ImputeGen: Pipeline for phasing and imputation analysis

User manual and guide

ImputeGen is a collection of bash scripts that automatically checks genotype inconsistencies and does filtering, phasing and imputation from lower density genotype panels (or low passing sequencing) to higher density panels (or whole genome sequecing level). ImputeGen also allows down-scaling panels from higher density to lower density level to test imputation methods. It is optimized to run in a local server machine or HPCC with slurm scheduler.

Note: The previous name of ImputeGen was ImputeMe. It has changed to avoid name clash with a homonymous project developed by Dr Lasse Folkersen group (https://impute.me).

Citation

Savegnago, R. P. ImputeGen: Pipeline for phasing and imputation analysis. 2021. GitHub repository, https://github.com/rodrigopsav/ImputeGen

Download citation

↥ back to top

Running ImputeGen first time: quick steps

```
# Download IVDP
git clone https://github.com/rodrigopsav/ImputeGen.git
wait

# Install IVDP dependencies
cd ImputeGen/install_imputegen_dependencies
./install_imputegen_dependencies.sh -d .
cd ..
wait

# Running ImputeGen in a local server (parFile1_50k_seq.txt)
./imputeGen.sh -p params/parFile1_50k_seq.txt

# Running ImputeGen in HPCC with slurm (parFile1_50k_seq.txt)
# ./imputeGen.sh -p params/parFile1_50k_seq.txt -c params/configSlurm.txt
```

↥ back to top

Download ImputeGen

```
git clone https://github.com/rodrigopsav/ImputeGen.git
```

↥ back to top

Install ImputeGen Dependencies

All ImputeGen dependencies (except minimac3 and impute5) are installed in two conda environments: imputegen and r-env4.0. Before run ImputeGen, you **MUST**

install the dependencies (ImputeGen/install_impute_dependencies/install_impute_dependencies.sh file) even if you have already installed the programs in your machine. To install ImputeGen dependencies, run:

```
cd ImputeGen/install_impute_dependencies
./install_imputegen_dependencies.sh -d <directory/to/install/imputegen/dependencies>
```

Please visit the official websites of the programs used on ImputeGen to learn more about them:

--**Bioinformatic programs** BCFtools Beagle Eagle GLIMPSE Minimac3 Plink1.9 Plink2 SHAPEIT4 HTSlib VCFtools

--**Programs for file manipulation** dos2unix ghostscript img2pdf pigz sed vcflib zlib

-- **R packages**

data.table dplyr ggplot2 gridextra ggrepel vcfR

↥ back to top

Install Minimac3 and impute5

Minimac3 and impute5 are not available as Conda package. It must be installed separately following the instructions on their websites: Minimac3 and impute5.

Installing minimac3

To install minimac3, use:

```
wget [ftp://share.sph.umich.edu/minimac3/Minimac3.v2.0.1.tar.gz](ftp://share.sph.umich.edu/minimac3/Minimac3.v2.0.1.tar.gz)
tar -xzvf Minimac3.v2.0.1.tar.gz
cd Minimac3/
make
```

After installing minimac3, copy it to the binary folder of imputegen conda environment:

```
conda activate imputegen
cp ./Minimac3/bin/Minimac3 $CONDA_PREFIX/bin/minimac3
```

NOTE: Minimac3 was designed to imputation analysis in human (i.e. 23 chromosomes). If you want to use Minimac3 with other species, please follow the instructions published in in this paper:

Al-Mamun HA, Bernardes PA, Lim D, Park B, Gondro C (2017). A guide to imputation of low density single nucleotide polymorphism data up to sequence level. Journal of Animal Breeding and Genomics, 1(1):59-68.

Installing impute5

1. Download impute5 by clicking "download" > "direct donwload"
2. Move impute5_*_static to binary folder of imputegen conda env:

```
conda activate imputegen
mv impute5_*_static $CONDA_PREFIX/bin/impute5
```

↥ back to top

Using conda environments without ImputeGen

You can use all the programs without ImputeGen, by activating the proper conda environment:

```
conda activate imputegen
conda activate r-env4.0
```

All the bioinformatic programs are in imputegen environment and R programs and its packages are in r-en4.0. To deactivate a conda environment, use:

```
conda deactivate
```

It is usual to activate only one conda environment at time to avoid incompatibilities. However, if you want to use more than one at once, type:

```
conda activate imputegen
conda activate --stack r-env4.0
```

NOTE: If you activate two conda envs at the same time, some programs cannot work properly.

↥ back to top

Run ImputeGen

To run ImputeGen in a local machine, type:

```
./imputegen.sh -p <imputegen parameter file>
```

and to run on a HPCC with slurm scheduler, type:

```
./imputegen.sh -p <imputegen parameter file> -c <slurm parameter file>
```

The parameter file is mandatory. The slurm config file will activate the jobs submission on HPCC with slurm.

