건강한 국민, 안전한 사회

한국인 참조패널 기반 imputation 분석 서비스

미래의료연구부 유전체연구기술개발과





국립보건연구원

한국인 참조패널 기반
Imputation 분석 서비스

목차

- Preprocessing input data
- Running Imputation Service
- Checking output data

Preprocessing Input Data

Introduction : Bcftools

- Utilities for variant calling and manipulating VCFs
- You can see a detailed description of bcftools in
  (https://samtools.github.io/bcftools/bcftools.html)

Commands	Mean	Basic usage
bcftools view	Subset, filter and convert VCF files	bcftools view FILE | less -S
bcftools sort	Sort VCF files	bcftools sort FILE
bcftools index	Index VCF	bcftools index FILE
bcftools concat	Concatenate VCF files from the same set of samples	bcftools concat FILE1 FILE2 bcftools concat --file-list

Preprocessing input data 1

- Checking the human genome reference version
- GRCh37 : b37, hs37, hg19, grch37, contig ID = 1 ... 22, etc.
- GRCh38 : b38, hs38, hg38, grch38, contig ID = chr1 ... chr22, etc.

```
bgzip -c KBAv2.0_KOGO_QC.vcf > KBAv2.0_KOGO_QC.vcf.gz

bcftools view KBAv2.0_KOGO_QC.vcf.gz | less -S
```

Preprocessing input data 2

National Institute of Health
Republic of Korea

- The task of Imputation requires sequential genomic positions
- Sorting VCF file by genomic position

```
bcftools sort KBAv2.0_KOGO_QC.vcf.gz \
--output-type z --output KBAv2.0_KOGO_QC_SORT.vcf.gz
```

Preprocessing input data 3

- The removal of variants with a low Minor Allele Frequency (MAF) benefits the computational and multiple testing correction burden in Association

```
bcftools view KBAv2.0_KOGO_QC_SORT.vcf.gz \
--exclude 'MAF < 0.01' \
--output-type z --output KBAv2.0_KOGO_QC_SORT_MAF0.1.vcf.gz

tabix -p vcf KBAv2.0_KOGO_QC_SORT_MAF0.1.vcf.gz
```

Preprocessing input data 4

- Imputation starts with the chromosome
- Splitting VCF file by chromosome

```
for chr in `seq 1 22`; do
  echo $chr;
  bcftools view KBAv2.0_KOGO_QC_SORT_MAF0.1.vcf.gz \
  --regions $chr \
  --output-type z  --output KBAv2.0_KOGO_QC_SORT_MAF0.1_CHR$chr.vcf.gz;
Done
```

Running Imputation Service

Access to Imputation Service website 1

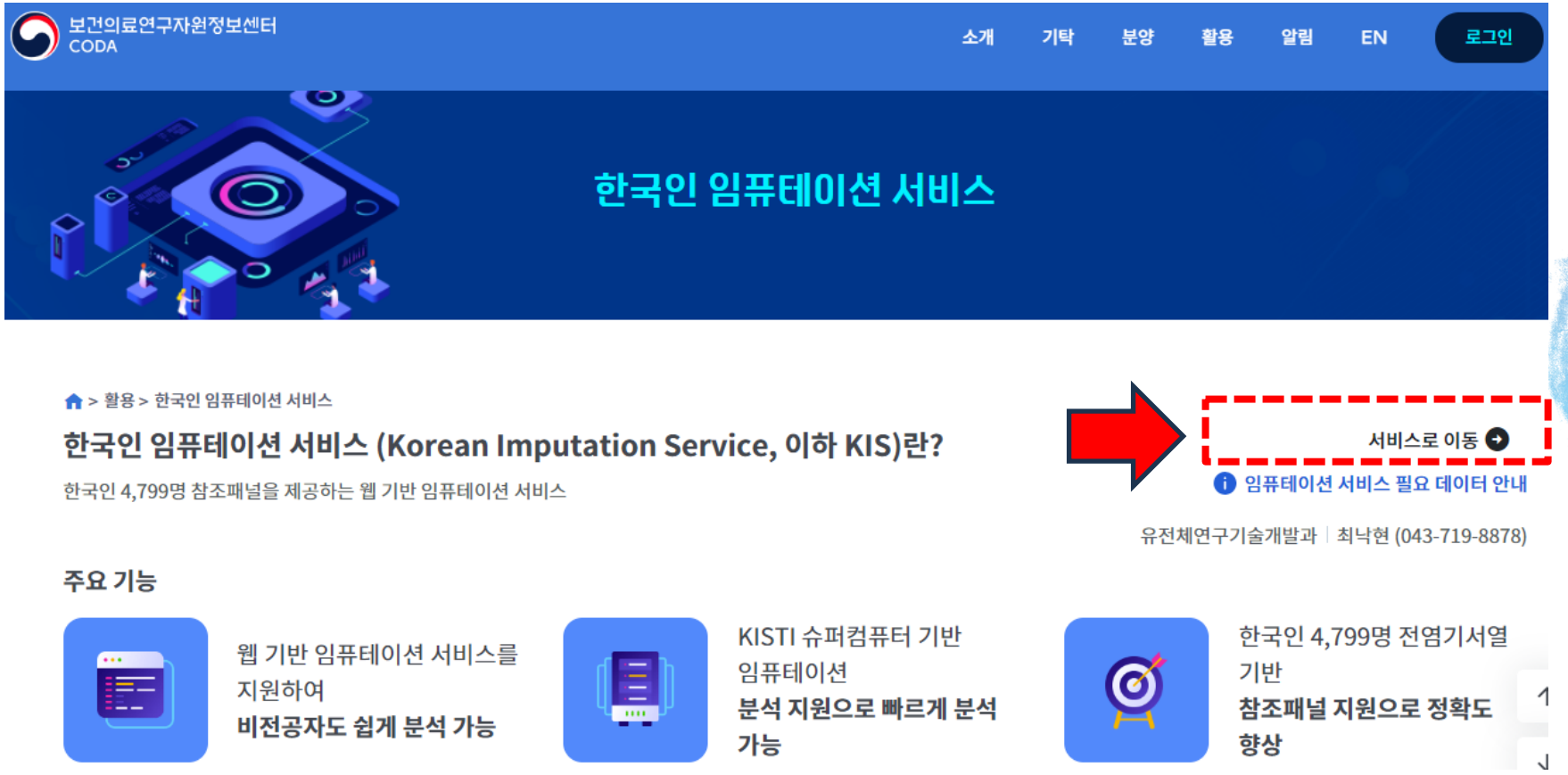
National Institute of Health
Republic of Korea

- https://coda.nih.go.kr/frt/index.do
- Click [활용] - [한국인 임퓨테이션 서비스]

Access to Imputation Service website 2

National Institute of Health
Republic of Korea

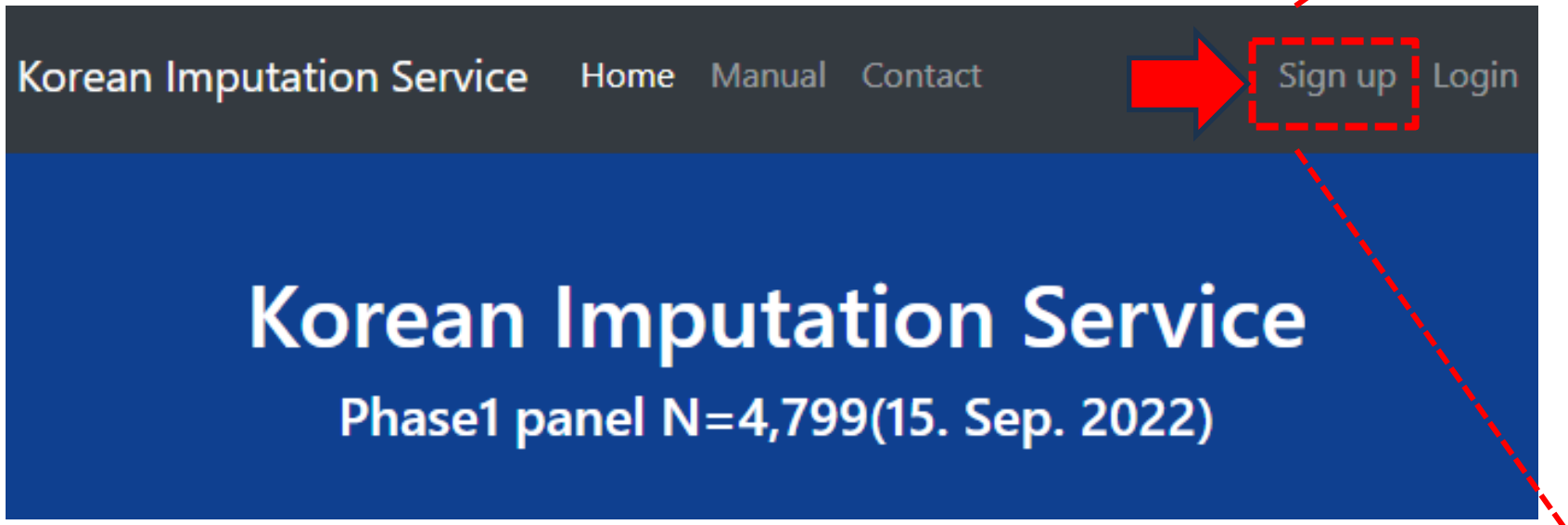
- Click [서비스로 이동]

