Imputation server

Christian Fuchsberger

University of Michigan

1000 Genomes

A Deep Catalog of Human Genetic Variation

Genotype imputation

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

- **Intuition:** apparently “unrelated” individuals share short stretches of haplotype

- **Approach:** identify the shared stretches and fill in “missing” genotypes

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

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

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

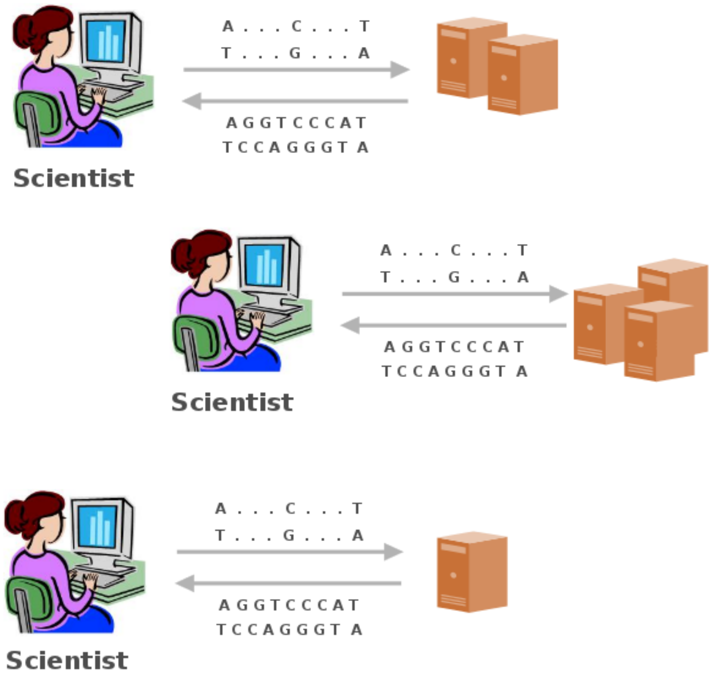
Larger panel increase imputation quality and # of variants imputed

Year	Reference Panel	# Haplotypes	# Variants	Imputation quality (r^2) 1-5% MAF
2007	HapMap 2 (CEU)	120	2.5M	.70
2010	1000 Genomes Phase 1	2,184	40M	.74
2014	1000 Genomes Phase 3	5,010	81M	.80

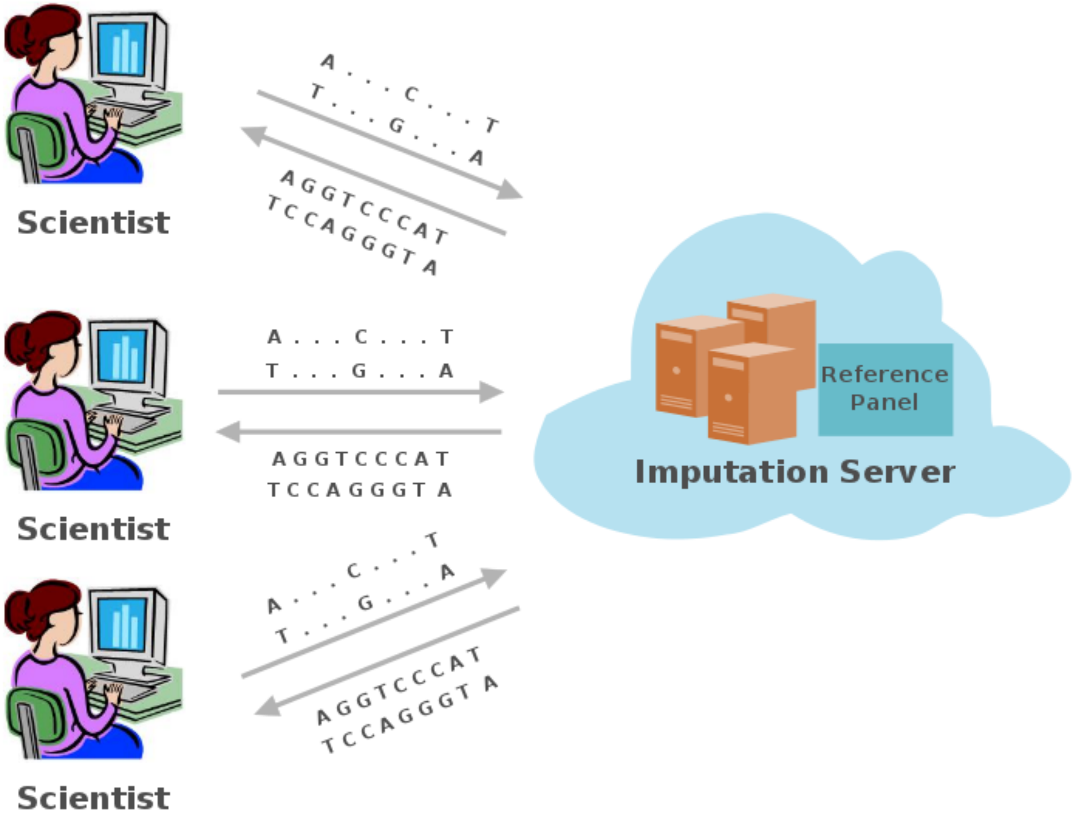
Imputation quality results are based on 1,004 sequenced Finnish samples, mimicking a 330k Illumina GWAS chip

