CEGA-InSel v1.1 User Manual

Shilei Zhao, Lianjiang Chi, Hua Chen

August 8, 2022

Contents

1 Introduction

CEGA-InSel is designed to detect natural selection using multilocus or genomic polymorphism and divergence data from two species. It can detect positive selection in a specific species lineage or balancing selection in one or both species. CEGA-InSel is especially useful for investigating natural selection in noncoding regions. CEGA-InSel implements a two-step maximum likelihood estimation of parameters. In the first step, the software estimates the global model parameters N_0, N_1, N_2 and T. After the global parameters are inferred, CEGA-InSel implements the second step to estimate the locus-specific parameters λ_1^l and λ_2^l and mutation rate μ^l.

If you have any issues or suggestions with the software, please get in touch with Shilei Zhao at zhaoshilei2018d@big.ac.cn.

If you use CEGA-InSel and publish your analysis, please cite the publication:

Shilei Zhao, Lianjiang Chi & Hua Chen. Inferring natural selection by comparative population genomic analysis across species (not published yet).

2 Installation

CEGA-InSel can run on Linux platforms. The CEGA-InSel software is freely available on http://github.com/ChenHuaLab/CEGA-InSel. Download the file “CEGA-InSel-1.1.tar.gz” from the release CEGA-InSel-v1.1.

Implement the following commands:

```
tar -zxvf CEGA-InSel-1.1.tar.gz
cd CEGA-InSel-1.1
make
```

After that, test the software by running the following command:

```
./CEGA-InSel -i1 testdata_species1.vcf -i2 testdata_species2.vcf -t 10 -o result.out
```

It will cost ~5 minutes to complete the program (Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GH), then the file named “result.out” will be generated. Note: Make sure you have at least ten threads, or you can change the number of threads used by the -t option.

3 Command line arguments

To run CEGA-InSel, enter the following command:

```
./CEGA-InSel [arguments]
```

Inputs:

-i1 population 1 genetic variant file (.vcf .vcf.gz .tped .hap).

-p1 population 1 position file (format: chr position, split by tab), only required for .hap (-i1) genetic variant file.

-i2 population 2 genetic variant file (.hap .vcf .vcf.gz .tped).

-p2 population 2 position file (format: chr position, split by tab), only required for .hap (-i2) genetic variant file.

-o set output file name.

Options:

-N0 (double) initial lower bound upper bound (defalut: 10000.0 100.0 1000000.0).

Set the initial value and range of haploid effective population size for common ancestor species. CEGA-InSel will estimate N0 under these constraints. For two species of long-term divergence time, providing additional information on N0 can help to infer global parameters more accurately. Especially, N0 can be fixed by setting the same values of the low bound and up bound.

-N1 (double) initial lower bound upper bound (defalut: 10000.0 100.0 1000000.0).

Similar to -N0. CEGA-InSel can infer N1 from data reasonably, and it is not recommended to set the constraints.

-N2 (double) initial lower bound upper bound (defalut: 10000.0 100.0 1000000.0).

Similar to -N0. CEGA-InSel can inter N2 from data reasonably, and it is not recommended to set the constraints.

-T (double) initial lower bound upper bound (defalut:10000.0 100.0 1000000.0).

Set the initial value and range of divergence time. CEGA-InSel will estimate T under these constraints. For two species of long-term divergence time, providing additional information on T can help to infer global parameters more accurately.

-t (int) thread number (default: 1).

-d (int) filtering windows with s1+s2+s12+D < this value (default: 0).

-mu (double) mutation rate (default:2.5e-8). Unit: per base per generation.

-op (int) 0: bfgs, 1: kmin_hj (default:0). Set optimization algorithm.

-ws (int) window_size step_size (default: 10000 1000).

Set the window size and step size (unit: bp). The first window of each chromosome starts from the first SNP.

-wf (file) window file (format: chr start (1-base, include) end (1-base, include) effective length, split by tab), if input, '-ws' is disable (default: null).

The argument -wf is a substitute for -ws. Set the windows by providing a detailed window information file. The window information file provides effective window sizes, denoting the remaining genomic length after filtering. See details on the file format in the “Input Files” section.

-wf_g (file) window file for estimating global parameters, format same to '-wf' (default: null).

Set the subset of the genomic regions applied in estimating the global parameters. If not input, CEGA-InSel will estimate the global parameters using the complete genomic information in input files. See details on the file format in the “Input Files” section.