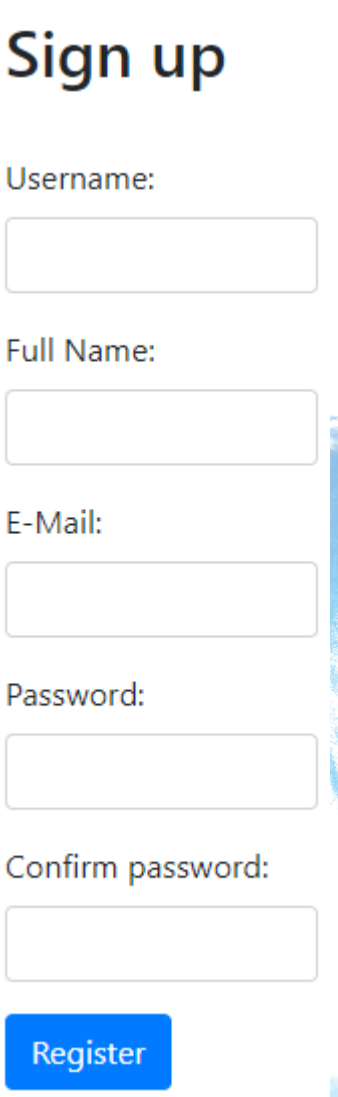


Sign up for Imputation Service

National Institute of Health
Republic of Korea

- Click [Sign up]
- Complete the blanks & Click [Register]

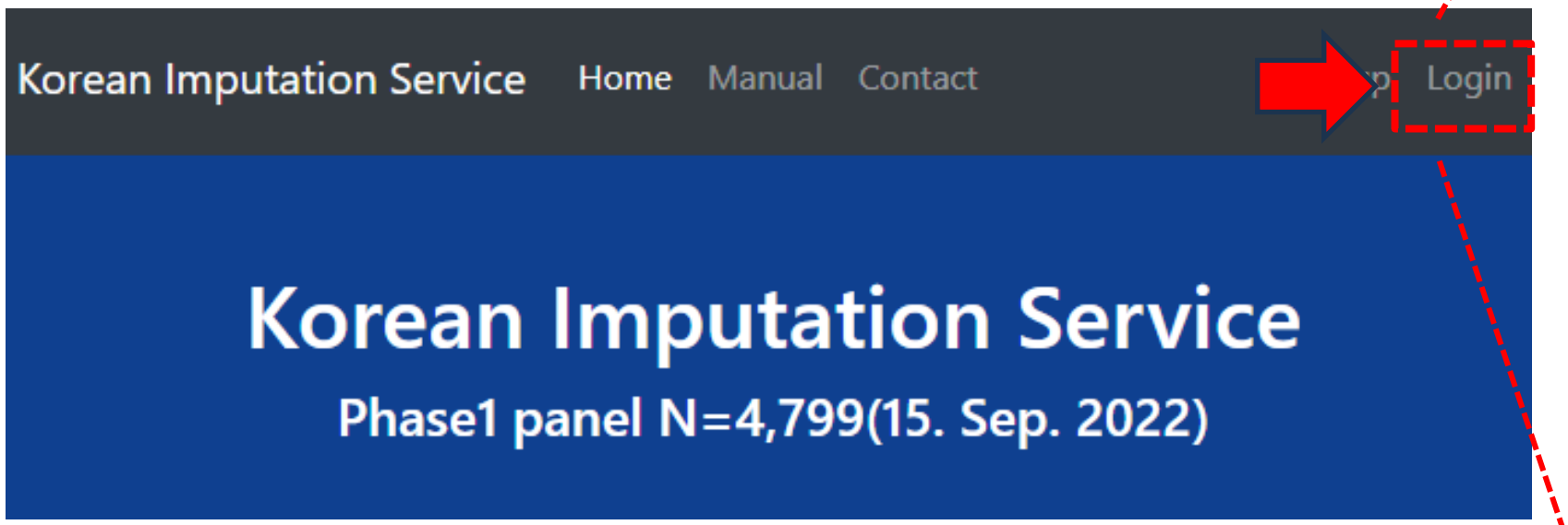


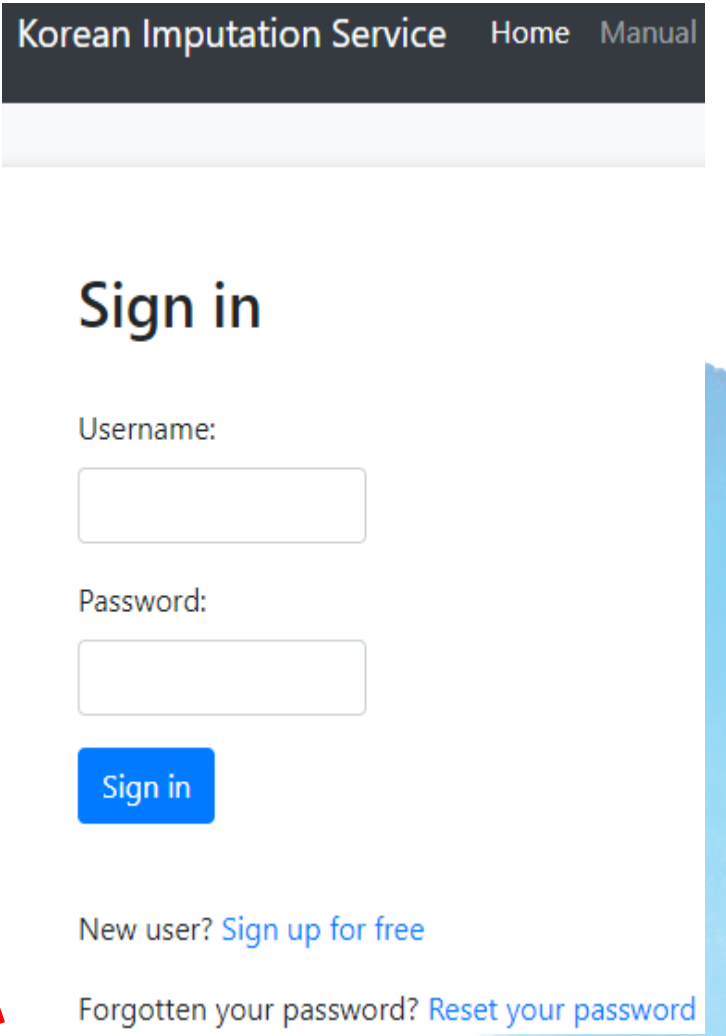


Login for Imputation Service

National Institute of Health
Republic of Korea

- Click [Login]
- Complete the blanks & Click [Sign in]

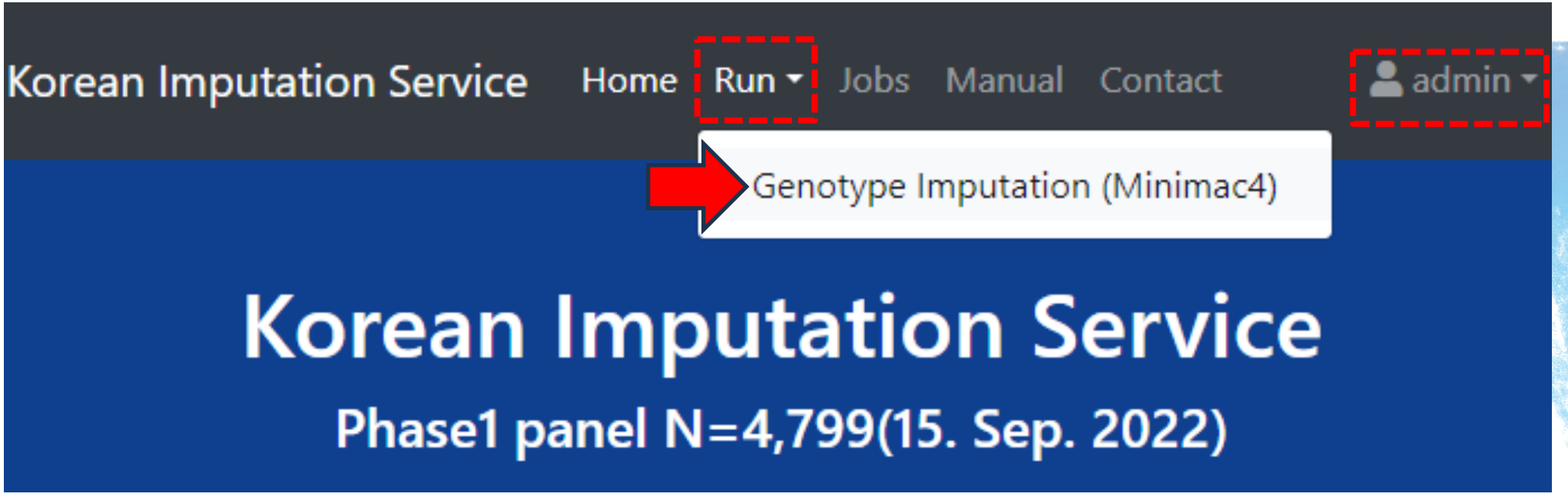




Start Imputation Service : [Run]