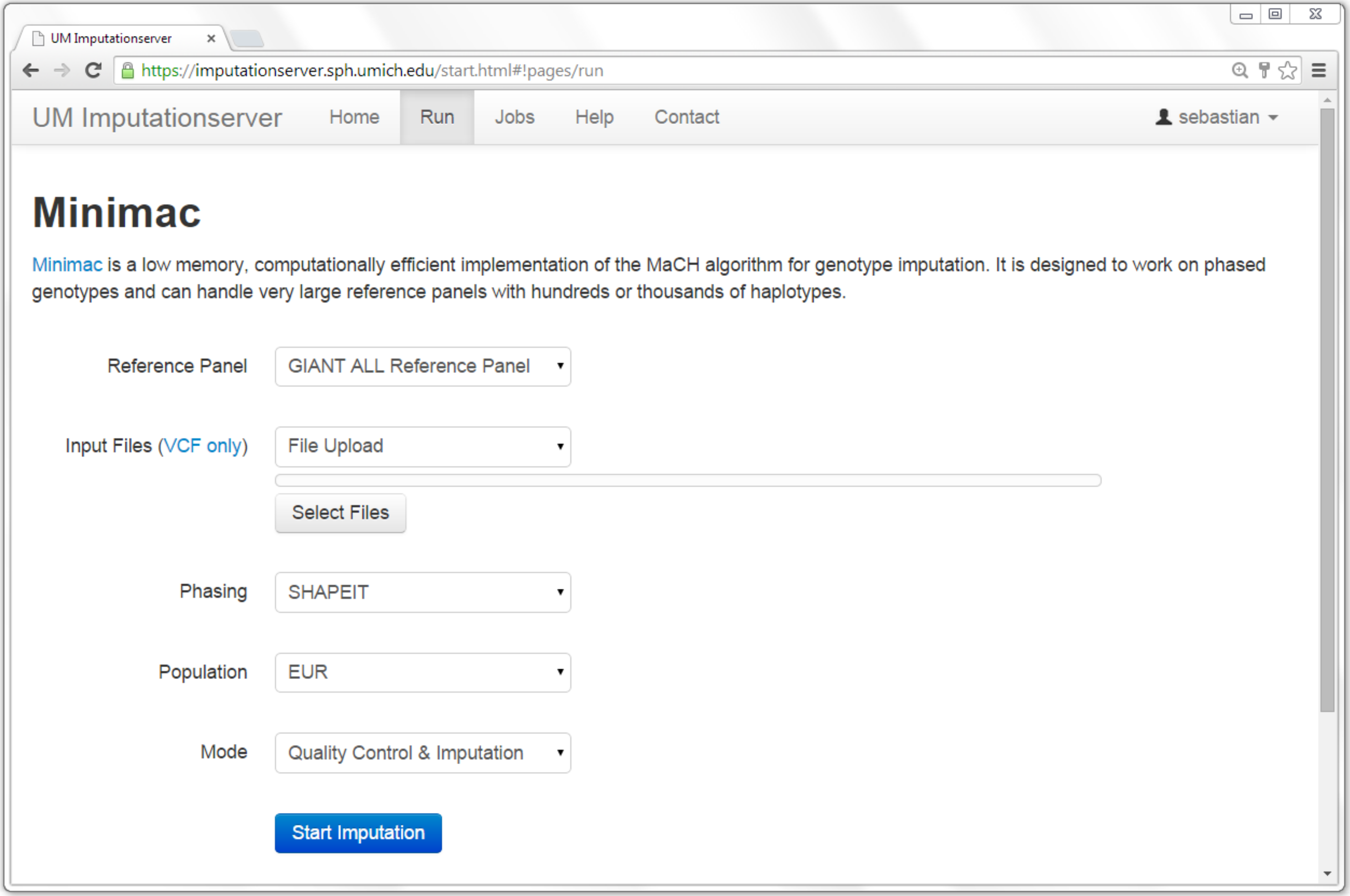
Challenges

1. Computational and analytical burden for many studies
2. Protocols are well developed, but many pitfalls: strand and allele matching, parameter settings,…



Solution: imputation web service





Minimac

Minimac is a low memory, computationally efficient implementation of the MaCH algorithm for genotype imputation. It is designed to work on phased genotypes and can handle