4 Input Files

4.1 vcf format

Use -vcf to specify a .vcf file (see details on https://github.com/samtools/hts-specs). A .vcf file contains three parts in the following order: (1) Meta-information lines (lines beginning with "##"). (2) One header line (line beginning with "#CHROM"). (3) Data lines contain marker and genotype data (one variant per line). The first nine columns contain information about the locus, and the file is organized in the following way:

<chr#> <physical position > <id > <reference allele > <alternate allele > <quality > <filter > <info > <format > <individual 1 genotype > ... <individual N genotype >

For example:

```
##fileformat=VCFv4.2
##FORMAT=<ID=GT,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=GP,Number=G,Type=Float,Description="Genotype Probabilities">
##FORMAT=<ID=PL,Number=G,Type=Float,Description="Phred-scaled Genotype Likelihoods">
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO    FORMAT  sample1 sample2 sample3
chr1  108  .   .   .   .   .   .   .   0/0   0/0   0/0
chr1  167  .   .   .   .   .   .   .   1/1   ./.   1/1
chr1  306  .   .   .   .   .   .   .   1/1   1/0   0/1
chr1  336  .   .   .   .   .   .   .   0/0   0/0   0/0
```

represents three diploid samples with variant information at four loci. The symbol "." is used to denote missing data. Columns in the red box are necessary for CEGA-InSel. The genetic data is not required to be phased.

CEGA-InSel accepts any chromosome symbols, such as chr1, 1, chromosome1, etc. Note that the format needs to be unified in all input files. Sex chromosomes need to run separately due to the different effective population sizes. For autosomes, it is recommended to run together to estimate the global parameters more precisely.

4.2 tped format

Use -tped to specify a .tped (transposed PLINK file, see details on http://pngu.mgh.harvard.edu/) file containing genetic variant information. The file is organized in the following way:

<chr#> <id > <genetic position > <physical position > <haploid copy 1> ... <haploid copy N>

For example,

```
chr1 rs108   0.000108 108  0   0   0   0   0   0
chr1 rs167   0.000167 167  1   1   9   9   1   1
chr1 rs306   0.000306 306  1   1   1   0   0   1
chr1 rs336   0.000336 336  0   0   0   0   0   0
```

represents six haploid samples with variant information at four loci. Any symbols except for "0" and "1" are interpreted as missing data.

4.3 hap and pos format

Use -hap to specify a .hap file containing genetic variant information. A .hap file organizes variant data with rows representing a single haploid copy from an individual and columns representing consecutive loci delimited by whitespace. For example,

```
0 1 1 0
0 1 1 0
0 1 1 0
0 1 0 0
0 1 0 0
0 1 1 0
```

represents six haploid samples with variant information at four loci. Any symbols except for "0" and "1" are interpreted as missing data.

Note that for .hap genetic variant file, CEGA-InSel expects an additional position file to provide physical position information. The position file contains two columns representing chromosome and physical position delimited by whitespace. The columns in the .hap file exactly correspond to rows in the position file.

For example,

```
chr1 108
chr1 167
chr1 306
chr1 336
```

represents the positions of four SNPs in the aforementioned .hap example file.

4.4 window file

Use -wf to specify a window file containing the range and the effective size of local windows. CEGA-InSel will detect selection signals for all the local windows. The file is organized in the following way:

<chr#> <start position> <end position> <effective window size>

For example,

```
chr1 1001     11000    8000
chr1 2001     12000    10000
chr1 3001     13000    10000
chr1 4001     14000    10000
chr1 5001     15000    10000
```

represents five overlapping local windows with a window size of 10kb and a step size of 1kb. The effective size of the first window is 8kb, indicating a 2kb length region within chromosome 1: 1001-11000 has been filtered from generating the genetic variant file.

4.5 window file for estimating global parameters

Use -wf_g to specify a window file containing a subset of the input genetic variant information for estimating global parameters. The file is organized in the following way:

<chr#> <start position> <end position> <effective window size>

For example,

```
chr1 10000001     120000000    105000000
chr1 125000001    240000000    110000000
```

represents a 225Mb region with an effective size of 215Mb to estimate the global parameters. Note the windows in this file should not be overlapped.

