**Practical 3 – Phasing and Imputation**

**Introduction**

In the association analysis practical, there was a particular SNP that generated a significant p-value. By utilising genotype imputation methods, we can investigate whether there are neighbouring SNPs that also share evidence of association with the phenotype. The focus in this practical will be on imputing genotypes within close proximity of the significant SNP, and then to re-run the association analyses undertaken in Practical 2 on the imputed dataset.

**Pre-phasing and imputation**

Prior to undertaking genotype imputation, we first of all need to ‘pre-phase’ our genotype data. The general idea of pre-phasing is to convert the genotype data into haplotype form. Doing this affords more efficient imputation as the reference panels on which we base our imputation are all in haplotype form as well. Haplotypes span across many locus so it is important that when focusing in on a certain region of the genome, (for example the region surrounding a significant SNP as we are in this practical), that we allow a sufficient buffer zone around the region. If we didn’t allow this buffer zone then we could interrupt haplotype construction with the locus of interest!

To undertake the pre-phasing we use a software package called SHAPEIT2. The webpage for SHAPEIT2 is available at <https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.htm>. The webpage includes a full description of the commands and options for running SHAPEIT2, and is a useful point of reference.

The command we will use to run pre-phasing in SHAPEIT2 includes the following elements (please note that there are many other, alternative and additional options that can be specified in the SHAPEIT2 command, details of which can be found on the webpage):

|  |  |
| --- | --- |
| -input-ped | This specifies the names of the input files (i.e. the original study files which include genotype data at SNPs that were actually genotyped), when the input files are in .ped/.map format. Input files can alternatively be in binary PLINK format (.bed/.bim/.fam) (if so, use ‘–input-bed’) or .sample/.gen format (if so, use ‘-input-gen’). |
| -input- map | This specifies the name of the genetic map file for the reference dataset to be used for the prephasing. |
| -output-max | This specifies the name of the output datasets, which will contain the haplotype data estimated by SHAPEIT. It will produce two files: a ‘.haps’ file and a ‘.sample’ file. The former file contains the genotypic data now in haplotype form whilst the latter contains the patient information, similar to the first six columns of the .ped input file. The --output-max option produces files that are easiest to read for IMPUTE2, the software programme we will use for the imputation of the dataset later. There are other output file types available and these can be explored on the SHAPEIT2 website. |
| -thread | This option can be used to speed up analysis when using a multi-core computer. If we have a 8-core computer, for example, we can specify this to be 8. |
| -input-from | Base pair position for the start of the region for which we wish to pre-phase. |
| -input-to | Base pair position for the end of the region for which we wish to pre-phase. |

The causal SNP we found in practical 2 had base pair position 64803579 and so a sensible region to pre-phase might be from 63Mb to 68Mb, which will include the immediate region surrounding the causal SNP, plus a buffer region at either end. In the course folder, you will find a .ped and .map file called ‘genstudy\_pp’, which is the ‘genstudy’ dataset used in the previous practicals, but with the genotype data thinned down to this region only. Each job submitted to our computer cluster will be run on a processor with 8 cores. Therefore, if we wish to pre-phase the ‘genstudy\_pp’ dataset, we can run the following SHAPEIT2 command:

**shapeit --input-ped genstudy\_pp.ped genstudy\_pp.map** **--input-map chr10.map --output-max genstudy.ph --thread 8 --input-from 63000000 --input-to 68000000**

* *Undertake prephasing of the ‘genstudy\_pp’ dataset, by using the command above.*

Once we have pre-phased our data, we can undertake genotype imputation and to do this we will use a software package called ‘IMPUTE2’. The command we will use to run IMPUTE2 includes the following elements (please note that there are many other, alternative and additional options that can be specified in the IMPUTE2 command, details of which can be found on the webpage, available at: https://mathgen.stats.ox.ac.uk/impute/impute\_v2.html):