very large reference panels with hundreds or thousands of haplotypes.

Reference Panel: GIANT ALL Reference Panel

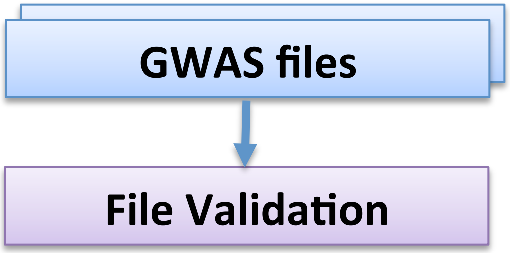
Input Files (VCF only): File Upload

Phasing: SHAPEIT

Population: EUR

Mode: Quality Control & Imputation

Imputation server workflow



1.) User uploads unphased or pre-phased genotypes in VCF format via web or secure ftp

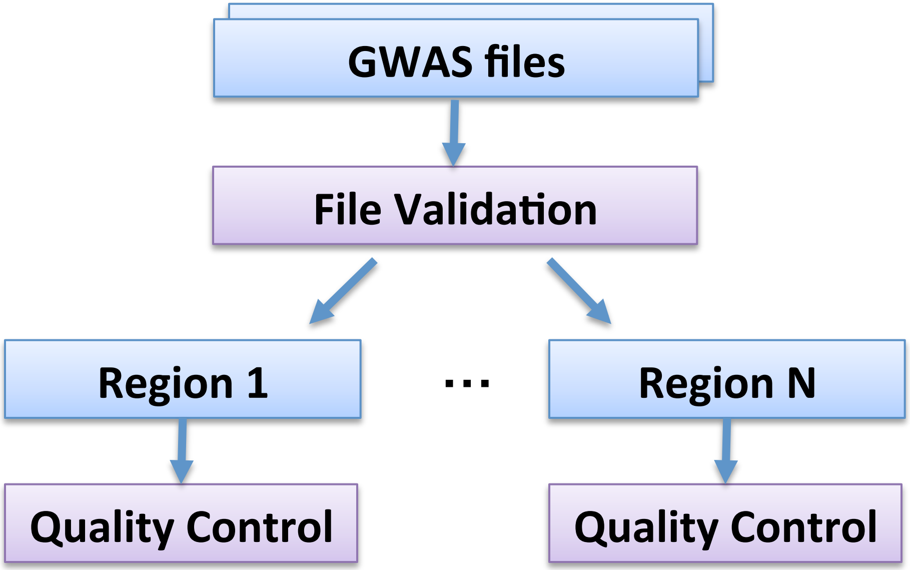
Key:

