IMPUTE 5

Introduction

IMPUTE 5 is a genotype imputation method that can scale to reference panels with millions of samples. This method continues to refine the observation made in the IMPUTE2 method, that accuracy is optimized via use of a custom subset of haplotypes when imputing each individual. It achieves fast, accurate, and memory-efficient imputation by selecting haplotypes using the Positional Burrows Wheeler Transform (PBWT). By using the PBWT data structure at genotyped markers, IMPUTE 5 identifies locally best matching haplotypes and long identical by state segments. The method then uses the selected haplotypes as conditioning states within the IMPUTE model.

imp5Converter is a program to convert VCF/BCF reference panels in imp5 file format. imp5 is a file format used by IMPUTE 5 to store reference panels and allows fast read of custom regions, without the need to use compression libraries like ZLIB.

Citation:

If you use IMPUTE 5 in your research, please cite the following publication:

S. Rubinacci, O. Delaneau, J. Marchini (2019) Genotype imputation using the Positional Burrows Wheeler Transform

Documentation

Example files to test IMPUTE 5 and imp5Converter can be found in the *test* directory. Examples shown below assume impute5 binary is called from the *test* directory.

1. IMPUTE5

1.1. Simple run

To run IMPUTE5 with default parameters, use the following command line:

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g target.vcf.gz --r 20 --o imputed.vcf.gz
```

All five options are mandatory and their descriptions are:

- `--h` specifies the haplotype reference panel in VCF/BCF/IMP5 format (must have .vcf[.gz]/.bcf/.imp5 extension). The file must be indexed (tabix/imp5 index).
- `--m` specifies the fine-scale recombination map for the region to be analyzed. Maps for humans can be found HERE.

- `--g` specifies the file containing target haplotypes for a study cohort that you want to impute in VCF/BCF format (must have .vcf[.gz]/.bcf extension). The file must be indexed (tabix index).
- `--r` specifies the target region or chromosome to be imputed (replaces IMPUTE4's -int parameter). Example -r 20 (whole chromosome 20) or -r 20:1000000-5000000 (region within chromosome 20). Buffer parameter will expand this region, if specified.
- `--o` specifies the output filename. Extensions might be added if necessary.

IMPUTE5 will consider as genotype markers only the markers in the intersection between `--g` and `--h`. In practice, it proceeds exactly the same way than `bcftool isec -c none` and therefore consider as genotype markers only the variants with chromosome ID, position, REF and ALT alleles that perfectly match between the two panels. Markers only in the target panel are discarded and markers only in the reference panel are considered imputed markers.

1.2. Log file

To record all the verbose that appear on the screen, use the `--l` option as follows:

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20 --o imputed.bcf --l imputed.log
```

We strongly recommend to use this option for any run.

1.3. Imputing a chunk of data

To impute the 1Mb region located in the genomic interval 2Mb-3Mb, use:

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20:2000000-3000000 --o imputed.vcf.gz
```

`--r` option is mandatory. Double check that the chromosome ID matches one of those specified in the VCF file. A common mistake is to use other specifications for the chromosome different from the one specified in the VCF/BCF file. A quick way to check it would be running:

```
bcftools view -H -G target.vcf.gz | head -n 1 | awk
'{print $1}'
```

Also, please verify that your reference and target panel present the same notation for the chromosome.

Each chunk of imputed data will be expanded by a buffer region, in order to help preventing imputation quality from deteriorating near the edges of the region. Markers in the buffer region will help the inference but do not appear in the output files. Larger buffers can improve accuracy, at the cost of longer running times. Buffer value is expressed in kb:

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20:2000000-3000000 --b 500 --o
imputed.vcf.gz
```

A typical imputation region varies from ~5Mb to 15Mb, depending on the data.

1.4. Output file format

IMPUTE5 file automatically detects the format of the input and output file by the extension. Input can be specified in three different file format: VCF[.gz]/ BCF/IMP5. Output can be specified in three file formats: VCF/BCF/BGEN.

1.4.1 BGEN output

You can choose the compression used by BGEN for the output file format using `--out-compr` parameter (values accepted: `no,zlib,zstd`).

