ELAI user manual

Yongtao Guan
Baylor College of Medicine

Version 1.01
6 April 2016
Bug fixes on 13 May 2021

Contents

1 Copyright

2 What ELAI Can Do

The software, Efficient Local Ancestry Inference (ELAI), is developed and maintained by Yongtao Guan (http://www.ncbi.nlm.nih.gov/pubmed/24388880). Please refer to the paper for the details of the statistical method. The ELAI is designed to perform local ancestry inference for admixed individuals. Comparing to existing methods to infer local ancestry, ELAI has the following advantages:

- Directly works with diploid data, no phasing required,
- No recombination map is required—the recombination rates are implicitly estimated,
- It has a high resolution and can detect local ancestry track length of a few tenth of a centi-Morgan (cM).
- The new version assign weights to training at cohort samples so that it applies to datasets of an arbitrary number of individuals (the previous version requires splitting a large sample into small subsets). The weighting scheme can be found at (http://www.ncbi.nlm.nih.gov/pubmed/26863142). Current implementation uses weights such that equivalently the training sample size (for each ancestry) is twice of the cohort sample size.

3 A simple example

Untar the downloaded file, one finds a subfolder called ./example, and executables elai-mac for Mac OS X and elai-lin for Linux. In the subfolder, those with suffix *.inp are input genotype files, the one with suffix *.pos is the position file, and the other two files *.truth and *.marker contain truth of the simulated admixture. A R-script, named r.2panel.R, is included to plot the example runs.

One may perform a testing run using the following command line:

```
./elai -g example/hap.ceu.chr22.inp -p 10 -g example/hap.yri.chr22.inp -p 11
-g example/admix-1cm.inp -p 1 -pos example/hgdp.chr22.pos
-s 20 -C 2 -c 10 -o test -mixgen 50 --exclude-nopos --exclude-miss1
```

This will generate four files with prefix 'test' in a newly created subfolder ./output. Using the R-script, r.2panel.R, one can plot the truth and the inferred local ancestries. Both the command line for test run and a simple instruction to run the R script is in note.txt.

4 Input file formats

The users should prepare genotype files for admixed cohort samples and training samples from source populations. All genotype files assume the BIMBAM format.

4.1 Genotype file format

Genotypes should be for bi-allelic SNPs, all on the same chromosome. The first two lines should each contain a single number. The number on the first line indicates the number of individuals; the number in the second line indicates the number of SNPs. Optionally, the third row can contain individual identifiers for each individual whose genotypes are included: this line should begin with the string IND, with subsequent strings indicating the identifier for each individual in turn. Subsequent rows contain the genotype data for each SNP, with one row per SNP. In each row the first column gives the SNPs "name" (which can be any string, but might typically be a rs-number), and subsequent columns give the genotypes for each individual in turn. Genotypes must be coded in ACGT while missing genotypes can be indicated by NN, ??, or 00 (zero). Example Genotype file, with 5 individuals and 4 SNPs:

```
5
4
IND, id1, id2, id3, id4, id5
rs1, AT, TT, ??, AT, AA
rs2, GG, CC, GG, CC, CG
rs3, CC, ??, ??, CG, GG
rs4, AC, CC, AA, AC, AA
```

Note that plink can convert genotype files from plink format to bimbam format. The option is `--recode-bimbam`.

4.2 Phased genotype file format

By default `ELAI` assumes that the genotypes in the Genotype file are *unphased*. If one has data where the phase information is known, or can be accurately estimated (e.g. from trio data, as in the HapMap data), then this can be specified by putting an "=" sign at the end of the first line, after the number of individuals. In this case, the order of the two alleles in each genotype becomes significant: the first allele of each genotype should correspond to the alleles along one haplotype, and the second allele of each genotype should correspond to the alleles along the other haplotype. For example, in the following input file, the haplotypes of the first individual are AGCA and TCCC:

```
5 =
4
IND, id1, id2, id3, id4, id5
```

```
rs1, AT, TT, ??, AT, AA
rs2, GC, CC, GG, CC, CG
rs3, CC, ??, ??, CG, GG
rs4, AC, CC, AA, AC, AA
```

Note: accidentally treating phased data as unphased has little harm except slower computation; accidentally treating unphased panel as phased is very harmful. Please make sure the genotypes are phased before you put "=" sign!

