MICHIGAN IMPUTATIONSERVER

Workshop ASHG 2019

Christian Fuchsberger

Cassandra N. Spracklen

Daniel Taliun

Lukas Forer

Sebastian Schönherr

Sarah Gagliano Taliun

Financial Disclosure for

ASHG 2019

The Michigan Imputation Server: data preparation, genotype imputation, and data analysis

- No financial relationships to disclose:

Christian Fuchsberger
Cassandra N. Spracklen
Daniel Taliun
Lukas Forer
Sebastian Schönherr
Sarah Gagliano Taliun

MICHIGAN IMPUTATIONSERVER

ASHG 2019 ANNUAL MEETING HOUSTON • OCTOBER 15-19

1.) Workshop Wi-Fi:

1. Select the SSID:
   ASHGWORKSHOP
2. Enter password:
   ASHGWORKSHOP
   (Password is case-sensitive)

2.) Interactive Polls during Workshop:

https://imputationserver.sph.umich.edu/poll

For App users: www.pollev.com/**ashg19**

3.) Imputation Server Playground:

https://imputationserver.sph.umich.edu/aws

Learning objectives

Participants will

1. Understand the principles of genotype imputation and the Imputation Server
2. Know the steps needed for proper Quality Control
3. Understand the various server settings
4. Be able to track imputation runs
5. Know how to download data and prepare for analysis
6. Know how to use the imputed data to perform a GWAS

Setup

- 5 Sessions
  - Lectures
  - Demos
  - Interactive Polls
- Question & Answer session at the end

Session 1

Imputation and the Server

Christian Fuchsberger

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

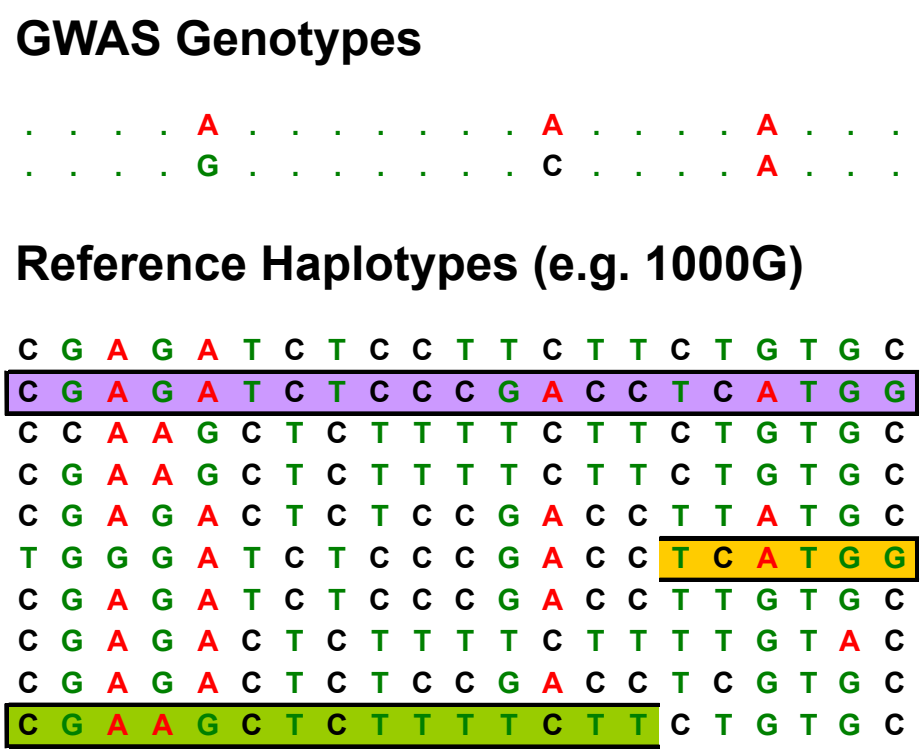
GWAS Genotypes

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

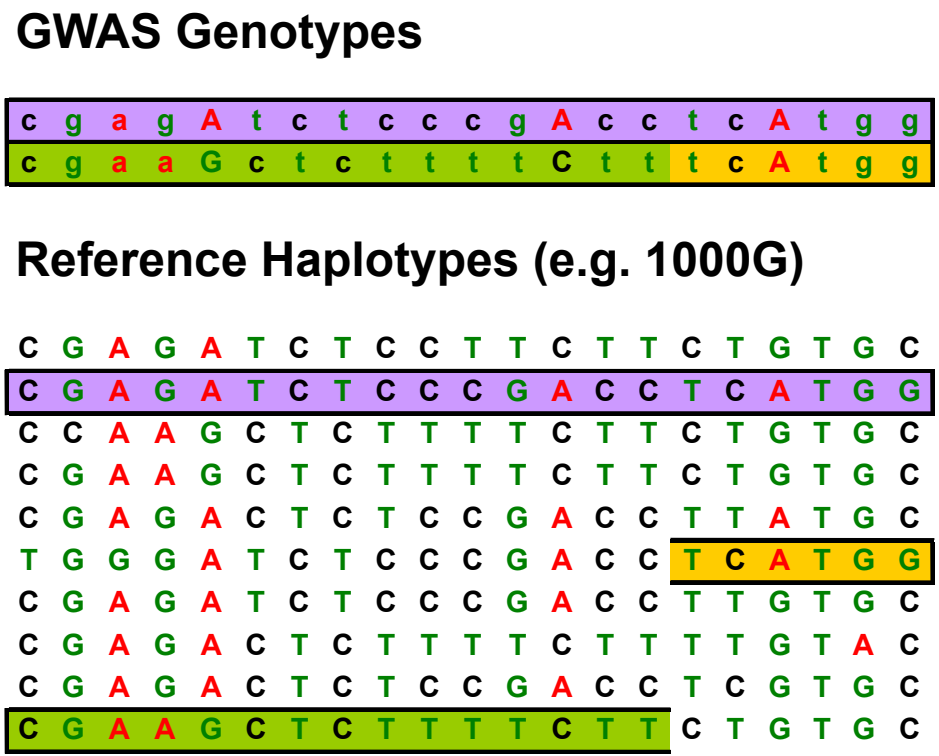
Reference Haplotypes (e.g. 1000G)

```
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
```

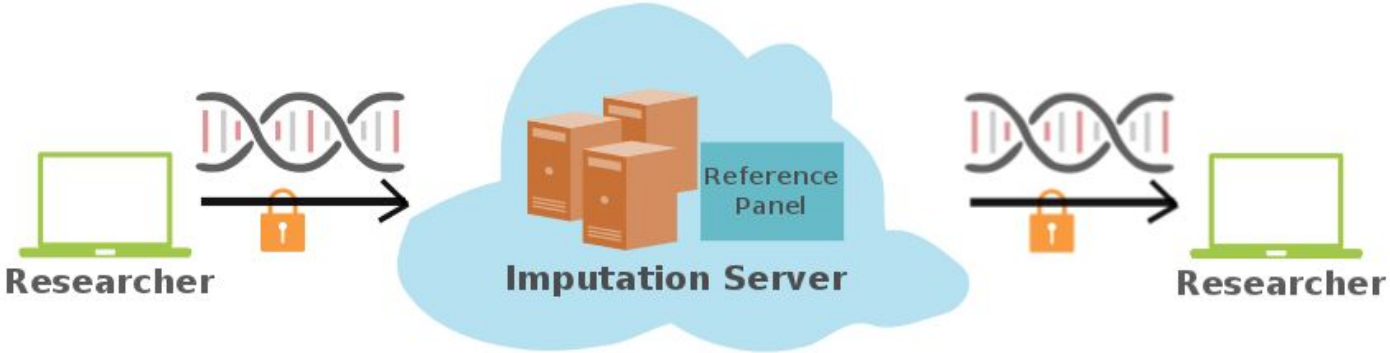
1. Identify match among reference



2. Impute


GWAS Genotypes
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. 1000G)
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

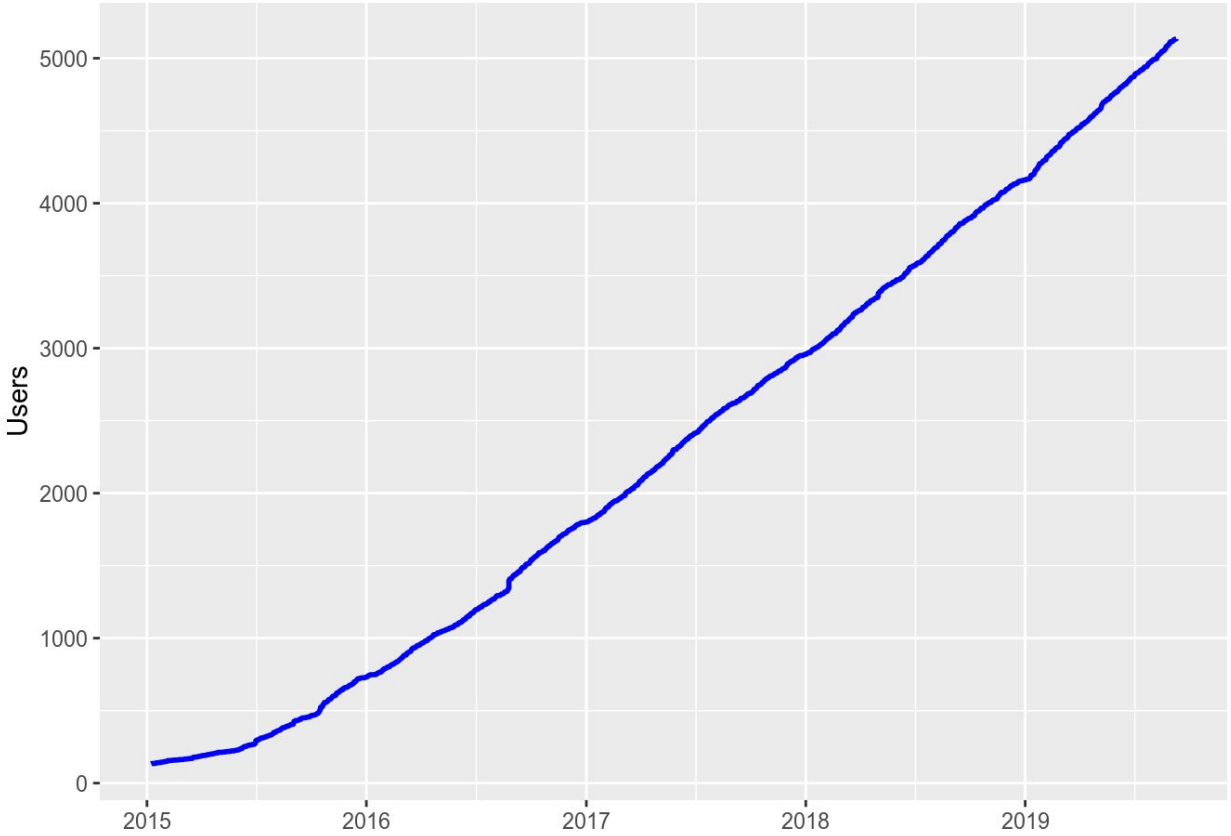
- Quality checks
- Pre-phasing
- Imputation
- Encryption

3.

