ImpG 1.0 User Manual

1. Introduction

We provide here documentation for ImpG 1.0 software package. Please report the version of ImpG that is used in your analysis, and please cite Pasaniuc et al “Fast and accurate imputation from summary association statistics” (submitted). Complete details of the methods implemented in this software package can be found in the manuscript. For comments, bugs, and/or suggestions please email Bogdan Pasaniuc (bpasaniuc@mednet.ucla.edu) and Huwenbo Shi (shihuwenbo@ucla.edu).

We start by describing file formats in Section 2 followed by description of binaries in Section 3. Section 4 and Section 5 are dedicated to pipelines for performing imputation into your GWAS. Section 4 describes the pipeline for imputation that includes the computation of weights from 1000 Genomes reference panels. Section 5 describes the pipeline that uses pre-computed weights for imputation.

2. File Formats

2.1 Z-score File

Z-score file must start with a line that contains column labels – SNP ID, SNP position, reference allele, alternative allele, and Z-score – followed by lines of data entries. Each field of data entries must be separated by white spaces.

Note:
1. Users should provide this file for SNPs that are typed in their studies.
2. Data entries must be sorted by SNP positions.

Example:

SNP ID	*SNP Position*	*Ref Allele*	*Alt Allele*	*Z-score*
rs7	*10007*	*A*	*G*	*0.482*
rs50	*10030*	*T*	*C*	*0.245*
rs30	*10050*	*A*	*C*	*0.4805*

2.2 Correlation Matrix File

Correlation matrix file must start with a line that contains SNP IDs, followed by lines of correlations between these SNPs. Each line should contain the correlation between SNP_i and SNP_{i+1}, SNP_{i+2}, ..., SNP_N.

Note:
1. If LD statistics of SNPs are available, users should use ImpG-SummaryLD to impute Z-scores for more accurate result.
2. If LD statistics are not available but individual level genotype data is available, GenLD can be used to compute the LD statistics, and then use ImpG-SummaryLD to impute Z-scores.
3. The order of SNPs must be consistent with the order of SNPs in the Z-score file.

Example:
rs1 rs2 rs3 rs4 rs5 rs6 rs7 rs8
0.15977990 0.20866970 -0.05191686 0.04777456 0.16038020 0.01434770 0.07128912
0.18975133 0.06952873 0.10955660 0.24012401 0.03171390 0.15052799
0.05660377 -0.06512569 0.03065676 0.06920502 0.06256269
0.09567849 0.19095356 -0.03197191 0.12271913
0.21148839 0.21398655 0.14377552
0.00868597 0.18976135
0.25713617

2.3 Genotype File

Genotype file must start with a line that contains SNP IDs, followed by lines of 0s, 1s, 2s that denote the number of reference allele at these SNPs, one line per individual.

Note:
1. If users use genotype file to compute the correlations between SNPs using GenLD, the SNP order must be consistent with the SNP order in Z-score file.

Example:
rs0 rs1 rs2 rs3 rs4 rs5 rs6 rs7 rs8 rs9 rs10 rs11 rs12 rs13
00111111222202
21001200212212
02002212101121

2.4 Imputed Z-score File

Imputed Z-score file starts with a line that contains column labels – SNP ID, SNP position, reference allele, alternative allele, Z-score, and r2pred of Z-score – followed by lines of data entries. Each field of data entries must be separated by white spaces.

Note:
1. Both ImpG-Summary and ImpG-SummaryLD generate this file.

Example:

SNP ID	*SNP Pos*	*Ref Allele*	*Alt Allele*	*Z-Score*	*r2pred*
rs58108140	*10583*	*G*	*A*	*0.100000*	*0.028840*
rs189107123	*10611*	*C*	*G*	*0.100000*	*0.328413*
rs180734498	*13302*	*C*	*T*	*0.020000*	*0.327089*

2.5 Haplotype File

Haplotype file must start with a line that contains column labels – SNP ID, 1000 Genome project sample IDs – followed by lines of 1s and 2s denoting the haplotype (1 for reference allele, 2 for alternative allele) for each individual at each SNP.

