Practical: Imputation

Objectives

1.Phase a set of subjects who have been directly genotyped

2.Impute phased haplotypes using the 1000 Genomes reference data

3.Carry out an association study on imputed data and interpret the results

Log In

On a windows machine, press on the start button and type "Putty" into the search box.

In Putty enter the Host Name "bluecrystalp3.acrc.bris.ac.uk" and click open.

A new window will pop up, type in your username, press enter, and similarly with your password. Don't worry if the password doesn't show up on the screen when you are typing.

The command line will now be visible which looks like: [username@newblue3 ~]$

Finally, run the following command to access a compute node: qsub -I -q teaching -l nodes=1:ppn=1, walltime=02:00:00. This will give you access to the unix command line.

You should be now able to complete the practical.

Data

Data (directly genotyped data, genetic maps, reference halplotypes) for this practical is available in pract6_Imputation/data.

We will be using the ‘clean’ GWAS dataset that you encountered in "Practical 3 : Genomewide association study in Plink".

We will be examining the transmembrane protein 18 (TMEM18) gene on the p telomere of chromosome 2 (2p25.3).

Scripts (files containing commands) can be found in pract6_Imputation/scripts. We will be looking at the input and output files of these scripts, the locations of which can be obtained from the the script (e.g. by applying the less or head unix command).

You can save your output to pract6_Imputation/output

We will not have enough time to phase and impute the data. If the program is taking too long and you are ready to move on, please press "control" and "z" together. This will stop the program and then refer to the ready made output is available in pract6_Imputation/results.

Exercise 1 phasing the first 5 megabases (mB) of chromosome 2

Take a look at the genetic map file (e.g. head -n 10 data/geneticMap/genetic_map_chr2_combined_b37.txt)

Navigate to the scripts folder (cd scripts).

Run the script phase.sh

Take a look at the output file results/geno_qc_TMEM18.phased.haps

Question 1:

How would the phasing algorithm use recombination rates ?

Question 2:

How many subjects and SNPs are in the sample data?

Question 3:

How many haplotypes, then, will be in the sample and how many SNPs?

Question 4:

What do the the runs of 0's and 1's represent in the output haplotype file? hint (check https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#hapsample)

Question 5:

How many rows and columns would you expect in this file? Confirm how many there actually are.

Exercise 2 Impute haplotypes using the 1000 Genomes reference data

Navigate to the scripts folder (cd scripts).

Take a look at the script impute.sh.

Take a look at the reference haplotype and legend file.

zcat ../data/haplotypes/ALL.chr2.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.nomono.haplotypes.gz | head -n 10

zcat ../data/legend/ALL.chr2.integrated_phase1_v3.20101123.snps_indels_svs.genotypes.nomono.legend.gz | head -n 10

Run the script impute.sh

Take a look at the log file impute2.log

Question 6:

What do the following options mean with respect to the imputation process?

-use_prephased_g

-known_haps_g

-m

-l

-h

-int

Question 7:

How many SNPs and samples are being used for imputation from the target data, from the reference data? (Hint check https://mathgen.stats.ox.ac.uk/impute/impute2_overview.html). How many samples and SNPs are in the output.

Question 8:

Describe and interpret the concordance table (hint https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#concordance_tables)

Question 9:

What value of info score should be filtered on? What proportion of SNPs are removed after filtering on this info score. Why is it a good idea to filter on info score (hint https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#-i and https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#info_metric_details)

Question 10:

What do the first eight columns represent in the imputation output. What is the most likely genotype of the first person at the first SNP (columns 6 to 8). What is the dosage of the C allele.

Exercise 3 run an association analysis on the imputed results

We started this practical using the output of practical 3 "Genomewide association study in Plink", a cleaned set of genotyped data (which has also being aligned to the forward strand). We will now re-examine the association on chromosome 2 after imputing the data.

check the GWAS significant associations in the observed data (in the results folder)

Run the script assoc.sh and take a look at the association results

Question 11:

What does "aligned to the forward strand" mean?

Question 12:

How many significant associations on chromosome 2 are there in the observed data? What are the names, reference alleles, betas and P values of these SNPs?

Question 13:

How many significant associations on chromosome 2 are there in the imputed data? For the two significant SNPs in the observed data what are there association statistics after imputation? What do the top associations in the imputed data look like?

Answer to questions

Answer 1:

The phasing algorithm can use recombination rates to help estimate the probable relationship between haplotypes. For example if two in phase haplotypes are are observed to be 1100 and 0001, then 1101 (formed by recombination) is more likely than 1111 (not possible by recombination or mutation alone).

Answer 2:

add plink as a module - "module add apps/plink-1.07"

obtain the path of the samples to be phased by taking a look the "phase.sh" script

run plink with no output - plink --bfile ../data/directlyGenotypedData/geno_qc_TMEM18 --noweb

There are 651 SNPs and 8,237 samples

Answer 3:

The number of haplotypes in the sample is twice the number of subjects, in this case 2*8237 = 16,474. There are still 651 SNPs.

Answer 4:

These 0's represent the first listed allele (column 4) and the 1's represent the second listed allele (column 5). Each sucessive pair of digits in each row represent a pair of genotypes (either 00, 01, 11) for one subject in the reference panel. Each column, (starting from column 6) represents a haplotype and each pair of successive columns represents two phased haplotypes of a subject.