National Institute of Health
Republic of Korea

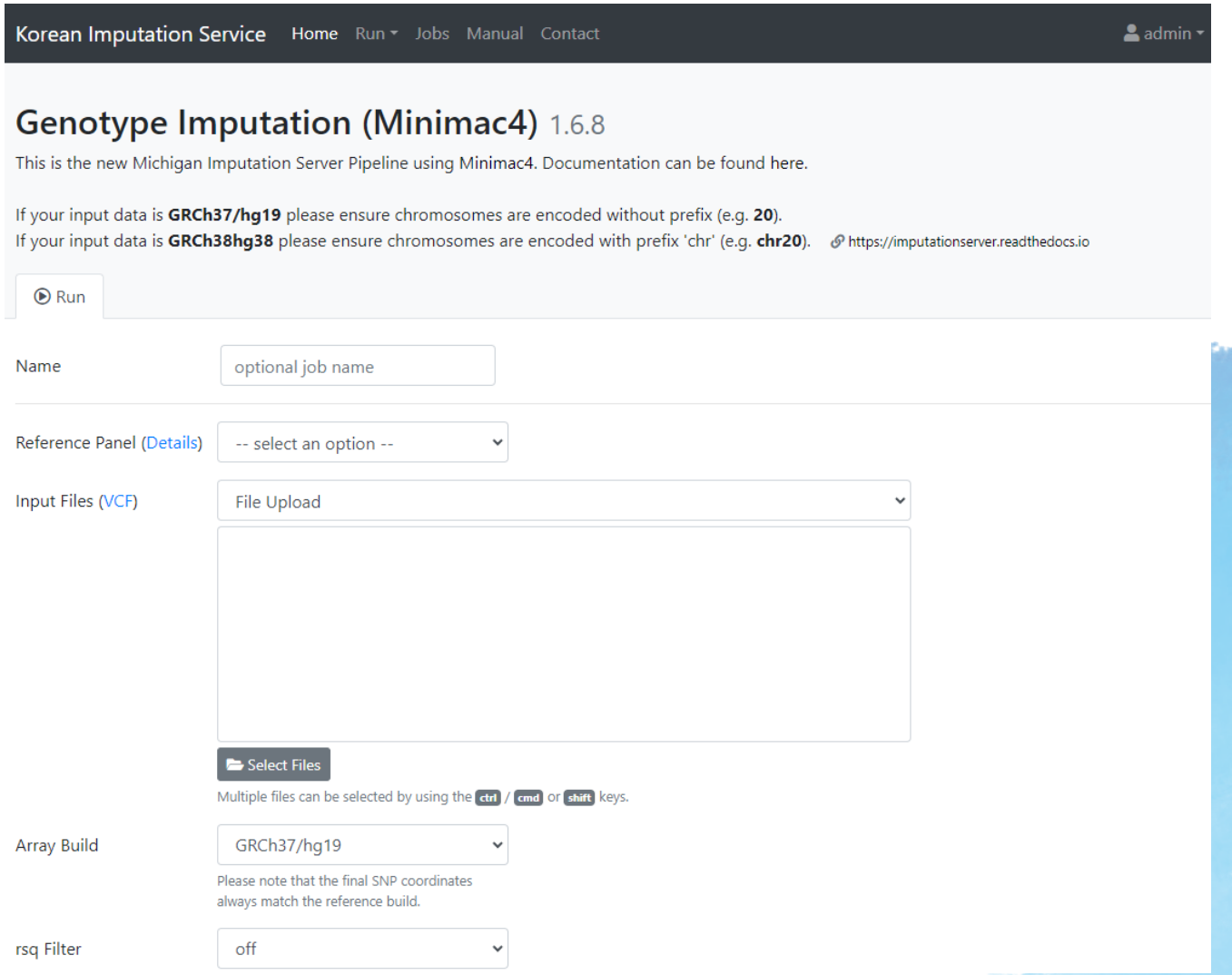
- Click [Run] & [Genotype Imputation (Minimac4)]



Start Imputation Service : Overview

National Institute of Health
Republic of Korea

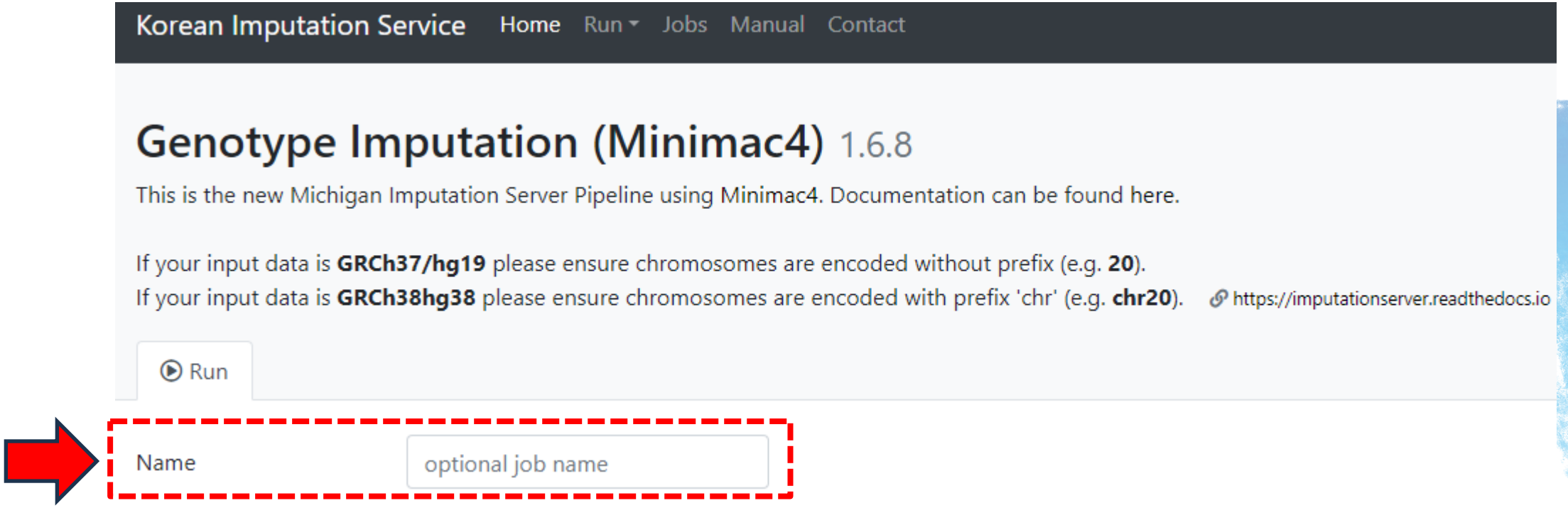
- Imputation service are based on the Michigan Imputation Server
- The service provide RUN, Details, Results, Logs information



Start Imputation Service : [Name]

National Institute of Health
Republic of Korea

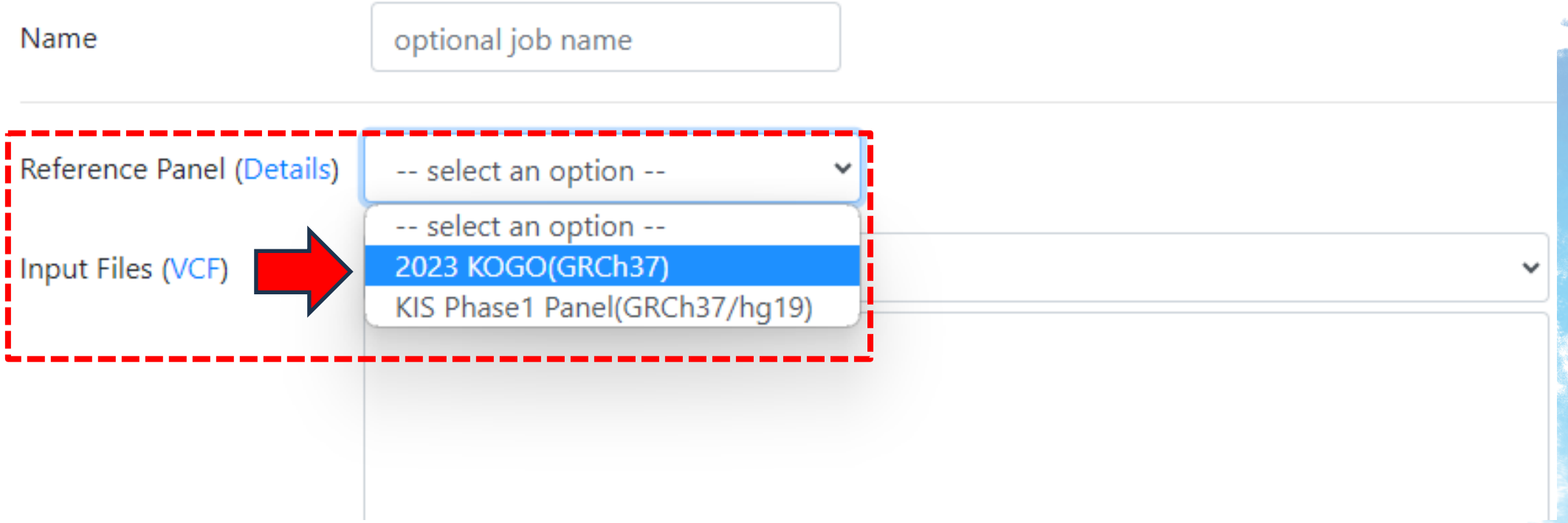
- Set study name



Start Imputation Service : [Reference Panel]

