genome → 3×10^9 base pairs

- **SNP**:
  - ~ 40 millions
  - Positions in the genome where some individuals have one nucleotide (e.g. A) and others have a different one (e.g. a G).
  - Vast majority bi-allelic (0 major allele / 1 minor allele)

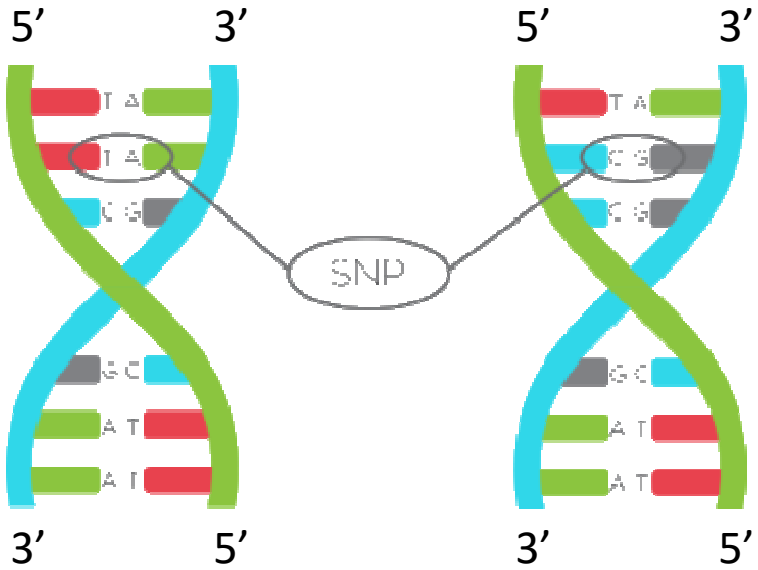
- **Minor allele frequency** (MAF)
  - Common variants → SNPs with MAF > 0.05
  - Rare variants → SNPs with MAF < 0.05

Basic concepts



- Human cells are diploid
  - two copies of each chromosome
  - one inherited from mother and one from father

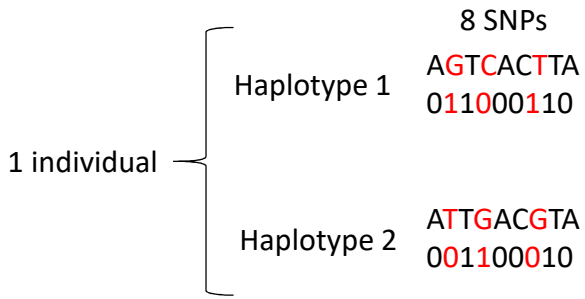
- **Strand**



Genotype strand 5’ → T C
Genotype strand 3’ → A G

Basic concepts

- **Haplotype** :
  - Description of SNP alleles on a chromosome region
  - 0/1 encoding:
    - 0 - for reference (major) allele
    - 1 - for alternative (minor) allele



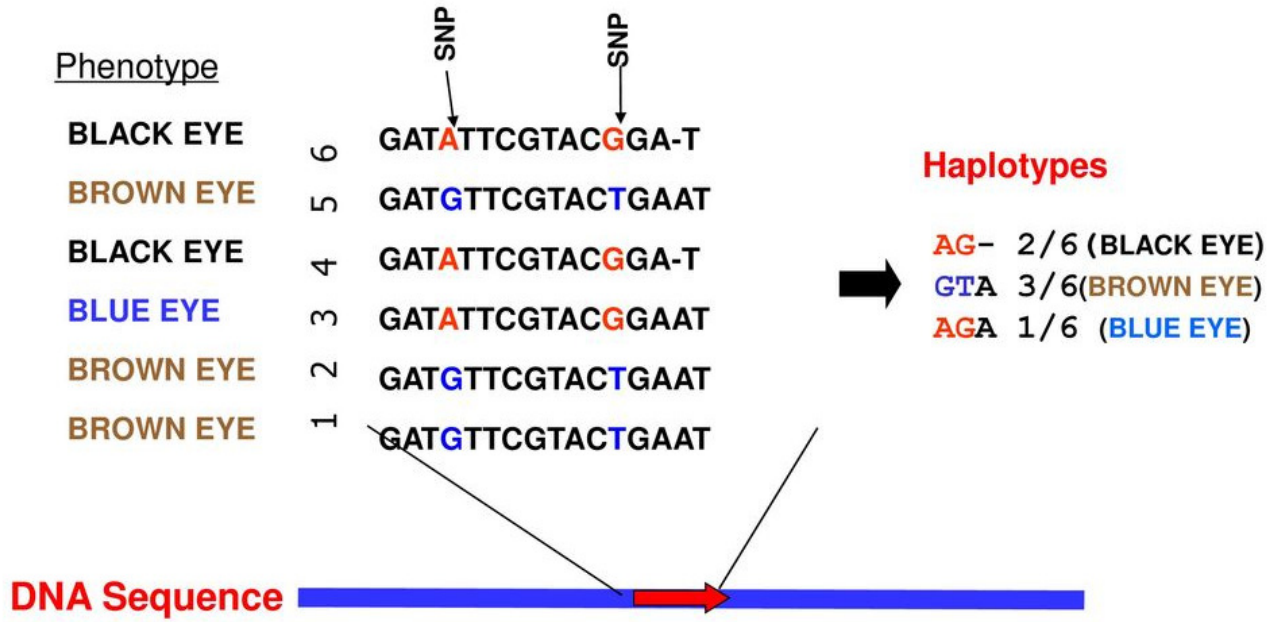
- **Genotype** :
  - Description of SNP alleles on both chromosome copies
  - 0/1/2 encoding:
    - 0 - both chromosome strands contain the major allele
    - 1 - the chromosomes contain different alleles (heterozygous)
    - 2 - both chromosome strands contain the minor allele

AA GT TT CG AA CC TG TT AA
0 1 2 1 0 0 1 2 0

Genotype for 8 SNPs and 1 individual

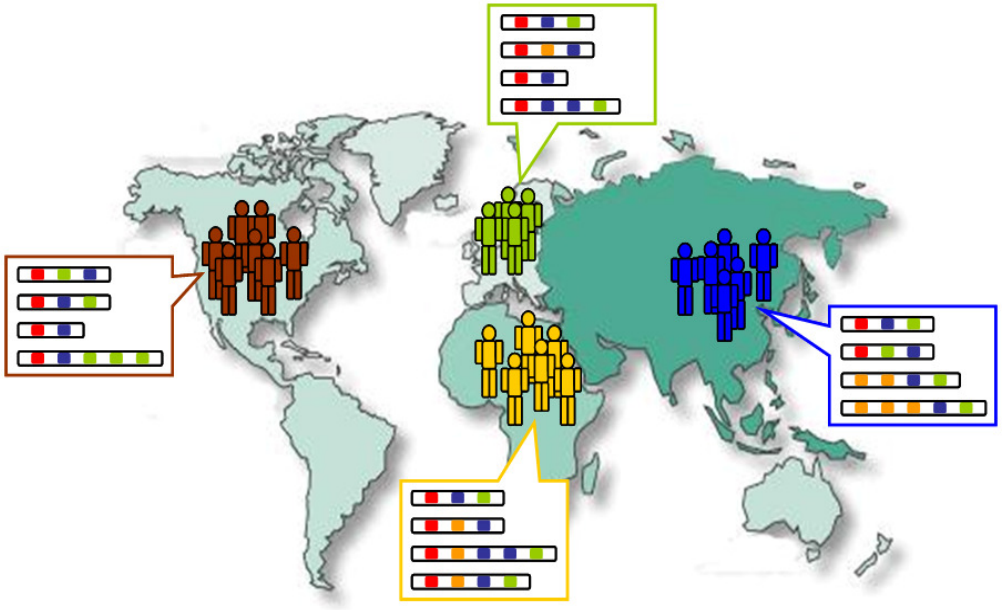
Basic concepts

- **Phenotype** → Characteristics or traits of an organism.



Reference Panels

Reference Panels



- 1000 genomes

  https://www.internationalgenome.org/

- Haplotype Reference Consortium (HRC)

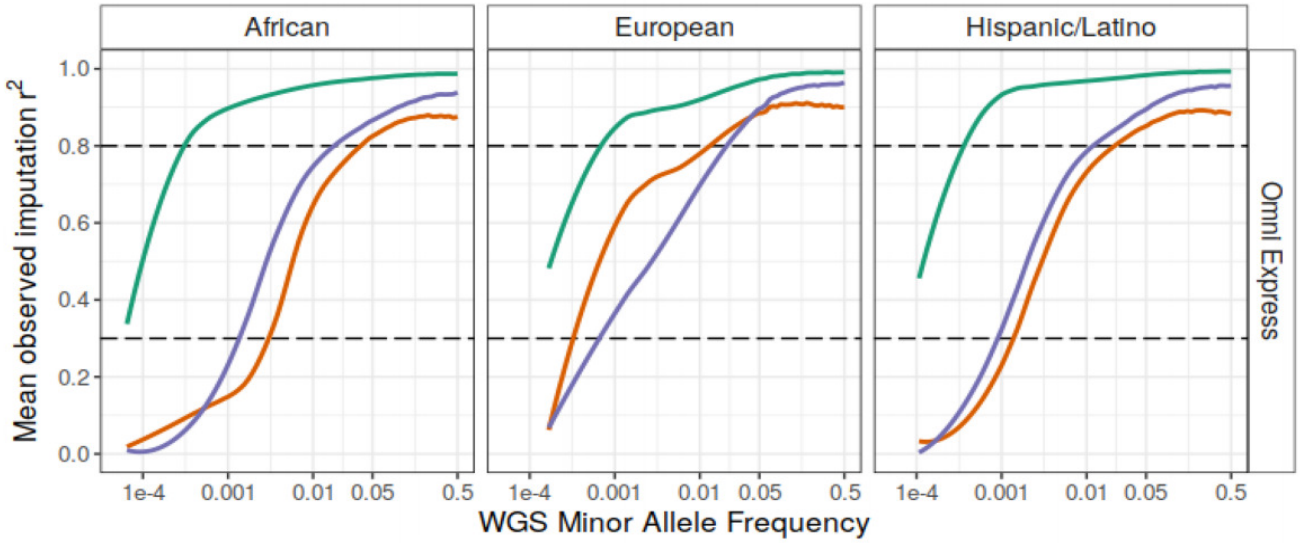
  http://www.haplotype-reference-consortium.org/

- Trans-Omics for Precision Medicine (TOPMed)

  https://www.nhlbiwgs.org/

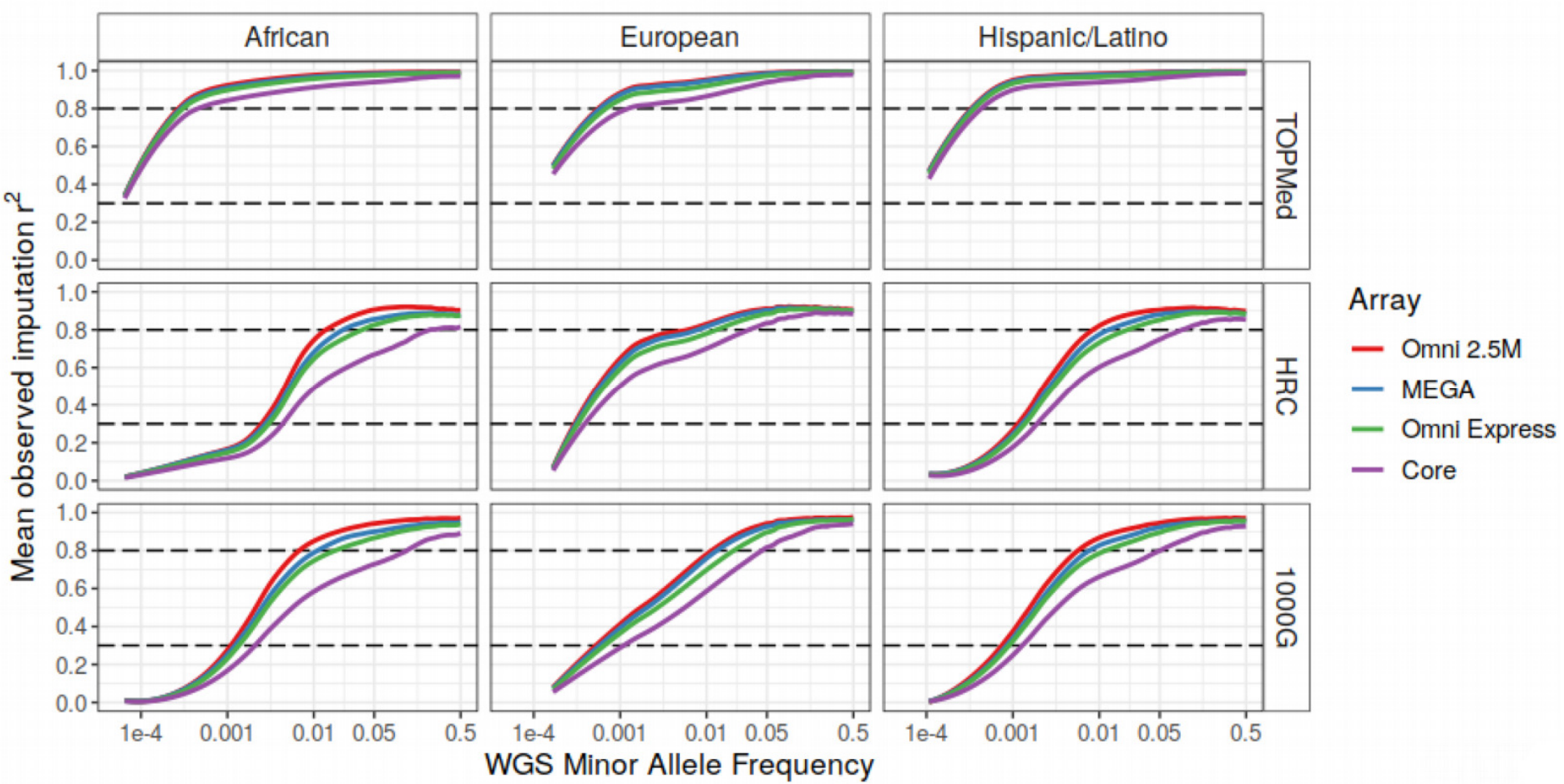
Reference Panels Comparison

	1000 Genomes	HRC 2016	TOPMed
Number of samples	2.504	32.470	97.256
Number of sites	~49M	~39M	~308M
Chromosomes	chr1 - chr22, chrX	chr1 - chr22, chrX	chr1 - chr22, chrX
Populations	Multiethnic	Mostly European	Multiethnic
Variants	SNPs + indels	SNPs	SNPs + indels
Genome build	hg19	hg19	hg38



https://imputationserver.readthedocs.io/en/latest/workshops/ASHG2020/Session6/

Reference Panel Comparison



https://imputationserver.readthedocs.io/en/latest/workshops/ASHG2020/Session6/

Reference Panel Comparison

TOPMed Reference Panel:

- Biggest number of samples and variants.
- Gives highest imputation quality in all populations.
- Rare variants, with MAF < 0.05, well imputed ($r^2 \geq 0.8$)
- Small effect of genotyping array.

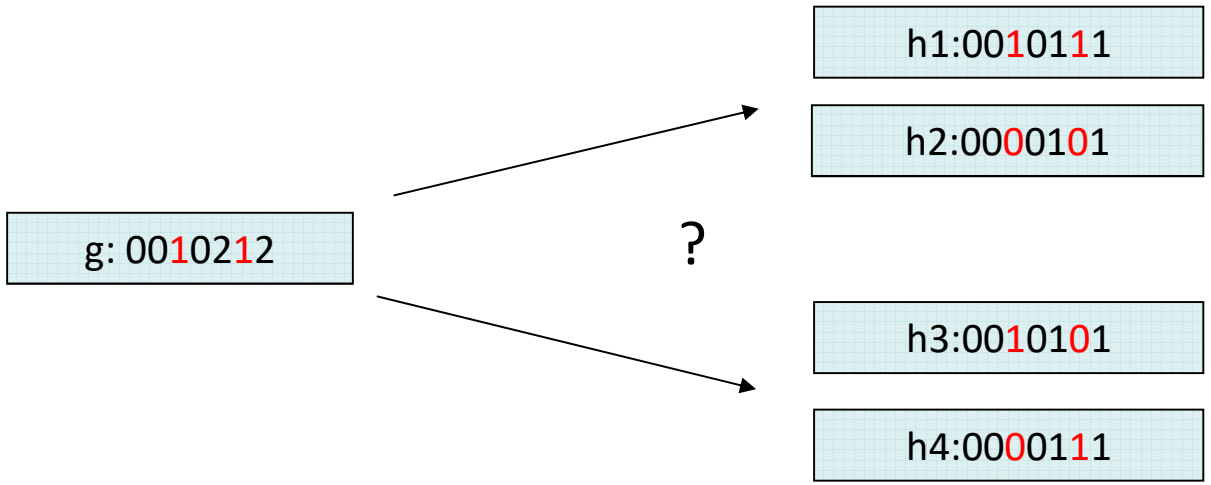
→ Asian population needs an specific reference panel.

