CEGA v1.3 User Manual

Shilei Zhao, Lianjiang Chi, Hua Chen

October 30, 2023

Contents

1 Introduction 2
2 Installation 2
3 Command line arguments 3
4 Input Files 4
4.1 vcf format 4
4.2 tped format 5
4.3 hap and pos format 5
4.4 window file 6
4.5 window file for estimating global parameters 6
5 Output File 7
5.1 CEGA-λ (-LRT 0, default) 7
5.2 CEGA-LRT (-LRT 1) 7
6 How to use CEGA 8
6.1 Basic usage 8
6.2 Analyzing population genomic data of humans and chimpanzees 9
7 Main changes 11
Reference 11

1 Introduction

CEGA 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 is especially useful for investigating natural selection in noncoding regions. CEGA 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 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 and publish your analysis, please cite the publication:

Zhao, S., Chi, L. & Chen, H. CEGA: a method for inferring natural selection by comparative population genomic analysis across species. Genome Biol 24, 219 (2023). https://doi.org/10.1186/s13059-023-03068-8

2 Installation

CEGA can run on Linux platforms. The CEGA software is freely available on http://github.com/ChenHuaLab/CEGA. Download the file "CEGA-1.3.tar.gz" from the release CEGA-v1.3.

Implement the following commands (gcc>=4.4):

```
1.  tar -zxvf CEGA-1.3.tar.gz
2.  cd CEGA-1.3
3.  make
```

After that, the software can be tested by running with the toy example:

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

It costs ~10 seconds to complete the program (Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GH), and generates the output file named "result.out". Note: the number of threads is set with the -t option.