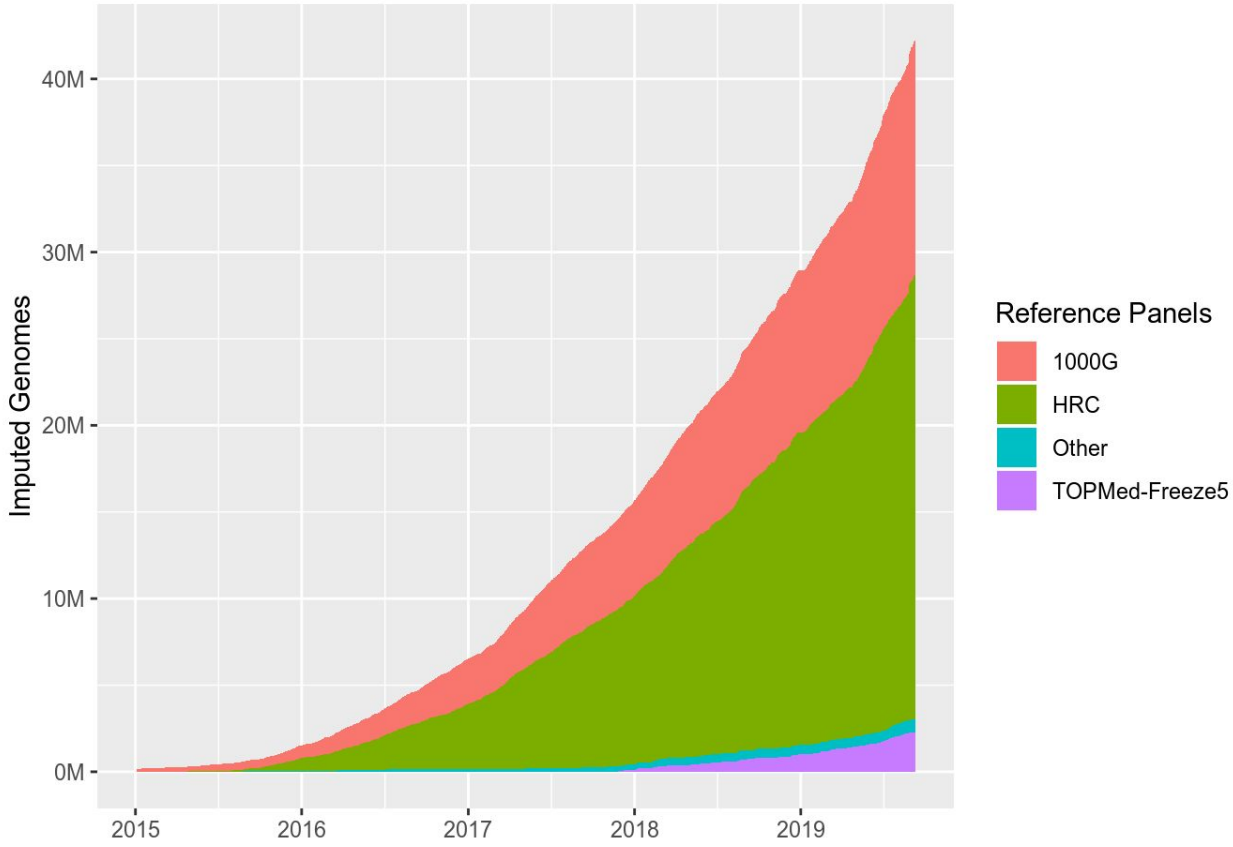
Download results

Have you ever used the Michigan Imputation Server

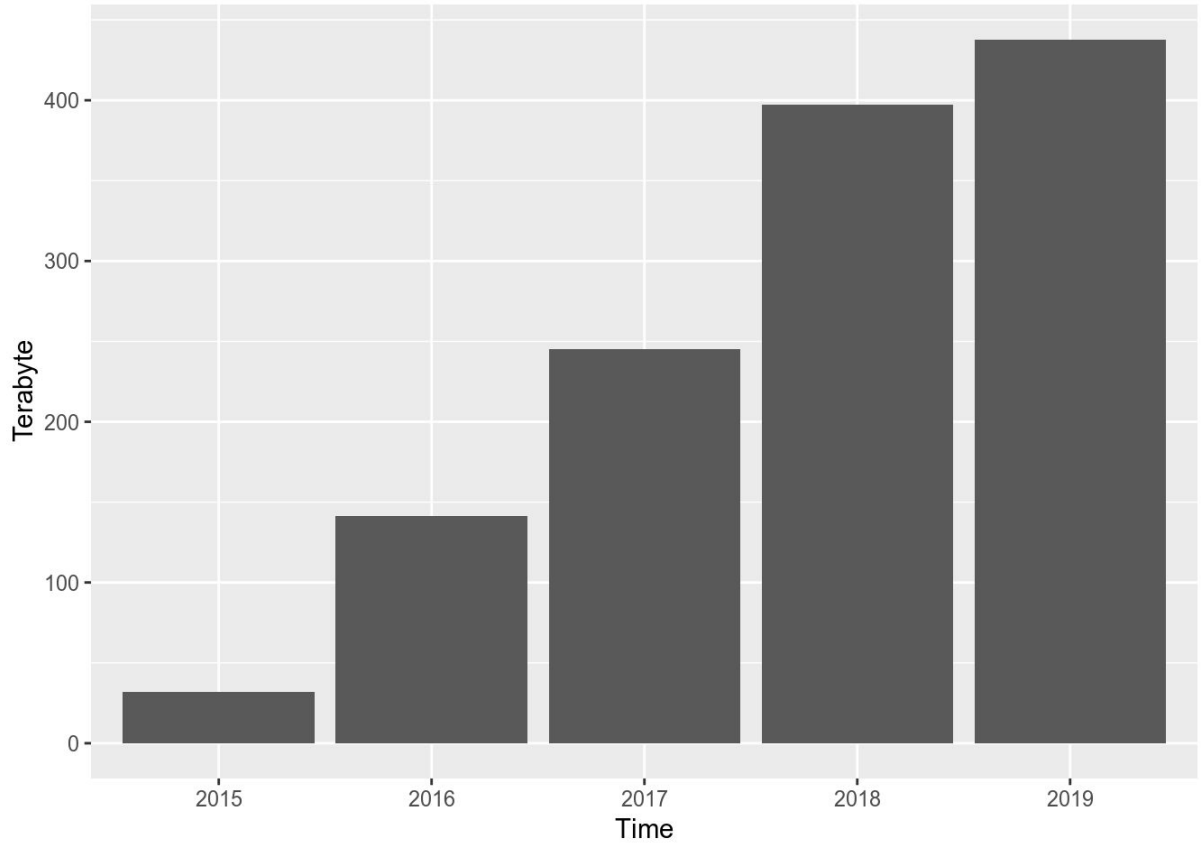
>5,000 users



>45M imputed genomes



>1 PB data delivered



Session 2

Run a job, Quality Control and Data Preparation

Sebastian Schönherr

Submit Your First Imputation Job

Genotype Imputation (Minimac4) 1.2.1

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

Optional Job Name

Reference Panel (Details): -- select an option --

Reference panel?
HapMap, 1000G Phase 1, 1000G Phase 3, HRC, Genome Asia

Input Files (VCF): File Upload

Where is your data located?
local, sftp, https, AWS S3,

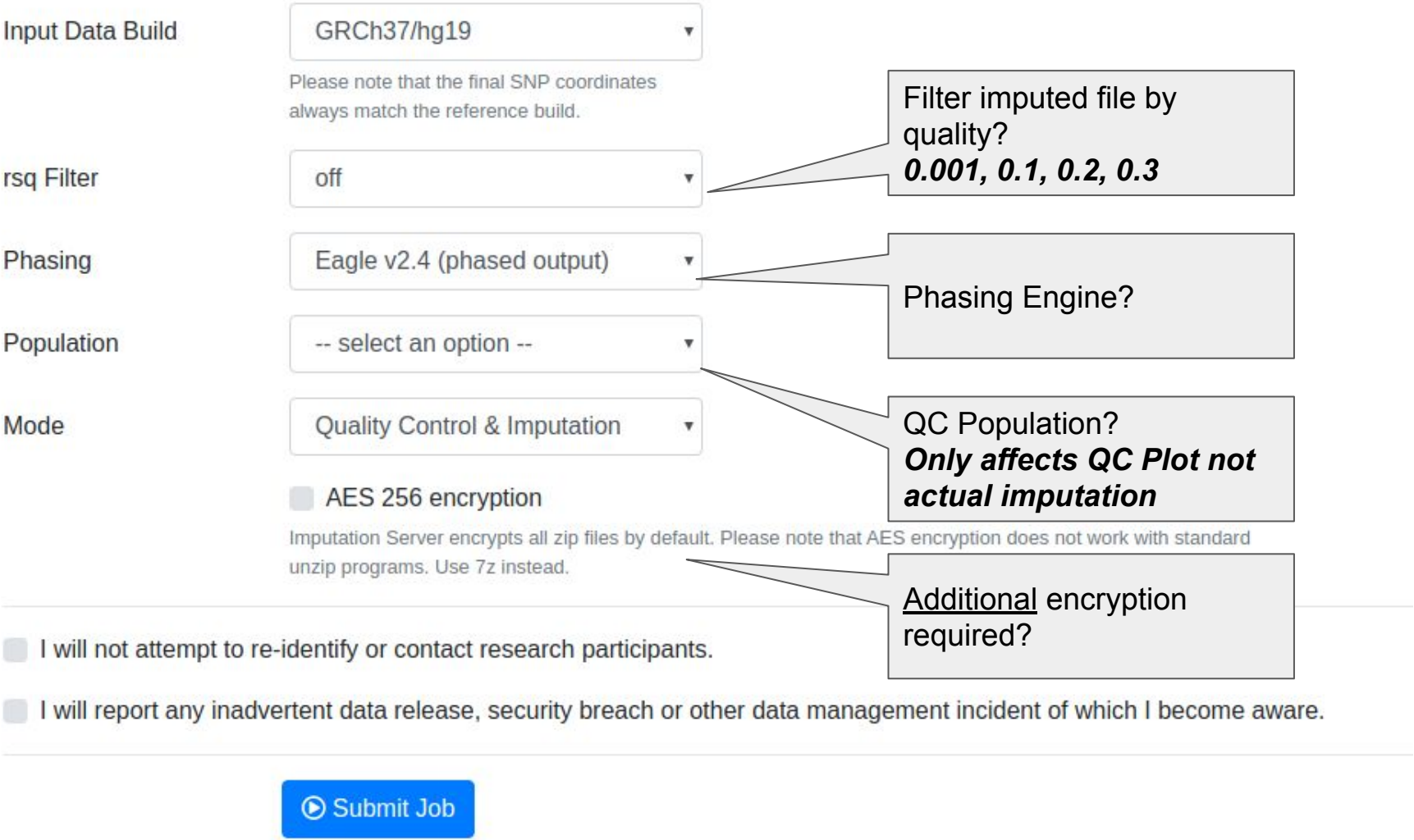
Select Files

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

Input Data Build: GRCh37/hg19

Please note that the final SNP coordinates always match the reference build.

What are the coordinates of your input data?
b37 vs b38


Input Data 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
Filter imputed file by quality?
0.001, 0.1, 0.2, 0.3
Phasing Engine?
QC Population?
Only affects QC Plot not actual imputation
Additional encryption required?

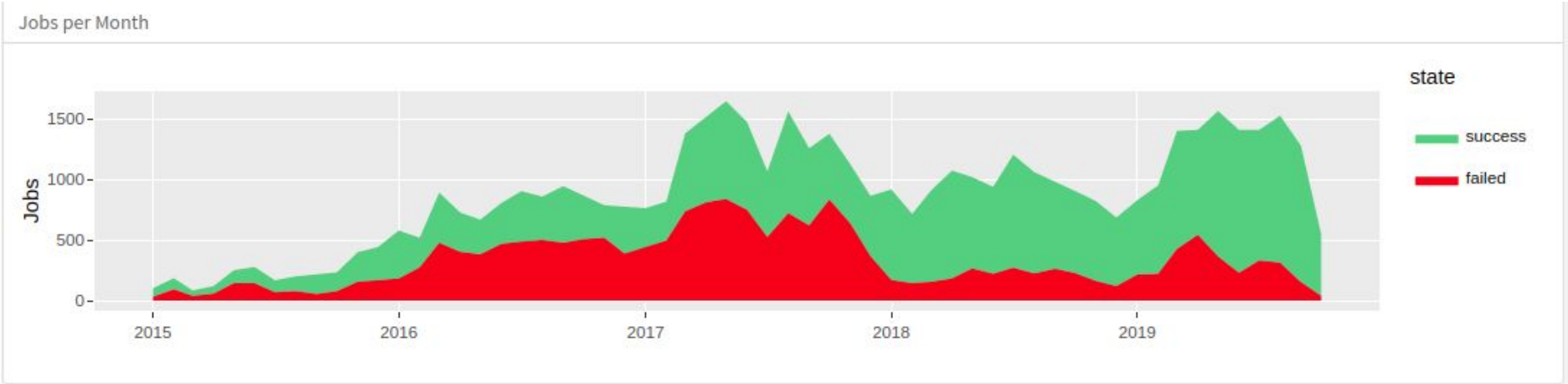
Did you run into QC Problems so far?

Yes

No

How many jobs are not passing QC?

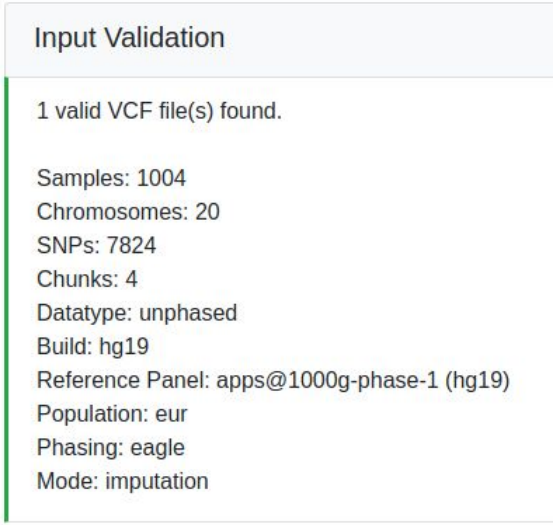
- 1/3 of all jobs
- Reason: Something wrong with your input data
- Imputation Server helps to identify problems



MIS QC: Helps to identify problems

-

	Imputation Server
Input	**VCF / chromosome**
Output	Imputed VCF / chromosome
File Validation & Statistics	
Basic SNP Filtering	
Fixes Strand Errors	
Updating Ref / Alt Assignment	
Removes SNPs with allele frequency difference	



MIS QC: Helps to identify problems

-

	Imputation Server
Input	**VCF / chromosome**
Output	Imputed VCF / chromosome
File Validation & Statistics	
Basic SNP Filtering	
Fixes Strand Errors	