Genotype imputation background

Phasing

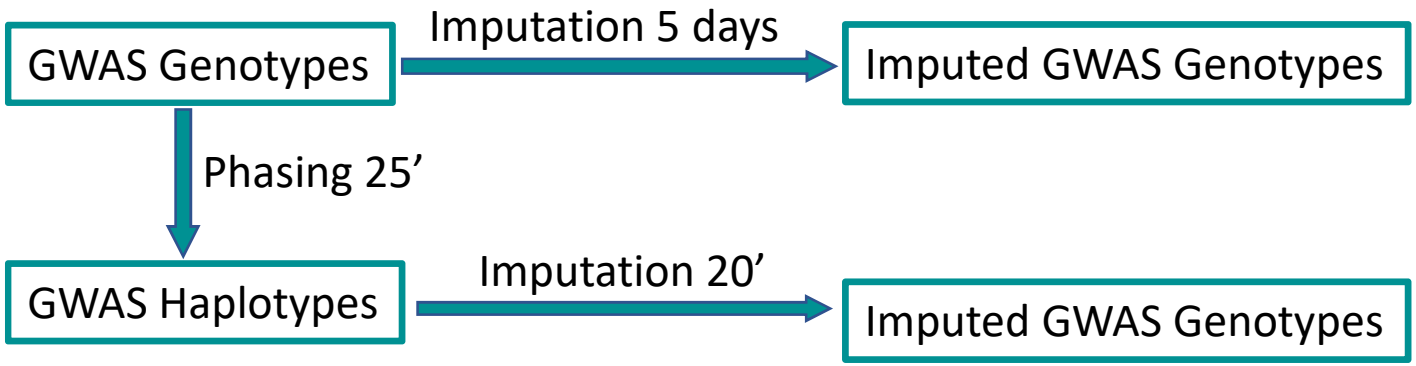
- In microarrays, each SNP is genotyped independently.

- For a genotype with k 1's there are 2^{k-1} possible pairs of haplotypes explaining it.



Phasing

- Why? It helps to make the process of imputation faster.



https://www.genome.gov/sites/default/files/genome-old/pages/Research/DER/ICHG-1000GenomesTutorial/Imputation_in_GWAS_Studies.pdf

- Computational approaches to genotype phasing

  - Statistical methods: PHASE, Phamily, PL, GERBIL, SHAPEIT, Eagle, …

  - Combinatorial methods: Parsimony, HAP, 2SNP, ENT, …

Phasing: example with eagle

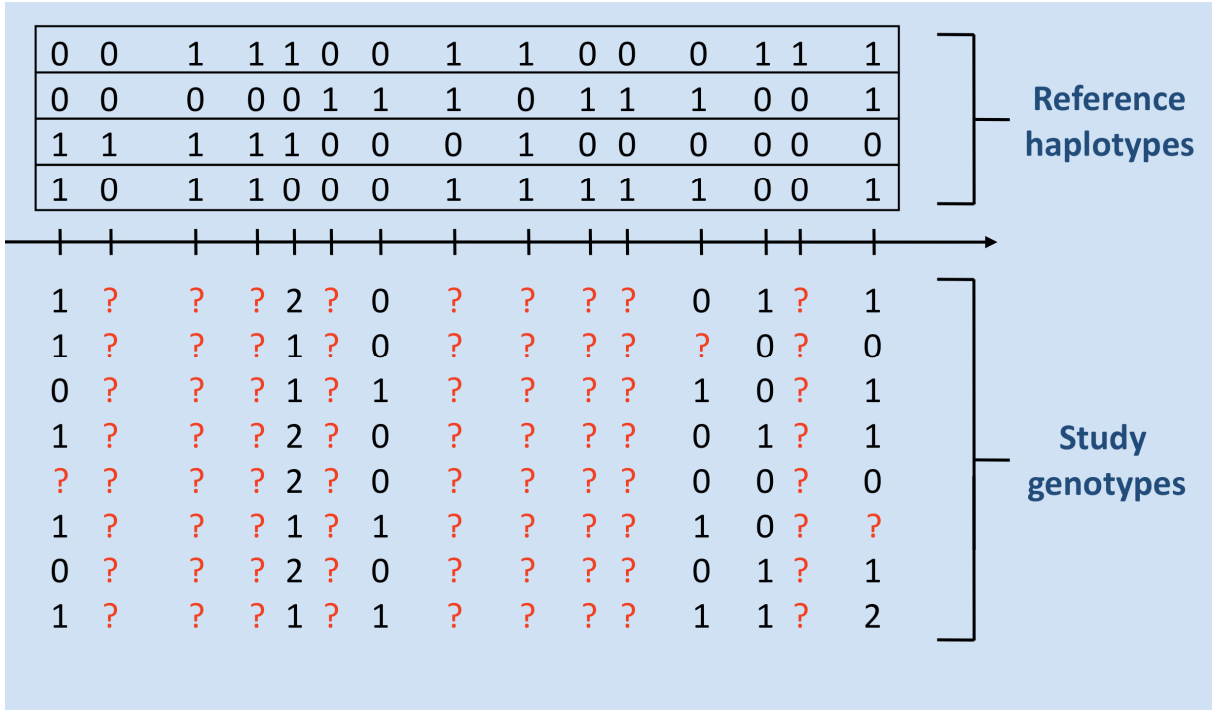
Data is usually broken into manageable chunks of ~20Mb each phased independently:

```
./eagle
--vcfRef HRC.r1-1.GRCh37.chr20.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 20000000
--chrom 20
```

Genotype imputation

Statistical inference of unobserved genotypes to:

- Increase the number of tested variants.
- Perform meta-analysis using different arrays.



Genotype imputation vs Whole Genome Sequencing

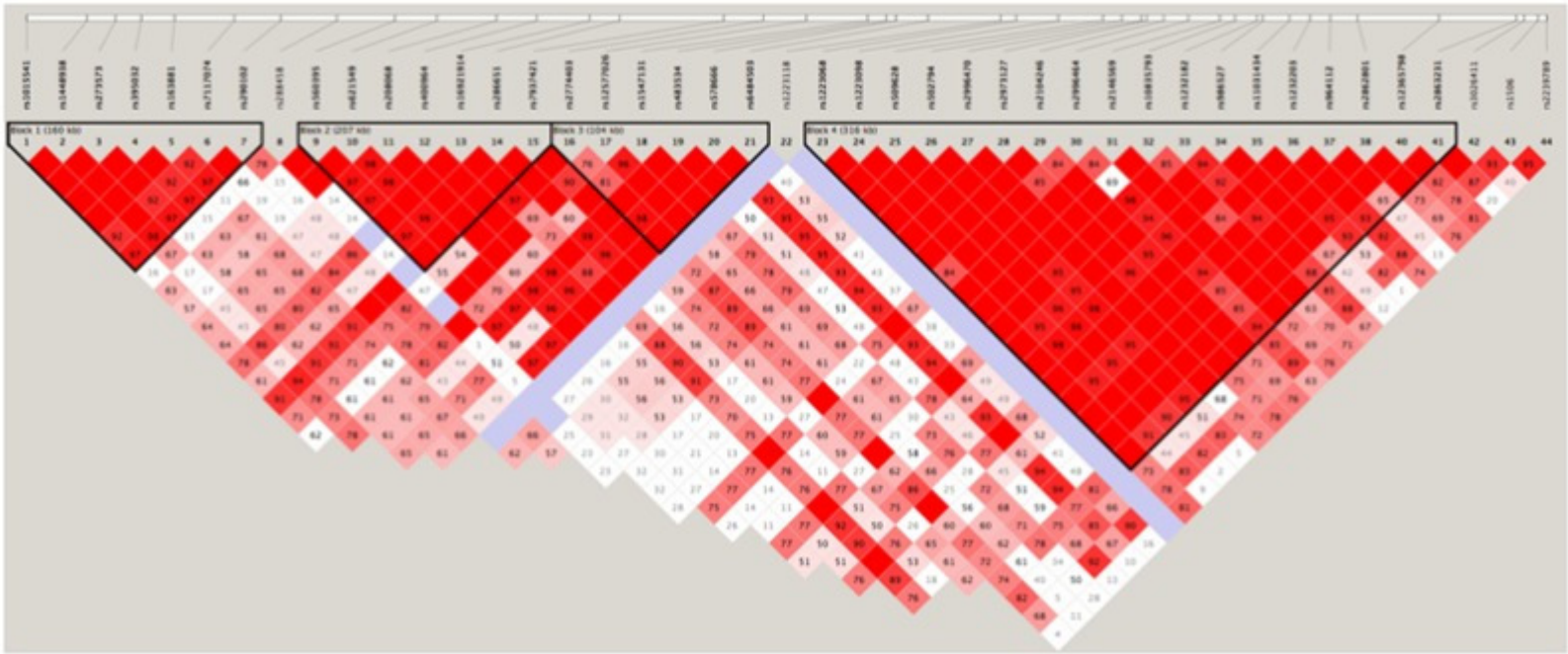
- Whole Genome Sequencing (WGS) is expensive, $600 to $1000 per genome.

- Genotyping:
  - is cheaper $50 to $200 per genome.
  - is limited to predetermined set of mostly common (MAF>0.05) variants.

- Genotyping + Imputation is good enough!!
  - The genotyped SNPs are linked (correlated) to ungenotyped SNPs.

Linkage disequilibrium

- Non-random association of alleles at different SNPs in a given population.
- Mutations that remain on the same haplotype throughout the generations are in LD.
- LD is lost with time and distance between SNPs.



Imputation: basic idea

Sequenced and phased reference haplotypes

Haplotype	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
1	1	0	1	1	0	0	1
2	0	1	1	1	0	0	1
3	0	0	1	1	1	0	0
4	0	1	1	0	0	0	1
5	1	0	1	0	0	1	0
6	1	1	0	1	1	0	0

Genotyped (microarray)

Haplotype	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
1	0	?	1	0	?	?	1
2	1	?	1	0	?	?	0

0: ref allele
1: alt allele

Imputation: basic idea

Sequenced and phased reference haplotypes

Haplotype	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
1	1		1	1			1
2	0		1	1			1
3	0		1	1			0
4	0		1	0			1
5	1		1	0			0
6	1		0	1			0

Genotyped

Haplotype	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
1	0	?	1	0	?	?	1
2	1	?	1	0	?	?	0

Imputation: basic idea

Sequenced and phased reference haplotypes

Haplotype	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
1	1	0	1	1	0	0	1
2	0	1	1	1	0	0	1
3	0	0	1	1	1	0	0
4	0	**1**	1	0	**0**	**0**	1
5	1	0	1	0	0	1	0
6	1	1	0	1	1	0	0

Genotyped

Haplotype	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
1	0	**1**	1	0	**0**	**0**	1
2	1	?	1	0	?	?	0

Imputation: basic idea

Sequenced and phased reference haplotypes

Haplotype	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
1	1		1	1			1
2	0		1	1			1
3	0		1	1			0
4	0		1	0			1
5	1		1	0			0
6	1		0	1			0

Genotyped

Haplotype	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
1	0	1	1	0	0	0	1
2	1	?	1	0	?	?	0

Imputation: basic idea

Sequenced and phased reference haplotypes

Haplotype	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
1	1	0	1	1	0	0	1
2	0	1	1	1	0	0	1
3	0	0	1	1	1	0	0
4	0	1	1	0	0	0	1
5	1	**0**	1	0	**0**	**1**	0
6	1	1	0	1	1	0	0

Genotyped

Haplotype	SNP 1	SNP 2	SNP 3	SNP 4	SNP 5	SNP 6	SNP 7
1	0	1	2	2	0	0	1
2	1	**0**	1	0	**0**	**1**	0

Quality control

before imputation

Quality Control Filters

- Variant filters
- Sample filters

a) Proportion of missing values : Remove SNPs with more than 5% of missing data.

b) Proportion of missing values: Remove samples with more than 10% of missing data.

c) Sex concordance: Check samples with no sex concordance

plink function --check-sex → compares sex assignments in the input dataset with those imputed from X chromosome inbreeding coefficients

Quality Control Filters

Variant filters

Sample filters

d) Heterozygosity: Remove samples with high or low heterozygosity.

Heterozygosity = having two different alleles of a specific SNP

The heterozygosity rate of an individual is the proportion of heterozygous genotypes.

- High heterozygosity means lots of genetic variability → low quality sample?
- Low heterozygosity means little genetic variability → inbreeding?

plink function **--het** → computes observed and expected autosomal homozygous genotype counts for each sample, and reports method-of-moments F coefficient estimates.

mean(proportion of heterozygous sites) ± 4 * sd(proportion of heterozygous sites)

Quality Control Filters

- Variant filters
- Sample filters

e) Duplicates and relatedness: Check related samples.

plink function --genome → invokes an identical by state (IBS) and identical by descent (IBD) computation and returns PI_HAT summary statistic.

PI_HAT > 0.8

Quality Control Filters

Variant filters

Sample filters

f) Hardy-Weinberg equilibrium

Biallelic marker having allele frequencies *p* and *q=1-p*, is in equilibrium if and only if:

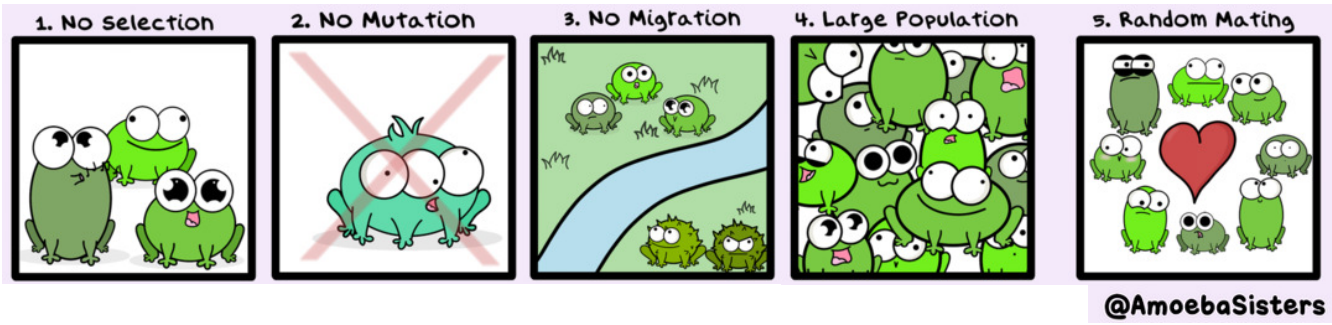
$P(AA) = p^2$

$P(Aa) = 2pq$

$P(aa) = q^2$

χ^2 test

Assumptions →



A disequilibrium can be expected in cases so HWE should only be investigated in controls.

plink function **--hardy** → returns HWE rates **HWE in controls p-value < 1e-04**

Quality Control Filters

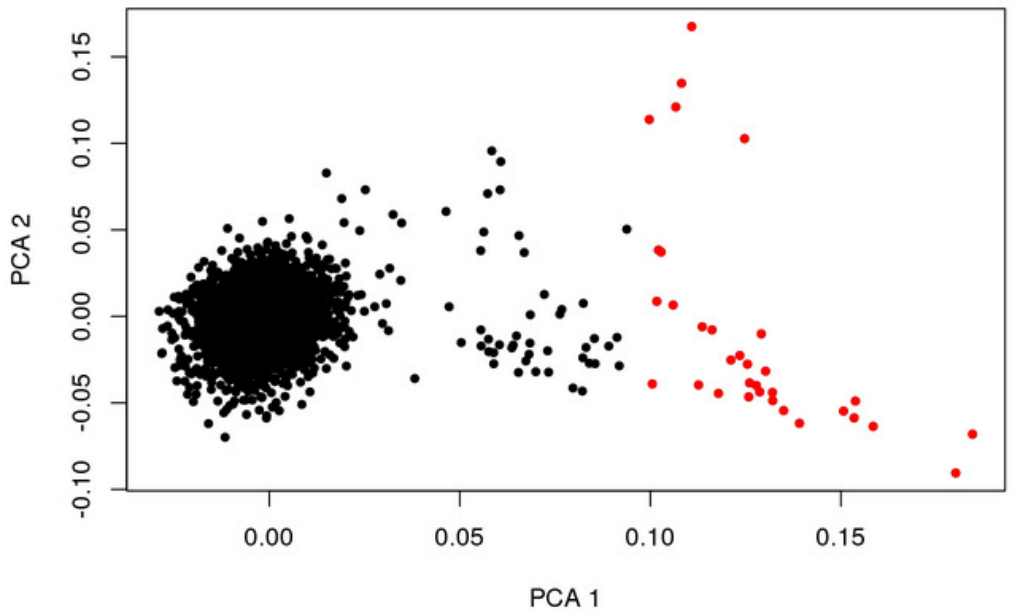
Variant filters

Sample filters

g) Ancestry: Check samples with a different ancestry

Principal Component Analysis (PCA) with 2,300 ancestry-informative marker SNPs.