4.3 SNP position file format

The file contains three columns: the first column is the SNP ID, the second column is its physical location, and the third column contains its chromosome number (optional). It is okay if the rows are unordered, `ELAI` will sort the SNPs based on their position and chromosome number. `ELAI` can take multiple position files as input, and duplicate entries are acceptable. If the genotype files contain SNPs across different chromosome, `ELAI` will sort SNPs based on its chromosome and position. However, we recommend users to run `ELAI` chromosome by chromosome. The following describes an example file, where the delimit "," can be changed to " ".

```
rs1, 1200,  1
rs2, 4000,  1
rs3, 3320,  1
```

In certain applications, one may insist on running multiple chromosomes together, then the best practice is to stitch the position file together, and add $100,000,000$ unto the SNP positions with each chromosome swtich. For example, three lines in position file [rs11, 1200, 1], [rs22, 2323, 2], and [rs33, 3400, 4] may become [rs11, 1200, 99], [rs22, 100,002323, 99], and [rs33, 200,003400, 99]. The $100,000,000$ increment is because `ELAI` used prior such that 100,000,000 corresponds to 1 Morgan.

5 Running `ELAI`

First some general comments:

- `ELAI` is a command line based program. The command should be typed in a terminal window, in the directory in which `ELAI` executable exists.
- The command line should be all on one line: the line-break (denoted by back-slash) in the example is only because the line is too long to fit the page.
- Unless otherwise stated, the "options" (`-g -p -pos -o`, etc.) are all case-sensitive.
- There are three key parameters to specify, number of upper clusters `-C`, number of lower clusters `-c`, and number of admixing generations `-mg`.

Now we illustrate how to use `ELAI` through examples.

1. **A minimal example for two-way admixture**

```
./elai -g source_pop1.txt -p 10 -g source_pop2.txt -p 11 -g admixed_pop.txt \
      -p 1  -pos position_file.txt -s 30 -o pref -C 2 -c 10 -mg 10
```

The command line will run EM 30 steps, uses 2 upper-layer clusters and 10 lower-layer clusters, assuming number of admixture generation is 10. The output files will start with "pref" in the `output` directory.

2. **A more complicated example for three-way admixture**

```
./elai -g source_pop1.txt -p 10 -g source_pop2.txt -p 11 -g source_pop3.txt \
-p 12 -g admixed_pop.txt -p 1 -pos position_file.txt  -s 30 -o pref -C 3 \
-c 15 -mg 20 -exclude-maf 0.01 --exclude-miss 0.05 --exclude-miss1 \
--exclude-nopos
```

This command line takes three training samples of different ancestral source populations (designated by -p 10, -p 11, and -p 12), and one admixed samples (designated by -p 1), merge them based on the SNP position files. A SNP will be excluded by `ELAI` if its minor allele frequency is < 0.01, or its missing proportion is > 0.05, or it is missed in one population, or its position is not recorded in the position file. `ELAI` will fit a model of 3 upper clusters and 15 lower clusters, by running 30 EM steps. The output files will start with pref.

3. **Use saved EM parameters.**

```
./elai -g source_pop1.txt -p 10 -g source_pop2.txt -p 11 -g admixed_pop.txt \
      -p 1  -pos position_file.txt -s 10 -o pref -C 2 -c 10 -mg 10  \
      -rem output/pref.em.txt
```

This command line will read EM parameters saved in the first example (by default), and continue to run 10 more steps.

4. **The population label -p**
`ELAI` assigns each individual an integer to designate its population label. The population label can be $0, 1, 9, 10, 11, 12, \ldots$. The training samples are labelled as $10, 11, 12, \ldots$, with each number represents an ancestry. It is important, however, that the labels for training samples start with 10 and do not skip an integer. In other words, if the number of ancestral populations is 2, then 10 and 11 will be used, neither 10 and 12, nor 11 and 12 is valid. Similarly, if the number of ancestral populations is 3, then $10, 11$, and 12 will be used. If `-p` is followed by a valid integer, then all individuals in the matching genotype file will be assigned a label of that integer. The `-p` can be followed by a filename; then that file must contain the same number of entries as the number of individuals in the matching genotype file, with one number occupying one row. Finally, an admixed cohort sample is labelled by 1, an un-labelled training sample is labelled by 0, and an sample that is to be excluded is labelled by 9.

6 Output Files

`ELAI` produces output files in a directory named `output/`. The directory will be created automatically if it does not exist. The names of the output files begin with "prefix," which can be specified by the `-o` option. We now describe the contents of these output files.

6.1 Log file: `prefix.log`

A log file contains the command line, details of the progress, and warnings generated. When sending in a bug report, it is important to include the log file as an attachment.

6.2 SNP information file: `prefix.snpinfo.txt`

This file contains 6 columns, with each SNP occupying a row. The columns are rsID, minor allele, major allele, minor allele frequency, chromosome, and position.