National Institute of Health
Republic of Korea

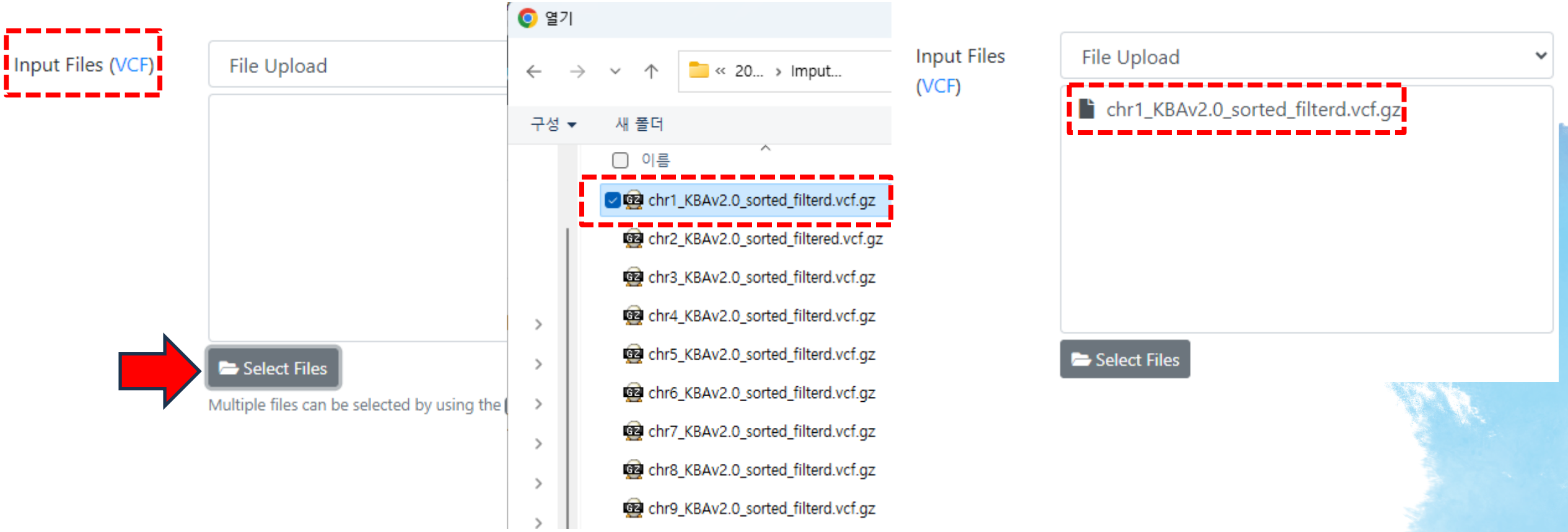
- 2024 KOGO Reference Panel was created by scaling down KIS Phase1 Panel(GRCh37/hg19)
- Select [Reference Panel] - [2024 KOGO(GRCh37)]



Start Imputation Service : [File Upload]

National Institute of Health
Republic of Korea

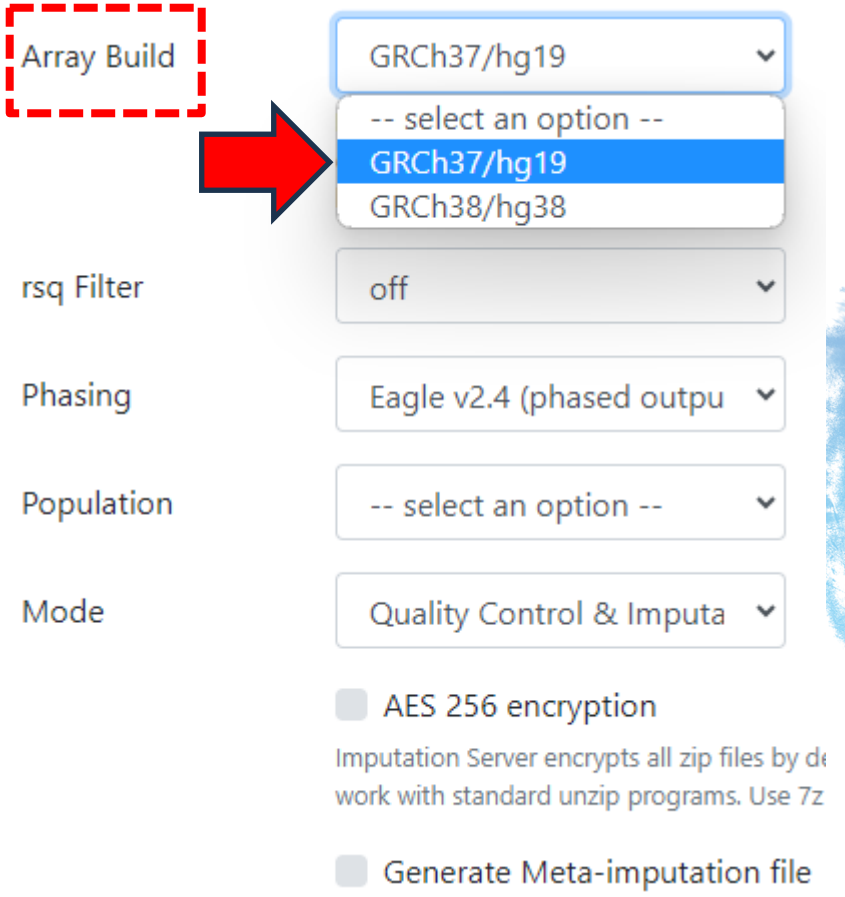
- Select input data



Start Imputation Service : [Array Build option]

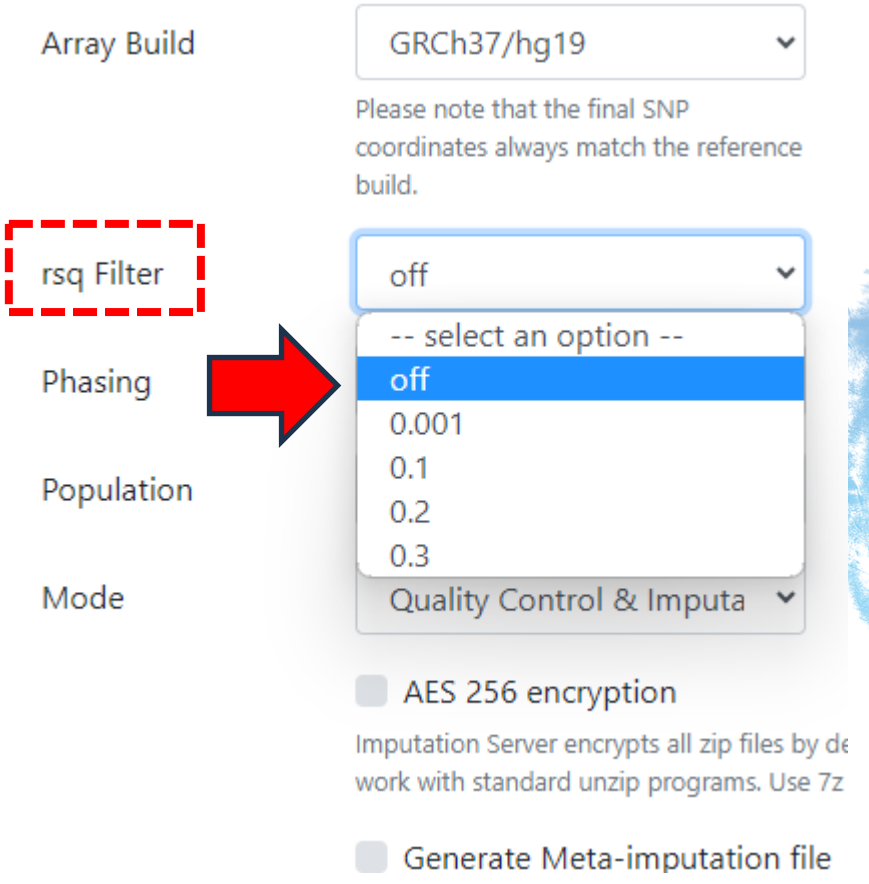
National Institute of Health
Republic of Korea