Each study will have different filtering thresholds!



Data preparation

before imputation

a) Filter in valid SNPs

- Remove SNPs outside autosomes or chrX

- Remove SNPs with invalid alleles

Valid alleles are A, C, T o G

- Remove multiple mapping SNPs

- Remove monomorphic SNPs with MAF < 0.00001

b) Liftover

Transform data from a genome version to another genome version.

Suppose our data is in hg19 and we want to impute using TOPMed panel as reference → **hg19 to hg38**

https://github.com/sritchie73/liftOverPlink

Download the chain file that allows the mass conversion of coordinates from one assembly to another.

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/liftOver/

```
liftOverPlink.py
-m plinkFile.map
-p plinkFile.ped
-o plinkFile_lifted
-c hg19ToHg38.over.chain.gz
```

c) Check bim

```
20  rs6078030    0.1781993   61098   A   G
20  GSA-rs6076506   0.1882266   63231   C   A
20  rs60263736  0.2246549   70980   A   G
20  rs892665    0.244747    75254   A   C
20  chr20-76786 0.251949    76786   A   G
20  rs1935386   0.2532204   87416   A   C
20  GSA-rs75507632  0.2337057   90814   G   A
20  rs13039134  0.2247925   92366   G   A
20  rs6052070   0.1985757   96931   G   A
20  rs6037772   0.1780502   100505  G   A
```

https://www.well.ox.ac.uk/~wrayner/tools/

- Checks:
  Strand, alleles, position, Ref/Alt assignments and frequency differences

- Produces:
  A set of plink commands to update or remove SNPs

- Updates:
  Strand, position, ref/alt assignment

- Removes:
  - A/T & G/C SNPs if MAF > 0.4
  - SNPs with differing alleles
  - SNPs with > 0.2 allele frequency difference
  - SNPs not in reference panel

c) Check bim

https://www.well.ox.ac.uk/~wrayner/tools/

1.- Convert ped/map to bed

```
plink --file plinkFile_lifted --make-bed --out plinkFile
```

2.- Create a frequency file

```
plink --freq --bfile plinkFile --out plinkFile
```

3.- Check bim

```
perl HRC-1000G-check-bim.pl -h
-r PASS.Variants.TOPMed_freeze5_hg38_dbSNP.tab.gz
-b plinkFile.bim -f plinkFile.frq
-p EUR
-o outputDir
```

c) Check bim

Run-plink.sh

```
plink --bfile plinkFile --exclude Exclude-plinkFile-HRC.txt --make-bed --out TEMP1
plink --bfile TEMP1 --update-map Chromosome-plinkFile-HRC.txt --update-chr --make-bed --out TEMP2
plink --bfile TEMP2 --update-map Position-plinkFile-HRC.txt --make-bed --out TEMP3
plink --bfile TEMP3 --flip Strand-Flip-plinkFile-HRC.txt --make-bed --out TEMP4
plink --bfile TEMP4 --a2-allele Force-Allele1-plinkFile-HRC.txt --make-bed --out plinkFile-updated
plink --bfile plinkFile-updated --real-ref-alleles --make-bed --chr 1 --out plinkFile-updated-chr1
plink --bfile plinkFile-updated --real-ref-alleles --recode vcf --chr 1 --out plinkFile-updated-chr1
plink --bfile plinkFile-updated --real-ref-alleles --make-bed --chr 2 --out plinkFile-updated-chr2
plink --bfile plinkFile-updated --real-ref-alleles --recode vcf --chr 2 --out plinkFile-updated-chr2
plink --bfile plinkFile-updated --real-ref-alleles --make-bed --chr 3 --out plinkFile-updated-chr3
plink --bfile plinkFile-updated --real-ref-alleles --recode vcf --chr 3 --out plinkFile-updated-chr3
plink --bfile plinkFile-updated --real-ref-alleles --make-bed --chr 4 --out plinkFile-updated-chr4
plink --bfile plinkFile-updated --real-ref-alleles --recode vcf --chr 4 --out plinkFile-updated-chr4
plink --bfile plinkFile-updated --real-ref-alleles --make-bed --chr 5 --out plinkFile-updated-chr5
plink --bfile plinkFile-updated --real-ref-alleles --recode vcf --chr 5 --out plinkFile-updated-chr5
```

• • •

```
plink --bfile plinkFile-updated --real-ref-alleles --recode vcf --chr 21 --out plinkFile-updated-chr21
plink --bfile plinkFile-updated --real-ref-alleles --make-bed --chr 22 --out plinkFile-updated-chr22
plink --bfile plinkFile-updated --real-ref-alleles --recode vcf --chr 22 --out plinkFile-updated-chr22
plink --bfile plinkFile-updated --real-ref-alleles --make-bed --chr 23 --out plinkFile-updated-chr23
plink --bfile plinkFile-updated --real-ref-alleles --recode vcf --chr 23 --out plinkFile-updated-chr23
rm TEMP*
```

d) Chromosome X heterozygosity in males

Setting as missing those heterozygous SNPs from male samples in chrX

```
plink
--bfile plinkFile-updated-chr23
--set-hh-missing
--chr-output chrM
--make-bed
--out plinkFile_chrX
```

e) Create and index vcf files for each chromosome

Save SNPs with the reference allele matching with reference panel

```
bim <- read.table("plinkFile-updated-chr20.bim")

write.table(bim[,c(2,6)], file = "snps.txt")
```

Extract data from each chromosome and save as vcf

```
plink
--bfile plinkFile-updated-chr20
--reference-allele snps.txt
--chr-output chrM
--recode vcf
--out  plinkFile_chr20
```

Create vcf.gz file

```
vcf-sort plinkFile_chr20.vcf | bgzip -c > plinkFile_chr20.vcf.gz
```

Index vcf.gz file

`bcftools index plinkFile_chr20.vcf.gz` → .csi (coordinate-sorted index) file is created.

Imputation

Algorithms

- IMPUTE4: https://jmarchini.org/impute-4/

- Minimac4: https://genome.sph.umich.edu/wiki/Minimac4

- Others: Beagle; MaCH, Ped_Pop, …

Algorithms

Method	Input data format	CPU seconds per sample *
Minimac4	individual=row; snp=column	3.97
Impute4	snp=row; individual=column	7.99

* Single-threaded CPU to impute chromosome 20 (1,718,742 markers) from 2,452 reference samples from the 1000 Genomes Project. Imputation analyses were run on a 2.6 GHz Intel Xeon E5-2630v2 computer with 128 GB of memory. CPU time is the sum of user and system time reported by the Unix time command. DOI: 10.1016/j.ajhg.2018.07.015

- Similar accuracy
- Different data formats
- Different processing time

a) Manually

Manual Imputation with Minimac

https://genome.sph.umich.edu/wiki/Minimac4

Data is usually broken into manageable chunks each phased independently:

```
Minimac4
--refHaps HRC.r1-1.GRCh37.chr20.shapeit3.mac5.aa.genotypes.m3vcf.gz
--haps chunk_20_0000000001_0020000000.phased.vcf
--chr 20
--start 1
--end 20000000
--window 500000
--prefix chunk_20_0000000001_0020000000
```

Manual Imputation with Impute

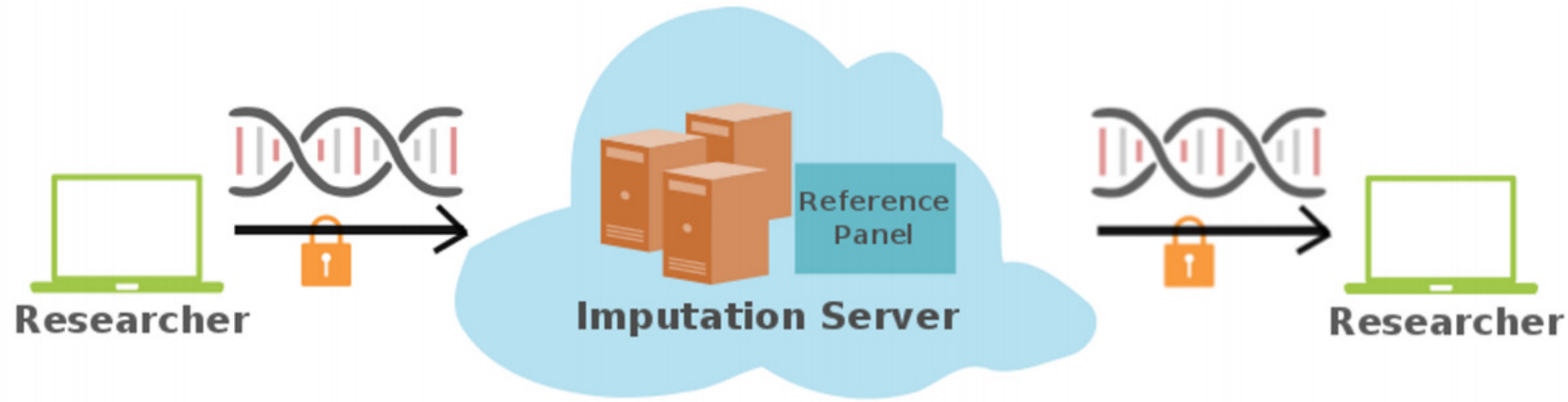
https://jmarchini.org/impute-4/

Data is usually broken into manageable chunks each phased independently:

```
impute4.1.2_r300.1
-h reference.hap.gz
-l reference.legend
-m chunk_20_0000000001_0020000000.phased.maps.txt
-g chunk_20_0000000001_0020000000.phased.haps.gz
-int 0 20000000
-o chunk_20_0000000001_0020000000
```

b) Server

Imputation in a server



1.

Upload GWAS data

2.

Server performs

- Quality checks
- Pre-phasing
- Imputation
- Encryption

3.

