EZImputer

User Guide and Cookbook for 1000 genome imputation.

Version 1.0

Hugues Sicotte, Naresh Prodduturi 2013

© Mayo Foundation, All Rights Reserved

Table of Contents

[Installation 5](#_Toc353363694)

[Downloading 5](#_Toc353363695)

[EXTERNAL tools dependencies 5](#_Toc353363696)

[Get\_impute\_reference.pl 6](#_Toc353363697)

[INPUT FILES 7](#_Toc353363698)

[PLINK COMMANDS 7](#_Toc353363699)

[DATA QC 8](#_Toc353363700)

[Upgrade\_inputmarkers\_to\_build37\_by\_DBSNP.pl 9](#_Toc353363701)

[Filter\_inputdata\_by\_impute\_ref.pl 11](#_Toc353363702)

[QC\_fwd\_structure.pl 13](#_Toc353363703)

[Output from sample qc file 14](#_Toc353363704)

[Create\_fwdstrd\_ind\_4ambiInputMarkers.pl 16](#_Toc353363705)

[Filter\_1000genome\_reference\_by\_maf.pl 17](#_Toc353363706)

[Phase\_Impute\_by\_parallel\_proc.pl 19](#_Toc353363707)

[Optional modules 21](#_Toc353363708)

[Filter out the input data markers based on the different 1000 genome population MAFs 21](#_Toc353363709)

[Filter the imputation reference 22](#_Toc353363710)

[Appendices 23](#_Toc353363711)

[Appendix A: Specifying the files with the pre-phased data. 23](#_Toc353363712)

[Appendix B: Ambiguous Markers 24](#_Toc353363713)

[Appendix C : Small region imputation 25](#_Toc353363714)

[Appendix D: Config File Parameters 26](#_Toc353363715)

[Appendix E [Output Files] 28](#_Toc353363716)

[Appendix F: Restart workflow at different stages [Job failed?] 30](#_Toc353363717)

[Appendix G: Links (Reference Files) 31](#_Toc353363718)

[Appendix H: WORKFLOW NOT SUPPORTING JOB SCHEDULER PARAMETERS SGE or PBS (RUNNING WORKFLOW MANUALLY) 32](#_Toc353363719)

[Appendix I: Downloading External Tools (and TOOLINFO FILE) 34](#_Toc353363720)

[EXAMPLES 40](#_Toc353363721)

[WHOLE GENOME IMPUTATION EXAMPLE 40](#_Toc353363722)

[SAMPLE REGION IMPUTATION EXAMPLE 45](#_Toc353363723)

[TWO PLATFORM DATA 47](#_Toc353363724)

The EZImputer workflow is a modular set of scripts for genome-wide imputation. It automates many of the steps regularly needed for an imputation project. We have extensively tested the following workflow and even created work-around to compensate for bugs in the various external tools we use. We provide a cookbook with fully worked out examples (full scripts are provided in the downloadable package).

The modules in this workflow should be run in the following sequence. Steps with a \* are optional, based on the specific project.

1. [Get\_impute\_reference](#_Get_impute_reference.pl).pl : To download impute2 1000 genome reference files.
2. [Upgrade\_inputmarkers\_to\_build37\_by\_DBSNP](#_Upgrade_inputmarkers_to_build37_by_).pl\*: To convert input data which is on a previous build (e.g. build 36) to the current build (e.g. build 37) coordinates using DBSNP files.
3. [Filter\_inputdata\_by\_impute\_ref](#_Filter_inputdata_by_impute_ref.pl).pl\*: To filter the input data markers based on the different 1000 genome populations minor allele frequency ranges.
4. [QC\_fwd\_structure](#_QC_fwd_structure.pl).pl\*: To perform initial sample & marker level QC, perform plink sex check, perform forward strand mapping, and run the STRUCTURE program to get the predicted ethnicity (compared to Hapmap samples) for the samples.
5. [Create\_fwdstrd\_ind\_4ambiInputMarkers](#_Create_fwdstrd_ind_4ambiInputMarker).pl\*: To generate the indicator file for ambiguous markers which is used by the impute workflow to deal with ambiguous markers in the two step process.
6. [Filter\_1000genome\_reference\_by\_maf](#_Filter_1000genome_reference_by_maf.).pl\*: To filter and create a new impute2 imputation reference after filtering out some SNPs based thresholds for the minor allele frequency in one or more reference populations.
7. [Phase\_Impute\_by\_parallel\_proc](#_Phase_Impute_by_parallel_proc.pl).pl : To perform phasing and imputation.

# Installation

## Downloading

Download the EZImputer from google code <https://code.google.com/p/ezimputer/>

Decide on an installation directory for ezimputer (you will need about 80MB for the Easy imputer tools and external tools).

**export EZIMPUTER=”full\_path\_to\_installation\_directory”**

**cd ${EZIMPUTER}**

# Untar the code( the name of the tar file may change based on the version you have)

**tar –zxvf ezimputer.tar.gz**

# You can then either install the required external tools yourself (see below) or run the following two commands.

## EXTERNAL tools dependencies

A number of external tools need to be installed prior to running the workflow.

plink,impute2,shapeit,check\_strands,gprobs\_metric,structure

To let the workflow know where each tool is installed, you need to specify its location in a tool info (text file with any name) configuration file

We provide a binary whenever the license terms allows, but you may want to download the latest versions and copy them in the bin directory. The following two scripts automate the installation.

**./install\_tools.sh ${EZIMPUTER}**

# and if there are no errors..

**./make\_tool\_info.sh ${EZIMPUTER} > tool\_info.txt**

Alternatively, you can install the tools manually and create the config file as follows in Appendix I: [Downloading External Tools (TOOLINFO FILE)](#_Appendix_I:_Downloading)

## Get\_impute\_reference.pl

This script downloads the impute2 reference files, untars them, and generates the reference file as a subdirectory in the path specified. The name of the downloaded dataset is to be used as the parameter for -IMPUTEREF\_VERSION for the various modules.

USAGE : perl $EZIMPUTER/Get\_impute\_reference.pl -OUT\_REF\_DIR <TARGET IMPUTE REFERENCE DIRECTORY> -DOWNLOAD\_LINK <DOWNLOAD LINK>

-OUT\_REF\_DIR -> Path to the impute reference directory (Output directory will be created if not existing)

-DOWNLOAD\_LINK-> Download link for the impute [see below or visit impute website to get the download link]

**Links for Impute2 1000 genome Reference data**

http://mathgen.stats.ox.ac.uk/impute/data\_download\_1000G\_phase1\_integrated.html

http://mathgen.stats.ox.ac.uk/impute/ALL\_1000G\_phase1integrated\_v3\_impute.tgz

**SAMPLE EXECUTION**

|  |
| --- |
| #Example command (all on one line)  perl $EZIMPUTER/Get\_impute\_reference.pl  -OUT\_REF\_DIR /home/EzImputer\_Sample\_project/impute\_ref  -DOWNLOAD\_LINK <http://mathgen.stats.ox.ac.uk/impute/ALL_1000G_phase1integrated_v3_impute.tgz>  #Will download the reference and create one file per chromosome.  #/home/EzImputer\_Sample\_project/impute\_ref/**ALL\_1000G\_phase1integrated\_v3**\_chr\*\_impute.hap.gz  #The script will echo both complete path to reference and reference keyword  IMPUTE REF DIRECTORY : /home/EzImputer\_Sample\_project/impute\_ref  IMPUTE REF KEYWORD : ALL\_1000G\_phase1integrated\_v3  #In this case, the “VERSION” parameter for many scripts will be **ALL\_1000G\_phase1integrated** |

# INPUT FILES

The input data for the scripts should be in plink transposed fileset format (The following description of this format is taken from the plink website: <http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml> )

A transposed fileset contains two text files:

one (TPED) containing SNP and genotype information where one row is a SNP;

one (TFAM) containing individual and family information, where one row is an individual.

The first 4 columns of a TPED file are the same as a standard 4-column MAP file. All genotypes are listed for all individuals for each particular SNP on each line. The TFAM file is just the first six columns of a standard plink PED file.

Example

*<------------- trans.tped ------------->* *<- trans.tfam ->*

1 snp1 0 5000650 A A A C C C A C C C C C 1 1 0 0 1 1

1 snp2 0 5000830 G T G T G G T T G T T T 2 1 0 0 1 1

3 1 0 0 1 1

4 1 0 0 1 2

5 1 0 0 1 2

6 1 0 0 1 2

This kind of format can be convenient to work with when there are very many more SNPs than individuals (i.e. WGAS data).