6.3 Mean local ancestry dosage: `prefix.ps21.txt`

This file contains the estimated ancestral allele dosages for each individual at each SNP. This file contains N lines, each admixed individuals occupies one line. Each line contains $S \times M$ entries, where S is the number of source populations and M is the number of markers. Let $j = S \times k + s$, then j-th column of a row is the k-th SNP's ancestry allele dosage of the s-th population. This file has the same format for haploid and diploid individuals. In R, one may use the following commands to scan and partition the ancestral allele dosages.

```
> yy=scan("output/prefix.ps21.txt");
> dim(yy)=c(S, M, N);
```

6.4 Joint distribution of local ancestry for diploid individuals: `prefix.ps22.txt`

[05/13/2021] This file will only be generated if the option `--ps2` is used. This file contains N lines, each admixed individuals has one line. Each line contains $S \times S \times M$ entries, where S is the number of source populations and M is the number of markers. If $S = 2$, then each individual at each marker has 4 entries, which is a column stacking of a 2 by 2 symmetric matrix whose entries sum to 1. If $S = 3$, then each individual at each marker has 9 entries, which is a column stacking of a 3 by 3 symmetric matrix whose entries sum to 1. (I understand I can record only upper-triangle of the symmetric matrix, but record all entries makes it easier to parse.).

7 Choice of parameters

For the EM steps (specified with -s), a number between 20 and 50 is recommended. For the upper layer number of clusters (specified with -C), please use 2 for African American, and 3 for Hispanics. Other numbers is possible, provided that you have the appropriate panel data sets. But please be careful with the interpretation. For the lower layer number of clusters (specified with -c), a number, $5 \times C$ is recommended. For the admixture generation (specified with -mg), please use 10 for African American, 20 for Hispanics, and 100 for Uyghurs. Other values of mixture generation can be used, and the inferred local ancestry should be averaged.

7.1 Multiple EM runs.

By default, `ELAI` runs a single EM run. It is recommended to run `ELAI` multiple times and average these results to achieve better estimates. It is important to use `-R` to specify a distinct random number (seed) for each run. If `-R` is omitted `ELAI` will use the machine time as the random seed. But when multiple jobs are run simultaneously, they may accidentally use the same random seed. Thus, `-R` is highly recommended.

7.2 EM steps `-s` and a fast linear approximation `-w`.

The number of EM steps is specified by `-s` and `-s 20` is recommended. This option fits the model using a quadratic algorithm which is accurate but slow. We developed a fast linear approximation to the quadratic algorithm, which is less accurate but fast. One may try it to get a quick glimpse to the data, but it is not yet recommended for serious studies. The linear speed can be achieved using `-w 20 -s 0`.

8 Appendix A: `ELAI` Options

Unless otherwise stated, *arg* stands for a string, *num* stands for a number.
File I/O related options:

- -g *arg* can use multiple times, must pair with -p.
- -p *arg* can use multiple times, must pair with -g. *arg* takes integer values, $1, 10, 11, 12, \ldots$.
- -pos *arg* can use multiple times. *arg* is a file name.
- -o *arg* *arg* will be the prefix of all output files, the random seed will be used by default.

EM Parameters:

- -s(step) *num* specify steps in EM run.
- -C *num* specify number of upper clusters.
- -c *num* specify number of lower clusters.
- -mg *num* specify number of mixture generations.
- -R *num* specify random seed, system time by default.
- -sem *num* save EM results to prefix.em.txt.
- -rem *file* read EM from a file.

Other options:

- -v(ver) print version and citation
- -h(help) print this help

- -exclude-maf *num* exclude SNPs whose maf is less than *num* , default 0.
- `--exclude-nopos` exclude SNPs that has no position information
- `--exclude-miss1` exclude SNPs that are missing in at least one file.
- `--silence` no terminal output.
- `--ps2` output joint distribution of inferred ancestry in file `pref.ps22.txt`. [05/13/2021]

9 Appendix B: `ELAI` source code

If you want to compile an executable from the source code, the first thing to do is to install a gsl library, which can be obtained from `http://www.gnu.org/software/gsl/`. Remember the path to which the gsl is installed and modify the Makefile, the one in the src directory, substituting the old path with the correct path. Then you may type make to compile.

10 Appendix C: What's new

- Minor bug fix on individual weights.
- Bug fix on underflow and outflow for non-human data. [05/13/2021]
- Option "–ps2" can output a joint distribution of the inferred ancestry at each marker. The output column stacks a symmetric matrix of C by C. [05/13/2021]