Section 2

Run a job, Data Preparation and Data Download

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

Learning objectives

Participants will learn

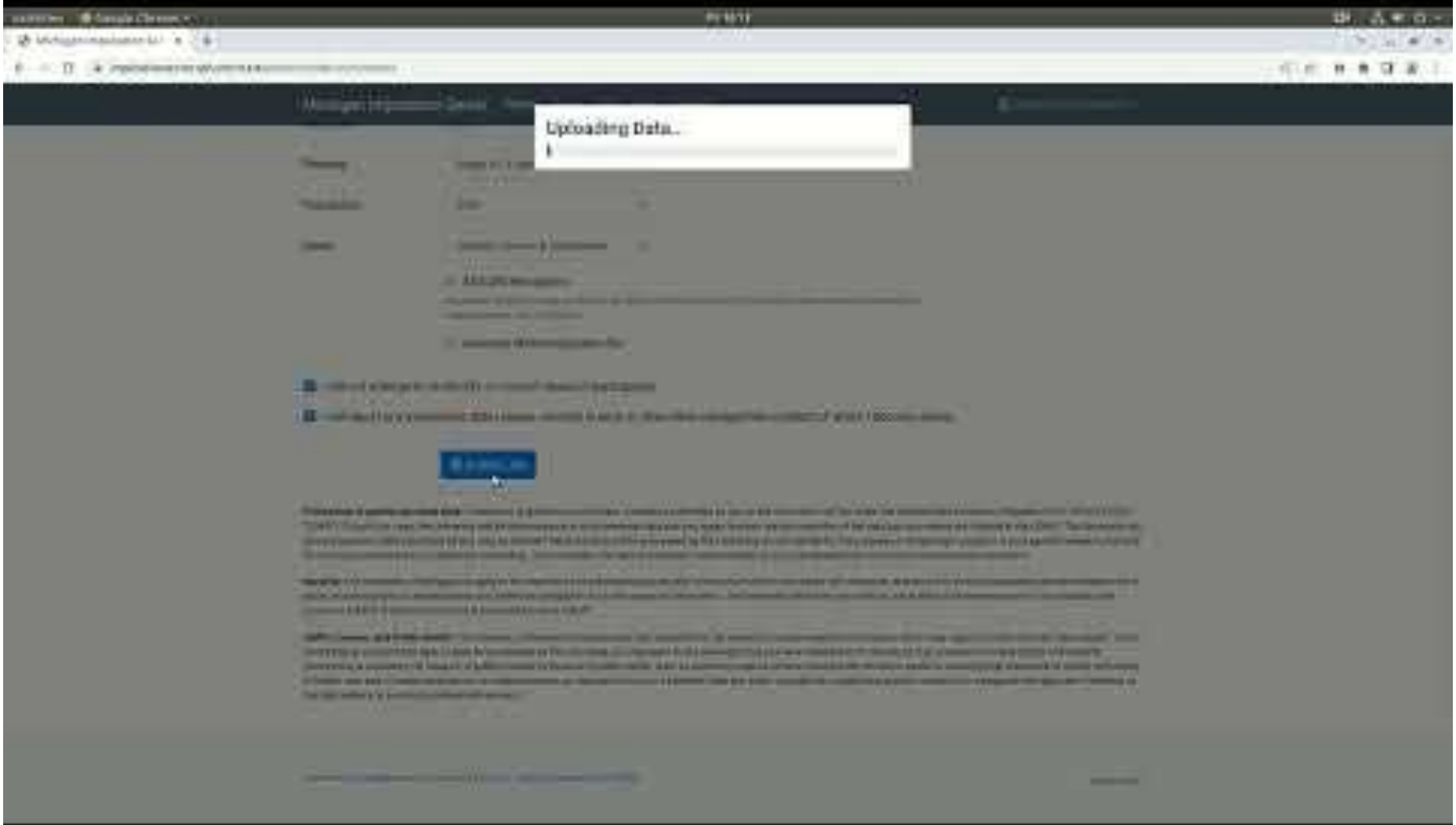
1. How to submit a job on Michigan Imputation Server (MIS)
2. How to prepare your GWAS data
3. Different ways to download final datasets