- Set [Array Build] for the build of your data according to the human genome reference version
- Input data (GRCh37/hg19) : chromosomes are encoded without prefix (e.g. 1 ... 22)
- Input data (GRCh38/hg38) : chromosomes are encoded with prefix (e.g. chr1 ... chr22)
- Server automatically updates the genome positions (liftOver) of your data



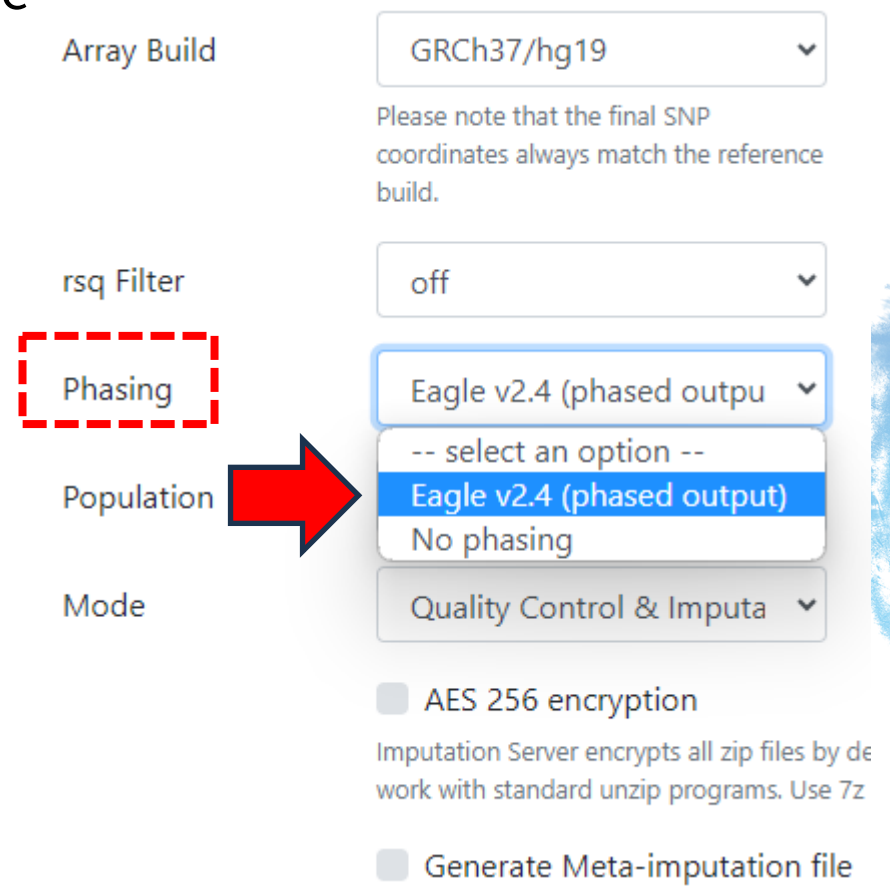
Start Imputation Service : [rsq Filter option]

- Set [rsq Filter] for output
- Purpose to minimize the file size
- Removing all imputed SNPs with a rsq-value(=imputation quality) smaller than the specified value
- Well-imputed variants (estimated rsq >= 0.8)



Start Imputation Service : [Phasing option]

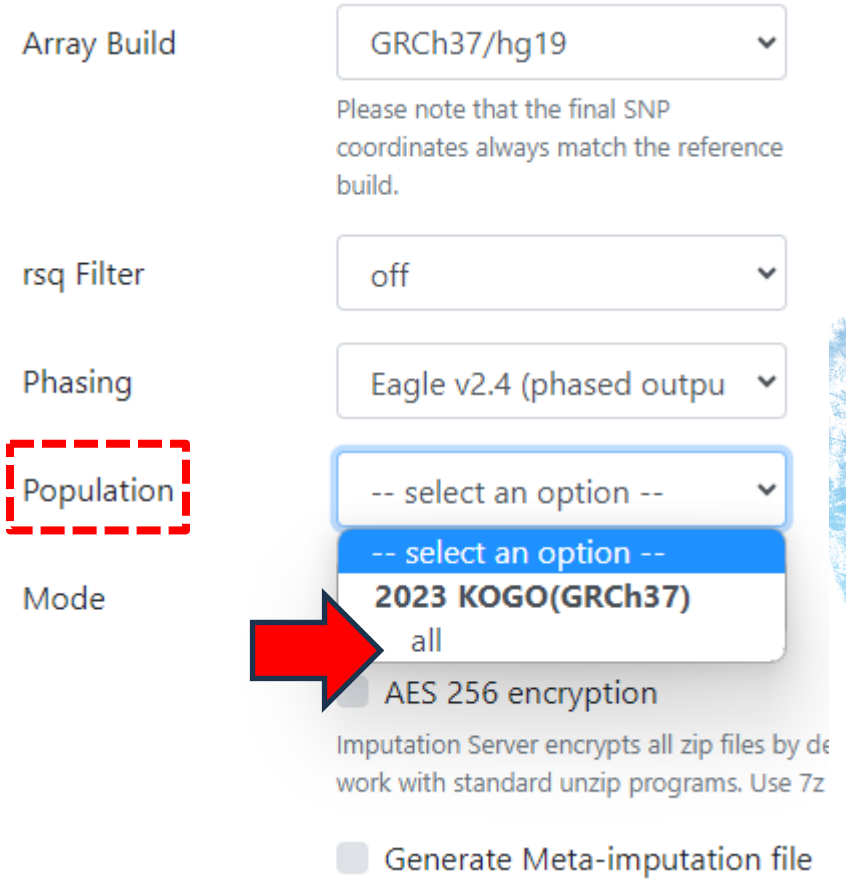
- Set [Phasing] for your data according to phase status
- Unphased data(0/0) : selecting [Eagle v2.4]
- Phased data(0|0) : selecting [No phasing]



Start Imputation Service : [Population option]

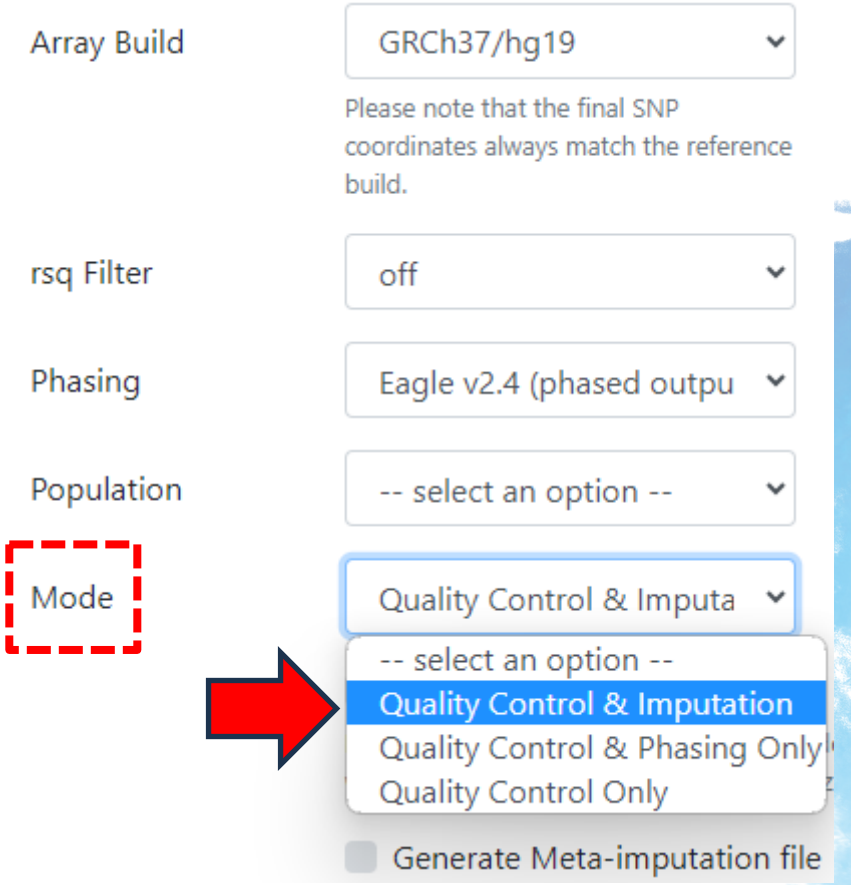
National Institute of Health
Republic of Korea

- Set [Population] for your data
- Selecting the population of your uploaded samples
- The option [all] is only supported



Start Imputation Service : [Mode option]

