Gimpute Beta

Junfang Chen and Dietmar Lippold

11 September 2017

1. Introduction

1.1 Goal

In order to ensure the reliability and reproducibility of GWAS data, we set up an efficient data pre-processing, imputation and post-imputation pipeline and documented detailed steps with this vignette. We are aiming at creating a manual in an easy-to-follow and user-friendly manner.

1.2 Required Tools

A list of tools are required:

Plink: https://www.cog-genomics.org/plink2

GTOOL: http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html

SHAPEIT: http://www.shapeit.fr/

IMPUTE2: https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

GCTA64: http://cnsgenomics.com/software/gcta/download.html

1.3 Data and Parameters

1.4 Terminology

2. Whole pipeline

Overview of the whole data processing pipeline

2.1 Conversion of chip data to matching genome build

All files named in this subsection are in the directory 1-conversion/.

Input files:

- Raw chip data of the instances in PLINK format: dataProcessResults/0-rawData/plinkFiles/*.#

- Files with meta information about the instances: dataProcessResults/0-rawData/sampleInfo/*.txt
- Conversion file for converting the chip data to a genome build (for its name see section 1.1.4).
- The file with the IDs of the excluded instances (for its name see section 1.1.4).
- The file with the IDs of the excluded probes of the chip (for its name see section 1.1.4).

Processing steps:

1. Use the raw chip data and the files with meta information to generate a file with the following meta data of the instances (in parentheses the names of the columns): family ID in the plink files (FID), individual ID in the plink files (IID), ID in the description files (descID), self identified ancestry (ance; e.g. EA or AA), sex (sex; 1 = male, 2 = female), age (age), group (group; 0 = control/unaffected, 1 = case/affected). All unknown and missing values are represented by the value NA. Lines with a missing value for FID or IID are not contained.
- Output file: 1_01_metaData.txt
2. From the raw chip data remove all excluded instances.
- Output files: 1_02_removedExclInst.%
3. Replace all values for affection (phenotype, group) and sex in the plink fam file by those values in the file 1_01_metaData.txt. For that the missing value for sex is represented by the value 0 (zero), the missing value for affection (group) is represented by the value -9.
- Output files: 1_03_replacedGroupAndSex.%
4. Remove instances which have the missing affection value (i.e. the value -9).
- Output files: 1_04_removedNoGroupId.%
5. Remove instances which have a value for ance (self identified ancestry) in the file 1_01_metaData.txt which is excluded from the dataset (for the excluded ancestry values see section 1.1.3).
- Output files: 1_05_removedWrongAnceInst.%
6. Remove the excluded probes of the chip and check that there are not two probes with the same name afterwards (if there is a name which is contained twice no output file is generated).
- Output files: 1_06_removedExclProbe.%
7. Remove probes which are not contained in the conversion file or which have missing values in the conversion file.
- Primary output files: 1_07_removedUnmapProbes.%
- Output file with the removed probes:1_07_probesUnmapped2ChipRef.txt
8. Remove probes which belong to a SNP or have a position (i.e. a base pair position and chromosome) in the conversion file which is the same as that of at least one other probe.

- Primary output files: 1_08_removedDoubleProbes.%
- Output file with the removed probes:1_08_probesDouble.txt

9. Use the conversion file to replace the probe names by SNP names, to update the SNP chromosomal location and the SNP base pair position and to convert all alleles to the positive strand. For that use the following mapping form chromosome names to numbers: X ! 23, Y !24, XY (pseudo-autosomal region of X) ! 25, MT (mitochondrial) !26.

- Output files: 1_09_updatedSnpInfo.%

10. Correct the chromosome of the SNPs in the pseudoautosomal region by using the plink option --split-x with the name of the genome build of the conversion file.

- Output files: 1_10_changedXyChr.%

11. Remove the SNPs of the chromosomes 24 (Y) and 26 (MT).

- Output files: 1_11_removedYMtSnp.%

2.2 Quality Control

All files named in this subsection are in the directory 2-QC/.

Input files:

- Genotype data of instances for a specific genome build:dataProcessResults/1-conversion/1_11_removedYMtSnp.#

Processing steps:

1. Determine for each SNP of the chromosome 23 from the genotype data the number of male instances which have the value one as the minor allele count for that SNP and remove all SNPs which number is higher than 0.5 % of the number of male instances.