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

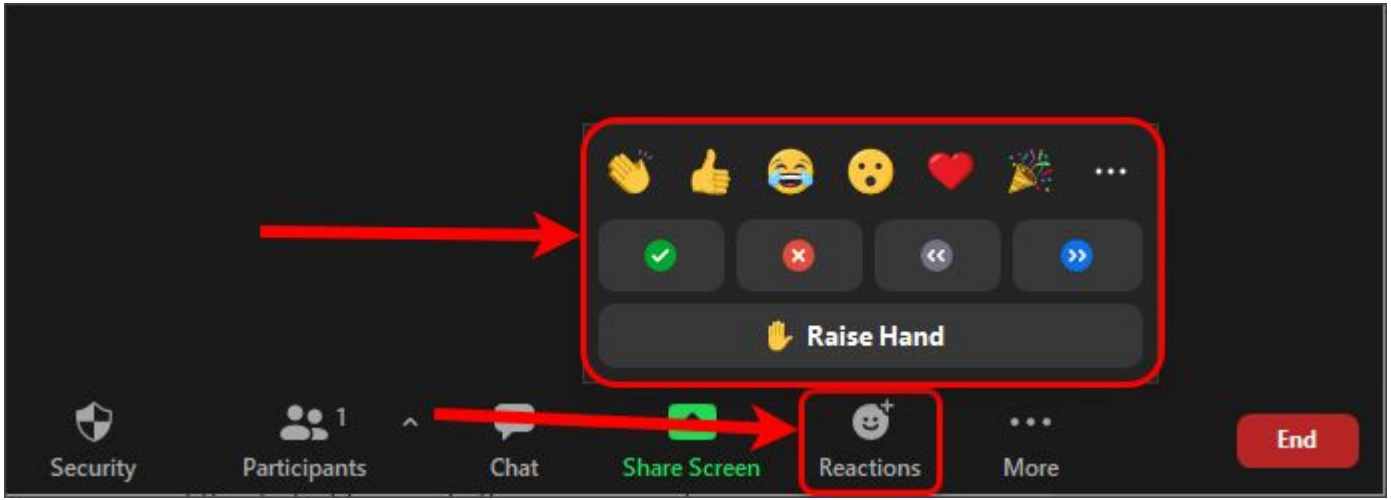
- Input Validation and Quality Control executed right after data upload
  - Immediate feedback to users
  - Jobs passing the QC are then added to a long-time queue
- MIS outputs SNP statistics and a QC Report for each job
  - Helps you to <u>identify problems</u>

Have you run into QC problems so far?

Put a “Yes” in the chat

Or

Raise Hand

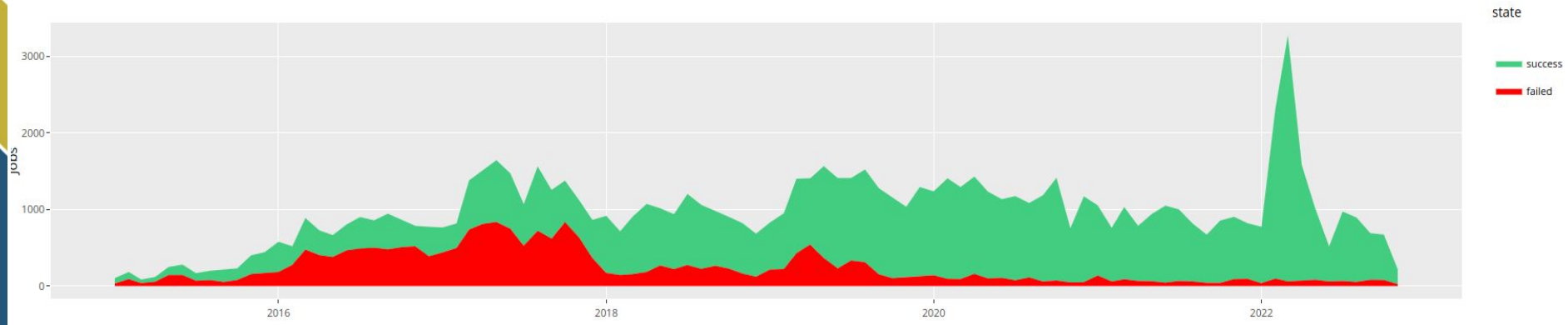


ASHG
American Society of Human Genetics

How many jobs are failing?