Updating Ref / Alt Assignment	
Removes SNPs with allele frequency difference	

Statistics:
Alternative allele frequency > 0.5 sites: 2,308
Reference Overlap: 99.95 %
Match: 7,816
Allele switch: 0
Strand flip: 0
Strand flip and allele switch: 0
A/T, C/G genotypes: 0
Filtered sites:
Filter flag set: 0
Invalid alleles: 0
Multiallelic sites: 0
Duplicated sites: 0
NonSNP sites: 0
Monomorphic sites: 0
Allele mismatch: 4
SNPs call rate < 90%: 0

Passing QC

Filtered sites:
Filter flag set: 0
Invalid alleles: 0
Multiallelic sites: 0
Duplicated sites: 57
NonSNP sites: 0
Monomorphic sites: 0
Allele mismatch: 158
SNPs call rate < 90%: 0

Excluded sites in total: 215
Remaining sites in total: 21,818
Typed only sites: 304

Failing QC

Filtered sites:
Filter flag set: 0
Invalid alleles: 0
Multiallelic sites: 0
Duplicated sites: 0
NonSNP sites: 0
Monomorphic sites: 0
Allele mismatch: 96
SNPs call rate < 90%: 0

Excluded sites in total: 212
Remaining sites in total: 2,540,780
Typed only sites: 1,637

Warning: 1 Chunk(s) excluded: < 3 SNPs
Warning: 1 Chunk(s) excluded: reference overlap < 50.0%
Remaining chunk(s): 153
Error: More than 100 obvious strand flips have been detected. Please check strand. Imputation cannot be started!

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

Email MIS team

Data are on the wrong build

Sample call rate was not checked

Must be a MIS problem, since imputation runs locally

Don't know

How to check & fix input files?

- HRC, 1000G & CAAPA: Imputation Preparation Tool (W. Wrayner)
- Checks for consistency between input data and a ref panel
- Required Input
  - Data in PLINK Binary Format (bim, bed, fam)
  - Frequence File (-f): allele frequency of input data
  - Reference Site List (-r)

```
perl HRC-1000G-check-bim.pl -b <bim file> -f
<Frequency file> -r <Reference panel> -h
```

https://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-check-bim-v4.2.11.zip