- Primary output files: 2_01_removedSnpHetX.%
- Output file with the number of instances with heterozygous alleles for each SNP of the chromosome 23 before SNP removal (each line contains a SNP name and the respective number, lines are sorted descending by number): 2_01_snpHetXInstNumberBefore.txt
- Output file with the number of instances with heterozygous alleles for each SNP of the chromosome 23 after SNP removal:2_01_snpHetXInstNumberAfter.txt

2. Determine for each male instance the number of SNPs of the chromosome 23 which have the value one as the minor allele count for that instance and remove all instances which number is higher than 15.

- Primary output files: 2_02_removedInstHetX.%
- Output file with the number of SNPs of the chromosome 23 with heterozygous alleles for each instance before instance removal (each line contains an instance ID and the respective number, lines are sorted descending by number): 2_02_instHetXSnpNumberBefore.txt

- Output file with the number of SNPs of the chromosome 23 with heterozygous alleles for each instance after instance removal: 2_02_instHetXSnpNumberAfter.txt
- Output file with the number of instances with heterozygous alleles for each SNP of the chromosome 23 after instance removal: 2_02_snpHetXInstNumberAfter.txt

3. Set all heterozygous alleles of SNPs of the chromosome 23 for males (i.e. when the SNP has the value one as the minor allele count) as missing.
- Output files: 2_03_setHeteroHaploMissing.%

4. Remove SNPs with missingness 0.05 (before instance removal).
- Output files: 2_04_removedSnpMissPre.%

5. Remove instances with missingness 0.02.
- Output files: 2_05_removedInstMiss.%

6. Remove instances with great autosomal heterozygosity deviation, i.e. with |Fhet| 0.2
- Output files: 2_06_removedInstFhet.%

7. Remove from each pair of related instances one of the instances from the dataset, whereat the number of removed instances should be as small as possible (when there are pairs with the same instance). The threshold for the plink option --king-table-filter is 0.11.
- Output files: 2_07_removedInstRelated.%
- Output file with the removed instances IDs and their kinship values: 2_07_instRemovedKinships.txt

8. Replace the paternal ID and maternal ID of instances (childs) by the value zero if the paternal ID and the maternal ID do not belong to any instance (parent) with the same family ID as the child.
- Output files: 2_08_removedParentIdsMiss.%

9. Remove SNPs with missingness 0.02 (after instance removal).
- Output files: 2_09_removedSnpMissPost.%

10. Remove SNPs with difference 0.02 of SNP missingness between cases and controls.
- Output files: 2_10_removedSnpMissDiff.%

11. Remove chrX SNPs with missingness 0.05 in females
- Output files: 2_11_removedSnpMissAllo.%

12. Remove autosomal SNPs with Hardy-Weinberg equilibrium p < 10???6 in controls.
- Primary output files: 2_12_removedSnpHweAutoCt.%
- Output file with HWE p-value for the autosomal SNPs before removing, sorted ascending by the p-values: 2_12_snpHweValuesAutoCt.txt
- Output file with the removed SNP names: 2_12_snpRemovedHweAutoCt.txt

13. Remove chrX SNPs with HWE p < 10???6 in female controls.
- Primary output files: 2_13_removedSnpHweAlloCt.%

- Output file with HWE p-value for the female chrX SNPs before removing, sorted ascending by the p-values:2_13_snpHweValuesAlloCt.txt
- Output file with the removed SNP names:2_13_snpRemovedHweAlloCt.txt

14. From the files 2_12_removedSnpHweAutoCt.# (not from the output of the last step) remove chrX SNPs with HWE p < 10???6 in all female instances.

