Documentation for *STRAT* software

Jonathan K. Pritchard

Department of Statistics
University of Oxford
1 South Parks Rd
Oxford, OX1-3TG, UK

`www.stats.ox.ac.uk/~pritch/home.html`

May 26, 2000

Contents

1 Introduction

It is well known that case-control studies are liable to suffer from high rates of false positives in the presence of population structure. However, in recent papers we have described methods for (1) testing whether population structure is a potential problem for a particular case-control sample (Pritchard and Rosenberg, 1999), and (2) performing statistically valid tests of association in the presence of population structure (Pritchard *et al.*, 2000b). Both papers make use of the idea that it is possible to learn about population structure by looking at a set of unlinked marker loci.

The program *STRAT* implements the method of Pritchard *et al.* (2000b) for performing tests of association in structured populations[1]. Briefly, it is assumed that there is a sample of unrelated case and control individuals from one or more subpopulations, possibly with admixture. Each individual is genotyped at a series of unlinked marker loci (eg SNPs or microsatellites). The goal is to perform statistically valid tests of association between the alleles at these marker loci, and the case-control phenotypes. We proceed as follows.

1. Apply the test described by Pritchard and Rosenberg (1999) to determine whether population structure is a problem for this particular case-control sample. If there is evidence for population structure, proceed to (2). This test is implemented in *STRAT* (but is easy to perform by other means too).

2. Apply the program *structure*, described by Pritchard *et al.* (2000a), to learn about the structure of the population that the sample has been drawn from, and to estimate the ancestry of the individuals in the sample. Proceed to (3).

3. Use *STRAT* to test for association at each locus, conditional on the ancestry of the individuals in the sample (as estimated by *structure*).

Background on these methods, and the underlying assumptions, are provided in the papers listed above. Questions, comments, and bug reports should be directed to me, at pritch@stats.ox.ac.uk.

[1] The programs *STRAT* and *structure* are available from `www.stats.ox.ac.uk/~pritch/home.html`.

2 Getting started

The first thing is to download both *structure* and *STRAT*, and the accompanying documentation. The *structure* documentation provides details on how to prepare the input data file[2]. Platform specific information about running the programs is provided on the web page.

3 Running the program

3.1 Assuming no population structure

You can run the program assuming no population structure as follows. Set the parameters describing the input file in *mainparams*, and set MAXPOPS=1. Then *STRAT* will produce Chi-square values, and simulated p-values under the assumption of no population structure. The program outputs the value of the test statistic suggested by Pritchard and Rosenberg (1999) for testing whether the case and control samples are mismatched (and hence population structure is a concern).

3.2 Correcting for population structure

In order to account for population structure, you first need to run *structure*, as described in the documentation for that program. I discuss the issue of how to choose MAXPOPS, below.

When running *structure*, set PRINTQHAT=1, in the file *extraparams*. This will produce an output file with the name OUTFILE_q, where OUTFILE represents the name of the output file set in *mainparams*.

Once you have run *structure* to create the output file of estimated ancestries, you can go on to run *STRAT*. *STRAT* will also look into the file *mainparams* to find the value of MAXPOPS, the name and format of the input file, and the name of the file of estimated ancestries (OUTFILE_q).

4 Parameters in the file STRATparams

STRAT reads some parameter values from *mainparams* (ie describing and naming the input data files, and setting MAXPOPS). These values will already have been set in order to run *structure*. You can also set several parameters in the file *STRATparams*. The default values of all of these

[2] The phenotypes in the input file are represented using the integers {0,1,...}–eg cases and controls would be represented by 0 or 1.

are probably fine at first (except NUMPHENS if you have more than two phenotypes).

NUMSIMSTATS (int) Number of simulated test statistics per locus. The *p*-value of the observed test statistic is estimated as the fraction of simulated test statistics that exceed the observed value, so obviously larger values of NUMSIMSTATS lead to more accurate estimates of the *p*-values, but slow down the program.

NUMPHENS (int) Number of different phenotypes in the data file.

POOLFREQ (int) Eliminate rare alleles that have fewer than POOLFREQ copies in the entire sample. The pooling is done as follows. Any alleles with fewer than POOLFREQ copies are pooled together. If the pooled group has fewer than POOLFREQ total copies, they are then pooled with the allele with the next fewest copies. POOLFREQ=0 prevents any pooling. This option is primarily here in order to make the Chi-square tests valid (when $K = 1$).

LOCUSxONLY (int) Run the test of association on one locus only, rather than on all the loci in the sample. This might be used to obtain more accurate *p*-values (with large values of NUMSIMSTATS) for loci that appear to be significant. All loci are tested for association if this is set to 0.

EMERROR (double) The allele frequencies are estimated using an EM algorithm. This algorithm is run for a number of steps until the change in allele frequencies in successive steps is very small, indicating that it is near convergence. The stopping rule used here is that if all the allele frequency estimates change by less than EMERROR between successive steps, the algorithm is considered to have converged. A value close to 0 is ideal here, but this part of the program is fairly time-consuming.

5 Program Output

The program outputs the following kinds of data. It prints lines in the following format to the screen and to OUTFILE_P:

```
63:  chisq= 5.796 1 df; TS = 0.41, p = 6.84000e-01
```

This line gives the locus-number in the input file (63), the value of the chi-square test of association assuming no population structure (5.796), the number of degrees of freedom (1), the value of the STRAT test-statistic (0.41), and the STRAT p-value (6.84000e-01). The p-value is represented in scientific notation: eg, the p-value here is 6.84×10^{-1}. Stars are printed at the end of the line for small p-values.