Note:
1. Each column in the haplotype file corresponds to one haplotype.
2. SNP order in haplotype file must be consistent with the SNP order in Z-score mapping file.

Example:

SNP ID	*HG00096 HG00096 HG00097 HG00097 HG00099 HG00099*
rs3	*112111*
rs5	*121111*
rs1	*111122*
rs1	*211211*

2.6 SNP Mapping File

SNP mapping file must start with a line that contains column labels – SNP ID, SNP position, reference allele, and alternative allele – followed by lines of data entries. Each field of data entries must be separated by white spaces.

Note:
1. This file is available on ImpG website, and is used to convert Z-scores provided by users such that they are consistent with the reference and alternative allele used by ImpG.
2. SNPs must be sorted by their positions.

Example:

SNP ID	*SNP Pos*	*Ref Allele*	*Alt Allele*
rs40	*10583*	*G*	*A*
rs23	*10611*	*C*	*G*
rs98	*13302*	*C*	*T*

2.7 Beta File

Beta file must start with a line that contains column labels – SNP ID, SNP position, typed SNP IDs, Z-score r2pred – followed by lines of data entries. Each field of data entries must be separated by white spaces.

Note:
1. Both ImpG-Summary-GenBeta and ImpG-SummaryLD-GenBeta generate this file. ImpG-Summary estimates the correlation matrix between SNPs using the reference panel, whereas ImpG-SummaryLD uses the correlation matrix provided by user together with the reference panel.
2. This file, along with the files in the following two sections, are used by ImpG internally.
3. Beta files for certain SNP arrays are available for download on ImpG website.

Example:
SNP ID SNP Pos rs7 rs30 rs50 r2pred
rs0 10000 0.00000000 0.00000000 0.00000000 0.00000000
rs1 10001 0.01117717 -0.01951444 -0.03313261 0.03459354
rs2 10002 0.19646947 -0.04732929 0.03914114 0.64998824
rs3 10003 0.29886462 -0.07330285 -0.03282211 0.76228474

2.8 SNP File

SNP file must start with a line that contains column labels – SNP ID and SNP position – followed by lines of data entries. Each field of data entries must be separated by white spaces.

Note:
1. SNP files contain SNPs that are used for constructing the sigma matrices in ImpG-Summary and ImpG-SummaryLD.
2. SNP files are generated and used by ImpG internally, and are available for download on ImpG website.

Example:
SNP ID SNP Pos
rs12564807 734462
rs3094315 752566
rs3131972 752721

3. Binaries

3.1 ImpG-Summary-GenBeta

ImpG-Summary-GenBeta computes the weights (beta) associated with the Z-scores of typed SNPs for one partition of a chromosome.

Input:
1. Reference haplotype file
2. ImpG SNP mapping file
3. Typed SNPs file

Output:
1. A beta file that contains weights associated with the Z-scores of typed SNPs for one partition of a window.
2. A SNP file that contains the SNPs used for constructing the sigma matrix. This file is generated and used by ImpG internally.

Usage:
-h (required) specify haplotype file
-m (required) specify SNP mapping file
-t (required) specify typed SNP file
-p (required) specify output file prefix
-f (optional) specify minimum MAF (0.01 by default)

Note:
SNPs with very low MAF may contribute to noise in constructing the correlation matrix. The -f option is used to filter out SNPs with very low MAF in constructing the correlation matrix so as to avoid false positives. Please think carefully about the value of the threshold.

Example:
The following command generates beta file for all the SNPs in snps.map, using typed SNPs in snps.typed and reference panel in ref.hap, and generates step1.beta and step1.snp, which store the betas and SNPs for constructing the sigma matrix.

./ImpG-Summary-GenBeta -h ref.hap -m snps.map -t snps.typed -p step1

3.2 ImpG-Summary

ImpG-Summary Imputes Z-scores for untyped SNPs based on the weights computed by *ImpG-Summary-GenBeta* and Z-scores of typed SNPs for one partition of a chromosome.

