Disclosure Slide

Financial Disclosure for:

Christian Fuchsberger
Sebastian Schönherr
Lukas Forer
Cassie Spracklen
Albert Smith
Sarah Hanks

We have nothing to disclose

Section 2

Run a job, Quality Control and Data Preparation

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

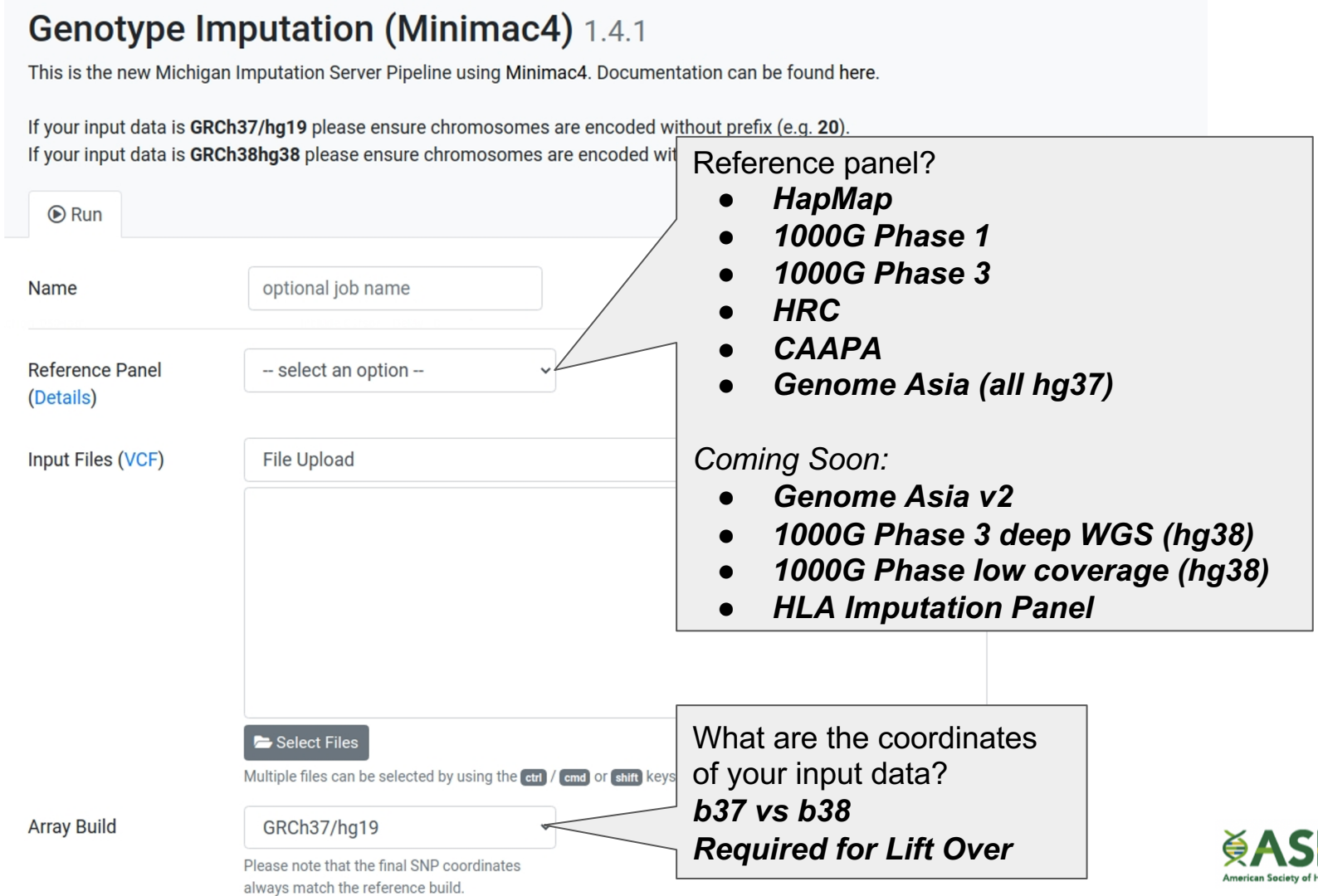
1. How to submit a job on Michigan Imputation Server
2. How to prepare their input data that they are passing our QC step