Use Preparation Tool before uploading your data

-

	Imputation Server	Preparation Tool
Input	**VCF / chromosome**	PLINK binary data
Output	Imputed VCF / chromosome	**VCFs / chromosome**
File Validation & Statistics		
Basic SNP Filtering		
Fixes Strand Errors		
Updating Ref / Alt Assignment		
Removes SNPs with allele frequency difference		

```
seb@genepi:~/ashg19$
```

```
seb@genepi:~/ashg19$ wget https://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-che
ck-bim-v4.2.11.zip
```

```
seb@genepi:~/ashg19$ wget https://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-che
ck-bim-v4.2.11.zip
--2019-10-13 21:02:34--  https://www.well.ox.ac.uk/~wrayner/tools/HRC-1000G-chec
k-bim-v4.2.11.zip
Resolving www.well.ox.ac.uk (www.well.ox.ac.uk)... 52.56.206.186
Connecting to www.well.ox.ac.uk (www.well.ox.ac.uk)|52.56.206.186|:443... connec
ted.
HTTP request sent, awaiting response... 200 OK
Length: 9674 (9.4K) [application/zip]
Saving to: ‘HRC-1000G-check-bim-v4.2.11.zip’

HRC-1000G-check-bim 100%[===================>]   9.45K  --.-KB/s    in 0.003s

2019-10-13 21:02:35 (3.27 MB/s) - ‘HRC-1000G-check-bim-v4.2.11.zip’ saved [9674/
9674]

seb@genepi:~/ashg19$
```

```
seb@genepi:~/ashg19$ unzip HRC-1000G-check-bim-v4.2.11.zip
Archive:  HRC-1000G-check-bim-v4.2.11.zip
  inflating: HRC-1000G-check-bim.pl
  inflating: LICENSE.txt
seb@genepi:~/ashg19$
```

```
seb@genepi:~/ashg19
--2019-10-14 09:36:15--  https://www.well.ox.ac.uk/~wrayner/tools/1000GP_Phase3_c
Resolving www.well.ox.ac.uk (www.well.ox.ac.uk)... 52.56.206.186
Connecting to www.well.ox.ac.uk (www.well.ox.ac.uk)|52.56.206.186|:443... connect
HTTP request sent, awaiting response... 200 OK
Length: 1548210555 (1.4G) [application/x-gzip]
Saving to: ‘1000GP_Phase3_combined.legend.gz’

1000GP_Phase3_combined.legend.gz                    0%[
```

```
seb@genepi:~/ashg19$ ls raw-22-filtered.*
raw-22-filtered.bed  raw-22-filtered.bim  raw-22-filtered.fam
seb@genepi:~/ashg19$
```

```
seb@genepi:~/ashg19$ plink --freq --bfile raw-22-filtered --out raw-22-filtered
```

```
seb@genepi:~/ashg19$ plink --freq --bfile raw-22-filtered --out raw-22-filtered
PLINK v1.90b6.10 64-bit (17 Jun 2019)          www.cog-genomics.org/plink/1.9/
(C) 2005-2019 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to raw-22-filtered.log.
Options in effect:
  --bfile raw-22-filtered
  --freq
  --out raw-22-filtered

7740 MB RAM detected; reserving 3870 MB for main workspace.
22337 variants loaded from .bim file.
5034 people (3027 males, 2007 females) loaded from .fam.
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 0.997355.
--freq: Allele frequencies (founders only) written to raw-22-filtered.frq .
seb@genepi:~/ashg19$
```

```
seb@genepi:~/ashg19$ perl HRC-1000G-check-bim.pl -b raw-22-filtered.bim -f raw-2
2-filtered.frq -r 1000GP_Phase3_combined.legend.gz -g -p EUR
```

```
 Script to check plink .bim files against HRC/1000G for
strand, id names, positions, alleles, ref/alt assignment
            William Rayner 2015-2018
                wrayner@well.ox.ac.uk

                     Version 4.2.9


Options Set:
Reference Panel:             1000G
Bim filename:                raw-22-filtered.bim
Reference filename:          1000GP_Phase3_combined.legend.gz
Allele frequencies filename: raw-22-filtered.frq
Chromosome flag set:         No
Allele frequency threshold:  0.2
Population for 1000G:        EUR


Reading 1000GP_Phase3_combined.legend.gz
Reference Panel is zipped
```

```
Skipped (XY, Y, MT) 0
Total in bim file 22337
Total processed 22280

Indels 0

SNPs not changed 2746
SNPs to change ref alt 16548
Strand ok 11313
Total Strand ok 19294

Strand to change 10268
Total checked 21976
Total checked Strand 21581
Total removed for allele Frequency diff > 0.2 331
Palindromic SNPs with Freq > 0.4 151


Non Matching alleles 244
ID and allele mismatching 244; where 1000G is . 0
Duplicates removed 57
```

```
seb@genepi:~/ashg19$ sh Run-plink.sh
```

```
Note: No phenotypes present.
--make-bed to raw-22-filtered-updated-chr22.bed +
raw-22-filtered-updated-chr22.bim + raw-22-filtered-updated-chr22.fam ... done.
PLINK v1.90b6.10 64-bit (17 Jun 2019)          www.cog-genomics.org/plink/1.9/
(C) 2005-2019 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to raw-22-filtered-updated-chr22.log.
Options in effect:
  --bfile raw-22-filtered-updated
  --chr 22
  --out raw-22-filtered-updated-chr22
  --real-ref-alleles
  --recode vcf

7740 MB RAM detected; reserving 3870 MB for main workspace.
21250 variants loaded from .bim file.
5034 people (3027 males, 2007 females) loaded from .fam.
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 0.997463.
21250 variants and 5034 people pass filters and QC.
Note: No phenotypes present.
```

```
seb@genepi:~/ashg19$ ls raw-22-filtered-updated-chr22.vcf
raw-22-filtered-updated-chr22.vcf
seb@genepi:~/ashg19$ bgzip raw-22-filtered-updated-chr22.vcf
```

Obvious QC 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).

Trickier QC Problems: Chromosome X

- MIS Workflow based on Minimac3
  - Separate males and females for imputation

```
Chromosome X check failed!
java.io.IOException: Something went wrong with the keepSamples male command
Error during manifest file creation.
```

- **Update July 2019:** MIS Workflow based on Minimac4:
  - No split required
  - Users can upload one single file and get an imputed Chromosome X file in return

Session 3

Tracking runs and downloading data

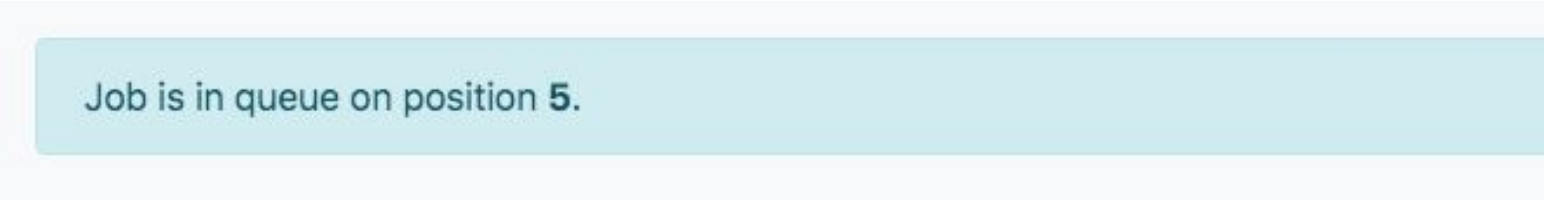
Lukas Forer

Life Cycle of an Imputation Job

- User uploads data
- Job is created
- Input Validation and Quality Control started