To read a transposed fileset (mydata.tped and mydata.tfam), using plink

plink --tfile mydata

http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#tr

# PLINK COMMANDS

1. Convert BED(Binary plink files) to TPED (Transpose plink files)

plink –bfile <INPUT BINARY FILE> --recode --transpose –out <OUTPUT TRANSPOSE FILE>

1. Convert PED/MAP(Ped plink files) to TPED/TFAM (Transpose plink files)

plink –file <INPUT PED FILE> --recode --transpose –out <OUTPUT TRANSPOSE FILE>

# DATA QC

The following QC should be performed by the user prior to running our tools.

1. Please provide unique samples ids in the tfam files (column 2).
2. Convert your platform specific marker Id’s to rsid’s based on platform annotation file.
3. Set mother and father id to missing (plink dataset: 3rd and 4th column in the tfam/fam file, i.e. the columns should be 0 in plink).
4. No duplicate rsid ‘s in the input dataset.
5. Make sure your input dataset doesn’t have duplicate positions for the same chromosome.
6. TPED file should be sorted according to chromosome and position.

sort -k1,1n -k4,4n sample.tped

# Upgrade\_inputmarkers\_to\_build37\_by\_DBSNP.pl

This script is used to convert input data which is on a previous human genome coordinate build (e.g. build 36) to the current build coordinates (e.g. build 37) using DBSNP files. This step can be skipped if the data is already mapped to the same build as the 1000 genomes reference files (currently build 37)

USAGE : perl $EZIMPUTER/Upgrade\_inputmarkers\_to\_build37\_by\_DBSNP.pl -DBSNP\_DIR <DBSNP DIR>

-DBSNP\_DOWNLOADLINK <DBSNP VERSION>

-INPUT\_FILE <INPUT TPED FILE>

- REMAPPED\_CURRENT\_BUILD <Filename with full path>

- NOTMAPPED\_OLD\_BUILD <Filename with full path>

**INPUT ARGUMENTS**

“-DBSNP\_DIR”->Target DBSNP directory for downloading DBSNP file (Output directory will be created if not existing)

“-DBSNP\_DOWNLOADLINK”->DBSNP download link

**DBSNP download links**

#for DBSNP version 137 ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human\_9606/database/organism\_data/b137\_SNPChrPosOnRef.bcp.gz

#for DBSNP version 135 ftp://ftp.ncbi.nih.gov/snp/organisms/human\_9606/database/b135\_archive/organism\_data/b135\_SNPChrPosOnRef\_37\_3.bcp.gz

-INPUT\_FILE -> Path to input TPED file

- REMAPPED\_CURRENT\_BUILD ->Full path and file name to current build output TPED file for successfully remapped SNPs.

- NOTMAPPED\_OLD\_BUILD -> Full path and file name to old build output TPED file for unsuccessfully remapped SNPs.

**SAMPLE EXECUTION**

|  |
| --- |
| #sample command, all one line  perl $EZIMPUTER/Upgrade\_inputmarkers\_to\_build37\_by\_DBSNP.pl -DBSNP\_DIR /home/EzImputer\_Sample\_project/DBDIR/ -DBSNP\_DOWNLOADLINK ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human\_9606/database/organism\_data/b137\_SNPChrPosOnRef.bcp.gz  -INPUT\_FILE /home/EzImputer\_Sample\_project/hapmap3\_r3\_b36\_fwd.consensus.qc.poly.tped  -REMAPPED\_CURRENT\_BUILD /home/EzImputer\_Sample\_project/hapmap3\_r3\_b36\_fwd.consensus.qc.poly\_build37.tped  - NOTMAPPED\_OLD\_BUILD /home/EzImputer\_Sample\_project/hapmap3\_r3\_b36\_fwd.consensus.qc.poly\_build36.tped |

# Filter\_inputdata\_by\_impute\_ref.pl

This script filters the input data markers based on the different 1000 genome population or subpopulations minor allele frequency ranges. This script also automatically filters out markers not on the reference since they are not needed for imputation.

Filtering extremely rare SNPs can speed up imputation. Additionally, it has been shown by Hancock et al (2012) [see article for citation] that for samples of multi-ethnicity origin, it is best to eliminate SNPs monomorphic in subpopulations relevant to those samples. For example, for imputing genotypes for African-American samples, exclude SNPs monomorphic in European or African populations, e.g. –POP AFR,CEU –LMAF 0.0001,0.0001 . We use 0.0001 instead of 0, since the –LMAF limit is inclusive.

USAGE : perl $EZIMPUTER/Filter\_inputdata\_by\_impute\_ref.pl

-POP <Different population : afr,amr,asn,eur>

-LMAF <Lower Limit Minor allele frequency, inclusive>

-UMAF <Upper Limit Minor allele frequency, inclusive>

-REF\_GENOME\_DIR <Reference Genome Directory>

-IMPUTEREF\_VERSION <IMPUTE REFERENCE directory name (e.g. ALL\_1000G\_phase1integrated\_v3 )>

-INPUT\_FILE <INPUT TPED FILE>

-OUTPUTFILE <OUTPUT TPED FILE>

**INPUT ARGUMENTS**

-POP -> The user can specify populations or sub populations. Multiple populations can be separated by ‘,’. Available populations are AFR, AMR, ASN, EUR and available sub populations are ASW,CEU,CHB,CHS,CLM,FIN,GBR,IBS,JPT,LWK,MXL, PUR,TSI,YRI .

-LMAF -> Lower limit of minor allele frequency for each population

-UMAF -> Upper limit of minor allele frequency for each population

-REF\_GENOME\_DIR-> Path to the IMPUTE2 Reference genome directory

-IMPUTEREF\_VERSION -> ALL\_1000G\_phase1integrated\_v3

[Note: This field is the file prefix of the impute reference files. Look at the files in the impute reference file folder and enter the root prefix (boldface part) of the filename **ALL\_1000G\_phase1integrated\_v3**\_chr9\_impute.hap.gz]

-INPUT\_FILE -> Full path and filename to the INPUT TPED FILE

-OUTPUTFILE -> Full path and filename to the OUTPUT TPED FILE

**SAMPLE EXECUTION**

|  |
| --- |
| #sample command, all one line  perl $EZIMPUTER/Filter\_inputdata\_by\_impute\_ref.pl  -POPULATION afr,eur  -LMAF 0.01,0.01  -UMAF 0.5,0.5  -REF\_GENOME\_DIR /home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute  -IMPUTEREF\_VERSION ALL\_1000G\_phase1integrated\_v3  -INPUT\_FILE /home/EzImputer\_Sample\_project/fwdStrandResults\_input.tped  -OUTPUT\_FILE /home/EzImputer\_Sample\_project/filter\_fwdStrandResults\_input.tped |

# QC\_fwd\_structure.pl

This script does following

1. Initial sample & marker level QC : Filter out samples and markers based on the sample level missing threshold and marker level missing threshold specified in the config file.
2. Plink sex check : Gender information is imputed using plink. If the gender is included in the input FAM file, this script also reports if the input gender information is not concordant with imputed one.
3. Reference strand mapping of input data: Maps the genotypes and SNP in the input data to the same strand as the reference panel in two steps:
4. The alleles and genotypes for non-ambiguous SNP will be made to match the alleles of the inpute2 reference strand, reverse-complementing the alleles and genotypes if needed.
5. Ambiguous SNPs (A/T or C/G SNPs) will be flipped to match the impute2 reference strand using beagle utility tools. In brief, after converting the Impute2 reference files to beagle format, the beagle utility tool rapidly imputes the genotype at the ambigous sites using nearby markers. By comparing these imputed genotypes to the input genotypes (and the reverse complement of the input genotypes), the programs is able to determine the relative strand of the input genotype needed to match the reference alleles.
6. If the input marker alleles don’t match reference alleles (tri allelic with reference) then the marker is ignored (not included in the output).
7. Ethnicity QC: Runs the structure program to get the population structure for the samples.

