# MA5106 Practical 3:

## Meta Analysis, Phasing, and Imputation

*Tyler Medina*
*ID 13101308*

In this practical, chromosome 22 data from a case-control study and a replicate study will be combined and analyzed for significant SNP association with a binary phenotype. Further data will be added through imputation to add power to the analysis. For both sets of data QC has already been run prior to analysis.

The tools SNPTEST2, META, QCTOOL, SHAPEIT2, and IMPUTE2 will be used via the HPCC in the command line, and the associated commands run for them will be shown.

---

## Part I: Meta Analysis

The first set of data to be analyzed comprises approximately 2000 individuals, half of which are case and half of which are control, based on a binary phenotype. The data for each of these groups is stored as a pair of files, ".gen" containing the SNP genotypes across all individuals and ".sample" containing phenotype and covariate information for each individual.

### Individual Information

The two excerpts below show excerpts from the .sample files for the case and control groups. The individual data includes:

- id_1: The individual's ID
- id_2: The individual's family ID
- missing: The amount of missing genotype information for the individual

The individuals' phenotype and covariate information is classed as follows, and is indicated as such in each column:

- P: phenotype, continuous
- B: phenotype, binary
- C: covariate, continuous
- D: covariate, discrete

In this analysis, "pheno1" will be used for association testing.

**Case Group:**

```
cases <- read.table("cases2.sample")
colnames(cases) <- c()
head(cases)
```

```
##
## 1  id_1  id_2 missing heterozygosity pheno1    pheno2    cov1 cov2
## 2     0     0       0              C      B         P       C    D
## 3 A1001 B1001       0       0.279731      1 -0.410022  2.2682    0
## 4 A1002 B1002       0       0.307783      1   1.17131 1.83839    0
## 5 A1003 B1003       0       0.273805      1  -2.32686 1.45202    1
## 6 A1004 B1004       0       0.337218      1   2.08926 1.17157    0
```

**Control Group:**

```
controls <- read.table("controls2.sample")
colnames(controls) <- c()
head(controls)
```

```
##
## 1 id_1 id_2 missing heterozygosity pheno1   pheno2      cov1 cov2
## 2    0    0       0              C      B        P         C    D
## 3   A1   B1       0       0.321217      0 -1.02693   1.37905    1
## 4   A2   B2       0       0.304425      0 -0.35523 0.0469157    0
## 5   A3   B3       0       0.319834      0 -1.14851  0.650471    0
## 6   A4   B4       0       0.289214      0 0.450384  -2.64719    1
```

---

### Genotype Data

The following table excerpts show the genotyping data for the SNP across all individuals in each group, which is contained in the .gen files. Each row contains the following:

- SNP name
- SNP reference ID
- SNP locus
- Allele "A"
- Allele "B"
- A sequence of bit triplets indicating the genotype for each individual, each indicating either:
  - Homozygous A (1 0 0)
  - Heterozygous (0 1 0)
  - Homozygous B (0 0 1)

Each row contains approximately 1000 triplets, so the following rows have been truncated to only the first two individuals i.e first 6 bits.

**Case Genotypes**

```
case_geno <- read.table("cases2.gen")
colnames(case_geno) <- c()
case_geno[c(1:5),c(1:11)]
```

```
##
## 1 SNP1  rs915677 14433758 A G 0 0 1 0 0 1
## 2 SNP2 rs9617528 14441016 C T 0 0 1 1 0 0
## 3 SNP3  rs140378 15257135 C G 1 0 0 1 0 0
## 4 SNP4  rs131564 15258423 C G 0 0 1 0 1 0
## 5 SNP5 rs5748616 15268900 C G 1 0 0 0 1 0
```

**Control Genotypes**

```
control_geno <- read.table("controls2.gen")
colnames(control_geno) <- c()
control_geno[c(1:5),c(1:11)]
```

```
##
## 1 SNP1  rs915677 14433758 A G 0 0 1 0 0 1
## 2 SNP2 rs9617528 14441016 C T 0 0 1 0 0 1
## 3 SNP3  rs140378 15257135 C G 1 0 0 1 0 0
## 4 SNP4  rs131564 15258423 C G 0 1 0 0 1 0
```

```
## 5 SNP5 rs5748616 15268900 C G 0 1 0 0 1 0
```

---

### Association Test

First, each SNP will be tested for significance association with the phenotype using the SNPTEST2 software tool:

```
./snptest_v2.5.2 -data cases2.gen cases2.sample controls2.gen controls2.sample \ -frequentist
1 -bayesian 1 -method score -pheno pheno1 -o snptest2.txt
```

- `-data`: Passes in the .gen and .sample files to use in the analysis
- `-frequentist 1`: Runs a frequentist association test, "1" specifying an additive genetic model
- `-bayesian 1`: Runs a Bayesian association test, "1" again specifying an additive genetic model
- `-method score`: Instructs the tool how to handle missing SNP information
- `-pheno pheno1`: Uses the "pheno1" phenotype information for the association test
- `-o snptest2.txt`: Writes the test results to a file called "snptest2.txt"

The resultant output displays a vast array of summary information per SNP. The excerpt below only shows 7 of the 53 columns of the output data, but includes minor allele frequencies for each group, the frequentist p-value, and the $\log_{10}$ Bayes factor.

