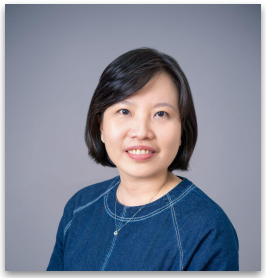
MICHIGAN IMPUTATIONSERVER

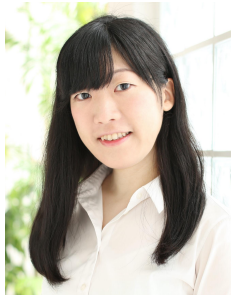
Christian Fuchsberger

Sebastian Schönherr



Xueling Sim

Lukas Forer



Saori Sakaue

Albert Smith

ASHG
American Society of Human Genetics

Disclosure Slide

Financial Disclosure for:

Christian Fuchsberger

Sebastian Schönherr

Lukas Forer

Xueling Sim

Albert Smith

Saori Sakaue

We have nothing to disclose

Setup

- 6 Sessions:
  (1) Intro, (2) Use the server and the Imputation Bot, (3) GWAS, (4) GWAS pipeline and PGS Server, (5) HLA Imputation, (6) TOPMed
  - Lectures
  - Demos
  - Interaction
    - PollEv.com/ashg
- Question & Answer session at the end
  - ASHG Q&A Tool during the session

ASHG
American Society of Human Genetics

Section 1

Imputation and the Server

Christian Fuchsberger
Eurac Research
cfuchsberger@eurac.edu

MICHIGAN IMPUTATIONSERVER

ASHG American Society of Human Genetics

Learning objectives

Participants will

1. Understand the principles of genotype imputation and the Michigan Imputation Server

Genotype imputation

Key method used in GWAS to

- Increase the number of tested variants
- Fine-mapping becomes more complete
- Meta-analysis using different arrays

0. Imputation setting

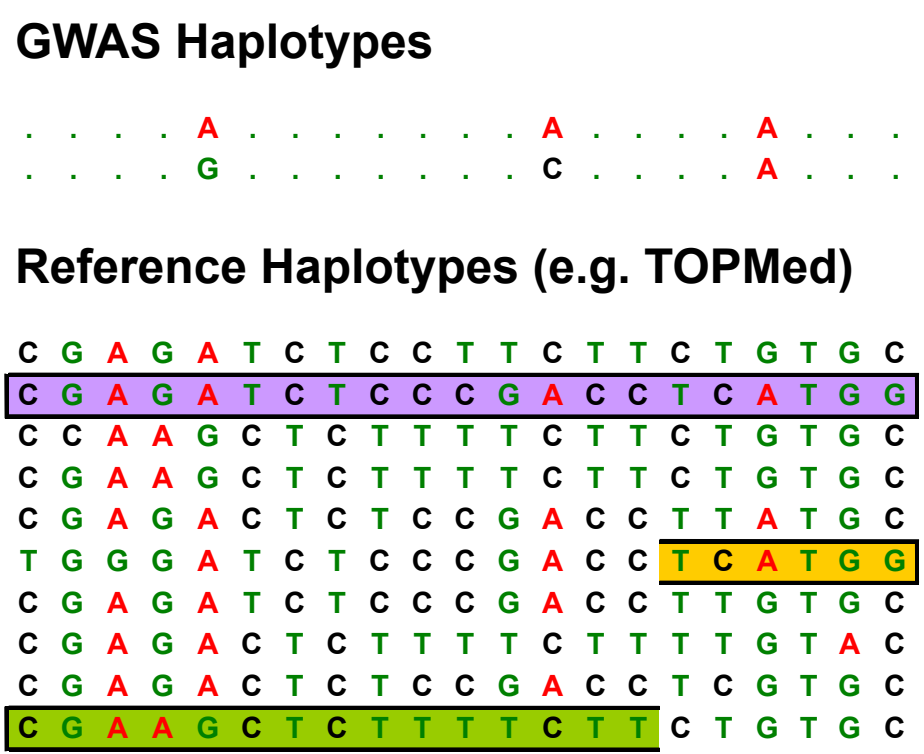
GWAS Haplotypes

```
. . . . A . . . . . . . A . . . . A . . .
. . . . G . . . . . . . C . . . . A . . .
```

Reference Haplotypes (e.g. TOPMed)

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

1. Identify match among reference



2. Impute

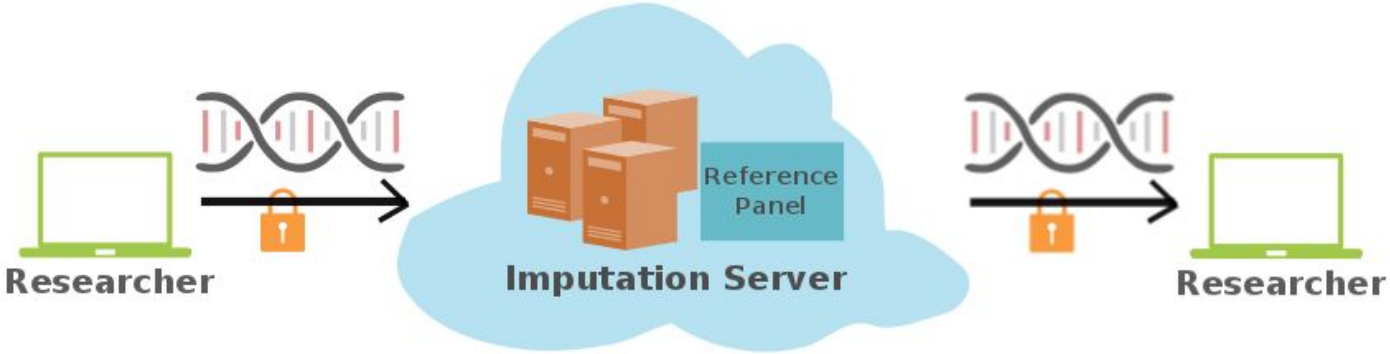
GWAS Haplotypes

```
c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t t t c A t g g
```

Reference Haplotypes (e.g. TOPMed)

```
C G A G A T C T C C T T C T T C T G T G C
C G A G A T C T C C C G A C C T C A T G G
C C A A G C T C T T T T C T T C T G T G C
C G A A G C T C T T T T C T T C T G T G C
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G G
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T T T T C T T T T G T A C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T T C T G T G C
```

ASHG 2014: imputation web service



1.

Upload GWAS data

2.

Server performs

- Quality checks
- Pre-phasing
- Imputation
- Encryption

3.

Download results

MICHIGAN IMPUTATIONSERVER

ASHG ANNUAL MEETING 2023 WASHINGTON, DC • NOVEMBER 1-5

1.) Online-Material

https://imputationserver.readthedocs.io/en/latest/workshops/ASHG2023/

2.) Interactive Polls:

https://pollev.com/ashg

ASHG American Society of Human Genetics

Have you ever used the Michigan Imputation Server

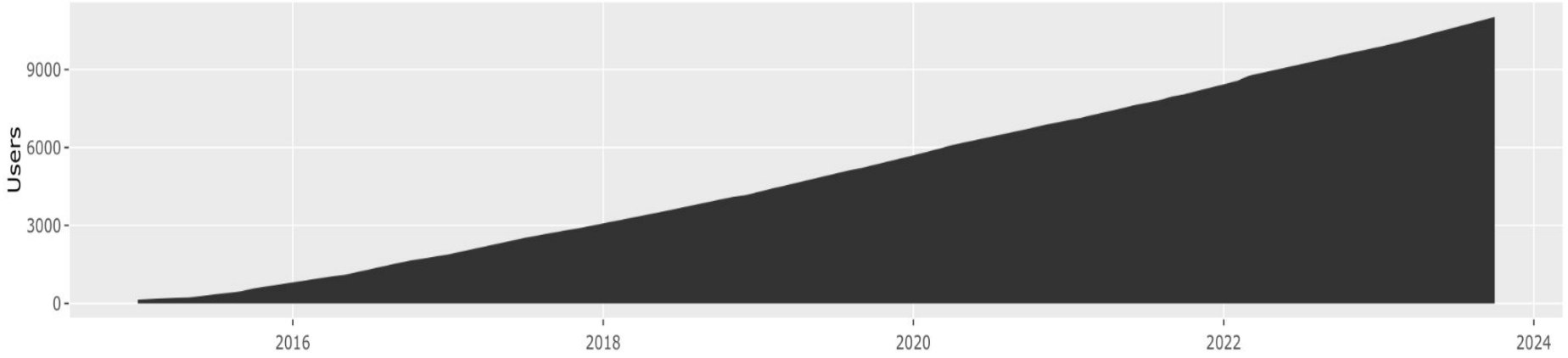
Yes

0%

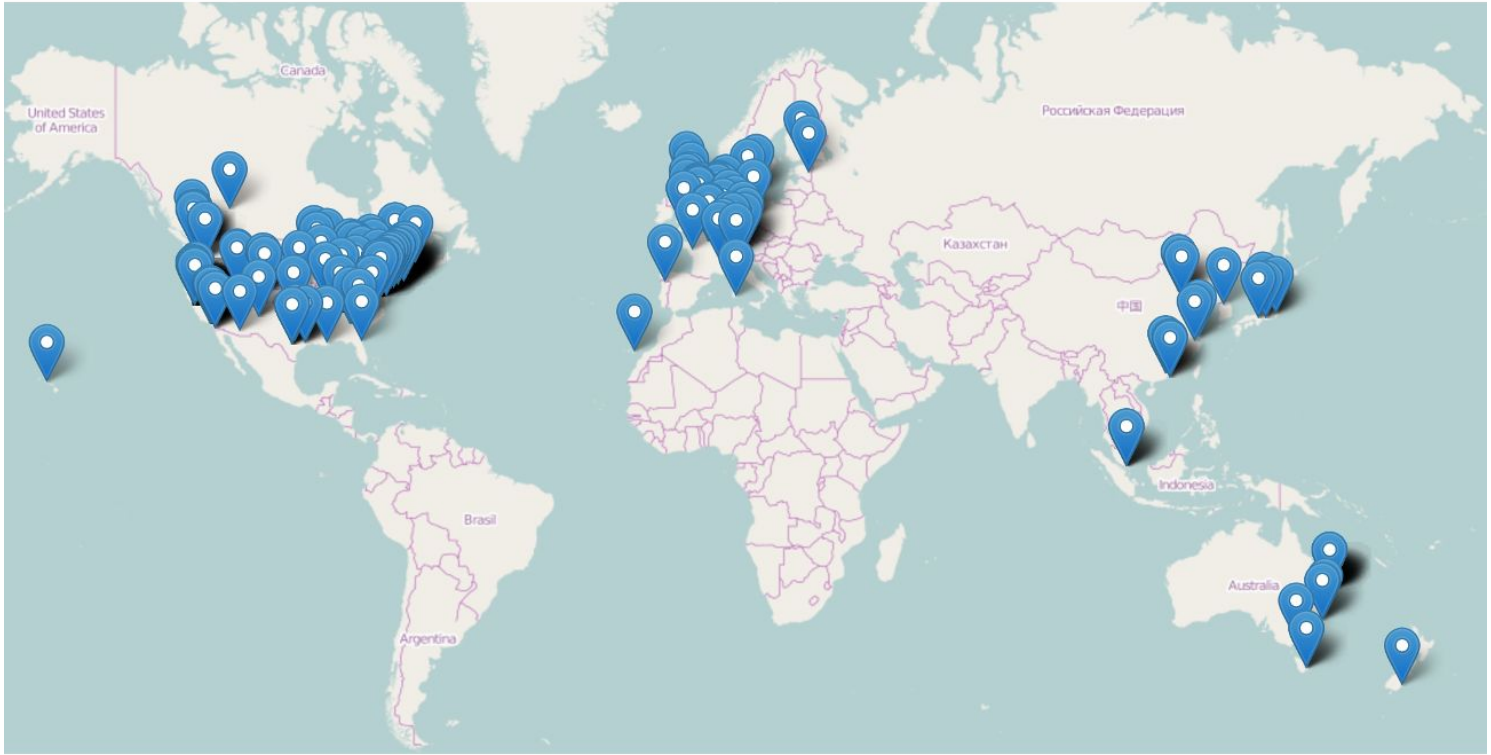
No

0%

>11,000 users

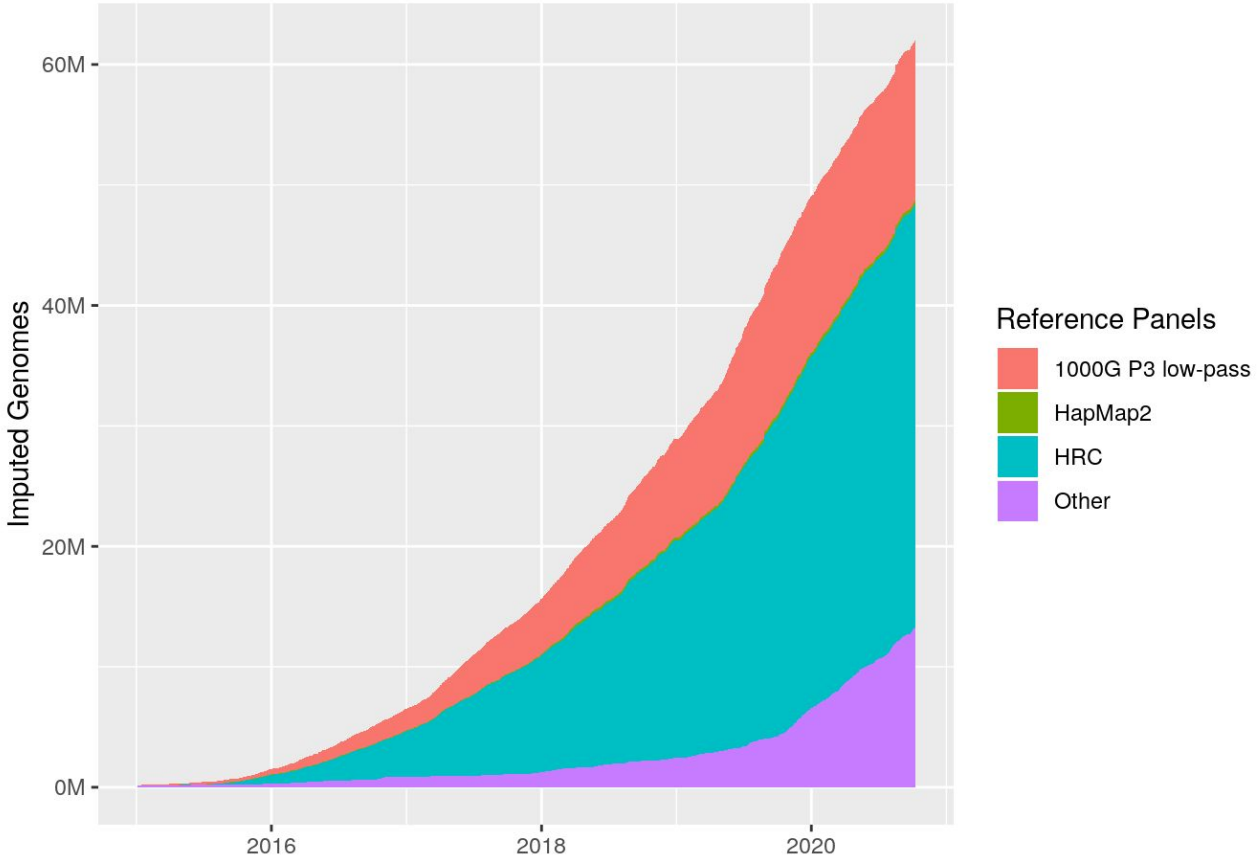


Used by researcher world-wide



ASHG
American Society of Human Genetics

>100M imputed genomes



Summary

Genotype imputation key method in GWAS

Michigan imputation server is easy to use and ensures high quality imputation

Cloud-services will accelerate genetic research so we can devote our time to more interesting tasks

More info and FAQ can be found here:
https://imputationserver.readthedocs.io

Section 2

Run a job, Data Preparation and Data Download

MICHIGAN IMPUTATIONSERVER

Sebastian Schönherr
Medical University of Innsbruck
sebastian.schoenherr@i-med.ac.at
@seppinho

ASHG
American Society of Human Genetics

Learning objectives

Participants will learn

1. How to submit a genotype imputation job
2. How to prepare your input data
3. Different ways to download final datasets

Run your first job on MIS or TMIS

https://imputationserver.sph.umich.edu (MIS)

or

https://imputation.biodatacatalyst.nhlbi.nih.gov (TOPMed Server)


Uploading Data

Recap

- 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
- MIS outputs SNP statistics and a QC Report for each job
  - Helps you to identify problems

Did you run into quality control problems so far?

Yes

0%

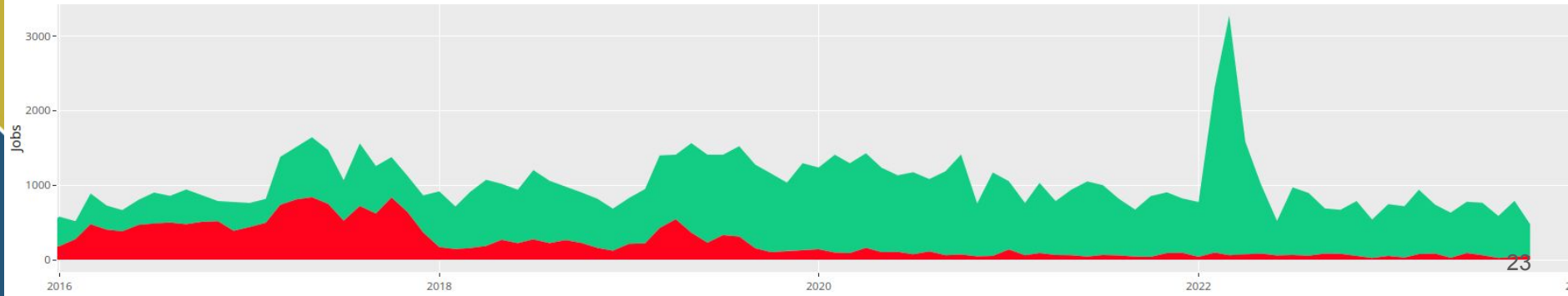
No

0%

How many jobs are failing?

- 40% in 2015; 20% in 2019; 7% 2020-2023
  - Reason for job failures: Something wrong with your input data ***or*** phasing/imputation issue on our side

Total # of jobs: **91,400** (Nov 23)
~1M in October 23



Failing Validation - Obvious Problems

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).

Failing QC - Trickier Problems

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!

Send Notification on Failure

We have sent an email to **sebastian.schoenherr@i-med.ac.at** with the error message.

How to fix input files?

Imputation Preparation Tool

- Developed by W. Rayner
- Works for all major reference panels (HRC, TOPMed, Asia, CAAPA, 1000G)
- Checks for consistency between input data and a reference panel
- Updates/removes SNPs, Updates strand, position and ref/alt assignment
- Input Data in PLINK Binary Format (bim, bed, fam)

https://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-bim-v4.3.0.zip

Execute Imputation Tool before uploading data

	Imputation Server	Preparation Tool
Input	**VCF / chromosome**	PLINK binary data
Output	Imputed VCF / chromosome	**VCFs / chromosome**
File Validation & Statistics		
Basic SNP Filtering		
Lift Over		
Fixes Strand Errors, Updating Ref / Alt Assignment		
Removes SNPs with allele freq difference, A/T & G/C SNPs if MAF > 0.4		

```
seb@seb-genepi:/data3/projects/ashg-imputation-tool$
seb@seb-genepi:/data3/projects/ashg-imputation-tool$ perl HRC-1000G-check-bim.pl -b study-raw-filtered.bim -f study.fr
q -r HRC.r1-1.GRCh37.wgs.mac5.sites.tab.gz -h


         Script to check plink .bim files against HRC/1000G for
        strand, id names, positions, alleles, ref/alt assignment
                    William Rayner 2015-2020
                        wrayner@well.ox.ac.uk

                               Version 4.3


Options Set:
Reference Panel:             HRC
Bim filename:                study-raw-filtered.bim
Reference filename:          HRC.r1-1.GRCh37.wgs.mac5.sites.tab.gz
Allele frequencies filename: study.frq
Plink executable to use:     plink

Chromosome flag set:         No
Allele frequency threshold:  0.2

Path to plink bim file: /data3/projects/ashg-imputation-tool
```

```
seb@seb-genepi:/data3/projects/ashg-imputation-tool$
seb@seb-genepi:/data3/projects/ashg-imputation-tool$ sh Run-plink.sh
PLINK v1.90b3.40 64-bit (16 Aug 2016)      https://www.cog-genomics.org/plink2
(C) 2005-2016 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to /data3/projects/ashg-imputation-tool/TEMP1.log.
Options in effect:
  --bfile /data3/projects/ashg-imputation-tool/study-raw-filtered
  --exclude /data3/projects/ashg-imputation-tool/Exclude-study-raw-filtered-HRC.txt
  --make-bed
  --out /data3/projects/ashg-imputation-tool/TEMP1

32074 MB RAM detected; reserving 16037 MB for main workspace.
1453472 variants loaded from .bim file.
5034 people (3027 males, 2007 females) loaded from .fam.
--exclude: 1392377 variants remaining.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 5034 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.997701.
1392377 variants and 5034 people pass filters and QC.
Note: No phenotypes present.
--make-bed to /data3/projects/ashg-imputation-tool/TEMP1.bed +
/data3/projects/ashg-imputation-tool/TEMP1.bim +
/data3/projects/ashg-imputation-tool/TEMP1.fam ... 30%
```

```
seb@seb-genepi:/data3/projects/ashg-imputation-tool$
seb@seb-genepi:/data3/projects/ashg-imputation-tool$ bgzip study-raw-filtered-updated-chr15.vcf
```

Error: No chunks passed the QC step. Imputation cannot be started!

Email MIS team
0%

Data are on the wrong build
0%

Too few variants overlap with the reference panel
0%

Must be a MIS problem, since imputation runs locally
0%

Don't know
0%

{"success":false,"message":"number of max downloads exceeded."}

Re-impute data since the job is already retired

0%

There is an internet problem - should try again later

0%

Email MIS team to increase download counter

0%

Don't know

0%

Uploaded my data 5 mins ago and my job is still waiting...

There is a problem with MIS - email MIS team to let them know

0%

There is a problem with my data

0%

MIS is very busy - check again later

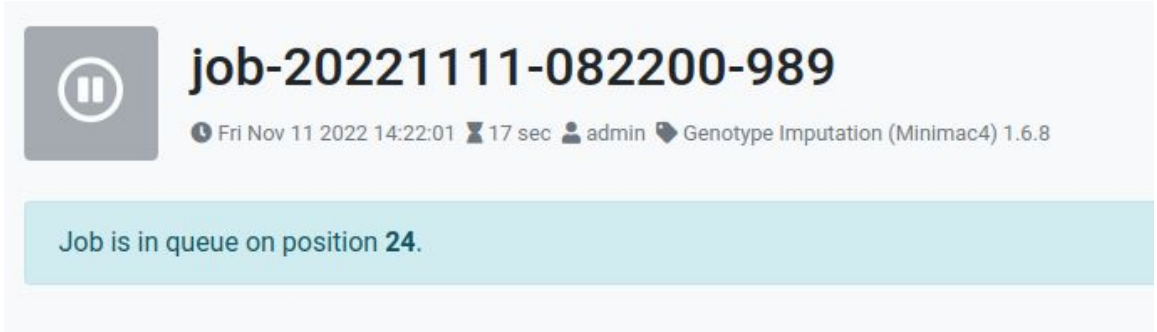
0%

Don't know

0%

Life Cycle of an Imputation Job

- Job passed Quality Control
- Job scheduled in imputation queue



- Waits until resources are available

Life Cycle of an Imputation Job

- Phasing and Imputation starts

Life Cycle of an Imputation Job

- Data is encrypted
- Email with one time password is sent to user

```
Dear Lukas,
the password for the imputation results is: ppO9Z0KeQvQMc

The results can be downloaded from https://imputationserver.sph.umich.edu/start.html#!jobs
/job-20190919-112230-581/results
```

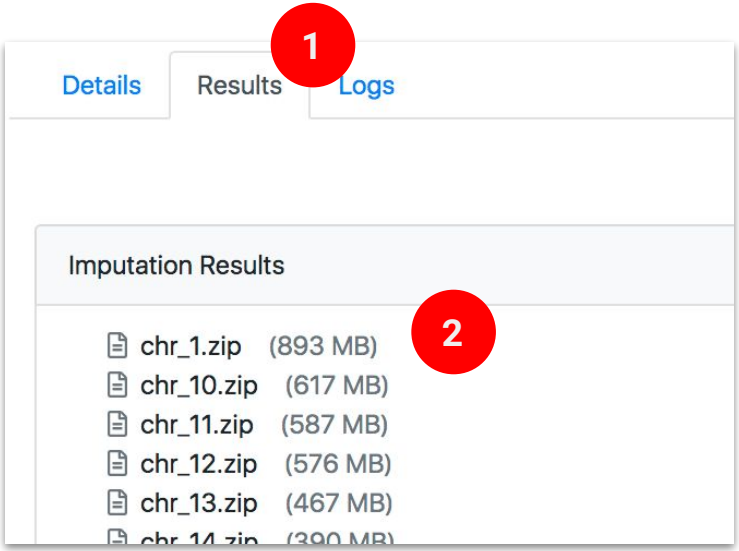
Life Cycle of an Imputation Job

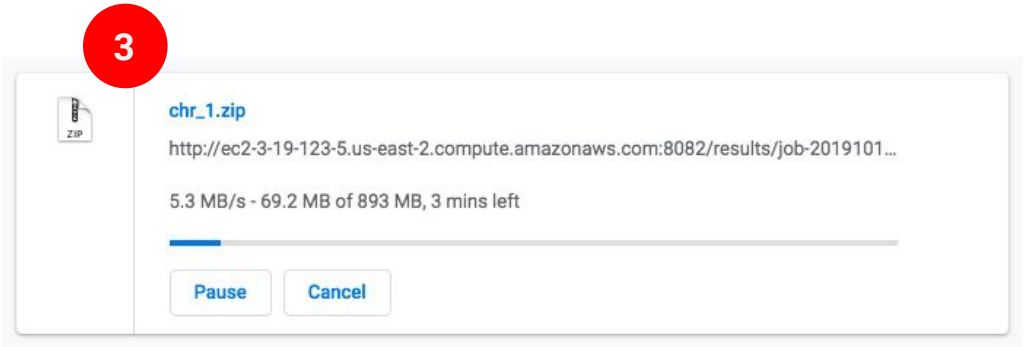
- After 7 days the job is retired
- All results are deleted
- We will send you an email 2 days before

```
Dear Lukas Forer,
Your job retires in 2 days! All imputation results will be deleted at that time.

Please ensure that you have downloaded all results from
https://imputationserver.sph.umich.edu/start.html#!jobs/job-20191011-124306-370
```

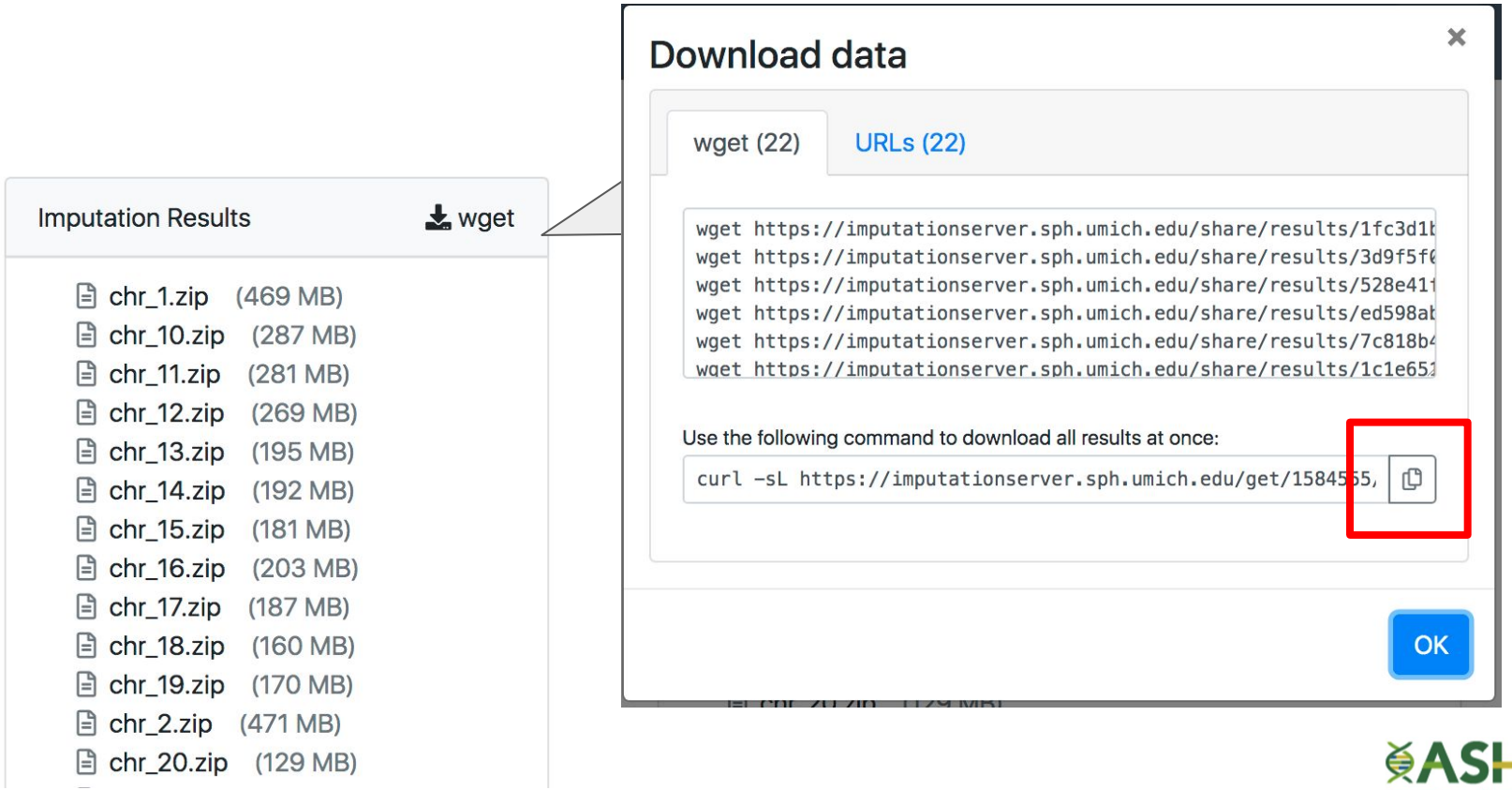
How to download the imputed genotypes?