USAGE: perl $EZIMPUTER/QC\_fwd\_structure.pl -run\_config <PATH TO THE RUN CONFIG FILE> -tool\_config <PATH TO THE TOOLINFO CONFIG FILE>

1. Create Run config file (see below)
2. You need a tool info file and helper tools configured (Appendix I)

**SAMPLE EXECUTION**

|  |
| --- |
| perl $EZIMPUTER/QC\_fwd\_structure.pl -run\_config $PWD/SAMPLEFILES/INPUTDATA/config\_QC\_FWD -tool\_config $EZIMPUTER/tool\_info.config |

**RUN Config file parameters (You should specify all options. [Do not include the comments within square bracket]) There should be no extra spaces in the config file.**

TPED=/home/EzImputer\_Sample\_project/hapmap3\_r3\_b36\_fwd.consensus.qc.poly\_build37.tped [PLINK F ORMAT TPED FILE FULL PATH]

TFAM=/home/EzImputer\_Sample\_project/hapmap3\_r3\_b36\_fwd.consensus.qc.poly.tfam [PLINK F ORMAT TFAM FILE FULL PATH]

TEMP\_FOLDER=/home/EzImputer\_Sample\_project/temp [The directory path where you want to create the temp files processing]

INNER\_DIR=temp179807890 [enter unique sub directory name you want to create in  ‘TEMP\_FOLDER’ :all the temp processing will be done in this directory, if you don’t want enter one the program will create one for you]

IMPUTE\_REF=/home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute [Impute reference path+”version” prefix]

OUTPUT\_FOLDER=/home/EzImputer\_Sample\_project [Forward Stranded plink files, Indicator files,markers.ignored and QC table will be send to OUTPUT\_FOLDER]

IMPUTEREF\_VERSION=ALL\_1000G\_phase1integrated\_v3 [Note: This field is the file prefix of the impute reference files. Look at the files in the impute reference file folder and enter the root prefix of the filename (bold in) **ALL\_1000G\_phase1integrated\_v3**\_chr9\_impute.hap.gz]

GENOTYPE\_PERCENT\_CUTOFF=0.05 [Marker level missing Threshold Percentage (0-1 values) : any Marker which have more missing value percentage will be dropped]

SAMPLE\_PERCENT\_CUTOFF=0.05 [Sample level missing Threshold Percentage (0-1 values) : any Sample which have more missing value percentage will be dropped]

BEAGLE\_REF\_DB=/home/EzImputer\_Sample\_project/BEAGLE/ [Directory that will contain imputation reference reformatted in beagle format. For mapping ambiguous markers to impute reference strand, beagle utility program (check\_strands\_16May11) is used and for this impute reference should be converted to beagle reference format (these files are generated when your running for the first time and for later runs these files are reused) ]

**OUTPUTFILES**

You can see output files in the output directory mentioned in the config parameter “OUTPUT\_FOLDER”.

**OUTPUT FILE DESCRIPTIONS**

1. sample.qc : Sample QC file
2. fwdStrandResults\_input.tped: Input plink tped file in the forward strand.
3. fwdStrandResults\_input.tfam: Input plink tfam sample description file.
4. fwdStrandResults\_input.ind: Indicator file for each marker in the tped file.
5. markers.ignored: Markers ignored in the process and not present in the fwdStrandResults\_input.tped.

### Output from sample qc file

**QC FILE COLUMN DESCRIPTION (per sample)**

* **MISSING\_GENO\_CLEAN** – Number of missing Genotypes after genotype level QC
* **NUM\_GENO\_CLEAN** - Total Number of Genotypes after genotype level QC
* **QC\_SAMPLE\_FLAG** – FLAG
  + ***1:*** Sample passed the QC and included
  + ***0****:* Sample failed the QC and removed
* **QC\_SAMPLE\_IND** – Warning FLAG [Sample Indicated as bad quality sample but not removed]
  + ***1:*** Sample passed the QC
  + ***0:*** Sample failed the QC
* **QC\_SEX\_IND** - FLAG for sex check QC
  + ***0:*** Plink not able to determine sex [chr 23 not available]
  + ***1:*** Provided sex information concordant with plink imputed one
  + ***2:*** Provided missing sex information and plink imputed the sex
  + ***3:*** Provided sex information is discordant with plink imputed one
  + ***NA:*** Samples are removed by QC
* **PLINK\_SEX\_IMP\_ESTIMATE** - The actual X chromosome inbreeding (homozygosity) estimate Plink\_Sex\_Imp\_Estimate
  + Usually float values or integers
  + 'nan' = if plink not able to determine sex[chr 23 not available]
* **EUR :** Percent sample to European population **(FROM STRUCTURE OUTPUT)**
* **ASN :** Percent sample to Asian population
* **YRI :** Percent sample to Yoruba population

**FORWARD STRAND INDICATOR FILE GENERATED BY THE FORWARD STRAND PROGRAM**

1. Marker ID
2. INDICATOR VALUES
3. ‘0’: Marker is not flipped as input alleles are on same strand as reference.
4. ‘1’: Marker is flipped to match input alleles to same strand reference.
5. ‘2’: Marker doesn’t exist on the reference.

# Create\_fwdstrd\_ind\_4ambiInputMarkers.pl

The QC module tries to correct the strand of ambiguous SNPs. For users who prefer to remove from the input data set any ambiguous SNPs, this script offers an alternative. It generates an indicator file that tells the imputation workflow to not use any ambiguous SNPs in the input data set. This indicator file is used by the imputation workflow to deal with ambiguous markers in the two step process (refer to [Phase\_Impute\_by\_parallel\_proc](#_Phase_Impute_by_parallel_proc.pl)  manual). Note that there are historically very few ambiguous SNPs on Illumina genome-wide platforms (on the order of 0-3%), but this number can be much higher on Affymetrix Genotyping arrays (~ 10-15%).

SAMPLE EXECUTION

|  |
| --- |
| #sample command, all one line  perl $EZIMPUTER/Create\_fwdstrd\_ind\_4ambiInputMarkers.pl -INPUT\_FILE <INPUT TPED FILE> -OUTPUTFILE <OUTPUT FORWARD STRAND INDICATOR FILE> |

# Filter\_1000genome\_reference\_by\_maf.pl

This script will create a new copy of the imputation reference after removing SNPs that fail to be within the specified limits of population specific Minor Allele Frequency. [An alternative, which is not supported by EZImputer, is to use the new reference is to extract the list of SNPs passing QC and use the ‘include’ option in the original impute2 software (this option can restrict the markers to be included in the imputation process), but this option will increase the running time when compared to using all the markers in the reference.]

USAGE : perl $EZIMPUTER/Filter\_1000genome\_reference\_by\_maf.pl

-POPULATION <Different population : AFR,AMR,ASN,EUR>

-LMAF <Lower Limit Minor allele frequency>

-UMAF <Upper Limit Minor allele frequency>

-REF\_GENOME\_DIR <Reference Genome Directory>

-OUTPUT\_DIR <OUTPUT DIR>

-IMPUTEREF\_VERSION <version of the impute reference files e.g. ALL\_1000G\_phase1integrated\_v3>

**INPUT ARGUMENTS**

-POPULATION -> You can specify populations or sub populations. Multiple populations can be separated by ‘,’. Available populations are AFR, AMR, ASN, EUR and available sub populations are ASW,CEU,CHB,CHS,CLM,FIN,GBR,IBS,JPT,LWK,MXL, PUR,TSI,YRI . When downloading the 1000 genome data, a file (e.g. ALL\_1000G\_phase1integrated\_feb2012.sample) is downloaded containing the population definitions that can be used by this script. This file will be downloaded in the directory specified by the OUT\_REF\_DIR argument of Get\_impute\_reference.pl

-LMAF -> Lower limit of minor allele frequency for each population

-UMAF -> Upper limit of minor allele frequency for each population

-REF\_GENOME\_DIR-> Path to the IMPUTE2 Reference genome directory

-IMPUTEREF\_VERSION ->ALL\_1000G\_phase1integrated\_v3

[Note: This field is the file prefix of the impute reference files. Look at the files in the impute reference file folder and enter the root prefix of the filename (bold in ) **ALL\_1000G\_phase1integrated\_v3**\_chr9\_impute.hap.gz]

-INPUT\_FILE -> PATH to the INPUT TPED FILE

-OUTPUT\_DIR -> PATH to the NEW REFERENCE FILES

**SAMPLE EXECUTION**

|  |
| --- |
| #sample command, all one line  perl $EZIMPUTER/Filter\_1000genome\_reference\_by\_maf.pl  -POPULATION eur  -LMAF 0.005  -UMAF 0.5  -REF\_GENOME\_DIR /home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute  -OUTPUT\_DIR /home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute\_filtered\_eur  -IMPUTEREF\_VERSION ALL\_1000G\_phase1integrated\_v3  #Note that after running this script, you have to update the config file with the new reference directory (IMPUTE\_REF) but not change the IMPUTEREF\_VERSION  IMPUTE\_REF=/home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute\_eur  IMPUTEREF\_VERSION= ALL\_1000G\_phase1integrated\_v3 |

# 

# Phase\_Impute\_by\_parallel\_proc.pl

This is the module that performs the imputation two steps

1. STEP 1 : Phasing
   1. Tool Used: SHAPEIT program
2. STEP 2 : Imputation
   1. Tool Used: IMPUTE2 program

You can perform

1. Phasing only
2. Imputation only, using data pre-phased from another phasing run. You would do that if you downloaded a more recent reference or were trying different filtering settings for the reference.
3. Phasing and Imputation

NOTE:

1. The workflow assumes that you are starting with tfam/tped files. These can be exported directly from Plink.
2. Please provide unique samples ids in the tfam files (column 2).
3. Convert your platform specific marker Id’s to rsid’s based on platform annotation file.
4. If your input data is build36 , upgrade it to build 37 because the workflow currently only supports build 37 reference samples.
5. Set mother and father id to missing (plink dataset: 3rd and 4th column in the tfam/fam file, i.e. the columns should be 0 in plink).
6. No duplicate rsid ‘s in the input dataset.
7. While the imputation imputes insertion/deletion genotypes, the workflow cannot accept insertion/deletions as input genotype data.
8. Make sure your input dataset doesn’t have duplicate positions for the same chromosome.
9. For imputing a single chromosome or small region, truncate the plink input data files to the desired region.
10. The imputation process may take a few days to complete depending on the request, so it is best practice to submit the job to a compute cluster (cluster managed with SGE or PBS are supported).
11. There should be no extra spaces in the config files.