NOTE: Do not use nohup and & to run ImputeGen in the background. After ./imputegen.sh starts, all the processes will be automatically sent to the backgroud (in a local machine).

↥ back to top

Check log files

↥ back to top

ImputeGen parameter file

The parameter file contain variables that you have to assign options to them:

- **REF_GENOME**: File directory for reference genome **(*.fa)**. The reference genome must be **uncompressed file** and not **(*.fa.gz)**.

  > **NOTE:**REF_GENOME=none will skip the step to check switched alleles (REF / ALT).

- **REF_PANEL_DIR**: Directory path with the reference panel for imputation.
- **REF_PANEL_PREFIX**: Prefix of file names of the reference vcf panels.
- **VAL_PANEL_DIR**: Directory path with the validation panel to calculate the imputation accuracy (only if you are testing imputation methods).
- **VAL_PANEL_PREFIX**: Prefix of file names of the validation vcf panels (only if you are testing imputation methods).
- **GENO_PANEL**: Low density panel to be imputed (vcf format)
- **BAM_DIR**: Directory path with the bam files

  > **NOTE:**It works with whatshap and stitch (*not implemented yet*)

- **PEDIGREE**: File directory for the pedigree file.

  > **NOTE:**It works with whatshap - PLINK-compatible PED files. (*not implemented yet*)

- **FINDHAPF90**: Directory path to findhapf90 input files (*not implemented yet*)
- **GENETIC_MAP**: path to genetic linkage map file. (Default: 1cMperMb)
- **OUTPUT_DIR**: Output directory folder (Default is the home directory if left blank: OUTPUT_DIR=).
- **OUTPUT_NAME**: Analysis name. Choose a single name without spaces (Default is user name if left blank: OUTPUT_NAME=).
- **PHASING_PROGRAM**: programs for phasing. Options: eagle, beagle, shapeit4, glimpse, stitch (*not implemented yet*), whatshap (*not implemented yet*), none

  > **NOTE:**PHASING_PROGRAM=none will skip the phasing step.

- **IMPUTE_PROGRAM**: programs for imputation. Options: minimac3, beagle, impute5, none

  > **NOTE:**IMPUTE_PROGRAM=none will skip the imputation step (use it if you already ran imputation before and something went wrong with the calculation of imputation accuracy).

  - **CHROM_SET**: Number of autosome pairs of the specie.

- **INCLUDE_CHROM_X**: =no
- **SELECT_CHROM**: Chromosome names in this format: [chrom] or [chrom]:[start]-[end]. If more than one region, separate them using ',' (e.g. CHROM=1,4,6:6000-7000,X,Y,MT).
- **MAF**: Minimum allele frequency for the filtering step (Use MAF=0 if vcf was already filtered for maf).
- **MISSING**: Maximum genotype-based missing rate to keep a locus (Use MISSING=0 if vcf was already filtered for missing rate genotypes).

  > **NOTE:**MAF=0 AND MISSING=0 will skip the filtering step.

- **DOWN_SCALING**: Proportion of samples to be used in the down-scaled panel from high density to low density level. It splits the high-density panel into training (reference population), test (imputation population), and validation (observed genotypes for te imputation population). (Default: 0)
- **THREADS**: Number of threads.

  > **NOTE**: it works only on local servers.

- **MEM**: Amount of memory.

  > **NOTE**: it works only on local servers.

- **BATCH**: Number of analysis at the same time, i.e. number of imputations to be done at the same time.

  > **NOTE**: it works only on local servers.

↥ back to top

Slurm parameter file

↥ back to top

ImputeGen Output

There are two main folder in ImputeGen output path: (1) a data folder with all the files created with ImputeGen, and (2) a log folder with log files for each chromosome. The data folder contain the subfolders:

- **accuracy**: It contains three subfolders:
  - *r2_maf*: It has the R2 and minimum allele frequency (MAF) only for imputed loci.
  - *genotypes*: **If DOWN_SCALING=yes**, it has the genotypes **only for imputed loci** from vcf files of Imputed folder in 0|0, 0|1, 1|0, 1|1 format. **If DOWN_SCALING=no**, it has **all the genotypes** from vcf files of Imputed folder in 0|0, 0|1, 1|0, 1|1 format.
  - *012*: **If DOWN_SCALING=yes**, it has the genotypes **only for imputed loci** from vcf files of Imputed folder in 012 format. **If DOWN_SCALING=no**, it has **all the genotypes** from vcf files of Imputed folder in 012 format.
- **data**:
  - *switchRefAlt*: It contains vcf files corrected for switched RE/ALT.
  - *filtered*: It contains the filtered vcf files.
  - *phased*: It contains the phased vcf files.
  - *imputed*: It contains the imputed vcf files.
  - *downScaled*: if DOWN_SCALING=yes, it contains the down-scaled panels from high to low density level to test imputation methods. It also contains the train, and validation sets used for imputation analysis and evaluation of its accuracy.
  - *Test set*: It contains the known SNPs on low density panel with a subset of individuals from the high density panel. The percentage of individuals in the test set is controlled by the **P** variable of parameter file.
  - *Train set*: It has *1-P* individuals and all known SNPs from high density panel.
  - *Validation set*: It has the same animals in the test set and the known SNPs from the high density panel.