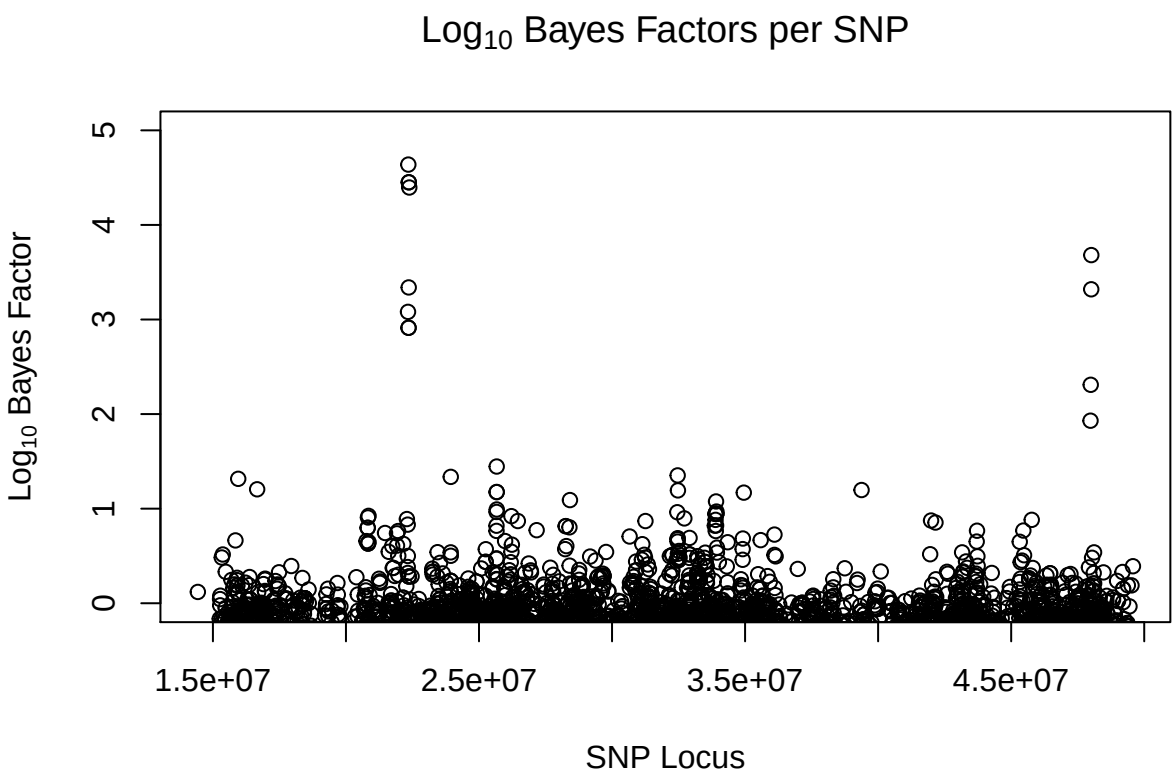
**SNPTEST2 Output:**

```
test1 <- read.table("snptest2.txt", header=T, as.is=T)
summary_t1 <- test1[c(1:5),c(1,2,4,34,35,46,50)]
colnames(summary_t1) <- c("Name", "RSID", "Position",
                          "Case MAF", "Control MAF", "Freq. P-Value", "Log10 Bayes")
summary_t1
```

```
##   Name      RSID Position  Case MAF Control MAF Freq. P-Value Log10 Bayes
## 1 SNP1  rs915677 14433758 0.0608875   0.0494845     0.1164610   0.1194090
## 2 SNP2 rs9617528 14441016 0.3338490   0.3226800     0.4636440  -0.3893160
## 3 SNP3  rs140378 15257135 0.0340557   0.0360825     0.7287940  -0.1646880
## 4 SNP4  rs131564 15258423 0.2858620   0.3025770     0.2546530  -0.2282160
## 5 SNP5 rs5748616 15268900 0.3126930   0.3371130     0.0992922   0.0440162
```

The data can then be visualized as a scatterplot of SNP locus versus $\log_{10}$ Bayes factor.

```
plot(test1$position, test1$bayesian_add_log10_bf, ylim = c(0, 5),
     main=expression("Log"[10]*" Bayes Factors per SNP"), xlab="SNP Locus",
     ylab=expression("Log"[10]*" Bayes Factor"))
```




As shown in the graph, the SNPs with a significant Bayes factor ($\log_{10}$BF $> 3$) are as follows:

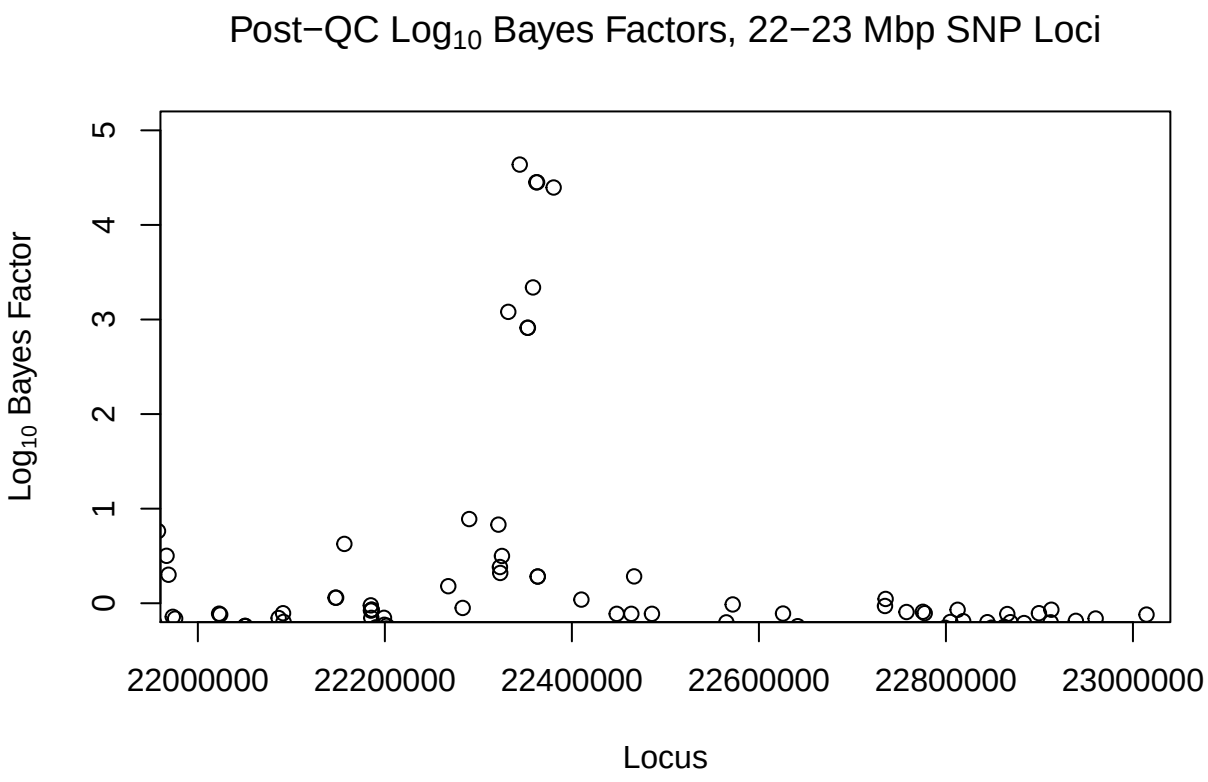
**Significant SNPs**

```
sig_test1 <- test1[which(test1$bayesian_add_log10_bf>3), c(1,2,4,50)]
sig_test1[order(-sig_test1[,4]),]
```

```
##      alternate_ids       rsid position bayesian_add_log10_bf
## 787          SNP829 rs12157657 22344229                4.63805
## 792          SNP834  rs7287369 22362015                4.45090
## 793          SNP835 rs13057362 22362771                4.45090
## 796          SNP838 rs17629956 22380471                4.39570
## 4854        SNP5033  rs2858613 48011290                3.68108
## 790          SNP832 rs11090280 22358412                3.33909
## 4853        SNP5032  rs2858612 48009185                3.31931
## 785          SNP827  rs5751704 22331975                3.08084
```

Two areas on chromosome 22 show significance. For clearer resolution, these areas can be graphed with a narrowed locus range. The graph below shows the region around the first group of significant SNPs.

```
plot(test1$position, test1$bayesian_add_log10_bf, ylim=c(0, 5),
     main=expression("Post-QC Log"[10]*" Bayes Factors, 22-23 Mbp SNP Loci"),
     xlab="Locus", ylab=expression("Log"[10]*" Bayes Factor"),
     xlim=c(22000000,23000000))
```




---

**Replicate Data Set**

The second set of data can be analyzed in the same way as the first set, beginning with a SNPTEST2 run:

```
./snptest_v2.5.2 -data rep.cases.gen rep.cases.sample rep.controls.gen rep.controls.sample
\ -frequentist 1 -bayesian 1 -method score -pheno pheno1 -o snptest3.txt
```

An excerpt of the output is shown below.