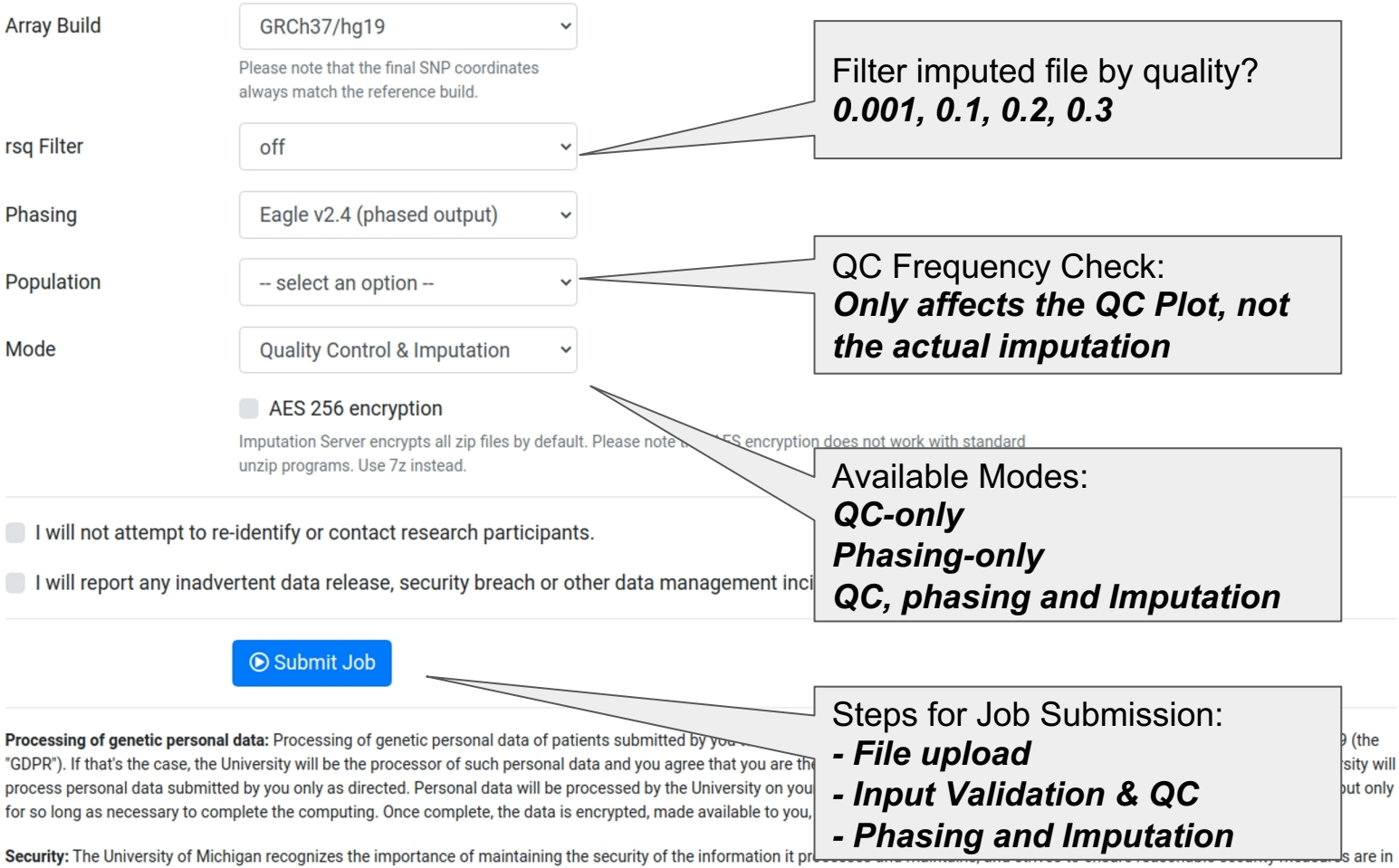
Options for job submission

- MIS Web Interface (This Section)
- API and Imputation Bot (Section 5)
- Docker Image (see Docs)
- Deploy on Amazon (see Docs)

