Choosing htSNPs

David Clayton

September 17, 2003

1. Datasets for 34 loci on 32, 48 or 96 subjects are in `seqxx.txt` (where `xx` is 32, 48 or 96). Locus names are in file "loci". To estimate haplotype distributions, we currently must use a stand-alone program — `snphap`. As an example, for the 48 subject dataset:

   ```
   snphap -nf loci -ss -mm 1000 seq48.txt
   ```

   This estimates haplotype frequencies using the EM algorithm, repeating the iteration 1000 times (`-mm 1000` option) from random starting points, and writing the best solution to results to a file (named, by default, `snphap.out`) in a "spreadsheet" format (`-ss` option) with variable names taken from the file `loci` (`-nf loci` option).

2. Read `snphap.out` into Stata (`.` denotes the Stata prompt — you don't type this. Cammands should be entered on a single line)

   ```
   . insheet using snphap.out
   ```

   You might like to browse the file. Each line gives a haplotype and its estimated probability.

3. Search for best set of SNPs using a crude (but fast) "step-up" search:

   ```
   . htstep mh15-jo22 [pw=probability], up cri(r2 min)
        until(0.8) ra(0.05)
   ```

   At each stage the programs adds the htSNP so as to maximize the *minimum* value of the R^2 which measures the proportion of variance of each remaining SNP "explained" by grouping on full htSNP haplotype (`r2 min` option). It stops when this index excedes 0.8 (`until` option). Loci with minor allele frequency less than 0.05 are ignored (`ra` option). Note that the expression in square brackets denotes the *probability weights*.

   In the current example this selects 8 htSNPs:

   ```
   mh15 mh14 ct52 ct41 ct44 jo27_1 jo26_2 jo23
   ```

4. Repeat this using a "step-down" search

```
. htstep mh15-jo22 [pw=probability], down cri(r2 min)
     until(0.8) ra(0.05)
```

   This selects 7 htSNPs:

```
mh14 mh2 ct41 ct44 jo27_1 jo26_2 jo23
```

5. The "consensus" htSNPs, selected by both methods are:

```
mh14 ct41 ct44 jo27_1 jo26_2 jo23
```

   We can now try an exhaustive subset search to find the smallest subset that we might add to this consensus set:

```
. htsearch mh15-jo22 [pw=probability],
    include(mh14 ct41 ct44 jo27_1 jo26_2 jo23)
    until(r2 min 0.8) ra(0.05)
```

   In this case this yields the same selection of 7 htSNPs as the step-down procedure :

```
mh14 ct41 ct44 jo27_1 jo26_2 jo23 mh2
```

6. A longer summary of the performance of these htSNPs can be obtained by

```
. haptag mh15-jo22,
     htsnps(mh14 ct41 ct44 jo27_1 jo26_2 jo23  mh2)
```

7. You might like to repeat this for the "Allelic" R^2 criterion in which each remaining SNP is predicted by regression on the htSNP alleles rather that their haplotypes (`r2a` criterion). Also try the different datasets to see how much agreement there is on the best set.You might also like to use `haptag` to see how well the htSNP set chosen with one dataset performs in another.