STEPS

(After the installation of the workflow and the external tools)

1. First step is to download the impute reference ([here](#_Get_impute_reference.pl)).
2. Create config file (see section on Config file at the end)
   1. For Phasing only mode
      1. set config parameter SHAPEITONLY=YES
   2. For Imputation only (If data is already phased),
      1. Set config parameter SHAPEITONLY=NO
      2. Set config parameter HAPS=<PATH to the tar file> (see below)
   3. For Phasing and Imputation
      1. Set config parameter SHAPEITONLY=NO
      2. Set config parameter HAPS=NA (see below)
3. [Downloading External tools & preparing tool info file](#TOOLINFO)
4. Invoke the script

**SAMPLE EXECUTION**

|  |
| --- |
| *#perl perl\_workflow\_impute.pl -run\_config <PATH TO THE RUN CONFIG FILE> -tool\_config <PATH TO THE TOOL CONFIG FILE>*  *#sample command, all one line*  *perl $EZIMPUTER/Phase\_Impute\_by\_parallel\_proc.pl –*run\_config $PWD/SAMPLEFILES/INPUTDATA/ config\_phase\_impute -tool\_config $PWD/SAMPLEFILES/INPUTDATA/tool\_info.config |

# Optional modules

## Filter out the input data markers based on the different 1000 genome population MAFs

This script filters the input data markers based on the different 1000 genome population minor allele frequency ranges and also automatically filters out markers not on the reference.

For example if you want to filter the input data based on the 1000 genome African and European population minor allele frequencies. You can run the program as explained below

perl $EZIMPUTER/Filter\_inputdata\_by\_impute\_ref.pl

-POPULATION afr,eur

-LMAF 0.01,0.01

-UMAF 0.5,0.5

-REF\_GENOME\_DIR /home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute

-IMPUTEREF\_VERSION ALL\_1000G\_phase1integrated\_v3

-INPUT\_FILE /home/EzImputer\_Sample\_project/fwdStrandResults\_input.tped

-OUTPUT\_FILE /home/EzImputer\_Sample\_project/filter\_fwdStrandResults\_input.tped

The output file generated by this program is located in the above ‘OUTPUT\_FILE’ parameter.

## Filter the imputation reference

This script will filter and recreate the new impute reference for imputation purpose based on population specific Minor Allele Frequency ranges.

For example if you want to filter the reference based on the minor allele frequency of European population.

perl $EZIMPUTER/Filter\_1000genome\_reference\_by\_maf.pl

-POPULATION eur

-LMAF 0.005

-UMAF 0.5

-REF\_GENOME\_DIR /home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute

-OUTPUT\_DIR /home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute\_filtered\_eur

-IMPUTEREF\_VERSION ALL\_1000G\_phase1integrated\_v3

The new reference will be created in the location specified in the PARAMETER ‘OUTPUT\_DIR’

# Appendices

## Appendix A: Specifying the files with the pre-phased data.

**PATH to the tar file:**

When data is already phased, the user needs to provide the previously phased data in a tar file. Here is how the workflow expects the tar-file of existing haplotype. This directory structure will be created by the workflow, it is only documented for the user who has already run the workflow once, and wants to use pre-phased data with a new release of the imputation reference.

Create a directory and subdirectories with names chromosome names from 1-23 (if chr X exist in your dataset, it is named 23)

For chromosomes 1-22, files needed

1. Haps file (Output from shapeit program)
2. Sample file (Output from shapeit program)

For chromosome 23, files needed

1. Haps file (Output from shapeit program)
2. TFAM file with SEX code non-missing (column 5)
3. Sample file (Output from shapeit program)

Tar & gzip compress directory : tar -zcf <PATH TO TAR.GZ FILE > -C <TARGET DIRECTORY> .

Ex: tar -zcf <OUTPUT DIRECTORY>/shapeit\_jobs.tar.gz -C <INPUT FOLDER>/SHAPEIT/ .

#To uncompress phased output tar file(example) :

# To uncompress tar –C <PATH TO THE TARGET DIRECTORY > -xzf <PATH TO THE TAR.GZ FILE>

Ex: tar -C <PATH TO THE TARGET DIRECTORY > / -xzf <PATH TO THE TAR.GZ FILE>/shapeit\_jobs.tar.gz

## Appendix B: Ambiguous Markers

**FORWARD STRAND INDICATOR FILE**

The imputation workflow will automatically flip alleles to match the alleles of the reference genotypes. However, for A/T and C/G SNPs, this is not possible (ambiguous SNPs). In this case, the user can provide a “forward strand indicator file” to the workflow to indicate how the alleles in the data should be flipped (for any SNP).

1. Tab delimited file
2. First column should be the marker id
3. Second column should be indicator values (‘0’ or ‘1’)
   1. values>0->Data for this marker is on the forward strand,
   2. values > 1 -> Data for this marker is not on the forward strand.
4. Ambiguous SNPs that do not have indicator values ‘0’ or ‘1’ are removed at the beginning of the imputation process , and re-imputed genome-wide. After imputation, the normalization process will compare the imputed genotypes with two alternative genotypes: original orientation and reverse-complemented orientation. It will replace the imputed values with original genotypes that match best the imputed genotypes (and keep missing values).

p.s. The program “QC\_fwd\_structure.pl” can pre-correct SNPs to the forward strand (this program will flip the markers to forward strand according to 1000 genome markers and platform specific probes) . This program generates the forward strand indicator file.

**DEALING WITH AMBIGUOUS MARKERS**

If you’re not sure about the strand information for ambiguous markers then create a FORWARD STRAND INDICATOR file for markers and give indicator value greater than 1. This is done with the script

## Appendix C : Small region imputation

**SINGLE CHROMSOME OR SMALL REGION IMPUTATION**

First, subset the plink input data files to the region you want to impute.

**SMALL REGION IMPUTATION**

For small region imputations, supply a truncated input dataset to the imputation workflow. The parameters you need to set in the config file are [do not enter the comments in brackets]

* CHR\_START\_INPUT=YES [Set this option to ‘YES’ if you want impute to small regions instead of whole chromosomes. So the imputation start position and stop position will be based on input data but not on the reference start and stop]
* SMALL\_REGION\_EXTN\_START=2000000[If you want to extend the imputation region on start end for small region imputation, you can specify here in unit bases. If you do not want any extension just set this option to ‘0’)]
* SMALL\_REGION\_EXTN\_STOP=3000000[If you want to extend the imputation region on stop end for small region imputation, you can specify here in unit bases. If you do not want any extension just set this option to ‘0’]

Note: The region will be extended by the buffer region as needed, but no result will be generated for the buffer region.

## Appendix D: Config File Parameters

There are two configuration files. The first one is used by the main imputation workflow, while the second one(tool info) is used by many scripts to know the location of tools.