Download results

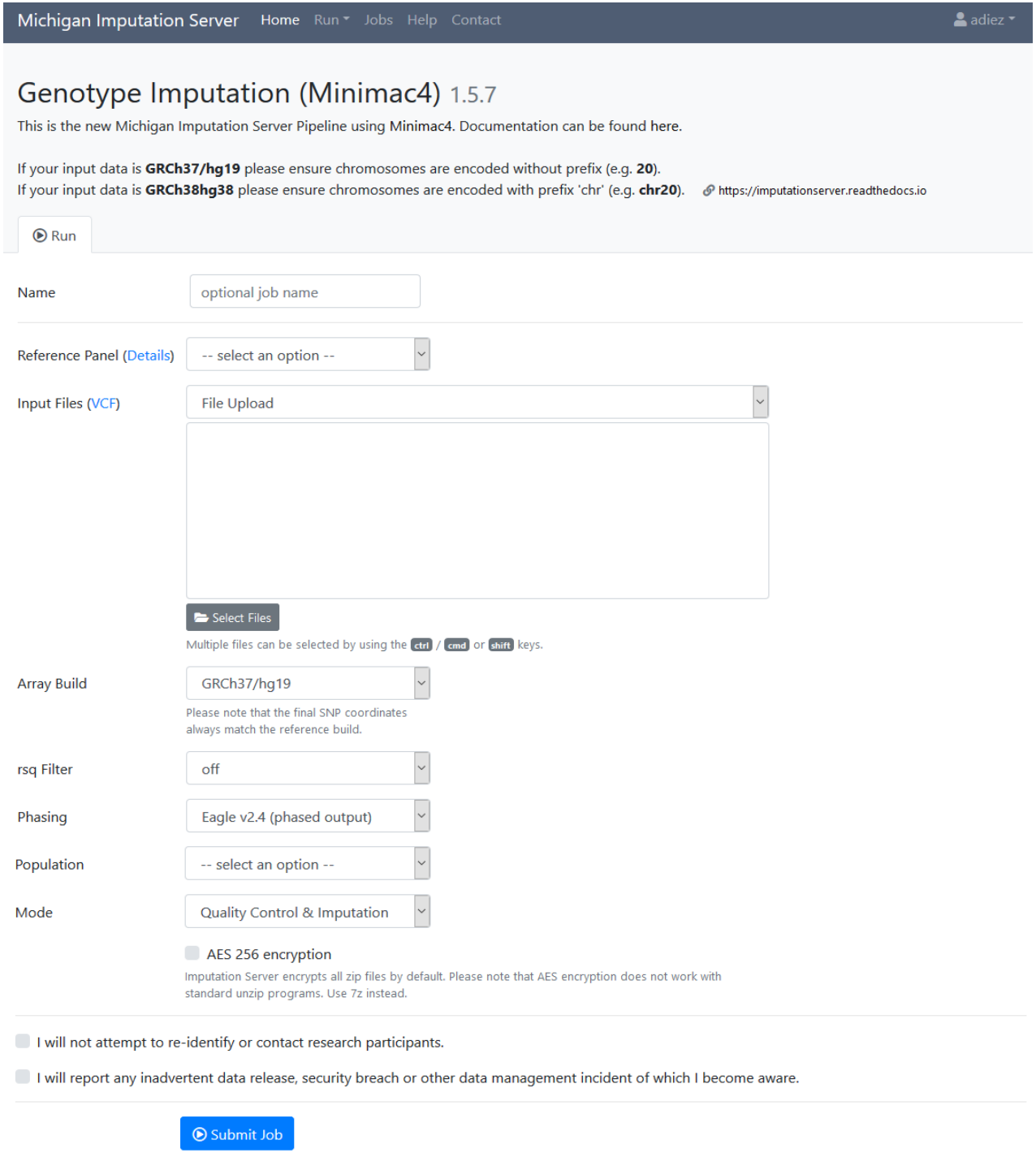
Imputation in a server

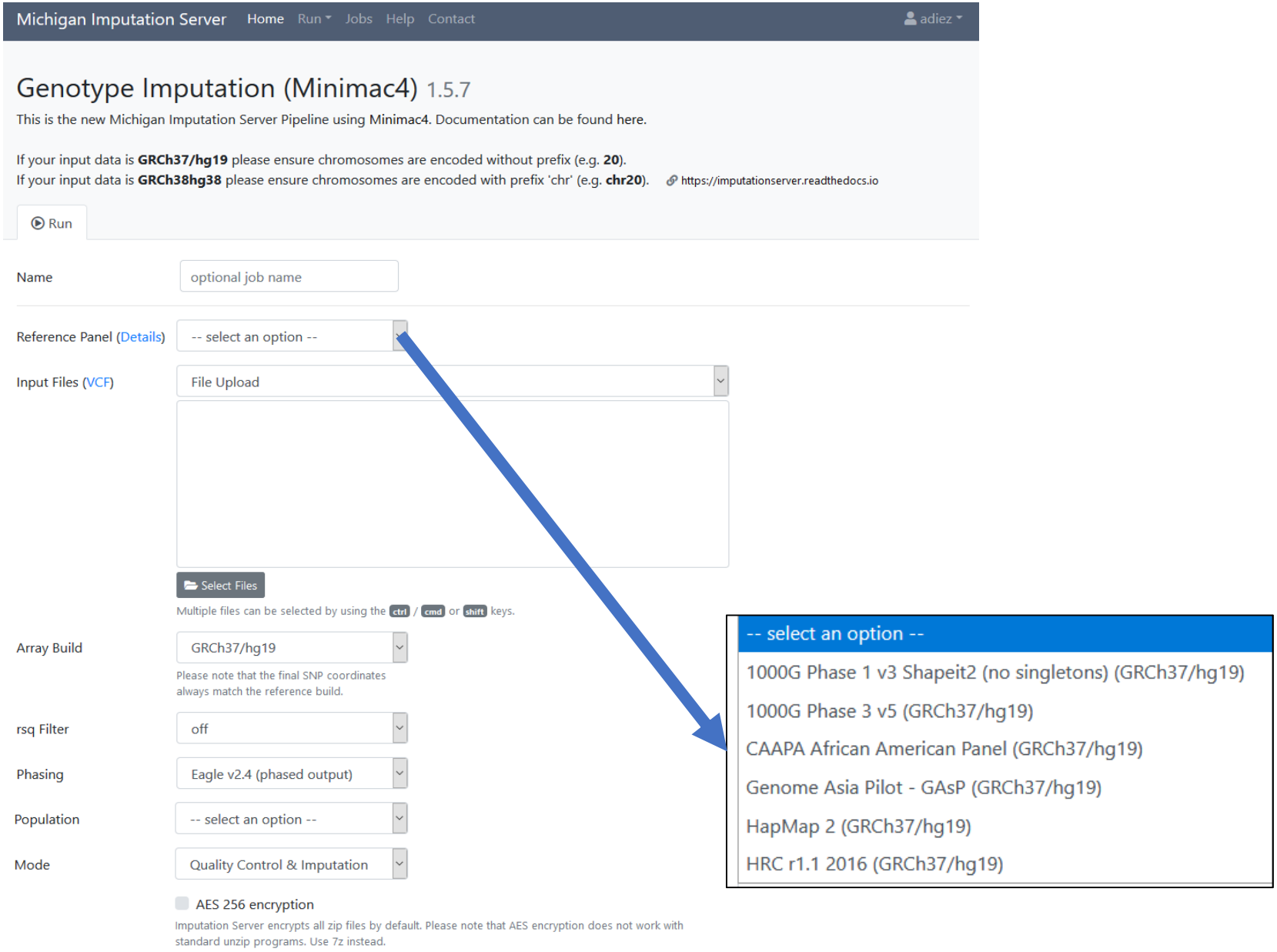
Submit a job

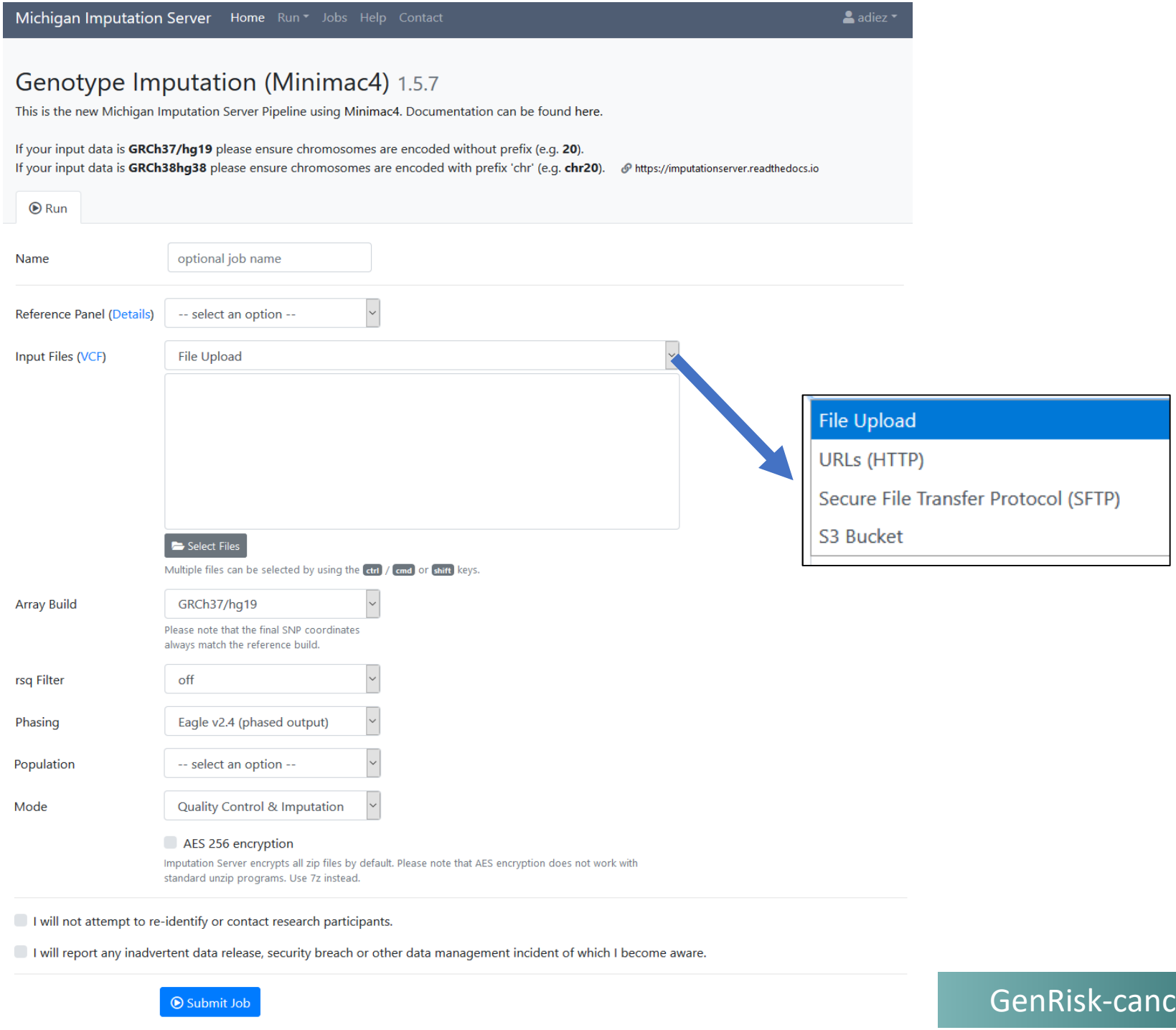
- Input Validation and Quality Control executed right after data upload
- Immediate feedback to users
- Jobs passing the QC are then added to a long-time queue for Phasing & Imputation
- Outputs SNP statistics and a QC Report for each job

Michigan Imputation Server

https://imputationserver.sph.umich.edu/index.html

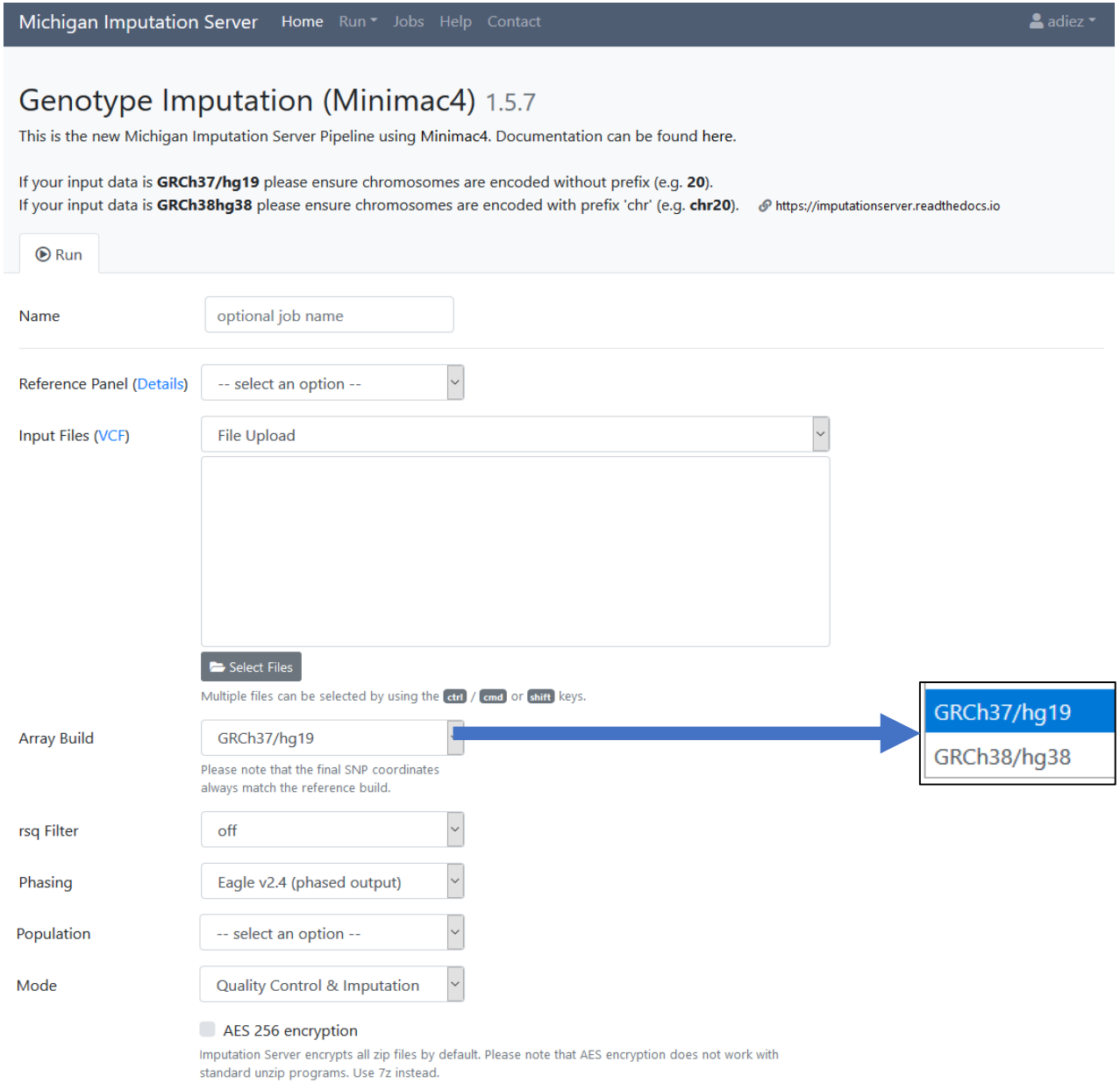

Michigan Imputation Server
Home
Run
Jobs
Help
Contact
adiez
Genotype Imputation (Minimac4) 1.5.7
This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found here.
If your input data is GRCh37/hg19 please ensure chromosomes are encoded without prefix (e.g. 20).
If your input data is GRCh38hg38 please ensure chromosomes are encoded with prefix 'chr' (e.g. chr20).
https://imputationserver.readthedocs.io
Run
Name
optional job name
Reference Panel (Details)
-- select an option --
Input Files (VCF)
File Upload
Select Files
Multiple files can be selected by using the ctrl / cmd or shift keys.
Array Build
GRCh37/hg19
Please note that the final SNP coordinates always match the reference build.
rsq Filter
off
Phasing
Eagle v2.4 (phased output)
Population
-- select an option --
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

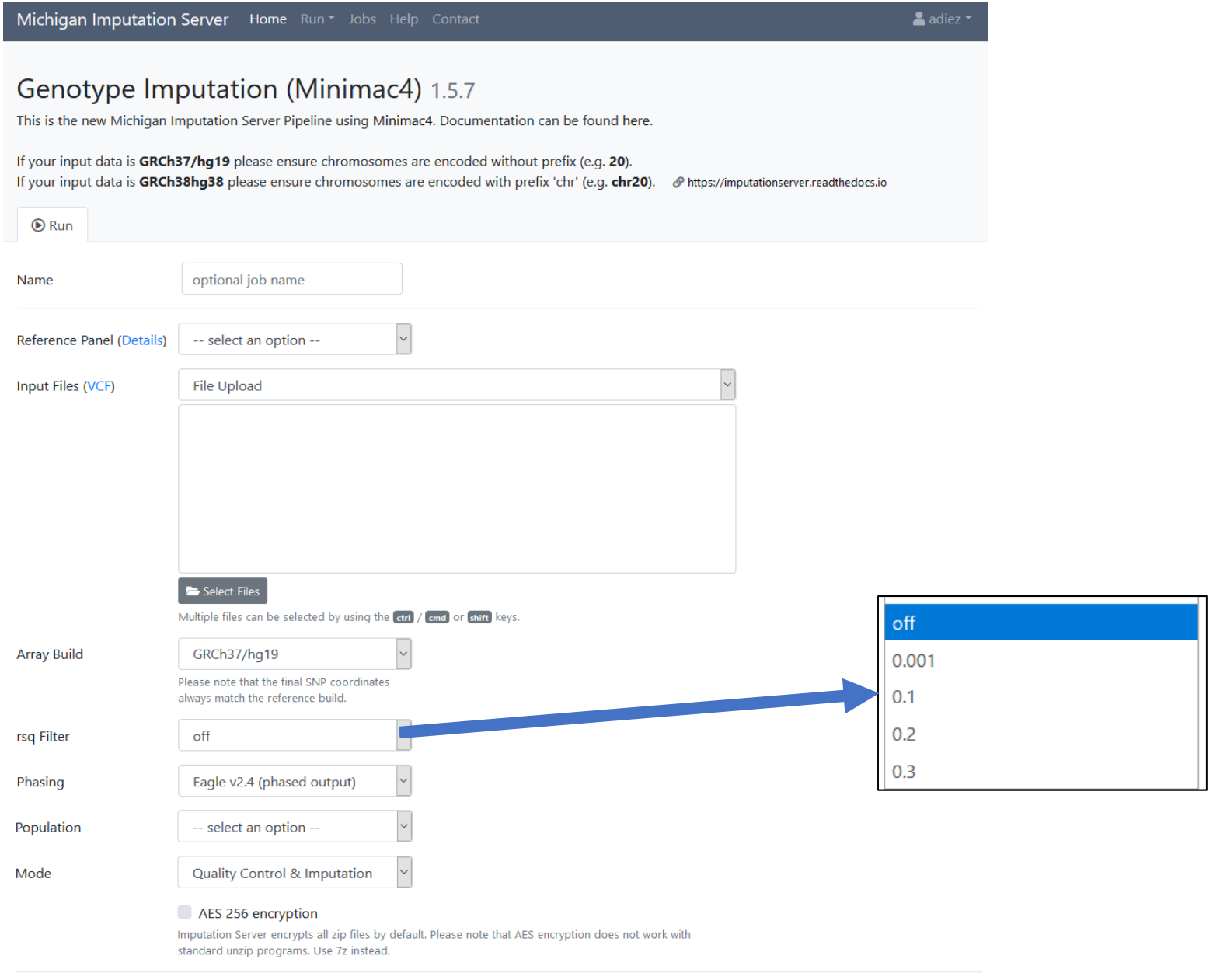

Michigan Imputation Server
Home
Run
Jobs
Help
Contact
adiez
Genotype Imputation (Minimac4) 1.5.7
This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found here.
If your input data is GRCh37/hg19 please ensure chromosomes are encoded without prefix (e.g. 20).
If your input data is GRCh38hg38 please ensure chromosomes are encoded with prefix 'chr' (e.g. chr20).
https://imputationserver.readthedocs.io
Run
Name
optional job name
Reference Panel (Details)
-- select an option --
Input Files (VCF)
File Upload
Select Files
Multiple files can be selected by using the ctrl / cmd or shift keys.
Array Build
GRCh37/hg19
Please note that the final SNP coordinates always match the reference build.
rsq Filter
off
Phasing
Eagle v2.4 (phased output)
Population
-- select an option --
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.
-- select an option --
1000G Phase 1 v3 Shapeit2 (no singletons) (GRCh37/hg19)
1000G Phase 3 v5 (GRCh37/hg19)
CAAPA African American Panel (GRCh37/hg19)
Genome Asia Pilot - GAsP (GRCh37/hg19)
HapMap 2 (GRCh37/hg19)
HRC r1.1 2016 (GRCh37/hg19)
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