- Primary output files: 2_14_removedSnpHweAlloAll.%
- Output file with HWE p-value for the females chrX SNPs before removing, sorted ascending by the p-values: 2_14_snpHweValuesAlloAll.txt
- Output file with the removed SNP names: 2_14_snpRemovedHweAlloAll.txt

15. Remove autosomal SNPs with Hardy-Weinberg equilibrium p < 10???10 in cases.

- Primary output files: 2_15_removedSnpHweAutoAll.%
- Output file with HWE p-value for the autosomal SNPs before removing, sorted ascending by the p-values: 2_15_snpHweValuesAutoAll.txt
- Output file with the removed SNP names: 2_15_snpRemovedHweAutoAll.txt

16. For 2_11_removedSnpMissAllo.# calculate the PCA (with plots) and remove the outliers from that dataset (for the definition of the outliers see section 1.1.3).

- Primary output files: 2_16_removedOutliersBeforeHwe.%
- Output file with the IDs of the instances and the values of their eigenvectors from the PCA before removal of instances: 2_16_eigenvecBeforeHwe-beforeRm.txt
- Output file with the plot of the first two PCs from the PCA before removal of instances: 2_16_eigenvecBeforeHwe-beforeRm.pdf
- Output file with the plot of the first two PCs from the PCA after removal of instances: 2_16_eigenvecBeforeHwe-afterRm.pdf
- Output file with the IDs of the removed instances and the values of their eigenvectors, sorted ascending by the PC values: 2_16_eigenvecBeforeHwe-removed.pdf

17. For 2_13_removedSnpHweAlloCt.# calculate the PCA (with plots) and remove the outliers from that dataset (for the definition of the outliers see section 1.1.3).

- Primary output files: 2_17_removedOutliersWithinHwe.%
- Output file with the IDs of the instances and the values of their eigenvectors from the PCA before removal of instances: 2_17_eigenvecWithinHwe-beforeRm.txt
- Output file with the plot of the first two PCs from the PCA before removal of instances: 2_17_eigenvecWithinHwe-beforeRm.pdf
- Output file with the plot of the first two PCs from the PCA after removal of instances: 2_17_eigenvecWithinHwe-afterRm.pdf
- Output file with the IDs of the removed instances and the values of their eigenvectors, sorted ascending by the PC values: 2_17_eigenvecWithinHwe-removed.pdf

18. For 2_15_removedSnpHweAutoAll.# calculate the PCA (with plots) and remove the outliers from that dataset (for the definition of the outliers see section 1.1.3).

- Primary output files: 2_18_removedOutliersAfterHwe.%
- Output file with the IDs of the instances and the values of their eigenvectors from the PCA before removal of instances: 2_18_eigenvecAfterHwe-beforeRm.txt
- Output file with the plot of the first two PCs from the PCA before removal of instances: 2_18_eigenvecAfterHwe-beforeRm.pdf
- Output file with the plot of the first two PCs from the PCA after removal of instances: 2_18_eigenvecAfterHwe-afterRm.pdf
- Output file with the IDs of the removed instances and the values of their eigenvectors, sorted ascending by the PC values: 2_18_eigenvecAfterHwe-removed.pdf

2.3 Lifting to the imputation reference

All files named in this subsection are in the directory 3-lifting/. In the following (time point) stands for one of the values Bh (before HWE tests), Wh (within HWE tests) or Ah (after HWE tests).

Input files:

- QC instance data before HWE tests for a specific genome build (for which see the conversion file in section 1.1.4): 2_16_removedOutliersBeforeHwe.#
- QC instance data within HWE tests for a specific genome build (for which see the conversion file in section 1.1.4): 2_17_removedOutliersWithinHwe.#
- QC instance data after HWE tests for a specific genome build (for which see the conversion file in section 1.1.4): 2_18_removedOutliersAfterHwe.#
- Imputation reference files (for its name see section 1.1.4).
- File with SNPs of the group of the dataset (for its name see section 1.1.3).

Processing steps: 1. If the genome build of the QC instance data is different from the genome build of the imputation reference files lift the data from the first genome build to the second. If the genome build is the same just copy the input files.

(a) Do that for the QC instance data before HWE tests.

- Output files: 3_1_liftedDatasetBh.%

(b) Do that for the QC instance data within HWE tests.

- Output files: 3_1_liftedDatasetWh.%

(c) Do that for the QC instance data after HWE tests.

- Output files: 3_1_liftedDatasetAh.%

2. Remove SNPs for which the name has a different position (i.e. combination of base pair position and chromosome) in the imputation reference files.

