CiBreed
Center for Integrated Breeding Research

HBIMPUTE

Increasing calling accuracy, coverage and read depth in sequence data by the use of haplotype blocks

Torsten Pook

15. OKTOBER 2020

UNIVERSITY OF GOETTINGEN
Animal Breeding and Genetics Group

HBimpute v0.2.18

1 General

HBimpute is an R-package to impute genomic data based on haplotype blocks derived by the R-package HaploBlocker (Pook et al. 2019). Methods are explained in the associated "Increasing calling accuracy, coverage and read depth in sequence data by the use of haplotype blocks" by Torsten Pook, Adnane Nemri, Eric Gerardo Gonzalez Segovia, Henner Simianer and Chris-Carolin Schön.

The developed approach and in particular the HBimpute step is patent pending by KWS SAAT SE & Co. KGaA and University of Goettingen (inventors: Torsten Pook & Adnane Nemri; application number EP20201121.9). This software is free for non-commercial use. If you are interested in using our software commercially (and/or HBimpute in general), please contact Torsten Pook (Torsten.pook@uni-goettingen.de).

2 Installation

The current version of HBimpute requires HaploBlocker v.1.5.0+ (available on github) and vcfR (available on CRAN). HaploBlocker additional requires a working version of RandomFieldsUtils (available on CRAN & github).

```
devtools::install_github("tpook92/HaploBlocker", subdir="pkg")
                    install.packages("vcfR")
  devtools::install_github("tpook92/HBimpute", subdir="pkg")
```

Furthermore the software BEAGLE (preferably version 5.0). This can be done by downloading a beagle.jar at https://faculty.washington.edu/browning/beagle/b5_0.html. On default we expect the beagle.jar to be placed in your R working directory and to be named "beagle5.jar". In case your .jar is placed in a different path you can specify this via the parameter **path_beaglejar**.

To check if you are correctly set up use the vcf-file "test.vcf" containing 2.209 SNPs for 340 DH-lines in maize via following lines code:

```
library(HBimpute)
impute(vcf = "test.vcf", out = "test_file", target_coverage = 0.9)
```

As this file contains mostly telemetric markers and default quality filters are quite conservative the resulting output VCF file will only contain 145 SNPs.

3 In-depth parameter explanations

3.1 General Parameters:

hetero

Default: FALSE; Alt: TRUE

Set this to TRUE in case you are working with non-homozygous material. In this case initial data will be phased and the two resulting haplotypes are using separately in the imputing procedure

chromo

Default: NULL; Alt: vector containing chromosome identifiers (e.g. 1:10)

On default all chromosomes in the vcf-file will be processed. To select a subset of chromosomes to consider use this parameter

out

Default: “out”, Alt: any character string

Choose the name of the output file

path_beaglejar

Default: „beagle5.jar“ Alt: provide the path to your BEAGLE Jar used to perform all beagle imputation sets.

Further parameter in BEAGLE that can directly be controlled in HBimpute the following will defaults chosen according to (Pook et al. 2020) :

beagle_core (default: 10). This corresponds to the nthreads parameter in BEAGLE
beagle_ne (default: 10.000). This corresponds to the ne parameter in BEAGLE

estimate_sv

Default: FALSE alt: TRUE

Set this to TRUE to active the detection of structural variation. All regions with average read-depth above **sv_cut1** (default: 1.3) will be called as duplications, everthing below **sv_cut2** (default: 0.7) as a deletion. Read-depth are smoothed via a kernel density function with bandwidth **sv_window** (default: 250.000bp). Reduce **sv_window** if the interest is in smaller structural variation events.

use_del / **use_cnv**

Default: FALSE, Alt: TRUE

Active this to add deletion / duplication being called in the output vcf-file. Marker positions will be 1bp after the SNP-position

3.2 Quality Filters:

max_hetero

Default: 0.01; Alt: numeric value between 0-1

All markers with higher share of heterozygous calls than this threshold will be removed. This parameter is only active for homozygous material.

maf

Default: 0, Alt: numeric value between {-1, 0-0.5}

All markers with lower minor allele frequency that this will be removed from the set. The default value of 0 will remove all fixated markers. Any negative value will preserve fixated markers.