Option 1: Web-Interface





Option 2: Batch Download



```
fantasia:~>
```

```
fantasia:~> curl -sL https://imputationserver.sph.umich.edu/get/1584555/675233d69
db57b793589b916f2a81cb8 | bash
```

```
fantasia:~> curl -sL https://imputationserver.sph.umich.edu/get/1584555/675233d69
db57b793589b916f2a81cb8 | bash

Downloading file chr_1.zip (1/22)...
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   185  100   185    0     0  21087      0 --:--:-- --:--:-- --:--:-- 30833
100  469M  100  469M    0     0   116M      0  0:00:04  0:00:04 --:--:--  167M

Downloading file chr_10.zip (2/22)...
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   185  100   185    0     0  23886      0 --:--:-- --:--:-- --:--:-- 37000
100  287M  100  287M    0     0  87.4M      0  0:00:03  0:00:03 --:--:--  138M

Downloading file chr_11.zip (3/22)...
```

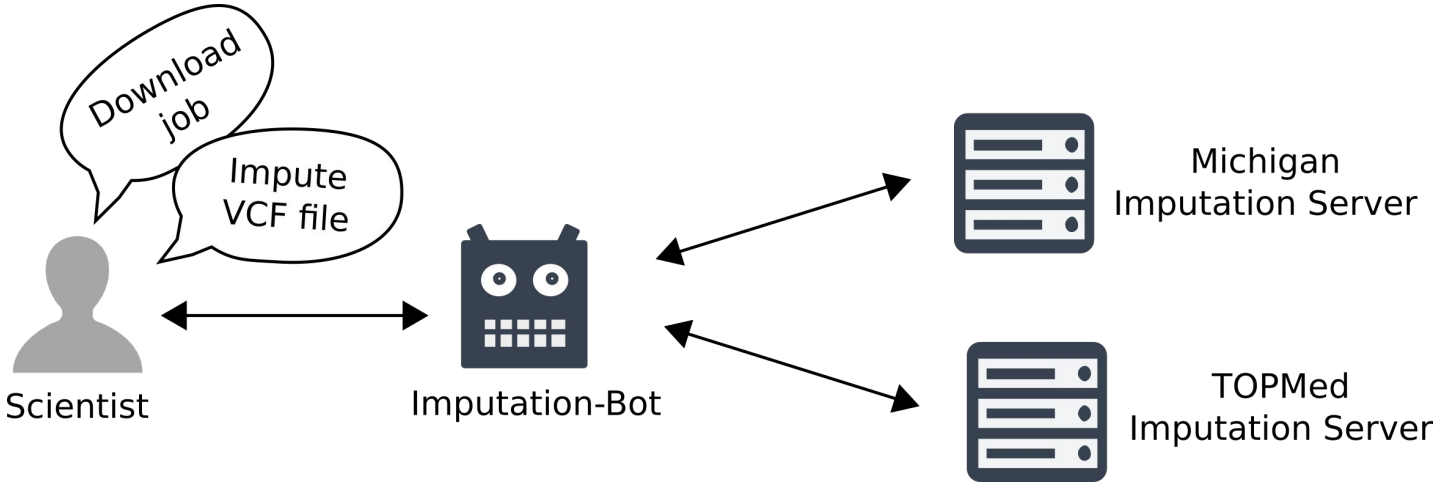
```
Downloading file chr_7.zip (20/22)...
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   185  100   185    0     0  24189      0 --:--:-- --:--:-- --:--:-- 37000
100  337M  100  337M    0     0   101M      0  0:00:03  0:00:03 --:--:--  160M

Downloading file chr_8.zip (21/22)...
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   185  100   185    0     0  24529      0 --:--:-- --:--:-- --:--:-- 37000
100  306M  100  306M    0     0  94.9M      0  0:00:03  0:00:03 --:--:--  152M

Downloading file chr_9.zip (22/22)...
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   185  100   185    0     0  22486      0 --:--:-- --:--:-- --:--:-- 37000
100  245M  100  245M    0     0  82.0M      0  0:00:02  0:00:02 --:--:-- 84.8M

All 22 file(s) downloaded.


fantasia:~>
```

Option 3: Use Imputation Bot

- Run everything on the command line

Imputation Bot

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



Have you ever used the MIS Application Program Interface (API)?

Yes

0%

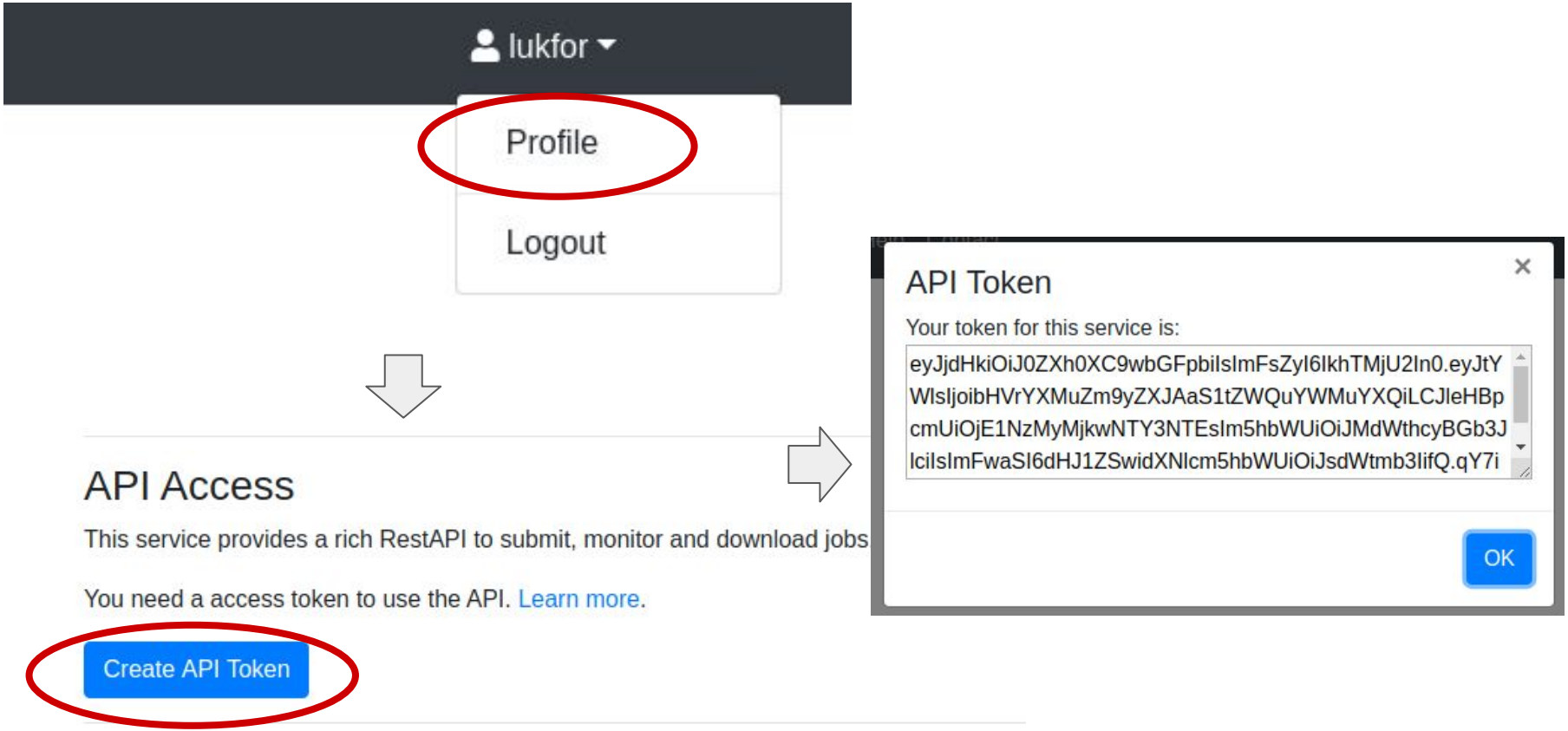
No

0%

What is an API?

0%

1. Enable API Access


lukfor
Profile
Logout
API Access
This service provides a rich RestAPI to submit, monitor and download jobs
You need a access token to use the API. Learn more.
Create API Token
API Token
Your token for this service is:
eyJjdHkiOiJ0ZXh0XC9wbGFpbiIsImFsZyI6IkhTMjU2In0.eyJtY
WlsIjoibHVrYXMuZm9yZXJAaS1tZWQuYWMuYXQiLCJleHBp
cmUiOjE1NzMyMjkwNTY3NTEsIm5hbWUiOiJMdWthcyBGb3J
lciIsImFwaSI6dHJ1ZSwidXNlcm5hbWUiOiJsdWtmb3lifQ.qY7i
OK

2. Install Imputation Bot

```
lukas@lukas-workstation:~/imputationbot$
```

```
lukas@lukas-workstation:~/imputationbot$ curl -sL imputationbot.now.sh | bash
```

```
lukas@lukas-workstation:~/imputationbot$ curl -sL imputationbot.now.sh | bash
Installing Imputation Bot v0.8.3...
Downloading Imputation Bot from https://github.com/lukfor/imputationbot/releases
/download/v0.8.3/imputationbot-installer.sh...
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   652  100   652    0     0   1895      0 --:--:-- --:--:-- --:--:--  1895
100 2995k  100 2995k    0     0   390k      0  0:00:07  0:00:07 --:--:--  518k
Verifying archive integrity...  100%   All good.
Uncompressing Make
script=self-extrabable
scriptargs=archives makeself.sh  100%

Imputation Bot v0.8.3 installation completed. Have fun!

lukas@lukas-workstation:~/imputationbot$
```

3. Configure Imputation Bot

```
lukas@lukas-workstation:~/imputationbot$ ./imputationbot add-instance
```

```
lukas@lukas-workstation:~/imputationbot$ ./imputationbot add-instance

Imputation Bot 0.8.3
https://imputationserver.sph.umich.edu
(c) 2019-2020 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2020-10-08T12:27:01Z

Imputationserver Url [https://imputationserver.sph.umich.edu]:
```

```
lukas@lukas-workstation:~/imputationbot$ ./imputationbot add-instance

Imputation Bot 0.8.3
https://imputationserver.sph.umich.edu
(c) 2019-2020 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2020-10-08T12:27:01Z

Imputationserver Url [https://imputationserver.sph.umich.edu]:
API Token [None]:
```

```
lukas@lukas-workstation:~/imputationbot$ ./imputationbot add-instance

Imputation Bot 0.8.3
https://imputationserver.sph.umich.edu
(c) 2019-2020 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2020-10-08T12:27:01Z

Imputationserver Url [https://imputationserver.sph.umich.edu]:
API Token [None]: eyJjdHkiOiJ0ZXh0XC9wbGFpbiIsImFsZyI6IkhTMjU2In0.eyJtYWlsIjoibH
VrYXMuZm9yZXJAaS1tZWQuYWMuYXQiLCJleHBpcmUiOjE2MDQxMzkzMzAwNzYsIm5hbWUiOiJMdWthcy
IsImFwaSI6dHJ1ZSwidXNlcm5hbWUiOiJsdWtmb3IifQ.1nOYSJr58WHXn3piOrv8Th-FZHJVxxUb-7K
VfUfadtA
```

```
lukas@lukas-workstation:~/imputationbot$ ./imputationbot add-instance

Imputation Bot 0.8.3
https://imputationserver.sph.umich.edu
(c) 2019-2020 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2020-10-08T12:27:01Z

Imputationserver Url [https://imputationserver.sph.umich.edu]:
API Token [None]: eyJjdHkiOiJ0ZXh0XC9wbGFpbiIsImFsZyI6IkhTMjU2In0.eyJtYWlsIjoibH
VrYXMuZm9yZXJAaS1tZWQuYWMuYXQiLCJleHBpcmUiOjE2MDQxMzkzMzAwNzYsIm5hbWUiOiJMdWthcy
IsImFwaSI6dHJ1ZSwidXNlcm5hbWUiOiJsdWtmb3IifQ.1nOYSJr58WHXn3piOrv8Th-FZHJVxxUb-7K
VfUfadtA

Hi Lukas




Imputation Bot is ready to submit jobs to https://imputationserver.sph.umich.edu
 (Genotype Imputation (Minimac4)) 1.4.1)

lukas@lukas-workstation:~/imputationbot$
```

4. Submit a Job

```
lukas@lukas-workstation:~/imputationbot$ ./imputationbot impute --files test-dat
a/chr20.R50.merged.1.330k.recode.vcf.gz --refpanel 1000g-phase-3-v5 --population
 eur
```

```
lukas@lukas-workstation:~/imputationbot$ ./imputationbot impute --files test-dat
a/chr20.R50.merged.1.330k.recode.vcf.gz --refpanel 1000g-phase-3-v5 --population
 eur

Imputation Bot 0.8.3
https://imputationserver.sph.umich.edu
(c) 2019-2020 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2020-10-08T12:27:01Z

Submitting job for 1000g-phase-3-v5 to Michigan Imputation Server...
  Parameters:
    refpanel: apps@1000g-phase-3-v5
    files:
      - test-data/chr20.R50.merged.1.330k.recode.vcf.gz
    build: hg19
    r2Filter: 0
    phasing: eagle
    population: eur
    aesEncryption: no

 Uploading files [100%]
```

```
Imputation Bot 0.8.3
https://imputationserver.sph.umich.edu
(c) 2019-2020 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2020-10-08T12:27:01Z

Submitting job for 1000g-phase-3-v5 to Michigan Imputation Server...
  Parameters:
    refpanel: apps@1000g-phase-3-v5
    files:
      - test-data/chr20.R50.merged.1.330k.recode.vcf.gz
    build: hg19
    r2Filter: 0
    phasing: eagle
    population: eur
    aesEncryption: no

  Uploading files [100%]

  Imputation job 'job-20201008-084828-935' submitted successfully
   Check the job progress on https://imputationserver.sph.umich.edu/index.html
#!jobs/job-20201008-084828-935
```

5. Download Data

```
bash-3.2$ ./imputationbot download job-20191012-152533-884 --password mkW5oPAB-7c
```

```
bash-3.2$ ./imputationbot download job-20191012-152533-884 --password mkW5oPAB-7c

Imputation Bot 0.8.3
https://imputationserver.sph.umich.edu
(c) 2019-2020 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2020-10-08T12:27:01Z

Downloading job job-20191012-152533-884...
  Downloading file job-20191012-152533-884/qcreport/qcreport.html (1/5)
```

```
bash-3.2$ ./imputationbot download job-20191012-152533-884 --password mkW5oPAB-7c

Imputation Bot 0.8.3
https://imputationserver.sph.umich.edu
(c) 2019-2020 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2020-10-08T12:27:01Z

Downloading job job-20191012-152533-884...
  Downloading file job-20191012-152533-884/qcreport/qcreport.html (1/5)
  Downloading file job-20191012-152533-884/statisticDir/snps-excluded.txt (2/5)
  Downloading file job-20191012-152533-884/statisticDir/typed-only.txt (3/5)
  Downloading file job-20191012-152533-884/local/chr_20.zip (4/5)
  Decrypting file job-20191012-152533-884/local/chr_20.zip...
```

```
bash-3.2$ ./imputationbot download job-20191012-152533-884 --password mkW5oPAB-7c

Imputation Bot 0.8.3
https://imputationserver.sph.umich.edu
(c) 2019-2020 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2020-10-08T12:27:01Z

Downloading job job-20191012-152533-884...
  Downloading file job-20191012-152533-884/qcreport/qcreport.html (1/5)
  Downloading file job-20191012-152533-884/statisticDir/snps-excluded.txt (2/5)
  Downloading file job-20191012-152533-884/statisticDir/typed-only.txt (3/5)
  Downloading file job-20191012-152533-884/local/chr_20.zip (4/5)
  Decrypting file job-20191012-152533-884/local/chr_20.zip...
  Downloading file job-20191012-152533-884/logfile/chr_20.log (5/5)

All data downloaded and stored in /Users/lukas/imputationbot/job-20191012-152533-

bash-3.2$
```

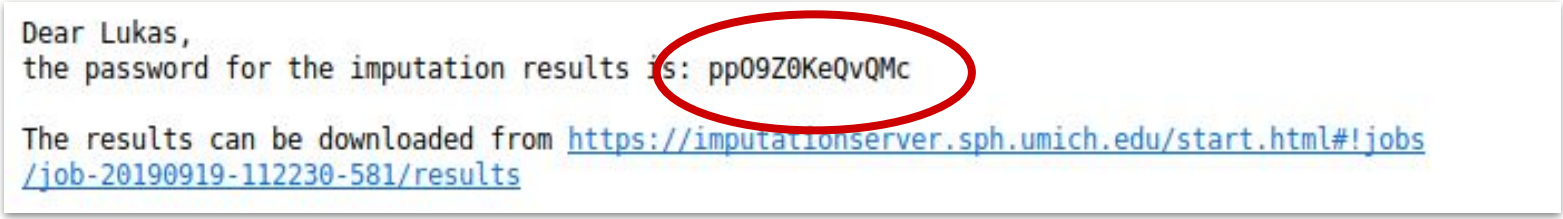
Summary

- Imputation bot automates interactions with Imputation Servers
- Simplifies multi-reference and multi-study imputation
- Available: https://github.com/lukfor/imputationbot

More info and FAQ can be found here:
https://imputationserver.readthedocs.io

Data Decryption

- All imputed genotypes are in **encrypted zip files** (e.g. chr_1.zip)
- We send you an email with a password



- You need this password to **decrypt** your genotypes
- Decryption with standard zip programs (e.g. WinZip, 7zip or gunzip)
- AES Encryption: Needs additional software to decrypt (e.g. 7z)

What is in each zip file?

chr_20.zip
├── chr20.dose.vcf.gz
├── chr20.empiricalDose.vcf.gz
└── chr20.info.gz

What is in each zip file?

chr_20.zip
- **chr20.dose.vcf.gz**
- chr20.info.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

What is in each zip file?

chr_20.zip
- **chr20.dose.vcf.gz**
- chr20.info.gz

#CHROM	POS	ID	REF	ALT
20	61795	20:61795:G:T	G	T
20	63231	20:63231:T:G	T	G
20	63244	20:63244:A:C	A	C

…

…

…

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

What is in each zip file?

chr_20.zip
- chr20.dose.vcf.gz
- **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	...

https://genome.sph.umich.edu/wiki/Minimac3_Info_File

What is in each zip file?

md5 checksum file

```
(base) seb@seb-laptop:~/ashg22$ cat results.md5
3ea13c00d323117e0b4648a683175d39 chr_11.zip
9ecb19e40d3f8a55f128c640333ab2ef chr_22.zip
161918ed598f32bcd88536399695b398 chr_12.zip
2709ee09f353c0b332686fdf40e9d062 chr_13.zip
```

chr_20.zip
- chr20.dose.vcf.gz
- **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	...

https://genome.sph.umich.edu/wiki/Minimac3_Info_File

Summary

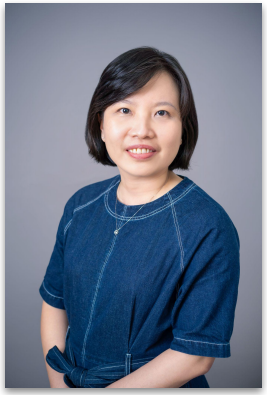
- MIS Web Interface provides a fast and reliable way to impute data
- MIS applies a strict Quality Control with the goal to return high quality imputation data
- Pre-Imputation tools available for data preparation
- Different options to download data

More info and FAQ can be found here:
https://imputationserver.readthedocs.io

ASHG
American Society of Human Genetics

Section 3

Performing GWAS using imputed data



Xueling Sim
National University of Singapore
ephsx@nus.edu.sg

MICHIGAN IMPUTATIONSERVER

Learning objectives

Participants will learn to:

- Identify and understand the use of variant imputation quality information following imputation in the MIS
- Distinguish between some of the available options for GWAS
- Troubleshoot common GWAS errors

Have you ever performed a GWAS?

Yes

0%

No

0%

I'm trying

0%

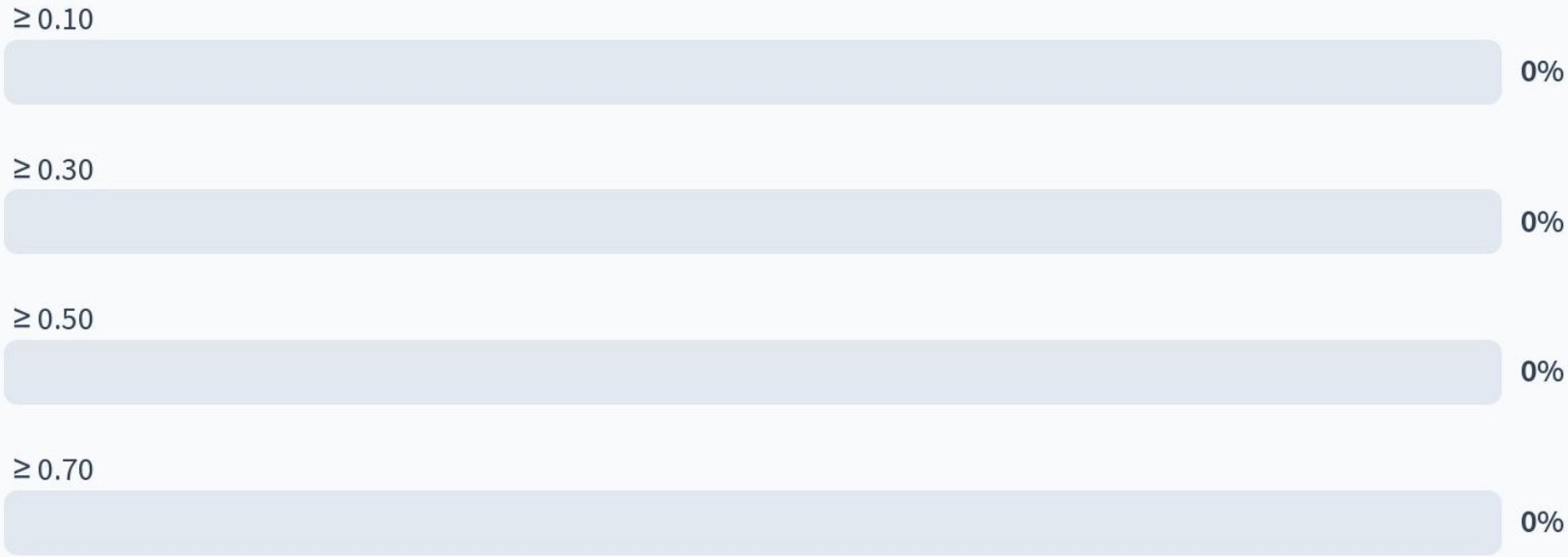
Imputation Quality

- For each variant, how confident can we be that the imputation dosages are sufficiently “accurate” for association analyses?
- Measure of confidence in imputed dosages: “Rsq” column [range 0-1]

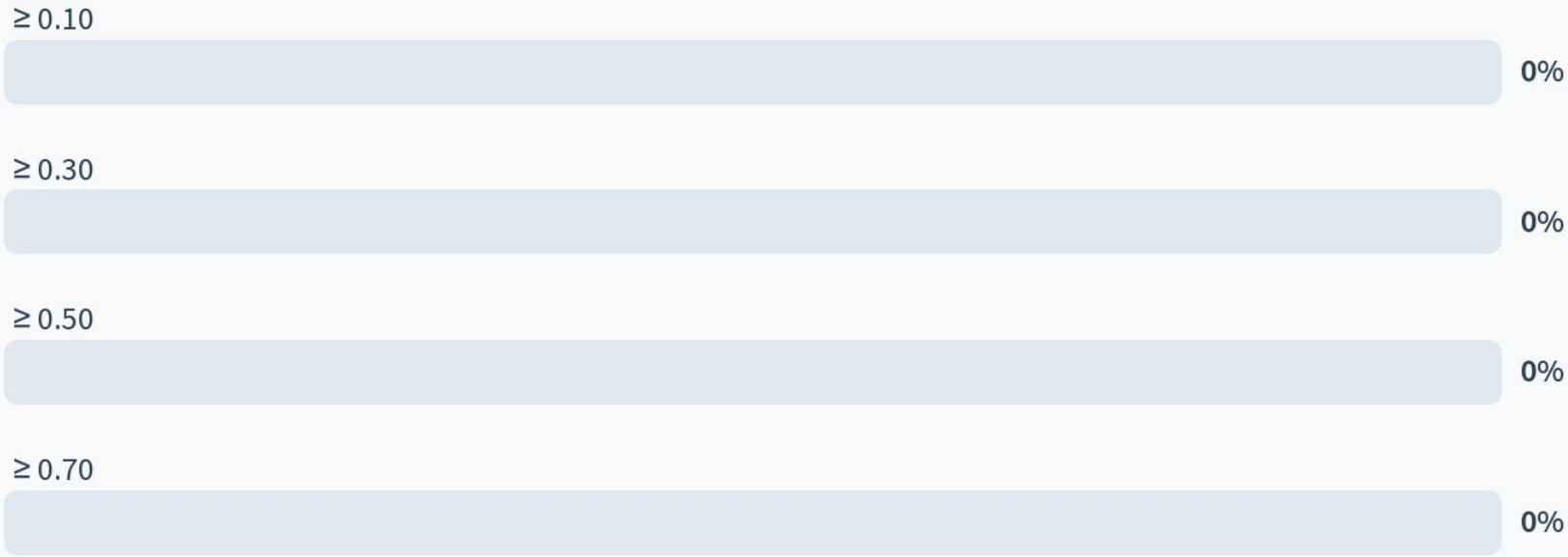
SNP	REF(0)	ALT(1)	ALT_Frq	MAF	AvgCall	Rsq	Genotyped	...
20:61795:G:T	G	T	0.26318	0.26318	0.88455	0.54658	Imputed	...
20:63231:T:G	T	G	0.03843	0.03843	0.98342	0.67736	Imputed	...
20:63244:A:C	A	C	0.16132	0.16132	0.91761	0.49907	Imputed	...