**Config file parameters for imputation (You should specify all options, but omit the comments within the square brackets)**

* TPED=/home/EzImputer\_Sample\_project/fwdStrandResults\_input.tped.gz [gzipped TPED FILE FULL PATH [Strictly gzip file is required]]
* TFAM=/home/EzImputer\_Sample\_project/fwdStrandResults\_input.tfam [TFAM FILE FULL PATH]
* FORWARDSTRAND\_IND=/home/EzImputer\_Sample\_project /fwdStrandResults\_input.ind [fwd strand indicator file if you have one or place ‘NA’  [no gzip files allowed] ] [If you want to create a indicator file specially to deal with ambiguous snps you can use this script [Create\_fwdstrd\_ind\_4ambiInputMarkers](#CFA)\*).
* TEMP\_FOLDER=/home/EzImputer\_Sample\_project/temp [The directory path where you want to create the temp files processing and output files]
* IMPUTE\_REF=/home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute[Impute reference path you want to use]
* IMPUTE\_WINDOW=5000000 [Impute window size you want to chop, should be between 2mb and 5mb]
* IMPUTE\_EDGE=125 [buffer region for the windows ,must be in KB ]
* HAPS=/data4/bsi/RandD/Workflow/temp/shapeit\_jobs.tar.gz [If you have pre-phased already done then give path to tar file or enter ‘NA’]
* [EMAIL=lastname.firstname@mayo.edu](mailto:EMAIL=lastname.firstname@mayo.edu) [Fill in user’s full e-mail address]
* SGE\_SHAPEIT\_MEM=15G [Expected memory taken by your shape it jobs ] [even if your shapeit jobs are done just give some numbers][optional parameter but should present when PBS=NO ]
* SGE\_SHAPEIT\_QUEUE=1-day [Queue you want submit for shapeit jobs ] [even if your shapeit jobs are done just give some numbers] [optional parameter but should present when PBS=NO ]
* SGE\_IMPUTE\_MEM=15G [Expected memory taken by your impute jobs ] [optional parameter but should present when PBS=NO ]
* SGE\_IMPUTE\_QUEUE=1-day  [Queue you want submit for impute jobs ] [optional parameter but should present when PBS=NO ]
* IMPUTEREF\_VERSION= ALL\_1000G\_phase1integrated\_v3

[Note: This field is the file prefix of the impute reference files. Look at the files in the impute reference file folder and enter the root prefix of the filename (bold in ) **ALL\_1000G\_phase1integrated\_v3**\_chr9\_impute.hap.gz]

* LOCALTEMP\_SHAPEIT=4G [Local temp space required by the shapeit sungrid engine jobs to compress result files]
* LOCALTEMP\_IMPUTE=1G [Local temp space required by the impute sungrid engine jobs to compress result files]
* LOCALTEMP= /home/EzImputer\_Sample\_project/temp/LOCALTEMP [Path to local temp space required by the impute sun grid engine jobs to compress result files. If your SGE doesn’t have enough space on the computational nodes you can point to the TEMP\_FOLDER ]
* INNER\_DIR=temp179807890 [enter unique sub directory name you want to create in  ‘TEMP\_FOLDER’ :all the temp processing will be done in this directory, if you don’t want enter one the program will create one for you]
* RESTART=NO [(1) ‘NO’ for initial run<no restarat>. (2) ‘SHAPEIT’ to restart the workflow if some phasing (shapeit) jobs fail at phasing state. (3) ‘IMPUTE’ to restart the workflow if some impute jobs fail at imputation state. (4) ‘POST’ to restart the workflow if some post processing jobs fail at post processing state(generation of r2 files, indicator files). Please remember and enter the same ‘INNER\_DIR” name as of the initial run for options ‘SHAPEIT’, ‘IMPUTE’ and for ‘POST’.]
* USERNAME=mXXXXX [To create unique directories in the local temp directory to identify the user]
* SHAPEITONLY=YES(IF you want just the prephasing set this option to ‘YES” or else use ‘NO’)
* PBS=NO (This should be set to YES if your running on PBS system instead of SUNGRID engine (For BORA set this option to YES if you are running jobs on iforge))
* PBS\_PARAM="-l walltime=10:05:00,-l nodes=1:ppn=1,-A normal,-A bf0"[Should be in double quotes “” and should include only ‘-l’ (wall clock times and number of nodes, make sure you select optimum time for each shape it and impute job each. Some shapeit<Phasing> jobs may take few days to complete) and ‘–A’ (account string and queue type) remaining parameters are filled by the workflow] [optional parameter but should present when PBS=YES ]
* CHR\_START\_INPUT=YES (Set this option to ‘YES’ if you want impute to small regions instead of whole chromosomes. So the imputation start position and stop position will be based on input data but not on the reference start and stop)
* SMALL\_REGION\_EXTN\_START=2000000 (If you want to extend the imputation on the start of the small region. Please specify bases in unit Bases) [optional parameter but should present when CHR\_START\_INPUT=YES] [optional parameter but should present when CHR\_START\_INPUT=YES]
* SMALL\_REGION\_EXTN\_STOP=2500000 (If you want to extend the imputation on the end of the small region. Please specify bases in unit Bases) [optional parameter but should present when CHR\_START\_INPUT=YES] [optional parameter but should present when CHR\_START\_INPUT=YES]
* WINDOW\_CUTOFF\_NUM\_MARKERS=200 (Minimum number of input data and reference markers to be present on a window for imputation. If the number of markers are less than cutoff then the window will be merge to next one to meet the cutoff).
* EDGE\_CUTOFF\_NUM\_MARKERS =50 (Minimum number of input data markers to be present on a buffer region for imputing a window. If the numbers of markers are less than cutoff then the edges will be extended up to 2 MB to meet the cutoff).
* SHAPEIT\_STATEPHASES=100(Default value suggested by the SHAPEIT manual is 100. To specify the number of conditioning haplotypes to be used during the phasing process. The running time of the algorithm increases only linearly with this number.)
* CHECK\_POINT\_STOP=SHAPEIT (If this workflow is not compatible with your job scheduler like SGE/PBS then you can run the workflow manually. use this option ([here](#_Appendix_H:_WORKFLOW)).)(optional parameter)

## Appendix E [Output Files]

**RESULTS**

The TEMPFOLDER/ INNER\_DIR (You specified this in the config file) will be your final output folder. All the below files are located in the individual chromosome directories (restricted to chromosomes present in the input data).For whole genome imputation you can see folders from 1-23 (From chromosome 1 to chromosome 23) and for small region imputation you can only see the chromosomes present in the input data.

1. **final.dosage.gz (Two probability file)** : Individual window two probability output files will be merged at the chromosome level. (One file per chromosome) .
2. **Combined\_impute\_results\_3\_prob\_ambi\_out.gz :** Three probability file.
3. **beagle\_r2.gz :** Output from beagle utility (gprobsmetrics.jar for calculating various R2 values) . See below section for header description.
4. **final.map.gz :** Plink map file is generated for all the markers. (COLUMN DESCRIPTION: chromosome, rsid, distance, position).
5. **Combined\_impute\_results\_3\_prob\_indels\_out.gz :** Imputed indels are separated.
6. **Combined\_impute\_results\_3\_prob\_ind.gz** : Indicator file(matrix) to differentiate observed and imputed marker(rows) and sample(columns).(Column 1 : RSID, Column2 : Dosage R2 value, From column 3 : sample 1, Column 4 : sample2 ……Rows are markers and Columns are samples. Sample order matches tfam file. Indicator “1”is imputed and Indicator “0” is observed).
7. Phasing output “.haps” files from all the chromosomes will be combined and tar compressed “shapeit\_jobs.tar.gz” and can be found in the SHAPEIT\_OUTPUT directory.

Copy the necessary output file and delete the temp folder. The only reason we retained all the temp files even after the run is successful to trace the unexpected errors easily.

**HEADERS FOR BEAGLE R2 FILES**

1) marker identifier

2) minor allele

3) minor allele frequency

4) allelic r-squared

5) dosage r-squared

6) HWE dosage r-squared

7) accuracy

8) missing score

Notes:

a) Allelic r-squared is the estimated squared correlation between the most likely allele dosage (rounded to the nearest integer) and the true allele dosage.

b) Dosage r-squared is the estimated squared correlation between the estimated allele dosage (0\*P(AA) + 1\*P(AB) + 2\*P(BB)) and the true allele dosage.

c) HWE dosage r-squared, is the estimated squared correlation between the estimated allele dosage and the true allele dosage when the variance of the true allele dosage is calculated from the estimated allele frequency.

d) Accuracy is the estimated genotype accuracy based on the posterior probability of the most likely genotype.

e) Missing score is the infinum of the C > 0 such that the proportion of genotype with probability < C is greater than or equal to (1.0 - C).

f) Monomorphic markers are assigned r-squared values of 0, and accuracy and missing score values of 1.

g) The squared correlation metrics can be derived using arguments found in Appendix 1 of "Browning BL and Browning SR, Am J Hum Genet 2009;84(2):210-23".

h) For comparison purposes, we report all r-squared metrics. The allelic r-squared metric is the default reported by BEAGLE and the dosage r-squared metric is the default reported by MACH.

## Appendix F: Restart workflow at different stages [Job failed?]

If you want to know at what stage the work flow got failed just check for the log file produced by the main job and log file directory present in the temp directory.

‘shapeit\_logfiles\_sungrid’ -> Workflow got failed at phasing stage.

‘impute\_logfiles\_sungrid’ -> Workflow got failed at imputation stage.

‘post\_logfiles\_sungrid’ -> Workflow got failed at postprocessing stage.