Run your first job using the web interface

https://imputationserver.sph.umich.edu


Genotype Imputation (Minimac4) 1.4.1
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 wit
Run
Name
optional job name
Reference Panel (Details)
-- select an option --
Input Files (VCF)
File Upload
Select Files
Multiple files can be selected by using the ctrl / cmd or shift keys
Array Build
GRCh37/hg19
Please note that the final SNP coordinates always match the reference build.
Reference panel?
HapMap
1000G Phase 1
1000G Phase 3
HRC
CAAPA
Genome Asia (all hg37)
Coming Soon:
Genome Asia v2
1000G Phase 3 deep WGS (hg38)
1000G Phase low coverage (hg38)
HLA Imputation Panel
What are the coordinates of your input data?
b37 vs b38
Required for Lift Over



ASHG
American Society of Human Genetics

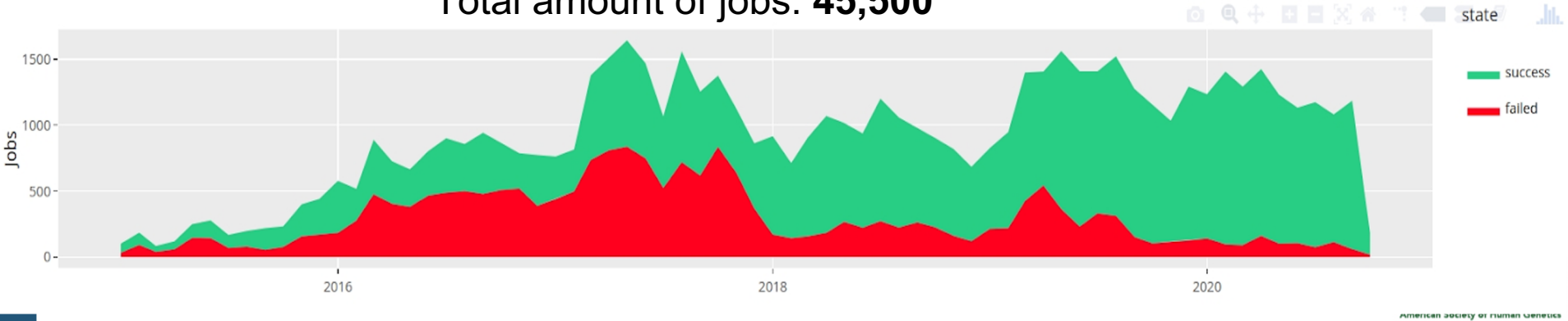
Submit a job

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

How many jobs are not passing QC?

- 2019: 20 % of all jobs failed
- 2020: 7-10 % jobs failed
- Reason for job failures: Something wrong with your input data ***or*** phasing/imputation issue on our side