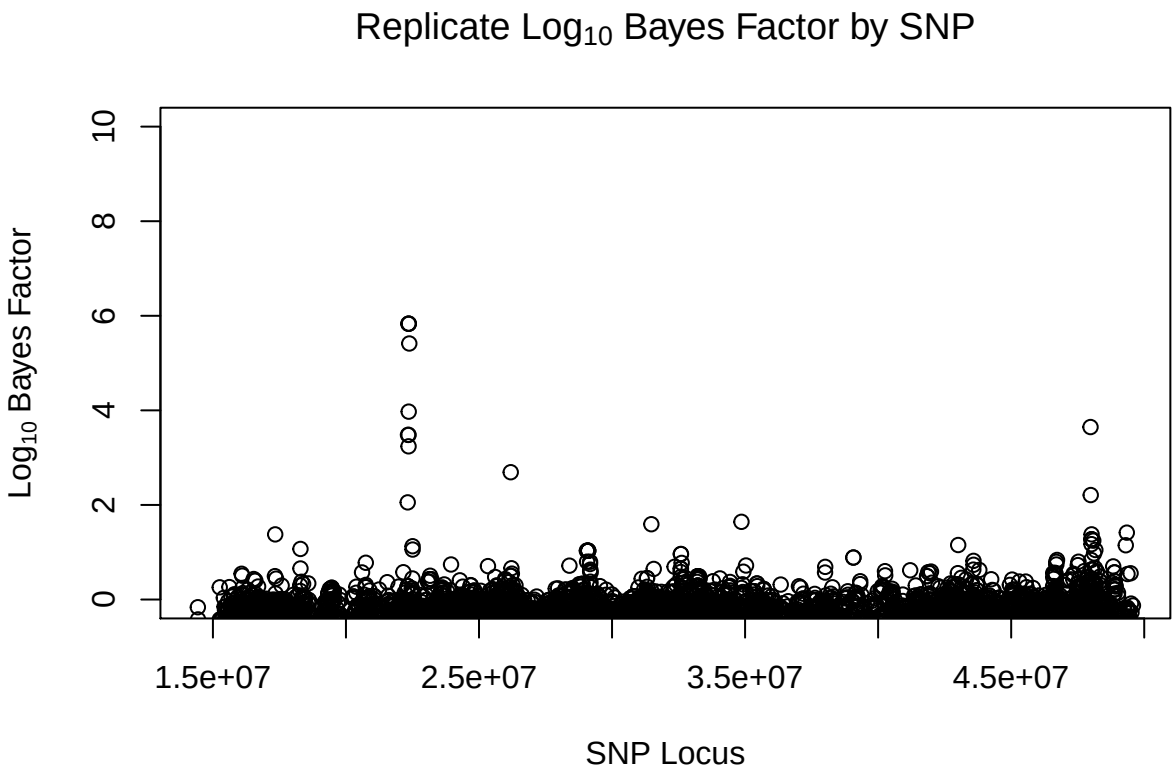
**SNPTEST2 Output, Second Study:**

```
rep_test <- read.table("snptest3.txt", header=T, as.is=T)
summary_rep <- rep_test[c(1:5),c(1,2,4,34,35,46,50)]
colnames(summary_rep) <- c("Name", "RSID", "Position", "Case MAF",
                           "Control MAF", "Freq. P-Value", "Log10 Bayes")
summary_rep
```

```
##   Name      RSID Position Case MAF Control MAF Freq. P-Value Log10 Bayes
## 1 SNP1  rs915677 14433758   0.0615      0.0555     0.4274060   -0.162910
## 2 SNP2 rs9617528 14441016   0.2380      0.2445     0.6344150   -0.420667
## 3 SNP3  rs140378 15257135   0.0595      0.0465     0.0643214    0.258161
## 4 SNP4  rs131564 15258423   0.2020      0.2050     0.8114660   -0.421778
## 5 SNP5 rs5748616 15268900   0.2170      0.2180     0.9382990   -0.441420
```

This data can be visualized with a scatterplot to ensure that the second dataset mirrors the original study.

```
plot(rep_test$pos, rep_test$bayesian_add_log10_bf, ylim=c(0, 10),
     main=expression("Replicate Log"[10]*" Bayes Factor by SNP"),
     xlab="SNP Locus", ylab=expression("Log"[10]*" Bayes Factor"))
```




The similarity of this plot to the original data plot gives evidence that the two data sets are suitable for use in a meta-analysis of the combined data. The significant SNPs in the second set are shown below.