- Primary output files: 3_2_removedSnpDiffNamePos.%

- Output file with the SNPs names for which a different position is contained in the imputation reference files: 3_2_snpDiffNamePos.txt

3. Remove SNPs for which the combination of base pair position and chromosome is not contained in the imputation reference files (ignoring the SNP name).

- Primary output files: 3_3_removedSnpMissPos.%
- Output file with the SNPs names for which the combination of base pair position and chromosome is not contained in the imputation reference files: 3_3_snpMissPos.txt

4. Remove SNPs for which the combination of base pair position and chromosome (ignoring the SNP name) has an allele which is not in the imputation reference files for that combination of base pair position and chromosome.

- Primary output files: 3_4_removedSnpDiffAlleles.%
- Output file with the removed SNP names: 3_4_snpDiffAlleles.txt
- Output file with the retained SNPs: 3_4_snpImpRefAlleles.txt

For every SNP in 3_4_snpImpRefAlleles.txt add the name of the current dataset, extended by - (e.g. nonGAIN-Ah), to the file with the SNPs of the group (for its format see section 1.1.5).

2.4 Imputation

All files named in this subsection are in the directory 4-imputation/.

Input files:

- Lifted instance data for the genome build of the imputation reference files: dataProcessResults/3-lifting/3_4_removedSnpDiffAlleles.#
- Imputation reference files (for its name see section 1.1.4).

Processing steps:

1. Remove monomorphic SNPs from the Lifted instance data.

- Primary output files: 4_1_removedMonoSnpBefore.%
- Output file with the removed monomorphic SNPs: 4_1_snpMonoRemoved.txt

2. Use the imputation reference files to generate pre-phase haplotypes by SHAPEIT and then do the imputation by using IMPUTE2.

- Output files: 4_2_imputedDataset.#

3. Remove imputed SNPs with (info < 0.6) where the value info comes from the files *.impute2_info from the imputation.

- Primary output files: 4_3_removedSnpInfoPostImp.#
- Output file with the removed SNP names: 4_3_snpRemovedInfoPostImp.txt
- Output file with the info scores of all SNPs, which consists of two columns, separated by whitespaces, the first the SNPs and the second the info scores: 4_3_snpImputedInfoScore.txt

4. Remove all SNPs which have the same position as a SNP in 4_1_snpMonoRemoved.txt has in Lifted instance data.

- Output files: 4_4_removedMonoSnpAfter.#

5. Add the monomorphic SNPs in 4_1_snpMonoRemoved.txt with their values from the lifted instance data.

- Output files: 4_5_addedMonoSnpAfter.#

6. Remove SNPs which have a non missing value for less then 20 instances.

- Primary output files: 4_6_removedSnpMissPostImp.#
- Output file with the removed SNP names: 4_6_snpRemovedMissPostImp.txt

2.5 Reduction and expansion

All files named in this subsection are in the directory 5-reductAndExpand/.

In the following stands for one of the values Bh, Wh or Ah.

Input files:

- The imputed dataset:dataProcessResults/4-imputation/4_6_removedSnpMissPostImp.#
- The SNPs before imputation:dataProcessResults/3-lifting/3_4_snpImpRefAlleles.txt
- The file with the SNPs with different alleles:dataProcessResults/3-lifting/3_4_snpDiffAlleles.txt
- The SNPs with missing positions: dataProcessResults/3-lifting/3_3_snpMissPos.txt
- The SNPs with different positions: dataProcessResults/3-lifting/3_2_snpDiffNamePos.txt
- The dataset before removing SNPs for imputation: dataProcessResults/3-lifting/3_1_liftedDataset.%

Processing steps:

1. Reduce the imputed dataset to the SNPs before imputation.

- Output files: 5_1_reducedToSpecific.%

2. Add the SNPs with different alleles with their values from the dataset before removing SNPs.

- Output files: 5_2_extSpecificDiffAllele.%

3. Add the SNPs with missing positions with their values from the dataset before removing SNPs.

- Output files: 5_3_extSpecificMissPos.%

4. Add the SNPs with different positions with their values from the dataset before removing SNPs.

- Output files: 5_4_extSpecificDiffPos.%