Manual for SPA Software

Wen-Yun Yang, John Novembre, Eleazar Eskin and Eran Halperin

Initial version: May 8, 2012
Revised version: Oct 15, 2012

Introduction

This software is an implementation of the SPatial Ancestral analysis (SPA) method introduced in the following paper

- A model-based approach for analysis of spatial structure in genetic data. Wen-Yun Yang, John Novembre, Eleazar Eskin and Eran Halperin. Nature Genetics. 2012

This software is also available at `http://genetics.cs.ucla.edu/spa`. For any suggestions and bug report, please send to Wen-Yun Yang (wenyun@ucla.edu)

SPA is a command line program written in C. The program is designed for analysis of large-scale SNP data. All commands involve typing `spa` at the command prompt followed by a number of options (starting with `--file-option` or `-option`).

Running SPA

SPA is a command-line program: clicking on an icon will get you nowhere. The correct way to run SPA should be in a command prompt or terminal window. Then type commands as described below

```
spa --file-option myfile -option value
```

The possible `--file-option` can be

- `--bfile myfile`, which expects three files in PLINK[1] BED file format: `myfile.bed`, `myfile.bim` and `myfile.tfam`

- `--pfile myfile`, which expects two files in PLINK PED file format: `myfile.ped` and `myfile.map`

[1] available at`http://pngu.mgh.harvard.edu/~purcell/plink/`

- `--tfile myfile`, which expects two files in PLINK TPED file format: `myfile.tped` and `myfile.tfam`
- `--gfile myfile`, which expects one file in genotype file format: `myfile`
- `--location-input myfile`, which expects one file containing individual geographical locations
- `--model-input myfile`, which expects one file containing slope function coefficients
- `--location-output myfile`, which expects output file name for individual geographical locations
- `--model-output myfile`, which expects output file name for slope function coefficients

The possible `-option` can be

- `-k dimension` : set the number of dimensions for geographical location (default 2)
- `-n generation` : set the number of generations for ancestral inference (default 1)
- `-e espilon` : set tolerance of termination criterion (default 0.01)
- `-r tradeoff` : set optimization epsilon tolerance (default $1e-6$). Larger value makes the program run faster but poor accuracy
- `-v verbose` : set the level of verbose (default 1)

File Format

A total of five file formats are supported by SPA. We will discuss their specific format as follows. Three of them basically contain the same genotype information in different format: PED fileset, TPED fileset and genotype file. Location file is used to contain location information for each individual, and model file is used to contain allele frequency slope function for each SNP.

BED fileset

The BED fileset used in SPA follows the same format as the original PLINK. The binary format of BED can significantly reduce the memory usage. The detail can be found in PLINK website.

PED fileset

The PED fileset used in SPA follows a similar format as the original PLINK. It includes two files: PED file ending with .ped and MAP file ending with .map.

The PED file is a white-space (space or tab) delimited file: the first six columns are mandatory:

```
Family ID
Individual ID
Paternal ID
Maternal ID
Sex (1=male; 2=female; other=unknown)
Phenotype
```

The IDs are alphanumeric: the combination of family and individual ID should uniquely identify a person. Genotypes (column 7 onwards) should also be white-space delimited: they can be any character (e.g. 1,2,3,4 or A,C,G,T or anything else) except 0 which is, by default, the missing genotype character. **All markers should be biallelic**. All SNPs must have two alleles specified. Either both alleles should be missing (i.e. 0) or neither. No header row should be given.

The MAP file is also a white-space (space or tab) delimited file with four columns. Each line of the MAP file describes a single marker.

```
chromosome (1-22, X, Y or 0 if unplaced)
rs\# or snp identifier
Genetic distance (morgans)
Base-pair position (bp units)
```

For example, PED and MAP files for 6 individuals with 2 SNPs (one row = one person) are shown in Figure 1

Transposed PED fileset

Another possible file-format is called a transposed PED fileset, containing two text files: one (.tped) containing SNP and genotype information where one row is a SNP; one (TFAM) containing individual and family information, where one row is an individual.