From a chr20.info.gz file

Minimally accepted Rsq value for common (MAF≥5%) variants?



Minimally accepted Rsq value for low frequency (MAF<5%) and rare variants?



Imputation Quality

- Minimal Rsq value for common variants
  - ≥ 0.30
- Minimal Rsq value for low frequency/rare variants
  - ≥0.50
- Before performing GWAS, remove variants that do not meet these thresholds
  - Suggested program: VCFtools
  - Saves computational time when performing GWAS

Which GWAS program(s) have you used?

Performing the GWAS

- Each program has its own input, output formats, and options
- Typical input files
  - Genotype file (.vcf; .bgen; .bed/.bim/.fam)
  - Phenotype/covariate file (.txt; .ped)
  - Some programs use separate phenotype and covariate files
  - Kinship/relationship matrix (EPACTS, SAIGE)

Available GWAS Programs

No File Reformatting (VCF from MIS)

- EPACTS
- Rvtests
- SNPTEST
- SAIGE

File Formatting Required

- BOLT-LMM
- BGENIE
- regenie
- PLINK

Each GWAS Program Has Strengths, Limitations

EPACTS/Rvtests

+ Many model options - single variant, gene-based
+ Chr X analyses
+ Phenotypic transformation (e.g inverse normal; Rvtests only)
+ Linear mixed model for sample relatedness (quantitative traits only)
+ Generate covariance matrices for downstream analyses (e.g conditional analyses; Rvtests only)

- Memory intensive
- Sample size ~≤20,000 (better ≤10,000)

EPACTS: https://genome.sph.umich.edu/wiki/EPACTS
Rvtests: https://genome.sph.umich.edu/wiki/Rvtests

Each GWAS Program Has Strengths, Limitations

SAIGE

+ Similar to Rvtests, but for very large sample sizes (e.g. biobanks)
+ Able to account for sample relatedness for binary traits
+ Designed to handle unbalanced number of cases and controls
+ Chr X analyses

- Should not be used to examine heritability (biased variance estimates)
- Computational time can vary widely between phenotypes and sample sizes
- Can be conservative for extremely unbalanced case and control ratio
- Odds ratios estimated to conserve computational time

SAIGE: https://github.com/weizhouUMICH/SAIGE

Each GWAS Program Has Strengths, Limitations

SNPTEST

+ Frequentist and bayesian methods supported
+ Chr X analyses

- Limited to unrelated individuals
- Computational intensive

SNPTEST: https://www.well.ox.ac.uk/~gav/snptest/#introduction

Each GWAS Program Has Strengths, Limitations

BOLT-LMM/BGENIE/Regenie

+ Great for very large sample sizes (e.g. biobanks)
+ Chr X analyses
+ Computationally efficient (Regenie)

- Requires files to be in BGEN or PLINK format
- Nextflow pipeline for regenie using VCF: https://github.com/genepi/nf-gwas
- Not optimal for extremely unbalanced case control ratio (especially with rare variants)

BOLT-LMM: https://data.broadinstitute.org/alkesgroup/BOLT-LMM/#x1-5600011
BGENIE: https://jmarchini.org/bgenie/
Regenie: https://github.com/rgcgithub/regenie

Each GWAS Program Has Strengths, Limitations

PLINK

+ Quick
+ Multiple versions; often as intermediary tool to the other programs
+ Can run on the command line (unix not required)
+ Chr X analyses

- Requires files to be in PLINK format (.bed/.bim/.fam)
- Limited model options

PLINK: https://www.cog-genomics.org/plink/2.0/

Summary of common GWAS analysis tools

	EPACTS	Rvtests	SNPTEST	SAIGE	BLOT-LMM	Bgenie	Regenie
Input VCF	Y	Y	Y				
Sample relatedness (Quantitative outcome)	Y	Y		Y	Y	Y	Y
Sample relatedness (Binary outcome)				Y		Y	Y
Case control imbalance				Y			Y
Large sample size (>20,000)				Y	Y	Y	Y

Which program(s) would be best?

A researcher new to genetic analyses and unfamiliar to the UNIX environment wants to perform a GWAS on total cholesterol using a cohort of 5,500 unrelated individuals.

EPACTS/Rvtests 0%

SAIGE 0%

BOLT-LMM 0%

PLINK 0%

Which program(s) would be best?
Researchers want to perform a GWAS using a cohort of 10,000 individuals with household based recruitment (i.e. includes related individuals).

EPACTS/Rvtests 0%

SAIGE 0%

BOLT-LMM 0%

PLINK 0%

Which program(s) would be best?
Researchers want to perform a GWAS using data from BioBank Japan (>200,000 individuals)

(A) EPACTS/Rvtests 0%

(B) SAIGE 0%

(C) BOLT-LMM 0%

(D) PLINK 0%

(E) Regenie/Bgenie 0%

Common Errors When Running a GWAS

- Wording of error messages vary by program, but the same issues will cause errors throughout all of the program
- [Unix] Errors independent of GWAS program
  - File permissions
    - Correct by changing file permissions
  - Directory/file not found
    - Correct by making sure all of the file locations and names are accurate
  - Not enough memory/time
    - Correct by restarting job with adequate memory/time allocation

Common Errors When Running a GWAS

- Common errors
  - IDs don’t match
    - Correct by ensuring that the ID in the phenotype, genotype, covariance, kinship matrix are consistent format in all files

Common Errors When Running a GWAS

- Common errors
  - IDs don’t match
    - Correct by ensuring that the ID in the phenotype, genotype, covariance, kinship matrix are consistent format in all files
  - File format(s) incorrect
    - Correct by making sure the format of all files are as the program is expecting (e.g. columns, delimiters, headers, file extension

Common Errors When Running a GWAS

- Common errors
  - IDs don't match
    - Correct by ensuring that the ID in the phenotype, genotype, covariance, kinship matrix are consistent format in all files
  - File format(s) incorrect
    - Correct by making sure the format of all files are as the program is expecting (e.g. columns, delimiters, headers, file extension
  - Improperly specified options/command
    - Correct by checking all needed options are specified, correct order, no typos

Common Errors When Running a GWAS

- Common errors
  - IDs don't match
    - Correct by ensuring that the ID in the phenotype, genotype, covariance, kinship matrix are consistent format in all files
  - File format(s) incorrect
    - Correct by making sure the format of all files are as the program is expecting (e.g. columns, delimiters, headers, file extension
  - Improperly specified options/command
    - Correct by checking all needed options are specified, correct order, no typos
  - Peripheral programs not available (e.g. R with EPACTS, SAIGE)
    - Correct by installing other peripheral programs

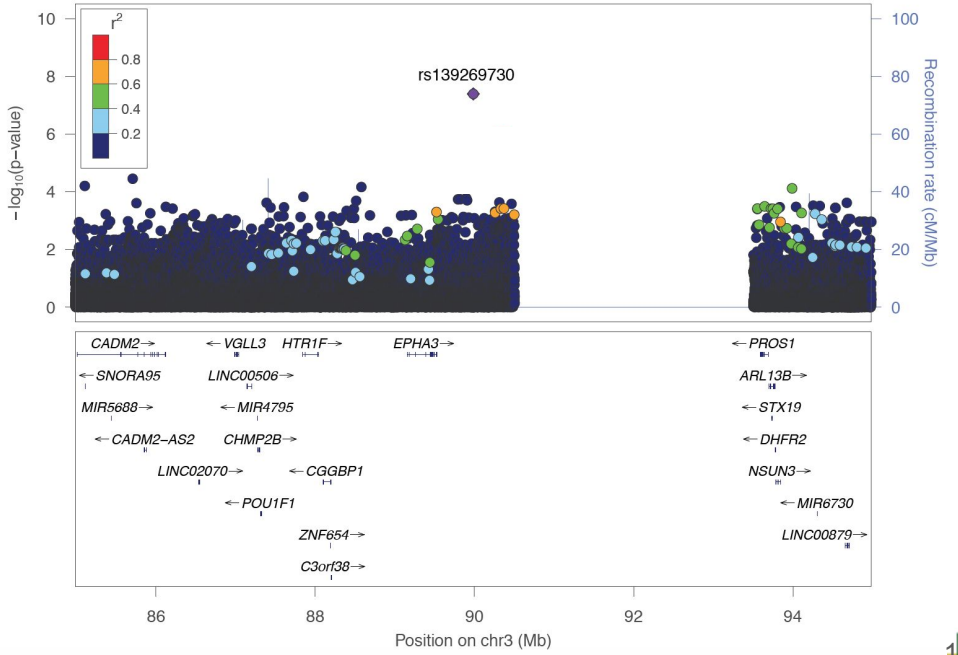
Common Errors When Running a GWAS

- Common errors
  - IDs don't match
    - Correct by ensuring that the ID in the phenotype, genotype, covariance, kinship matrix are consistent format in all files
  - File format(s) incorrect
    - Correct by making sure the format of all files are as the program is expecting (e.g. columns, delimiters, headers, file extension
  - Improperly specified options/command
    - Correct by checking all needed options are specified, correct order, no typos
  - Peripheral programs not available (e.g. R with EPACTS, SAIGE)
    - Correct by installing other peripheral programs
  - Invalid estimate (e.g. heritability in BOLT-LMM)
    - Sample too related and/or sample size too small
    - Correct by using a different program

Interpreting GWAS Results

- GWAS results must be carefully reviewed for:
  - Imputation quality!
  - Genomic inflation
  - False positives

- Replication datasets
- PheWas



Summary

- Variants must be filtered post-imputation to remove those with poor imputation quality
- There are many GWAS programs available, each with their own strengths and limitations - so be sure to pick one that fits your analyses needs
- As these GWAS programs are widely used or adopted by consortia, there are tutorials and help-pages available

More info and FAQ can be found here:
https://imputationserver.readthedocs.io

Sex chromosomes

- Humans have two sex chromosomes, X and Y, that in combination determine the sex of an individual.
  - To ensure balanced expression, one of the female's X-chromosome is randomly selected to undergo inactivation(either paternal or maternal determined at early embryonic development).
  - Pseudoautosomal regions (PAR1 and PAR2) are short regions on the X- and Y-chromosomes that are inherited in an autosomal rather than a sex-linked manner

X-chromosome - genotype calls and QC

- Genotype calls on X-chromosomes
  - One copy in males, two copies in females
- Clarity of genotype calls on X-Chromosome
  - 0/1/2 allele coding in females; 0/1 or 0/2 allele coding in males
  - 0/0.5/1 allele coding in females; 0/0.5 allele coding in males
  - Comparing standard error with autosomal data
- Analysis plan to be specific on allelic coding
- Develop QC checks before association analyses, e.g., similar allele frequencies between males and females (differential calls bias)

X-chromosome coding on DRAGEN

Ploidy Support

The small variant caller currently only supports either ploidy 1 or 2 on all contigs within the reference except for the mitochondrial contig, which uses a continuous allele frequency approach (see *Mitochondrial Calling*). The selection of ploidy 1 or 2 for all other contigs is determined as follows.

- If *--sample-sex* is not specified on the command line, the Ploidy Estimator determines the sex. If the Ploidy Estimator cannot determine the sex karyotype or detects sex chromosome aneuploidy, all contigs are processed with ploidy 2.
- If *--sample-sex* is specified on the command line, contigs are processed as follows.
  - For female samples, DRAGEN processes all contigs with ploidy 2 and marks variant calls on chrY with a filter PloidyConflict.
  - For male samples, DRAGEN processes all contigs with ploidy 2, except for the sex chromosomes. DRAGEN processes chrX with ploidy 1, except in the PAR regions, where it is processed with ploidy 2. chrY is processed with ploidy 1 throughout.

DRAGEN detects sex chromosomes by the naming convention, either X/Y or chrX/chrY. No other naming convention is supported.

https://support-docs.illumina.com/SW/DRAGEN_v39/Content/SW/DRAGEN/PloidySupport_fDg.htm

X-chromosome - imputation

Chromosome X Pipeline

Additionally to the standard QC, the following per-sample checks are executed for chrX:

- Ploidy Check: Verifies if all variants in the nonPAR region are either haploid or diploid.
- Mixed Genotypes Check: Verifies if the amount of mixed genotypes (e.g. 1/.) is < 10 %.

For phasing and imputation, chrX is split into three independent chunks (PAR1, nonPAR, PAR2). These splits are then automatically merged by Michigan Imputation Server and are returned as one complete chromosome X file. Only Eagle is supported.

https://imputationserver.readthedocs.io/en/latest/pipeline/#chromosome-x-pipeline

Interpretation of association

- Biological considerations:
  - Assumption of total X-inactivation but there is evidence of escape and compensation mechanisms
  - Different inheritance/activation in different tissues?
  - Sex-biased gene expression? Unclear if it is real differences in gene expression or influence of hormones.
- Methodology development
  - Statistical methods to estimate the inactivation when estimating the effects; adjusting for sex

Section 4

nf-gwas, Imputation Bot and PGS Server

Lukas Forer
Medical University of Innsbruck
lukas.forer@i-med.ac.at
@lukfor

MICHIGAN IMPUTATIONSERVER

ASHG American Society of Human Genetics

Learning objectives

Participants will

1. Learn how to run a GWAS pipeline
2. Learn how to use Imputation bot to automate job submission
3. Learn how to use the PGS Server extension

Summary of common GWAS analysis tools

	EPACTS	Rvtests	SNPTEST	SAIGE	BLOT-LMM	Bgenie	regenie
Input VCF	Y	Y	Y				
Sample relatedness (Quantitative outcome)	Y	Y		Y	Y	Y	Y
Sample relatedness (Binary outcome)				Y		Y	Y
Case control imbalance				Y			Y
Large sample size (>20,000)				Y	Y	Y	Y

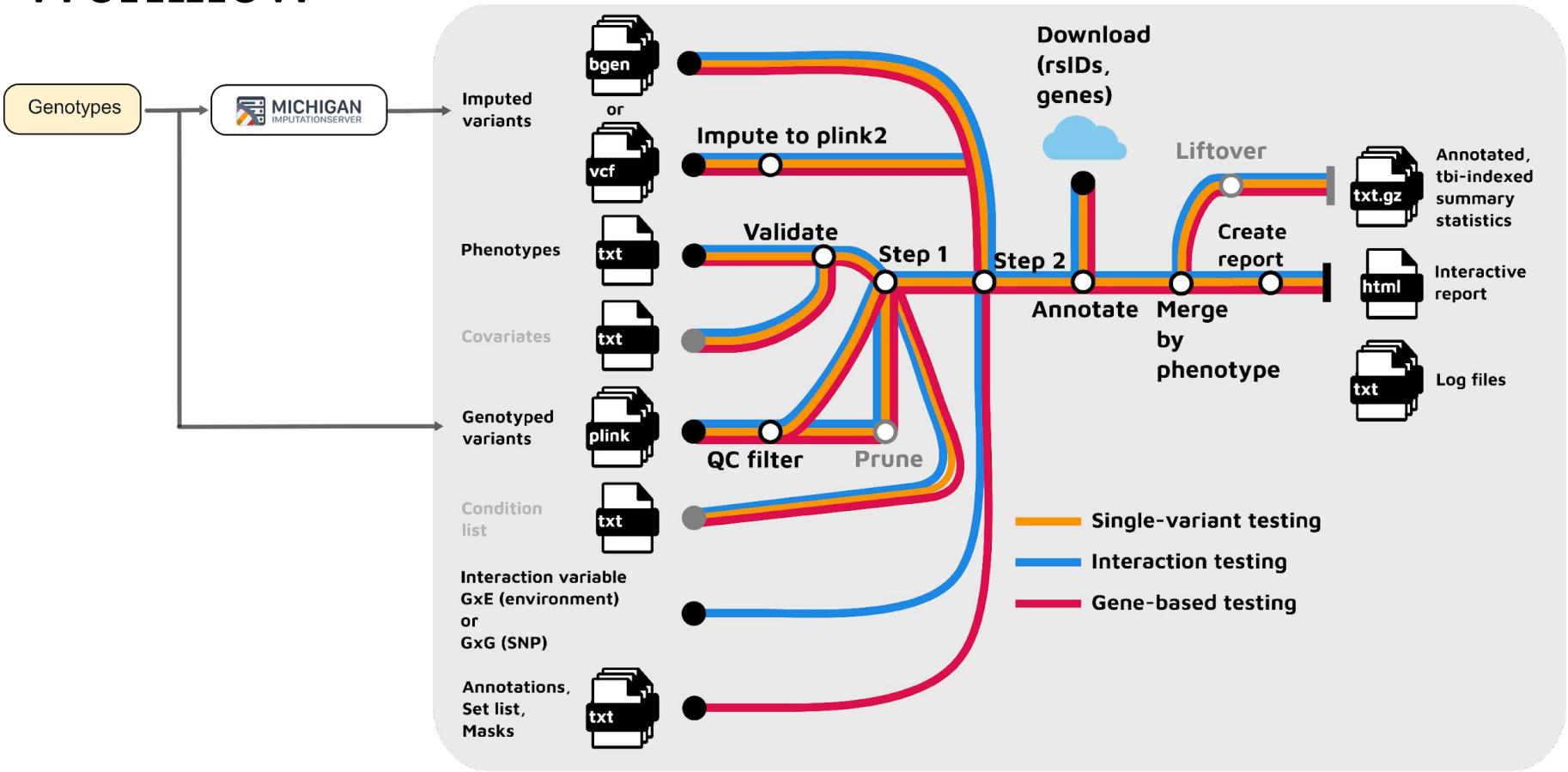
nf-gwas

- A GWAS pipeline based on REGENIE and works with VCF input
- Includes several pre- and post-processing steps

- Based on Nextflow
  - Allows to build a portable, reproducible, scalable pipeline
  - Runs on clusters (e.g. SLURM) or in the cloud (e.g. AWS Batch)

Workflow

nf-gwas



Easy to configure

1

my-gwas.config

```
params {

  project                       = 'test-gwas-additive'
  genotypes_prediction          = "$projectDir/tests/input/pipeline/example.{bim,bed,fam}"
  genotypes_association         = "$projectDir/tests/input/pipeline/example.vcf.gz"
  genotypes_association_format  = 'vcf'
  phenotypes_filename           = "$projectDir/tests/input/pipeline/phenotype.txt"
  phenotypes_columns            = 'Y1,Y2'
  regenie_test                  = 'additive'
}
```

2

```
nextflow run genepi/nf-gwas -r v1.0.0 -profile docker -c my-gwas.config
```

3

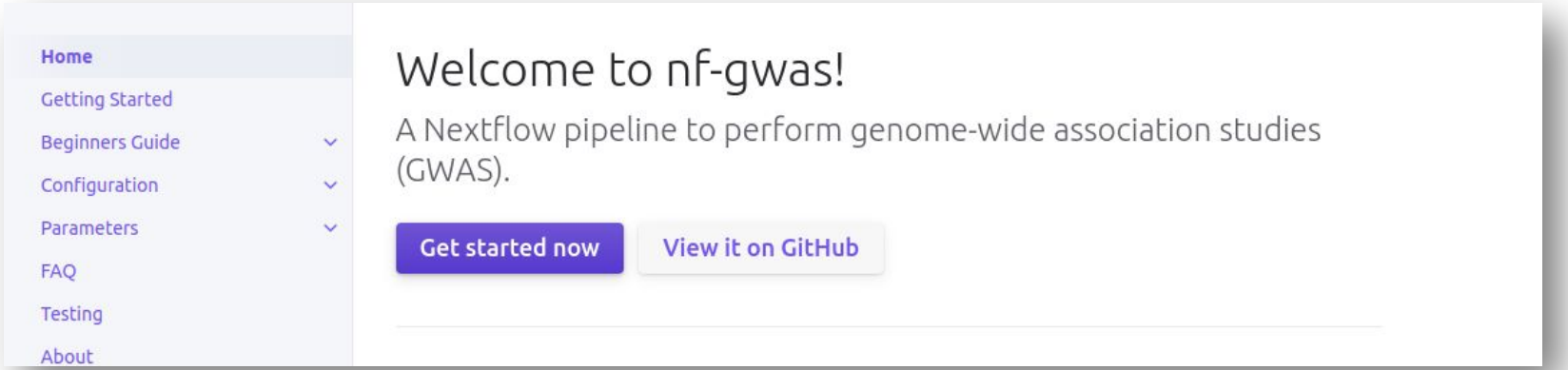
```
[47/6fe762] process > NF_GWAS:SINGLE_VARIANT_TESTS:INPUT_VALIDATION:VALIDATE_PHENOTYPES              [100%] 1 of 1 ✔
[52/de9397] process > NF_GWAS:SINGLE_VARIANT_TESTS:CHUNKING:CHUNK_ASSOCIATION_FILES (2)             [100%] 2 of 2 ✔
[a5/25f8fb] process > NF_GWAS:SINGLE_VARIANT_TESTS:CHUNKING:COMBINE_MANIFEST_FILES                  [100%] 1 of 1 ✔
[93/0ed76b] process > NF_GWAS:SINGLE_VARIANT_TESTS:QUALITY_CONTROL:QC_FILTER_GENOTYPED (1)          [100%] 1 of 1 ✔
[c5/ba3069] process > NF_GWAS:SINGLE_VARIANT_TESTS:REGENIE:REGENIE_STEP1:REGENIE_STEP1_RUN (1)      [100%] 1 of 1 ✔
[60/2aa54f] process > NF_GWAS:SINGLE_VARIANT_TESTS:REGENIE:REGENIE_STEP1:REGENIE_LOG_PARSER_STEP1 (1) [100%] 1 of 1 ✔
[18/1a395e] process > NF_GWAS:SINGLE_VARIANT_TESTS:REGENIE:REGENIE_STEP2:REGENIE_STEP2_RUN (example)  [100%] 2 of 2 ✔
[f8/83a76d] process > NF_GWAS:SINGLE_VARIANT_TESTS:REGENIE:REGENIE_STEP2:REGENIE_LOG_PARSER_STEP2   [100%] 1 of 1 ✔
[d6/7554d5] process > NF_GWAS:SINGLE_VARIANT_TESTS:ANNOTATION:ANNOTATE_RESULTS (2)                  [  0%] 0 of 2
[-        ] process > NF_GWAS:SINGLE_VARIANT_TESTS:MERGE_RESULTS                                    -
[-        ] process > NF_GWAS:SINGLE_VARIANT_TESTS:FILTER_RESULTS                                   -
[-        ] process > NF_GWAS:SINGLE_VARIANT_TESTS:REPORTING:REPORT                                 -
[-        ] process > NF_GWAS:SINGLE_VARIANT_TESTS:REPORTING:REPORT_INDEX
```

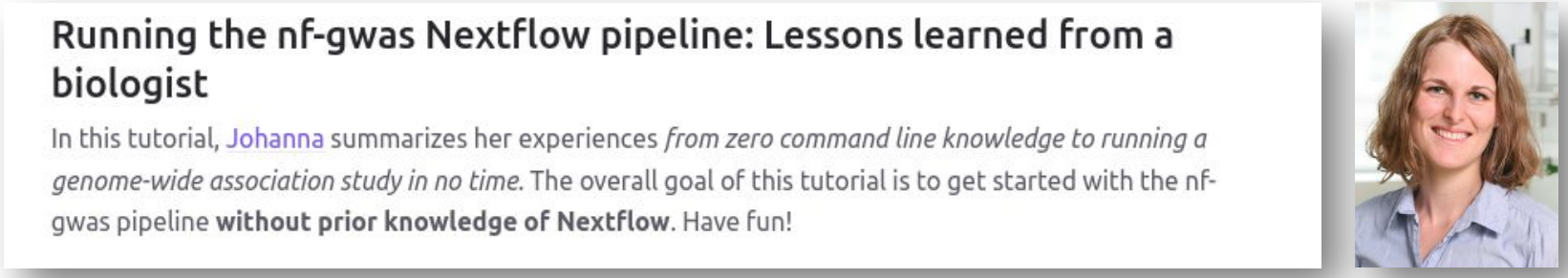
Interactive Manhattan Plot

Annotated Top Hits

CHROM	GENPOS	ID	ALLELE0	ALLELE1	A1FREQ	INFO	N	TEST	EXTRA	BETA	SE	CHISQ	LOG10P
6	161010118	rs10455872	G	A	0.933377	1	371458	ADD	NA	-1.41724	0.00471156	90481.2	19650.3
6	160985526	rs118039278	A	G	0.933278	0.991973	371458	ADD	NA	-1.42123	0.00472746	90379.4	19628.2
6	160997118	rs74617384	T	A	0.933446	1	371458	ADD	NA	-1.40973	0.00471335	89457.2	19427.9
6	161005610	rs55730499	T	C	0.932027	0.996391	371458	ADD	NA	-1.39503	0.00467706	88965.4	19321.2
6	161089307	rs56393506	T	C	0.828746	0.959932	371458	ADD	NA	-0.872161	0.00317397	75507	16398.7
6	161103805	rs374071816	A	T	0.92189	0.894006	371458	ADD	NA	-1.09158	0.00458159	56764.4	12328.7
6	161108144	rs2315065	A	C	0.914042	0.951733	371458	ADD	NA	-0.986453	0.00424719	53944.8	11716.4
6	161123451	rs4252185	C	T	0.915634	0.936269	371458	ADD	NA	-0.966992	0.00431546	50210	10905.4
6	160986915	rs6938647	C	A	0.195123	0.972351	371458	ADD	NA	-0.492756	0.00296913	27542.7	5983.14
6	161030554	6:161030554_GT_G	G	GT	0.989319	0.868419	371458	ADD	NA	0.0120399	17374.9	3775.14	
6	160978997	rs12210186	G	A	0.82095	0.970278	371458	ADD	NA	-0.393531	0.00304729	16677.5	3623.69
6	161086716	rs76735376	A	G	0.989979	0.890853	371458	ADD	NA	-1.54276	0.0122597	15835.5	3440.84
6	160750643	rs6911381	T	C	0.849502	0.991031	371458	ADD	NA	-0.405608	0.0032378	15693.3	3409.95
6	160750469	rs6933264	C	A	0.849495	0.991177	371458	ADD	NA	-0.405451	0.00323758	15683.2	3407.76
6	160746512	rs3103347	T	C	0.849697	0.994174	371458	ADD	NA	-0.40407	0.0032347	15604.3	3390.62
6	160745238	rs3103348	C	T	0.849689	0.998686	371458	ADD	NA	-0.401404	0.00322719	15470.9	3361.65
6	160740836	rs3125050	C	A	0.850559	0.996888	371458	ADD	NA	-0.402226	0.00323813	15429.4	3352.65
6	160963010	6:160963010_AGAG_A	A	AGAG	0.820499	0.995646	371458	ADD	NA	0.00300438	15405.7	3347.5	
6	160726432	rs3127584	A	G	0.849847	0.999463	371458	ADD	NA	-0.400297	0.0032272	15385.6	3343.13

Documentation





https://genepi.github.io/nf-gwas/

Extension: PGS Server

Motivation

Extend Imputation Server to simplify the application of polygenic risk scores (PGS) to imputed genotypes

- PGS: aggregates the effects of many genetic variants into a single number which predicts genetic predisposition for a phenotype

- Available since April 2022
- More than 1,200 submitted jobs

Have you ever used the PGS tool on the MIS before?