Input:
1. Beta file and SNP file generated by ImpG-Summary-GenBeta

2. ImpG SNP mapping file
3. Z-scores file for typed SNPs

Output:
1. Imputed Z-scores for all the SNPs in one partition of a chromosome

Usage:
-p (required) specify input file prefix
-m (required) specify SNP mapping file
-t (required) specify typed SNP file
-o (required) specify output file name

Example:
The following command imputes Z-scores for all the SNPs in snps.map using beta file generated by ImpG-Summary-GenBeta and stores the imputed result in imp.zsc.

./ImpG-Summary -p step1 -m snps.map -t snps.typed -o imp.zsc

3.3 GenLD

GenLD generates the correlation matrix from individual-level genotype data.

Input:
1. Z-score file that contains the Z-score for typed SNPs
2. Genotype file for the typed SNPs for a number of individuals

Output:
1. Correlation matrix file for the SNPs

Usage:
-t (required) specify typed SNPs file
-g (required) specify genotype file
-o (required) specify output file name

Example:
The following command generates correlation matrix for all pairs of SNPs in snps.typed using individual-level genotype data in snps.geno, and outputs the matrix in ld.mat.

./GenLD -t snps.typed -g snps.geno -o ld.mat

3.4 ImpG-SummaryLD-GenBeta

ImpG-SummaryLD-GenBeta computes the weights associated with Z-scores of typed SNPs if LD statistics is available.

Input:
1. Reference haplotype file
2. ImpG SNP mapping file
3. Z-score file for typed SNPs
4. Correlation matrix (LD statistics) for typed SNPs

Output:
1. A beta file that contains weights associated with the Z-scores of typed SNPs for one partition of a window.
2. A SNP file that contains the SNPs used for constructing the sigma matrix. This file is generated and used by ImpG internally.

Usage:
-h (required) specify haplotype file
-m (required) specify SNP mapping file
-t (required) specify typed SNP file
-l (required) specify LD statistics file
-p (required) specify output file prefix -p
-f (optional) specify minimum MAF (0.01 by default)

Note:
SNPs with very low MAF may contribute to noise in constructing the correlation matrix. The -f option is used to filter out SNPs with very low MAF in constructing the correlation matrix so as to avoid false positives. Please think carefully about the value of the threshold.

Example:
The following command generates beta file for all the SNPs in snps.map, using typed SNPs in snps.typed, reference panel in ref.hap, and LD statistics in ld, and generates ld_step1.beta and ld_step1.snp, which store the betas and SNPs for constructing the sigma matrix.

./ImpG-SummaryLD-GenBeta -h ref.hap -m snps.map -t snps.typed -l ld.mat -p ld_step1

3.5 ImpG-SummaryLD

ImpG-SummaryLD Imputes Z-scores for untyped SNPs based on the weights computed by *ImpG-SummaryLD-GenBeta* and Z-scores of typed SNPs for one partition of a chromosome.

Input:
1. Beta file and SNP file generated by *ImpG-SummaryLD-GenBeta*

2. ImpG SNP mapping file
3. Z-scores file for typed SNPs

Output:
1. Imputed Z-scores for all the SNPs in one partition of a chromosome

Usage:
-p (required) specify input file prefix
-m (required) specify SNP mapping file
-t (required) specify typed SNP file
-o (required) specify output file name

Example:
The following command imputes Z-scores for all the SNPs in snps.map using beta file generated by *ImpG-SummaryLD-GenBeta* and stores the imputed result in imp.zsc.

./ImpG-SummaryLD -p ld_step1 -m snps.map -t snps.typed -o imp.zsc

4. Complete Pipelines

This section provides complete instructions on using ImpG binaries and ImpG utilities for imputing Z-scores across the whole genome starting from reference panels of haplotypes. Please refer to Section 5 for imputation pipeline that uses pre-computed weights for different genotyping platforms available at ImpG website.