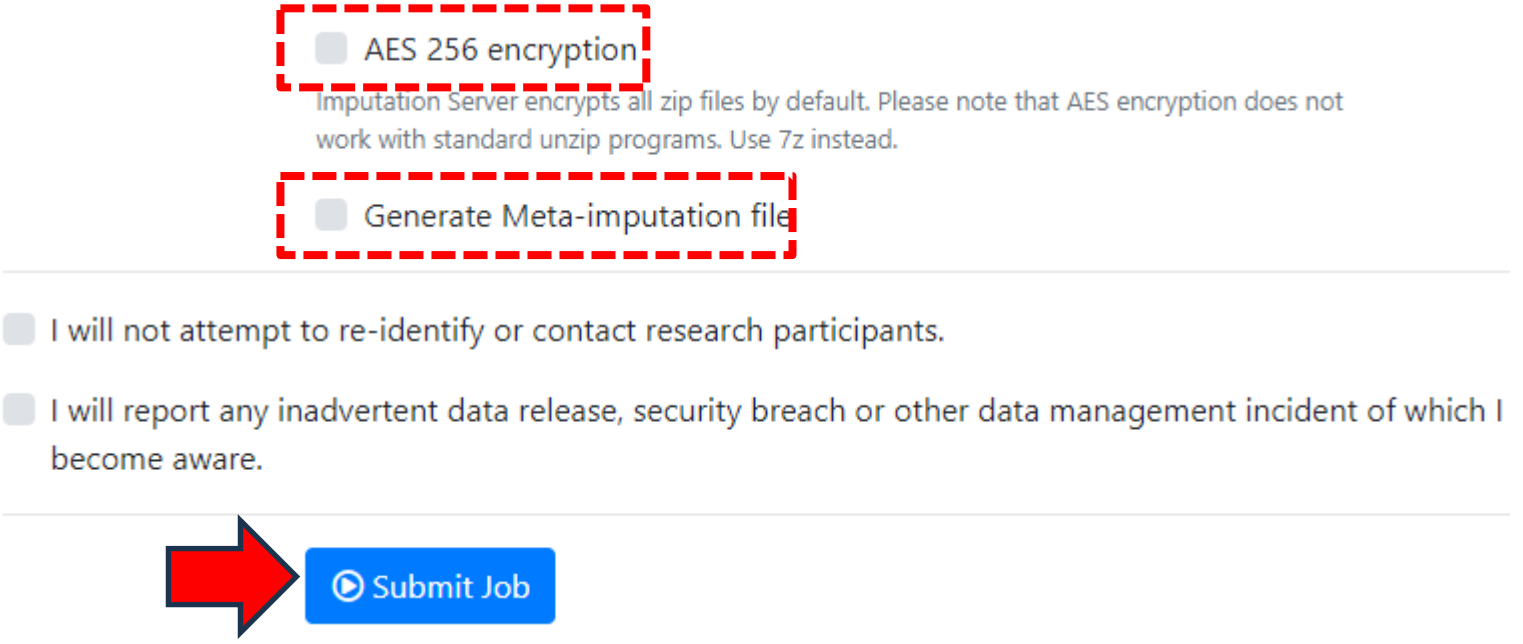
- Set [Mode] for your data
- Mode : [Quality Control & Imputation],
  [Quality Control & Phasing Only],
  [Quality Control Only]
- Selecting [Quality Control & Imputation]



Start Imputation Service : [Other options & Submit Job]

National Institute of Health
Republic of Korea

- [AES 256 encryption] : Output is encrypted by default. If you want to use AES 256 encryption, tick this checkbox
- [Generate Meta-imputation file] : Generating empirical.dose.vcf.gz for meta-imputation
- [Submit Job] : Start your imputation job

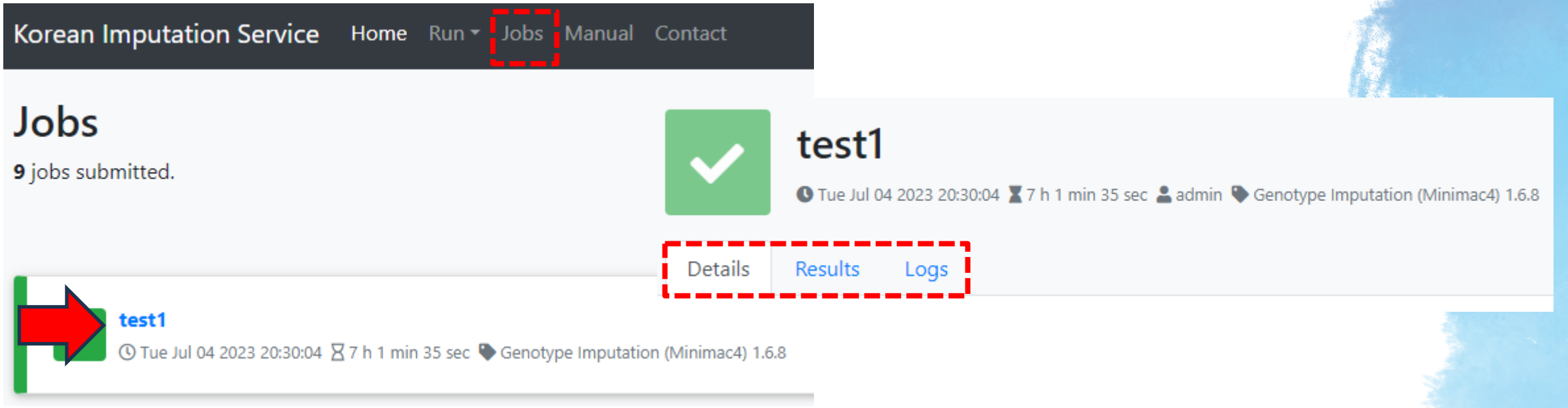


Checking Output Data

Job Menu

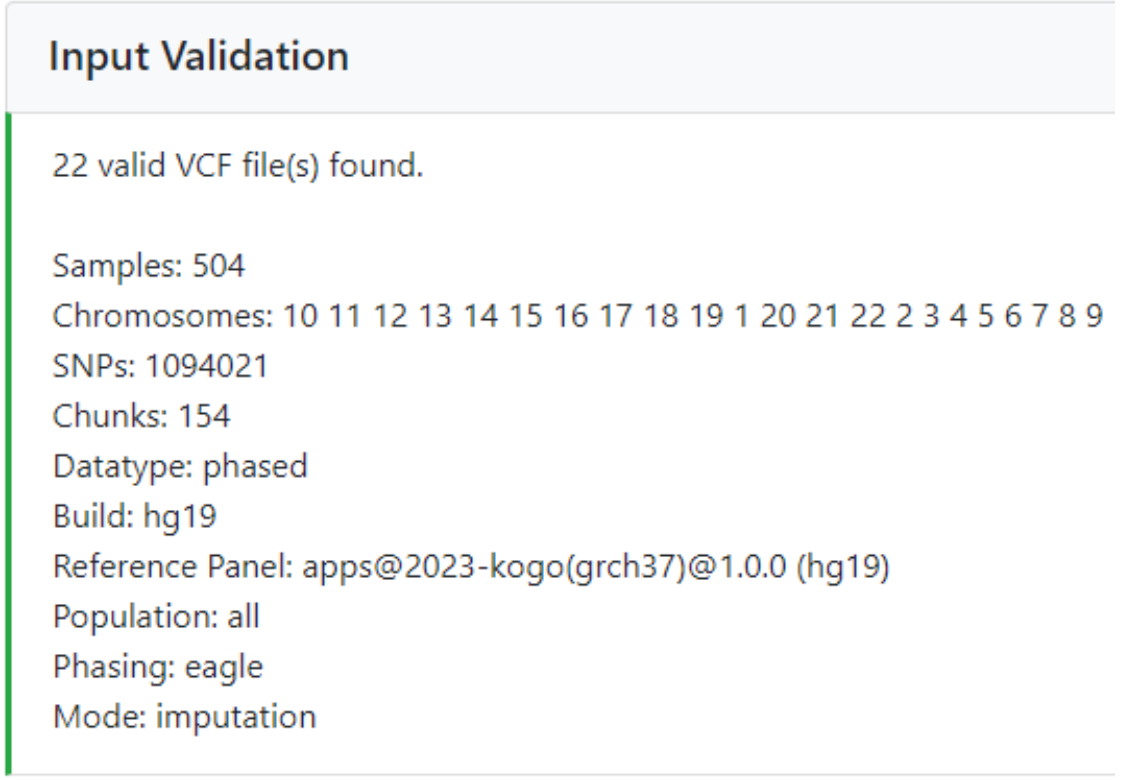
National Institute of Health
Republic of Korea

- Click [Job] - Your study name
- [Details] : Showing current status of your job
- [Results] : Showing output of your job
- [Logs] : : Showing current workflow of your job



Details : [Input Validation]