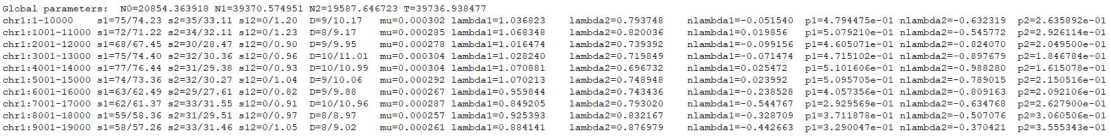
In practice, the window files in sections 4.4 and 4.5 are more complicated than the above examples.

5 Output File

CEGA-InSel produces one file as output. The output file contains 12 columns organized in the following way:

<window position> <polymorphic sites within species 1> <polymorphic sites within species 2> <shared polymorphic sites of both species 1 and 2> <divergent sites> <mutation rate> <lambda1> <lambda2> <normalized lambda1> <p-value 1> <normalized lambda2> <p-value 2>.

For example:



```
Global parameters:  N0=20854.363918 N1=39370.574951 N2=19587.646723 T=39736.938477
chr1:1-10000     s1=75/74.23 s2=35/33.11 s12=0/1.20  D=9/10.17   mu=0.000302 lambda1=1.036823    lambda2=0.793748    nlambda1=-0.051540  p1=4.794475e-01 nlambda2=-0.632319  p2=2.635892e-01
chr1:1001-11000 s1=72/71.22 s2=34/32.11 s12=0/1.23  D=8/9.17    mu=0.000285 lambda1=1.068348    lambda2=0.820036    nlambda1=0.019856   p1=5.079210e-01 nlambda2=-0.545772  p2=2.926114e-01
chr1:2001-12000 s1=68/67.45 s2=30/28.47 s12=0/0.90  D=9/9.95    mu=0.000278 lambda1=1.016474    lambda2=0.739392    nlambda1=-0.099156  p1=4.605071e-01 nlambda2=-0.824070  p2=2.049500e-01
chr1:3001-13000 s1=75/74.40 s2=32/30.36 s12=0/0.96  D=10/11.01  mu=0.000304 lambda1=1.028240    lambda2=0.719849    nlambda1=-0.071474  p1=4.715102e-01 nlambda2=-0.897679  p2=1.846784e-01
chr1:4001-14000 s1=77/76.44 s2=31/29.38 s12=0/0.93  D=10/10.99  mu=0.000304 lambda1=1.070881    lambda2=0.696732    nlambda1=0.025472   p1=5.101606e-01 nlambda2=-0.988280  p2=1.615078e-01
chr1:5001-15000 s1=74/73.36 s2=32/30.27 s12=0/1.04  D=9/10.06   mu=0.000292 lambda1=1.070213    lambda2=0.748948    nlambda1=0.023992   p1=5.095705e-01 nlambda2=-0.789015  p2=2.150516e-01
chr1:6001-16000 s1=63/62.49 s2=29/27.61 s12=0/0.82  D=9/9.88    mu=0.000267 lambda1=0.959844    lambda2=0.743436    nlambda1=-0.238528  p1=4.057356e-01 nlambda2=-0.809163  p2=2.092106e-01
chr1:7001-17000 s1=62/61.37 s2=33/31.55 s12=0/0.91  D=10/10.96  mu=0.000287 lambda1=0.849205    lambda2=0.793020    nlambda1=-0.544767  p1=2.929569e-01 nlambda2=-0.634768  p2=2.627900e-01
chr1:8001-18000 s1=59/58.36 s2=31/29.51 s12=0/0.97  D=8/8.97    mu=0.000257 lambda1=0.925393    lambda2=0.832167    nlambda1=-0.328709  p1=3.711878e-01 nlambda2=-0.507076  p2=3.060506e-01
chr1:9001-19000 s1=58/57.26 s2=33/31.46 s12=0/1.05  D=8/9.02    mu=0.000261 lambda1=0.884141    lambda2=0.876979    nlambda1=-0.442663  p1=3.290047e-01 nlambda2=-0.370421  p2=3.555343e-01
```

The first row represents the estimation of global parameters.

The windows with p1<0.01 (corresponding to nlambda1<-2.3263) indicates species one has 99% confidence under positive selection in the window. The windows with 1-p1<0.01 (corresponding to nlambda1>2.3263) indicates species one has 99% confidence under balancing selection in the window. The same as p2 and nlambda2.