Answer 5:

There should be 651 rows in this file, one for each marker and 16,474 columns one for each haplotype plus 5 for the chromosome, SNP name, position, first allele, second allele, giving 16,479 in total.

This can be confirmed using some simpl-ish unix functions

```
wc -l ../results/geno_qc_TMEM18.phased.haps
```

```
head -n 1 ../results/geno_qc_TMEM18.phased.haps | sed 's/ /\n/g' | wc -l
```

Answer 6:

-use_prephased_g - this implements imputation using genetic data that has already been phased into two haplotypes per person

-m - the indicates the genetic map which contains recombination rates

-l - this indicates the legend file which contains positions and alleles for the markers in the reference data

-h - this indicates the reference data haplotypes

-int - this gives the region currently to be imputed

Answer 7:

There are 651 SNPs and 8,237 samples in the target data as before. There are 49,970 non-overlapping SNPs and 1092 subjects (2184 haplotypes) in the reference data. There are 50,621 (49,970 + 651) SNPs and 8,237 samples in the output.

Answer 8:

For this table only, each directly genotyped SNP is treated as missing, then imputed, and the agreement between a subjects genotype and its imputed version is summarized across all subjects. This is done for all directly genotyped SNPs and the results summarized in the concordance table. The table has three columns. The first indicates (ranks) the quality of the imputation on a scale of 0 to 1 for each genotype, the second indicates the number of genotypes with this level of quality and the third column indicates the amount of concordance between these genotypes and their imputed counter parts.

The concordance table is useful to spot a problem with the imputation. The number in the top right hand corner of the table gives the percentage of all genotypes which match their imputed counterparts. This should be over 95%.

Answer 9:

The info score ranges from 0 to 1, where 1 indicates an imputation with near certainty. It is typically to filter on an info score of 0.5 using Impute2 output, but this can vary from imputation to imputation and can be investigated by examining whether any inflation of test statistics from an association analysis are determined by what info score is used to filter on.

If we filter on a score of 0.5 we get

```
awk '{ if ($7 > 0.5) print }' ../results/geno_qc_TMEM18.phased.haps.impute2_info | wc -l
```

23638 SNPs. In other words about 50% of our SNPs are below this level.

It is a good idea to remove poorly imputed SNPs as they are unlikely to represent the true genotypic values and an association signal they represent may be unreliable.

Answer 10:

```
head -n 1 ../results/geno_qc_TMEM18.phased.haps.impute2 | cut -d ' ' -f1-8
```

The first 8 colums represent SNP id which is left blank at present, rsid, base pair position, the first allele, the second allele, the probability that the first person is homozygous for the first allele, the probability that the first person is heterozygous, the probability that the first person is homozygous for the second allele.

The first person is most likely a carrier of the CC genotype.

The corresponding dosage of the C allele would be 2.

Answer 11:

A SNP will indicate two possible bases at a genomic location, e.g. T/G. Each base pairs with a complementary base on the DNA strand. In this case T binds with A and G binds with C. Therefore it would be just as informative to identify the possible bases at this SNP as A/C. The difference here is that one is on the forward strand of DNA and one on the backwards strand of DNA. It is important that the target data and the reference data are coded on the same strand, to avoid the phasing and imputation algorithms resulting in an error.

Answer 12:

```
awk '{ if ($9 < 5.e-8) print }' ../results/bmi_clean.assoc.linear.add
```

There are two significant SNPs: rs2867125 (beta -0.6, reference allele T, P value 1.6e-09) and rs7561317 (beta -0.6, reference allele A, P value 1.6e-09).

Answer 13:

We use unix to print out relevant association statistics from the imputed results. awk '{ if (($9 > 0.5) && ($21 < 5.e-8) && ($19 > 0.01) && ($19 < 0.99)) print }' ../results/BMIphenImputedResults.txt | grep -v 'model_not_fit' | wc -l

There are now 218 associations. Columns 2, 6, 9, 19, 21, 23 are marker id, reference allele, info score, MAF, P value and beta respectively

```
awk '{ if ( ( $9 > 0.5) && ($21 < 5.e-8) && ($19 > 0.01) && ($19 < 0.99)) print $2" "$6" "$9" "$19" "$21" "$23 }' ../results/BMIphenImputedResults.txt | egrep 'rs2867125|rs7561317'
```

This gives the results below. These results match those in the observed data as expected.

```
rs2867125 C 1 0.167658 1.60884e-09 0.612093
```

```
rs7561317 G 1 0.167658 1.63548e-09 0.612093
```

Lets take a look at the top associations

```
awk '{ if ( ( $9 > 0.5) && ($21 < 5.e-8) && ($19 > 0.01) && ($19 < 0.99)) print }' ../results/BMIphenImputedResults.txt | grep -v 'model_not_fit' | awk '{ print $2" "$6" "$9" "$19" "$21" "$23}' | sort -g -k 5 | head
```

There are many SNPs with P values lower or close to those in the observed data. Imputation quality is also generally high. Each would be a good candidate for futher examination (e.g. protein coding changes, eqtl site? etc).