  The train and test sets are used to impute the test set (down-scaled SNP panel) to the train set level. The validation set is used to calculate imputation accuracy metrics comparing the known genotypes on validation set with imputed genotypes on test set.
- **params**:
- **log**:
- **imputeGen_*_report.zip**:
- **Imputed**: It contains the imputed vcf files.

↥ back to top

ImputeGen Examples

ImputeGen comes with an example folder with high density and low density SNP panels for the last autosomal chromosome of cattle (chromosome 29). The HD and LD panels are based on a 777K and 50K bovine SNP panels, respectively. The HD panel has 100 individuals and the LD panel has 20 individuals that is not in HD panel. The 120 individuals are from SRA run, which were downloads with SRAtoolkit.

Example 1

Check switched alleles (REF_GENO) filtering step (MAF=0.005 and MISSING=0.2), phasing with eagle, imputation with beagle

Imputation from low density to high density panel. Both HD and LD panels will be filtered using VCFtools, phased with Eagle, and imputed using Beagle. The filtering parameters are: **MAF**=0.01, **CALLRATE_HD**=0.7 and **CALLRATE_LD**=0.7. To run:

```
./ImputeGen_run.sh -p examples/parEx1.txt
```

↥ back to top

Example 2

Check switched alleles (REF_GENO) filtering step (MAF = 0.005 and MISSING = 0.2), phasing with beagle, imputation with beagle

Same as example 1 but down-scaling the high density to low density panel (**DOWN_SCALING=yes**). So, considering 100 individuals in high density panel, imputation will use 70 individuals for the reference population and 30 for imputation (**P=0.3**). To run:

```
./ImputeGen_run.sh -p examples/parEx2.txt
```

↥ back to top

Example 3

Check switched alleles (REF_GENO) filtering step (MAF = 0.005 and MISSING = 0.2), phasing with shapeit4, imputation with beagle

Same as example 2 but using minimac3 for imputation (**IMPUTE_PROGRAM=minimac3** and **MINIMAC3=/home/work/rps/Bio_prog/minimac3/bin/minimac3**). To run:

```
./ImputeGen_run.sh -p examples/parEx3.txt
```

↥ back to top

Example 4

Check switched alleles (REF_GENO) filtering step (MAF = 0.005 and MISSING = 0.2), phasing with glimpse, imputation with beagle

Do not run examples/parEx4.txt with the data in examples folder. This is only to show how it would be an analysis in which both HD and LD panels had already filtered and phased before. To skip these steps on ImputeGen, choose **PHASE_HD=no** and **PHASE_LD=no**.

↥ back to top

Example 5

Check switched alleles (REF_GENO) filtering step (MAF = 0.005 and MISSING = 0.2), phasing with eagle, imputation with minimac3

↥ back to top

Example 6

Check switched alleles (REF_GENO) filtering step (MAF = 0.005 and MISSING = 0.2), phasing with eagle, imputation with impute5

↥ back to top

Example 7

Skip switched alleles (REF_GENO=none), skip filtering step (MAF=0 and MISSING=0), phasing with eagle, imputation with minimac3

Example 8

Skip switched alleles (REF_GENO=none), skip filtering step (MAF=0 and MISSING=0), phasing with eagle, and skip imputation (IMPUTE_PROGRAM=none)

Example 9

Downsampling Reference panel, with 70% train, 30% test (DOWNSAMPLING=0.3), check switched alleles (REF_GENO) filtering step (MAF = 0.005 and MISSING = 0.2), phasing with eagle, imputation with minimac3

↥ back to top

Caveats

- **Running ImputeGen**
  -Call ImputeGen from its folder. Avoid run it out of ImputeGen folder.
  -Set up carefully all directories path in ImputeGen parameter files. Use absolute path to indicate them or relative path if the data and the output are inside ImputeGen folder.
- **ImputeGen parameter file**
  -Keep all the vertical bar "|" before variables in ImputeGen parameter file.

↥ back to top

Troubleshooting

↥ back to top