**Significant SNPs in Second Study**

```
sig_rep <- rep_test[which(rep_test$bayesian_add_log10_bf>3), c(1,2,4,50)]
sig_rep[order(-sig_rep[,4]),]
```

```
##      alternate_ids       rsid position bayesian_add_log10_bf
## 829         SNP829 rs12157657 22344229               5.83257
## 834         SNP834  rs7287369 22362015               5.83257
## 835         SNP835 rs13057362 22362771               5.83257
## 838         SNP838 rs17629956 22380471               5.41393
## 832         SNP832 rs11090280 22358412               3.97334
## 5026       SNP5026  rs5770018 47981564               3.64691
## 827         SNP827  rs5751704 22331975               3.48156
## 831         SNP831  rs2330578 22352975               3.48156
## 830         SNP830   rs738785 22352631               3.23944
```

Of these SNPs, the following are found to be significant in both the original and replicate studies:

**Replicated Significant SNPs**

```
rep_snps <- sig_test1[sig_test1$alternate_ids %in% sig_rep$alternate_ids,1:3]
rep_snps
```

```
##     alternate_ids       rsid position
## 785        SNP827  rs5751704 22331975
## 787        SNP829 rs12157657 22344229
## 790        SNP832 rs11090280 22358412
## 792        SNP834  rs7287369 22362015
## 793        SNP835 rs13057362 22362771
## 796        SNP838 rs17629956 22380471
```

All SNPs that share significance in both studies are found in the same region between 22 and 23 Mbp on the chromosome. Because the data is found to be quite similar between the two sets, the META software package can be used to combine the results from the two SNPTEST2 results into one meta-set:

```
./meta_v1.7 --cohort snptest2.txt snptest3.txt --snptest --method 1 --output meta.txt
```

- `--cohort snptest2.txt snptest3.txt`: Instructs META which files to combine
- `--snptest`: Informs META that the files are SNPTEST2 outputs
- `--method 1`: Sets the meta-analysis method, "1" indicating a fixed-effects inverse-variance method
- `--output meta.txt`: Writes the results to a file called "meta.txt"

An excerpt of the results are shown below.

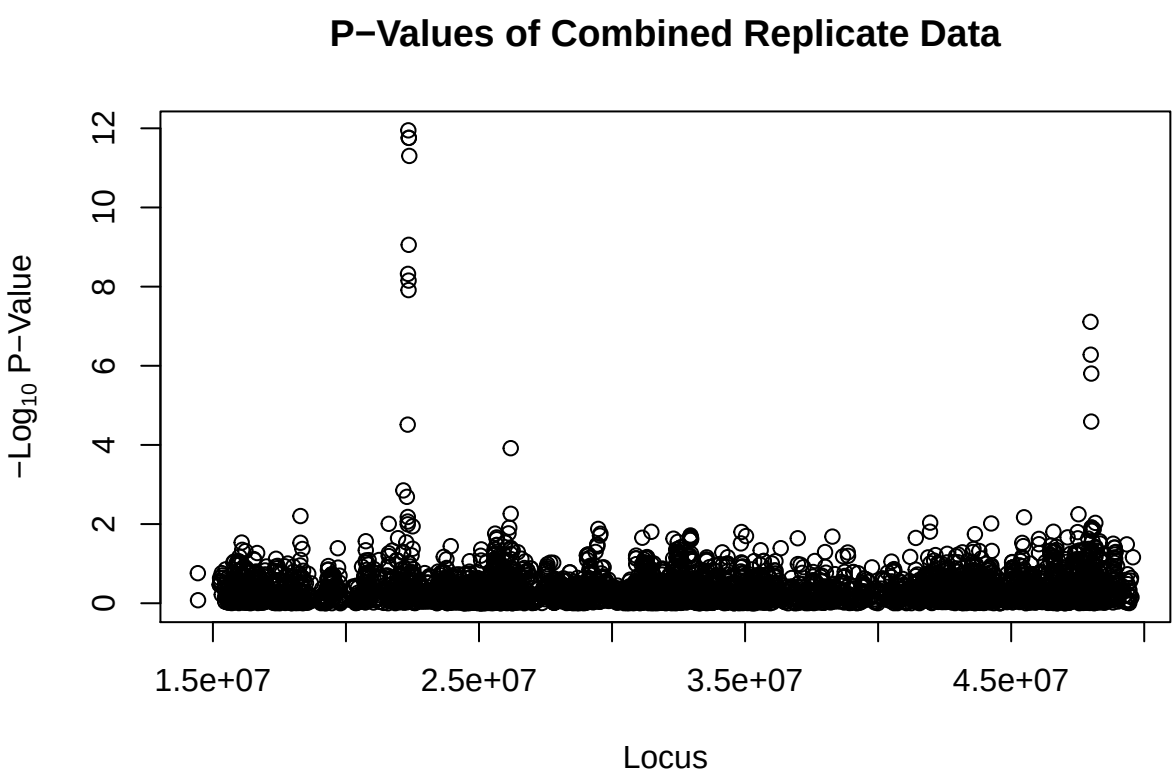
**META Results**

```
metaset <- read.table("meta.txt", header=T, as.is=T)
metaset[c(1:5),c(2,3,6)]
```

```
##        rsid      pos  P_value
## 1  rs915677 14433758 0.173742
## 2 rs9617528 14441016 0.839328
## 3  rs140378 15257135 0.337879
## 4  rs131564 15258423 0.343520
## 5 rs5748616 15268900 0.231078
```

Although META does not use the Bayes factors from the SNPTEST2 results, the meta-analysis p-values can be plotted for each SNP.

```
plot(metaset$pos, -log10(metaset$P_value), main="P-Values of Combined Replicate Data",
     xlab="Locus", ylab=expression("-Log"[10]*" P-Value"))
```




Once again, the same areas of significance are represented in the plot. From the evidence presented thus far, it is clear that the two sets show replicated data, and thus increase confidence in the findings.