- 40% in 2015; 20% in 2019; 7% 2020-2022
  - Reason for job failures: Something wrong with your input data ***or*** phasing/imputation issue on our side

Total amount of jobs: **91,400** (Nov 22)
> 2M in July 22



MIS QC: Input Validation & Statistics

-

	Imputation Server
Input	**VCF / chromosome**
Output	Imputed VCF / chromosome
File Validation & Statistics	
Basic SNP Filtering	
Lift Over	

Input Validation

4 valid VCF file(s) found.

Samples: 51471
Chromosomes: 11 12 13 14
SNPs: 72808
Chunks: 26
Datatype: unphased
Build: hg19
Reference Panel: apps@1000g-phase-3-v5 (hg19)
Population: eur
Phasing: eagle
Mode: imputation

MIS QC: Basic SNP Filtering

-

	Imputation Server
Input	**VCF / chromosome**
Output	Imputed VCF / chromosome
File Validation & Statistics	
Basic SNP Filtering	
Lift Over	

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

ASHG
American Society of Human Genetics

MIS QC: Lift Over Step

-

	Imputation Server
Input	**VCF / chromosome**
Output	Imputed VCF / chromosome
File Validation & Statistics	
Basic SNP Filtering	
Lift Over	

Quality Control

Uploaded data is hg38 and reference is hg19.