3 Command line arguments

To run CEGA, enter the following command:

```
./CEGA [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 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 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 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 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 10000000.0).

Set the initial value and range of divergence time **(Unit: generation)**. CEGA 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.

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

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 to specify neutral genome region 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 will estimate the global parameters using the complete genomic information in input files. See details on the file format in the “Input Files” section.

-LRT (int)1: implement CEGA-LRT, 0: implement CEGA-λ (default:0)

CEGA provides two methods to do significant test: the distribution of λ (CEGA-λ), and the likelihood ratio test (CEGA-LRT). By default, CEGA implement CEGA-λ.

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. The genetic data is not required to be phased.

CEGA 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 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 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 produces one file as output. The format of output file is depend on the option -LRT.

5.1 CEGA-λ (-LRT 0, default)

If you do the significant test by the distribution of λ, the output file will contain 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=19587.646127	T=39736.938477					
chr1:23-10022	s1=75	s2=35	s12=0	D=9	mu=0.000302	lambda1=1.036823	lambda2=0.793747	nlambda1=-0.054438	p1=4.782929e-01	nlambda2=-0.660911	p2=2.543347e-01
chr1:1023-11022	s1=72	s2=33	s12=0	D=8	mu=0.000282	lambda1=1.085360	lambda2=0.802976	nlambda1=0.056407	p1=5.224914e-01	nlambda2=-0.626717	p2=2.654223e-01
chr1:2023-12022	s1=68	s2=30	s12=0	D=9	mu=0.000278	lambda1=1.016474	lambda2=0.739392	nlambda1=-0.102945	p1=4.590034e-01	nlambda2=-0.875564	p2=1.906336e-01
chr1:3023-13022	s1=75	s2=32	s12=0	D=10	mu=0.000304	lambda1=1.028240	lambda2=0.719849	nlambda1=-0.074745	p1=4.702089e-01	nlambda2=-0.958813	p2=1.688264e-01
chr1:4023-14022	s1=77	s2=31	s12=0	D=10	mu=0.000304	lambda1=1.070881	lambda2=0.696732	nlambda1=0.024031	p1=5.095859e-01	nlambda2=-1.061924	p2=1.441351e-01
chr1:5023-15022	s1=73	s2=32	s12=0	D=9	mu=0.000291	lambda1=1.055961	lambda2=0.752931	nlambda1=-0.009938	p1=4.960354e-01	nlambda2=-0.819858	p2=2.061486e-01
chr1:6023-16022	s1=62	s2=29	s12=0	D=9	mu=0.000266	lambda1=0.945506	lambda2=0.748086	nlambda1=-0.282582	p1=3.887486e-01	nlambda2=-0.839616	p2=2.005619e-01
chr1:7023-17022	s1=63	s2=33	s12=0	D=10	mu=0.000288	lambda1=0.861510	lambda2=0.788013	nlambda1=-0.519417	p1=3.017350e-01	nlambda2=-0.682470	p2=2.474708e-01
chr1:8023-18022	s1=59	s2=31	s12=0	D=8	mu=0.000257	lambda1=0.925393	lambda2=0.832167	nlambda1=-0.336722	p1=3.681631e-01	nlambda2=-0.522443	p2=3.006811e-01
chr1:9023-19022	s1=58	s2=33	s12=0	D=8	mu=0.000261	lambda1=0.884141	lambda2=0.876979	nlambda1=-0.452727	p1=3.253725e-01	nlambda2=-0.372932	p2=3.545994e-01
chr1:10023-20022	s1=55	s2=35	s12=0	D=7	mu=0.000251	lambda1=0.865267	lambda2=0.976694	nlambda1=-0.508188	p1=3.056607e-01	nlambda2=-0.079099	p2=4.684771e-01
chr1:11023-21022	s1=54	s2=34	s12=0	D=9	mu=0.000266	lambda1=0.777989	lambda2=0.888192	nlambda1=-0.786875	p1=2.156774e-01	nlambda2=-0.337357	p2=3.679238e-01
chr1:12023-22022	s1=57	s2=33	s12=0	D=9	mu=0.000269	lambda1=0.826614	lambda2=0.849857	nlambda1=-0.626861	p1=2.653751e-01	nlambda2=-0.461965	p2=3.220533e-01
chr1:13023-23022	s1=52	s2=35	s12=0	D=8	mu=0.000256	lambda1=0.780652	lambda2=0.958825	nlambda1=-0.777778	p1=2.183500e-01	nlambda2=-0.128256	p2=4.489731e-01
chr1:14023-24022	s1=47	s2=35	s12=0	D=8	mu=0.000247	lambda1=0.716143	lambda2=1.000819	nlambda1=-1.010356	p1=1.561624e-01	nlambda2=-0.014896	p2=4.940574e-01

The first row represents the estimation of global parameters.