ciberesp

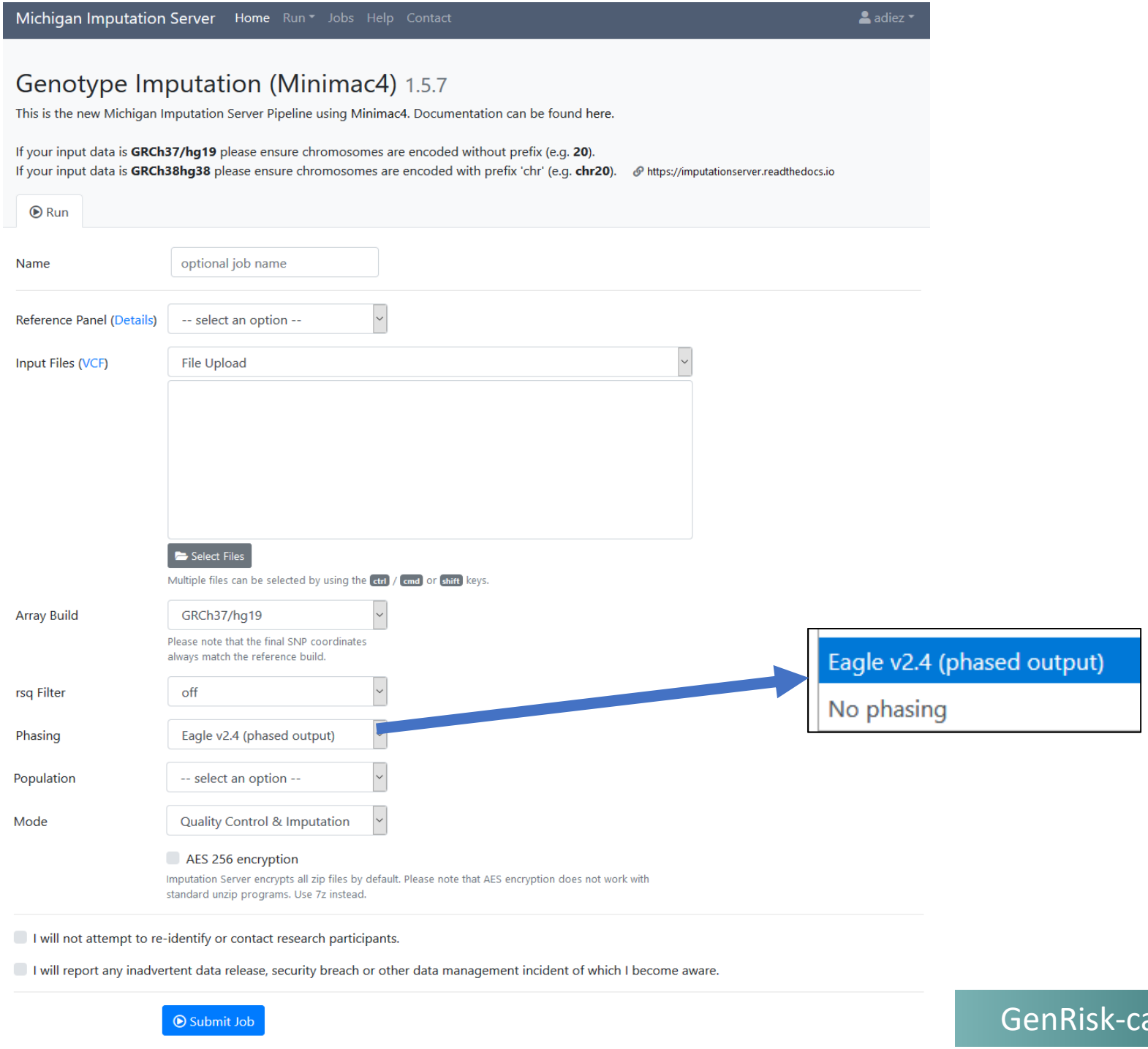
GenRisk-cancer

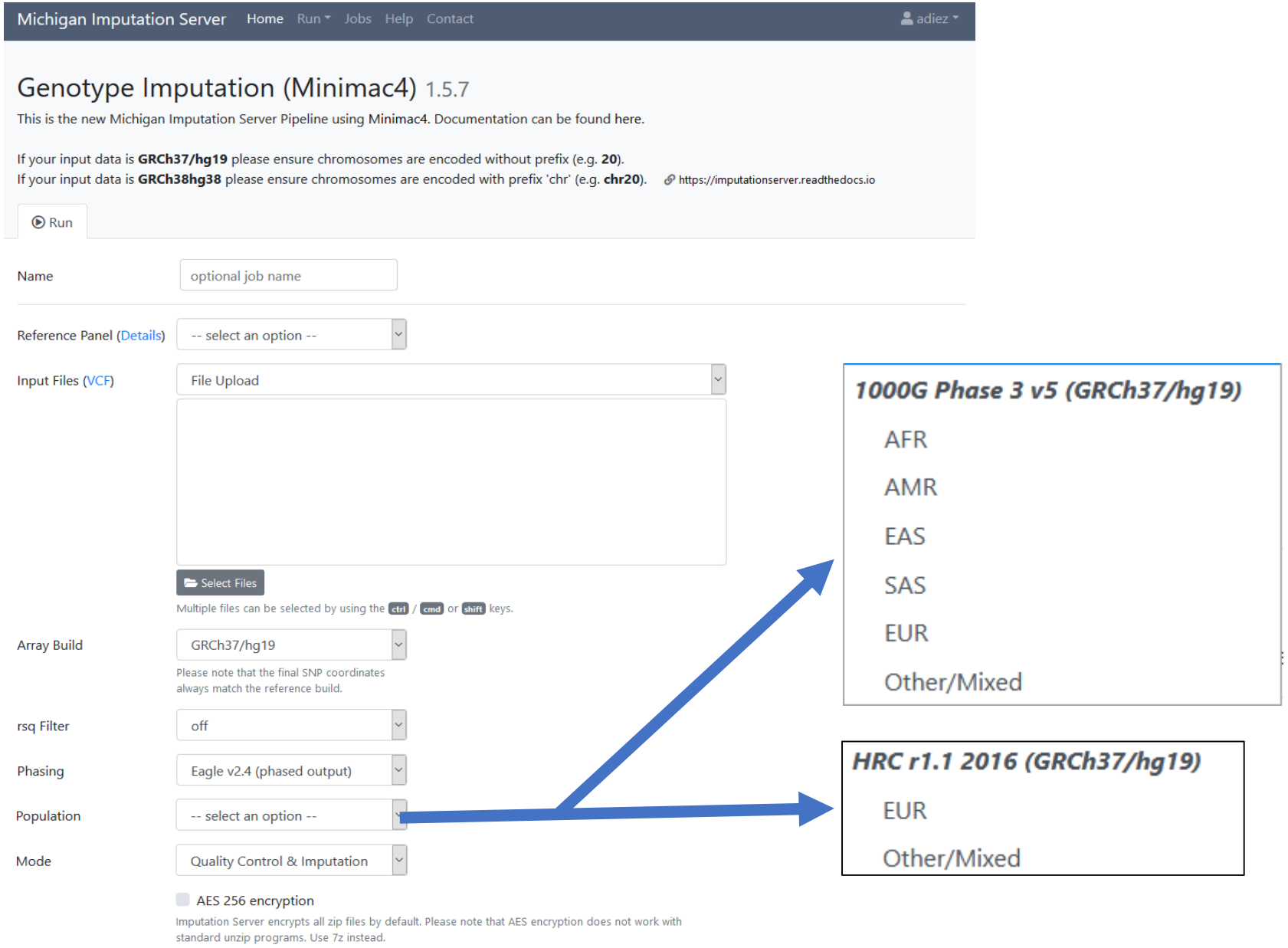

Michigan Imputation Server
Home
Run
Jobs
Help
Contact
adiez
Genotype Imputation (Minimac4) 1.5.7
This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found here.
If your input data is GRCh37/hg19 please ensure chromosomes are encoded without prefix (e.g. 20).
If your input data is GRCh38hg38 please ensure chromosomes are encoded with prefix 'chr' (e.g. chr20).
https://imputationserver.readthedocs.io
Run
Name
optional job name
Reference Panel (Details)
-- select an option --
Input Files (VCF)
File Upload
Select Files
Multiple files can be selected by using the ctrl / cmd or shift keys.
Array Build
GRCh37/hg19
Please note that the final SNP coordinates always match the reference build.
GRCh37/hg19
GRCh38/hg38
rsq Filter
off
Phasing
Eagle v2.4 (phased output)
Population
-- select an option --
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

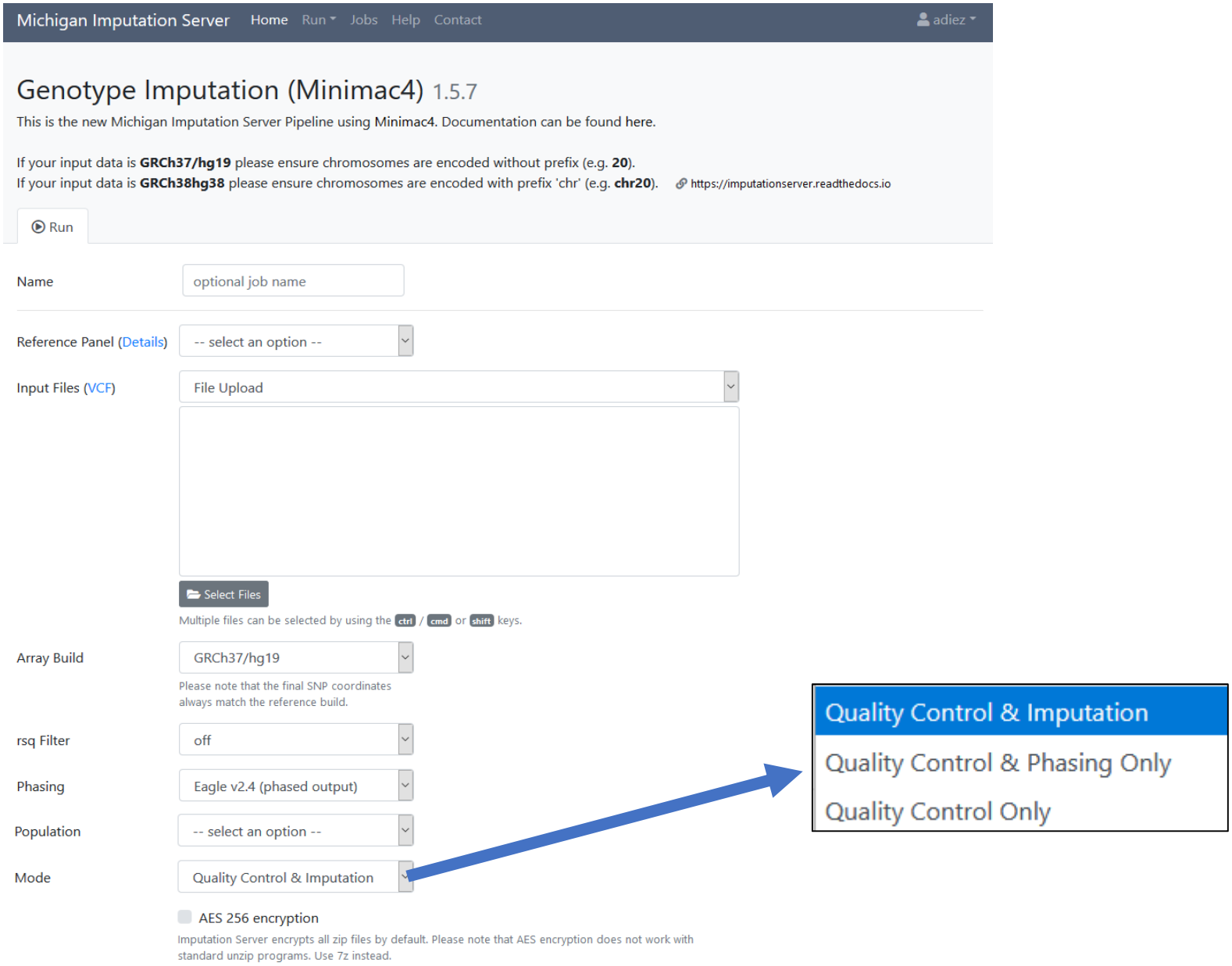


ciberesp

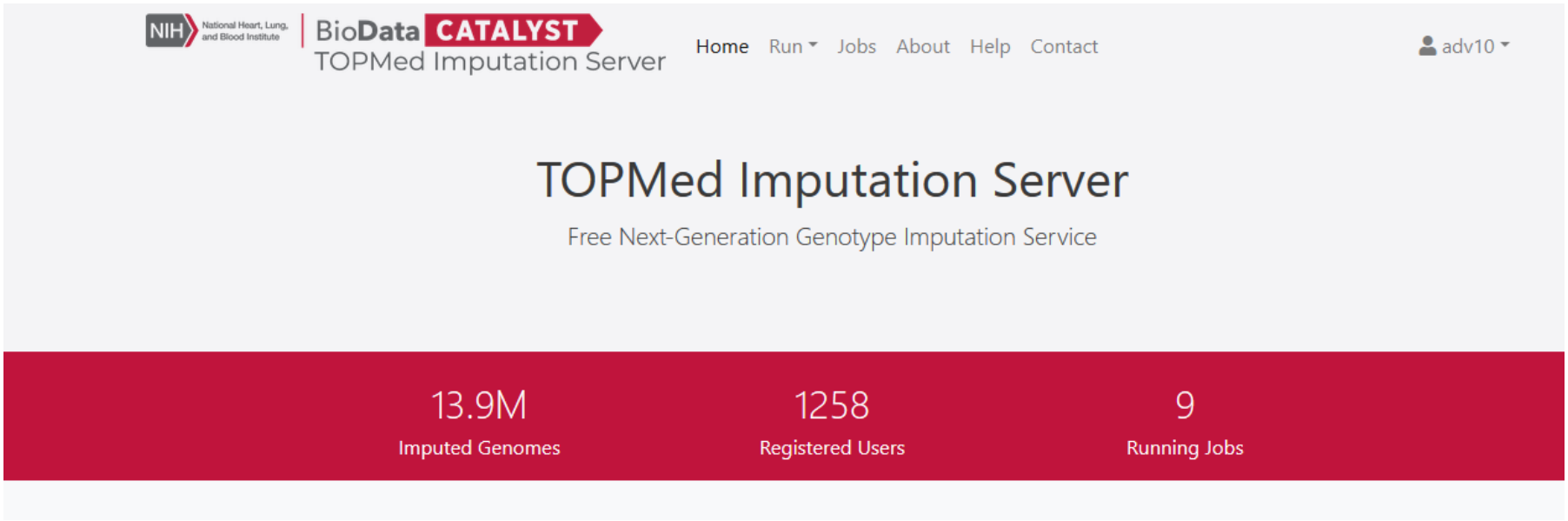
GenRisk-cancer


Michigan Imputation Server
Home
Run
Jobs
Help
Contact
adiez
Genotype Imputation (Minimac4) 1.5.7
This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found here.
If your input data is GRCh37/hg19 please ensure chromosomes are encoded without prefix (e.g. 20).
If your input data is GRCh38hg38 please ensure chromosomes are encoded with prefix 'chr' (e.g. chr20).
https://imputationserver.readthedocs.io
Run
Name
optional job name
Reference Panel (Details)
-- select an option --
Input Files (VCF)
File Upload
Select Files
Multiple files can be selected by using the ctrl / cmd or shift keys.
Array Build
GRCh37/hg19
Please note that the final SNP coordinates always match the reference build.
rsq Filter
off
Phasing
Eagle v2.4 (phased output)
Population
-- select an option --
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
Eagle v2.4 (phased output)
No phasing

