

# Documenation for `utility.multiSNP.R`

Matthew Stephens

July 30, 2007

The file `utility.multiSNP.R` contains R code to plot a summary of BIMBAM output when computing multi-SNP BFs. The problem this code attempts to solve is that when performing a multi-SNP analysis there may be large numbers of different combinations of SNPs that could be associated with the phenotype, and sifting through and interpreting these results can be challenging. The approach we take is to

- Summarise each SNP by its (marginal) posterior probability of affecting phenotype.
- Group SNPs together into clusters of SNPs that tend not to occur together in the model. (this can be thought of as analagous to grouping SNPs into “bins” of SNPs that are good proxies for one another (i.e. highly correlated), since when two SNPs are highly correlated parsimony tends to favour not including both in the model.
- Output, for each cluster, the probability that at least one SNP from that cluster affects phenotype.

**Note that this code is all based on the assumption that there *is at least one SNP in the region affecting phenotype*.** That is, it is attempting to *explain* observed associations, rather than test for the presence of associations. The overall BF for the region (in the `summary` file), and the individual SNP BFs, should be used to check that there is actually good evidence for something interesting in the region before over-interpreting the results of this analysis!

The code contains two functions:

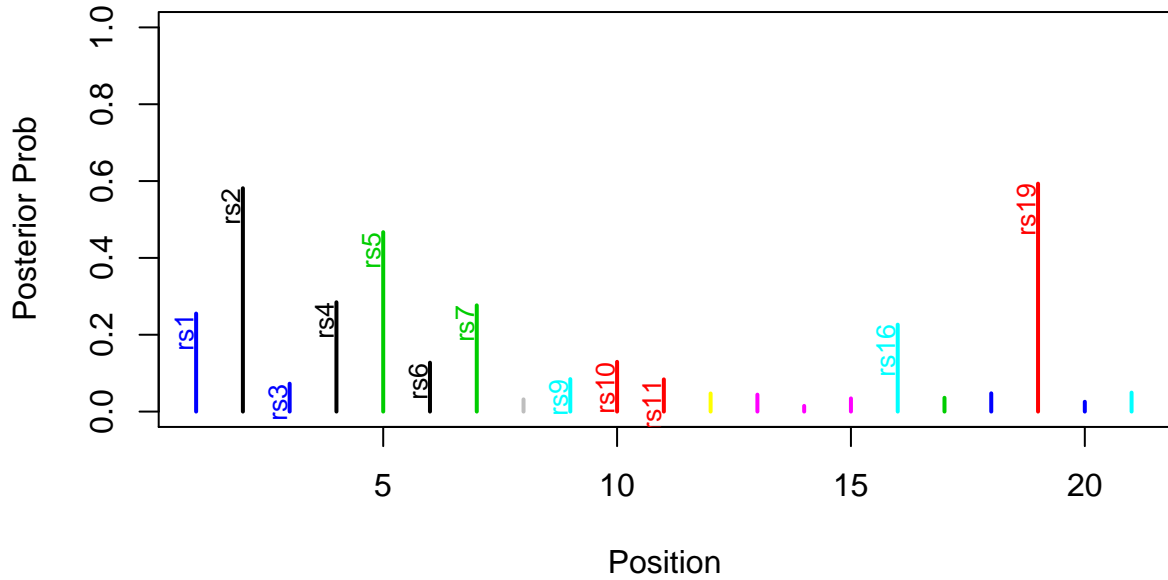
- `cluster.multiSNP` takes as input the `.multi` and a `.single` file, produced by BIMBAM , and performs some calculations to produce an object.
- `plot.multiSNP` plots this object.

Here’s an example of using the code:

```
source("utility.multiSNP.R")
mSNP = cluster.multiSNP("output/outputpref.multi.txt","output/outputpref.single.txt")
pdf("plotfile.pdf",width=6.5,height=8)
plot.multiSNP(mSNP)
dev.off()
```

And here’s an example plot (see caption for interpretation):

### Marginal Probability of Each SNP



### Probability for each cluster of SNPs

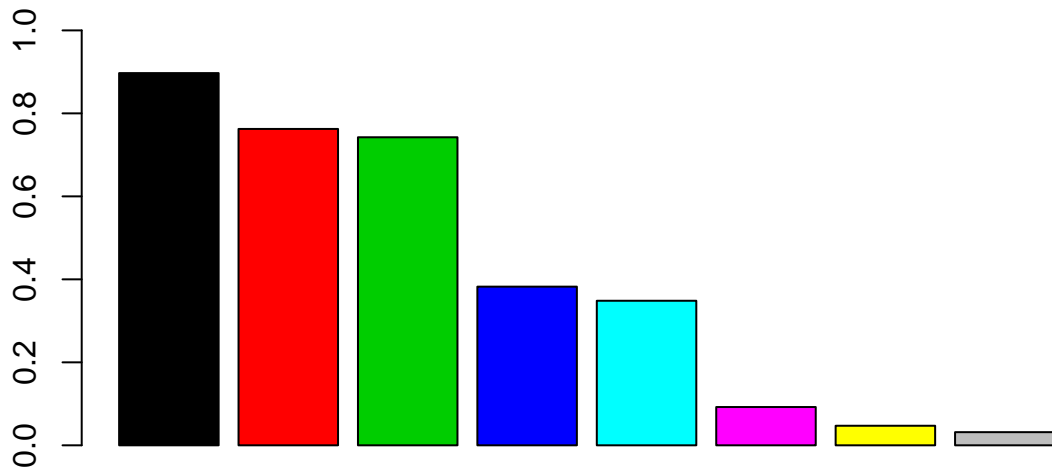


Figure 1: Top: the marginal posterior probability for each SNP, with colors indicating the cluster membership of each SNP; Bottom: the probability, for each cluster, that at least one SNP in that cluster affects phenotype. In this example, there is a good chance that there is at least one black SNP in the model (most likely rs2, rs4, or rs6, with rs2 being the most likely, since it has the largest posterior probability); at least one red SNP (rs10, rs11 or rs19, with rs19 being the most likely), and at least one green SNP (rs5 or rs7). There are a number of other SNPs with non-negligible probability of affecting phenotype. This is a relatively complex example, where it is clear that there may be many SNPs affecting phenotype, but with considerable uncertainty about which ones are responsible.