4.1 Data Used by ImpG

Reference panels of haplotypes are used to compute the weights of typed SNPs as follows. For simplicity, data is assumed to be in Beagle format as follows.

1000 Genome Project Study Sample Information File:
http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3/phase1_integrated_calls.20101123.ALL.panel

SNP Mapping File:
http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3/ALL.chr{1..22}.phase1_release_v3.20101123.filt.markers

Phased Haplotype File:
http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3/ALL.chr{1..22}.phase1_release_v3.20101123.filt.bgl.gz

4.2 ImpG-Summary Pipeline

4.2.1 Extract 1000 Genomes Study Samples of A Specific Population

ExtractPop of ImpG utilities extracts the 1000 genome project study sample IDs of a specific population.

Input:
1. A file that contains the information of the 1000 Genome project study samples, including sample ID, population, etc.
Example:
http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3/phase1_integrated_calls.20101123.ALL.panel

Output:
1. A file that contains a list of 1000 Genome project study sample IDs of the specified population, one ID per line.
Example:
HG00096
HG00097
HG00099
HG00100

Usage:
-f (required) specify the 1000 Genome project study sample information file
-c (required) specify the code of the super population, i.e. EUR for Europeans

Example:
The following command extracts the European study sample IDs from the 1000 Genome project.

./ExtractPop -f all.panel -c EUR > eur.panel

4.2.2 Generate SNP Mapping Files For Partitions of A Chromosome

A partition of a chromosome consists of a window and two buffers at two sides of the window on one chromosome. Default window size is 1Mb. Default buffer size is 0.25Mb on each side of the window leading to 1.5Mb in total.

GenMaps of ImpG utilities partitions the SNP mapping file of a chromosome into smaller SNP mapping files, each containing SNP mappings in a partition of one chromosome.

Input:
1. SNP mapping file for all the SNPs on one chromosome
Example:
http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3/ALL.chr{1..22}.phase1_release_v3.20101123.filt.markers

Output:
1. Smaller SNP mapping files, each containing SNP mappings of SNPs in a partition of one chromosome.

Usage:
-m (required) specify input SNP mapping file for all SNPs on one chromosome
-p (required) specify the prefix for output per partition SNP mapping files
-w (optional) specify window size (1Mb by default)
-b (optional) specify total buffer size (0.5Mb by default, 0.5/2Mb on each side of the window)
-s (required) specify map sequence output file name

Example:
The following command partitions the SNP mapping file chr1.map, into smaller SNP mapping files, and stores these files in the directory chr1_maps, with "chr1" as the prefix. The names of the smaller mapping files are stored in chr1_maps.names.

./GenMaps -m chr1.map -p chr1_maps/chr1 -s chr1_maps.names

4.2.3 Generate Haplotype Reference Panel Files for Partitions of A Chromosome

GenHaps in the ImpG utilities generates haplotypes for each partition of one chromosome using the 1000 Genome data for a specific population.

Input:
1. The per partition SNP mapping files for all the partitions generated by GenMaps
2. A file that contains the file names of the per partition SNP mapping files
3. A list of 1000 Genome project study sample IDs for a specific population
4. A file that contains the phased haplotypes for the 1000 Genome project for all the individuals
Example:
http://bochet.gcc.biostat.washington.edu/beagle/1000_Genomes.phase1_release_v3/ALL.chr{1..22}.phase1_release_v3.20101123.filt.bgl.gz

Output:
1. Multiple files, each containing the haplotypes of a specific population for a partition of one chromosome

Usage:
-d (required) specify the directory that contains the mapping files
-s (required) specify the file that contains the names of the mapping files
-a (required) specify the file that contains all the phased haplotypes
-p (required) specify the file that contains the list of study sample IDs
-o (required) specify the output directory of the haplotype files

Example:
The following command generates haplotypes for chromosome partitions stored in chr1_maps/. The names of mapping files are listed in maps.names. The phased haplotypes are stored in phased_haps. Only haplotypes for Europeans are extracted. The per partition reference panels are stored in chr1_haps.