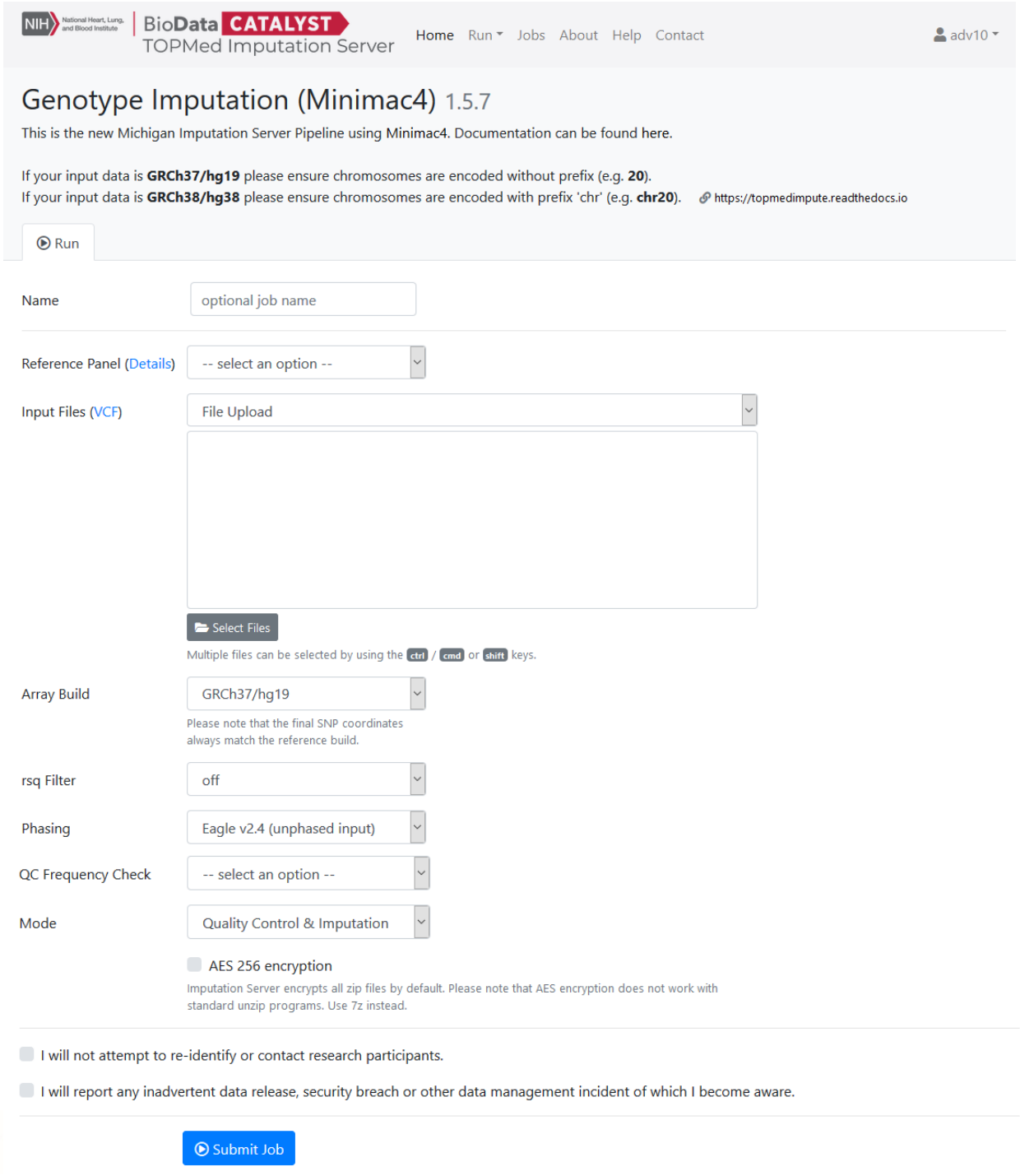

Michigan Imputation Server
Home
Run
Jobs
Help
Contact
adiez
Genotype Imputation (Minimac4) 1.5.7
This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found here.
If your input data is GRCh37/hg19 please ensure chromosomes are encoded without prefix (e.g. 20).
If your input data is GRCh38hg38 please ensure chromosomes are encoded with prefix 'chr' (e.g. chr20).
https://imputationserver.readthedocs.io
Run
Name
optional job name
Reference Panel (Details)
-- select an option --
Input Files (VCF)
File Upload
Select Files
Multiple files can be selected by using the ctrl / cmd or shift keys.
Array Build
GRCh37/hg19
Please note that the final SNP coordinates always match the reference build.
rsq Filter
off
Phasing
Eagle v2.4 (phased output)
Population
-- select an option --
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.
1000G Phase 3 v5 (GRCh37/hg19)
AFR
AMR
EAS
SAS
EUR
Other/Mixed
HRC r1.1 2016 (GRCh37/hg19)
EUR
Other/Mixed
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


Michigan Imputation Server
Home
Run
Jobs
Help
Contact
adiez
Genotype Imputation (Minimac4) 1.5.7
This is the new Michigan Imputation Server Pipeline using Minimac4. Documentation can be found here.
If your input data is GRCh37/hg19 please ensure chromosomes are encoded without prefix (e.g. 20).
If your input data is GRCh38hg38 please ensure chromosomes are encoded with prefix 'chr' (e.g. chr20).
https://imputationserver.readthedocs.io
Run
Name
optional job name
Reference Panel (Details)
-- select an option --
Input Files (VCF)
File Upload
Select Files
Multiple files can be selected by using the ctrl / cmd or shift keys.
Array Build
GRCh37/hg19
Please note that the final SNP coordinates always match the reference build.
rsq Filter
off
Phasing
Eagle v2.4 (phased output)
Population
-- select an option --
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
Quality Control & Imputation
Quality Control & Phasing Only
Quality Control Only