I have not programmed anything to compute p-values from chi-square test statistics, since these are easy to obtain either from tables (Rohlf and Sokal, 1995), or from a standard statistical package such as SPlus.

The program also prints a summary of the empirical distribution of all the p-values computed so far. Here are the top three lines (out of ten) from one example:

```
0.00---0.05:  0.013 0.010 0.010 0.007 0.007   0.0467 0.0467 -0.0033
0.05---0.10:  0.013 0.000 0.017 0.017 0.003   0.0500 0.0967 -0.0033
0.10---0.15:  0.013 0.013 0.017 0.017 0.013   0.0733 0.1700  0.0200
```

The first two columns show the range of p-values on that line. The next five columns show the fraction of p-values in intervals of 0.01. For example, consider the second line here. This records what proportion of the p-values were in the range (0.05–0.10) in intervals of 0.01—for example 0.003 of the observations were in the range (0.09–0.10). The last three columns give (1) the row sum (ie, the fraction of observations falling in that interval of 0.05), (2) the cumulative sum (the fraction of observations up to and including that row), and (3) the difference between the cumulative sum, and the expected value. Large positive values in the last column indicate an excess of small p-values.

A summary of the estimated allele frequencies is printed into the file OUTFILE_fr. These are in the following format.

```
Locus 2:  estimated allele frequencies
Allele 0
0.479 (0.477 0.482)
0.115 (0.099 0.125)
Allele 1
0.521 (0.523 0.518)
0.885 (0.901 0.875)
```

This example shows the allele frequencies at a biallelic marker, assuming two populations and two phenotypes. Each row gives the frequencies in a different population. The first number in each row gives the overall frequency, and the numbers in parentheses give the frequencies among individuals of each phenotype. So for example, the overall frequency of allele 0 is 0.479 and 0.115 in populations 1 and 2 respectively. The estimated frequency of allele 0 is 0.477 and 0.482, among individuals with phenotypes 0 and 1, respectively, in population 1.

6 Model validation, and choosing MAXPOPS

A critical assumption of this method is that *structure* provides an adequate representation of the population structure, and estimates the ancestry of individuals (Q) sufficiently well. The examples in Pritchard *et al.* (2000b) are encouraging in this regard, showing that the estimation doesn't have to be perfect for the test to have acceptable properties.

It isn't easy to say in advance how many loci are necessary to obtain satisfactory estimates of Q, as this will vary among problems. I think that for realistic, non-trivial problems in humans, this will be perhaps 100-150 microsatellites (or a larger number of SNPs), and probably rather more in some cases. In general, the inference is more difficult in admixed populations than in discrete populations, and assigning individuals accurately is (obviously) more difficult when the populations have very similar allele frequencies. The number of individuals in the sample is also important because this impacts the accuracy of the allele frequency estimates. If there are individuals without phenotypic information these may be used in *structure* to improve the inference of population structure, and then removed from the data set before applying *STRAT*. Individuals from source populations will be particularly helpful when analyzing admixed populations. Prior population information can be used in *structure* to assist the assignments.

An important issue is how to choose MAXPOPS, the number of populations. In some cases there will be external information that can help here. When there is not, I suggest that MAXPOPS be chosen using two criteria. The first is that *structure* provides an approximate method for helping to choose the best number of populations. However, this method is only intended as a guide, and often over-estimates the number of populations. There is likely to be a cost in power to using a larger value of MAXPOPS than necessary.

I think that a more useful approach to choosing MAXPOPS (and to validating the use of *STRAT* in general) is to use the property that under the null hypothesis the p-values should be uniformly distributed between 0 and 1. Assuming that only a quite small proportion of markers are genuinely associated with disease loci, we should find that the empirical distribution of p-values is close to uniform. An excess of small p-values indicates that either the model doesn't adequately capture the population structure, or the estimation of Q isn't good enough. A Kolmogorov-Smirnov test can be used to test the fit (eg Rohlf and Sokal, 1995). I would be inclined to use the smallest value of MAXPOPS that produces an appropriate distribution of p-values.

6.1 Running *structure* on closely related populations

The program *structure* implements two models of allele frequencies. The default model assumes that the underlying allele frequencies in different populations are independent (and hence usually quite different). A second model assumes that the allele frequencies are all quite close to the mean frequency in the sample (set FREQSCORR=1 in *extraparams*). I have found that for human data sets in which the populations are from different continents, eg with admixture, the first model works well. However, when considering closely related populations (eg Chinese and Japanese, in one example), *structure* does not perform well under the default model, often grouping everybody into one population. Applying the model with FREQSCORR=1, however, produces accurate assignments. Also, in such cases, if there is little admixture, the model without admixture (NOADMIX=1) tends to produce better results.

References

Pritchard, J. K. and Rosenberg, N. A. (1999) Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Gen.*, **65**, 220–228.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000a) Inference of population structure using multilocus genotype data. Genetics (June, in Press). Available at `http://www.stats.ox.ac.uk/~pritch/papers/strucabs.html`.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. and Donnelly, P. (2000b) Association mapping in structured populations. AJHG (July, in Press).

Rohlf, F. J. and Sokal, R. R. (1995) *Statistical Tables.* New York: W. H. Freeman and Company, third edition.