For example, to output a BGEN file compressed using ZSTD you will use:

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20 --out-compr zstd --o imputed.bgen
```

to output a BGEN file compressed using ZLIB you will use:

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20 --out-compr zlib --o imputed.bgen
```

to output a BGEN file compressed with no compression you will use:

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20 --out-compr no --o imputed.bgen
```

1.4.1 VCF/BCF output

VCF and BCF file format contains phased genotypes in the GT field. At imputed markers the VCF/INFO field will contain:

- IMP flag, denoting that the marker is imputed;
- INFO field: containing the IMPUTE INFO score at the variant

The VCF/FORMAT by default has:

- GT, containing the most likely genotype;
- DS, containing the output genotype dosage;

The option **`--out-gp-field`** can be used to output the genotype probabilities in the GP field, while **`--out-ap-field`** can be used to output ALT haplotype probabilities in the FORMAT/AP field.

To output a file vcf.gz file using ZLIB library (vcf.gz) the command uses the following options:

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20 --o imputed.vcf.gz
```

The option **`--out-gp-field`** can be used in combination with VCF/BCF output to add GP (genotype probability) format in the VCF INFO field of imputed markers.

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20 --out-gp-field --o imputed.vcf.gz
```

1.5. Parallelization

1.5.1 Parallelization by chunk

IMPUTE5 parallelise by chunks so that different imputation regions can be imputed at the same time on a different process.

To do this, you just need to run a IMPUTE5 job per imputation region, by exploiting the `--r` parameter, for example running in bash script:

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20:2000001-3000000 --o
imputed.00.vcf.gz &

impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20:3000001-4000000 --o
imputed.01.vcf.gz &

wait
```

will run two different impute5 jobs in parallel on two different imputation regions of size 1Mb.

1.5.1 Parallelization by multi-threading

A single chunk can also be multi-threaded. Multi-threading is only performed on parts of the algorithm (e.g. HMM calculations), therefore is not as efficient as parallelization by chunk.

Multi-threaded parallelization imputes in parallel several individuals, therefore is useful is the number of target samples is large.

To run impute5 on a chunk in parallel, run:

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20 --threads 4 --o imputed.vcf.gz
```

This will run imputation for the chunk using four threads.

1.6. Tuning the PBWT based selection

Reducing the number of conditioning neighbours in the PBWT can be achieved using the `--pbwt-depth` option (called `L` in the paper). The default value is 8.

Decreasing it results in faster runs at the cost of some accuracy.

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20 --o imputed.bgen --pbwt-depth 16
```

To change how frequently the PBWT selection is performed you can use the option `--pbwt-cm`. The default value is 0.02 (cM), meaning that the selection is performed once every 0.02 cM. Decreasing this value will result in more states selected and a possible increase in accuracy. However, the selection step and the HMM will be slower.

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20 --o imputed.bgen --pbwt-cm 0.005
```

You can also change the selection algorithm by using **`--div-select`** turning on SHAPEIT4's diverge select algorithm:

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g
target.vcf.gz --r 20 --o imputed.bgen --div-select
```

1.7. Other options

The parameter **`--ohapcopy`** outputs a list of files (one for each target haplotype) containing the expected amount of sequence (in cM) copied from each reference haplotype in the list of copying states of the target haplotype. The output is in CSV file format, tab separated (gzipped – can be read using zcat).

`impute5 --h reference.bcf --m chr20.b37.gmap.gz --g target.vcf.gz --r 20 --o imputed.vcf.gz` **`--ohapcopy`** hapcopy.list

The first four columns of the output CSV file indicate (i) the ID of the reference haplotype, from 0 to N-1 (ii) the name of the target sample (iii) a boolean value indicating which of the two reference haplotypes. Other columns of the file output the value of the hapcopy (in cM) for each target haplotype. The file has a header, describing each column of the CSV file and the target samples/haplotype.

Since the **`--ohapcopy`** can be used to identify shared segments, it might be appropriate to use the same dataset as target and reference panel in specific situations. For this reason it is possible to ban from the copy list of each target haplotype the reference haplotype sharing the same sample ID (string) using the option **`--ban-hapid`**. Normally no specific order in the target or reference panel is required. However, an exception is made for the case *pbwt-depth* > *N* (where *N* is the number of haplotypes in the reference panel), where we make the assumption that the reference and

target samples have the same order in the both the datasets and the filter is applied at the level of orderings, instead of sample names (target haplotypes 0 and 1 would ban haplotypes 0 and 1 in the reference panel, assuming they refer to the same sample).