The first 4 columns of a TPED file are the same as a standard 4-column MAP file. Then all genotypes are listed for all individuals for each particular SNP on each line. The TFAM file is just the first six columns of a standard PED file. In other words, we have just taken the standard PED/MAP file format, but swapped all the genotype information between files, after rotating it 90 degrees. For example, the example PED/MAP fileset can be equivalently represented as TPED/TFAM fileset at in Figure 1.

```
<---- myfile.ped ---->
1 1 0 0 1  1  A A  G T
2 1 0 0 1  1  A C  T G
3 1 0 0 1  1  C C  G G
4 1 0 0 1  2  A C  T T
5 1 0 0 1  2  C C  G T
6 1 0 0 1  2  C C  T T
```

```
<--- myfile.map --->
1 snp1   0   5000650
1 snp2   0   5000830
```

```
<------------- myfile.tped ------------->
1 snp1 0 5000650 A A A C C C A C C C C C
1 snp2 0 5000830 G T G T G G T T G T T T
```

```
 <- myfile.tfam ->
1  1  0  0  1  1
2  1  0  0  1  1
3  1  0  0  1  1
4  1  0  0  1  2
5  1  0  0  1  2
6  1  0  0  1  2
```

```
<- myfile.geno ->
2  1
1  1
0  2
1  0
0  1
0  0
```

Figure 1: Examples for PED/MAP, TPED/TFAM and genotype file sets.

Genotype file

Genotype file contains a matrix with elements only 0, 1 and 2. One row is an individual and one column is a SNP. The elements 0, 1 and 2 indicates the number of minor alleles for each individual at each SNP. The missing value can be any other integer number, usually -1 is used as missing genotype. For example, the equivalent genotype file is given in Figure 1. The minor alleles for the first and second SNP are A and G, respectively.

However, as genotype file does not contain any information about individual and SNP, our software SPA will implicitly assume an individual identifier i i 0 0 0 for the i-th individual, and a SNP identifier 0 SNPj 0 0 1 2 for the j-th SNP. Thus, when one is dealing with the known location training or known model training, the location and model file should use the same individual and SNP identifiers for appropriate matching.

Location file

The input and output location files follow the same format: one line is an individual. The first six columns are

```
Family ID
Individual ID
Paternal ID
Maternal ID
Sex
Phenotype
```

The seventh column and onward contain the predicted location for each individual. An example of location file for 2 dimension geographical locations is given in Figure 2.

In the case of admixed individuals, each individual would have two locations. The corresponding location file would be one more location in the same format appended afterward the first location.

Model file

The input and output model files follow the same format: one line is a SNP. The first six columns are

```
chromosome (1-22, X, Y or 0 if unplaced)
rs\# or snp identifier
Genetic distance (morgans)
Base-pair position (bp units)
Minor allele
Major allele
```

Then the seventh column and onward are slope function coefficients. The number of coefficient columns depend on the number of dimensions. If the expected dimension of geographical location is K, then the number of coefficient columns are $K+1$: K columns for coefficient

```
<---- myfile.loc ----->
1 1 0 0 1 0 0.23424   -0.43045
2 1 1 3 1 0 0.92378   -0.23442
3 1 0 0 0 1 1.20334   0.23234
4 1 0 0 1 0 -0.23435  -0.95965
5 1 0 0 0 1 -0.23675  0.43485
6 1 4 5 0 0 -1.03294  0.30438

<--------------------- myfile.model ---------------------->
1 snp1   0   5000650  C  A  0.93454 0.21351 0.02342 1.09772
1 snp2   0   5000830  T  G  0.23432 0.34849 0.95853 0.92843
```

Figure 2: Example for location and model files

a, 1 column for coefficient b. The last column is SPA score for selection signal detection. An example of model file for 2 dimension geographical locations is given in Figure 2.

Typical Usage

SPA is designed to perform a couple of functions in one program. Each function can be realized by different combination of parameters. Below we will introduce some typical usages of SPA.