---

## Part II: Phasing and Imputation

Imputation will be used to add previously sequenced haplotype information to this dataset. To begin this process, the original study data must be phased to generate haplotype information about the data.

Because the information about whether an individual is case or control is essentially stored within the "pheno1" column, the case and control sets can be combined into a single dataset using QCTOOL without losing information:

```
./qctool -g cases2.gen -s cases2.sample -g controls2.gen -s controls2.sample \-og cohort1.gen
-os cohort1.sample -omit-chromosome
```

- `-g cases2.gen`: Sets a .gen file to be combined
- `-s cases2.sample`: Sets a .sample file to be combined
- `-og cohort1.gen`: Sets the output .gen file
- `-os cohort1.sample`: Sets the output .sample file
- `-omit-chromosome`: Instructs the tool to not output a chromosome column in the resultant data

The output cohort1.sample and cohort1.gen files are much the same as the originals, but contain both the case and control groups.

**Cohort Individual Data**

```
cohort1_sam <- read.table("cohort1.sample")
colnames(cohort1_sam) <- c()
cohort1_sam[c(1:7,972:977),]
```

```
##
## 1    id_1  id_2 missing heterozygosity pheno1    pheno2       cov1 cov2
## 2       0     0       0              C      B         P          C    D
## 3   A1001 B1001       0       0.279731      1 -0.410022     2.2682    0
## 4   A1002 B1002       0       0.307783      1   1.17131    1.83839    0
## 5   A1003 B1003       0       0.273805      1  -2.32686    1.45202    1
## 6   A1004 B1004       0       0.337218      1   2.08926    1.17157    0
## 7   A1005 B1005       0       0.303042      1  0.484458 -0.0489582    0
## 972    A1    B1       0       0.321217      0  -1.02693    1.37905    1
## 973    A2    B2       0       0.304425      0  -0.35523  0.0469157    0
## 974    A3    B3       0       0.319834      0  -1.14851   0.650471    0
## 975    A4    B4       0       0.289214      0  0.450384   -2.64719    1
## 976    A5    B5       0       0.300277      0 0.0569129 -0.0911498    1
## 977    A6    B6       0       0.276571      0   1.00102   0.494835    1
```

**Cohort SNP Data**

```
cohort1_gen <- read.table("cohort1.gen")
colnames(cohort1_gen) <- c()
cohort1_gen[c(1:7),c(1:11)]
```

```
##
## 1 SNP1  rs915677 14433758 A G 0 0 1 0 0 1
## 2 SNP2 rs9617528 14441016 C T 0 0 1 1 0 0
## 3 SNP3  rs140378 15257135 C G 1 0 0 1 0 0
## 4 SNP4  rs131564 15258423 C G 0 0 1 0 1 0
## 5 SNP5 rs5748616 15268900 C G 1 0 0 0 1 0
## 6 SNP6 rs4010554 15274264 A C 0 0 1 0 1 0
## 7 SNP7 rs4010550 15280134 A G 1 0 0 0 1 0
```

The combined genotyping information from the new set will now provide additional sequence data for the phasing tool to work with.

---

**Phasing**

Haplotype information will now be estimated by phasing using the SHAPEIT2 tool. Because significant SNPs were found in the 22-23 Mbp region, only this region will be phased to speed up analysis.

```
./shapeit_v2 --input-gen cohort1.gen cohort1.sample --input-map genetic_map_chr22_combined_b36.txt
\   --output-max cohort1.21.5-23.5Mb.haps cohort1.haps.sample --thread 4 \   --input-from
21500000 --input-to 23500000
```

- `--input-gen`: Specifies that the files are in .gen/.sample format, and which files to use
- `--input-map`: Specifies the genetic map of SNP locations and recombination values to use, which is provided *a priori*
- `--output-max`: Writes the haplotype information to the specified file
- `--thread`: Takes advantage of threading on multi-core processors to speed calculation time
- `--input-from`: Sets the lower bound of the base-pair region to phase
- `--input-to`: Sets the upper bound of the base-pair region to phase

Note that although the desired region is 22-23 Mbp, the phasing region is extended past these boundaries by 0.5 Mbp in each direction as a buffer, since phasing accuracy drops off as the tool processes each end.