Yes

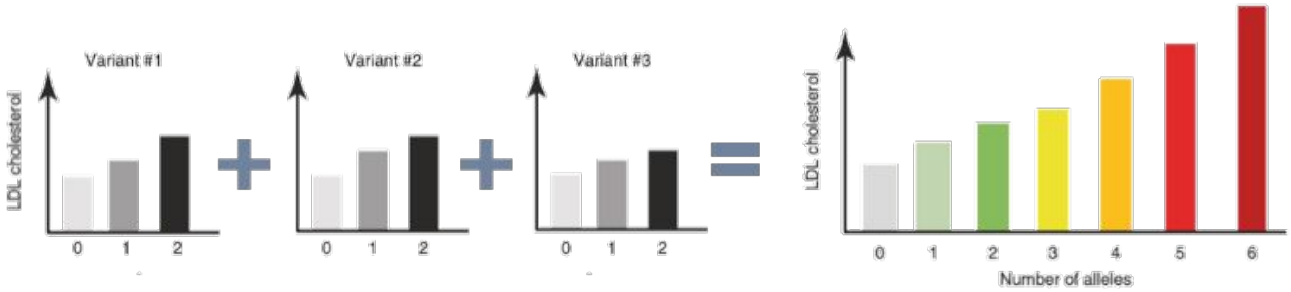
0%

No

0%

Polygenic Scores (PGS)

Combines the estimated effect of many genetic variants on an individual's phenotype



$$\mathrm{PRS}_i = \sum_{j=1}^{n} x_{ij}\beta_j$$

counting the number of effect alleles for each individual and multiplying the count by the corresponding effect size

A higher PGS indicates a greater accumulation of risk-associated genetic variants in an individual

ASHG
American Society of Human Genetics

PGS Repositories

PGS Catalog

- The PGS Catalog is an open database of published polygenic scores (PGS)
- Each PGS in the Catalog is consistently annotated with relevant metadata
- Scores> 3,000
- http://pgscatalog.org

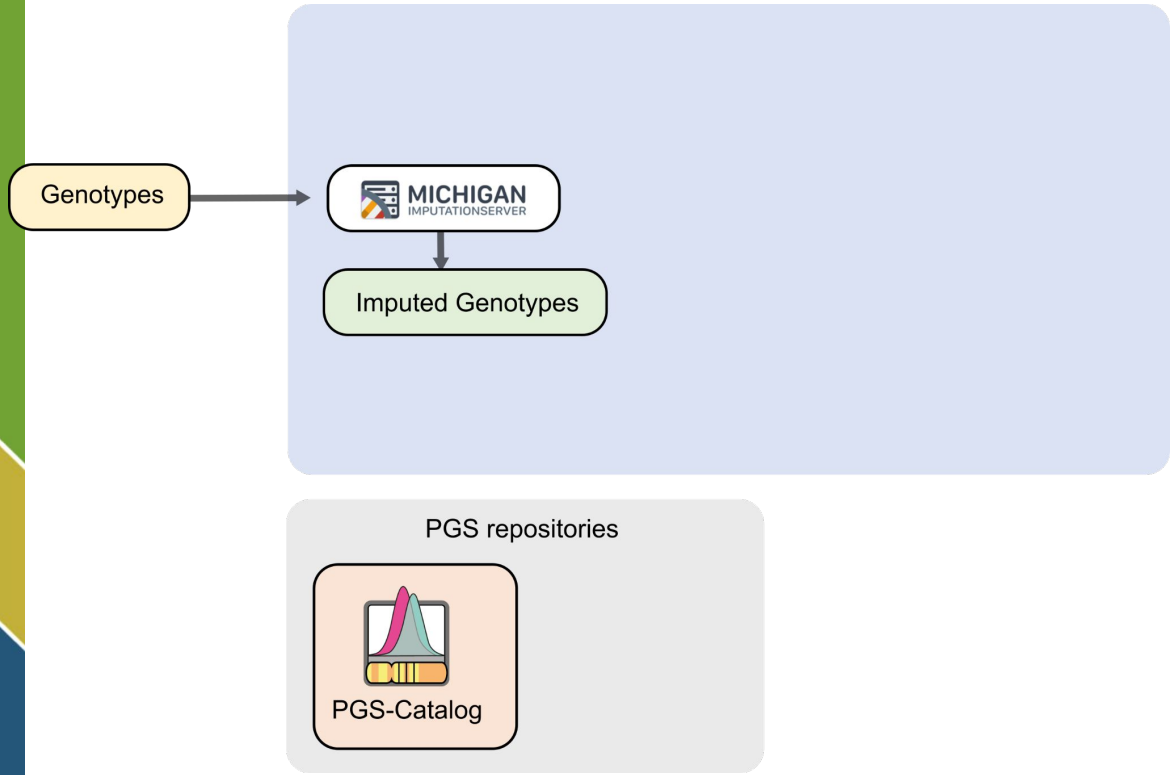
Cancer PRSWeb

- Scores: 305
- https://prsweb.sph.umich.edu

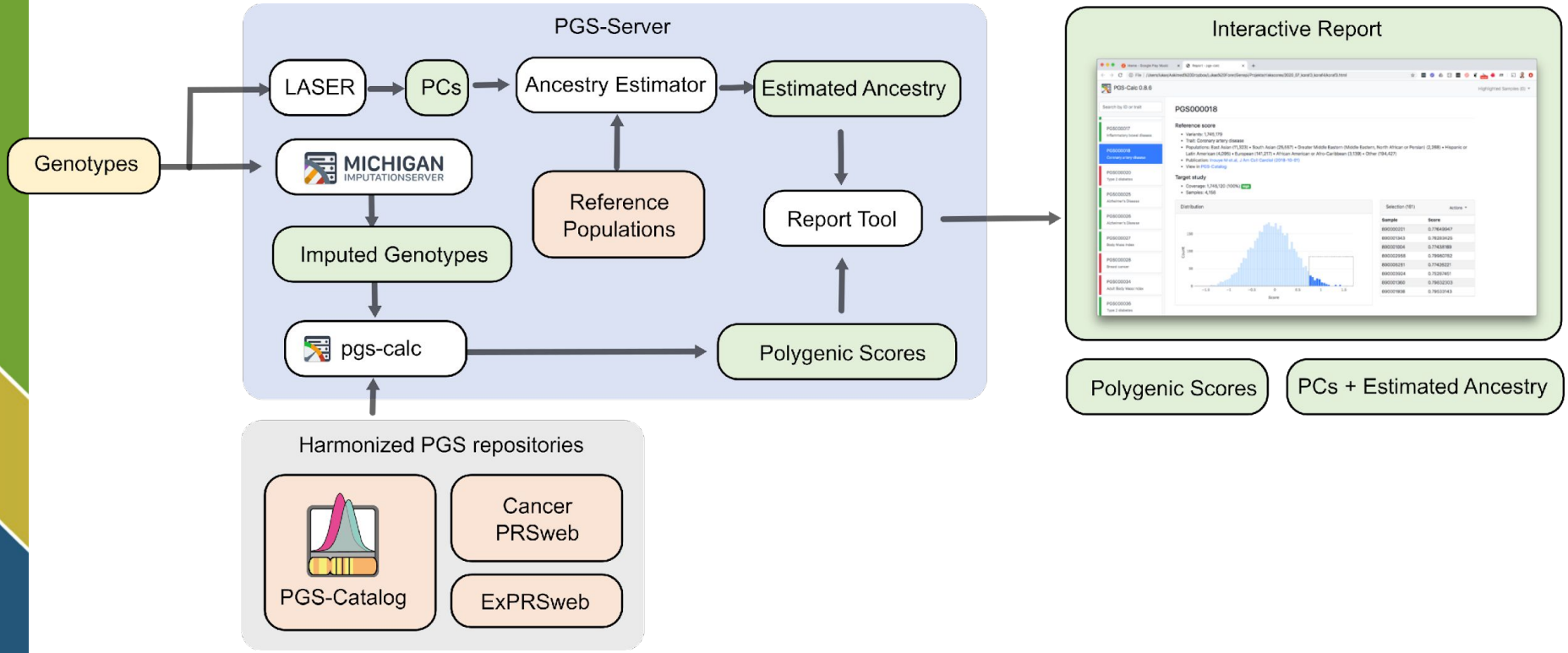
ExPRSWeb

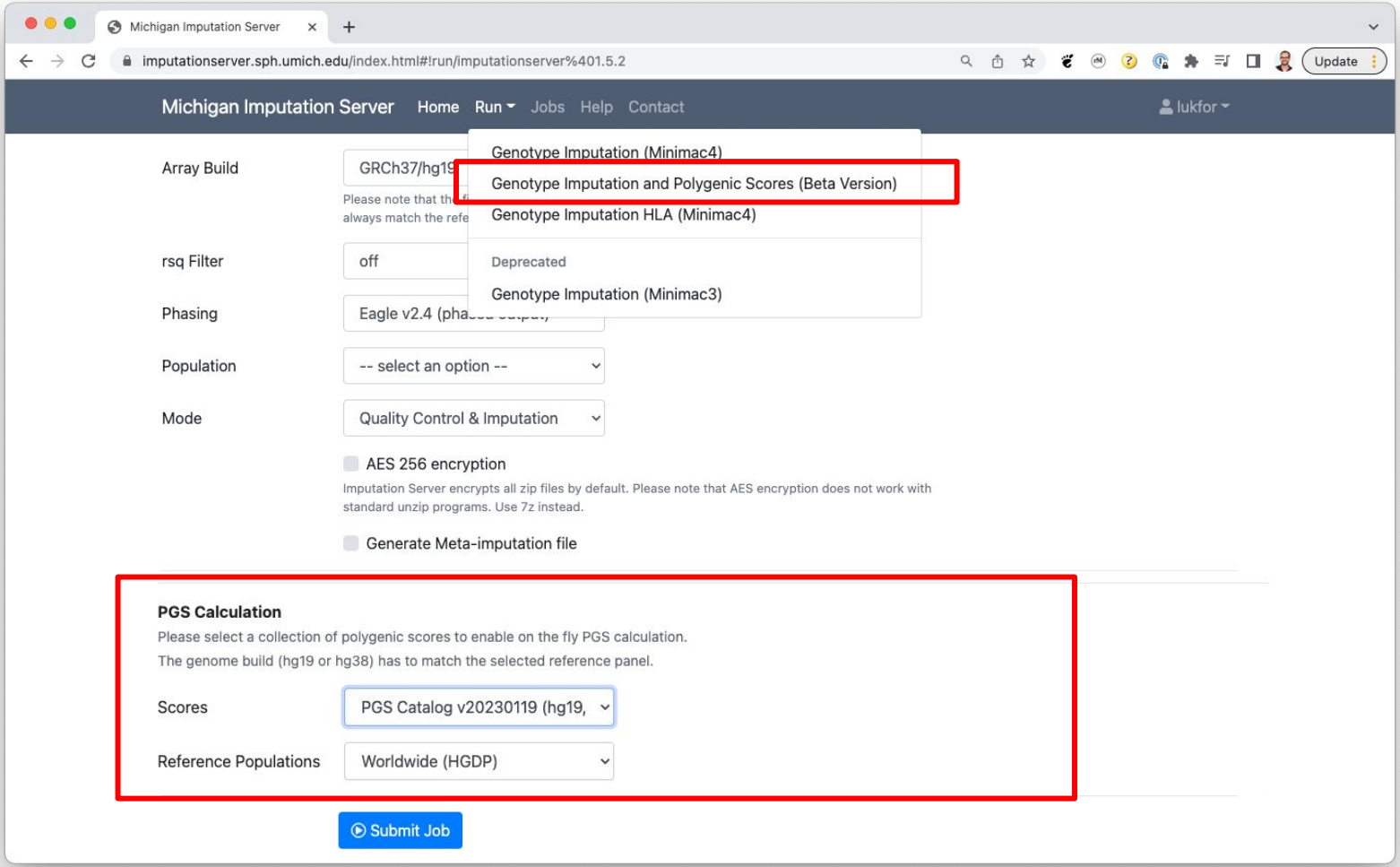
- Scores: 514
- https://exprsweb.sph.umich.edu

Problem: PGS on Imputed Genotypes

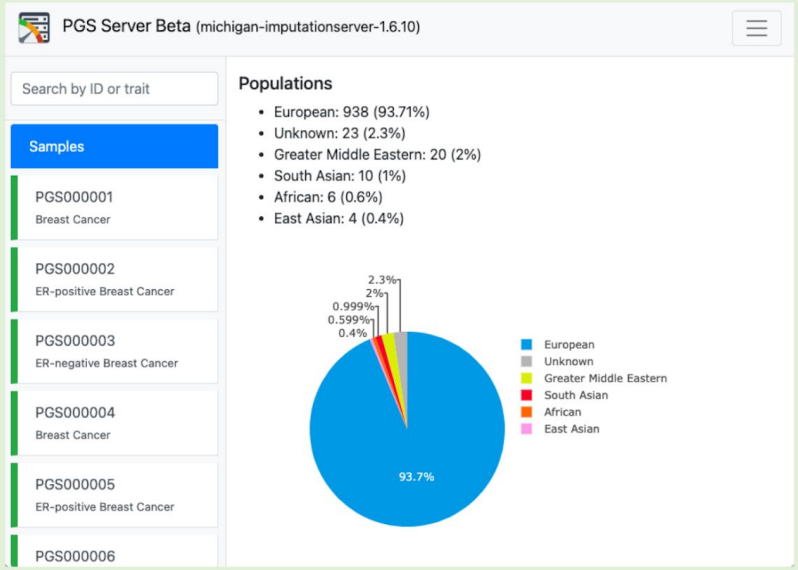


PGS Server Workflow




Michigan Imputation Server
imputationserver.sph.umich.edu/index.html#!run/imputationserver%401.5.2
Update
Michigan Imputation Server
Home
Run
Jobs
Help
Contact
lukfor
Genotype Imputation (Minimac4)
Genotype Imputation and Polygenic Scores (Beta Version)
Genotype Imputation HLA (Minimac4)
Deprecated
Genotype Imputation (Minimac3)
Array Build
GRCh37/hg19
Please note that th
always match the refe
rsq Filter
off
Phasing
Eagle v2.4 (pha
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.
Generate Meta-imputation file
PGS Calculation
Please select a collection of polygenic scores to enable on the fly PGS calculation.
The genome build (hg19 or hg38) has to match the selected reference panel.
Scores
PGS Catalog v20230119 (hg19,
Reference Populations
Worldwide (HGDP)
Submit Job

Estimated Ancestries



Report with Population Check

PGS000027

Excluded **23 sample(s)** (Unknown) due to ancestry mismatch.
Excluded **20 sample(s)** (Greater Middle Eastern) due to ancestry mismatch.
Excluded **10 sample(s)** (South Asian) due to ancestry mismatch.
Excluded **4 sample(s)** (East Asian) due to ancestry mismatch.

Reference score

- Variants: 2,100,302
- Variants ignored: 0
- Trait: Body Mass Index
- Mapped Trait(s):
  - EFO_0004340: body mass index
- Samples: 238,944
- Population:
  - European: 99.1%
  - Hispanic or Latin American: 0.5%
  - African: 0.4%
- Publication: Khera AV et.al, Cell (2019-04-01)
- View in PGS-Catalog

Target study

- Coverage: 2,100,302 (100%) high
- Samples: 1,001

PGS-Calc 0.8.6

Highlighted Samples (0)

Search by ID or trait

PGS000017
Inflammatory bowel disease

PGS000018
Coronary artery disease

PGS000020
Type 2 diabetes

PGS000025
Alzheimer's Disease

PGS000026
Alzheimer's Disease

PGS000027
Body Mass Index

PGS000028
Breast cancer

PGS000034
Adult Body Mass Index

PGS000036
Type 2 diabetes

PGS000018

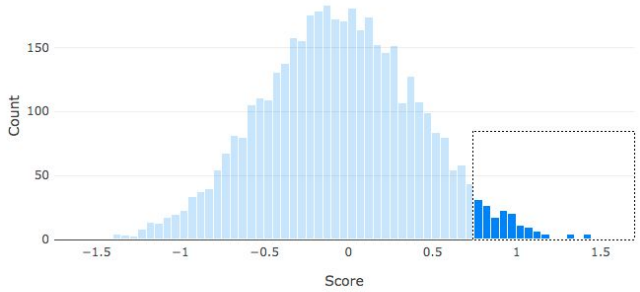
Reference score

- Variants: 1,745,179
- Trait: Coronary artery disease
- Populations: East Asian (11,323) • South Asian (25,557) • Greater Middle Eastern (Middle Eastern, North African or Persian) (2,268) • Hispanic or Latin American (4,095) • European (141,217) • African American or Afro-Caribbean (3,139) • Other (194,427)
- Publication: Inouye M et.al, J Am Coll Cardiol (2018-10-01)
- View in PGS-Catalog

Target study

- Coverage: 1,745,120 (100%) high
- Samples: 4,158

Distribution



Selection (161) Actions

Sample	Score
890000201	0.77649947
890001343	0.76283425
890001004	0.77438169
890002958	0.79980782
890005251	0.77426221
890003924	0.75287451
890001360	0.79832303
890001938	0.79533143

Conclusion

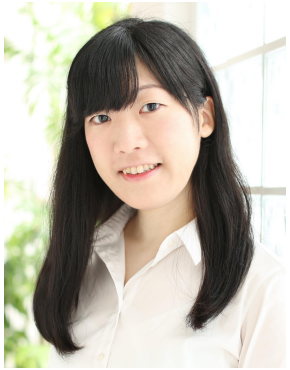
- PGS became a key role in genetic research

- Extension of Imputation Server
  - Calculates scores automatically after imputation
  - Supports PGS Catalog, CancerPRSWeb and ExPRSWeb
  - Creates interactive reports

Taking the burden out of genetic analysis

Section 5

HLA Imputation



Saori Sakaue
Broad Institute
ssakaue@broadinstitute.org
@saorisakaue

MICHIGAN IMPUTATIONSERVER

ASHG American Society of Human Genetics

Learning objectives

Participants will

1. Learn variations within HLA
2. Learn how to impute HLA alleles and amino acid sequences
3. Learn how to associate HLA alleles with disease

Genotype Imputation is great, but not perfect

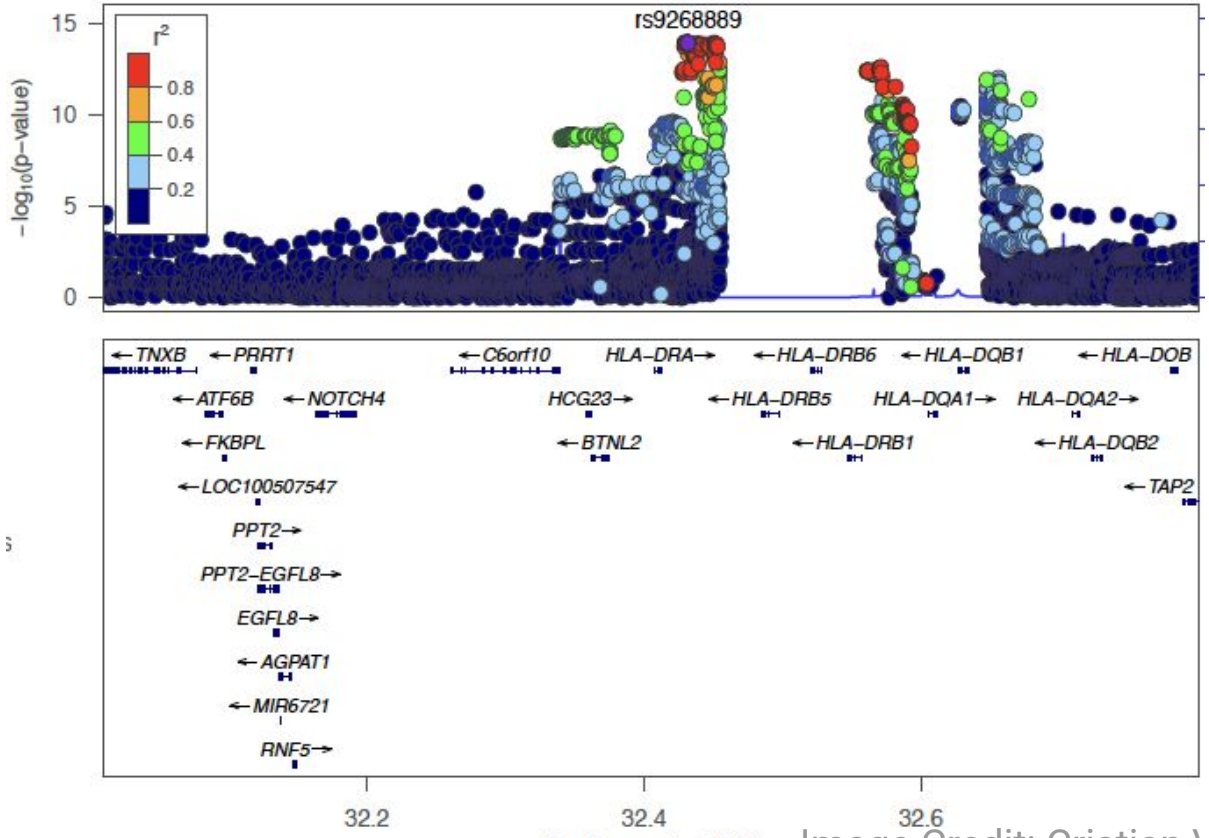


Image Credit: Cristian Valencia

Are you planning to impute the HLA region?

No

0%

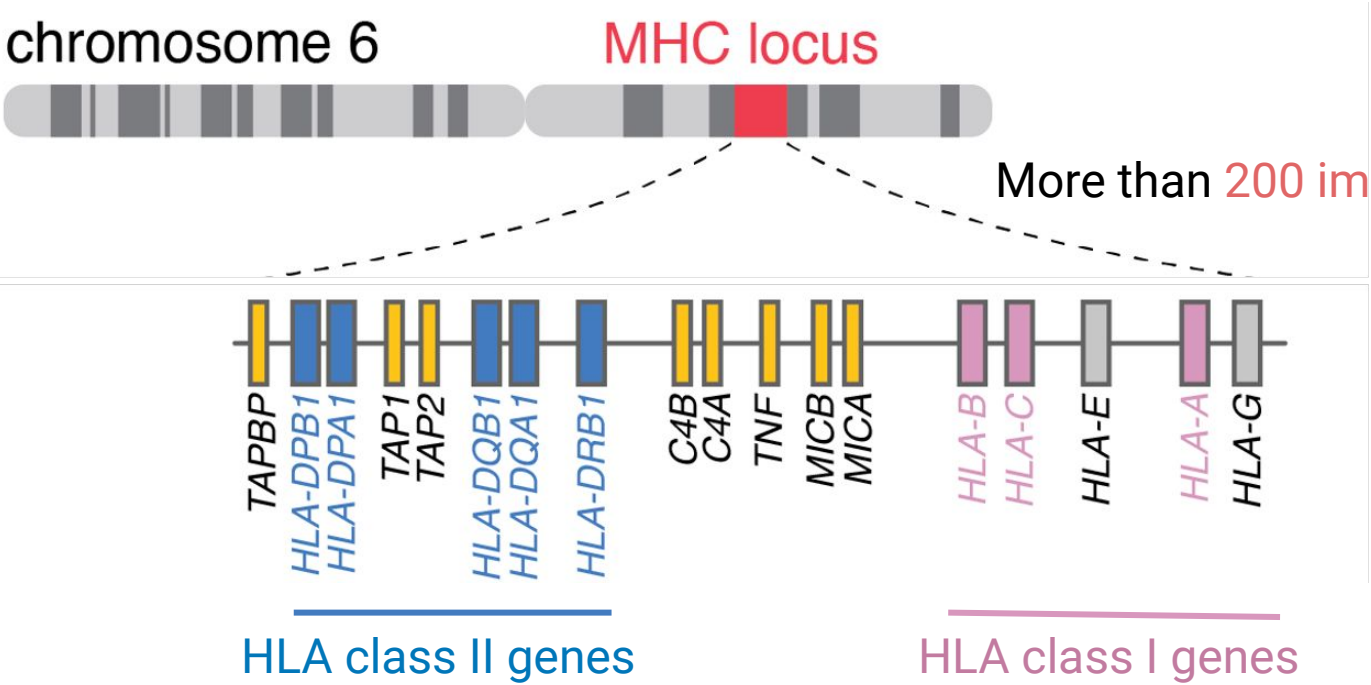
Yes

0%

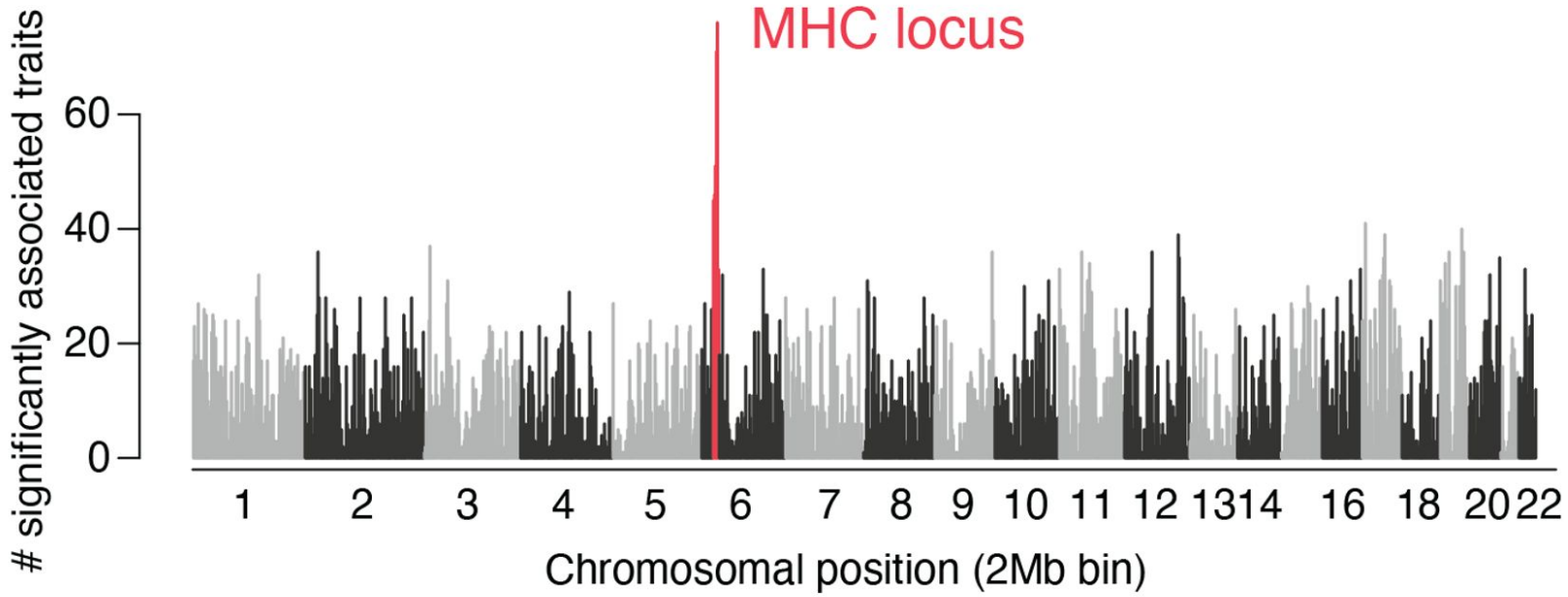
Already imputed!