TOPMed Imputation Server

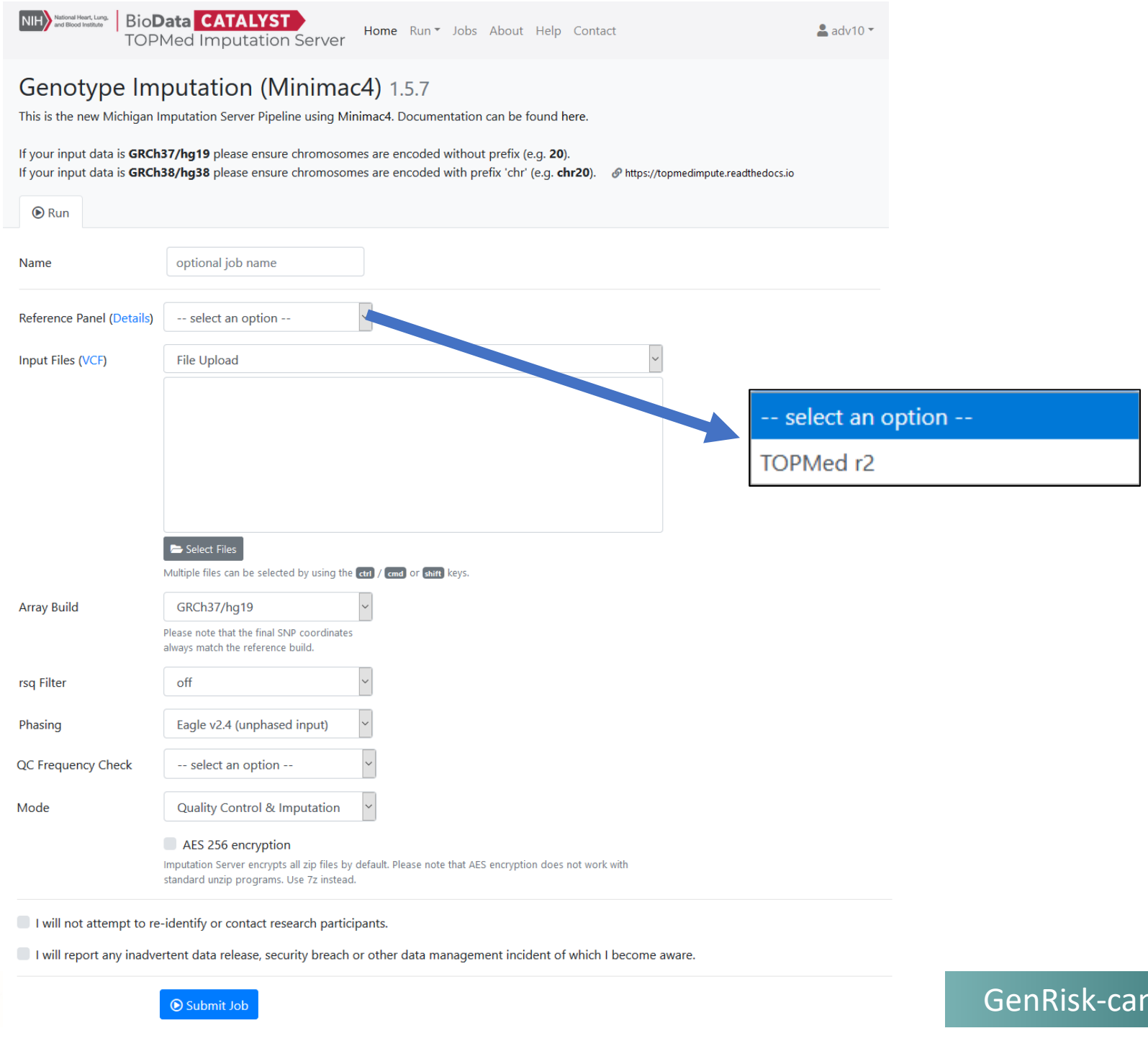


https://imputation.biodatacatalyst.nhlbi.nih.gov



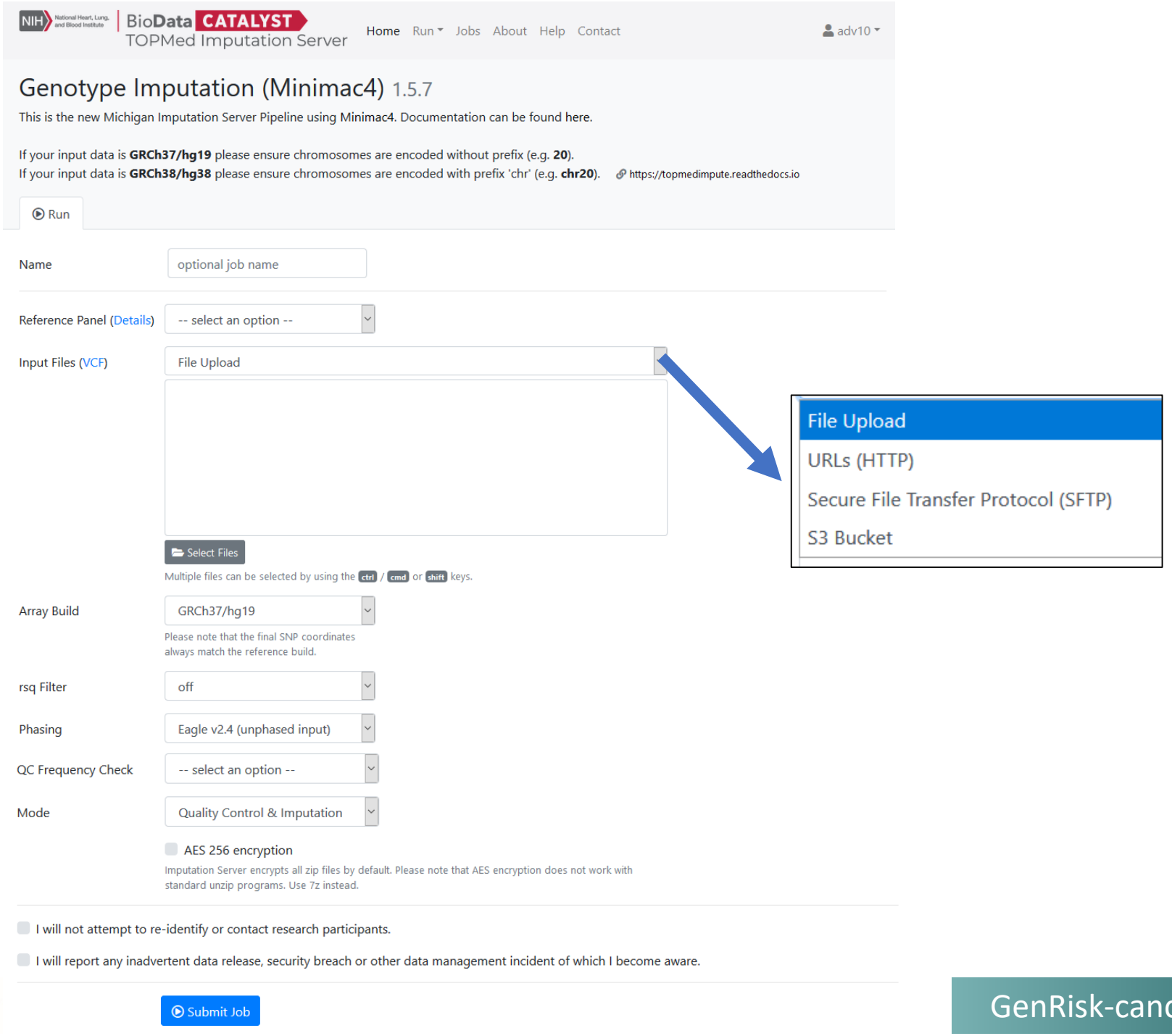
ciberesp

GenRisk-cancer



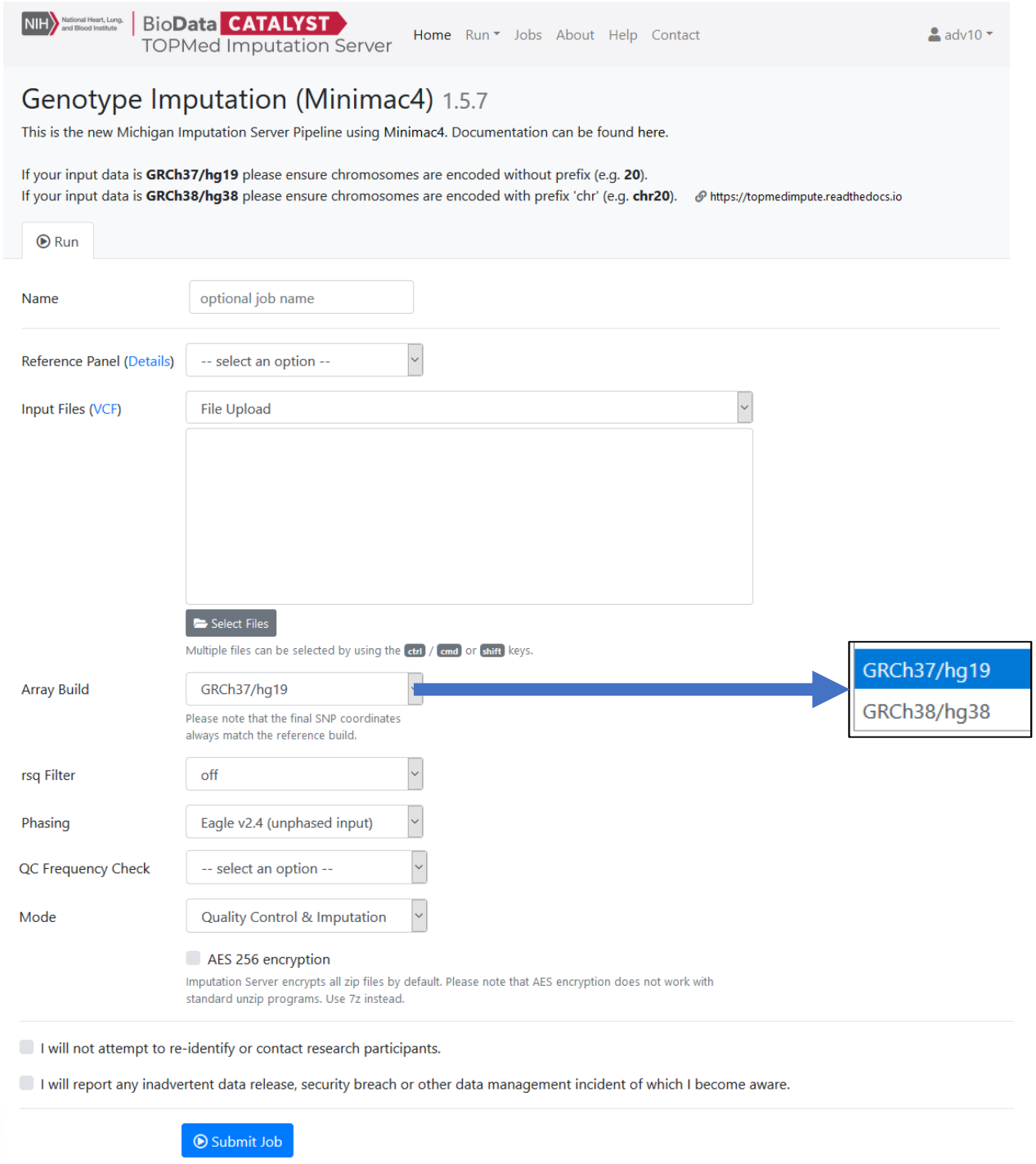
ciberesp

GenRisk-cancer



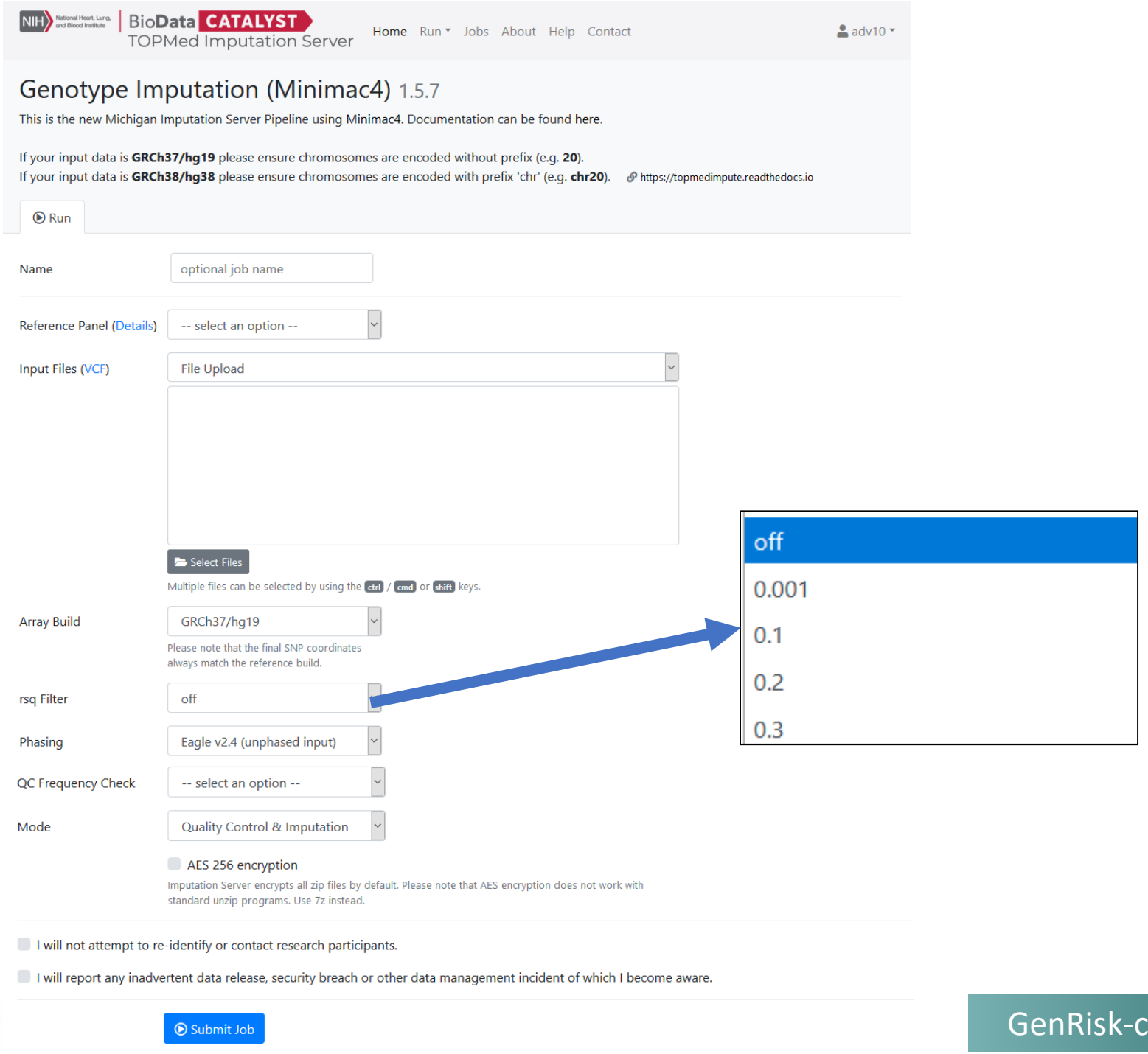
ciberesp

GenRisk-cancer



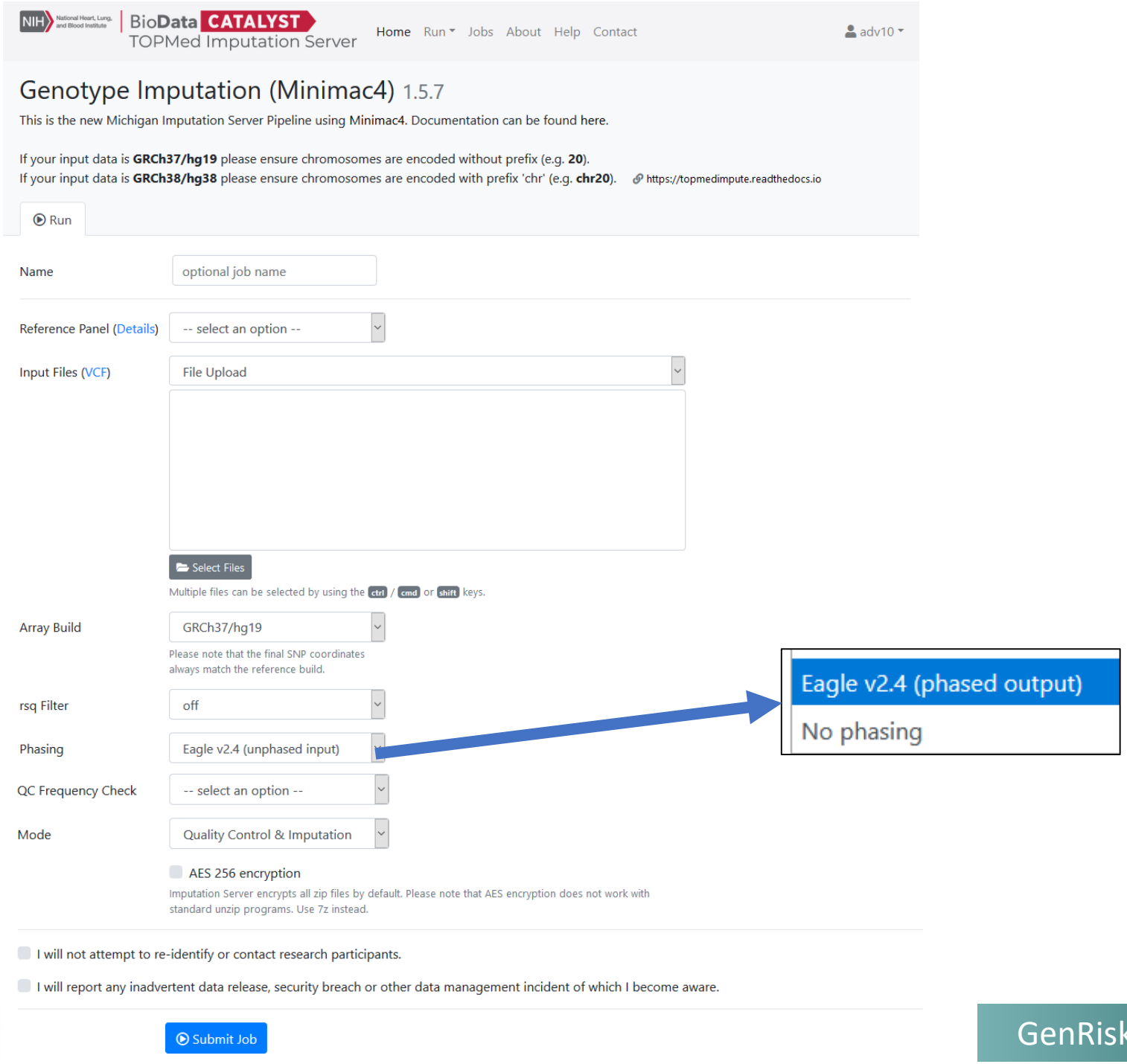
ciberesp

GenRisk-cancer



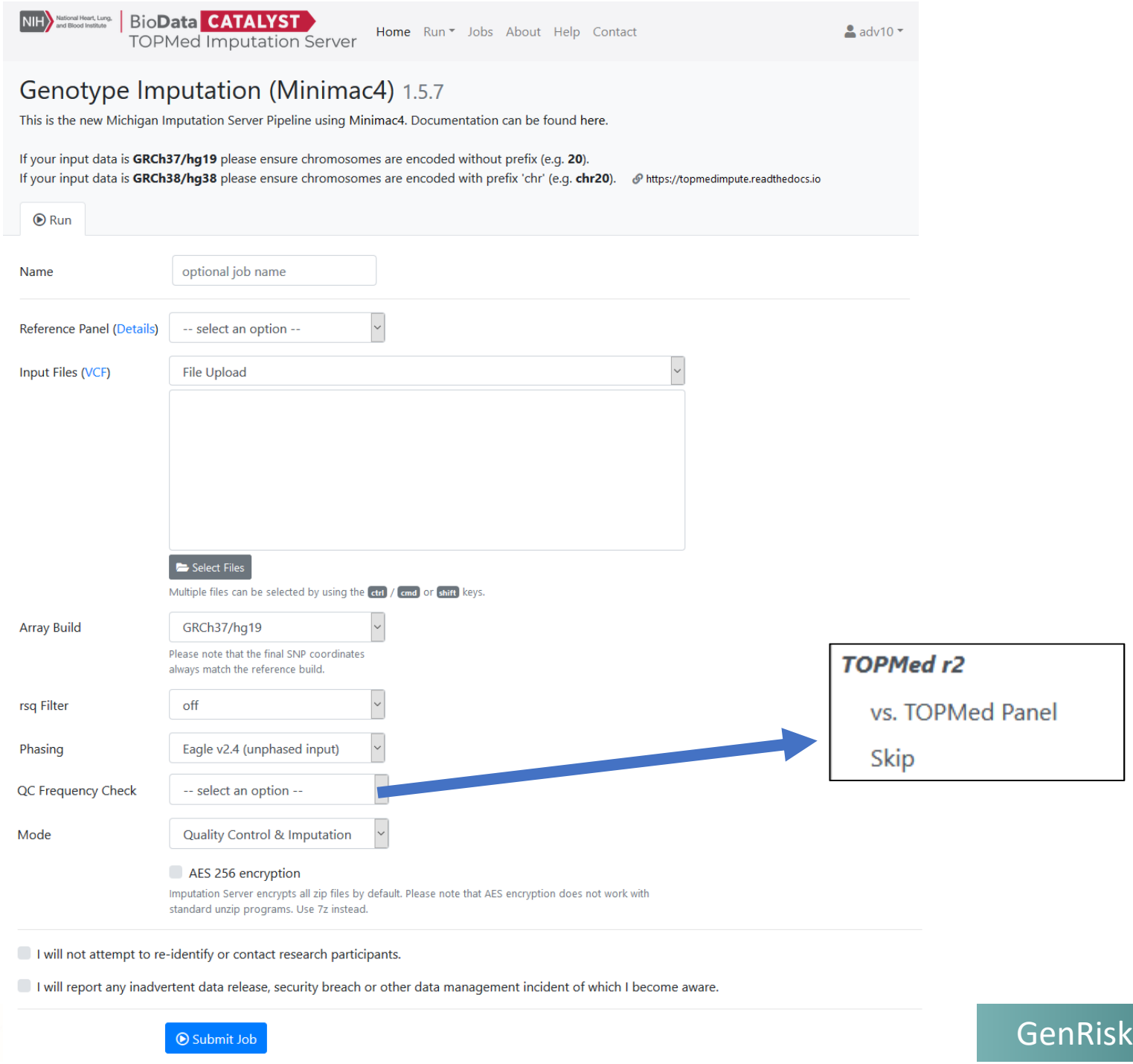
ciberesp

GenRisk-cancer



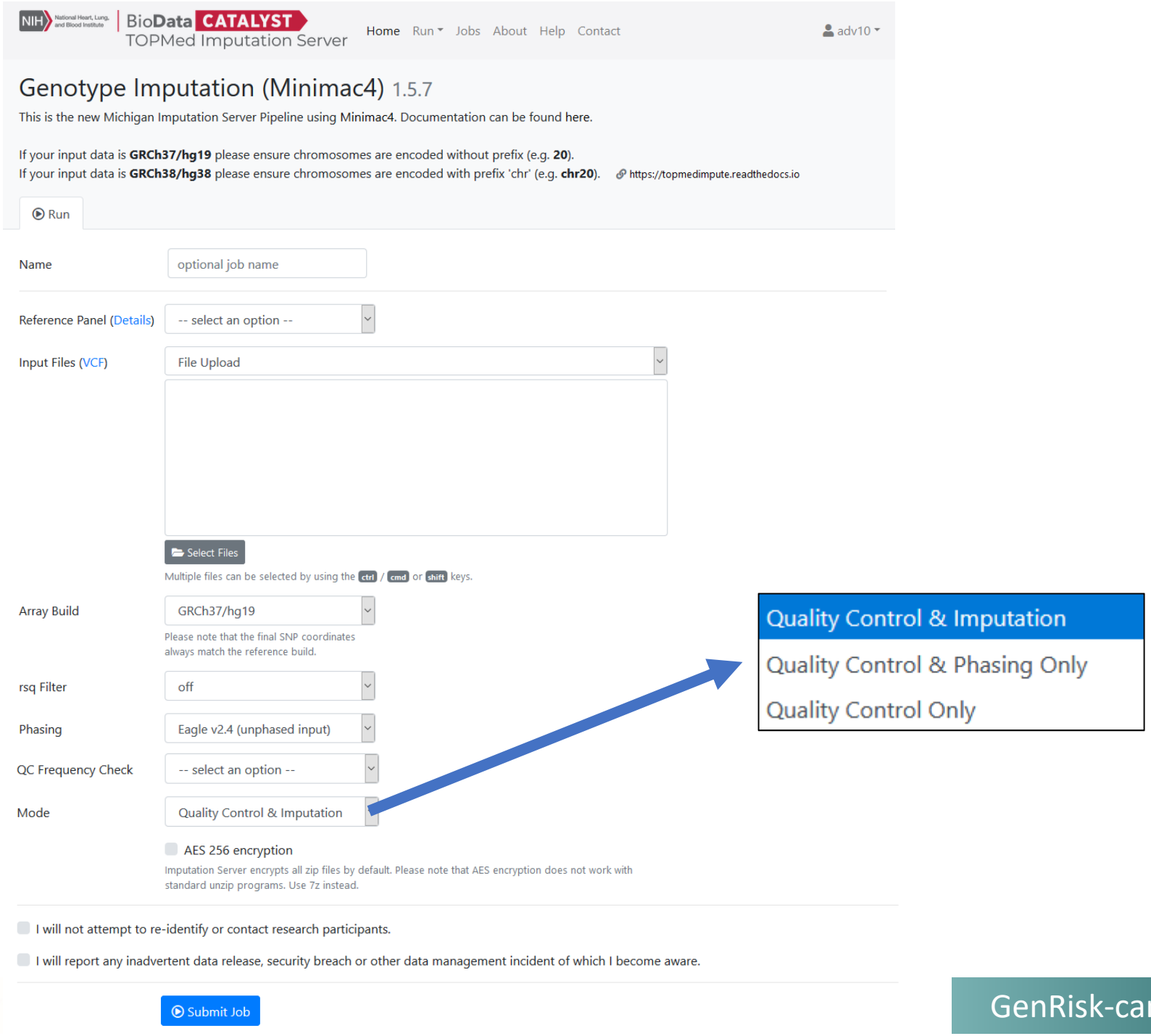
ciberesp

GenRisk-cancer



ciberesp

GenRisk-cancer



ciberesp

GenRisk-cancer

Output from the server

Input errors:

Input Validation

The provided VCF file is malformed. Error during index creation: [tabix] was bgzip used to compress this file? (see Help).

Input Validation

The provided VCF file contains more than one chromosome. Please split your input VCF file by chromosome (see Help).

Input Validation

Unable to parse header with error: Your input file has a malformed header: We never saw the required CHROM header line (starting with one #) for the input VCF file (see Help).

Input errors:

Excluded sites in total: 695
Remaining sites in total: 185,791
See snps-excluded.txt for details
Typed only sites: 397
See typed-only.txt for details

Warning: 2 Chunk(s) excluded: reference overlap < 50.0% (see chunks-excluded.txt for details).
Remaining chunk(s): 40
Error: More than 100 obvious strand flips have been detected. Please check strand. Imputation cannot be started!