While the output cohort1.haps.sample file contains the same information as the input cohort1.sample file, the cohort1.21.5-23.5Mb.haps file now contains the estimated haplotype information for the given region, which includes a total of 245 SNPs. For each SNP, the file shows:

- SNP ID
- RSID
- Position
- Allele A
- Allele B
- A sequence of paired bits regarding haplotype information, with one pair for each individual's genotype at that SNP
  - 0 = allele A, 1 = allele B
  - The first digit refers to the allele on the first haplotype
  - The second digit refers to the allele on the second haplotype

The excerpt shows the haplotypes of the first 5 SNPs for the first 3 individuals. For example, individual 1's estimated haplotypes are C-C-C-T-A and C-C-T-T-A for the first 5 SNPs in the region.

**Cohort Haplotypes**

```
haps <- read.table("cohort1.21.5-23.5Mb.haps")
summary_haps <- haps[c(1:5),c(1:11)]
colnames(summary_haps) <- c()
summary_haps
```

```
##
## 1 SNP710  rs383391 21552208 C T 0 0 0 0 0 0
## 2 SNP711 rs6003387 21581602 C G 0 0 0 1 0 0
## 3 SNP712 rs2854097 21585556 C T 0 1 1 0 1 1
## 4 SNP713  rs390407 21588231 C T 1 1 1 1 1 1
## 5 SNP714  rs451115 21588369 A G 0 0 0 0 0 1
```

---

**Imputation**

In the final step, the phased data can now be imputed with additional SNP information using the IMPUTE2 tool:

```
./impute_v2.3.2 -use_prephased_g -known_haps_g cohort1.21.5-23.5Mb.haps \ -m genetic_map_chr22_combined_
-h genotypes_chr22_CEU_r22_nr.b36_fwd_phased_by_snp \ -l genotypes_chr22_CEU_r22_nr.b36_fwd_legend_by_sn
-int 22000000 23000000 \ -Ne 15000 -buffer 500 -o cohort1.22-23Mb.imputed.gen
```

- `-use_prephased_g`: Specifies that the input files have been prephased
- `-known_haps_g`: Specifies the pre-phased file
- `-m`: As in the phasing step, the map file of SNP locations and recombination rates
- `-h`: Reference file of known haplotypes from the IMPUTE2 site
- `-l`: Another reference file with legend information for the known haplotype file
- `-int`: The interval to impute
- `-Ne`: A parameter that adjusts linkage equilibrium patterns that IMPUTE2 uses.
- `-buffer`: The length of the buffer on each side of the imputed region
- `-o`: Writes the data to the specified file

IMPUTE2 produces an imputed .gen file containing additional estimated SNP information for the haplotypes provided. From approximately 250 SNPs, the imputed set now contains over 750, more than tripling the

data in the study. SNPs added by imputation are noted in the table by their lack of SNP ID and estimated haplotype values.

### Imputed Haplotypes

```
imp <- read.table("cohort1.22-23Mb.imputed.gen")
colnames(imp) <- c()
imp[c(10:15),c(1:9)]
```

```
##
## 10    ---  rs7510999 22020560 C T 0.998 0.002 0.000 0.998
## 11    ---  rs6003640 22021471 C T 0.006 0.989 0.004 0.018
## 12 SNP777  rs8139654 22022738 C T 1.000 0.000 0.000 1.000
## 13    ---  rs9624127 22023618 A T 0.998 0.002 0.000 0.998
## 14 SNP778 rs12165974 22023911 A G 1.000 0.000 0.000 1.000
## 15    ---  rs6003643 22028819 G T 0.000 0.032 0.967 0.053
```

Finally, the imputed data can be run through SNPTEST2 for higher-powered association testing with the additional SNP information:

```
./snptest_v2.5.2 -data cohort1.22-23Mb.imputed.gen cohort1.sample \ -frequentist 1 -bayesian
1 -method score -pheno pheno1 -o snptest4.txt
```