0%

HLA genes are within MHC (major histocompatibility complex) locus on chromosome 6

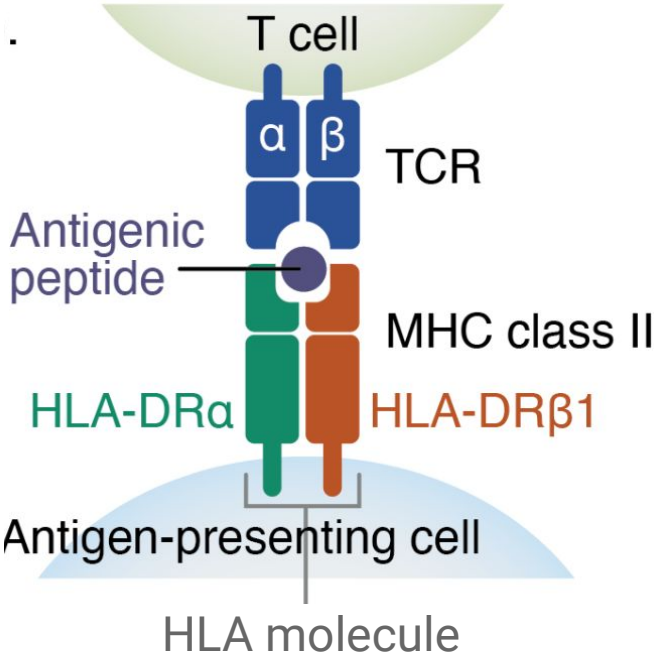


MHC locus confers the largest number of associations of any locus genome-wide

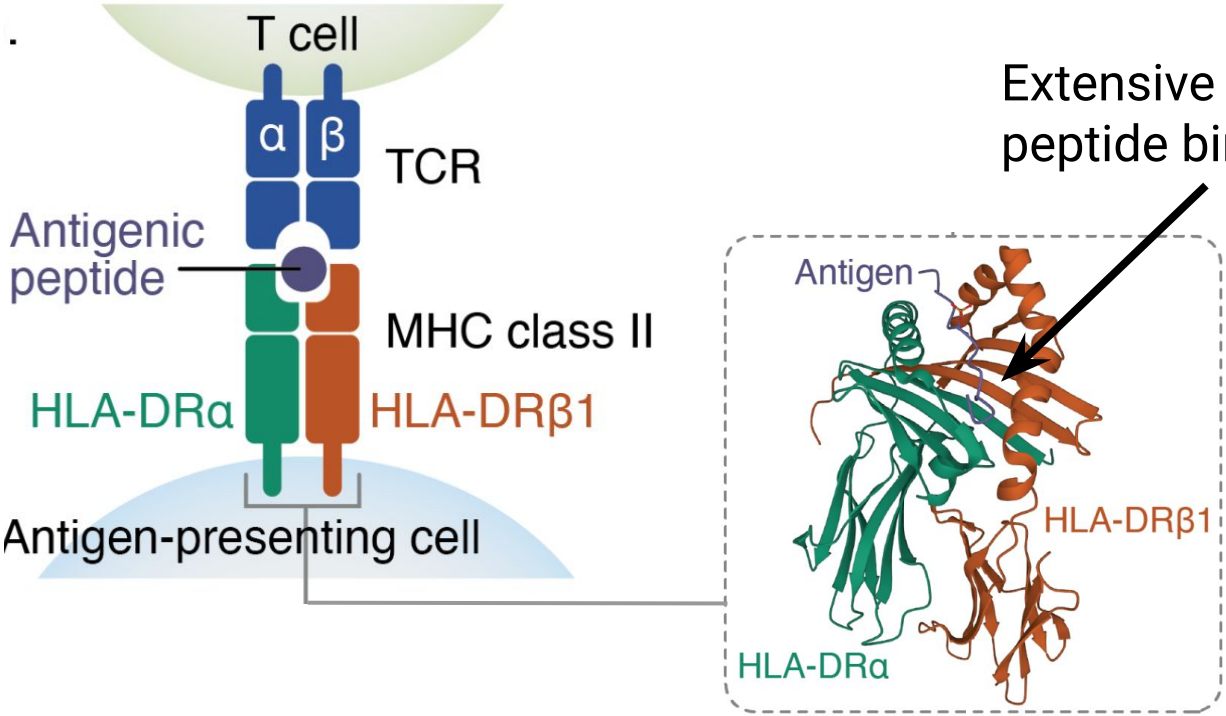


In UK Biobank, Sakaue et al. Nat Genet 2021, Sakaue et al. Nature Protocols 2023

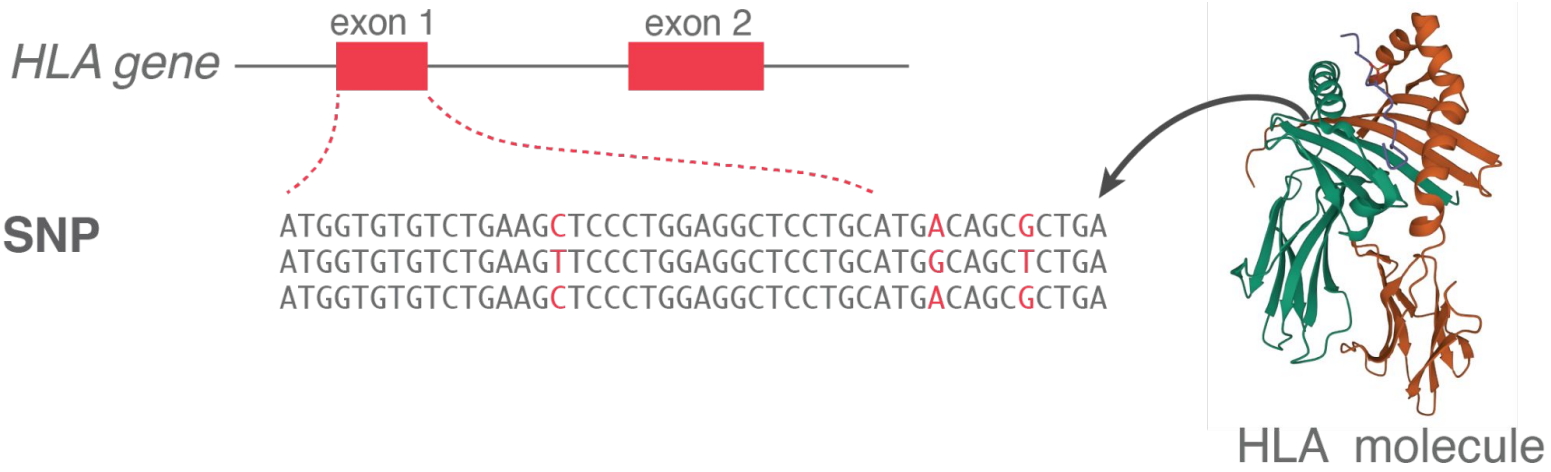
HLA presents antigens to T cells



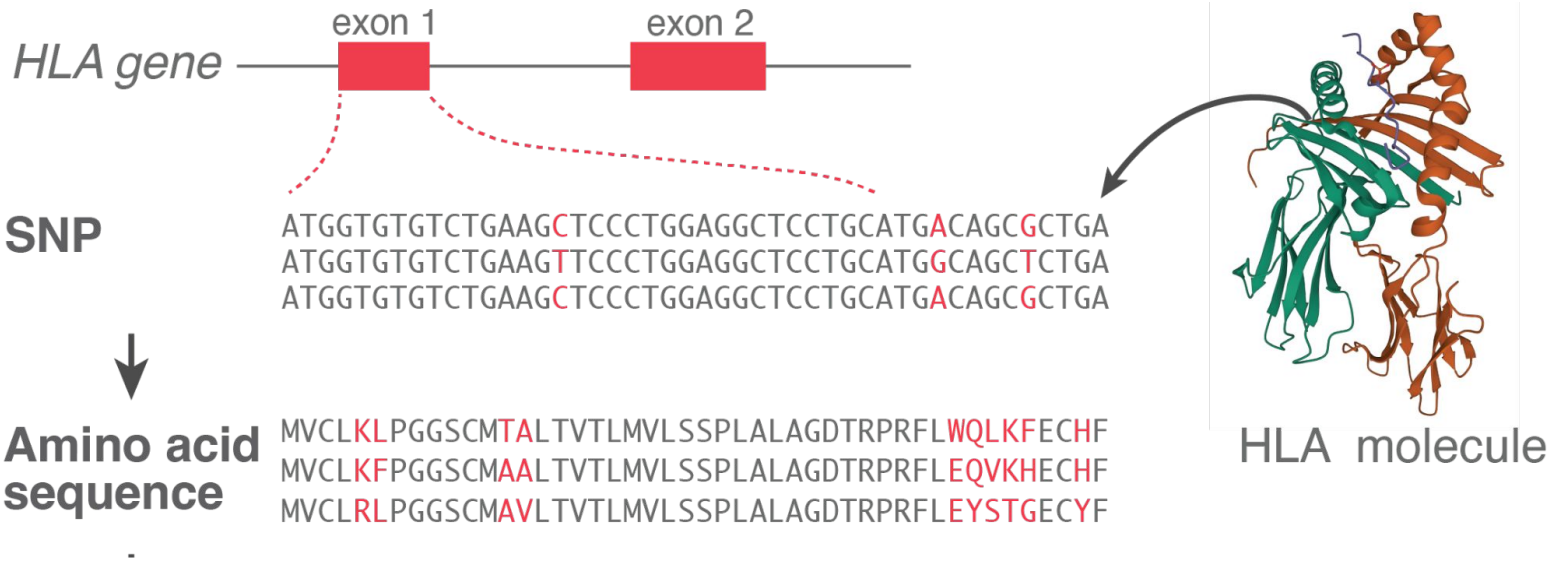
HLA presents antigens to T cells, and is extremely polymorphic



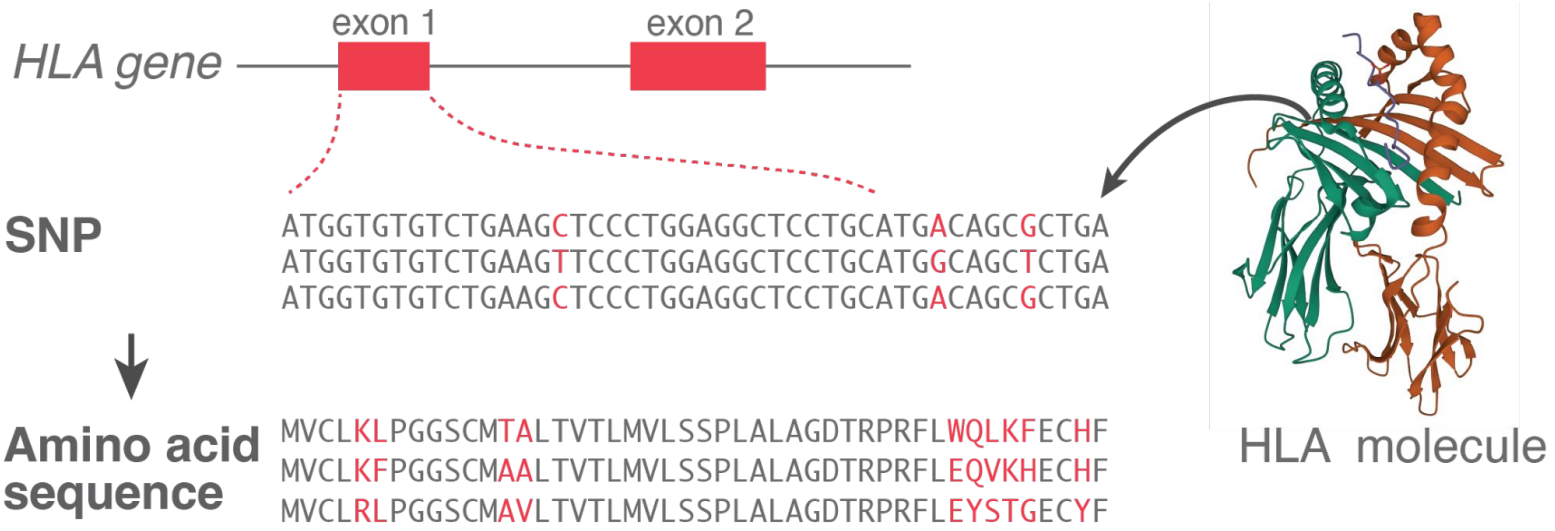
HLA variations determine HLA “alleles” and function



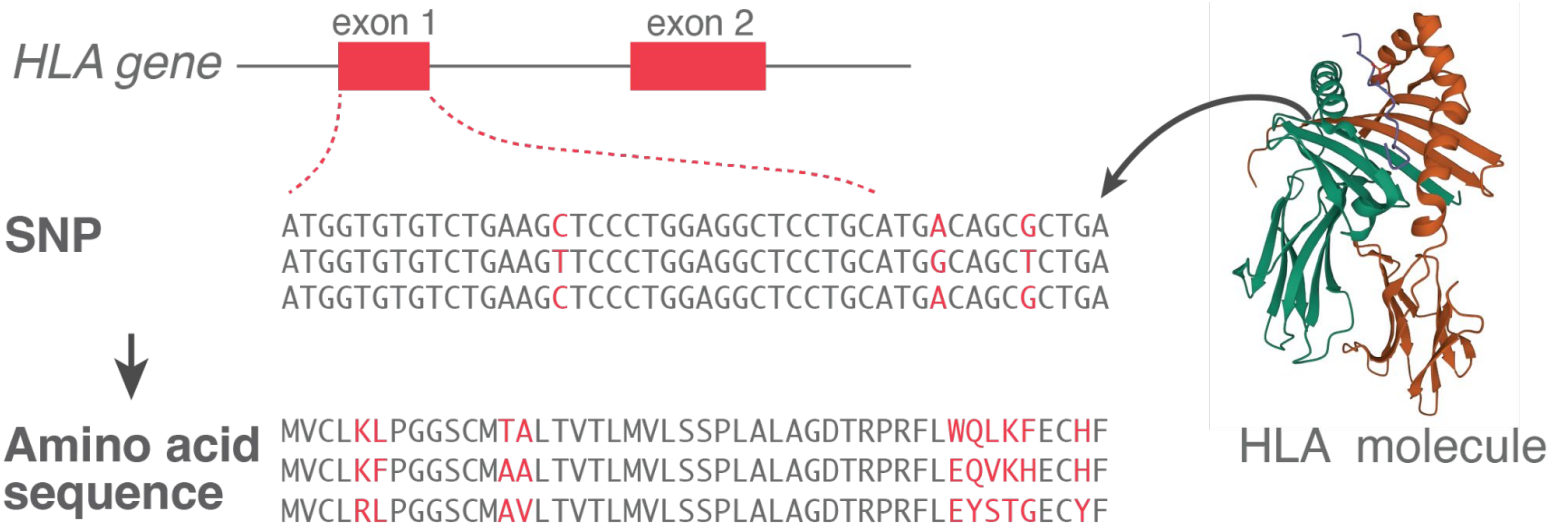
HLA variations determine HLA “alleles” and function



HLA variations determine HLA “alleles” and function

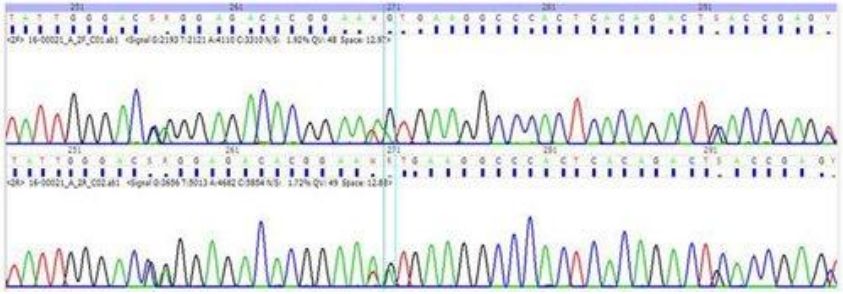


HLA variations determine HLA “alleles” and function

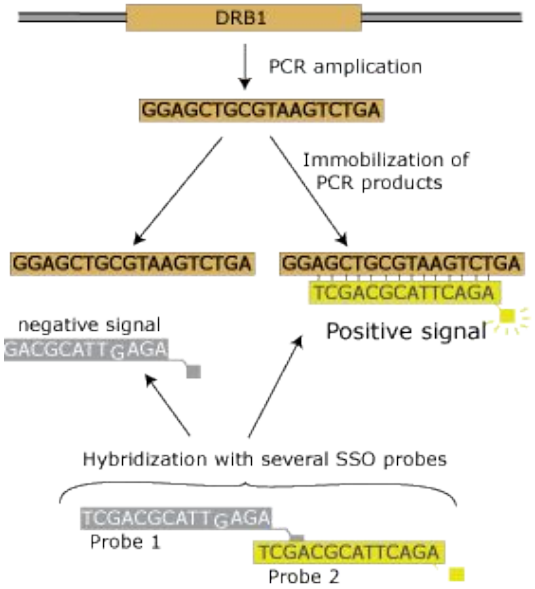


Direct HLA typing is challenging and costly

Sanger sequence-based typing (SBT)



Sequence-specific oligonucleotide probe hybridization (SSOP)



http://medweb4.unige.ch/immunologie/home/HSC/donor/HLA_typing/SSO.php

Genotype imputation in a nutshell

	Genotype imputation
Known	Genotyped SNPs
Given	Reference haplotype with whole-genome SNPs
Unknown (Impute!)	Untyped SNPs

HLA imputation in a nutshell

	Genotype imputation	HLA imputation
Known	Genotyped SNPs	Genotyped SNPs
Given	Reference haplotype with whole-genome SNPs	Reference haplotype with HLA alleles and amino acid sequences
Unknown (Impute!)	Untyped SNPs	Untyped HLA alleles and amino acid sequences

What field of HLA alleles describe non-synonymous substitutions?

Field 1

0%

Field 2

0%

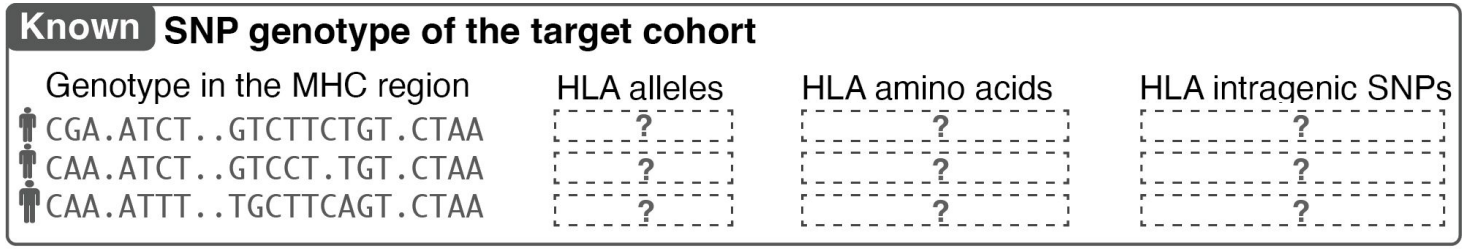
Field 3

0%

Field 4

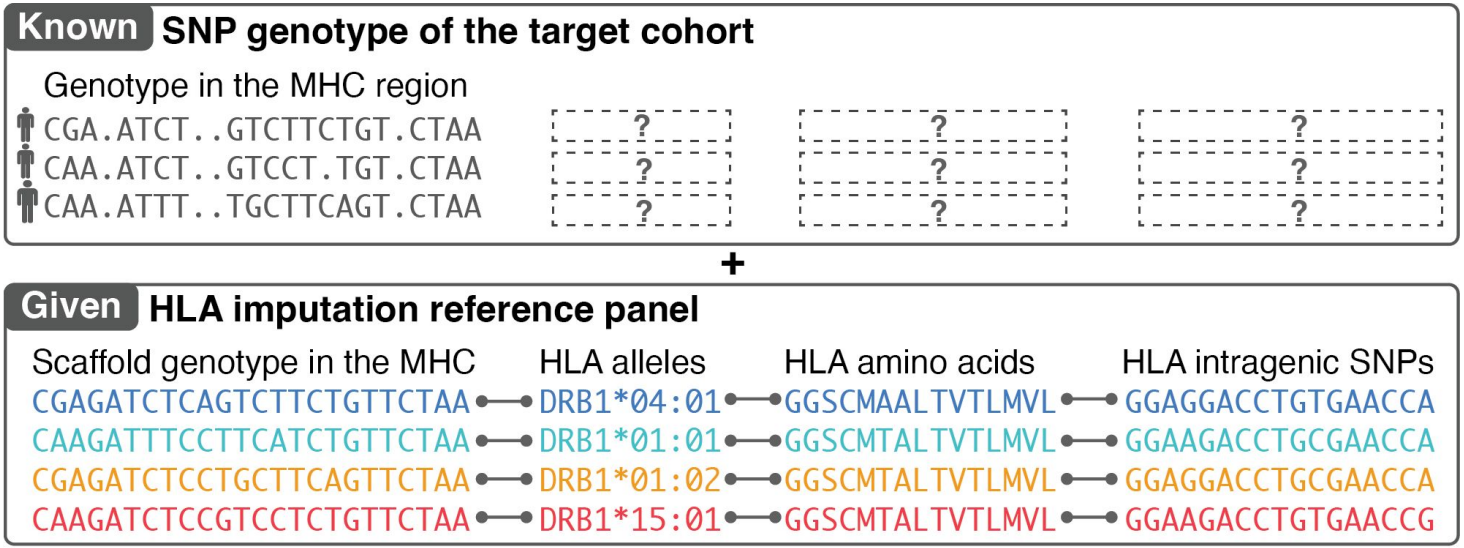
0%

HLA imputation in a nutshell



QCed Plink *.{bed,bim,fam}

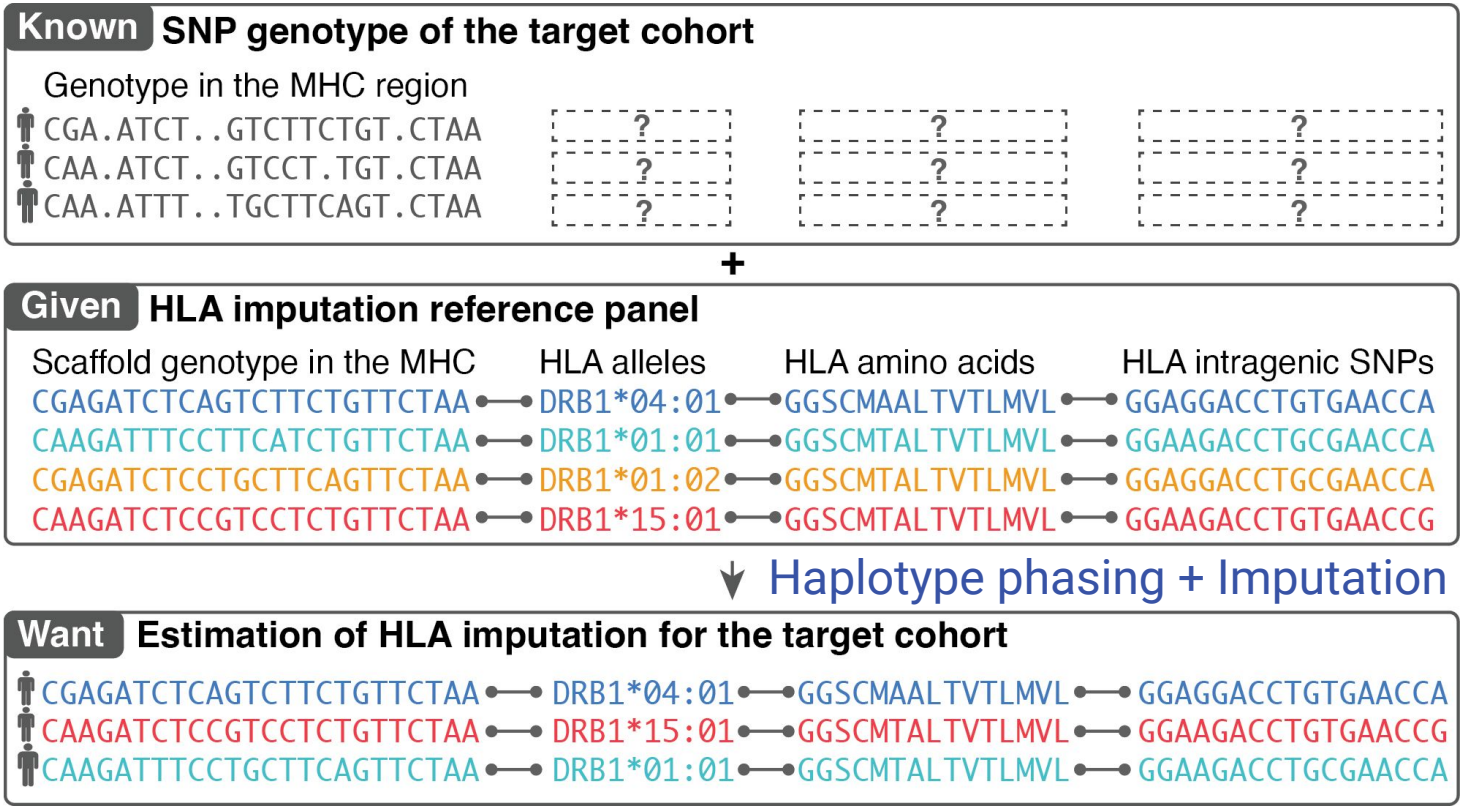
HLA imputation in a nutshell



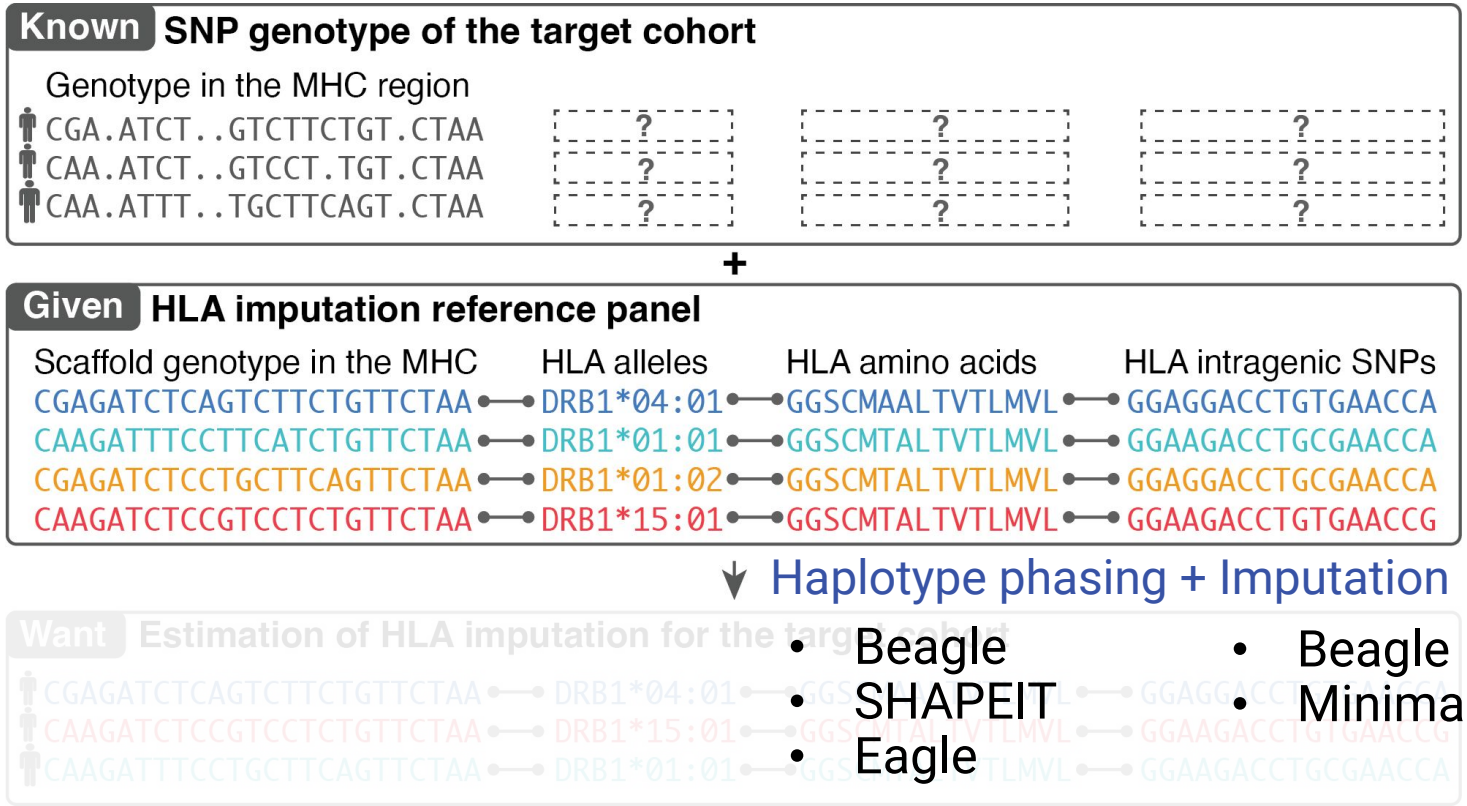
Multi-ancestry panel
(EUR, EAS, SAS, AFR, LAT)