6 How to use CEGA-InSel

6.1 Basic usage

CEGA-InSel is a command-line tool. All supported command-line flags are provided in section 3. The basic execution examples under different formatted genetic variants input files are:

(1) .vcf format

```
./CEGA-InSel -i1 testdata_species1.vcf -i2 testdata_species2.vcf -ws 10000 1000 -t 10 -o result.out
```

(2) .tped format

```
./CEGA-InSel -i1 testdata_species1.tped -i2 testdata_species2.tped -ws 10000 1000 -t 10 -o result.out
```

(3) .hap format

```
./CEGA-InSel -i1 testdata_species1.hap -i2 testdata_species2.hap -p1 testdata_species1.pos -p2 testdata_species2.pos -ws 10000 1000 -t 10 -o result.out
```

Note: When .hap format input files are used, position files must also be provided. We also provided the test files of “wf_10kb_1kb.txt” (for -wf) and “wf_g.txt” (for -wf_g).

6.2 Analyzing population genomic data of humans and chimpanzees

We applied CEGA-InSel to whole-genome sequencing data of nine Homo sapiens and nine Pan troglodytes [1, 2]. Considering the running time, we cut out the 40Mb genome segments (Chr6: 10,000,001-50,000,000) as an example. The command line is:

```
./CEGA-InSel -i1 MHC_human.vcf.gz -i2 MHC_chimpanzee.vcf.gz -N0 30000.0 24000.0 42000.0 -mu 2.5e-8 -t 30 -d 50 -wf wf_MHC.txt -wf_g wf_g_MHC.txt -o result_MHC.out
```

Here, human and chimpanzee genomes are aligned to the same reference genome. Genome segments with tandem repeats, segmental duplication, genomic gaps, and structural variants may cause false positive selection signals. A strict filtering strategy can help to avoid artifact bias in analyzing genomic data [2]. To do this, we first removed the SNPs on the filtering regions from .vcf data files of both species. Then, we prepared the file “wf_MHC.txt” to specify local windows and their effective size (length of the window after excluding the filtering regions). The whole genome was divided into 39,991 windows with a window size of 10 kb and a step size of 1 kb. We excluded the windows with an effective size <2kb from the window file. The selection signals will be detected among the remaining 37,814 windows.

We prepared the file “wf_g_MHC.txt” to specify regions for estimating the global parameters. We excluded the following genome segments from estimating global parameters: (1) regions prone to under selection (such as gene regions and their flanking regions of 10kb); (2) CpG islands with more shared polymorphic sites that are recurrent mutations from identical by state processes rather than identical by descent processes; (3) tandem repeats, segmental duplication, genomic gaps, and structural variants, etc.

For two species with long-term divergence time (such as T>10Ne generations), the estimation of N0 and T may not be accurate as a result of less information from shared polymorphic sites. The additional information on N0 or T can help to estimate the global parameters more precisely. Here, we used the arguments “-N0 30000.0 24000.0 42000.0” to restrict N0 from 24000.0 to 42000.0 according to the previous research [3]. **Warning:** Avoid setting narrow bounds for both N0 and T at the same time, or it may affect the optimization.

We also filtered 568 windows with s1+s2+s12+D<50 by setting "-d 50" considering less information. After completing the running, the results of 37,246 windows will be recorded on the file "result_MHC.out". Figure 1 and Figure 2 show the balancing selection signals in the MHC region (between the bash lines) from the "result_MHC.out" file.