max_depth

Default: 10, Alt: integer value

All allele calls with more than **max_depth** read supporting this call will be set the **max_depth** – reads to reduce the effect of extremely high read depth (and thereby high weighed in the imputation).

quali_filter

Default: TRUE, Alt: FALSE

Activating this will enable the filtering of the imputed dataset after the HBimpute step to remove marker with appeared low marker quality. Default are set conservative and should only remove extreme cases

max_na

Default: 0.5, Alt: numeric value between 0-1

All markers with more than **max_na** of all variants not called after the HBimpute step are removed from the set as this indicated low marker quality.

min_depth

Default: 0.5, Alt: numeric value between 0-1

All markers with an average read depth below **min_depth** of the average read depth are removed from the set as this indicated low marker quality.

3.3 Haplotype library

hb_data // **hb_map**

Default: NULL alt: haplotype dataset / vector of physical positions (only one chromosome supported!)

To derive the haplotype library not based on the genetic dataset itself but another dataset (HB-array) provide input here.

The following parameter all directly correspond to parameter in HaploBlocker to shape the structure of you haplotype library. For extended documentation readers are referred to the “Guidelines to HaploBlocker”.

window_size

Default: 20, Alt: any integer value

target_coverage

Default: NULL, Alt: numeric value: 0-1 (we usually recommend 0.9 – 0.95).

min_majorblock

Default: 5000, Alt: any integer value

3.4 Expert parameter

Following parameter all should have reasonable input, but depending on the application changing some of those setting could make sense.

min_confi

Default: 4, alt: numeric value 1-Inf

This parameter is to set the ratio between read for one variant and reads for all other variants to call a genetic variant.

hetero_is_missing

Default: TRUE Alt: FALSE

This parameter is only active when (**hetero** == FALSE) and control if heterozygous entries will be set to NA.

overwrite_call

Default: FALSE, Alt: TRUE

Activicate this to always preserve the original base call from the input vcf. Even if HBimpute is seeing clear indication for another variant. Use **overwrite_call_min_depth** (default: 1) to only replace those alleles with at least a given number of reads supporting this call.

overwrite_na

Default: TRUE, Alt: FALSE

Active this to replace HBimpute called of NA (usually due to calling threshold not exceeded) by the original call from the input vcf. Use **overwrite_na_min_depth** (default: 1) to only replace those alleles with at least a given number of reads supporting this call.

3.5 Parameters mostly used in testing:

geno, depth, allele, lines, posi, zero_two_coding

Default: NULL

These inputs can substitute the use of a vcf-file. They come with no advantage besides potentially loading times than for a vcf-file.

del_freq / cutoff / cnv_freq / cnv_min ((Old structural variation module))

Default: 0.1, 0.9999, 0.1, 2

Only markers with at least **del_freq /cnv_freq** of the individuals with identified structural variation are included in the new data panel.

Deletions are called based on the number of present reads being below the cutoff-quantile of the binomial distribution.

CNVs are called if the local read-depth (standardized by the number of individuals considered) in the given marker is higher than **cnv_min**.

ref_panel

Default: NULL Alt: path to an additional vcf-file containing a reference panel

This code is currenlty only written for heterozygous material and not generalized for multiple chromosomes! If you want to use this – contact me so I can update the code!

Basic idea of a reference panel is to use read from individuals not included in the dataset.

extended_output

Default: FALSE Alt: TRUE

Activating this will store an additional RData file containing information on read-depth, estimated structural variation and the haplotype library.

References

Pook, Torsten; Mayer, Manfred; Geibel, Johannes; Weigend, Steffen; Cavero, David; Schoen, Chris C.; Simianer, Henner (2020): {Improving imputation quality in BEAGLE for crop and livestock data}. In: *G3: Genes, Genomes, Genetics* 10 (1), S. 177–188.

Pook, Torsten; Schlather, Martin; de los Campos, Gustavo; Mayer, Manfred; Schoen, Chris Carolin; Simianer, Henner (2019): {HaploBlocker}. {Creation of subgroup specific haplotype blocks and libraries}. In: *Genetics*, 1045-1061.

4 Acknowledgements

This package was developed in the context of the Project “MAZE – Accessing the genomic and functional diversity of maize to improve quantitative traits” (Grant ID 031B0195).

Special thanks to project partners from KWS, TUM and University of Hohenheim for providing the genetic data to test and develop the methods on.