- Data in VCF format
- Genome build 37
- Forward strand
- Basic quality controlled

Decisions:

- Reference panel to use
- Pre-phasing: SHAPEIT2 / HAPI-UR
- Population to use for QC

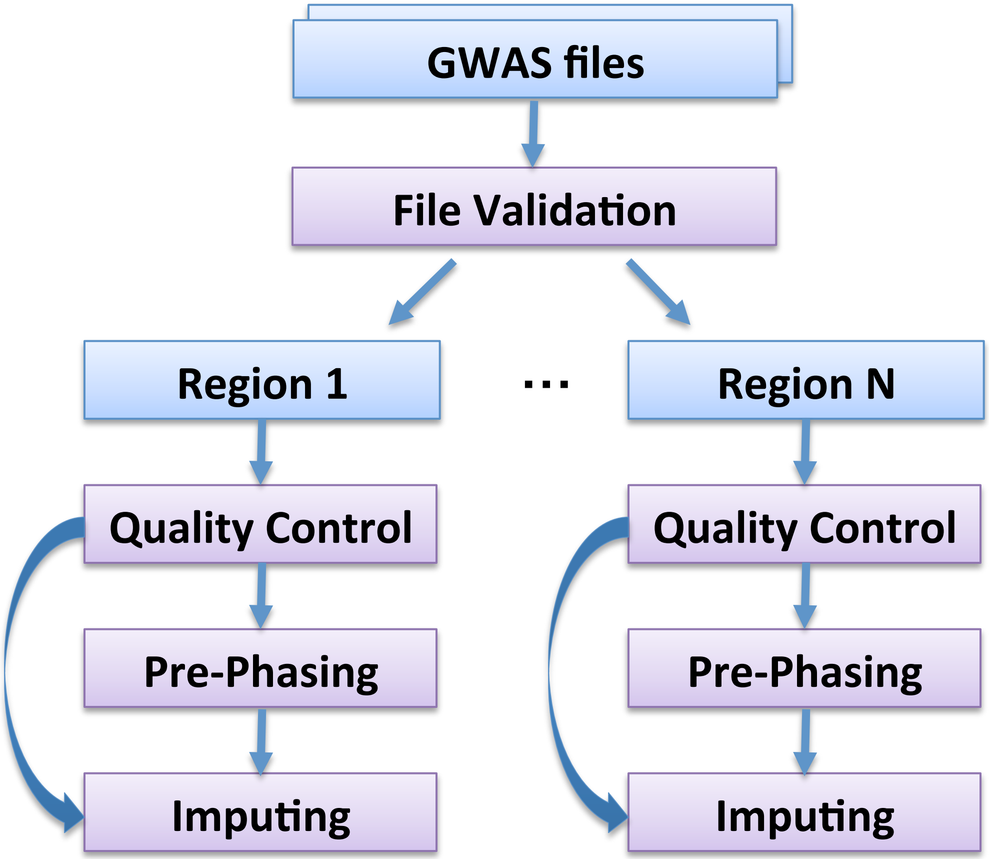
Imputation server workflow



2.) Input files are split into 10Mb ±500kb regions and checked:

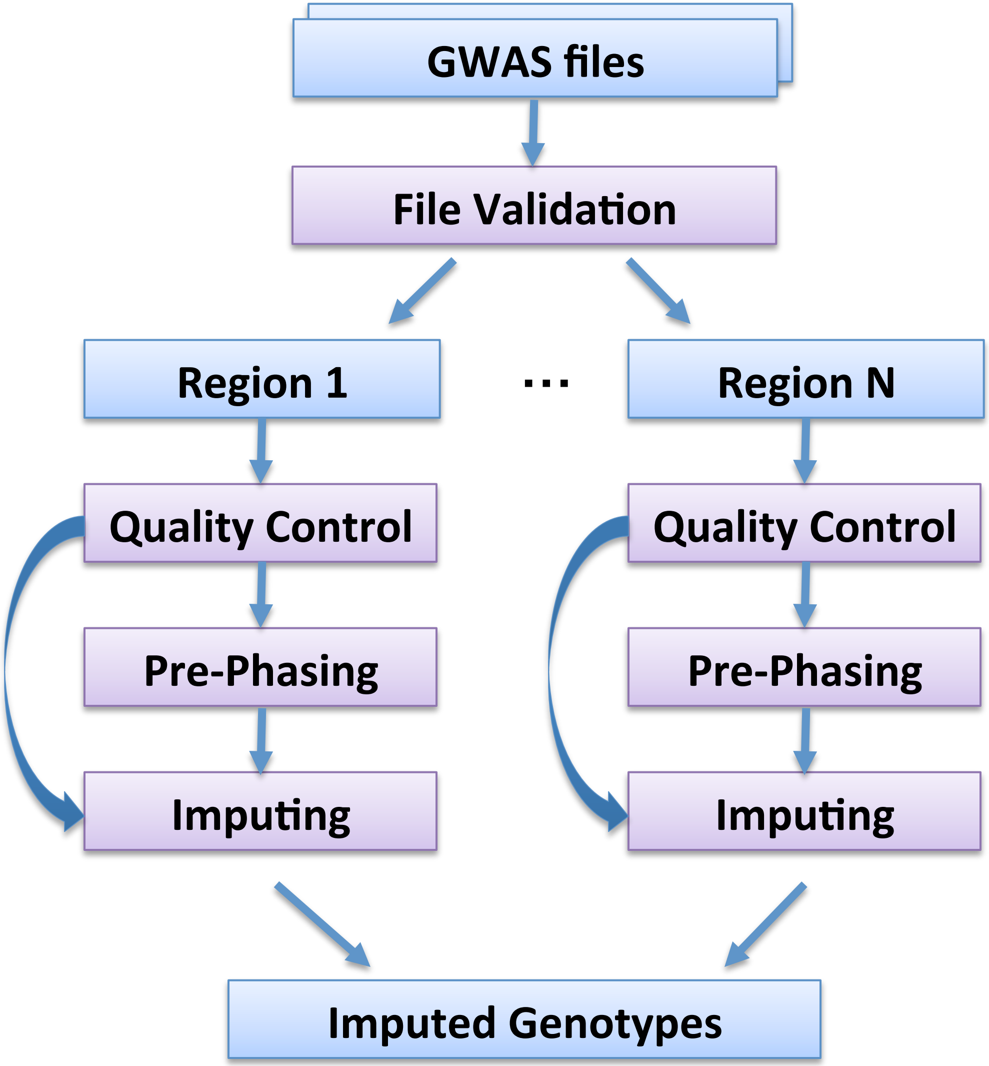
- Reference panel overlap
- Allele match
- Call rate
- …

Imputation server workflow



3.) Regions are pre-phased with SHAPEIT2 or HAPI-UR (if needed) and then imputed with minimac

Imputation server workflow



4.) Finally, regions are combined and encrypted with a one time password

Results are deleted after 7 days.

>138,000 genomes imputed to date

- Summary
  - >30 studies (N=55 to 52,189)
  - 90% of the studies used 1000 Genomes Phase 1
  - Maximum runtime start to finish: 15 days
- Current server throughput

Panel	# Haplotypes	# Samples/day
HapMap 2 (CEU)	120	150,000
1000 Genomes Phase1	2,184	15,000
1000 Genomes Phase 3	5,008	7,000

Summary

- Genotype imputation is a key step in association analysis
- Larger reference panel will improve imputation quality
- Free web service for genotype imputation:

  https://imputationserver.sph.umich.edu
- Open source, to enable set up of additional imputation servers
- >138,000 genomes imputed to date

Acknowledgments

- Cloud framework
  - Lukas Forer
  - Sebastian Schönherr

- Imputation (minimac)
  - Sayantan Das
  - Goncalo Abecasis
  - David Hinds
- Pre-phasing
  - HAPI-UR: Amy Williams
  - SHAPEIT2: Olivier Delaneau, Jean-Francois Zagury, and Jonathan Marchini