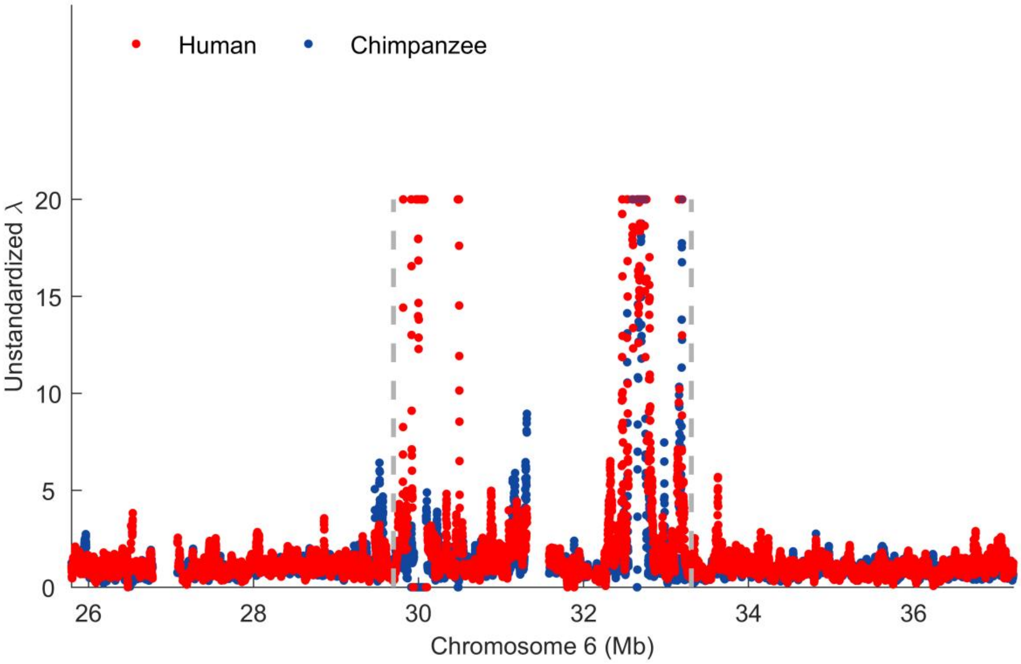


Figure 1. The unstandardized λ values. Values larger than 20 were set to 20 for better illustration.

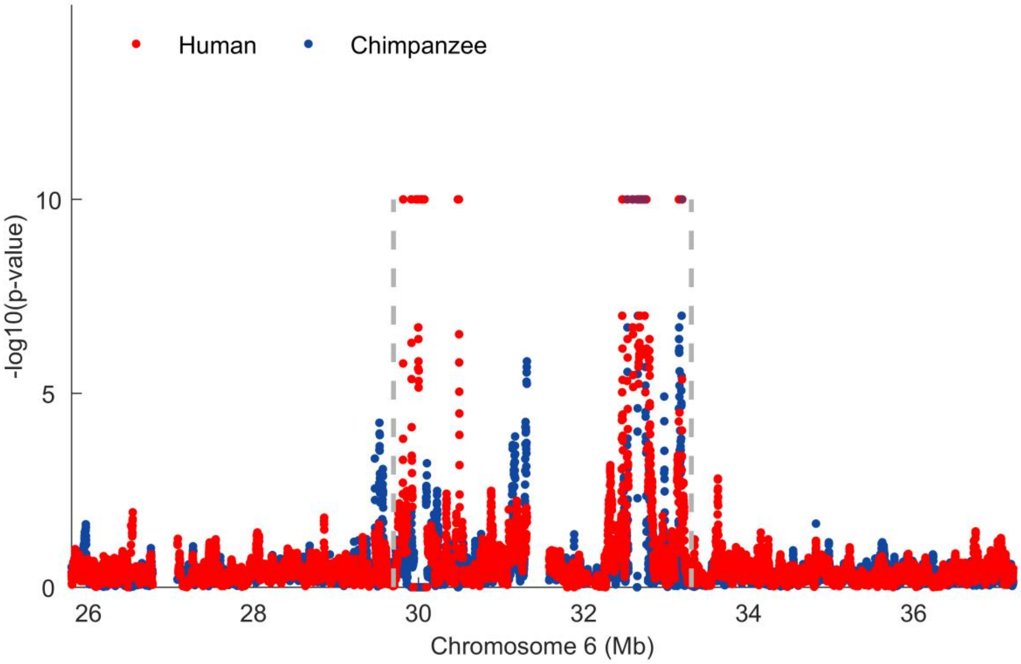


Figure 2. The significant values. Here p-values equal to 1-p1 and 1-p2, since we aim to show balancing selection signals. Values larger than 10 were set to 10 for better illustration.

Reference

1. Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, Veeramah KR, Woerner AE, O'Connor TD, Santpere G, et al: **Great ape genetic diversity and population history.** *Nature* 2013, **499:**471-475.
2. Cagan A, Theunert C, Laayouni H, Santpere G, Pybus M, Casals F, Prufer K, Navarro A, Marques-Bonet T, Bertranpetit J, Andres AM: **Natural Selection in the Great Apes.** *Mol Biol Evol* 2016, **33:**3268-3283.
3. Yang Z: **Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci.** *Genetics* 2002, **162:**1811-1823.