|  |  |
| --- | --- |
| -use\_prephased\_g | This option simply tells IMPUTE2 to undertake imputation using pre-phased haplotypes (as opposed to phasing the genotypes first and then imputing, which is an alternative, but sub-optimal option). |
| -known\_haps\_g | This specifies the name of the input data file, which is the ‘.haps’ file that was output from SHAPEIT2. |
| -m | This specifies the reference panel ‘.map’ file on which we want to base our imputation. This file should have three columns: physical position (in base pairs), recombination rate between current position and next position in map (in cM/Mb), and genetic map position (in cM). Reference panel files are available to download at <https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference>, in the necessary format for IMPUTE2. |
| -h | This specifies the reference panel ‘.hap’ file on which we want to base our imputation, and is a file of known haplotypes. The file should have one row per SNP and one column per haplotype. All alleles must be coded as 0 or 1. Again, reference panel files are available to download at https://mathgen.stats.ox.ac.uk/impute/impute\_v2.html#reference, in the necessary format for IMPUTE2. |
| -l | This specifies the reference panel ‘.leg’ file, which is a legend file with information about the SNPs in the ‘.hap’ file. Each file should have four columns: rsID, physical position (in base pairs), allele 0, and allele 1. The last two columns specify the alleles underlying the 0/1 coding in the corresponding ‘.hap’ file; these alleles can take values in {A,C,G,T}. Again, reference panel files are available to download at https://mathgen.stats.ox.ac.uk/impute/impute\_v2.html#reference, in the necessary format for IMPUTE2. |
| -int | This specifies the genomic interval to use for inference, as specified by the lower and upper boundaries in terms of their base pair position. That is, the region within which we wish to impute genotypes. |
| -Ne | This specifies the "effective size" of the population (commonly denoted as ‘N’e in the population genetics literature) from which your study dataset was sampled. This parameter scales the recombination rates that IMPUTE2 uses to guide its model of linkage disequilibrium patterns. Modern imputation analyses typically involve reference panels with significant ancestral diversity, which can make it hard to determine the "ideal" -Ne value for a particular study. Fortunately, imputation accuracy is highly robust to different -Ne values and it is suggested that setting -Ne to 20000 is adequate in the majority of modern imputation analyses. |
| -buffer | This specifies the length of buffer region (in kb) to include on each side of the analysis interval specified by the ‘-int’ option. SNPs in the buffer regions inform the inference but do not appear in output files. Using a buffer region helps prevent imputation quality from deteriorating near the edges of the analysis interval. Larger buffers may improve accuracy for low-frequency variants (since such variants tend to reside on long haplotype backgrounds) at the cost of longer running times. Anything from 250kb to 1000kb is advisable. |
| -o | This specifies the name of the output file – that is, the file that will contain the imputed genotypes. |

After running SHAPEIT2 on our ‘genstudy\_pp’ dataset, we obtained a ‘genstudy.ph.haps’ datafile which contained phased haplotypes for our specified region. We can now run IMPUTE2 on this .haps datafile to impute genotypes at SNPs not originally genotyped, but included on our chosen reference panel. For the purpose of this practical, we will use the reference panel from the 1000 Genomes Project

(<http://www.1000genomes.org>). To do this, we can use the following IMPUTE2 command:

**impute -use\_prephased\_g -known\_haps\_g genstudy.ph.haps -m chr10.map -h chr10.haps.gz -l chr10.leg.gz -int 64300000 65300000 -Ne 20000 -buffer 500 -o genstudy\_imp.gen**

Please note that the interval we specify under the option ‘*-int’ (64300000 to 65300000)*  is narrower than the one we specified in the SHAPEIT2 command. This is because the imputation process will produce genotypes rather than haplotypes. Therefore we can narrow down to a 1Mb region of interest.

* *Undertake imputation of the ‘genstudy.ph,haps’ dataset, by using the command above. This will create a ‘genstudy\_imp.gen’.*

Whilst IMPUTE2 is running a lot of information will be posted to screen. All sections of this monitor IMPUTE2’s progress, however, some also display interesting information about the data you are imputing.

In the ‘data processing’ section there are two sections of interest. When IMPUTE2 is reading ‘Panel 2’ haplotypes this refers to our study data. You will see that 474 patients are detected as well as 2394 SNPS in the analysis interval+buffer region. If you compare this to the next header down for ‘Panel 0’ haplotypes (from 1000 Genomes reference panel) you’ll see there are nearly twenty times the amount of SNPs! This highlights the level to which genotype data is imputed and also why extra scrutiny may be required with determining significance from GWAS.

In the ‘data summary’ section under the ‘analysis region’ header there are type 0, type 1, type 2 and type 3 SNPs. There is a brief explanation of what each of these means at the top of the section. Crucially, there should be type 2 SNPs, as this means there are SNPs where the reference haplotypes and patient haplotypes overlap. These overlaps allow imputation to happen as denser reference haplotypes can be probabilistically matched to the sparser patient haplotypes.

After the Imputation has run, in the screen output there is a section called ‘Imputation accuracy assessment’. The interval here denotes the probability range at which a SNP genotype has been called and the concordance with the patient genotypes. These figures can be improved by increasing the number of patients in the study or with better match reference haplotypes.

Notice that IMPUTE2 produces ‘.gen/.sample’ format files from SHAPEIT2’s haplotype files. There are no options to output your imputed data in any other way from IMPUTE2. If you require the data to be reformatted into different forms then the programme “GTOOL” (<http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html> ) can perform these actions.

**Re-running our univariate analyses of association**

Previously, in Practical 2 when we were dealing only with actually genotyped SNPs, our .gen file (‘genstudy\_qc.gen’) included the first five compulsory columns, followed by three columns for each genotyped SNP. The three columns represented an individual’s probability of having the ‘AA’, ‘AB’ and ‘BB’ genotypes respectively. As we did not have any imputed SNPs in Practical 2, the probability for each genotype was either 0 or 1. However, now that we are dealing with imputed SNPs, the probability for each genotype can be any number between 0 and 1, representing our genotype uncertainty.

* *Take a look at the first row of the file ‘genstudy\_imp.gen’ by typing the following command:*

**head -n 1 genstudy\_imp.gen**

You will notice that there are number other than 0 and 1 in the columns representing genotype probabilities for the SNPs.