**Restart Workflow at Phasing Phase [SHAPEIT]**

If one of your phasing jobs got failed, first check for the error log file in the ‘shapeit\_logfiles\_sungrid’ directory check for the exact reason, correct the input files if required and rerun the workflow by changing the parameter “RESTART=SHAPEIT”

**Restart Workflow at Imputation Phase**

If one of your imputation jobs got failed, first check for the error log file in the ‘impute\_logfiles\_sungrid’ directory check for the exact reason, correct the input files if required and rerun the workflow by changing the parameter “RESTART=IMPUTE”

**Restart Workflow at Post Processing Phase**

If one of your post processing jobs got failed, first check for the error log file in the ‘post\_logfiles\_sungrid’ directory check for the exact reason, correct the input files if required and rerun the workflow by changing the parameter “RESTART=POST”

## Appendix G: Links (Reference Files)

**1000 GENOME REFERENCE**

1000 genome imputation reference is available at

<http://mathgen.stats.ox.ac.uk/impute/data_download_1000G_phase1_integrated.html>

**DBSNP download links**

#for DBSNP version 137 ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human\_9606/database/organism\_data/b137\_SNPChrPosOnRef.bcp.gz

#for DBSNP version 135 ftp://ftp.ncbi.nih.gov/snp/organisms/human\_9606/database/b135\_archive/organism\_data/b135\_SNPChrPosOnRef\_37\_3.bcp.gz

**Impute2 Reference links**

http://mathgen.stats.ox.ac.uk/impute/data\_download\_1000G\_phase1\_integrated.html

<http://mathgen.stats.ox.ac.uk/impute/ALL_1000G_phase1integrated_v3_impute.tgz>

## Appendix H: WORKFLOW NOT SUPPORTING JOB SCHEDULER PARAMETERS SGE or PBS (RUNNING WORKFLOW MANUALLY)

**RUNNING WORKFLOW MANUALLY**

If your job scheduler parameters are different from config parameters and so you cannot run the workflow automatically. Then you can still use the imputation work flow manually as follows

1. **Phasing**
2. Create the config files ad mentioned in the manual.
3. select the config file parameter “CHECK\_POINT\_STOP=SHAPEIT” & “RESTART=NO”.
4. Run the workflow and the array job script “ArrayJob\_shapeit.csh” in the temp folder will be generated.
5. Edit the SGE or PBS parameters accordingly in this script “ArrayJob\_shapeit.csh” and submit the script to job scheduler.
6. Check for the files “.haps” and “.sample” in each chromosome directory and check the log files in the directory “shapeit\_logfiles\_sungrid” to make sure all jobs got successful without any error.
7. Remove the config parameter “CHECK\_POINT\_STOP=SHAPEIT” and select option “RESTART=SHAPEIT” and SHAPEITONLY=YES .
8. Resubmit the workflow and you can see “shapeit\_jobs.tar.gz” output file in the directory “SHAPEIT\_OUTPUT” in the temp directory.
9. **Imputation**
10. Copy the above shapeit tar file “shapeit\_jobs.tar.gz” to some direcoty.
11. Create the config files ad mentioned in the manual.
12. select the config file parameter “HAPS=<PATH TO FILE/ shapeit\_jobs.tar.gz”,“CHECK\_POINT\_STOP=IMPUTE” & “RESTART=NO”.
13. Run the workflow and the array job script “ArrayJob\_shapeit.csh” in the temp folder will be generated.
14. Edit the SGE or PBS parameters accordingly in this script “ArrayJob\_impute.csh” and submit the script to job scheduler.
15. Check the log files in the directory “impute\_logfiles\_sungrid” to make sure all jobs got successful without any error.
16. **Post Processing**
17. Create the config files as mentioned in the manual.
18. select the config file parameter “HAPS=<PATH TO FILE/ shapeit\_jobs.tar.gz”, & “RESTART=POST”. Remove the parameter “CHECK\_POINT\_STOP=IMPUTE” from the config file.
19. Run the workflow and the array job script “ArrayJob\_post.csh” in the temp folder will be generated.
20. Edit the SGE or PBS parameters accordingly in this script “ArrayJob\_post.csh” and submit the script to job scheduler.
21. Check the log files in the directory “post\_logfiles\_sungrid” to make sure all jobs got successful without any error.

## Appendix I: Downloading External Tools (and TOOLINFO FILE)

**We provide two scripts to install the required tools and create the tool info file. If you need to install the tools separately (perhaps some of the tools are already installed in your environment), the instructions are below. However, you can also just run the install\_tools.sh followed by the make\_tool\_info.sh (both tools only require the path to the ezimputer directory).**

**export EZIMPUTER=~/EZIMPUTER**

**Manually downloading tool and tool\_info.txt config file.**

Download the following the tools, unzip them and place them in a common directory. Create a tool info config file for the programs to execute.

1. **PLINK(Version v1.07):** Plink is a free, open-source whole genome association analysis toolset. You can download plink from <http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml> .
2. **STRUCTURE(Version 2.3.2.1):** The program *structure* is a free software package for using multi-locus genotype data to investigate population structure. You can download structure from <http://pritch.bsd.uchicago.edu/structure_software/release_versions/v2.3.4/html/structure.html> .
3. **SHAPEIT(Version V1.X):** SHAPEIT is a fast and accurate haplotype inference software. You can download SHAPEIT from <http://www.shapeit.fr/> .Make sure you download the version specified I.e. V1.x as newer versions (V2) require code changes in the workflow. We tested the newer version and found it to crash.
4. **IMPUTE: IMPUTE (version 2.3)** (also known as **IMPUTE2**) is a genotype imputation and haplotype phasing program. You can download IMPUTE from <http://mathgen.stats.ox.ac.uk/impute/impute_v2.html#download> .
5. **CHECK\_STRAND\_UTILITY(Beagle Utilities):** This is a collection of Beagle utility python scripts to perform strand-checking of input files from different genotyping platforms. You can download from <http://faculty.washington.edu/sguy/beagle/strand_switching/strand_switching.html> .
6. **GPROBSMETRICS(Beagle Utilities):** The gprobsmetrics utility calculates per-marker statistics from Beagle format genotype probabilities. You can download from file.[**http://faculty.washington.edu/browning/beagle\_utilities/utilities.html#gprobsmetrics**](http://faculty.washington.edu/browning/beagle_utilities/utilities.html#gprobsmetrics) **.**

SAMPLE TOOL INFO CONFIG FILE

Here is the sample tool info config file (see below detailed description) [a copy is included as tool\_info.txt in package . and you can use the install\_tools.sh and make\_tool\_info.sh script to create a copy with your own path for the tools.]

PLINK=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/PLINK/plink

STRUCTURE=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/STRUCTURE/console/structure

STRUCTURE\_PARAM=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/STRUCTURE/console/extraparams

SHAPEIT=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/SHAPEIT/shapeit.v1.ESHG.linux.x86\_64

IMPUTE=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/IMPUTE/impute\_v2.3.0\_x86\_64\_static/impute2

CHECK\_STRAND=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/CHECK\_STRAND/check\_strands\_16May11/check\_strands.py

GPROBS=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/GPROBS/gprobsmetrics.jar

PERL=/usr/local/biotools/perl/5.10.0/bin/perl

PYTHON=/usr/local/biotools/python/2.7/bin/python

JAVA=/usr/java/latest/bin/java

QSUB=/home/sge6\_2/bin/lx24-amd64/qsub

SH=/bin/bash

**DETAILED DESCRIPTION TO DOWNLOAD THE TOOLS**

export EZIMPUTER= “/home/EzImputer\_Sample\_project”

#Create main tools directory

mkdir $EZIMPUTER/EXTERNALTOOLS

#Change directory

cd $EZIMPUTER/EXTERNALTOOLS

**#PLINK**

#Create main plink directory

mkdir $EZIMPUTER/EXTERNALTOOLS/PLINK

#Change to plink main directory

cd $EZIMPUTER/EXTERNALTOOLS/PLINK

#Download the plink package

wget <http://pngu.mgh.harvard.edu/~purcell/plink/dist/plink-1.07-x86_64.zip>

#Uncompresszip files

unzip plink-1.07-x86\_64.zip

#The plink binary is now in $EZIMPUTER/EXTERNALTOOLS/PLINK/plink-1.07-x86\_64/plink

# when you enter the path in the too info file, you must replace the value of

# $EZIMPUTER with the actual path, e.g.

echo $EZIMPUTER/EXTERNALTOOLS/PLINK/plink-1.07-x86\_64/plink

**#STRUCTURE**

#Create main STRUCTURE directory

mkdir $EZIMPUTER/EXTERNALTOOLS/STRUCTURE

#Change to STRUCTURE directory

cd $EZIMPUTER/EXTERNALTOOLS/STRUCTURE

#Download the STRUCTURE tool

wget http://pritch.bsd.uchicago.edu/structure\_software/release\_versions/v2.3.4/release/structure\_linux\_console.tar.gz

tar -zxvf structure\_linux\_console.tar.gz

cd console

#test the structure executable with absolute path (change the permission if necessary)

$EZIMPUTER/EXTERNALTOOLS/STRUCTURE/console/structure

#Once you run the executable you should be able to see below

----------------------------------------------------

STRUCTURE by Pritchard, Stephens and Donnelly (2000)

and Falush, Stephens and Pritchard (2003)

Code by Pritchard, Falush and Hubisz

Version 2.3.4 (Jul 2012)

----------------------------------------------------

Reading file "mainparams".

datafile is

infile

Reading file "extraparams".