./GenHaps -d chr1_maps/ -s maps.names -a phased_haps -p eur.panel -o chr1_haps/

4.2.4 Partition Typed SNPs

ImpG provides two scripts for partitioning typed SNPs. *GenMaps-Anno partitions SNPs* from an annotation file. *GenMaps-Typed* partitions SNPs from a Z-score file for all SNPs.

4.2.4.1 Partition SNPs from Annotation File

GenMaps-Anno in the ImpG utilities partitions the SNPs in the annotation file according to the partitions generated by *GenMaps*. This script works for any annotation file that contains the list of SNPs in the first column.

Input:
1. The SNP mapping files generated by *GenMaps*.
2. A file that specifies the names of the mapping files.
3. The SNP array annotation file, which contains the SNPs in the first column.
4. The directory to store the output files.

Output:
1. SNP mapping files for typed SNPs for each partition on the chromosome.

Note:
1. Z-scores are initialized to 0. Users need to modify this value (and/or the reference/alternative allele) in the imputation step.

Usage:
-m (required) specify the directory that contains the input mapping files for each partition
-s (required) specify the file that contains the names of the per partition mapping files

-a (required) specify the annotation file, the first column should contain the SNPs
-d (required) specify output directory for the mapping files for the typed SNPs
-t (optional) specify the threshold for filtering out sparse partitions (default value is 10)
-o (required) specify name of the file that contains the mapping files for the typed SNPs

Example:
The following command splits the SNPs in annotation_file according to the partitions in the chr1_maps/ directory, which contain mapping files with names specified maps.name.1, generates mapping files in typed directory, and stores the file names in maps.names.2.

*./GenMaps-Anno -m chr1_maps/ -s maps.names.1 -a annotation_file -d typed/ *
-o maps.names.2

4.2.4.2 Partition SNPs from Z-score File for All SNPs

GenMaps-Typed in the ImpG utilities partitions the SNPs in the Z-score file for all SNPs according to the partitions generated by *GenMaps*. This script works for any annotation file that contains the list of SNPs in the first column.

Input:
1. The SNP mapping files generated by *GenMaps*.
2. A file that specifies the names of the mapping files.
3. The Z-score file for all SNPs, which contains SNP IDs, SNP positions, reference allele, alternative allele, and Z-scores.
4. The directory to store the output files.

Output:
1. SNP mapping files for typed SNPs for each partition on the chromosome.

Usage:
-m (required) specify the directory that contains the input mapping files for each partition
-s (required) specify the file that contains the names of the per partition mapping files
-y (required) specify the Z-score file for all SNPs
-d (required) specify output directory for the mapping files for the typed SNPs
-t (optional) specify the threshold for filtering out sparse partitions (default value is 10)
-o (required) specify name of the file that contains the mapping files for the typed SNPs

Example:
The following command splits the SNPs in zsc_file according to the partitions in the chr1_maps/ directory, which contain mapping files with names specified maps.name.1, generates mapping files in typed directory, and stores the file names in maps.names.2.

*./GenMaps-Typed -m chr1_maps/ -s maps.names.1 -y zsc_file -d typed/ *
-o maps.names.2

4.2.5 Generate Beta Files for All the Partitions of One Chromosome

ImpG-Summary-GenBeta-Chr in the ImpG utilities generates the beta files for all the partitions of one chromosome.

Usage:
-b (required) specify path to ImpG-Summary-GenBeta binary
-s (required) specify the names of mapping files for all the partitions
-d (required) specify the directory that contains all the data for one chromosome
-u (optional) specify the suffix of imputation data directories (empty by default)

Example:
The following command applies ImpG-Summary-GenBeta on all the partitions of chromosome 1. The names of the mapping files for all the partitions are specified in maps.names. All the data for chromosome 1 is stored in chr1 directory.

