Practical on haplotype inference

All input files are available at

http://www.stat.washington.edu/stephens/sisg/

You will need to save the input files example1.txt, example2.txt and example3.txt to your local computer in the same directory as PHASE lives.

Then open an MSDOS window, and change to the directory where PHASE lives.

Example 1

The input file example1.txt contains an input file for the PHASE software. Here is what it looks like, with some comments

```
3 (this is the number of individuals)
5 (number of SNPs)
P 100 200 300 400 500 (positions of SNPs)
SSSSS (types of loci: all 5 are SNPs)
#1   (individual label - this is data for individual 1)
11111   (5 columns, 2 rows, each column is 1 genotype)
11111   (this case is homozygous for 1 allele at all SNPs)
#2
00000
00000
#3
00000  (this one is heterozygous at all SNPs)
11111
```

Before running PHASE on this file, here are some things to think about:

- How many of the individuals have ambiguous haplotypes?
- What would you guess for the haplotypes of the ambiguous individual(s)?
- How confident would you be? Very confident? Less confident? (What might make you more confident?)

- Try running PHASE on this file using

  ```
  PHASE example1.txt example1.out
  ```

  Take a look at the output files example1.out, and example1.out_pairs, and see if PHASE's answers correspond to your intuition.

  You might also like to look at the output file example1.out_freqs which contains estimates of the population haplotype frequencies.

- Try running PHASE again:

  ```
  PHASE example1.txt example1.2.out
  ```

  Do you get the same answers? You should, because by default PHASE uses the same random numbers to simulate the Markov chain.

Changing the Seed

To change the random numbers used, you need to use the -S option to set the ”seed”:

PHASE -S332554 example1.txt example1.3.out

The answers should look similar, but not identical.

Performing multiple runs with different seeds is a helpful way to check that you are running the algorithm long enough to get reliable results.

Example 2

The file example2.txt contains another small input file. This one was created by putting together individuals by randomly pairing haplotypes taken from a pool containing equal numbers of the 8 possible 3-SNP haplotypes.

- What would you expect to happen for this input file?
- Try running PHASE a few times, with different seeds, and compare the results.
- As well as estimating the haplotypes for each individual (eg in the _pairs file) PHASE also estimates population haplotype frequencies: see the _freqs file. Compare these estimates with the description of how the input file was created.
- One way to estimate haplotype frequencies is to first estimate the haplotypes for each individual and then to count up how many times each haplotype occurs in these estimates. But this is the

kind of 2-stage procedure that should be avoided. Can you think of another way?

- By default PHASE estimates recombination rates, and uses conditional distributions based on these rates within a model known as the coalescent. One can also specify, among other things, that PHASE is to use a Dirichlet prior for the haplotype frequencies (use the -ME option), or to assume a coalescent prior with no recombination (-MS). The main reason one might want to use these options instead of the default is that for big problems they can run appreciably quicker. Try running PHASE with these options on this data set, and compare the results (eg the _freqs file) with the results from the default.

Example 3

In addition to estimating haplotypes, PHASE can also be used to estimate recombination rates in a region, and to assess whether the region contains a recombination ”hotspot”. By invoking the -MR1 1 option, PHASE assumes that the recombination rate in the region is constant, with the possible exception of a single recombination hotspot somewhere in the gene. PHASE makes various assumptions about the location, width and intensity of the hotspot by making assumptions about the prior distributions of these quantities. It then outputs a _hotspot file, which contains a sample from the posterior distribution for these quantities.

The input file example3.txt contains data from a gene CD36. Run PHASE using the -MR1 1 option:

./PHASE -MR1 1 example3.txt example3.out

and examine the contents of the file example3.out_hotspot.

The following code can be used to read the _hotspot file into R and plot summaries. (Or you could use Excel if you prefer.) Try performing multiple runs of PHASE and comparing results (eg the posterior distribution of the hotspot intensity) for different seeds.

```
h = read.table("example3.out_hotspot")
hist(h[,4]) # h[,4] contains samples from the posterior dist
plot(h[,4])
mean(h[,4]>1) # returns the proportion of samples where the
              # (note that intensity 1 corresponds to "no ho
hist(h[,3]-h[,2]) # h[,3] and h[,2] are samples from the lim
plot(h[,3]-h[,2])
hist(h[,1]) # h[,1] is an estimate of the recombination para
sum(h[,4]>1)
```