ImputeMe: Pipeline for phasing and imputation analysis

User manual and guide

ImputeMe was designed for phasing and imputation analysis. It automatically filter and phase the high and low density panels using VCFtools and Eagle, respectively and does the imputation with Beagle or Minimac3. ImputeMe allows down-scaling panels from high density to low density level to test imputation methods. It is possible do the whole process for several chromosomes or regions at once. The following softwares must be installed before using ImputeMe:

Download ImputeMe

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

Install ImputeMe dependencies

ImputeMe uses a set of bioinformatic tools to filter, phasing, imputation and down-scaling SNP panels. Most of bioinformatic programs (except minimac3) are managed in a conda environment. Before run ImputeMe, you must install them using install_imputeme_dependencies.sh file inside install_imputeme_dependencies folder even if you have already installed the programs in your machine. The install_imputeme_dependencies.sh will install all the programs in specific conda environments created during the installation. To do that, run:

```
./install_imputeme_dependencies.sh -f path/to/install/dependencies/programs
```

The bash script uses conda functions to install the most recent versions of the programs. Unfortunately, errors sometimes occur installing the most updated version of the programs due to many reasons. If you detect some errors in the installation using the aforementioned command line, use the following command lines to remove the environments from the previous installation and to install the versions of the programs that were used originally to develop ImputeMe:

```
conda env remove --name imputeme
./install_imputeme_dependencies_versions.sh -f path/to/install/dependencies/programs
```

Only Minimac3 is not available as Conda package. It must be installed separately following the instructions on Minimac3 website. 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.

Please visit the official developer websites of each program used on ImputeMe to learn more about them: BCFtools Beagle Eagle Minimac3 Plink SHAPEIT4 HTSlib VCFtools

All these programs (except Minimac3) are installed with install_imputeme_dependencies.sh file. You can use them in any routine without ImputeMe. To access the ImputeMe conda environment, use:

```
conda activate imputeme
```

To deactivate the environment, use:

```
conda deactivate
```

Run ImputeMe

To run, use the command line:

```
./ImputeMe_run.sh -p parFile.txt
```

-p: parameter file name (can be any name)

ImputeMe parameter file

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

- **INPUT_HD**: File path for the high density panel.
- **INPUT_LD**: File path for the low density panel.
- **OUTPUT_DIR**: Directory path for the outputs of the analysis (Default is the home directory if left blank: OUTPUT_DIR=).
- **OUTPUT_NAME**: Name for analysis. Choose a single name without spaces (Default is user name if left blank: OUTPUT_NAME=).
- **PHASE_HD**: Phase HD panel? PHASE_HD=yes or PHASE_HD=no (Default yes).
- **PHASE_LD**: Phase LD panel? PHASE_LD=yes or PHASE_LD=no (Default yes).

- **IMPUTE_PROGRAM**: Choose the program for imputation analysis (IMPUTE_PROGRAM=beagle or IMPUTE_PROGRAM=minimac3).
- **MINIMAC3**: Path for minimac3 program (If you choose IMPUTE_PROGRAM=beagle, use MINIMAC3=none).
- **CHROM**: Chromosome names in this format: [chrom] or [chrom]:[start]-[end]. If more than region, separate then using ',' (e.g. CHROM=1,4,6:6000-7000,X,Y,MT).
- **CHROMX**: Number of chromosome X.
- **USE_HD_PHASE_LD**: Use HD panel to phase LD panel? USE_HD_PHASE_LD=yes if your hd panel has more samples than the low density panel. Otherwise, USE_HD_PHASE_LD=no
- **MAF**: Mimimum allele frequency for filtering (Use MAF=0 if vcf was already filtered for maf). This option only works if PHASE_HD=yes and PHASE_LD=yes.
- **CALLRATE_HD**: SNP call rate of high density panel for filtering (Use CALLRATE_HD=0 if HD panel was already filtered for call rate). This option only works if PHASE_HD=yes.
- **CALLRATE_LD**: SNP call rate of low density panel for filtering (Use CALLRATE_LD=0 if LD panel was already filtered for call rate). This option only works if PHASE_LD=yes.
- **DOWN_SCALING**: Down-scaling HD panel to LD panel level to test imputation method? DOWN_SCALING=yes or DOWN_SCALING=no
- **P**: Proportion of samples to be used in the test set. This option only works if DOWN_SCALING=yes.
- **THREADS**: Number of threads for analysis.
- **MEM**: Amount of memory for analysis.
- **BATCH**: Number of analysis at the same time, i.e. number of imputations to be done at the same time.

ImputeMe Output

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

- **hd**: it contains the phased high density panel for each chromosome.
- **ld**: it contains the phased low density panel for each chromosome.
- **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.
- **Imputed**: It contains the imputed vcf files.
- **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.

ImputeMe Examples

ImputeMe 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

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:

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

Example 2

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:

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

Example 3

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

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

Example 4

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 ImputeMe, choose **PHASE_HD=no** and **PHASE_LD=no**.