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 jobs 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
```

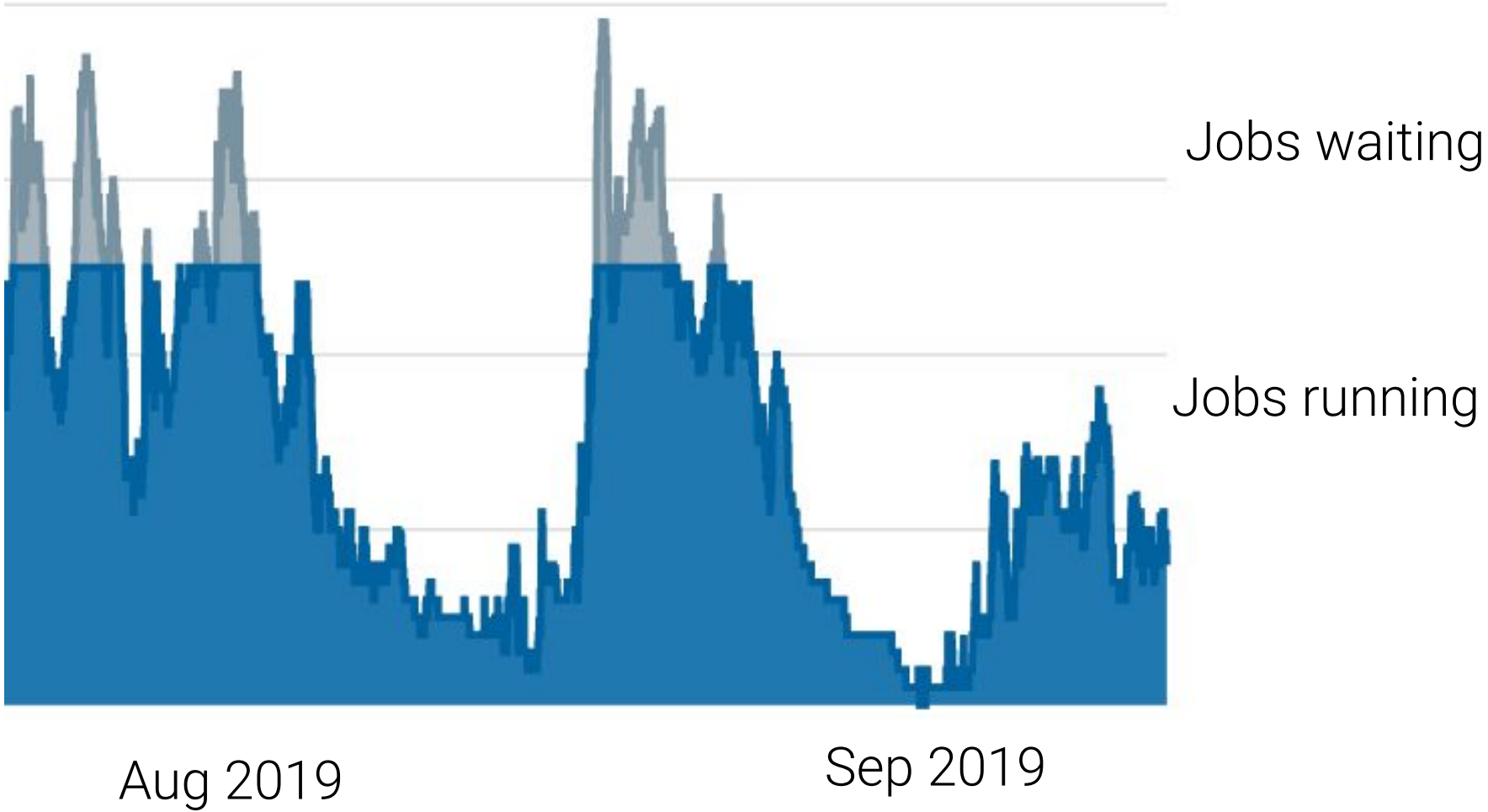
Uploaded my data 5 mins ago and my job is still waiting...

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

There is a problem with my data

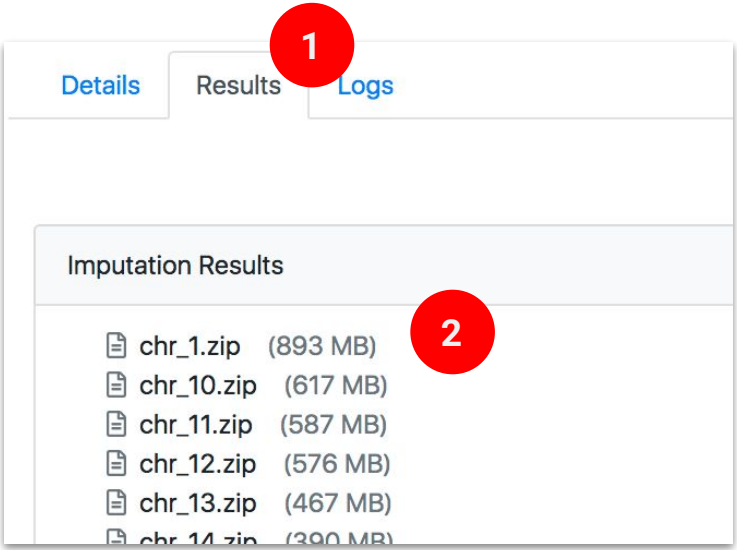
MIS is very busy - check again later

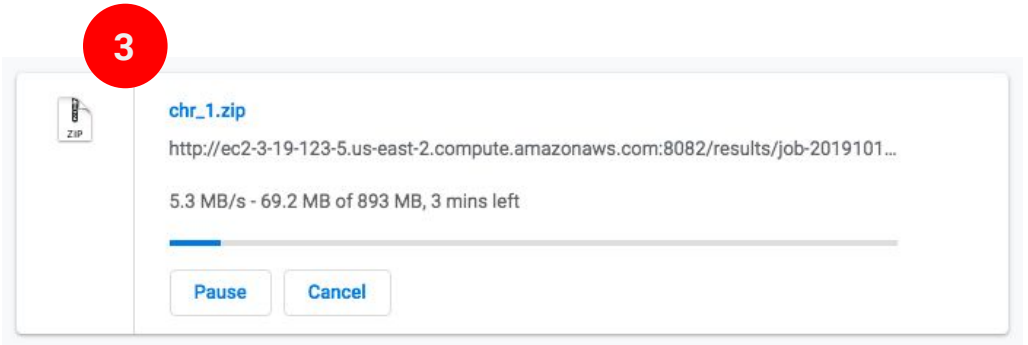
Don't know


Jobs waiting
Jobs running
Aug 2019
Sep 2019

How to get the imputed genotypes?

Option 1: Web-Interface





How to get the imputed genotypes?

Option 1: Web-Interface

+ easiest way to download data
- requires manual interaction

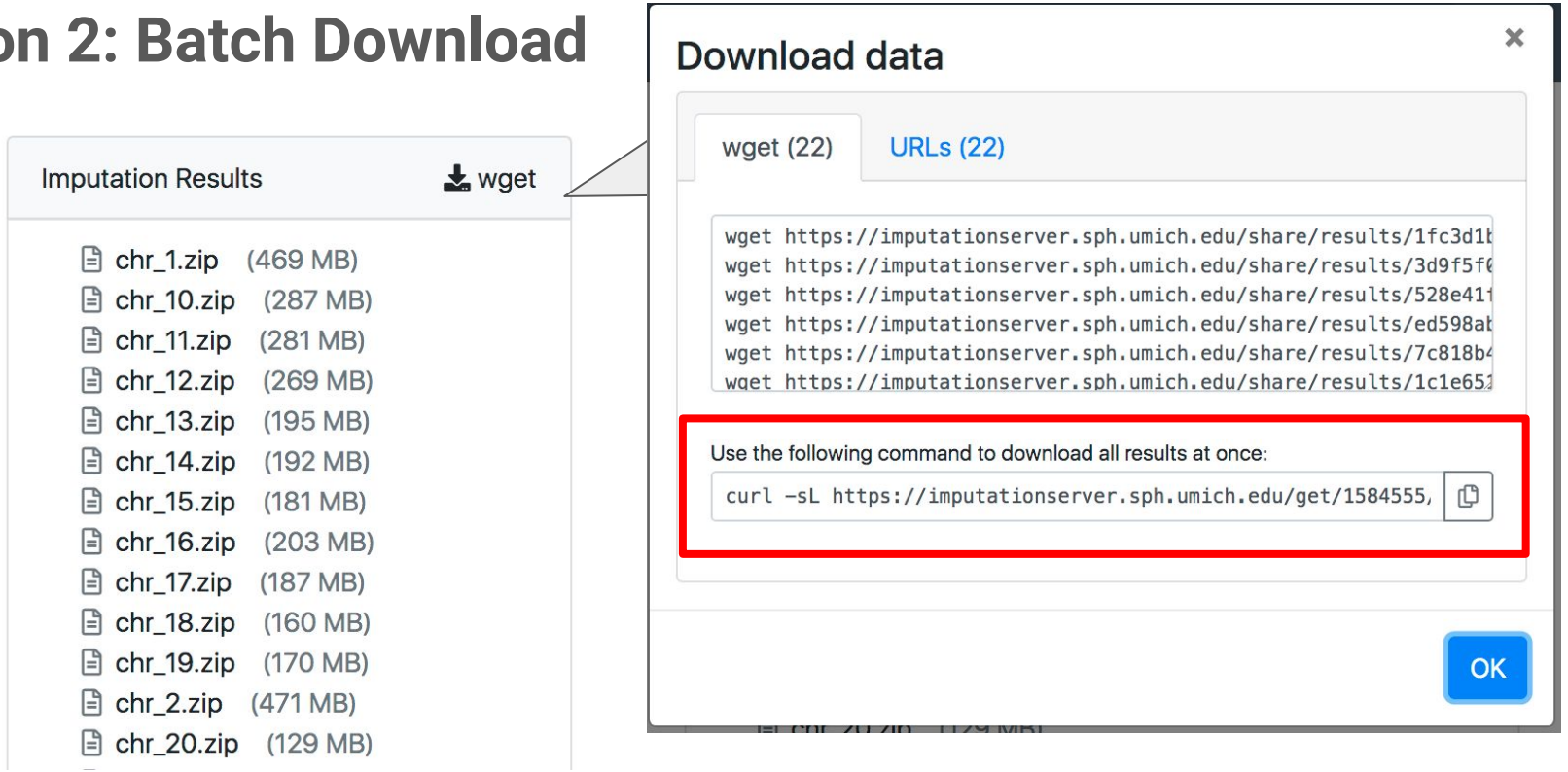
How to get the imputed genotypes?

Option 2: Batch Download

+ most flexible way to download data
+ no browser needed
- requires a basic knowledge of the command line

How to get the imputed genotypes?

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
  0     0    0     0    0     0      0      0 --:--:--  0:00:01 --:--:--     0
```

```
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:~>
```

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

Re-impute data since the job is already retired

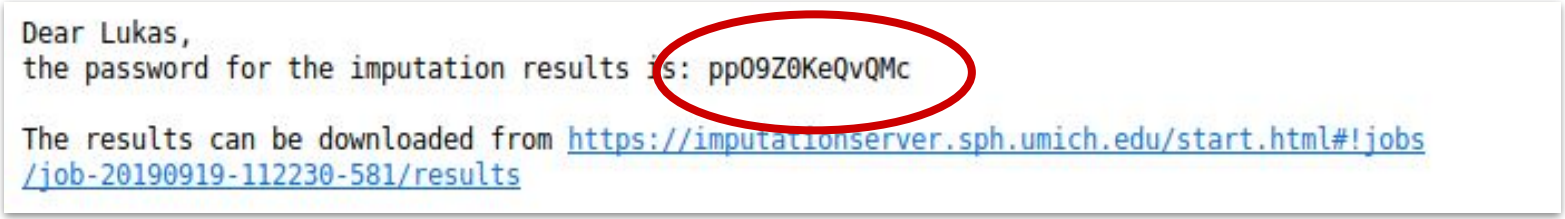
There is a internet problem - should try again later

Email MIS team to increase download counter

Don't know

Data Decryption

- All imputed genotypes are in **encrypted zip files** (e.g. chr_1.zip)
- We sent 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.info.gz

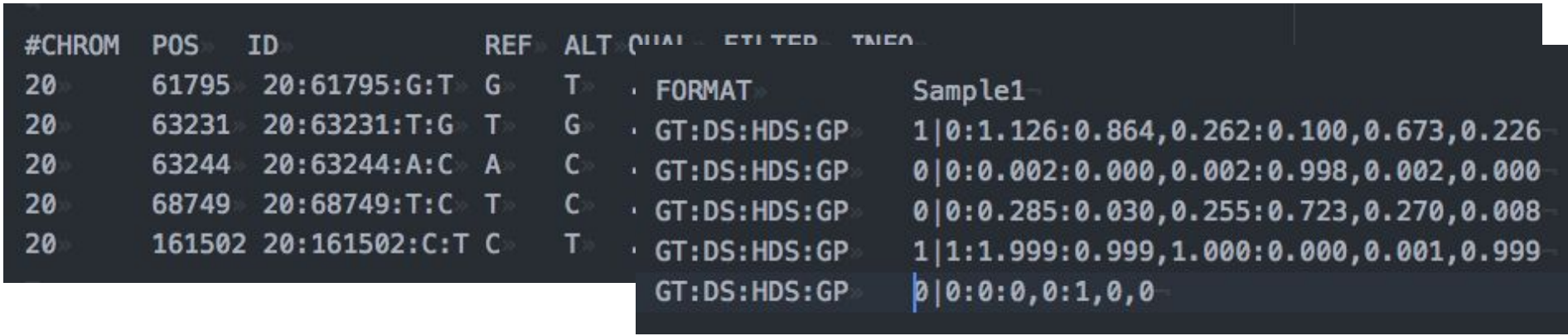
What is in each zip file?