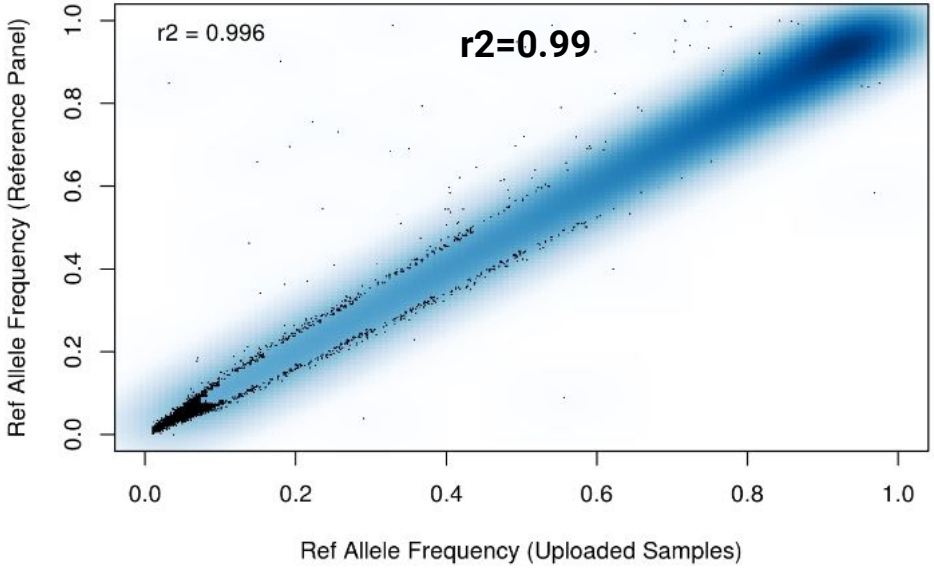
Lift Over

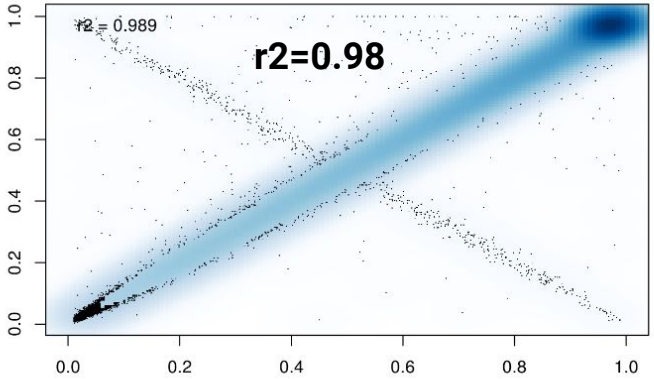
Calculating QC Statistics

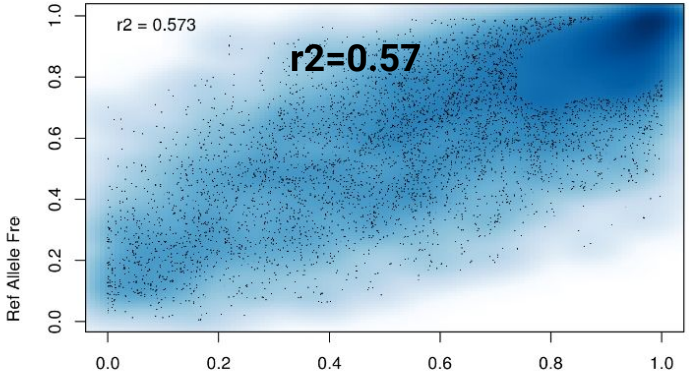
Statistics:
Alternative allele frequency > 0.5 sites: 2
Reference Overlap: 100.00 %

ASHG
American Society of Human Genetics

MIS QC Report: Allele Frequency Check







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?

ASHG
American Society of Human Genetics

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		

ASHG
American Society of Human Genetics

```
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%
```