Note: RANDOMIZE is set to 1. The random number generator will be initialized using the system clock, ignoring any specified value of SEED.

Unable to open the file infile.

Exiting the program due to error(s) listed above.

#If you the program is giving the error “/lib/libc.so.6: version `GLIBC\_2.7' not found”. You need to #download the source and recompile the program

mkdir $EZIMPUTER/EXTERNALTOOLS/STRUCTURE/console/source/

cd $EZIMPUTER/EXTERNALTOOLS/STRUCTURE/console/source/

Get the source package

wget <http://pritch.bsd.uchicago.edu/structure_software/release_versions/v2.3.4/structure_kernel_source.tar.gz>

#uncompress the package

tar -zxvf structure\_kernel\_source.tar.gz

cd structure\_kernel\_src/

#compile

make

Test the new executable

$EZIMPUTER/EXTERNALTOOLS/STRUCTURE/console/source/structure\_kernel\_src/structure

**#SHAPEIT**

#Create main SHAPEIT directory

mkdir $EZIMPUTER/EXTERNALTOOLS/SHAPEIT

#Change to SHAPEIT directory

cd $EZIMPUTER/EXTERNALTOOLS/SHAPEIT

#Download SHAPEIT tool package

wget http://www.shapeit.fr/script/get.php?id=16

#Untar the downloaded package

tar -zxvf shapeit.v1.ESHG.linux.x64.tar.gz

#Try SHAPEIT

$EZIMPUTER/EXTERNALTOOLS/SHAPEIT/shapeit.v1.ESHG.linux.x86\_64 --h

You should be able to see

Phaser options:

--help Produce licence message

--licence Produce licence message

-v [ --version ] Produce version details

-L [ --output-log ] arg (=shapeit\_05032013\_14h55m40s\_b4340e74-e467-4ed8-9bc9-4dee02807b9a.log)

Log file

--exclude-snp arg File containing all the positions of

the SNPs to exclude in input files

--include-snp arg File containing all the positions of

#…(and many more lines)

**#IMPUTE**

#Create main IMPUTE directory

mkdir $EZIMPUTER/EXTERNALTOOLS/IMPUTE

#Change to IMPUTE directory

cd $EZIMPUTER/EXTERNALTOOLS/IMPUTE

#Download IMPUTE tool package

wget http://mathgen.stats.ox.ac.uk/impute/impute\_v2.3.0\_x86\_64\_static.tgz

#Untar the downloaded package

tar -zxvf impute\_v2.3.0\_x86\_64\_static.tgz

#change directory

cd impute\_v2.3.0\_x86\_64\_static

#Try Impute

impute2

#You should be able to see

======================

IMPUTE version 2.2.2

======================

Copyright 2008 Bryan Howie, Peter Donnelly, and Jonathan Marchini

Please see the LICENCE file included with this program for conditions of use.

The seed for the random number generator is 1997316289.

Command-line input: impute2

ERROR: You must specify a valid interval for imputation using the -int argument.

**#CHECK\_STRAND**

#Create main CHECK\_STRAND directory

mkdir $EZIMPUTER/EXTERNALTOOLS/CHECK\_STRAND

#Change to CHECK\_STRAND directory

cd $EZIMPUTER/EXTERNALTOOLS/CHECK\_STRAND

#Download CHECK\_STRAND package

wget <http://faculty.washington.edu/sguy/beagle/strand_switching/check_strands_16May11.tar.gz>

#uncompress the package

tar -zxvf check\_strands\_16May11.tar.gz

#change directory

cd check\_strands\_16May11

# test program

python check\_strands.py

# You should be able to see

usage: python check\_strands.py infileprefixes outfileprefix

outfiles: check\_strands.py.markers check\_strands.py.log

completed combining marker files

done checking frequencies

1. **GPROBS**

#Create main GPROBS directory

mkdir $EZIMPUTER/EXTERNALTOOLS/GPROBS

#Change to GPROBS directory

cd $EZIMPUTER/EXTERNALTOOLS/GPROBS

#Download GPROBS package

wget http://faculty.washington.edu/browning/beagle\_utilities/gprobsmetrics.jar

#Once you are done with downloading next step is to create the tool info config file

**TOOL Config file parameters (You should specify all options)**

**PLINK=/PATH\_TO/plink** [Full Path and filename to plink executable]

**STRUCTURE=/PATH\_TO/STRUCTURE/structure** [Full Path and filename to structure executable]

**STRUCTURE\_PARAM=/PATH\_TO/STRUCTURE/structure.extraparams** [Full path and filename to structure param file (you will get when you download structure)]

**CHECK\_STRAND=/PATH\_TO/check\_strands\_16May11/check\_strands.py** [Full Path and filename to CHECK\_STRANDS PYTHON SCRIPT]

**PERL=/PATH\_TO/perl** [Full Path and filename to PERL]

**PYTHON=/PATH\_TO/python** [Full Path and filename to PYTHON]

**SH=/PATH\_TO/bash** [Full Path and filename to bash, usually /bin/bash]

**SH=/PATH\_TO/bash**[Path to SH/CSH/TSH]

# EXAMPLES

The examples below are included as complete scripts in the downloadable packages. You just have to edit the EZIMPUTER variable to the location of your own installation.. and alternatively substitute your own datasets, location for your own temp space, and your own email.

## WHOLE GENOME IMPUTATION EXAMPLE

**SAMPLE DATA:** You can download build 36 HAPMAP data from <http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/hapmap3_r3/plink_format/>

#Define a directory for EZimputer

export EZIMPUTER=”full path of your own directory”

# note for non bash shell, can try set EZIMPUTER=”full path of your own directory”

#Create a project directory

mkdir $EZIMPUTER

cd $EZIMPUTER

wget <http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/hapmap3_r3/plink_format/hapmap3_r3_b36_fwd.consensus.qc.poly.map.gz>

wget

<http://hapmap.ncbi.nlm.nih.gov/downloads/genotypes/hapmap3_r3/plink_format/hapmap3_r3_b36_fwd.consensus.qc.poly.ped.gz>

#Uncompress(gunzip) them and convert the plink files to plink transpose files (this may take ~1 hour)

gunzip hapmap3\_\*.gz

plink --file hapmap3\_r3\_b36\_fwd.consensus.qc.poly --transpose --recode --out hapmap3\_r3\_b36\_fwd.consensus.qc.poly

# Download and prepare external tools ([here](#TOOLINFO))

#Get the imputation reference by executing the script [Get\_impute\_reference](#GIR).pl.

mkdir $EZIMPUTER/impute\_ref

perl $EZIMPUTER/Get\_impute\_reference.pl -OUT\_REF\_DIR /home/EzImputer\_Sample\_project/impute\_ref -DOWNLOAD\_LINK http://mathgen.stats.ox.ac.uk/impute/ALL\_1000G\_phase1integrated\_v3\_impute.tgz

#Upgrade the hapmap markers from build 36 to build 37 by using the script #[Upgrade\_inputmarkers\_to\_build37\_by\_DBSNP](#UIB)\*.

perl $EZIMPUTER/Upgrade\_inputmarkers\_to\_build37\_by\_DBSNP.pl -DBSNP\_DIR $EZIMPUTER/DBDIR/ -DBSNP\_DOWNLOADLINK ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human\_9606/database/organism\_data/b137\_SNPChrPosOnRef.bcp.gz

-INPUT\_FILE $EZIMPUTER/hapmap3\_r3\_b36\_fwd.consensus.qc.poly.tped -REMAPPED\_CURRENT\_BUILD /home/EzImputer\_Sample\_project/hapmap3\_r3\_b36\_fwd.consensus.qc.poly\_build37.tped -REMAPPED\_OLD\_BUILD $EZIMPUTER/hapmap3\_r3\_b36\_fwd.consensus.qc.poly\_build36.tped .

#Perform QC by executing the script [QC\_fwd\_structure](#QFS)\* .