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
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

What is in each zip file?

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



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1
20	61795	20:61795:G:T	G	T	.	.	.	GT:DS:HDS:GP	1|0:1.126:0.864,0.262:0.100,0.673,0.226
20	63231	20:63231:T:G	T	G	.	.	.	GT:DS:HDS:GP	0|0:0.002:0.000,0.002:0.998,0.002,0.000
20	63244	20:63244:A:C	A	C	.	.	.	GT:DS:HDS:GP	0|0:0.285:0.030,0.255:0.723,0.270,0.008
20	68749	20:68749:T:C	T	C	.	.	.	GT:DS:HDS:GP	1|1:1.999:0.999,1.000:0.000,0.001,0.999
20	161502	20:161502:C:T	C	T	.	.	.	GT:DS:HDS:GP	0|0:0:0,0:1,0,0

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

Session 4

Performing GWAS using imputed data

Cassie Spracklen

Have you ever performed a GWAS?

Yes

No

I'm trying

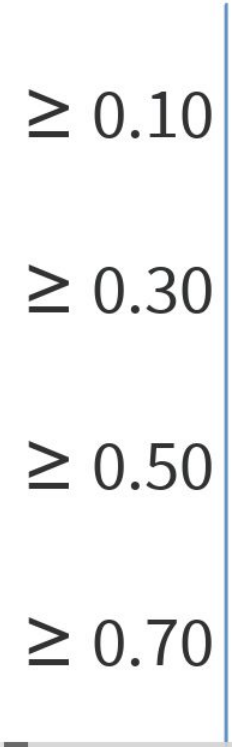
Imputation Quality

- How confident can we be that the imputation is accurate for a particular variant?
- “Rsq” column
  - Range 0-1

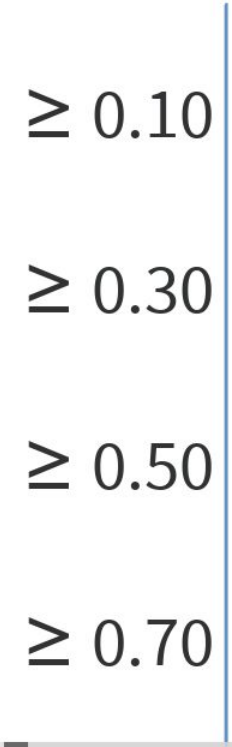
SNP	REF(0)	ALT(1)	ALT_Frq	MAF	AvgCall	Rsq	Genotyped		LooRsq	EmpR	EmpRsq	Dose0	Dose1
16:60164		G	T	0.00610	0.00610	0.99390	0.00134	Imputed	-	-	-	-	-
16:60216		T	C	0.00036	0.00036	0.99964	0.01577	Imputed	-	-	-	-	-
16:60232		T	C	0.00438	0.00438	0.99562	0.00086	Imputed	-	-	-	-	-
16:60288		C	A	0.00316	0.00316	0.99684	0.00332	Imputed	-	-	-	-	-
16:60291		T	C	0.64586	0.35414	0.74930	0.23490	Imputed	-	-	-	-	-
16:60301		C	A	0.00051	0.00051	0.99949	0.00161	Imputed	-	-	-	-	-
16:60375		A	AC	0.05541	0.05541	0.94485	0.03813	Imputed	-	-	-	-	-
16:60404		G	C	0.00332	0.00332	0.99668	0.00152	Imputed	-	-	-	-	-
16:60441		G	A	0.00018	0.00018	0.99982	0.00038	Imputed	-	-	-	-	-

From a chr16.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?

PLINK

EPACTS

Rvtests

BOLT-LMM

SNPTEST

SAIGE

OTHER

I'VE NEVER PERFORMED A GWAS

Available GWAS Programs

No File Reformatting (VCF from MIS)

- EPACTS
- Rvtests
- SAIGE
- SNPTEST

File Formatting Required

- PLINK
- BOLT-LMM

Each GWAS Program Has Strengths, Limitations

EPACTS/Rvtests

+ Many model options
+ Chr X analyses
+ Automatically programs for parallel processing (EPACTS only)
+ Can transform your phenotype file (e.g inverse normal; Rvtests only)
+ Produce covariance matrices for downstream 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)
+ Designed to handle unbalanced number of cases and controls

Chr X analyses unknown

- Should not be used to examine heritability (biased variance estimates)
- Computational time can vary widely between phenotypes and sample sizes
- Can be conservative when 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

PLINK

+ Quick
+ 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/plink2/

Each GWAS Program Has Strengths, Limitations

BOLT-LMM

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

- Requires files to be in BGEN or PLINK format
- Linear mixed models only (quantitative traits); need to convert to log OR for binary traits
- Not optimal for extremely unbalanced case control ratio (especially with rare variants)

BOLT-LMM: https://data.broadinstitute.org/alkesgroup/BOLT-LMM/#x1-5600011

Performing the GWAS

- Each program has its own code, options

- Typical input files (format varies by program)
  - 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)

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

SAIGE

BOLT-LMM

PLINK

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

SAIGE

BOLT-LMM

PLINK

Which program(s) would be best?

Researchers want to perform a GWAS using data from BioBank Japan (>200,000 individuals)

EPACTS/Rvtests

SAIGE

BOLT-LMM

PLINK

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

- Straight-forward errors
  - 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

- Additional 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

- Additional 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

- Additional 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

- Additional common errors
  - Not loaded other programs (e.g. R with EPACTS, SAIGE)
    - Correct by loading other needed programs

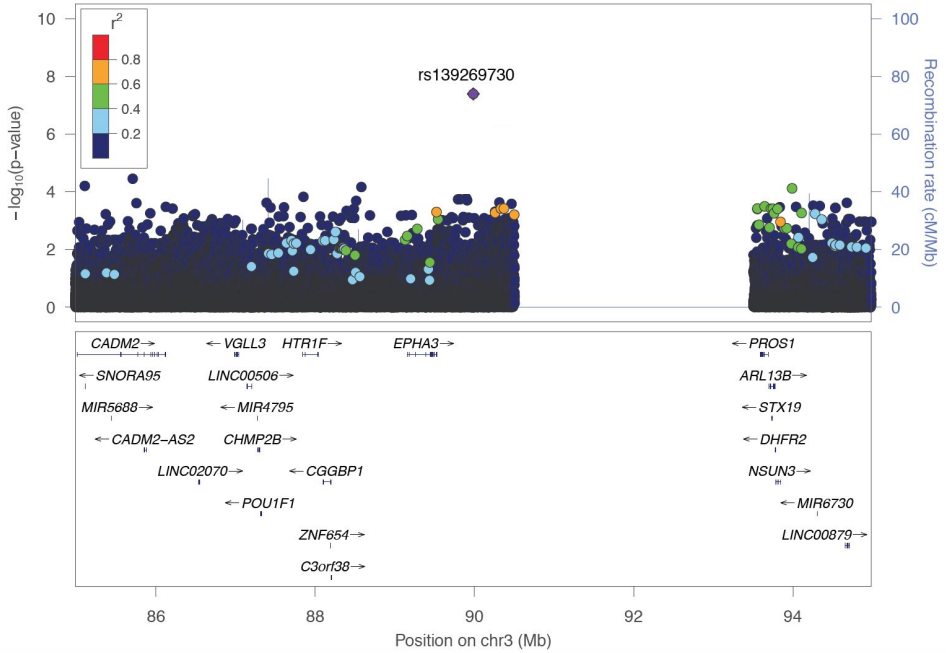
Common Errors When Running a GWAS

- Additional common errors
  - Not loaded other programs (e.g. R with EPACTS, SAIGE)
    - Correct by loading other needed programs
  - Invalid heritability estimate (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



Session 5

Outlook: Imputation Bot and TOPMed

Lukas Forer and Daniel Taliun

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

Yes

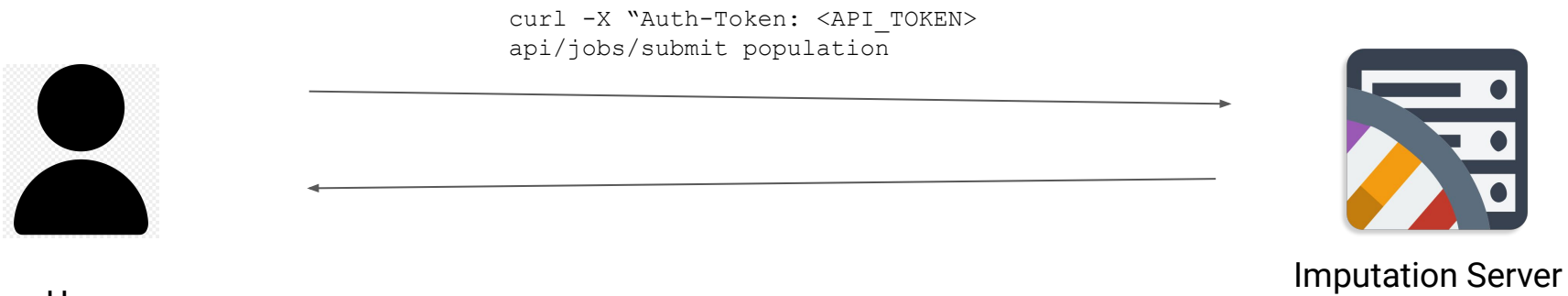
No

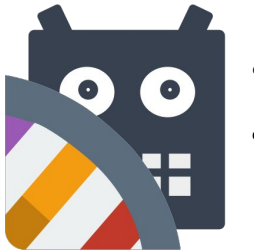
What is an API?