Input errors:

Filtered sites:
Filter flag set: 0
Invalid alleles: 0
Multiallelic sites: 0
Duplicated sites: 0
NonSNP sites: 0
Monomorphic sites: 688
Allele mismatch: 0
SNPs call rate < 90%: 0

Excluded sites in total: 688
Remaining sites in total: 1,325,650
See snps-excluded.txt for details

Pre-phasing and Imputation

Chr 11 Chr 22 Chr 12 Chr 13 Chr 14 Chr 15
Chr 16 Chr 17 Chr 18 Chr 19 Chr 1 Chr 2
Chr 3 Chr 4 Chr 5 Chr 6 Chr 7 Chr 8
Chr 9 Chr 20 Chr 10 Chr 21

Download Results

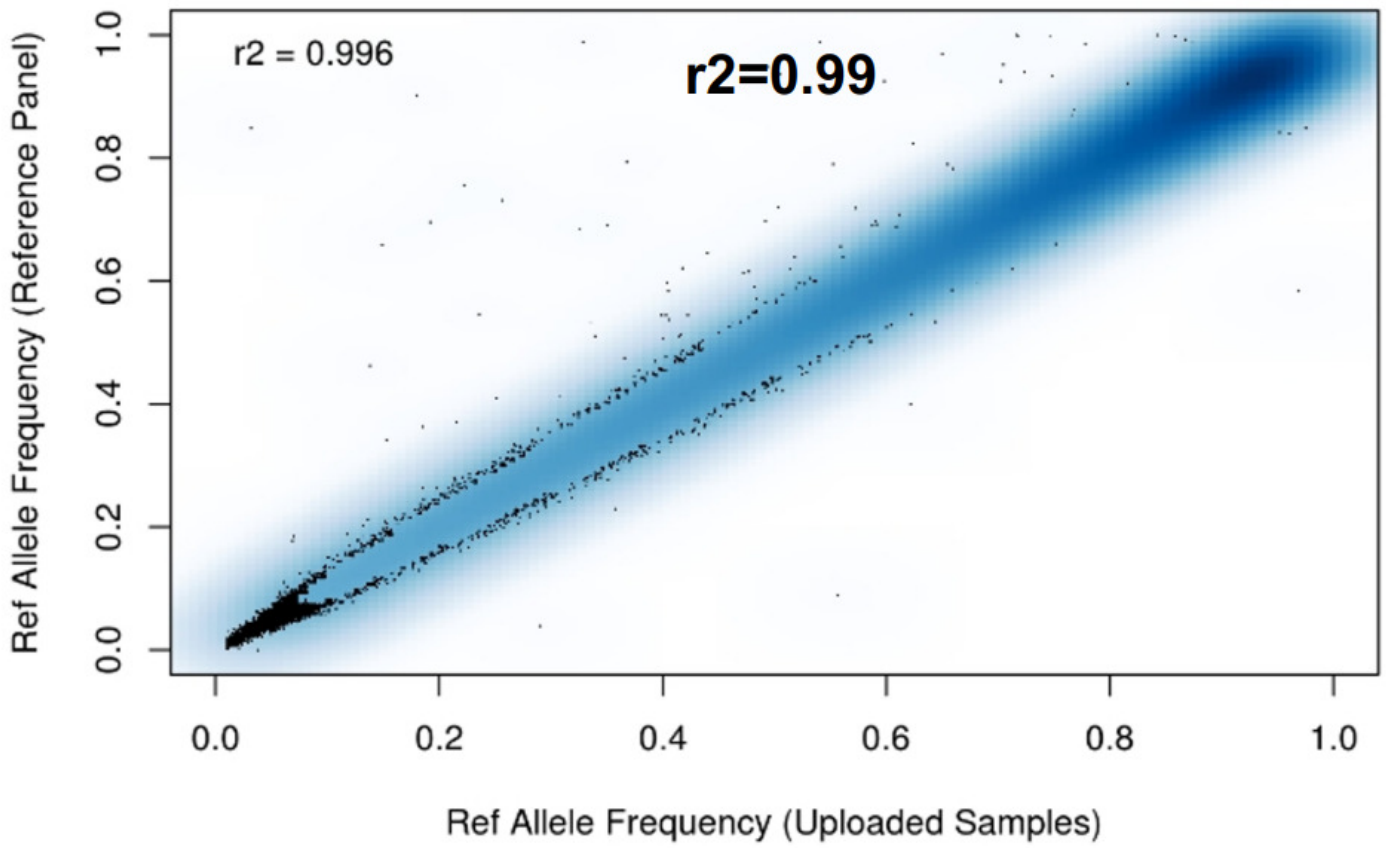
- An email with a password is sent

Dear Anna Díez-Villanueva,
the password for the imputation results is: NgB4ZipiS4frUH

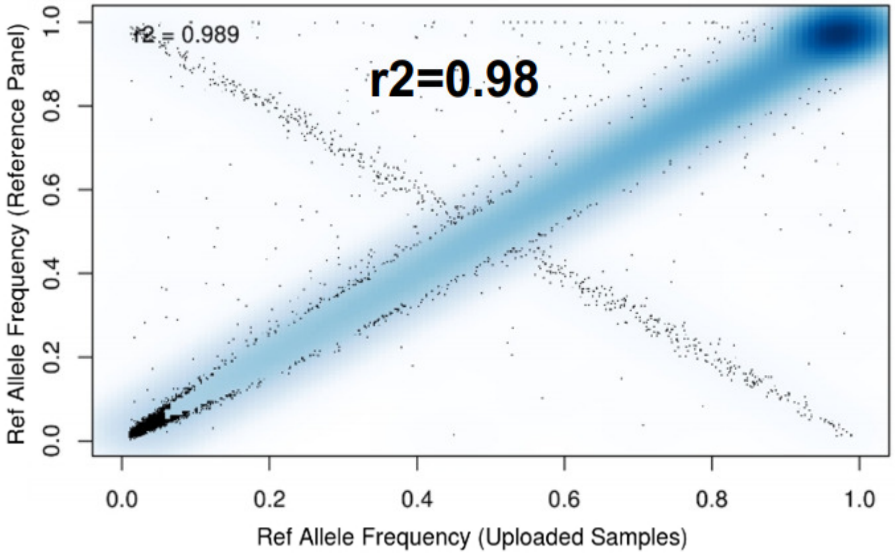
The results can be downloaded from https://imputation.biodatacatalyst.nhlbi.nih.gov/start.html#!jobs/job-20210318-160032-233/results

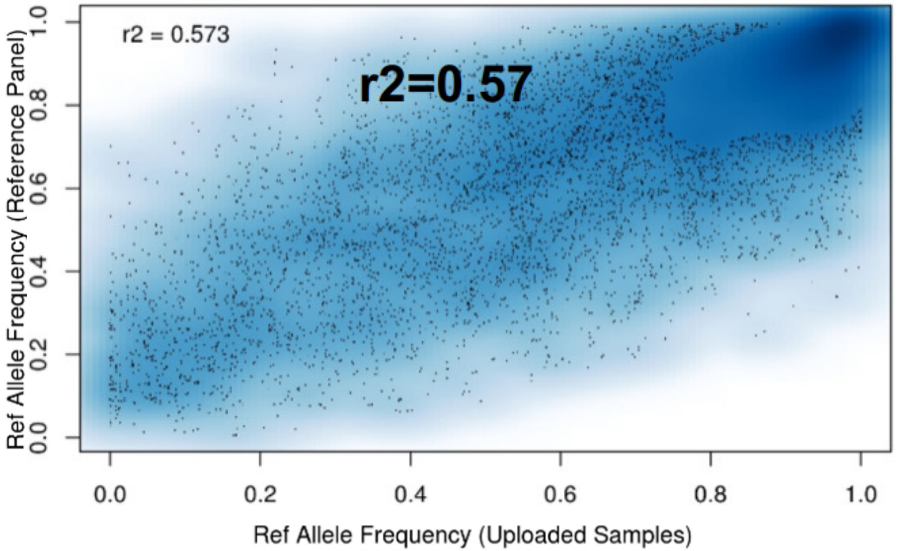
- All imputed genotypes are in encrypted zip files (e.g. chr_1.zip)

QC Report: Allele Frequency Check

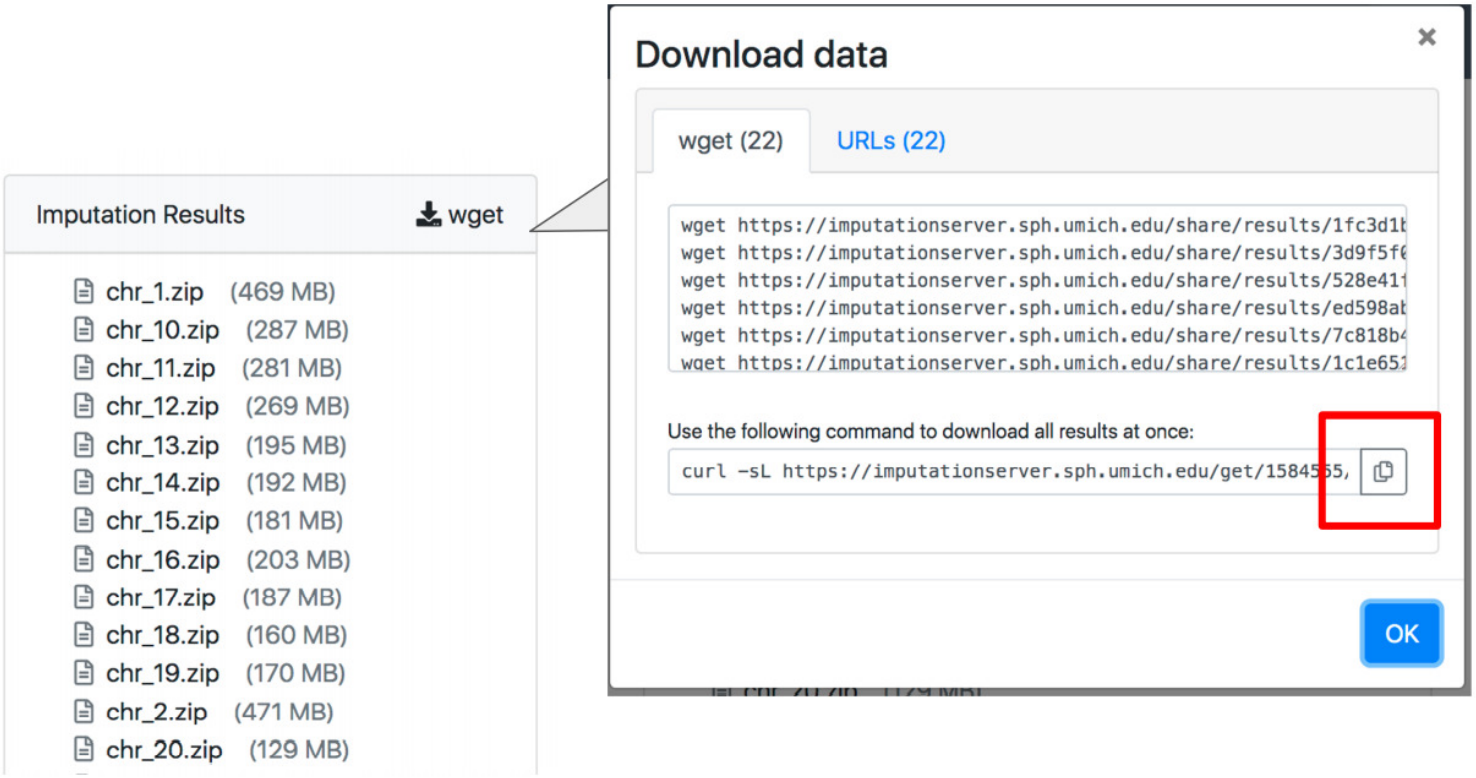


QC Report: Allele Frequency Errors





Download Results



chr_20.zip → chr20.dose.vcf.gz
chr_20.zip → chr20.info.gz

Info File

Example: chr20.info.gz

SNP	REF(0)	ALT(1)	ALT_Frq	MAF	AvgCall	Rsq	Genotyped	...
20:61795:G:T	G	T	0.26318	0.26318	0.88455	0.54658	Imputed	...
20:63231:T:G	T	G	0.03843	0.03843	0.98342	0.67736	Imputed	...
20:63244:A:C	A	C	0.16132	0.16132	0.91761	0.49907	Imputed	...

Rsq

- Estimated value of the squared correlation between imputed genotypes and true, unobserved genotypes.
- Observed dosage variance / Expected dosage variance, given observed allele frequency and assuming Hardy-Weinberg equilibrium.

Minimal Rsq value for common variants 0.30
Minimal Rsq value for rare variants 0.50

Dosage File

Example: chr20.dose.vcf.gz

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	61795	20:61795:G:T	G	T	.	PASS	AF=0.26318;MAF=0.26318 R2=0.54658;IMPUTED
20	63231	20:63231:T:G	T	G	.	PASS	AF=0.03843;MAF=0.03843 R2=0.67736;IMPUTED
20	63244	20:63244:A:C	A	C	.	PASS	AF=0.16132;MAF=0.16132 R2=0.49907;IMPUTED
20	68749	20:68749:T:C	T	C	.	PASS	AF=0.59894;MAF=0.40106 R2=0.98392;TYPED
20	161502	20:161502:C:T	C	T	.	PASS	AF=0.05882;MAF=0.05882 TYPED_ONLY

FORMAT	Sample1
GT:DS:GP	1|0:1.126:0.100,0.673,0.226
GT:DS:GP	0|0:0.002:0.998,0.002,0.000
GT:DS:GP	0|0:0.285:0.723,0.270,0.008
GT:DS:GP	1|1:1.999:0.000,0.001,0.999
GT:DS:GP	0|0:0:1,0,0

GT → genotype
DS → dosage
GP → genotype posterior probabilities

c) Bot

Imputation bot

GitHub - lukfor/imputationbot: Automate interactions with the imputation servers

- Automate remote imputation
- Submit and monitor jobs from the command line
- Different commands can easily be combined
- Improves automation

Practical Case

https://mybinder.org/v2/gh/victor-moreno/ImputationTutorial/master?urlpath=rstudio

HapMap 3 samples

https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html

- 1.397 samples from 11 human populations.

- Assembly hg18.

ASW	African ancestry in Southwest USA
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CHB	Han Chinese in Beijing, China
CHD	Chinese in Metropolitan Denver, Colorado
GIH	Gujarati Indians in Houston, Texas
JPT	Japanese in Tokyo, Japan
LWK	Luhya in Webuye, Kenya
MXL	Mexican ancestry in Los Angeles, California
MKK	Maasai in Kinyawa, Kenya
TSI	Toscani in Italia
YRI	Yoruba in Ibadan, Nigeria

→ ~100 random samples selected for the practical. (~80 European and ~20 African)

Plink files: BED, BIM, FAM

bim → SNP annotations

1. Chromosome code
2. Variant identifier
3. Position in morgans or centimorgans
4. Base-pair coordinate (1-based)
5. Allele 1 (minor)
6. Allele 2 (major)

```
20   rs6078030    0.1781993    61098   A   G
20   GSA-rs6076506   0.1882266   63231   C   A
20   rs60263736   0.2246549    70980   A   G
20   rs892665     0.244747     75254   A   C
20   chr20-76786  0.251949     76786   A   G
20   rs1935386    0.2532204    87416   A   C
20   GSA-rs75507632  0.2337057   90814   G   A
20   rs13039134   0.2247925    92366   G   A
20   rs6052070    0.1985757    96931   G   A
20   rs6037772    0.1780502    100505  G   A
```

fam → sample information

1. Family ID ('FID')
2. Within-family ID ('IID'; cannot be '0')
3. Within-family ID of father ('0' if father isn't in dataset)
4. Within-family ID of mother ('0' if mother isn't in dataset)
5. Sex code ('1' = male, '2' = female, '0' = unknown)
6. Phenotype value ('1' = control, '2' = case, '-9'/'0'/non-numeric = missing data if case/control)

```
S-1-0154 S-1-0154 0 0 1 2
A-0-0147 A-0-0147 0 0 2 1
S-1-0110 S-1-0110 0 0 1 2
A-0-0201 A-0-0201 0 0 2 1
S-1-0095 S-1-0095 0 0 1 2
A-0-0152 A-0-0152 0 0 2 1
A-0-0178 A-0-0178 0 0 2 1
F-0-0250 F-0-0250 0 0 1 1
```

bed → binary biallelic genotype data

Plink files: PED, MAP

map → SNP annotations

1. Chromosome code
2. Variant identifier
3. Position in morgans or centimorgans
4. Base-pair coordinate

```
20  rs6078030     0.0 80457
20  GSA-rs6076506    0.0 82590
20  rs60263736    0.0 90339
20  rs892665      0.0 94613
20  rs1935386     0.0 106775
20  GSA-rs75507632   0.0 110173
20  rs13039134    0.0 111725
20  rs6052070     0.0 116290
20  rs6037772     0.0 119864
```

ped → sample information

One line per sample. The first six fields are the same as those in a .fam file. The seventh and eighth fields are allele calls for the first variant; the 9th and 10th are allele calls for the second variant; and so on.

```
S-1-0154    S-1-0154    0   0   1   2   A A A A G G A C C C G A G G A A A A G G G G A A A A A A
A-0-0147    A-0-0147    0   0   2   1   A G A A G G C C A C A A G A A A A A A G G G G A A A A A A
S-1-0110    S-1-0110    0   0   1   2   G G A A G G A C A C A A G A A A A A A G G A G A A A A A A
A-0-0201    A-0-0201    0   0   2   1   G G A A G G C C A C A A G A A A A A A G G G G A A A A A A
S-1-0095    S-1-0095    0   0   1   2   G G A A G G A C C C A A G G A A A A A G G G G G A A A A A
A-0-0152    A-0-0152    0   0   2   1   G G A A G G A C C C A A G G A A A A A G G G G A A A A A A
A-0-0178    A-0-0178    0   0   2   1   G G A A G G C C A C A A G A A A A A A G G A G A A A A A A
F-0-0250    F-0-0250    0   0   1   1   G G A A G G A C C C A A G A G A A A A A G G G G A A A A A G A
S-1-0037    S-1-0037    0   0   2   2   G G A A G G A A C C A A G G A A A A A G G G G A A A A A A A
S-1-0009    S-1-0009    0   0   2   2   G G A A G G A C C C A A A A A A A A A A G G G G A A A A A A
```