Luo et al. Nat Genet 2021, Sakaue et al. Nat Protocols 2023

HLA imputation in a nutshell

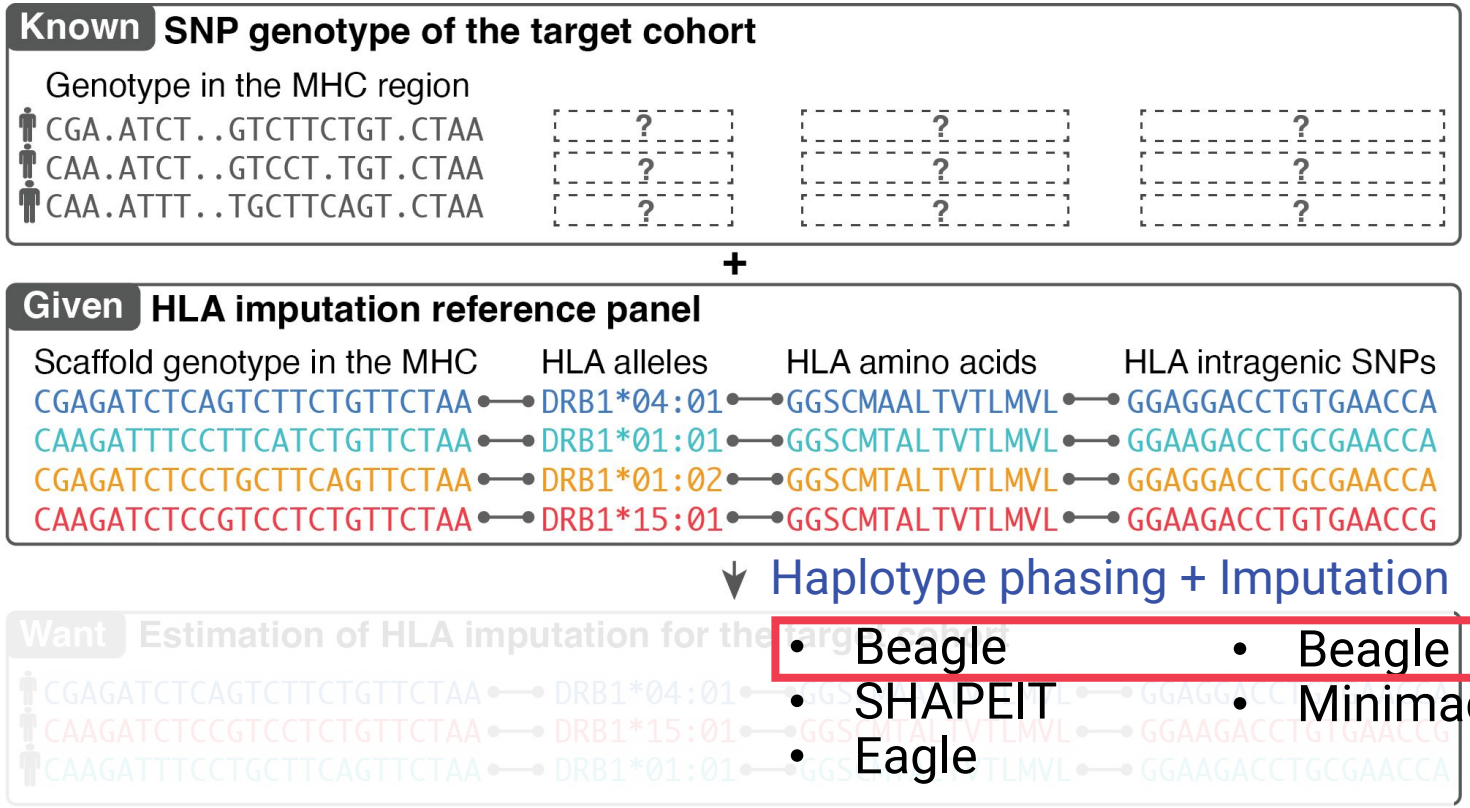


HLA imputation in a nutshell



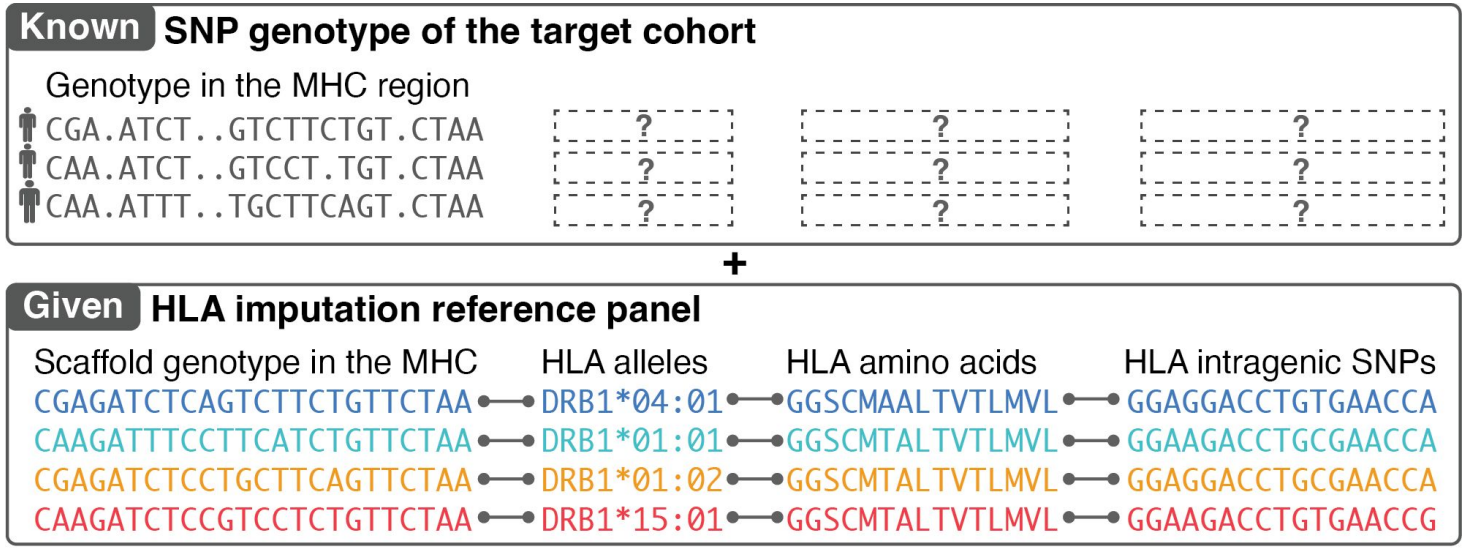
Browning & Browning. AJHG 2007, Dalaneau et al. Nat Methods 2013, Loh et al. Nat Genet 2016, Das et al. Nat Genet 2016

HLA imputation in a nutshell



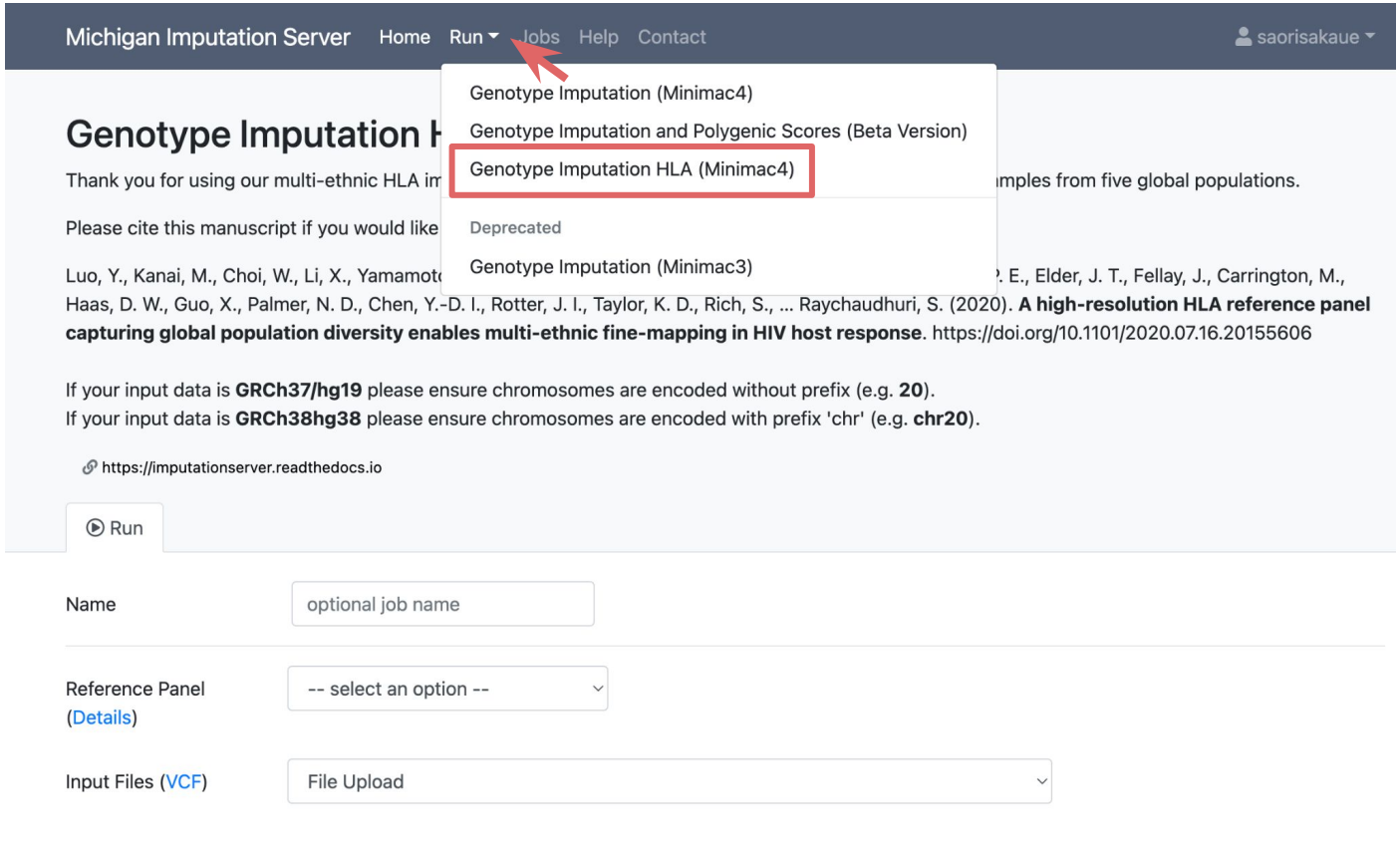
Jia et al. PLoS One 2013

HLA imputation in a nutshell

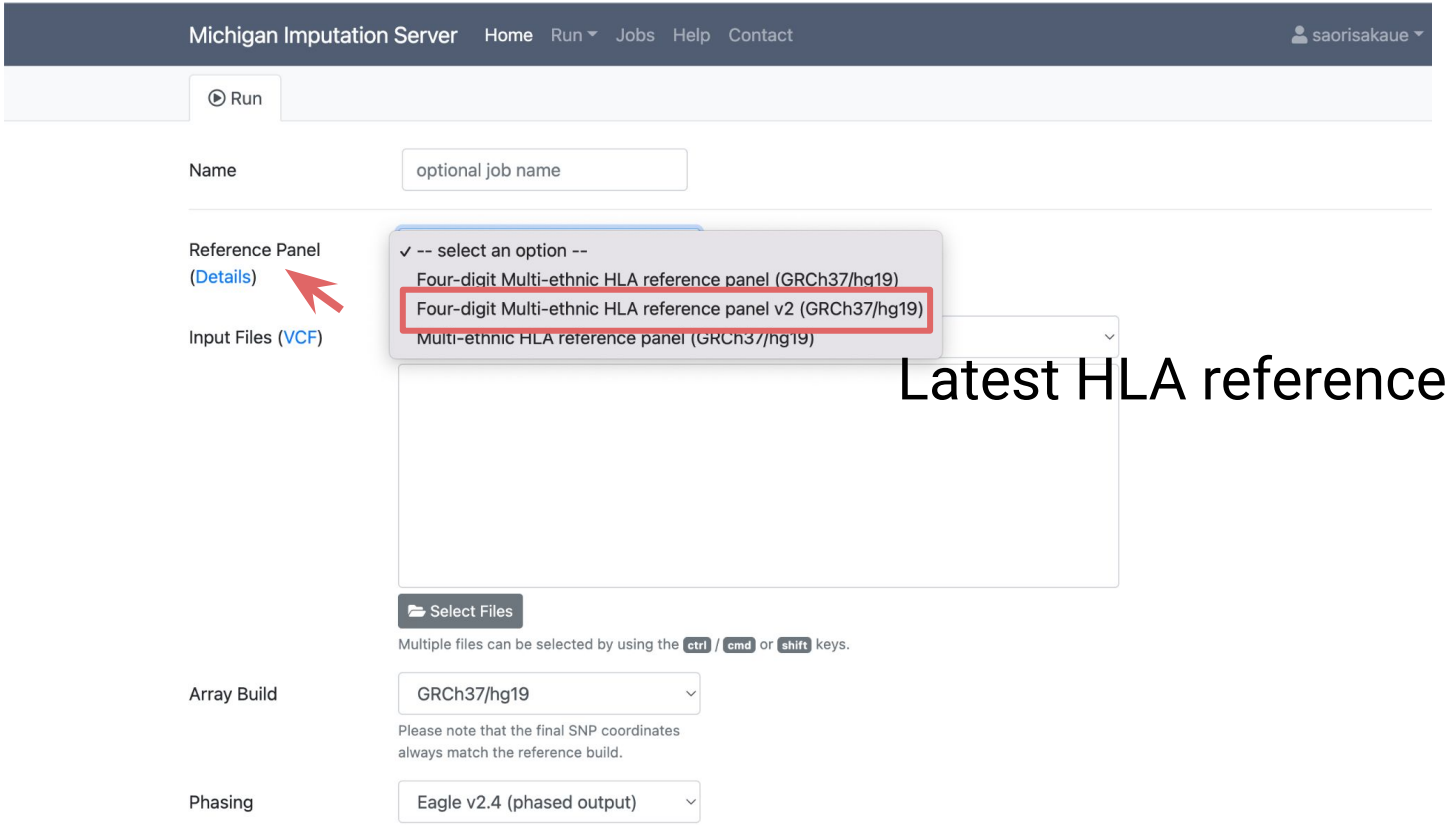


Sakaue et al. Nat Protocols 2023

HLA imputation in MIS

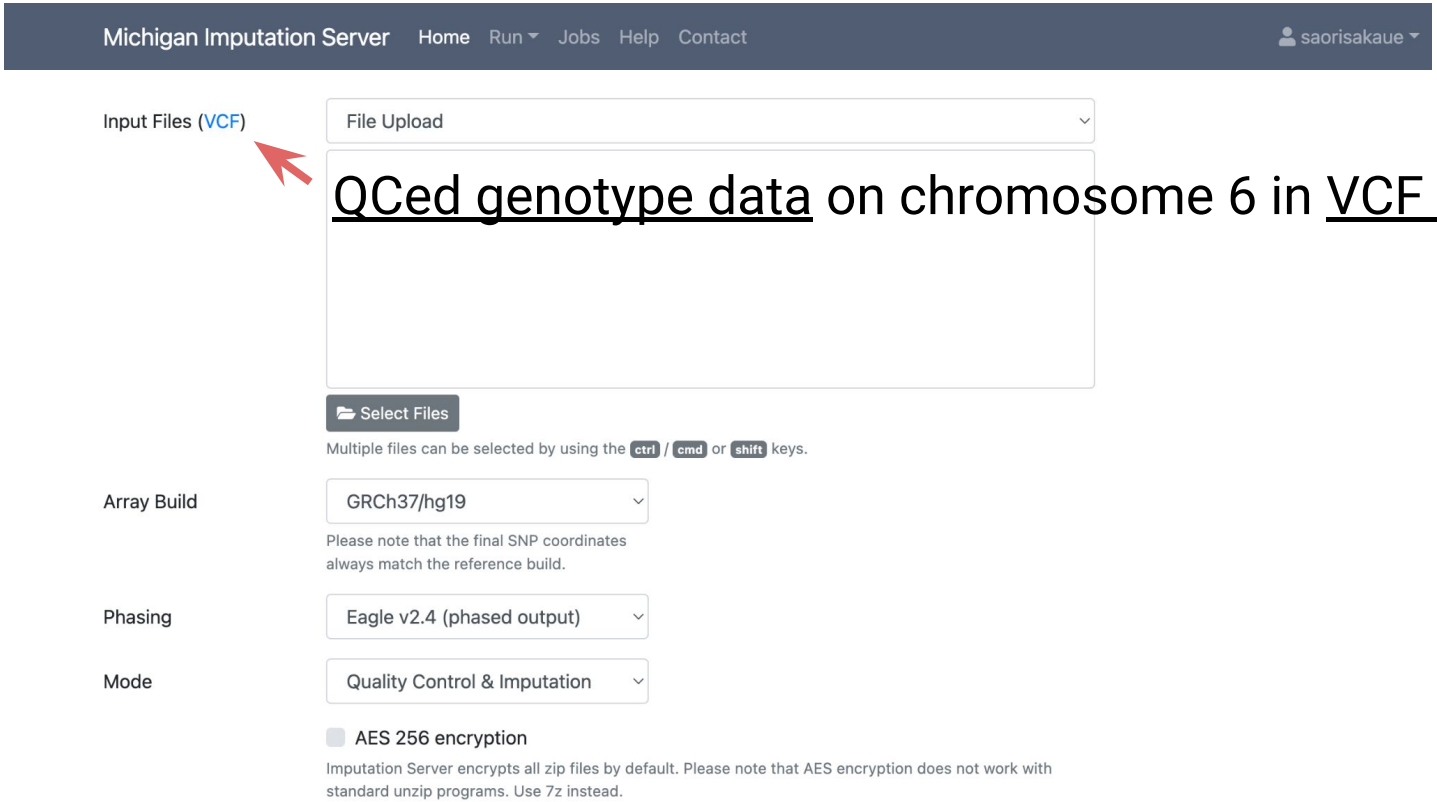


HLA imputation in MIS

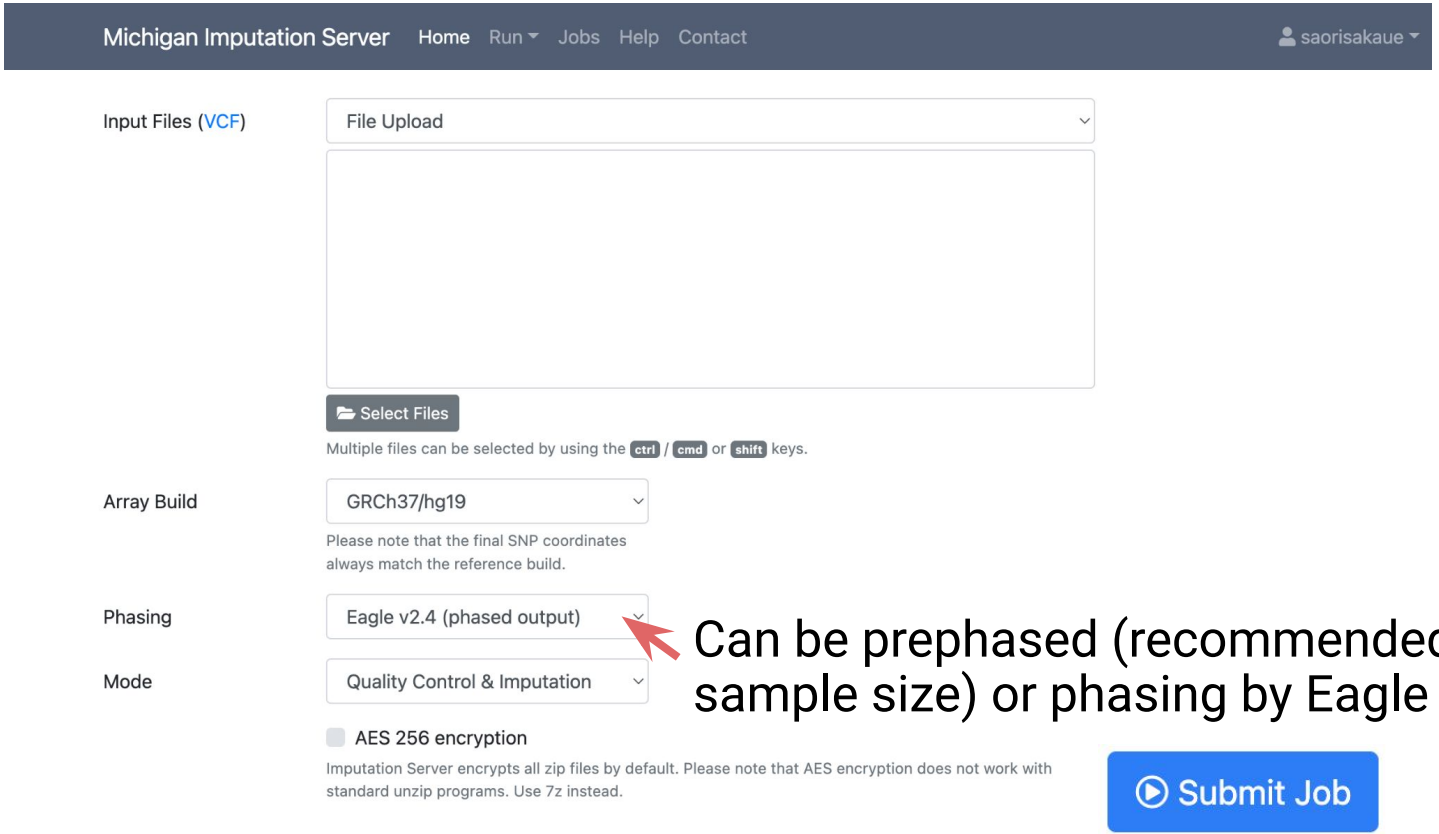


Latest HLA reference panel

HLA imputation in MIS

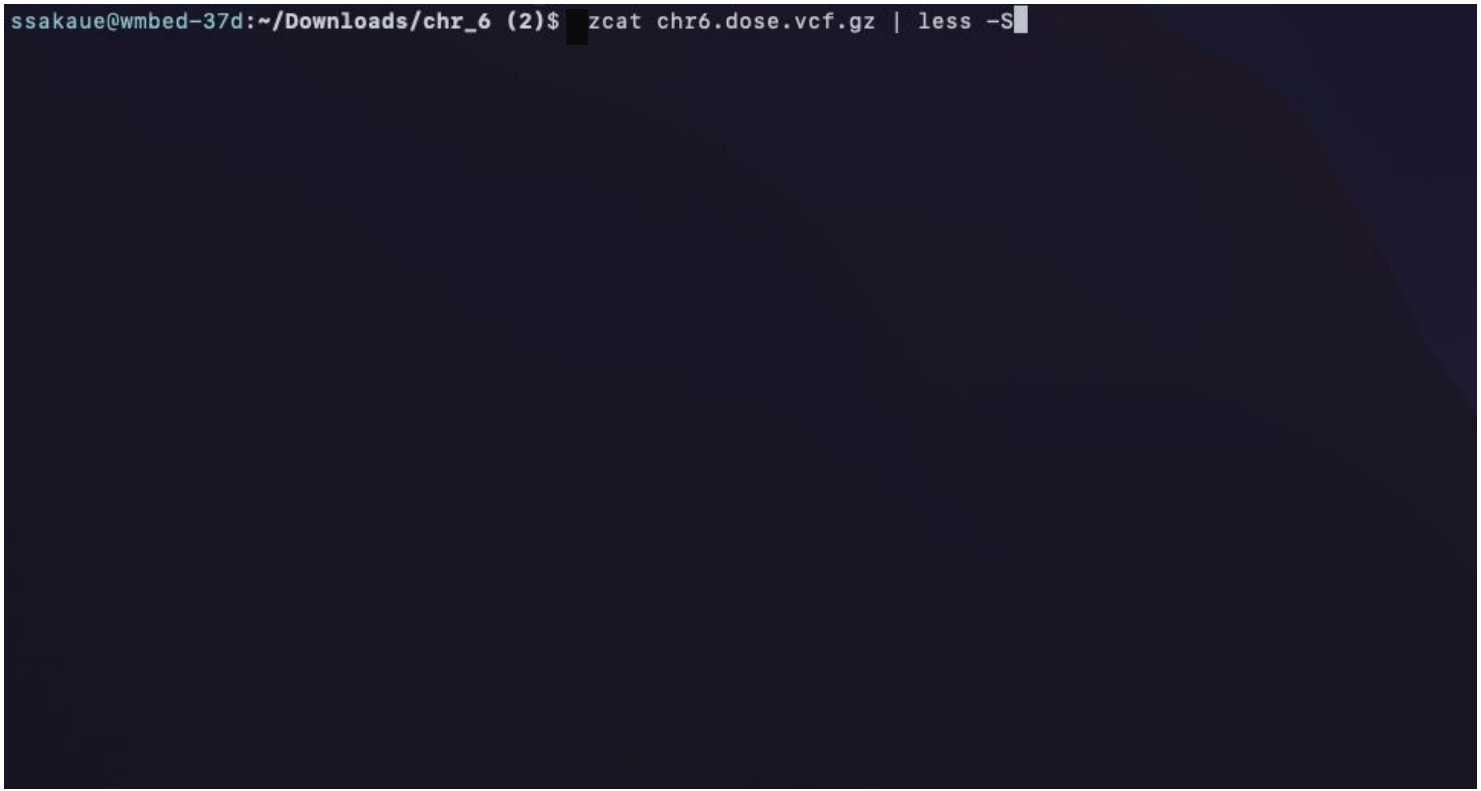


HLA imputation in MIS



Can be prephased (recommended in small sample size) or phasing by Eagle at MIS

Let’s look at the output from MIS!