Imputation Server API

Low level API access (since 2018)

- most flexible way to submit and monitor jobs programmatically
- requires basic knowledge of HTTP requests
- needs scripts to combine different commands





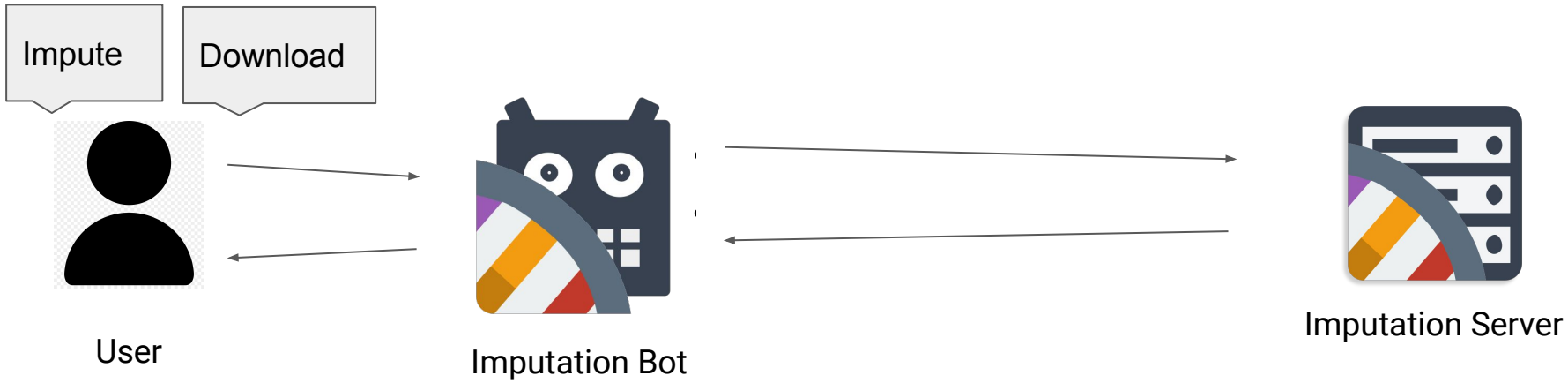
Imputation Bot

https://github.com/lukfor/imputationbot

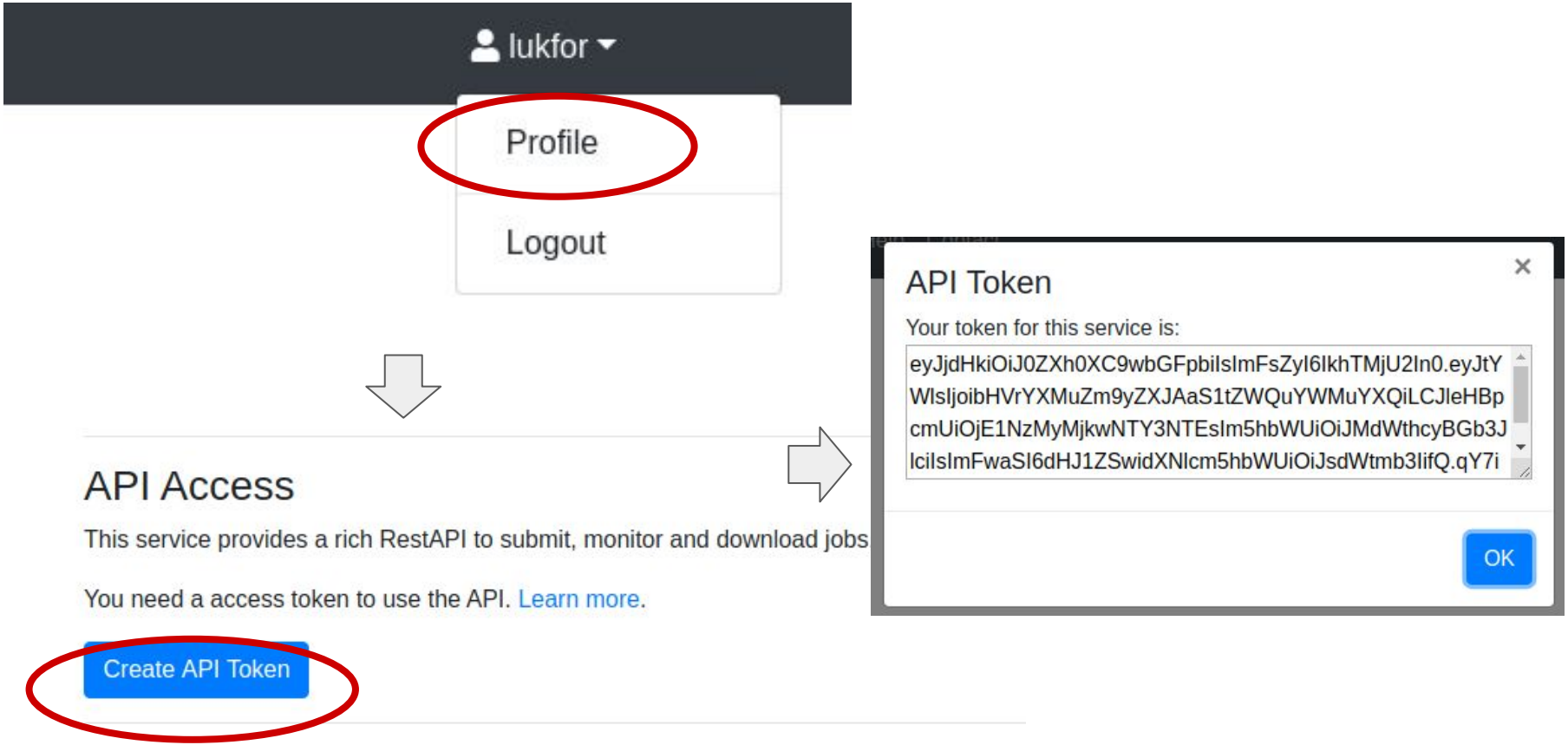
Imputation Bot

High level API access

- easiest way to submit and monitor jobs from the command line
- different commands can easily be combined
- doesn't require any prior knowledge of HTTP requests or API endpoints



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

```
bash-3.2$
```

```
bash-3.2$ curl -sL imputationbot.now.sh | bash
```

```
bash-3.2$ curl -sL imputationbot.now.sh | bash
Installing Imputation Bot v0.1.2...
Downloading Imputation Bot from https://github.com/lukfor/imputationbot/releases/
download/v0.1.2/imputationbot-installer.sh...
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   617    0   617    0     0   1446      0 --:--:-- --:--:-- --:--:--  1448
100 2283k  100 2283k    0     0  1139k      0  0:00:02  0:00:02 --:--:-- 1702k
Verifying archive integrity...  100%   All good.
Uncompressing Make
script=self-extrabable
scriptargs=archives makeself.sh  100%

Imputation Bot v0.1.2 installation completed. Have fun!

bash-3.2$
```

3. Configure Imputation Bot

```
bash-3.2$ ./imputationbot configure
```

```
bash-3.2$ ./imputationbot configure

Imputation Bot 0.1.2
https://imputationserver.sph.umich.edu
(c) 2019 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2019-10-12T12:17:51Z

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

```
bash-3.2$ ./imputationbot configure

Imputation Bot 0.1.2
https://imputationserver.sph.umich.edu
(c) 2019 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2019-10-12T12:17:51Z

Imputationserver Url [https://imputationserver.sph.umich.edu]: http://ec2-3-19-12
3-5.us-east-2.compute.amazonaws.com:8082
API Token [None]:
```

```
bash-3.2$ ./imputationbot configure

Imputation Bot 0.1.2
https://imputationserver.sph.umich.edu
(c) 2019 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2019-10-12T12:17:51Z

Imputationserver Url [https://imputationserver.sph.umich.edu]: http://ec2-3-19-12
3-5.us-east-2.compute.amazonaws.com:8082
API Token [None]: eyJjdHkiOiJ0ZXh0XC9wbGFpbiIsImFsZyI6IkhTMjU2In0.eyJtYWlsIjoibHV
rYXMuZm9yZXJAaS1tZWQuYWMuYXQiLCJleHBpcmUiOjE1NzM0NzU4NTYxOTEsIm5hbWUiOiJMdWthcyBG
b3JlciIsImFwaSI6dHJ1ZSwidXNlcm5hbWUiOiJsdWtmb3IifQ.bfBElLtt98HrUOrZImCLHLYdJ_yMOK
1Ti5hGiIUG9Ec
```

```
bash-3.2$ ./imputationbot configure

Imputation Bot 0.1.2
https://imputationserver.sph.umich.edu
(c) 2019 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2019-10-12T12:17:51Z

Imputationserver Url [https://imputationserver.sph.umich.edu]: http://ec2-3-19-12
3-5.us-east-2.compute.amazonaws.com:8082
API Token [None]: eyJjdHkiOiJ0ZXh0XC9wbGFpbiIsImFsZyI6IkhTMjU2In0.eyJtYWlsIjoibHV
rYXMuZm9yZXJAaS1tZWQuYWMuYXQiLCJleHBpcmUiOjE1NzM0NzU4NTYxOTEsIm5hbWUiOiJMdWthcyBG
b3JlciIsImFwaSI6dHJ1ZSwidXNlcm5hbWUiOiJsdWtmb3IifQ.bfBElLtt98HrUOrZImCLHLYdJ_yMOK
1Ti5hGiIUG9Ec

Hi Lukas Forer



Imputation Bot is ready to submit jobs to 'Michigan Imputation Server (Minimac4)'


bash-3.2$
```

4. Submit a Job

```
bash-3.2$ ./imputationbot impute --files test-data/chr20.R50.merged.1.330k.recode
.vcf.gz --refpanel 1000g-phase-3-v5  --population eur
```

