

ADMIXTURE 1.02 Software Manual

David H. Alexander John Novembre Kenneth Lange

November 17, 2009

Contents

1 Quick start	1
2 Reference	3
2.1 Do I need to thin the marker set for linkage disequilibrium?	3
2.2 How many markers do I need to supply to ADMIXTURE?	4
2.3 How do I change the random seed?	4
2.4 Optimization Methods	4
2.5 Termination criterion	4
2.6 Acceleration methods	5
2.7 Bootstrapping	5
3 Citation and further information	6

1 Quick start

ADMIXTURE is a program for estimating ancestry in a model-based manner from large autosomal SNP genotype datasets, where the individuals are unrelated (for example, the individuals in a case-control association study). Support for other kinds of markers (microsatellites, etc.) is forthcoming.¹

ADMIXTURE’s input is binary PLINK (.bed), ordinary PLINK (.ped), or EIGENSTRAT (.geno) formatted files and its output is simple space-delimited files containing the parameter estimates.

To use ADMIXTURE, you need an input file and an idea of K , your belief of the number of ancestral populations. Note: you should also have the associated support files alongside your main input file, in the same directory. For example, if your primary input file is a “.bed” file, you should have the associated “.bim” (binary marker information file) and “.fam” (pedigree stub file) files in the same directory. If your primary input file is a “.ped” or “.geno” file, a corresponding PLINK style “.map” file should be in the same directory.

Here is how to run:

If you have an binary PED (.bed) formatted file in your current directory

```
% ls
hapmap.bed hapmap.bim hapmap.fam
```

¹Do not try to trick ADMIXTURE by converting microsatellites into pseudo-SNPs. Please wait for us to provide the correct solution.

and you believe that the individuals in the sample derive their ancestry from three ancestral populations then run admixture like this:

```
% admixture hapmap.bed 3
```

ADMIXTURE will start running. Hopefully it will finish soon, and it will then output some estimates:

```
% ls
hapmap.bed hapmap.bim hapmap.fam hapmap.3.Q hapmap.3.F
```

There is an output file for each parameter set: Q (the ancestry fractions), and F (the allele frequencies of the inferred ancestral populations). Note that the output filenames have ‘3’ in them. This indicates the number of populations (K) that was assumed for the analysis. This filename convention makes it easy to run analyses using different values of K in the same directory, without output files being overwritten.

If you have a PLINK .ped “12” coded file—generated by a command like

```
% plink --file hapmap --recode12 --out hapmap
```

—then you follow basically the same instructions:

```
% admixture hapmap.ped 3
```

Note that ADMIXTURE infers the format of your input file based on the file extension (.bed, .ped, or .geno).

You can run ADMIXTURE on a input file located in another directory, but its output will always be put in the current working directory. This is by design. For example:

```
% pwd
/home/daalexander/AdmixtureRuns
% ls
(no output—current directory is empty)
% ls ~/Data
hapmap.ped hapmap.map
% admixture ~/Data/hapmap.ped 3
(wait for it to run)
% ls
hapmap.3.Q hapmap.3.F
% ls ~/Data
hapmap.ped hapmap.map
```

If you also wanted standard errors, then instead of the last command you should have used

```
% admixture -B ~/Data/hapmap.ped 3
```

This will perform point estimation and then will also use a bootstrapping procedure to calculate the standard errors. Note that (point-estimation & bootstrapping) takes considerably longer than point-estimation alone, so you will have to be patient. Eventually it will finish, yielding point estimates and standard errors:

```
% ls  
hapmap.3.Q hapmap.3.F hapmap.3.Q_se
```

The “_se” file is in the same unadorned file format as the point estimates.

2 Reference

2.1 Do I need to thin the marker set for linkage disequilibrium?

We tend to believe this is a good idea, since our model does not explicitly take LD into consideration, and since enormous data sets take more time to analyze. It is impossible to “remove” all LD, especially in recently-admixed populations, which have a high degree of “admixture LD”. Two approaches to mitigating the effects of LD are to include markers that are separated from each other by a certain genetic distance, or to thin the markers according to the observed sample correlation coefficients. The easiest way is the latter, using the `--indep-pairwise` option of PLINK. For example, if we start with a file `rawData.bed`, we could use the following commands to prune according to a correlation threshold and store the pruned dataset in `prunedData.bed`:

```
% plink --bfile rawData --indep-pairwise 50 10 0.1
```

(output indicating number of SNPs targeted for inclusion/exclusion)

```
% plink --bfile rawData --extract plink.prune.in --make-bed --out prunedData
```

Specifically, the first command targets for removal each SNP that has an R^2 value of greater than 0.1 with any other SNP within a 50-SNP sliding window (advanced by 10 SNPs each time). The second command copies the remaining (untargetted) SNPs to `prunedData.bed`.