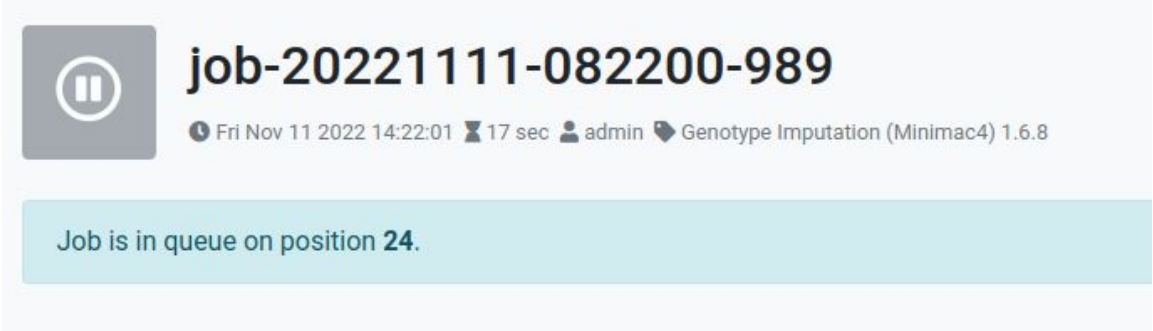
ASHG
American Society of Human Genetics

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

ASHG
American Society of Human Genetics

Life Cycle of an Imputation Job

- Job passed Quality Control
- Job scheduled in imputation queue



- <u>Waits until resources are available</u>

ASHG
American Society of Human Genetics

Life Cycle of an Imputation Job

- Phasing and Imputation starts

ASHG
American Society of Human Genetics

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
```

ASHG
American Society of Human Genetics

Life Cycle of an Imputation Job

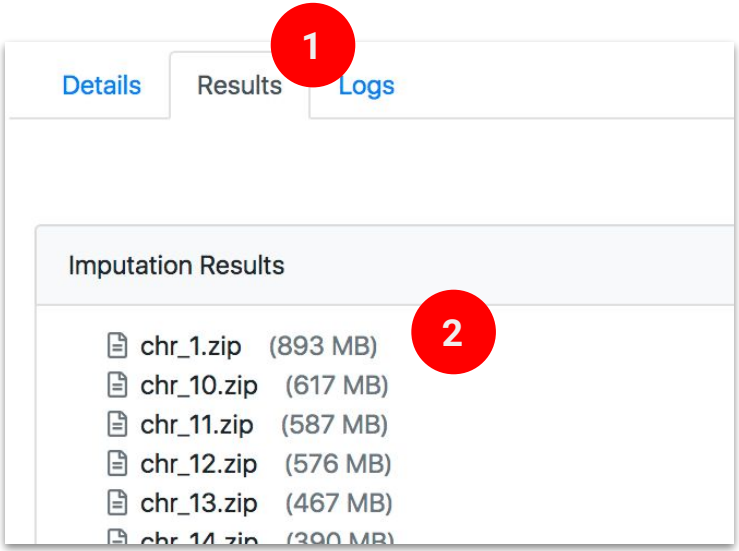
- After 7 days the job is retired
- All results are deleted
- We will send you an email 2 days before

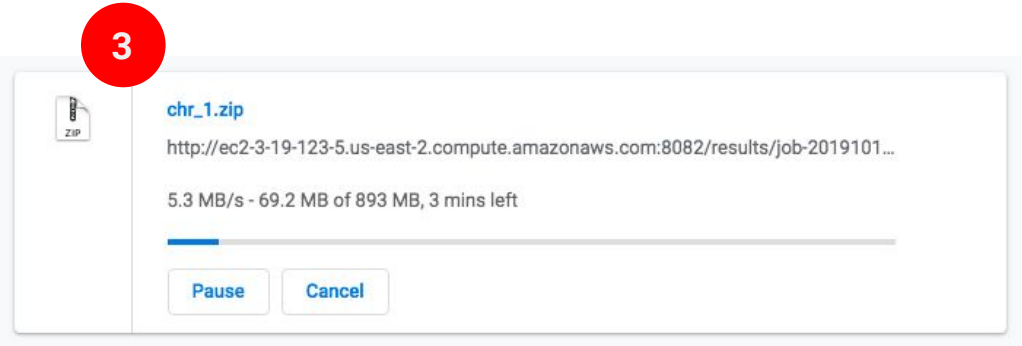
```
Dear Lukas Forer,
Your job retires in 2 days! All imputation results will be deleted at that time.

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

ASHG
American Society of Human Genetics

How to download the imputed genotypes?

Option 1: Web-Interface





ASHG
American Society of Human Genetics

Option 2: Batch Download

Imputation Results wget

- chr_1.zip (469 MB)
- chr_10.zip (287 MB)
- chr_11.zip (281 MB)
- chr_12.zip (269 MB)
- chr_13.zip (195 MB)
- chr_14.zip (192 MB)
- chr_15.zip (181 MB)
- chr_16.zip (203 MB)
- chr_17.zip (187 MB)
- chr_18.zip (160 MB)
- chr_19.zip (170 MB)
- chr_2.zip (471 MB)
- chr_20.zip (129 MB)

Download data

wget (22) URLs (22)

```
wget https://imputationserver.sph.umich.edu/share/results/1fc3d1b
wget https://imputationserver.sph.umich.edu/share/results/3d9f5f0
wget https://imputationserver.sph.umich.edu/share/results/528e411
wget https://imputationserver.sph.umich.edu/share/results/ed598ab
wget https://imputationserver.sph.umich.edu/share/results/7c818b4
wget https://imputationserver.sph.umich.edu/share/results/1c1e651
```

Use the following command to download all results at once:

```
curl -sL https://imputationserver.sph.umich.edu/get/1584565,
```

OK

ASHG American Society of Human Genetics

```
fantasia:~>
```

ASHG
American Society of Human Genetics

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

ASHG
American Society of Human Genetics

```
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
- Checkout Session 4

Data Decryption

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

```
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
```

- 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)

ASHG
American Society of Human Genetics

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

ASHG
American Society of Human Genetics

What is in each zip file?

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

#CHROM	POS	ID	REF	ALT	FORMAT	Sample1
20	61795	20:61795:G:T	G	T	GT:DS:GP	1|0:1.126:0.100,0.673,0.226
20	63231	20:63231:T:G	T	G	GT:DS:GP	0|0:0.002:0.998,0.002,0.000
20	63244	20:63244:A:C	A	C	GT:DS:GP	0|0:0.285:0.723,0.270,0.008

…

…

…

ASHG
American Society of Human Genetics

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?

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

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
```

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