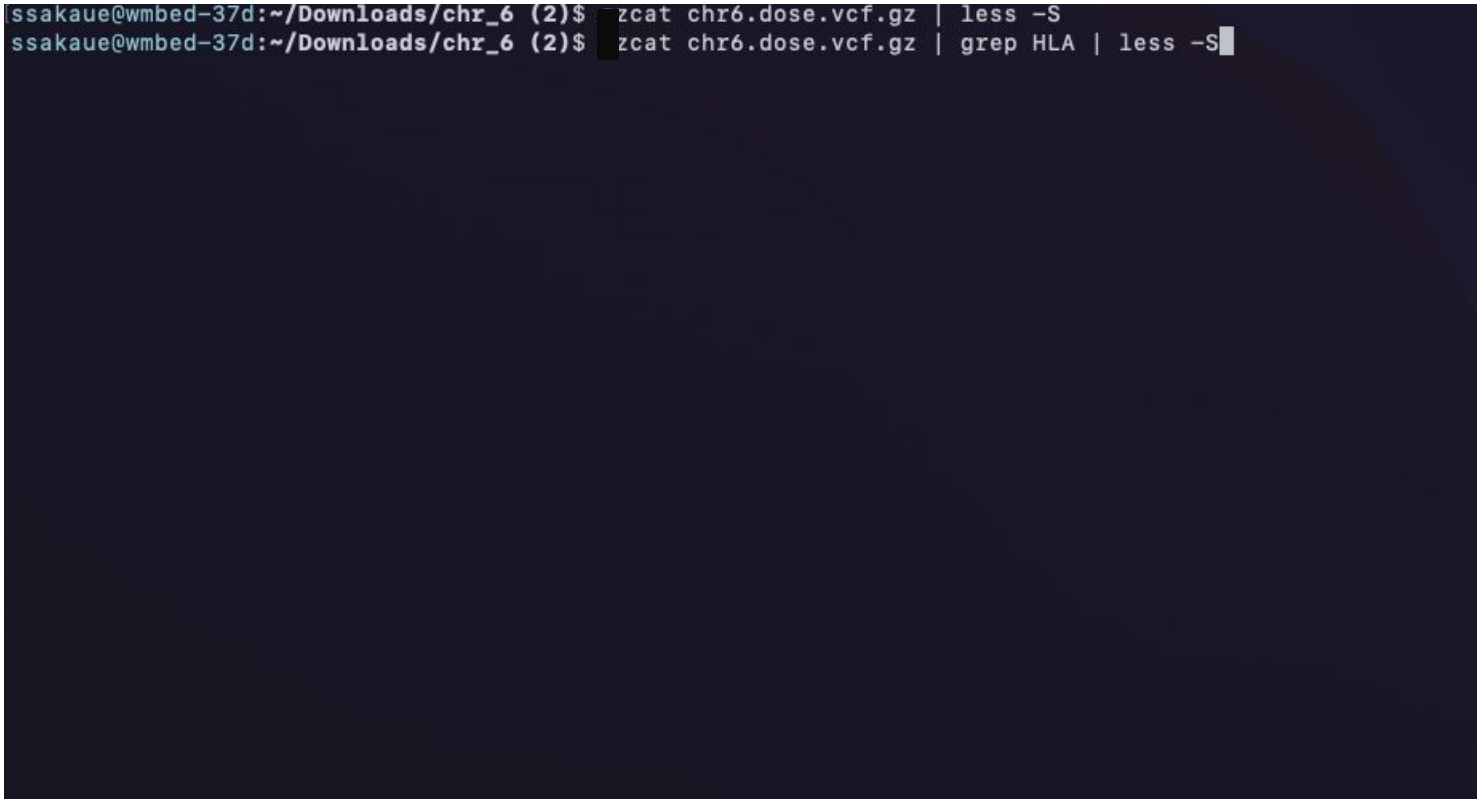


Let's look at the output from MIS!

```
##fileformat=VCFv4.1
##filedate=2022.11.14
##contig=<ID=6>
##INFO=<ID=AF,Number=1,Type=Float,Description="Estimated Alternate Allele Frequency">
##INFO=<ID=MAF,Number=1,Type=Float,Description="Estimated Minor Allele Frequency">
##INFO=<ID=R2,Number=1,Type=Float,Description="Estimated Imputation Accuracy (R-square)">
##INFO=<ID=ER2,Number=1,Type=Float,Description="Empirical (Leave-One-Out) R-square (available only for genotyped variants)">
##INFO=<ID=IMPUTED,Number=0,Type=Flag,Description="Marker was imputed but NOT genotyped">
##INFO=<ID=TYPED,Number=0,Type=Flag,Description="Marker was genotyped AND imputed">
##INFO=<ID=TYPED_ONLY,Number=0,Type=Flag,Description="Marker was genotyped but NOT imputed">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Estimated Alternate Allele Dosage : [P(0/1)+2*P(1/1)]">
##FORMAT=<ID=HDS,Number=2,Type=Float,Description="Estimated Haploid Alternate Allele Dosage ">
##FORMAT=<ID=GP,Number=3,Type=Float,Description="Estimated Posterior Probabilities for Genotypes 0/0, 0/1 and 1/1 ">
##pipeline=michigan-imputationserver-1.5.8
##imputation=minimac4-1.0.2
##phasing=eagle-2.4
##panel=apps@multiethnic-hla-panel-4digit-v2@1.0.0
##r2Filter=0.0
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  010061321010_R01C01_10007854    010061321010_R02C01_10020793    010061321010_
6       27970031        rs149946        G       T       .       PASS    AF=0.22479;MAF=0.22479;R2=0.99214;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6       27976200        rs9380032       G       T       .       PASS    AF=0.02975;MAF=0.02975;R2=0.97885;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6       27979188        rs4141691       A       G       .       PASS    AF=0.11754;MAF=0.11754;R2=0.94642;IMPUTED       GT:DS:HDS:GP    0|0:0.003:0.0
6       27979625        rs10484402      A       G       .       PASS    AF=0.04041;MAF=0.04041;R2=0.92706;IMPUTED       GT:DS:HDS:GP    0|0:0.003:0.0
6       27981673        rs9368540       G       A       .       PASS    AF=0.03706;MAF=0.03706;R2=0.98634;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6       27984726        rs74505854      A       C       .       PASS    AF=0.00719;MAF=0.00719;R2=0.94391;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6       27984907        rs17765055      T       C       .       PASS    AF=0.04632;MAF=0.04632;R2=0.99897;ER2=0.97293;TYPED     GT:DS:HDS:GP    0|0:0
6       27986199        rs72848791      C       T       .       PASS    AF=0.04361;MAF=0.04361;R2=0.99732;ER2=0.97465;TYPED     GT:DS:HDS:GP    0|0:0
6       27986529        rs9368544       A       C       .       PASS    AF=0.04631;MAF=0.04631;R2=0.99885;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6       27998258        rs149990        G       A       .       PASS    AF=0.11782;MAF=0.11782;R2=0.99974;ER2=0.99765;TYPED     GT:DS:HDS:GP    0|0:0
6       27999044        rs9368545       A       T       .       PASS    AF=0.04627;MAF=0.04627;R2=0.99825;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6       27999421        rs16893573      C       T       .       PASS    AF=0.02852;MAF=0.02852;R2=0.98601;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6       28001003        rs17708949      A       C       .       PASS    AF=0.03124;MAF=0.03124;R2=0.98465;IMPUTED       GT:DS:HDS:GP    0|0:0:0,0:1,0
6       28001610        rs149942        T       C       .       PASS    AF=0.26963;MAF=0.26963;R2=0.99868;ER2=0.98731;TYPED     GT:DS:HDS:GP    0|0:0
6       28002388        rs149943        G       A       .       PASS    AF=0.11781;MAF=0.11781;R2=0.99984;ER2=0.99889;TYPED     GT:DS:HDS:GP    0|0:0
6       28003271        rs183926        T       A       .       PASS    AF=0.00996;MAF=0.00996;R2=0.99551;ER2=0.95540;TYPED     GT:DS:HDS:GP    0|0:0
```

Imputed SNPs within MHC

Let’s look at the output from MIS!



Imputed HLA alleles

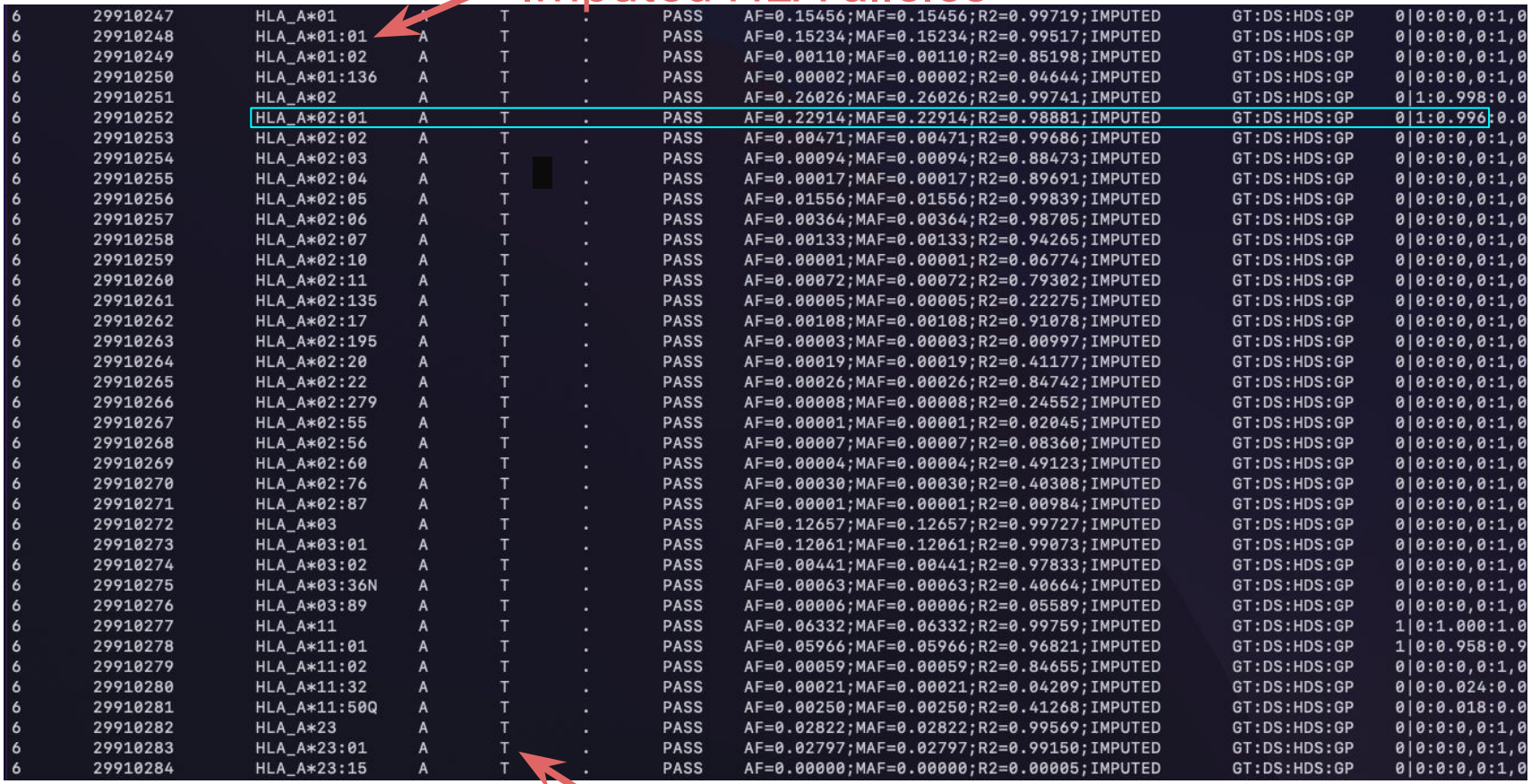
6	29910247	HLA_A*01	A	T	.	PASS	AF=0.15456;MAF=0.15456;R2=0.99719;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910248	HLA_A*01:01	A	T	.	PASS	AF=0.15234;MAF=0.15234;R2=0.99517;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910249	HLA_A*01:02	A	T	.	PASS	AF=0.00110;MAF=0.00110;R2=0.85198;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910250	HLA_A*01:136	A	T	.	PASS	AF=0.00002;MAF=0.00002;R2=0.04644;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910251	HLA_A*02	A	T	.	PASS	AF=0.26026;MAF=0.26026;R2=0.99741;IMPUTED	GT:DS:HDS:GP	0|1:0.998:0.0
6	29910252	HLA_A*02:01	A	T	.	PASS	AF=0.22914;MAF=0.22914;R2=0.98881;IMPUTED	GT:DS:HDS:GP	0|1:0.996:0.0
6	29910253	HLA_A*02:02	A	T	.	PASS	AF=0.00471;MAF=0.00471;R2=0.99686;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910254	HLA_A*02:03	A	T	.	PASS	AF=0.00094;MAF=0.00094;R2=0.88473;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910255	HLA_A*02:04	A	T	.	PASS	AF=0.00017;MAF=0.00017;R2=0.89691;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910256	HLA_A*02:05	A	T	.	PASS	AF=0.01556;MAF=0.01556;R2=0.99839;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910257	HLA_A*02:06	A	T	.	PASS	AF=0.00364;MAF=0.00364;R2=0.98705;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910258	HLA_A*02:07	A	T	.	PASS	AF=0.00133;MAF=0.00133;R2=0.94265;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910259	HLA_A*02:10	A	T	.	PASS	AF=0.00001;MAF=0.00001;R2=0.06774;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910260	HLA_A*02:11	A	T	.	PASS	AF=0.00072;MAF=0.00072;R2=0.79302;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910261	HLA_A*02:135	A	T	.	PASS	AF=0.00005;MAF=0.00005;R2=0.22275;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910262	HLA_A*02:17	A	T	.	PASS	AF=0.00108;MAF=0.00108;R2=0.91078;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910263	HLA_A*02:195	A	T	.	PASS	AF=0.00003;MAF=0.00003;R2=0.00997;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910264	HLA_A*02:20	A	T	.	PASS	AF=0.00019;MAF=0.00019;R2=0.41177;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910265	HLA_A*02:22	A	T	.	PASS	AF=0.00026;MAF=0.00026;R2=0.84742;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910266	HLA_A*02:279	A	T	.	PASS	AF=0.00008;MAF=0.00008;R2=0.24552;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910267	HLA_A*02:55	A	T	.	PASS	AF=0.00001;MAF=0.00001;R2=0.02045;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910268	HLA_A*02:56	A	T	.	PASS	AF=0.00007;MAF=0.00007;R2=0.08360;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910269	HLA_A*02:60	A	T	.	PASS	AF=0.00004;MAF=0.00004;R2=0.49123;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910270	HLA_A*02:76	A	T	.	PASS	AF=0.00030;MAF=0.00030;R2=0.40308;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910271	HLA_A*02:87	A	T	.	PASS	AF=0.00001;MAF=0.00001;R2=0.00984;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910272	HLA_A*03	A	T	.	PASS	AF=0.12657;MAF=0.12657;R2=0.99727;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910273	HLA_A*03:01	A	T	.	PASS	AF=0.12061;MAF=0.12061;R2=0.99073;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910274	HLA_A*03:02	A	T	.	PASS	AF=0.00441;MAF=0.00441;R2=0.97833;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910275	HLA_A*03:36N	A	T	.	PASS	AF=0.00063;MAF=0.00063;R2=0.40664;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910276	HLA_A*03:89	A	T	.	PASS	AF=0.00006;MAF=0.00006;R2=0.05589;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910277	HLA_A*11	A	T	.	PASS	AF=0.06332;MAF=0.06332;R2=0.99759;IMPUTED	GT:DS:HDS:GP	1|0:1.000:1.0
6	29910278	HLA_A*11:01	A	T	.	PASS	AF=0.05966;MAF=0.05966;R2=0.96821;IMPUTED	GT:DS:HDS:GP	1|0:0.958:0.9
6	29910279	HLA_A*11:02	A	T	.	PASS	AF=0.00059;MAF=0.00059;R2=0.84655;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910280	HLA_A*11:32	A	T	.	PASS	AF=0.00021;MAF=0.00021;R2=0.04209;IMPUTED	GT:DS:HDS:GP	0|0:0.024:0.0
6	29910281	HLA_A*11:50Q	A	T	.	PASS	AF=0.00250;MAF=0.00250;R2=0.41268;IMPUTED	GT:DS:HDS:GP	0|0:0.018:0.0
6	29910282	HLA_A*23	A	T	.	PASS	AF=0.02822;MAF=0.02822;R2=0.99569;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910283	HLA_A*23:01	A	T	.	PASS	AF=0.02797;MAF=0.02797;R2=0.99150;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910284	HLA_A*23:15	A	T	.	PASS	AF=0.00000;MAF=0.00000;R2=0.00005;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0

REF ALT

"T": Presence of the allele

"A": Absence of the allele

Imputed HLA alleles



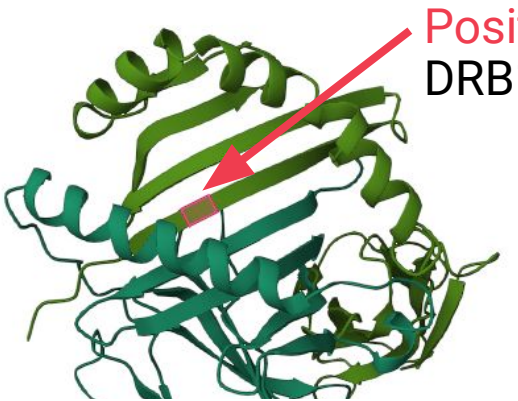
6	29910247	HLA_A*01	A	T	.	PASS	AF=0.15456;MAF=0.15456;R2=0.99719;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910248	HLA_A*01:01	A	T	.	PASS	AF=0.15234;MAF=0.15234;R2=0.99517;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910249	HLA_A*01:02	A	T	.	PASS	AF=0.00110;MAF=0.00110;R2=0.85198;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910250	HLA_A*01:136	A	T	.	PASS	AF=0.00002;MAF=0.00002;R2=0.04644;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910251	HLA_A*02	A	T	.	PASS	AF=0.26026;MAF=0.26026;R2=0.99741;IMPUTED	GT:DS:HDS:GP	0|1:0.998:0.0
6	29910252	HLA_A*02:01	A	T	.	PASS	AF=0.22914;MAF=0.22914;R2=0.98881;IMPUTED	GT:DS:HDS:GP	0|1:0.996:0.0
6	29910253	HLA_A*02:02	A	T	.	PASS	AF=0.00471;MAF=0.00471;R2=0.99686;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910254	HLA_A*02:03	A	T	.	PASS	AF=0.00094;MAF=0.00094;R2=0.88473;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910255	HLA_A*02:04	A	T	.	PASS	AF=0.00017;MAF=0.00017;R2=0.89691;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910256	HLA_A*02:05	A	T	.	PASS	AF=0.01556;MAF=0.01556;R2=0.99839;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910257	HLA_A*02:06	A	T	.	PASS	AF=0.00364;MAF=0.00364;R2=0.98705;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910258	HLA_A*02:07	A	T	.	PASS	AF=0.00133;MAF=0.00133;R2=0.94265;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910259	HLA_A*02:10	A	T	.	PASS	AF=0.00001;MAF=0.00001;R2=0.06774;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910260	HLA_A*02:11	A	T	.	PASS	AF=0.00072;MAF=0.00072;R2=0.79302;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910261	HLA_A*02:135	A	T	.	PASS	AF=0.00005;MAF=0.00005;R2=0.22275;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910262	HLA_A*02:17	A	T	.	PASS	AF=0.00108;MAF=0.00108;R2=0.91078;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910263	HLA_A*02:195	A	T	.	PASS	AF=0.00003;MAF=0.00003;R2=0.00997;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910264	HLA_A*02:20	A	T	.	PASS	AF=0.00019;MAF=0.00019;R2=0.41177;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910265	HLA_A*02:22	A	T	.	PASS	AF=0.00026;MAF=0.00026;R2=0.84742;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910266	HLA_A*02:279	A	T	.	PASS	AF=0.00008;MAF=0.00008;R2=0.24552;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910267	HLA_A*02:55	A	T	.	PASS	AF=0.00001;MAF=0.00001;R2=0.02045;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910268	HLA_A*02:56	A	T	.	PASS	AF=0.00007;MAF=0.00007;R2=0.08360;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910269	HLA_A*02:60	A	T	.	PASS	AF=0.00004;MAF=0.00004;R2=0.49123;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910270	HLA_A*02:76	A	T	.	PASS	AF=0.00030;MAF=0.00030;R2=0.40308;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910271	HLA_A*02:87	A	T	.	PASS	AF=0.00001;MAF=0.00001;R2=0.00984;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910272	HLA_A*03	A	T	.	PASS	AF=0.12657;MAF=0.12657;R2=0.99727;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910273	HLA_A*03:01	A	T	.	PASS	AF=0.12061;MAF=0.12061;R2=0.99073;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910274	HLA_A*03:02	A	T	.	PASS	AF=0.00441;MAF=0.00441;R2=0.97833;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910275	HLA_A*03:36N	A	T	.	PASS	AF=0.00063;MAF=0.00063;R2=0.40664;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910276	HLA_A*03:89	A	T	.	PASS	AF=0.00006;MAF=0.00006;R2=0.05589;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910277	HLA_A*11	A	T	.	PASS	AF=0.06332;MAF=0.06332;R2=0.99759;IMPUTED	GT:DS:HDS:GP	1|0:1.000:1.0
6	29910278	HLA_A*11:01	A	T	.	PASS	AF=0.05966;MAF=0.05966;R2=0.96821;IMPUTED	GT:DS:HDS:GP	1|0:0.958:0.9
6	29910279	HLA_A*11:02	A	T	.	PASS	AF=0.00059;MAF=0.00059;R2=0.84655;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910280	HLA_A*11:32	A	T	.	PASS	AF=0.00021;MAF=0.00021;R2=0.04209;IMPUTED	GT:DS:HDS:GP	0|0:0.024:0.0
6	29910281	HLA_A*11:50Q	A	T	.	PASS	AF=0.00250;MAF=0.00250;R2=0.41268;IMPUTED	GT:DS:HDS:GP	0|0:0.018:0.0
6	29910282	HLA_A*23	A	T	.	PASS	AF=0.02822;MAF=0.02822;R2=0.99569;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910283	HLA_A*23:01	A	T	.	PASS	AF=0.02797;MAF=0.02797;R2=0.99150;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0
6	29910284	HLA_A*23:15	A	T	.	PASS	AF=0.00000;MAF=0.00000;R2=0.00005;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0

REF ALT

"T": Presence of the allele

"A": Absence of the allele

How are imputed HLA variants associated with disease?



Disease ↔ ?

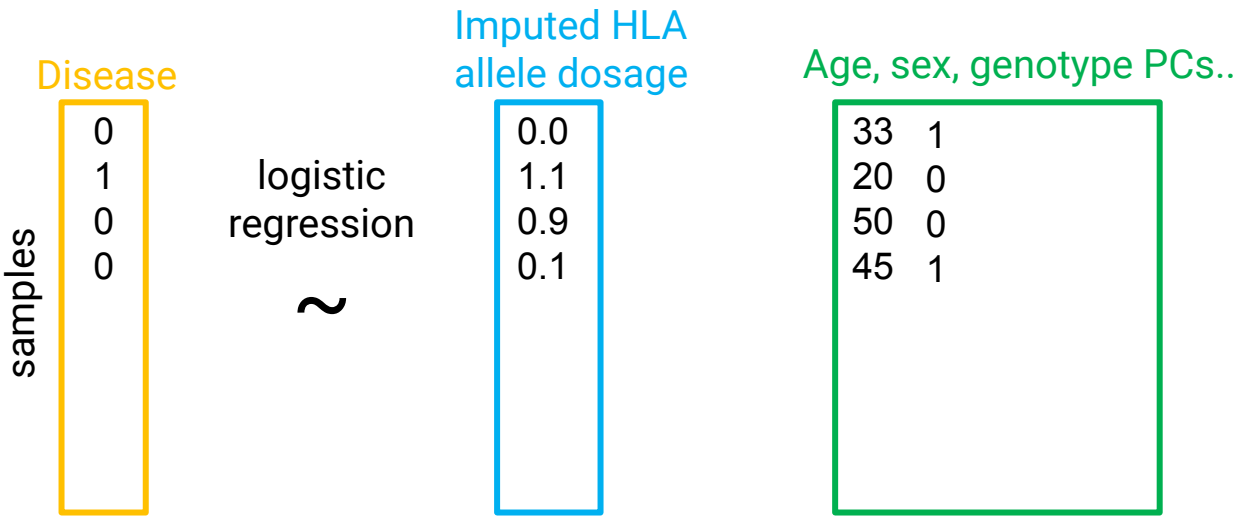
1 HLA alleles — **2 HLA amino acid positions (Omnibus test)**

```
HLA-DRB1*01:01 ⟷ GGSCMAALTVTLMVLSSP
HLA-DRB1*04:01 ⟷ GGSCMTALTVTLMVLSSP
HLA-DRB1*15:01 ⟷ GGSCMTALTVTLMVLSSP
```

3 HLA haplotypes (Conditional haplotype test)

Single-marker test to ask which HLA allele affects the disease

$$log(odds_i) = \beta_0 + \beta_a g_{a,i} + \sum_k \beta_k covariate_{k,i}$$



Omnibus test to ask which HLA amino acid position affects the disease

Full model: $log(odds_i) = \beta_0 + \sum_k \beta_k covariate_{k,i} + \sum_{m=1}^{M-1} \beta_m AM_{m,i}$

Reduced model: $log(odds_i) = \beta_0 + \sum_k \beta_k covariate_{k,i}$

ANOVA

HLA-DRB1

pos 11

RFLWQLKFECH
RFLEYSTSECH
RFLEQVKHECH
RFLWQGKYKCH
RFLKQDKFECH
RFLWQPKRECH

M = 6 possible amino acid residues

M - 1

	L	S	V	G	D	(P)
	0.1	0.0	0.9	0.0	0.0	1.0
	0.9	0.0	0.0	0.0	1.0	0.1
	0.0	1.0	0.0	0.9	0.0	0.1
	0.0	0.0	1.9	0.0	0.1	0.0
	2.0	0.0	0.0	0.0	0.0	0.0

Omnibus test to ask which HLA amino acid position affects the disease

Full model: $log(odds_i) = \beta_0 + \sum_k \beta_k covariate_{k,i} + \sum_{m=1}^{M-1} \beta_m AM_{m,i}$

Reduced model: $log(odds_i) = \beta_0 + \sum_k \beta_k covariate_{k,i}$

ANOVA(Full model, Reduced model)

How much does this amino acid position's polymorphism increase the explained variance for the trait?