This approach is imperfect but seems to work well in practice. Please read our paper for more information.

2.2 How many markers do I need to supply to ADMIXTURE?

This depends on how genetically-differentiated your populations are, and on what you plan to do with the estimates. It has been noted elsewhere [1] that the number of markers needed to resolve populations in this kind of analysis is inversely proportional to the genetic distance (F_{ST}) between the populations.

It is also noted in that paper that more markers are needed to perform adequate GWAS correction than are needed to simply observe the population structure.

As a rule of thumb, we have found that 10,000 markers suffice to perform GWAS correction for continentally separated populations (for example, African, Asian, and European populations $F_{ST} > .05$) while more like 100,000 markers are necessary when the populations are within a continent (Europe, for instance, $F_{ST} < 0.01$).

2.3 How do I change the random seed?’

To change the random seed to 12345, for example, you can use:

```
% admixture -s 12345 myFile.ped 3
```

2.4 Optimization Methods

The default optimization method used by ADMIXTURE is a block relaxation algorithm. An alternative method, an EM algorithm (identical to that implemented by the program FRAPPE) is also available. To use this alternative algorithm, use the `-m` switch to choose the *method*:

```
% admixture -m EM ~/Data/hapmap3.ped 3
```

The convergence of the block relaxation algorithm is generally much faster, so there is no particular reason to use the EM algorithm.

2.5 Termination criterion

The default termination criterion is to stop when the log-likelihood increases by less than $\epsilon = 10^{-4}$ between iterations. A different termination criterion can be specified. A termination criterion is defined as either (1) a convergence criterion ϵ for the log-likelihood ($\epsilon < 1$), or (2) a maximum number N of iterations, $N \geq 1$. ϵ or N is given after the flag `-C` in order to specify a desired termination criterion. For example, to stop when the log-likelihood change between iterations falls below 0.1, one could use

```
% admixture -C 0.1 ~/Data/hapmap3.ped 3
```

Or, to stop the algorithm after 10 iterations one could use

```
% admixture -C 10 ~/Data/hapmap3.ped 3
```

Note that this “overloaded” use of `-C` flag makes it impossible to use an ϵ greater than one. This doesn’t seem a major limitation to us at present, since in general a small value is desired for ϵ .

The `-C` flag designates the *major termination criterion*, i.e. the criterion for stopping the point estimation algorithm. In our paper we point out that we use a looser convergence criterion for re-estimation within bootstrap resamples. This *minor termination criterion* is specified in the same manner, but using the `-c` flag (lowercase). The default value for this is 3 (i.e., 3 iterations).

2.6 Acceleration methods

By default, ADMIXTURE uses the quasi-Newton convergence acceleration method described in the paper, with $q = 3$ secant conditions. To use quasi-Newton acceleration with a different number of secant conditions—for example, 2—use the `-a` flag as follows:

```
% admixture -a qn2 ~/Data/hapmap3.ped 3
```

To use the “squared extrapolation” techniques of Roland and Varadhan (referenced in the paper), use `-a [sqs1 | sqs2 | sqs3]`, according to which of their three methods you want to use.

To turn off acceleration entirely, use `-a none`.

2.7 Bootstrapping

ADMIXTURE estimates parameter standard errors using bootstrapping. As described in “Quick Start”, the basic way to get bootstrap standard errors is to include the `-B` flag when you run ADMIXTURE:

```
% admixture -B myInput.geno 3
```

This uses the default of 200 bootstrap replicates. Perhaps you want more? Try

```
% admixture -B2000 myInput.geno 3
```

to use 2000 bootstrap replicates. Naturally, this will take longer to run. Don’t forget that you need to have a PLINK `.map` file in place in order to do bootstrapping.

3 Citation and further information

If you use ADMIXTURE, please cite our forthcoming paper (information coming soon—email Dave).

Contact Dave at dalexander@ucla.edu with questions.

References

- [1] N. Patterson, AL Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, 2006.