Unsupervised Mapping

SPA would be able to perform unsupervised learning, where the only input is a genotype file (or PED/MAP or TPED/TFAM fileset). Then SPA is able to discover the spatial structure and slope function for allele frequency. SPA follows an alternative maximization procedure to maximize the log-likelihood. To perform this kind of analysis, the typical command would be

```
spa --gfile myfile.gen --location-output myfile.loc --model-output myfile.model
```

or

```
spa --pfile myfile --location-output myfile.loc --model-output myfile.model
```

or

```
spa --tfile myfile --location-output myfile.loc --model-output myfile.model
```

which correspond to genotype file, PED fileset and TPED fileset, respectively. Those three commands would be equivalent as long as the three files essentially contain the same genotype information. We would omit the last two commands for other part of this manual, but please keep in mind that they are basically equivalent but just in different format. The other two parameters indicate the output files to contain the analysis results. The file `myfile.lob` and `myfile.model` contains the geographical coordinates for individuals and slope function coefficients for SNPs, respectively.

By default, the dimension of spatial structure analysis is two, which is typical for samples from a local region within a country of a continent (e.g. Europe, Brazil). For samples from all over the world, the dimension should be three. Thus, the below command should be used

```
spa -k 3 --gfile myfile.gen --location-output myfile.loc --model-output myfile.model
```

where `-k 3` specifies the number of dimensions in the analysis.

Note that SPA does not support analysis with more than three dimensions, since we are focused on spatial structure.

Known geographical locations

In some cases, the geographical locations for individuals are known beforehand. Thus, we do not need to let SPA to infer those locations from scratch. SPA allows the input of geographical locations. The corresponding command would be

```
spa -gfile myfile.gen --location-input known.loc --model-output myfile.model
```

where the output file would be `myfile.model` containing all slope function coefficients. It is not necessary to arrange individuals in the same order between location file and genotype file, and it is not necessary to have the same number of individuals between them. SPA can figure it out automatically.

Ancestral inference

Another important application of SPA is the ancestral location inference. Given an individual's genotype and all slope functions of allele frequency, SPA is able to predict the geographical origin for a non-admixed individual, or two geographical origins for an admixed individual. We can use `-n generation` to specify how many origins we expect. By default, we have `-n 1`, which would predict one origin for an individual. An illustrative example is given below.

```
spa --gfile myfile.gen --model-input known.model --location-output myfile.loc
```

For prediction of two origins, we can use `-n 2` to specify that for SPA

```
spa -n 2 --gfile myfile.gen --model-input known.model --location-output myfile.loc
```

The output file `myfile.loc` would contain one or two origins for each individual in the above two cases, respectively.

A couple of precomputed model files are provided in our website, which can be used to infer individual geographical origins for any PLINK files you have. For example, we provide a file `europe.model` trained from POPRES data set. This model file can be used to infer locations for all individuals in any PLINK file you have. A typical command would be

```
spa --bfile myfile --model-input europe.model --location-output myfile.loc
```

It is not necessary to arrange the SNPs in the same order between `myfile.bed` and `europe.model`. SPA will figure it out automatically, as long as the rs numbers are correctly given.

Note that we have not implemented the prediction of two origins in three dimensional space. Thus, the parameter setting `-n 2 -k 3` would not be accepted yet.

Identifying loci with steep allele frequency gradients

As described in the file format for model file, the last column would be SPA score for each SNP. The SPA score would be available whenever the parameter `-model-output` is specified. Typically, the score can be obtained through both unsupervised scenario

```
spa --gfile myfile.gen --location-output myfile.loc --model-output myfile.model
```

and the scenario of known geographical locations

```
spa --gfile myfile.gen --location-input known.loc --model-output myfile.model
```

Higher SPA scores correspond to steeper inferred allele frequency gradients.

For 23andMe User

Since version 1.10, SPA software supports prediction of ancestral origin for 23andMe users. Given the genotype file generated from 23andMe, one can run the following command to get the prediction

```
spa --mfile 23andMeFile.txt --model-input europe.model --location-output 23andMe.loc
```

The result of the above program run will be two files: `23andMe.loc` and `23andMe.loc.html`. One can double click the HTML file `23andMe.loc.html`, which will redirect to google map to show the predicted ancestral origin. Note that the prediction is usually within 500 kilometers around the truth.