*./ImpG-Summary-GenBeta-Chr -b ImpG/ImpG-Summary-GenBeta -s chr1/maps.names *
-d chr1/

ImpG-Summary-GenBeta-Chr assumes the following way of organizing data for one chromosome.

chr1/
    haps/
        chr1.1-1250000.map.haps
        chr1.750000-2250000.map.haps
        ...
    maps/
        chr1.1-1250000.map
        chr1.750000-2250000.map
        ...
    typed/
        chr1.1-1250000.map.typed
        chr1.750000-2250000.map.typed
        ...
    betas/
        ...
    imp/
        ...

- *haps* contains the haplotype reference panel data for all the partitions of one chromosome.
- *maps* contains the SNP mapping files for all the partitions of one chromosome.
- *beta* is used to store the beta files for all the partitions of one chromosome.
- *imp* is used to store the imputed Z-scores for SNPs in all the partitions of one chromosome.
- To make imputation on different data sets with the same reference panel convenient, each directory that contains imputation data can have a suffix at the end, e.g. typed_LDL, betas_LDL, imp_LDL.

4.2.6 Impute Z-scores

ImpG-Summary-Chr in the ImpG utilities imputes Z-scores for SNPs in all the partitions of one chromosome.

Usage:
-b (required) specify the path to ImpG-Summary binary file
-d (required) specify data directory for one chromosome
-s (required) specify the file that contains the file names of the per partition SNP mapping files
-u (optional) specify the suffix of imputation data directories (empty by default)

Example:
The following command applies ImpG-Summary on all the partitions of chromosome 1. The names of the mapping files for all the partitions are specified in maps.names. All the data for chromosome 1 is stored in chr1 directory.

./ImpG-Summary-Chr -b ImpG/ImpG-Summary -d chr1/ -s chr1/maps.names

4.2.7 Merge Z-scores

MergeZsc in the ImpG utilities merges the Z-scores files computed by ImpG-Summary into one file.

Usage:
-d (required) specify directory that contain imputed Z-scores files
-s (required) specify the file that contains the file names of the per partition SNP mapping files
-o (required) specify output file name
-w (optional) specify window size (1Mb by default)
-b (optional) specify total buffer size (0.5 Mb by default)
-u (optional) specify the suffix of imputation data directories (empty by default)

Example:
The following command merges Z-score files in the zsc_dir directory, and saves the Z-scores for all SNPs in all.impz. The names of the mapping files for all the partitions are specified in maps.names.

./MergeZsc -d zsc_dir -s maps.names -o all.impz

5. Pipelines Using ImpG with Pre-computed Weights

This section provides instructions on using ImpG binaries and ImpG utilities for imputing Z-scores across the whole genome with pre-computed weights.

5.1 ImpG Web Resources

Reference Panels, SNP Mapping Files and pre-computed Beta Files can be downloaded from ImpG website.

5.1 ImpG-Summary Pipeline

5.1.1 Create Directories and Download Data

ImpG-Summary assumes the following way of organizing data. Users need to download the ImpG web resource data, and double check that the following directory structure holds. When downloading beta files, users need to make sure that they are downloading the beta files that match their SNP array.

genome/
- *chr1/*
  - *haps/*
    - *...*
  - *maps/*
    - *...*
  - *typed/*
    - *...*
  - *betas/*
    - *...*
  - *imp/*
    - *…*
- *chr2/*
  - *…*
- *chr3/*

…

chr4/

…

5.1.2 Compute Z-scores for Typed SNPs

Users need to provide *ImpG-Summary* the Z-scores of the SNPs that are typed in their GWA studies and make sure that the SNP array they use match the beta file they download.

Use *GenMaps-Typed* (introduced in 4.2.4) to partition the typed SNPs according to the mapping files in the maps directory for each chromosome.

Then replace the reference allele, alternative allele, and Z-scores with the ones used and computed by the users.

Put the Z-score files for typed SNPs in the typed directory for each chromosome.

Note:
ImpG automatically checks the encoding (reference and alternative allele) used for computing Z-scores. Users don't need to manually convert the encoding of their Z-scores.

5.1.3 Impute Z-scores

This step is identical to 4.2.6.

5.1.4 Merge Z-scores

This step is identical to 4.2.7.