```
bash-3.2$ ./imputationbot impute --files test-data/chr20.R50.merged.1.330k.recode
.vcf.gz --refpanel 1000g-phase-3-v5  --population eur

Imputation Bot 0.1.2
https://imputationserver.sph.umich.edu
(c) 2019 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2019-10-12T12:17:51Z

Parameters:
  refpanel: apps@1000g-phase-3-v5@2.0.0
  files:
    - test-data/chr20.R50.merged.1.330k.recode.vcf.gz
  build: hg19
  r2Filter: 0
  phasing: eagle
  population: eur
  aesEncryption: no
```

```
Imputation Bot 0.1.2
https://imputationserver.sph.umich.edu
(c) 2019 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2019-10-12T12:17:51Z

Parameters:
  refpanel: apps@1000g-phase-3-v5@2.0.0
  files:
    - test-data/chr20.R50.merged.1.330k.recode.vcf.gz
  build: hg19
  r2Filter: 0
  phasing: eagle
  population: eur
  aesEncryption: no

Imputation job submitted successfully

 Check the job progress on http://ec2-3-19-123-5.us-east-2.compute.amazonaws.co
m:8082/index.html#!jobs/job-20191012-152533-884


bash-3.2$
```

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 0.1.3
https://imputationserver.sph.umich.edu
(c) 2019 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2019-10-13T07:19:51Z

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 0.1.3
https://imputationserver.sph.umich.edu
(c) 2019 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2019-10-13T07:19:51Z

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 0.1.3
https://imputationserver.sph.umich.edu
(c) 2019 Lukas Forer, Sebastian Schoenherr and Christian Fuchsberger
Built by lukas on 2019-10-13T07:19:51Z

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$
```

TOPMed Imputation Reference Panel

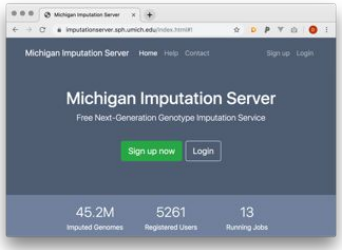
NHLBI

TOPMed

Trans-**O**mics of **P**recision **Med**icine Program (TOPMed) — NHLBI's large scale genomics program

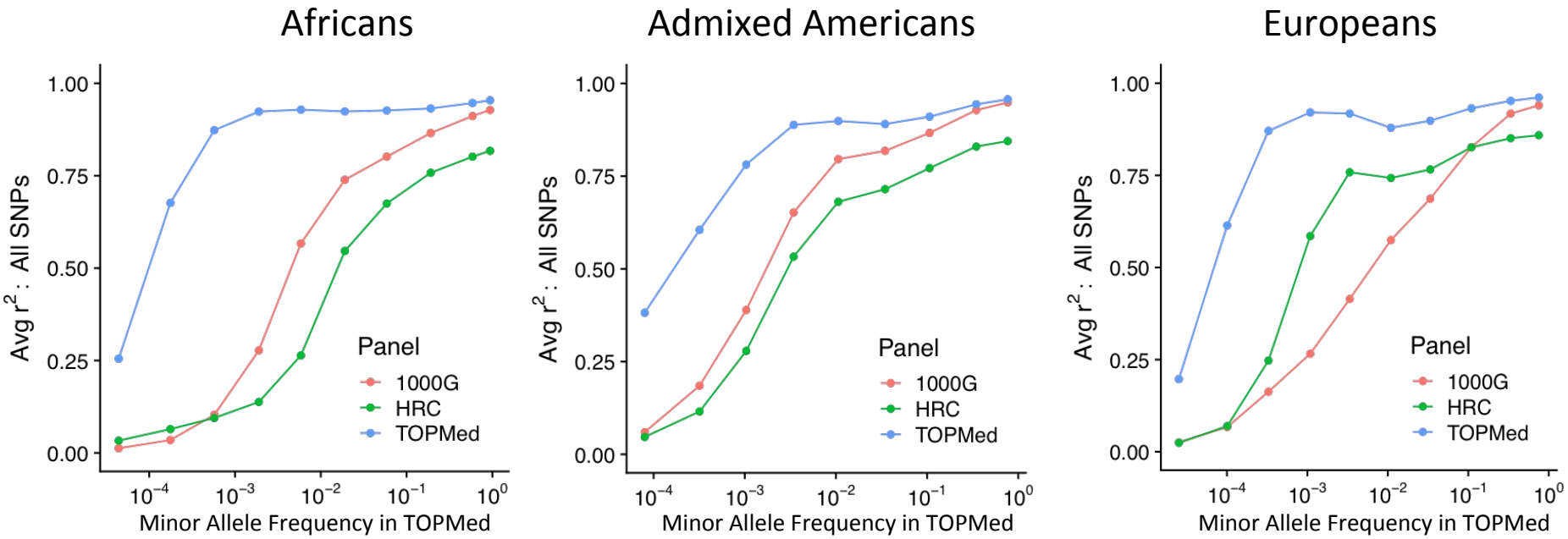
Currently generating deep (40X) whole genome sequence data for samples from >80 studies

Highly diverse group — Europeans, Africans, and Hispanic/Latino are well represented

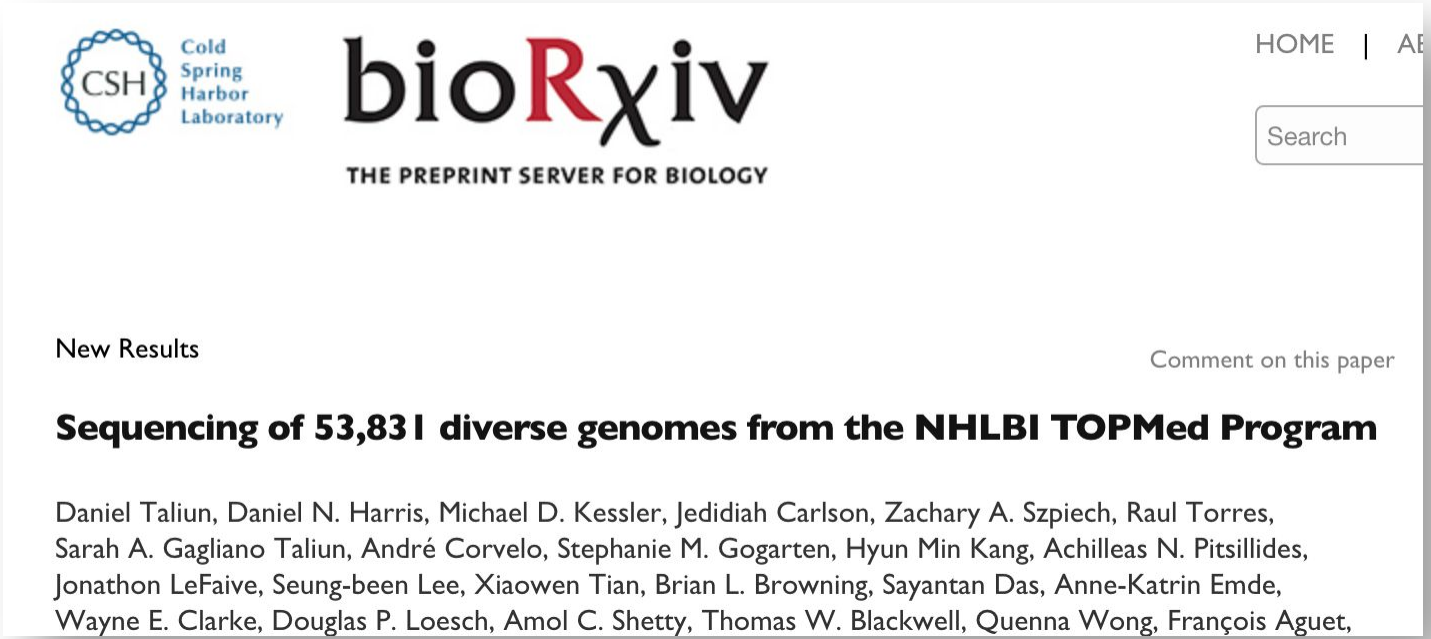


Expect to enable imputation for 10s of millions of variants down to <0.1% allele frequency

Imputation Accuracy with TOPMed Reference



More in TOPMed Flagship Paper



Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | AB

Search

New Results

Comment on this paper

Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program

Daniel Taliun, Daniel N. Harris, Michael D. Kessler, Jedidiah Carlson, Zachary A. Szpiech, Raul Torres, Sarah A. Gagliano Taliun, André Corvelo, Stephanie M. Gogarten, Hyun Min Kang, Achilleas N. Pitsillides, Jonathon LeFaive, Seung-been Lee, Xiaowen Tian, Brian L. Browning, Sayantan Das, Anne-Katrin Emde, Wayne E. Clarke, Douglas P. Loesch, Amol C. Shetty, Thomas W. Blackwell, Quenna Wong, François Aguet,

Session 6

Your Questions

What did happen with my rs numbers?

1	1:748878	0	748878	T	G
1	1:751756	0	751756	T	C
1	1:752566	0	752566	A	G
1	1:752721	0	752721	G	A
1	1:752894	0	752894	C	T
1	1:753405	0	753405	A	C
1	1:753474	0	753474	G	C
1	1:753541	0	753541	G	A
1	1:754182	0	754182	G	A
1	1:754192	0	754192	G	A
1	1:754334	0	754334	C	T
1	1:754503	0	754503	A	G
1	1:754964	0	754964	T	C

What reference panel should I use?

- Europeans: HRC (SNP only)
- Multiethnic studies / InDels+SVs: 1000 Genomes Phase 3
- African Americans: CAAPA
- Asians: Genome Asia

- For reproducibility reasons
  - HapMap & 1000 Genomes Phase 1

What should we prioritize?

Thank you!

For follow-up questions: tomorrow, 10-1030am
in front of the Grand Ballroom