The values of nlambda follow the standard normal distribution. The windows with p1<0.01 (corresponding to nlambda1<-2.3263) indicate species 1 has 99% confidence under positive selection. The windows with 1-p1<0.01 (corresponding to nlambda1>2.3263) indicate species 1 has 99% confidence under balancing selection.

We recommend the users calculate the p-value by the one-tailed test using the standard normal distribution to avoid numerical problems.

5.2 CEGA-LRT (-LRT 1)

The null hypothesis for the likelihood ratio test is: λ_1^l, $\lambda_2^l = 1$, and μ^l is free (denote the likelihood as $L(\theta_0)$). To test if species 1 is under selection, the alternative hypothesis is set to be: $\lambda_2^l = 1$, λ_1^l and μ^l are free (denote the likelihood as $L(\theta_1)$). To test if species 2 is under selection, the alternative hypothesis is: $\lambda_1^l = 1$, λ_2^l and μ^l are free (denote the likelihood as $L(\theta_2)$).

If you do the significant test by the likelihood ratio test (set the option: -LRT 1), 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> <LLR1> <LLR2> <p1(LRT)> <p1(LRT)>.

For example:

Global parameters:	N0=20854.363918		N1=39370.574951		N2=19587.647915	T=39736.938477					
chr1:23-10022	s1=75	s2=35	s12=0	D=9	mu=0.000291	lambda1=1.081407	lambda2=0.790201	LLR1=0.053834	LLR2=0.529230	p1(LRT)=8.165220e-01	p2(LRT)=4.669307e-01
chr1:1023-11022	s1=72	s2=33	s12=0	D=8	mu=0.000277	lambda1=1.134713	lambda2=0.794457	LLR1=0.128638	LLR2=0.473154	p1(LRT)=7.198479e-01	p2(LRT)=4.915399e-01
chr1:2023-12022	s1=68	s2=30	s12=0	D=9	mu=0.000262	lambda1=1.069342	lambda2=0.737931	LLR1=0.035797	LLR2=0.794745	p1(LRT)=8.499343e-01	p2(LRT)=3.726692e-01
chr1:3023-13022	s1=75	s2=32	s12=0	D=10	mu=0.000287	lambda1=1.087195	lambda2=0.717402	LLR1=0.060074	LLR2=1.031325	p1(LRT)=8.063783e-01	p2(LRT)=3.098478e-01
chr1:4023-14022	s1=77	s2=31	s12=0	D=10	mu=0.000289	lambda1=1.144174	lambda2=0.690685	LLR1=0.150954	LLR2=1.269364	p1(LRT)=6.976252e-01	p2(LRT)=2.598860e-01
chr1:5023-15022	s1=73	s2=32	s12=0	D=9	mu=0.000279	lambda1=1.112738	lambda2=0.747804	LLR1=0.093941	LLR2=0.765676	p1(LRT)=7.592253e-01	p2(LRT)=3.815580e-01
chr1:6023-16022	s1=62	s2=29	s12=0	D=9	mu=0.000245	lambda1=0.985810	lambda2=0.752776	LLR1=0.001618	LLR2=0.665245	p1(LRT)=9.679155e-01	p2(LRT)=4.147142e-01
chr1:7023-17022	s1=63	s2=33	s12=0	D=10	mu=0.000260	lambda1=0.886947	lambda2=0.799726	LLR1=0.130476	LLR2=0.452059	p1(LRT)=7.179387e-01	p2(LRT)=5.013588e-01
chr1:8023-18022	s1=59	s2=31	s12=0	D=8	mu=0.000240	lambda1=0.950820	lambda2=0.839291	LLR1=0.020197	LLR2=0.250694	p1(LRT)=8.869878e-01	p2(LRT)=6.165867e-01
chr1:9023-19022	s1=58	s2=33	s12=0	D=8	mu=0.000243	lambda1=0.900562	lambda2=0.888312	LLR1=0.091318	LLR2=0.117277	p1(LRT)=7.625085e-01	p2(LRT)=7.320071e-01
chr1:10023-20022	s1=55	s2=35	s12=0	D=7	mu=0.000238	lambda1=0.868271	lambda2=0.991688	LLR1=0.165434	LLR2=0.000565	p1(LRT)=6.842019e-01	p2(LRT)=9.810330e-01
chr1:11023-21022	s1=54	s2=34	s12=0	D=9	mu=0.000238	lambda1=0.788326	lambda2=0.907606	LLR1=0.503124	LLR2=0.080435	p1(LRT)=4.781309e-01	p2(LRT)=7.767081e-01
chr1:12023-22022	s1=57	s2=33	s12=0	D=9	mu=0.000243	lambda1=0.842852	lambda2=0.865214	LLR1=0.254915	LLR2=0.179686	p1(LRT)=6.136356e-01	p2(LRT)=6.716433e-01
chr1:13023-23022	s1=52	s2=35	s12=0	D=8	mu=0.000233	lambda1=0.784602	lambda2=0.980148	LLR1=0.509869	LLR2=0.003338	p1(LRT)=4.751955e-01	p2(LRT)=9.539269e-01
chr1:14023-24022	s1=47	s2=35	s12=0	D=8	mu=0.000221	lambda1=0.716085	lambda2=1.027025	LLR1=0.960685	LLR2=0.005729	p1(LRT)=3.270143e-01	p2(LRT)=9.396668e-01

In the output file, LLR1=2[$ln(L(\theta_1))$-$ln(L(\theta_0))$], and LLR2=2[$ln(L(\theta_2))$-$ln(L(\theta_0))$]. After adjustment, LLR follows the Chi-squared distribution with 1 degree of freedom (see details in the published paper). The windows with p1(LRT)<0.01 and lambda1<1 indicate species 1 has 99% confidence under positive selection. The windows with p1(LRT)<0.01 and lambda1>1 indicate species 1 has 99% confidence under balancing selection.

6 How to use CEGA

6.1 Basic usage

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

(2) .tped format

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

(3) .hap format