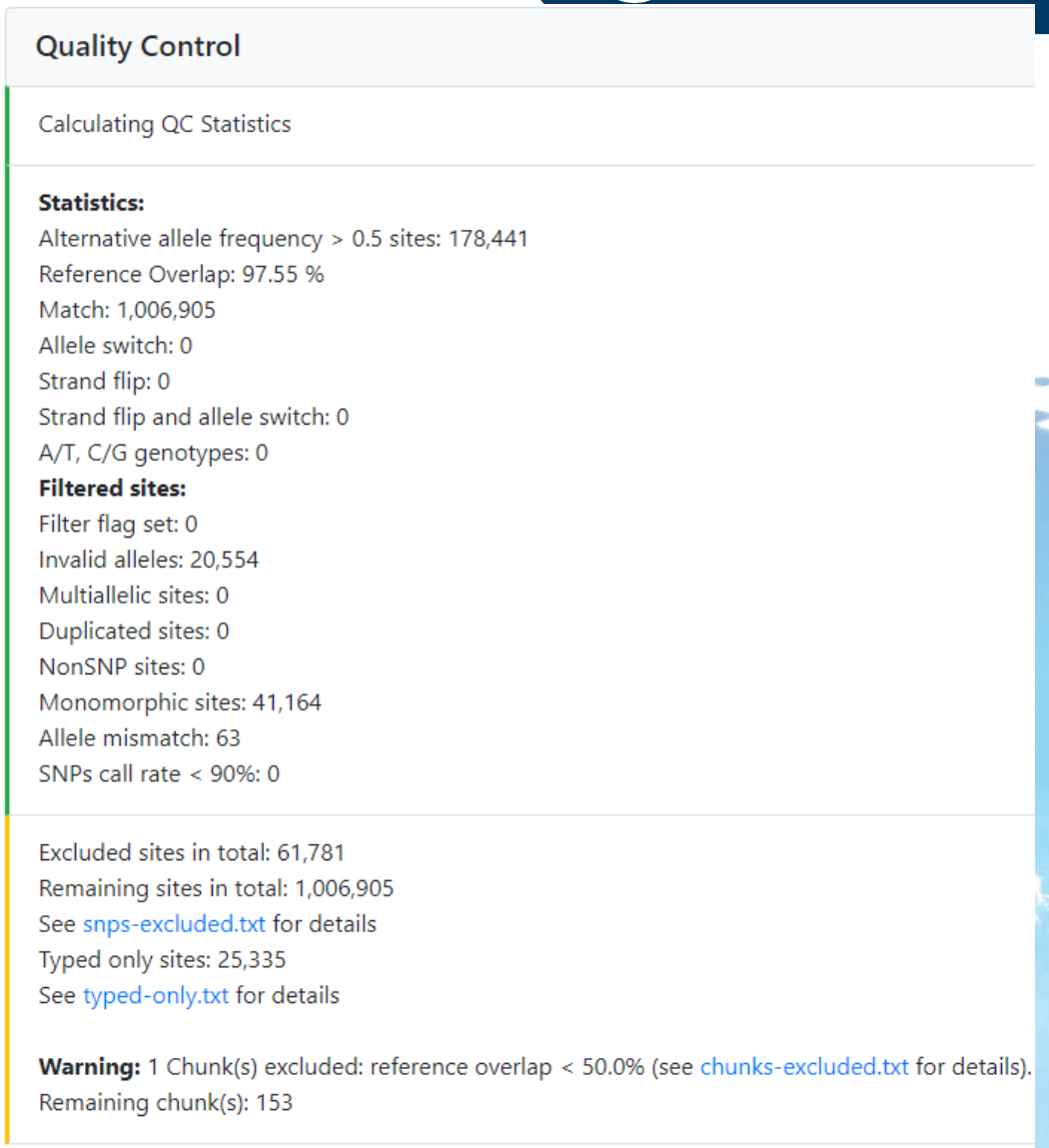
- [Input Validation] : Check uploaded files
- Calculate some basic statistics such as amount of samples, chromosomes and SNPs



Details : [Quality Control]

National Institute of Health
Republic of Korea

- [Quality Control] : Check each variant and exclude it (Invalid alleles, duplicates, indels, monomorphic sites, allele mismatch between reference panel and uploaded data, SNP call rate < 90%)
- All filtered variants are listed in files (chunks-excluded.txt, snps-excluded.txt)



Quality Control

Calculating QC Statistics

Statistics:
Alternative allele frequency > 0.5 sites: 178,441
Reference Overlap: 97.55 %
Match: 1,006,905
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: 20,554
Multiallelic sites: 0
Duplicated sites: 0
NonSNP sites: 0
Monomorphic sites: 41,164
Allele mismatch: 63
SNPs call rate < 90%: 0

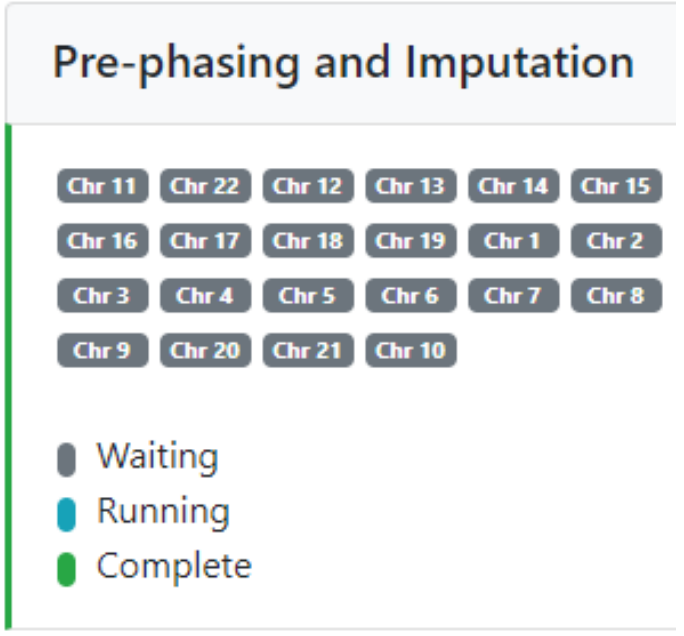
Excluded sites in total: 61,781
Remaining sites in total: 1,006,905
See snps-excluded.txt for details
Typed only sites: 25,335
See typed-only.txt for details

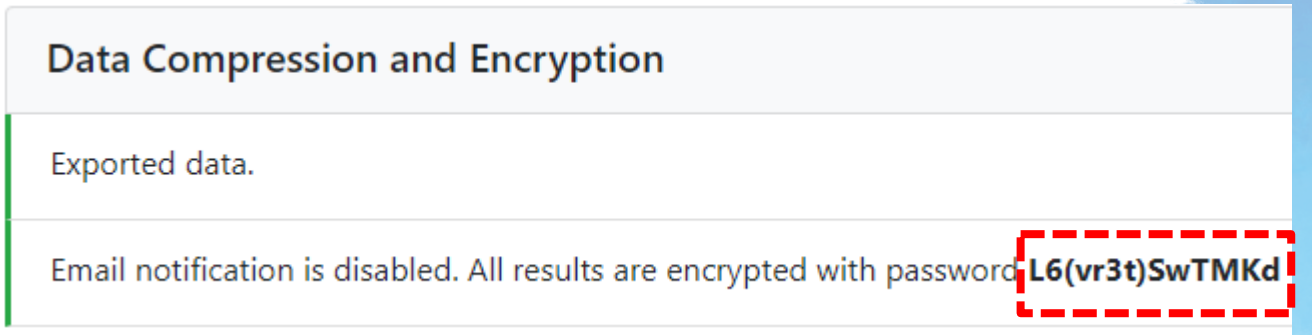
Warning: 1 Chunk(s) excluded: reference overlap < 50.0% (see chunks-excluded.txt for details).
Remaining chunk(s): 153

Details : [Pre-phasing & Imputation & Data Compression]

National Institute of Health
Republic of Korea

- [Pre-phasing] : Execute Eagle2 algorithm for each 2Mb chunk with 5Mb window size
- [Imputation] : Execute Minimac4 algorithm for each 2Mb chunk with 500 kb window size
- [Data Compression and Encryption] : Keep your password for output