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g target.vcf.gz --r 20 --o imputed.bgen --ohapcopy --ban-hapid
```

One of the parameters in the IMPUTE model is *ne*, the effective population size. To change the effective population size value you can use **--ne** option.

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g target.vcf.gz --r 20 --o imputed.bgen --ne 11000
```

The parameter **--nothreshold** allows to skip the thresholding step at the end of the HMM, and keep all the states for imputation. If this parameter is specified imputation might be more accurate, however, the Li and Stephens state probabilities are typically very sparse and this option might make imputation a lot slower.

```
impute5 --h reference.bcf --m chr20.b37.gmap.gz --g target.vcf.gz --r 20 --o imputed.bgen --no-threshold
```

1.8. Option summary

The full list of options can be obtained by running the command:

```
impute5 --help
```

This should output this list of options:

Input

Option		Default value	Description
--h	STRING	-	Haplotype reference panel in VCF/BCF/IMP5 format (must have .vcf[.gz]/.bcf/.imp5 extension). The file must be indexed (tabix/imp5 index).
--m	STRING	-	Fine-scale recombination map for the region to be analyzed.
--g	STRING	-	File containing target haplotypes for a study cohort that you want to impute in VCF/BCF format (must have .vcf[.gz]/.bcf extension). The file must be indexed (tabix index).
--r	STRING	-	Region or chromosome to be imputed (replaces IMPUTE4's -int parameter). Example -r 20 (whole chromosome 20) or -r 20:1000000-5000000 (region within chromosome 20). Buffer parameter will expand this region, if specified.
--b	INT	250	Length of buffer region (in kb) to include on each side of the analysis window specified by the -r option. SNPs in the buffer regions inform the inference but do not appear in output files

State selection

Option	Type	Default value	Description
--pbwt-depth	INT	4	Depth of PBWT indexes to condition on
--pbwt-cm	FLOAT	0.02	Distance in cM where the selection is performed
--ohapcopy	STRING	hapcopy.list	Output a file for each target haplotype. Each file contains the expected amount of sequence (in cM) copied from each reference haplotype in the list of copying states. The output is in CSV file format.
--ban-hapid	NA	NA	Ban two haplotypes with the same sample ID from the state selection. To be used only when the target and the reference panel share (even partially) the same set of individuals. In the case the --pbwt-depth parameter is set to a value >= the number of the haplotypes in the reference panel, the option would assume that the panels list the samples in the same order (no sample name check is performed in this case).
--div-select	NA	NA	Use divergence arrays to select states.

Output

Option	Type	Default value	Description
--l	STRING	-	Location of the log file to be written. If not specified, only console output will be

			generated.
--o	STRING	impute5.out.bgen	Specifies output file name.
--out-compr	STRING	zstd	Specifies the compression of the output file for BGEN file format (to be used with --o [name].bgen. Accepted values: [no, zlib, zstd]
--out-gp-field	NA	-	Print GP field (Genotype probabilities) if output is in VCF/BCF format.
--out-ap-field	NA	-	Print AP field (ALT haplotype probabilities) if output is in VCF/BCF format.

Other parameters

Option	Type	Default value	Description
--help	NA	-	Produces help message, listing all the accepted arguments
--threads	INT	1	Number of threads
--no-threshold	NA	-	Specifies if need to use all forward backward states and not apply a threshold
--ne	FLOAT	20000	Effective population size.

2. imp5Converter

imp5Converter converts a reference panel in VCF/BCF format to the imp5 file format. An imp5 file contains a region within a chromosome. Typically we want to create a IMP5 file for each chromosome in order to perform imputation on different chunks of the chromosome. An imp5 file is also complemented by an index (.imp5.idx), that allows IMPUTE 5 to have random access to the region of the chromosome.

2.1. Simple run

To convert the full reference panel chromosome 20 to an imp5 file you can simply use:

```
imp5Converter --h reference.vcf.gz --r 20 --o
reference.imp5
```

The output is a file named reference.imp5 and an index file named reference.imp5.idx. We can now call IMPUTE 5 and pass the imp5 file as a reference panel (IMPUTE 5's --h option).

2.2. Option summary

The full list of options can be obtained by running the command:

```
imp5Converter --help
```

This should output this list of options:

Input

Option		Default value	Description
--h	STRING	-	Haplotype reference panel in VCF/BCF format (must have .vcf[.gz]/.bcf extension). The file must be indexed (tabix index).
--r	STRING	-	Region containing the whole imputation region (typically a whole chromosome). Example -r 20 (whole chromosome 20).

Output

Option	Type	Default value	Description
--o	STRING	out.ref.imp5	Specifies output file name ('.imp5' suffix will be added automatically if not specified.)

Other parameters

Option	Type	Default value	Description
--help	NA	-	Produces help message, listing all the accepted arguments
--threads	INT	1	Number of threads used for decompression of the input file

Contact

Please join the OXSTATGEN mailing list and then post any questions there

https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=OXSTATGEN