To run univariate analyses of association on a dataset which includes imputed genotypes, we can again use the SNPTest programme as we did in Practical 2, except that when dealing with an imputed dataset the element ’method’ within the SNPTest command needs to be specified differently. The ‘method’ element controls the way in which genotype uncertainty is accounted for in the analyses of association, bearing in mind that imputation can only provide a ‘best estimate’ of what the true genotype at a SNP would be. Options for the element ‘method’ are as follows:

|  |  |
| --- | --- |
| threshold | Assumes a particular genotype at a SNP only if accuracy of the genotype call is above a given threshold. The calling threshold is controlled by the flag -call\_thresh and the default calling threshold is 0.9. |
| expected | Uses expected genotype counts, also known as genotype dosages, in the analyses of association. |
| score | Uses a missing data likelihood score test in the analyses of association. |
| ml | Again uses missing data likelihood approach, but with multiple Newton-Raphson iterations to estimate the parameters in the missing data likelihood. |
| em | Again uses missing data likelihood approach, but with an EM algorithm to estimate the parameters in the missing data likelihoodl. |

So, to run a univariate analysis of association on our imputed dataset, assuming the threshold method we would use the following command:

**snptest -data genstudy\_imp.gen genstudy\_qc.sample –o genstudy\_qc\_univariate\_add.imp.out -pheno diseased -frequentist 1 -method threshold**

* *Run an univariate analysis of association between each SNP in the imputed dataset ‘genstudy\_imp’ and the phenotype ‘diseased’, assuming an additive mode of inheritance, and using the threshold method to account for genotype uncertainty, by using the command above.*

The SNPtest output file ‘genstudy\_qc\_univariate\_add.imp.out ‘ can now be scrutinised in much the same way as we did for the output file in practical 2 – i.e. we can produce Manhattan and QQ plots of the resulting p-values – except for one additional and important QC step which needs to occur post-imputation. This QC step involves filtering out SNPs with poor imputation accuracy. The SNPtest output file has a column labelled ‘INFO’, which contains, for each SNP, a number indicating how accurately it has been imputed. The value of INFO will range from 0 (no certainty in imputation) to 1 (perfect accuracy in imputation/actually genotyped SNP). Poorly imputed SNPs are likely to contribute to false positive results whilst removing a large number of imputed SNPs will distort the overall results. Therefore, as a QC step we need to identify an appropriate, yet not too stringent, threshold for the level of accuracy that is allowed. Unfortunately, there is no exact science to choosing the threshold; the SNPtest website (https://mathgen.stats.ox.ac.uk/impute/impute\_v2.html) views INFO thresholds between 0.3 and 0.5 as reasonable. For our analysis we will utilise an INFO threshold of 0.4.

This additional QC step can be implemented very easily by changing the command we used in STEP 1 of the process to re-format the SNPtest output ready for creating Manhattan plots, as per Practical 2 (see under STEP 1, page 5 of Practical 2). STEP 1 needs to change to the following:

STEP 1: Add chromosome number to each row and filter based on INFO score:

**awk 'NR>11&$9>0.4{$1=10;print}' genstudy\_qc\_univariate\_add.imp.out >genstudy\_qc\_new.imp.out**

* *Run STEP 1 using the command above, and then modify STEP 2 and STEP 3 from Practical 2 as appropriate to reformat your SNPtest output ready to create a Manhattan plot.*
* *Utilise QuickManhattan.sh in the same way as in Practical 2 to generate your plot. What is different about this plot compared to Practical 2? Does this plot make you feel that the signal seen previously is a truly positive result?*
* *Extract a list of the SNPs giving the lowest p-values, so that you can investigate them further, by modifying the code you used in Practical 2 as necessary.*

**Re-running our multiple regression analyses of association**

The imputed data can also be run through SNP association analyses using a multiple regression approach. In SNPtest, to adjust our analysis for the covariates age and family, assuming an additive mode of inheritance and the threshold method, the command for running a multiple regression analysis is::

**snptest -data genstudy\_qc.gen genstudy\_qc.sample -o genstudy\_qc\_adjusted\_add.imp.out -pheno diseased -frequentist 1 -method threshold -cov\_names age family**

* *Run multiple regression analysis of association between each SNP and your phenotype ‘diseased’, assuming an additive mode of inheritance by running the above command.*
* *Take a look at the first 2 lines of your output file by typing the following command:*

**head -20 genstudy\_qc\_adjusted\_add.imp.out**

* *Prepare a Manhattan plot and QQ-plot of the results from your multiple regression analysis by adjusting the code used for plotting the univariate results. Remember that you need to filter out SNP with an INFO metric <0.4, as for the univariate analysis undertaken above.*
* *Obtain a list of the lowest p-values from this analysis, and compare them with your lowest p-values from the univariate analysis.*
* *Overall, do you feel that the signal is a true positive or false positive? What possible further actions either statistically or otherwise could you take to reassure you further?*
* *If you have time, experiment with undertaking a multiple regression analysis of association on the imputed dataset, but this time using the ‘expected’ method in SNPtest.*