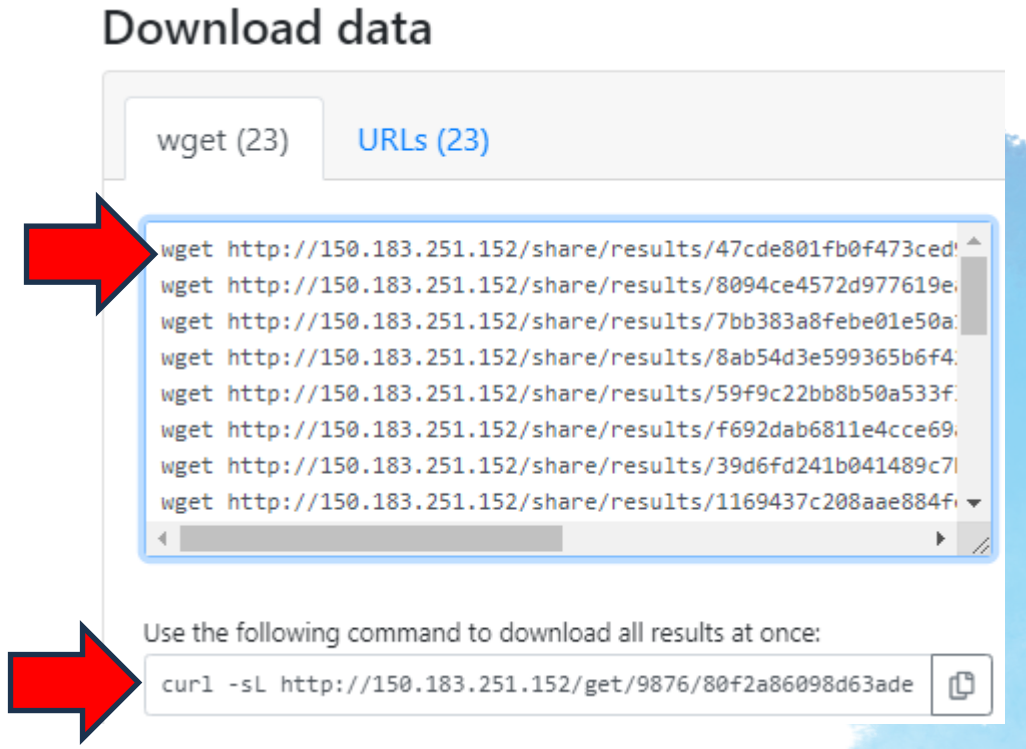




Results : [Imputation Results]

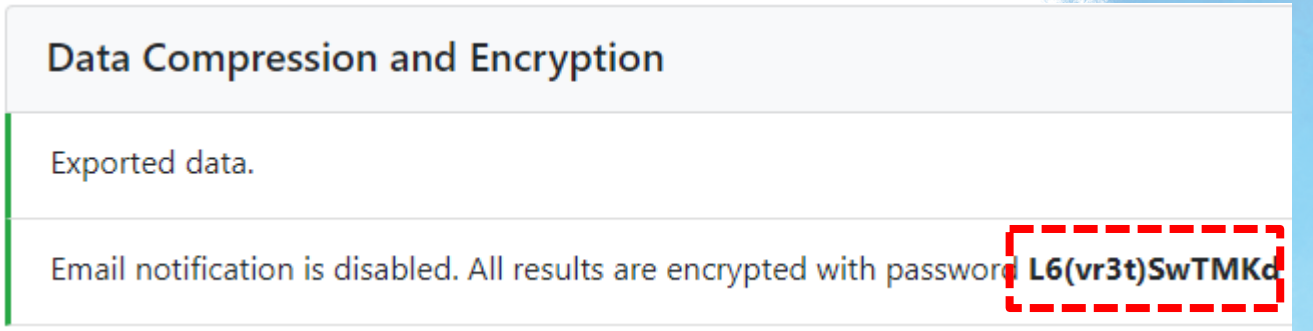
National Institute of Health
Republic of Korea

- [Imputation Results] : Download your output
- Click each output
- Copy commands & paste in command window



Output Decompression

- Decompress output with password using unzip
- chr#.dose.vcf.gz : Output for association and other analysis
- chr#.empircalDose.vcf.gz : Output for meta-imputation with MetaMinimac2
- chr#.info.gz : Summary statistics from chr#.dose.vcf.gz




```
unzip -P c6hsGfDZOtIjr2 chr_21.zip
```

Decompressed Output 1

National Institute of Health
Republic of Korea

- chr#.dose.vcf.gz : Imputation output for association and other analysis

```
##fileformat=VCFv4.1
##filedate=2023.7.4
##contig=<ID=1>
##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-1.6.8
##imputation=minimac4-1.0.2
##phasing=n/a
##panel=apps@2023-kogo(grch37)@1.0.0
##r2Filter=0.0
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00403	HG004
1	10682	1:10682:A:(	A	G	.	PASS	**AF=0.00323;MAF=0.00323;R2=0.17328;IMPUTED**	**GT:DS:HDS:GP**	0|0:0.004:0.003,0.001:0.996,0.004,0.000	0|0:0.0(
1	10687	1:10687:G:(	G	C	.	PASS	AF=0.00023;MAF=0.00023;R2=0.00428;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0,0	0|0:0:0,
1	54548	1:54548:C:T	C	T	.	PASS	AF=0.00012;MAF=0.00012;R2=0.01052;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0,0	0|0:0:0,
1	69610	1:69610:C:T	C	T	.	PASS	**AF=0.01091;MAF=0.01091;TYPED_ONLY**	GT:DS:HDS:GP	0|0:0:0,0:1,0,0	0|0:0:0,
1	71886	1:71886:T:(	T	C	.	PASS	AF=0.00036;MAF=0.00036;R2=0.02017;IMPUTED	GT:DS:HDS:GP	0|0:0:0,0:1,0,0	0|0:0:0,

Decompressed Output 2

National Institute of Health
Republic of Korea

- chr#.empircalDose.vcf.gz : Output for meta-imputation with MetaMinimac2

##fileformat=VCFv4.1															
##filedate=2023.7.4															
##contig=<ID=1>															
##INFO=<ID=TYPED,Number=0,Type=Flag,Description="Marker was genotyped AND imputed">															
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotyped alleles from Array">															
##FORMAT=<ID=LDS,Number=1,Type=String,Description="Leave-one-out Imputed Dosage : Estimated Haploid Alternate Allele Dosage assuming site was NOT genotyped ">															
##pipeline=michigan-imputationserver-1.6.8															
##imputation=minimac4-1.0.2															
##phasing=n/a															
##panel=apps@2023-kogo(grch37)@1.0.0															
##r2Filter=0.0															
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	HG00403	HG00404	HG00406	HG00407	HG00409	HG00410	HG0
1	713250	1:713250:(	G	C	.	PASS	TYPED	GT:LDS	0|0:0.001|0.001	0|0:0.000|0.000	0|0:0.001|C	0|0:0.000|C	0|0:0.000|C	0|0:0.001|C	0|1:(
1	729957	1:729957:1	T	C	.	PASS	TYPED	GT:LDS	0|0:0.000|0.000	0|0:0.000|0.000	0|0:0.000|C	0|0:0.000|C	0|0:0.000|C	0|0:0.000|C	0|1:(
1	738539	1:738539:1	T	C	.	PASS	TYPED	GT:LDS	0|0:0.000|0.000	0|0:0.000|0.000	0|0:0.000|C	0|0:0.000|C	0|0:0.000|C	0|0:0.000|C	0|1:(
1	739132	1:739132:/	A	C	.	PASS	TYPED	GT:LDS	0|0:0.002|0.008	0|0:0.001|0.002	0|0:0.005|C	0|0:0.000|C	0|0:0.002|C	0|0:0.000|C	0|0:(

- Integrate the imputed results using MetaMinimac2 (https://github.com/yukt/MetaMinimac2)

In terminal

MetaMinimac2 -i PanelA.imputed:PanelB.imputed -o A_B.meta.run

Decompressed Output 3

- chr#.info.gz : Summary statistics from chr#.dose.vcf.gz

SNP	REF(0)	ALT(1)	ALT_Frq	MAF	AvgCall	Rsq	Genotyped	LooRsq	EmpR	EmpRsq	Dose0	Dose1
1:52955:A:G	A	G	0.00023	0.00023	0.99977	0.00865	Imputed	-	-	-	-	-
1:54548:C:T	C	T	0.00012	0.00012	0.99988	0.01052	Imputed	-	-	-	-	-
1:69610:C:T	C	T	0.01091	0.01091	-	-	Typed_Only	-	-	-	-	-
1:71886:T:C	T	C	0.00036	0.00036	0.99964	0.02017	Imputed	-	-	-	-	-

In terminal

pigz -d ./chr#.info.gz

Rscript output.R Start

```
[1] "The number of imputed variants ..."
# A tibble: 3 × 2
# Groups:   Genotyped [3]
  Genotyped        n
  <fct>        <int>
1 Typed_Only   25325
2 Genotyped  1006901
3 Imputed     581332
[1] "The number of WellImputed variants per AF ..."
# A tibble: 8 × 3
# Groups:   WellImputed, Range [8]
  WellImputed Range            n
  <chr>       <fct>        <int>
1 Imputed     AF<=0.1     283230
2 Imputed     0.1<AF<=0.5  82793
3 Imputed     0.5<AF<=5    32601
4 Imputed     5<AF         11913
5 WellImputed AF<=0.1       4858
6 WellImputed 0.1<AF<=0.5  23797
7 WellImputed 0.5<AF<=5    33976
8 WellImputed 5<AF        108164
```

Preprocessing output for Association

- Merge all output into one file for Association

```
bcftools concat --file-list merge_list.txt \
--output-type z --output KBAv2.0AB_KOGO_IMP.vcf.gz --threads 10
bcftools index KBAv2.0AB_KOGO_IMP.vcf.gz
```

건강한 국민, 안전한 사회

국립보건연구원