```
./CEGA -i1 ./testdata/testdata_species1.hap -i2 ./testdata/testdata_species2.hap -p1 ./testdata/testdata_species1.pos -p2 ./testdata/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 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 -i1 ./testdata/MHC_human.vcf.gz -i2 ./testdata/MHC_chimpanzee.vcf.gz -N0 30000.0 24000.0
42000.0 -mu 2.5e-8 -t 30 -d 50 -wf ./testdata/wf_MHC.txt -wf_g ./testdata/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.

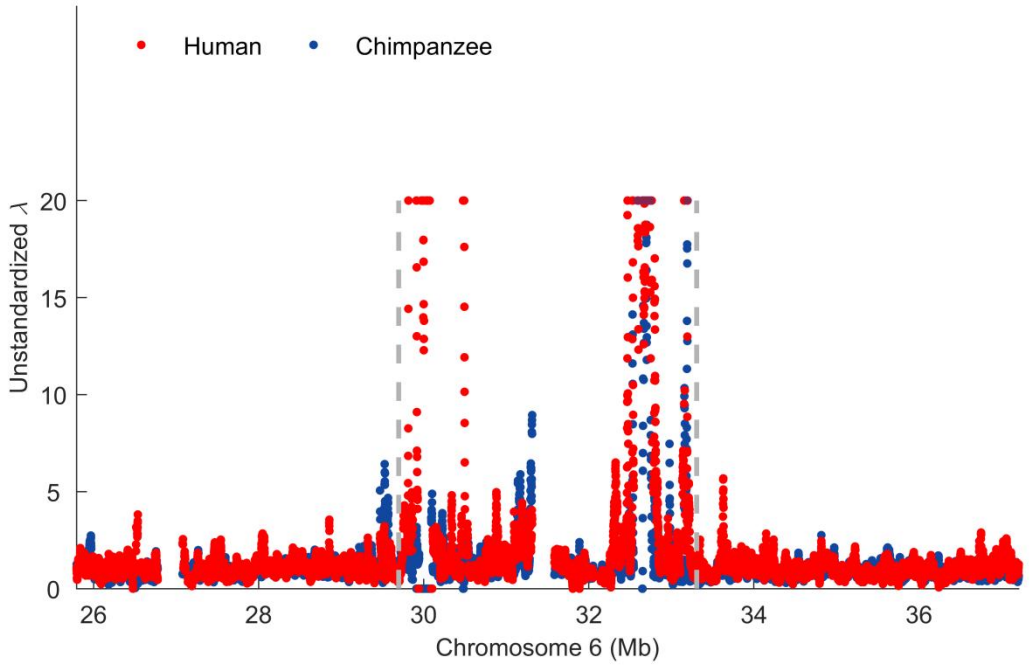

Figure 1. The unstandardized λ values. Values larger than 20 were set to 20 for better illustration.

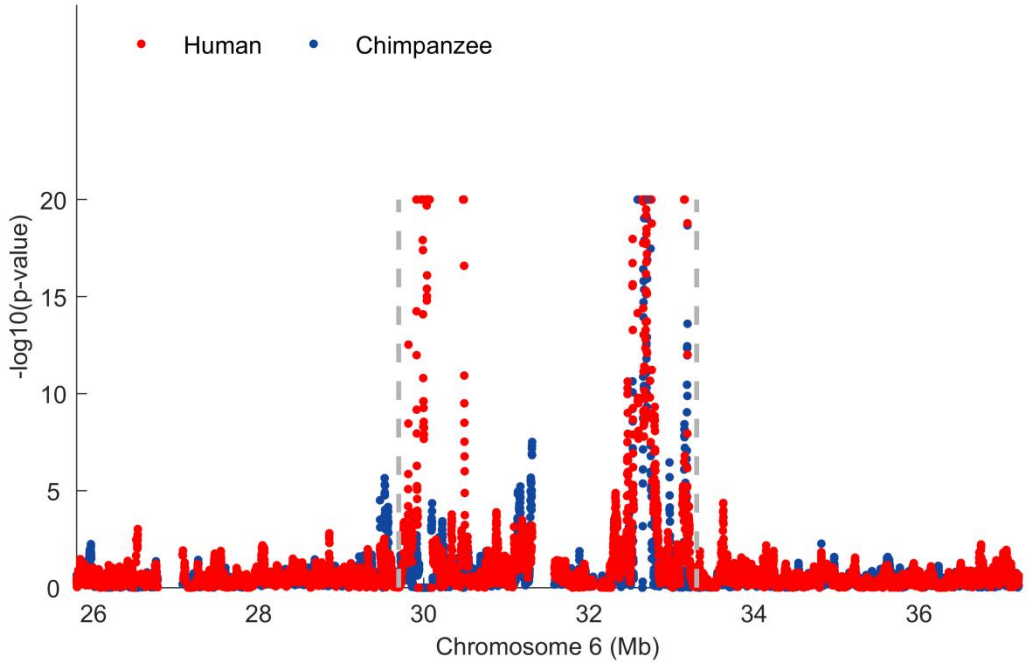

Figure 2. The significant test by CEGA-λ. Here p-values equal to 1-p1 and 1-p2, since we aim to show balancing selection signals. Values larger than 20 were set to 20 for better illustration.

7 Main changes

Main changes of CEGA v1.3 compared with CEGA v1.2:

(1) CEGA v1.3 optimizes memory usage.

(2) CEGA v1.3 eliminates the use of Advanced Vector Extensions (AVX), making it applicable on Linux platforms that do not support AVX.

Main changes of CEGA v1.2 compared with CEGA v1.1:

(1) CEGA v1.2 adds the option -LRT for implementing the likelihood ratio test as an alternative to the significant test (set -LRT 1 for CEGA-LRT).

(2) The running speed of CEGA v1.2 is improved.

(3) The lower bound of λ is changed from 10^{-6} to 10^{-4}. The upper bound of divergence time T is changed from 10^{6} to 10^{7}.

(4) The normalization of λ is changed.

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.