Total amount of jobs: **45,500**



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

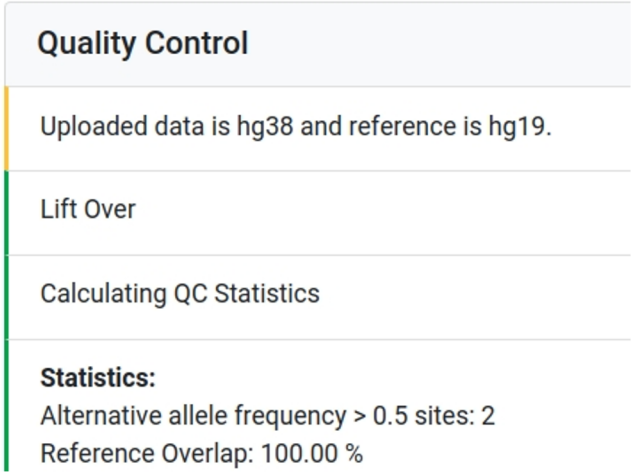
MIS QC: Basic SNP Filtering

	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

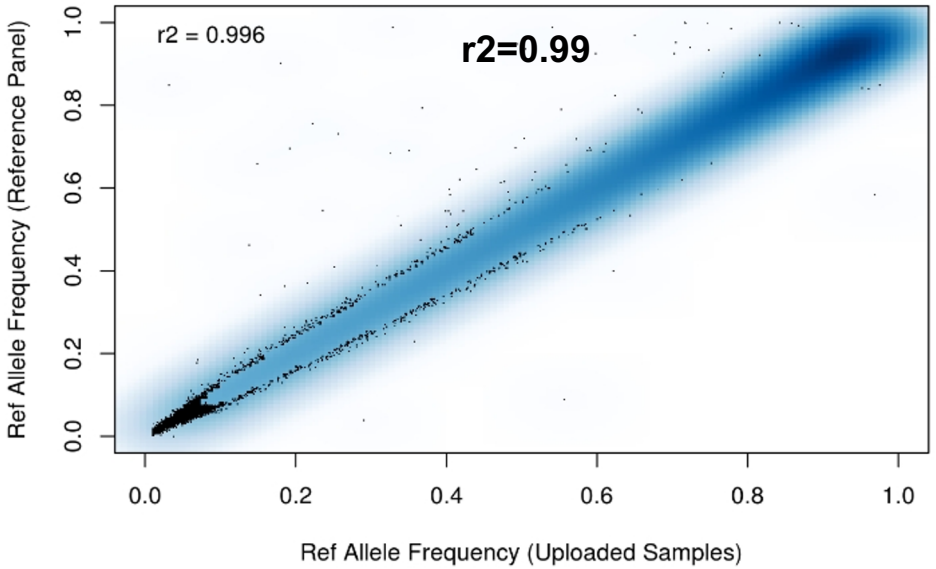
MIS QC: Lift Over Step

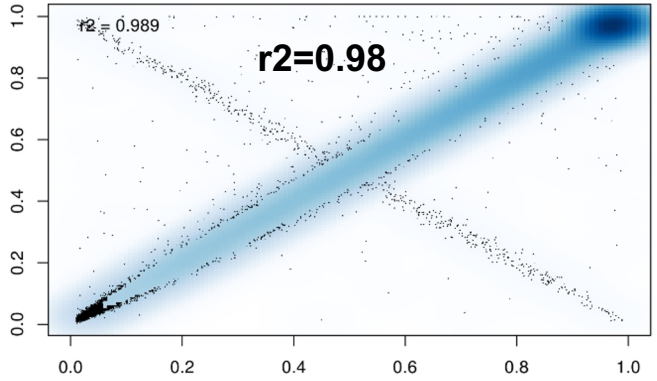
	Imputation Server
Input	**VCF / chromosome**
Output	Imputed VCF / chromosome
File Validation & Statistics	
Basic SNP Filtering	
Lift Over	

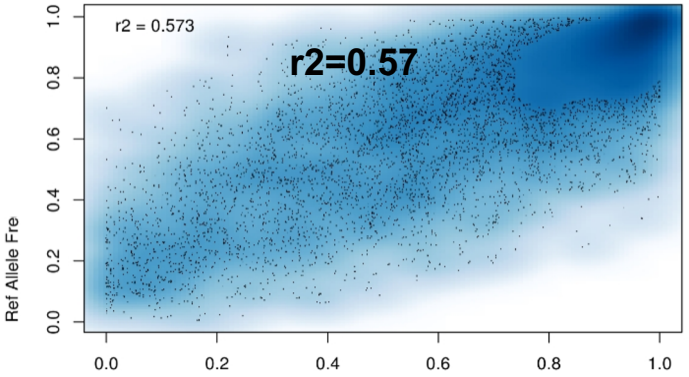


ASHG
American Society of Human Genetics

MIS QC Report: Allele Frequency Check







Passing QC

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

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

Pre-phasing and Imputation

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

ASHG
American Society of Human Genetics

Failing QC - 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. Wrayner
- 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, SNP with differing alleles		

ASHG
American Society of Human Genetics

```
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
```

ASHG
American Society of Human Genetics

```
        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
```

ASHG
American Society of Human Genetics

```
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
```

Summary

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

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

ASHG
American Society of Human Genetics