# Make sure you have already generated a tool info file (when you installed the program)

#Create run config file, download the externals tools and prepare the tool info config file and execute the script.

#RUN INFO CONFIG FILE

# using a text editor, create following file (run\_info.config)

# This will be created automatically by the script in the distribution

TPED=/home/EzImputer\_Sample\_project/hapmap3\_r3\_b36\_fwd.consensus.qc.poly\_build37.tped

TFAM=/home/EzImputer\_Sample\_project/hapmap3\_r3\_b36\_fwd.consensus.qc.poly.tfam

OUTPUT\_FOLDER=/home/EzImputer\_Sample\_project

TEMP\_FOLDER=/home/EzImputer\_Sample\_project/temp

IMPUTEREF\_VERSION=ALL\_1000G\_phase1integrated\_v3

GENOTYPE\_PERCENT\_CUTOFF=0.05

SAMPLE\_PERCENT\_CUTOFF=0.05

IMPUTE\_REF=/home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute

INNER\_DIR=temp1070516921

BEAGLE\_REF\_DB=/home/EzImputer\_Sample\_project/BEAGLE/

HOW TO EXECUTE

/usr/bin/perl $EZIMPUTER/QC\_fwd\_structure.pl -run\_config run\_info.config -tool\_config tool\_info.config

The program will generate the following files

fwdStrandResults\_input.tfam, fwdStrandResults\_input.tped, fwdStrandResults\_input.ind, sample.qc, markers.ignored

Samples and markers are dropped which do not the meet the QC criteria.Perform the whole genome imputation once the QC is performed ([Phase\_Impute\_by\_parallel\_proc](#PIPP)).

Create run config file, download the externals tools and prepare the tool info config file and excute the script.

Compress the tped file

gzip /data2/bsi/RandD/Arraybased\_RND/Easy\_imputer\_test/fwdStrandResults\_input.tped

RUN INFO CONFIG FILE (run\_info.config)

TPED=/home/EzImputer\_Sample\_project/fwdStrandResults\_input.tped.gz

TFAM=/home/EzImputer\_Sample\_project/fwdStrandResults\_input.tfam

FORWARDSTRAND\_IND=/home/EzImputer\_Sample\_project /fwdStrandResults\_input.ind

TEMP\_FOLDER= /home/EzImputer\_Sample\_project/temp

IMPUTE\_REF=/home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute

IMPUTE\_WINDOW=5000000

IMPUTE\_EDGE=125

HAPS=NA

EMAIL=youremail@yourworkplace.edu

SGE\_SHAPEIT\_MEM=5G

SGE\_SHAPEIT\_QUEUE=1-day

SGE\_IMPUTE\_MEM=10G

SGE\_IMPUTE\_QUEUE=1-day

IMPUTEREF\_VERSION=ALL\_1000G\_phase1integrated\_v3

LOCALTEMP\_SHAPEIT=4G

LOCALTEMP\_IMPUTE=1G

INNER\_DIR=temp1070516922

RESTART=NO

USERNAME=\*\*\*\*\*\*\*\*

SHAPEITONLY=NO

LOCALTEMP=/home/EzImputer\_Sample\_project/temp/LOCALTEMP

SHAPEIT\_STATESPHASE=100

PBS=NO

CHR\_START\_INPUT=NO

WINDOW\_CUTOFF\_NUM\_MARKERS=200

EDGE\_CUTOFF\_NUM\_MARKERS=50

TOOL INFO CONFIG FILE (tool\_info.config)

PLINK=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/PLINK /plink

SHAPEIT=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/SHAPEIT/shapeit.v1.ESHG.linux.x86\_64

IMPUTE=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/IMPUTE/impute\_v2.3.0\_x86\_64\_static/impute2

GPROBS=/home/EzImputer\_Sample\_project/EXTERNALTOOLS/GPROBS/gprobsmetrics.jar

PERL=/usr/local/biotools/perl/5.10.0/bin/perl

JAVA=/usr/java/latest/bin/java

QSUB=/home/sge6\_2/bin/lx24-amd64/qsub

SH=/bin/bash

HOW TO EXECUTE

/usr/bin/perl $EZIMPUTER/Phase\_Impute\_by\_parallel\_proc.pl -run\_config run\_info.config -tool\_config tool\_info.config

OUTPUTFILES

Please check temp folder (/home/EzImputer\_Sample\_project/temp/) for the results once the job gets completed.

Out file description can be found [here](#IMPUTERESULTS).

## SAMPLE REGION IMPUTATION EXAMPLE

For example dataset you can extract small region from the whole genome dataset (fwdStrandResults\_input.tped && fwdStrandResults\_input.tfam plink files) using plink. You can keep the same forward strand indicator file as the workflow only considers markers which are present in the input files.

plink --tfile fwdStrandResults\_input --chr 2 --from-kb 3500 --to-kb 6000 --transpose --recode --out small\_region\_fwdStrandResults\_input

Compress tped file

Gzip small\_region\_fwdStrandResults\_input.tped

Create config file with changes to the parameters “CHR\_START\_INPUT”,” SMALL\_REGION\_EXTN\_START” and “SMALL\_REGION\_EXTN\_STOP”.

RUN INFO CONFIG FILE (run\_info.config)

TPED=/home/EzImputer\_Sample\_project/small\_region\_fwdStrandResults\_input.tped.gz

TFAM=/home/EzImputer\_Sample\_project/ small\_region\_fwdStrandResults\_input.tfam

FORWARDSTRAND\_IND=/home/EzImputer\_Sample\_project /fwdStrandResults\_input.ind

TEMP\_FOLDER= /home/EzImputer\_Sample\_project/temp

IMPUTE\_REF=/home/EzImputer\_Sample\_project/impute\_ref/ALL\_1000G\_phase1integrated\_v3\_impute

IMPUTE\_WINDOW=5000000

IMPUTE\_EDGE=125

HAPS=NA

EMAIL=prodduturi.naresh@mayo.edu

SGE\_SHAPEIT\_MEM=5G

SGE\_SHAPEIT\_QUEUE=1-day

SGE\_IMPUTE\_MEM=10G

SGE\_IMPUTE\_QUEUE=1-day

IMPUTEREF\_VERSION=ALL\_1000G\_phase1integrated\_v3

LOCALTEMP\_SHAPEIT=4G

LOCALTEMP\_IMPUTE=1G

INNER\_DIR=temp1070516922

RESTART=NO

USERNAME=m081429

SHAPEITONLY=NO

LOCALTEMP=/home/EzImputer\_Sample\_project/temp/LOCALTEMP

SHAPEIT\_STATESPHASE=100

PBS=NO

CHR\_START\_INPUT= YES

SMALL\_REGION\_EXTN\_START= 2000000

SMALL\_REGION\_EXTN\_STOP= 2000000

WINDOW\_CUTOFF\_NUM\_MARKERS=200

EDGE\_CUTOFF\_NUM\_MARKERS=50

TOOL INFO CONFIG FILE (tool\_info.config)

Should have been created when installing ezimputer.

HOW TO EXECUTE

/usr/bin/perl $EZIMPUTER/Phase\_Impute\_by\_parallel\_proc.pl -run\_config run\_info.config -tool\_config tool\_info.config

OUTPUTFILES

Please check temp folder (/home/EzImputer\_Sample\_project/temp/2) for the results once the job gets completed. As the data contains only of chr 2 so all the impute dosage files can be found in the folder ‘2’.

Out file description can be found [here](#IMPUTERESULTS).

## TWO PLATFORM DATA

If your data is made of two (or more) different platform (for Example: AFFY & ILLUMINA), you can impute from the merged datasets, but you must QC and forward map the alleles separately first. Follow the below steps

1. Separate the dataset (plink transpose files) in to individual platform specific datasets.
2. If the datasets are on build 36 you can convert them to build 37 using the script [Upgrade\_inputmarkers\_to\_build37\_by\_DBSNP](#UIB)\* .
3. Run the QC step separately on each platform specific data ([QC\_fwd\_structure](#QFS)\*).
4. Combine the datasets. If you have same samples in two different platform you can keep the samples with high quality using the sample qc file generated by the QC step. (i.e sample with low missing values)
5. Run the imputation on the combined dataset ([Phase\_Impute\_by\_parallel\_proc](#PIPP)).

**Citations:**

If you like this package, the manuscript has been sumitted.

EZimputer: A workflow for parallelized genome wide imputation.

Hugues Sicotte,Naresh Prodduturi, Julie A. Johnson,Yaxiong Lin, Paul A. Decker,Jeannette Eckel-Passow, Martha E. Matsumoto, Robert B. Jenkins, Shannon K. McDonnell, Mariza de Andrade, and Jean-Pierre A. Kocher

<https://code.google.com/p/ezimputer/>