### Significant Imputed Association Test Results

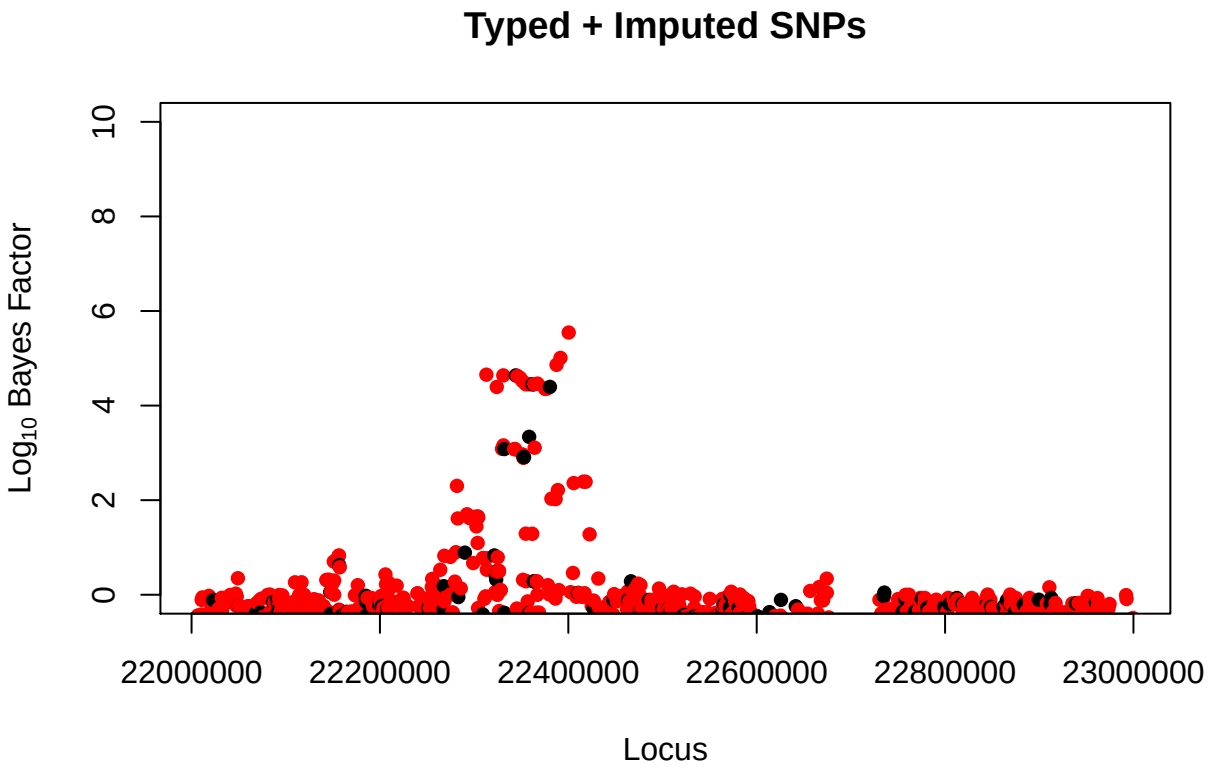
```
imp_test <- read.table("snptest4.txt", header=T, as.is=T)
sig_imp_test <- imp_test[which(imp_test[,46]>3), c(1,2,4,46)]
colnames(sig_imp_test) <- c("Name", "RSID", "Position", "Log10 Bayes")
sig_imp_test <- sig_imp_test[order(-sig_imp_test$`Log10 Bayes`),]
sig_imp_test
```

```
##       Name       RSID Position Log10 Bayes
## 378    ---   rs738787 22400416     5.54452
## 377    ---  rs2330580 22391698     5.00813
## 374    ---  rs8143019 22387485     4.86210
## 294    ---  rs2330575 22312999     4.65403
## 317    --- rs12166768 22330993     4.63855
## 325 SNP829 rs12157657 22344229     4.63805
## 327    ---  rs1109403 22346138     4.62398
## 329    --- rs12169702 22348850     4.58386
## 332    ---   rs762263 22351462     4.51648
## 362    --- rs12170464 22367306     4.46457
## 356    --- rs17629631 22363869     4.45285
## 352 SNP834  rs7287369 22362015     4.45090
## 353 SNP835 rs13057362 22362771     4.45090
## 342    ---  rs2051194 22355007     4.44940
## 346    ---  rs9624291 22358351     4.44940
## 369 SNP838 rs17629956 22380471     4.39570
## 302    ---  rs6003828 22324026     4.39353
## 366    ---  rs6003847 22377657     4.36128
## 365    ---  rs7286137 22375645     4.35683
## 364    ---  rs7291110 22375321     4.35613
## 347 SNP832 rs11090280 22358412     3.33909
## 318    ---  rs6519465 22331094     3.15784
## 357    ---   rs738786 22364288     3.10915
## 316    ---  rs7288276 22329627     3.08535
## 320 SNP827  rs5751704 22331975     3.08084
```

```
## 322    ---  rs5759950 22342624     3.08014
## 323    ---   rs762262 22343178     3.08014
## 324    ---  rs2330555 22343288     3.08014
```

From SNPTEST2, the imputated dataset contains 28 significant SNPs, up from 8 and 9 in the original and second data sets. These can be visualized on the following graph, which shows the imputed SNPs in red, and the original SNPs in black:

```
plot(imp_test$pos, imp_test$bayesian_add_log10_bf, ylim=c(0, 10), main="Typed + Imputed SNPs",
     col=1+1*(imp_test[,1]=="---"), pch=16, xlab="Locus",
     ylab=expression("Log"[10]*" Bayes Factor"))
```




---

Ultimately, the process of imputation has greatly increased the density of significant SNPs in the region implicated in the original study, which will allow for a much more in-depth analysis of the region and provides much more confidence that the area in question is indeed associated with the case phenotype.