→ Determine the single most significant amino acid position

Summary

1. HLA amino acid sequences and alleles characterize antigen presentation and disease risk within HLA.
2. HLA amino acid sequences and alleles can be accurately imputed from genotyped SNPs by MIS.
3. Imputed HLA alleles can be used to fine-map causal disease mechanisms.

Summary

1. HLA amino acid sequences and alleles characterize antigen presentation and disease risk within HLA.
2. HLA amino acid sequences and alleles can be accurately imputed from genotyped SNPs by MIS.
3. Imputed HLA alleles can be used to fine-map causal disease mechanisms.

Sakaue et al. *Nature Protocols* 2023

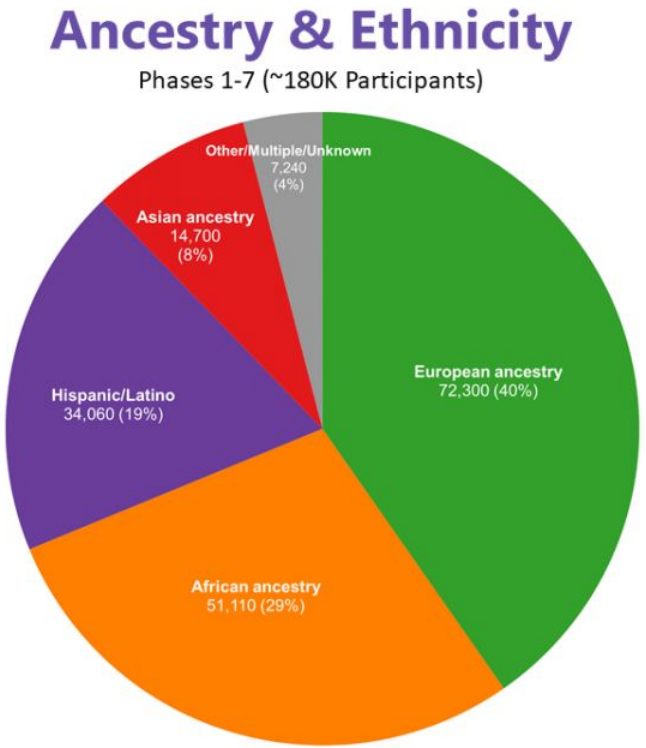
Section 6

The TOPMed Imputation Server

Albert Smith
University of Michigan
albertvs@umich.edu
@avsmith

TOPMed Program

- Trans-Omics for Precision Medicine (TOPMed) Program
- A Precision Medicine Initiative sponsored by National Heart, Lung and Blood Institute
- Integrating whole-genome sequencing and other omics data
- >180k participants from >90 studies



TOPMed Imputation

- Current reference panel based on TOPMed Freeze 8 Calls
- Michigan Imputation Server ported to Amazon Web Services
- Released April 2020
- https://imputation.biodatacatalyst.nhlbi.nih.gov
- Registration as before, open access to TOPMed panel

TOPMed Panel

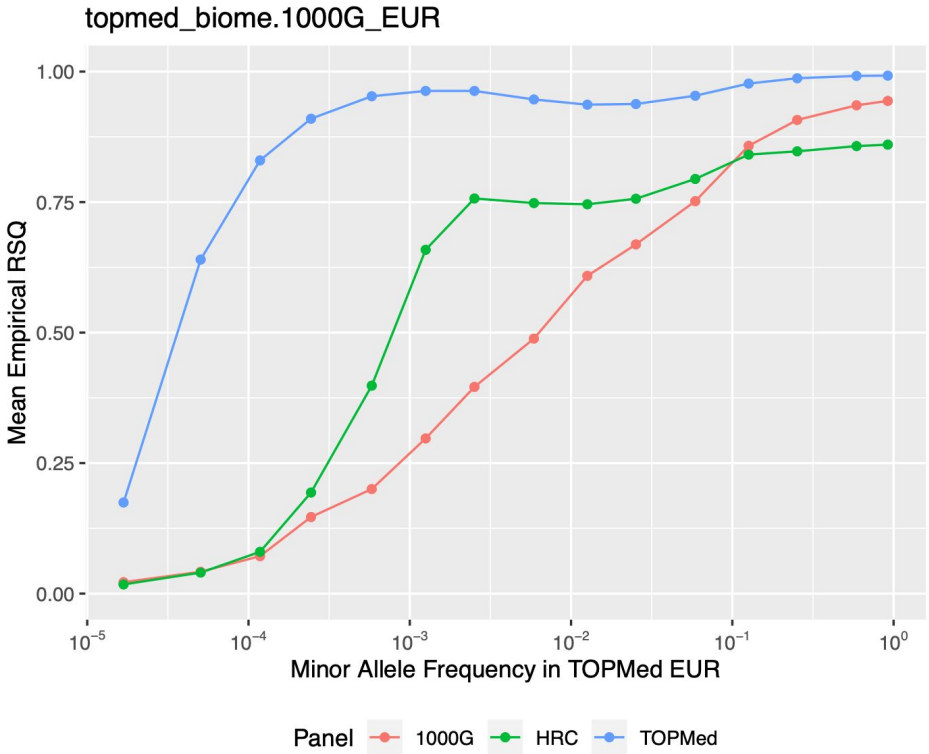
Variation type	Non-reference allele frequency bins				
	(0, 0.005]	(0.005, 0.01]	(0.01, 0.05]	(0.05, 1)	Totals
SNVs	270,352,495	3,365,284	5,330,340	7,020,861	286,068,980
Insertions	5,462,262	74,150	130,506	148,595	5,815,513
Deletions	15,406,052	185,606	297,186	333,748	16,222,592
Totals	**291,220,809**	**3,625,040**	**5,758,032**	**7,503,204**	**308,107,085**

Panel based on TOPMed Freeze 8

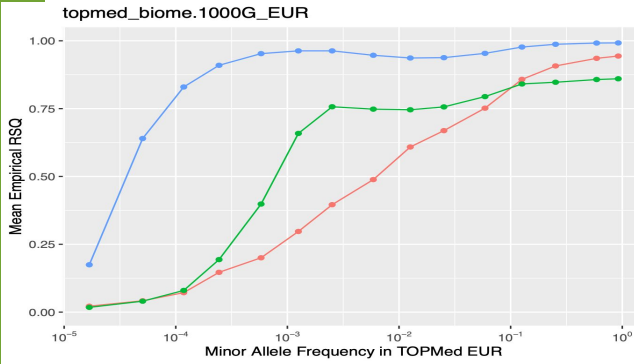
TOPMed Panel Compared

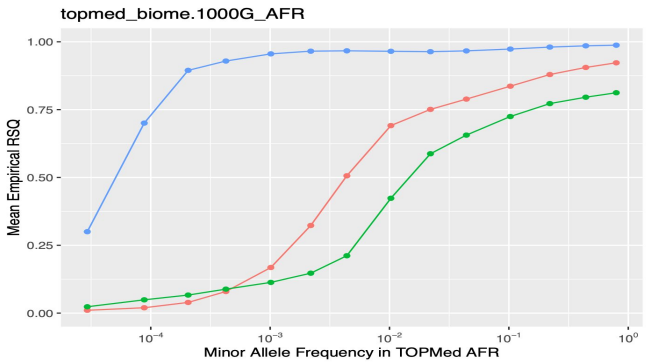
	TOPMed_r2	HRC	1000G Genomes
N samples	97K	39K	2,500
Ancestry	Multiethnic	European	Multiethnic
N variants	308M	39M	88M
Avg. depth	38X	8X	4X
Genome build	b38	b37	b37

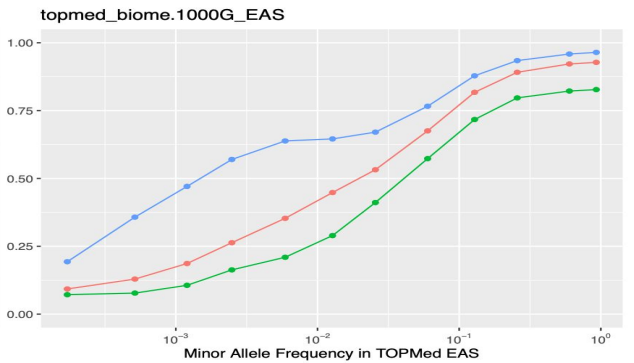
Imputation Panel Quality

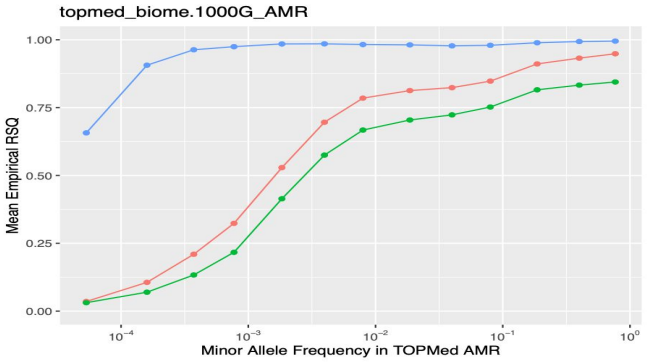


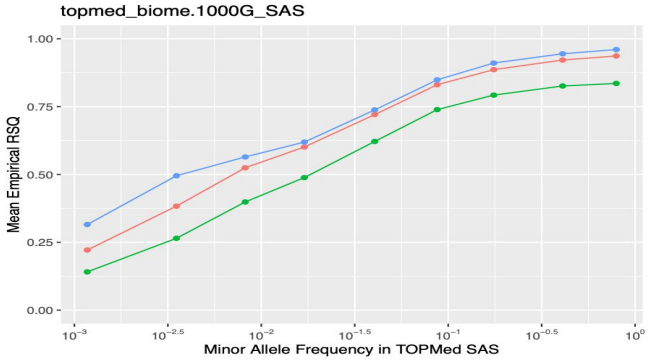
Imputation Panel Quality





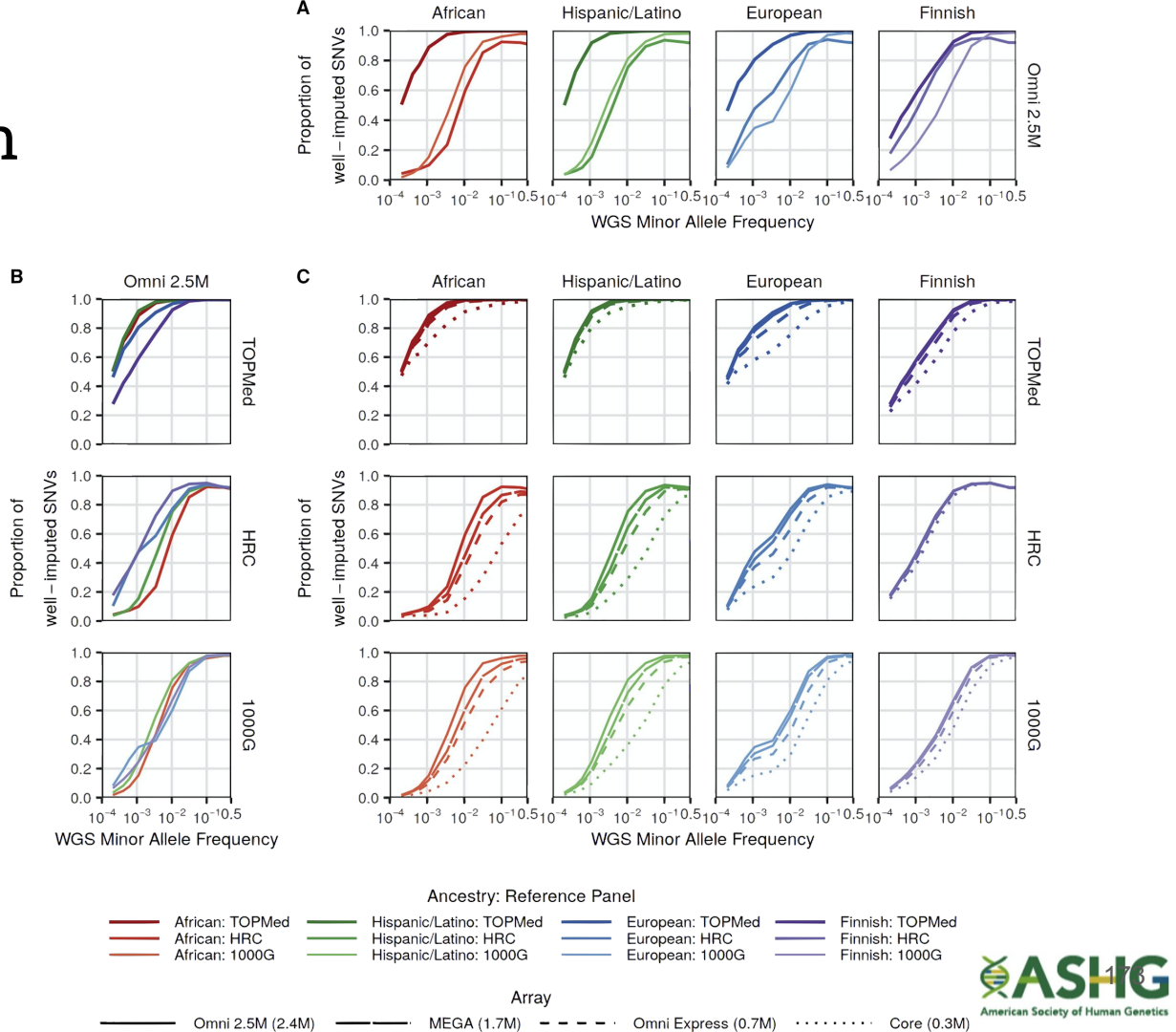


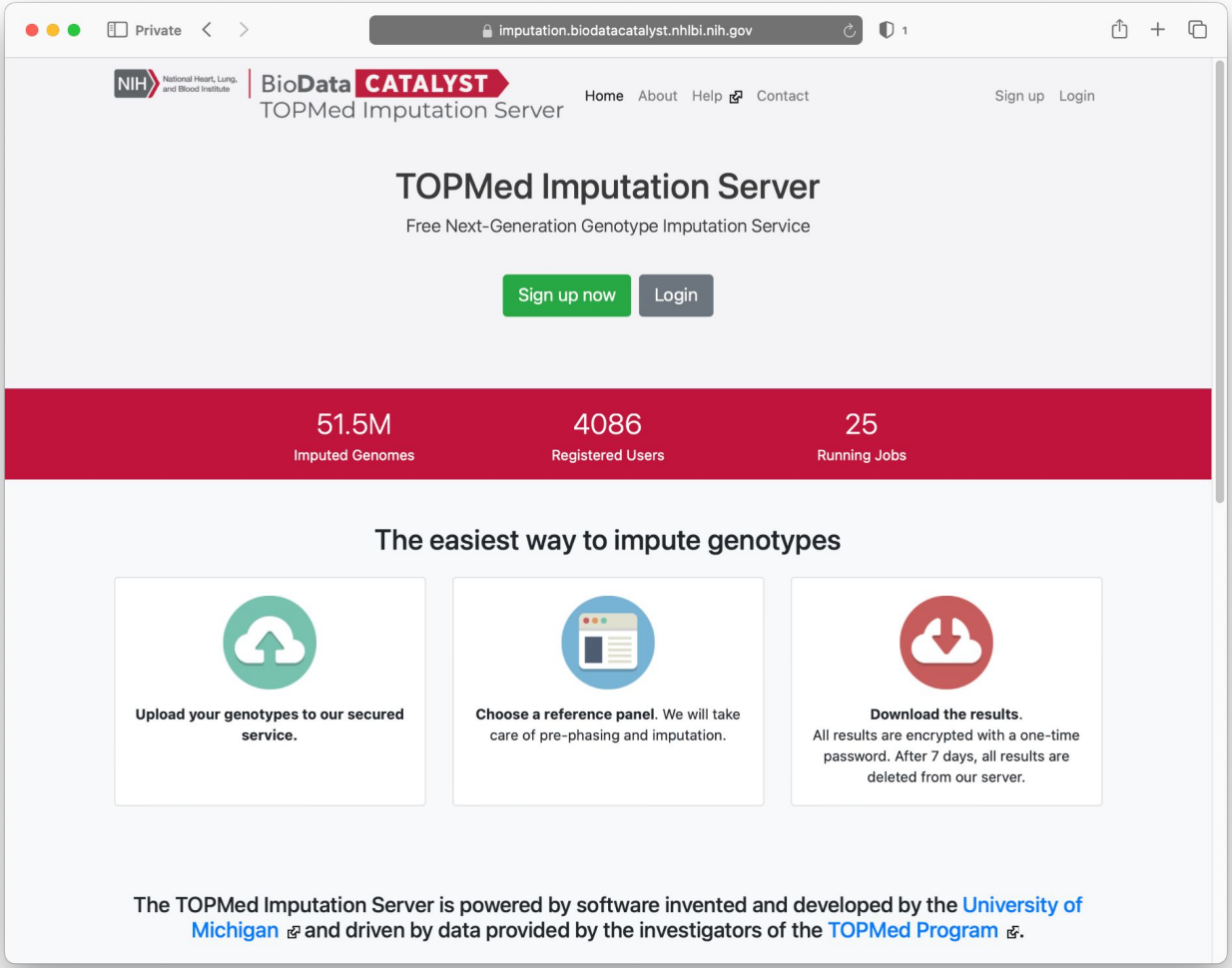


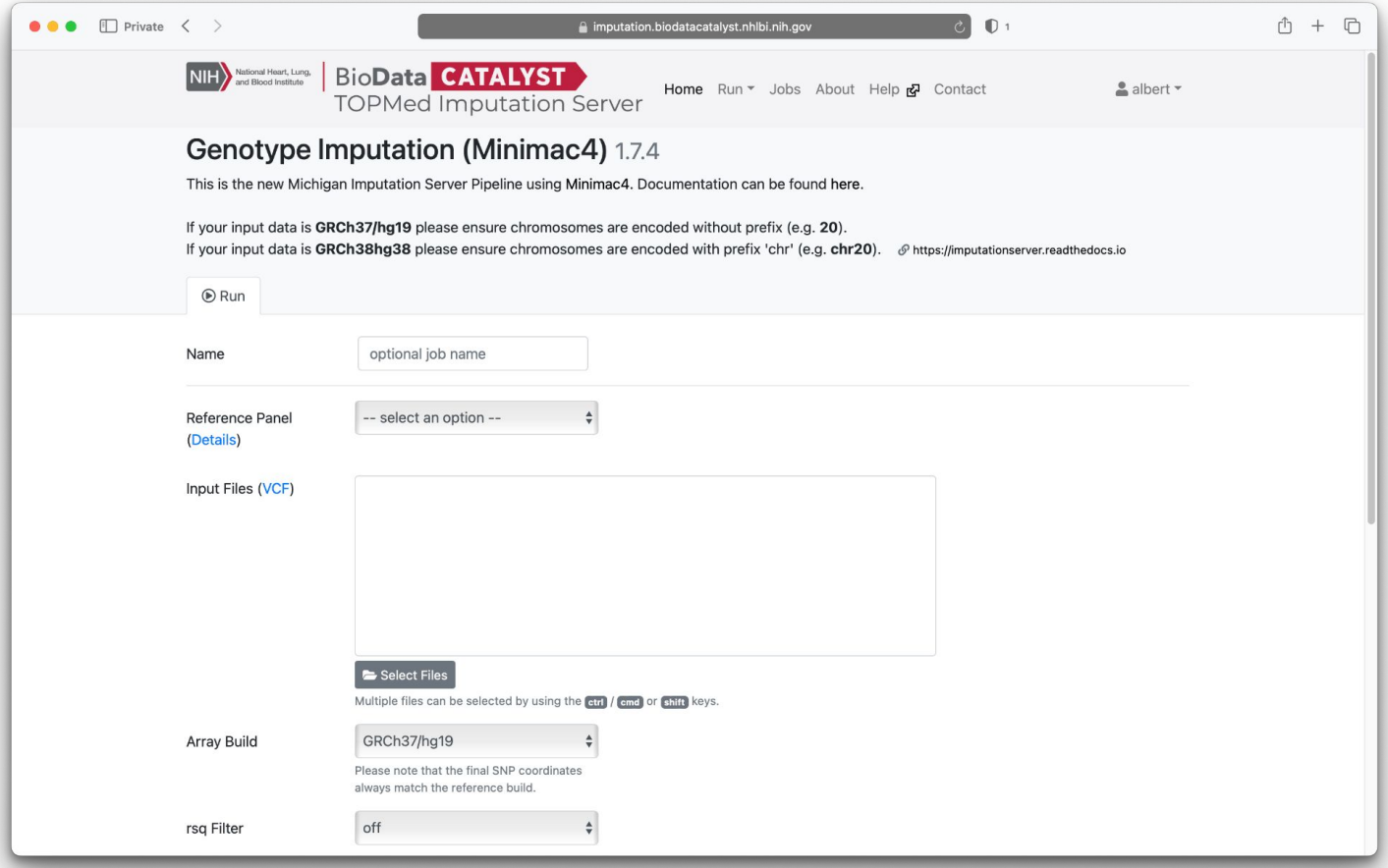


TOPMed Imputation Compared to WGS

- Proportion well imputed (r2 > 0.8) down to MAF:
  - 0.14% in African
  - 0.11% in Hispanic/Latino
  - 0.35% in European
  - 0.85% in Finnish
- Similar performance for arrays with >700k variants
- Source: Hanks et al. https://doi.org/10.1016/j.ajhg.2022.07.012

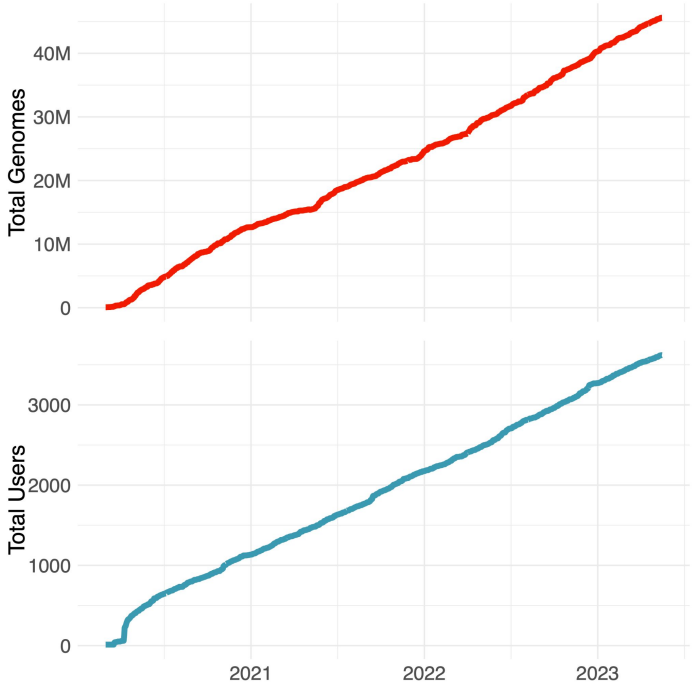






TOPMed Imputation

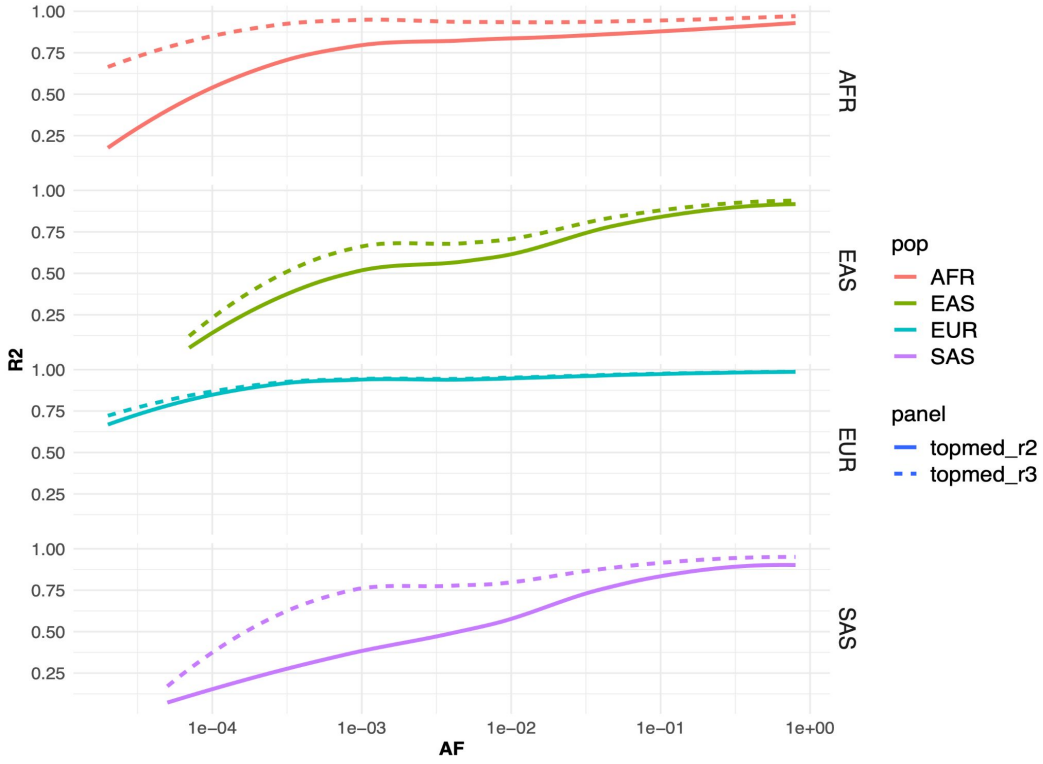
- 52M genomes imputed
- Largely supplanted 1000g & HRC imputation
- Particularly benefits ethnically diverse cohorts
- Satisfying GDPR-related concerns of European users remains a challenge



ASHG
American Society of Human Genetics

Updated TOPMed Panel

- Existing panel has better coverage for EUR, AFR and AMR samples
- Expanded panel developed (97k => 143k)
- Targeted improvement for SAS and EAS samples
- To be released very soon



Imputation Resources

- Michigan Imputation Server

  https://imputation.sph.umich.edu/
- TOPMed Imputation Server

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

  https://imputationserver.readthedocs.io/
  https://topmedimpute.readthedocs.io/
- TOPMed Imputation Contact

  imputationserver@umich.edu

I would like to see next

Gene expression imputation 0%

GWAS analysis 0%

More documentation 0%

Low-pass Imputation 0%

Single-sample Imputation 0%

I would like to see next

Nobody has responded yet.

Hang tight! Responses are coming in.

Your questions

If we run out of time, please send your questions to

Christian - cfuchsb@umich.edu

Thank you!