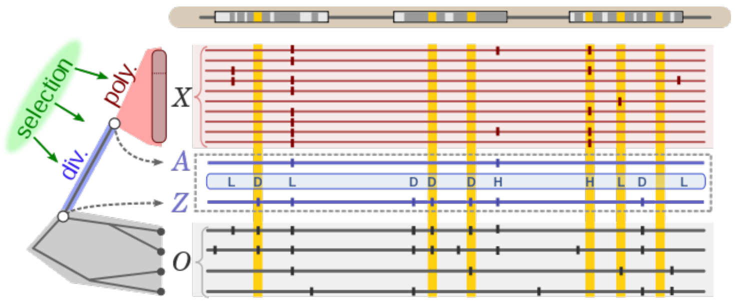
INSIGHT

Inference of Natural Selection from Interspersed Genomically coHerent elemenTs

version 1.0

User Manual



Contents

1. About INSIGHT
2. Download and Install
3. Package Contents
4. Compiling – GSL Dependencies
5. Input File for INSIGHT-EM
6. Simple Running Example
7. Other Running Modes and Options
    7.1 Full list of options
    7.2 Verbose output
    7.3 Altering the halting conditions
    7.4 Logging EM progress
    7.5 Initial parameter values
    7.6 Fixing parameters and likelihood ratio tests (LRTs)
    7.7 Approximate standard errors
    7.8 Posterior expected counts
    7.9 Conditioning on known ancestral states
    7.10 Running EM to estimate proportion between β_1 and β_3
8. Errors and Bugs

1. About INSIGHT

INSIGHT is a software package for inferring signatures of natural selection from a collection of short interspersed genomic elements based on observed patterns of polymorphism and divergence. INSIGHT directly contrasts patterns of polymorphism and divergence within a genomic element with patterns observed in flanking neutral sites, thus accounting for genome-wide variation in mutation rates and genealogical backgrounds and buffering the effect of demography on patterns of polymorphism. INSIGHT uses a full probabilistic model that considers a mixture of weak and strong negative selection, strong positive selection, and neutral drift acting on the elements of interest.

The core component of INSIGHT is an EM algorithm that produces maximum likelihood estimates of the model parameters that describe the influence of selection on polymorphism and divergence. Inferred values of interest include the fraction of sites under selection (ρ), the number of divergences driven by positive selection (D_p), and the number of polymorphisms under weak negative selection (P_w). The algorithm was designed and implemented by Ilan Gronau and Leonardo Arbiza at the Siepel lab in Cornell. More information on INSIGHT can be found in (Gronau et al., 2012) available online on ArXiv.

Gronau I, Arbiza L, Mohammed I, Siepel A. Inference of Natural Selection from Interspersed Genomic Elements Based on Polymorphism and Divergence. Submitted. Preprint available from ArXiv e-prints (2012). **http://arxiv.org/abs/1109.6381**.

This user manual provides basic information for users of the software, as well as several examples. Users should make sure to carefully read the manual before trying out the software. For questions, comments, and feature requests for INSIGHT, please contact Ilan Gronau at:

Ilan Gronau <ig67@cornell.edu>

Good luck,

Ilan.

2. Download and Install

1. Download the INSIGHT source code from the website
   http://compgen.bscb.cornell.edu/INSIGHT
2. Unzip the downloaded file
   ```
   ==> tar -xvzf INSIGHT-v1_0.tar.gz
   ```
3. Move to the unzipped directory
   ```
   ==> cd INSIGHT/
   ```
4. Compile INSIGHT
   ```
   ==> make
   ```
5. The INSIGHT binary (`INSIGHT-EM-v1_0`) will be in the bin/ subdirectory.
6. Post-Install Test Run
   ```
   ==> bin/INSIGHT-EM-v1_0 samples/GATA2_TFBS-f15.ins
   ```

Note: on request, we can also supply executables for multiple platforms such as Mac OS and Windows.

3. Package Contents

After extraction of the tar file, the INSIGHT directory will contain the following components:

File / directory	description
INSIGHT_Manual.pdf	This user manual
Makefile	A make file for compiling and linking INSIGHT-EM (just run 'make')
samples/	Samples directory
scripts/	Scripts directory (currently empty, will contain data processing scripts in future releases)
bin/	Empty directory where the INSIGHT-EM executable is placed during compilation
obj/	Empty directory where all object (.o) files are placed during compilation
src/	Source directory where all source files are placed.
src/Utils.h	header file for Utils.c
src/SumLogs.h	Header file for SumLogs.c
src/NumericOpt.h	Header file for NumericOpt.c
src/bfgs.h	Header file for bfgs.c
src/Utils.c	Module with various general utility functions
src/SumLogs.c	Module for processing and maximizing "sum-of-logs" functions (main component in expected ln-likelihood)
src/NumericOpt.c	Module with various numerical optimization procedures
src/bfgs.c	Module with BFGS algorithm implementation taken from HMMld **(translated from fortran and not used in current implementation !!)**
src/INSIGHT-EM.c	Main source file with implementation of EM

4. Compiling – GSL Dependencies

Compilation of INSIGHT-EM can be done by simple execution of the Makefile. Just enter 'make' in the root INSIGHT directory. INSIGHT-EM uses numerical optimization procedures implemented in GSL (GNU Scientific Library). In order to successfully compile and link the program, you will have to make sure the GSLDIR variable in the Makefile is set appropriately. The default installation directory for GSL is /usr/local/, which is the default value for GSLDIR. If you install GSL in another directory, make sure to adjust GSLDIR to point to that directory.

NOTE: we do intend to make INSIGHT-EM a self-sufficient package, by merging in the relevant GSL code, but for now, it requires installation of the entire GSL package.

```
==> make

Compliling source file src/INSIGHT-EM.c   -->  gcc -fstack-protector-all -Wall -O3 -I/usr/local//include/ -c src/INSIGHT-
EM.c -o obj/INSIGHT-EM.o

Compliling source file src/Utils.c        -->  gcc -fstack-protector-all -Wall -O3 -I/usr/local//include/ -c src/Utils.c -o
obj/Utils.o

Compliling source file src/SumLogs.c      -->  gcc -fstack-protector-all -Wall -O3 -I/usr/local//include/ -c src/SumLogs.c
-o obj/SumLogs.o

Compliling source file src/NumericOpt.c   -->  gcc -fstack-protector-all -Wall -O3 -I/usr/local//include/ -c
src/NumericOpt.c -o obj/NumericOpt.o

Compliling source file src/bfgs.c         -->  gcc -fstack-protector-all -Wall -O3 -I/usr/local//include/ -c src/bfgs.c -o
obj/bfgs.o

Building   executable bin/INSIGHT-EM-v1.0 -->  gcc obj/INSIGHT-EM.o obj/Utils.o obj/SumLogs.o obj/NumericOpt.o obj/bfgs.o
-fstack-protector-all -Wall -O3 -I/usr/local//include/ -L/usr/local//lib/  /usr/local//lib/libgsl.a
/usr/local//lib/libgsl.a -lm -o bin/INSIGHT-EM-v1.0
```

INSIGHT-EM Compilation

5. Input File for INSIGHT-EM

An INSIGHT input file is expected to have exactly one line in each of the following two formats:

- **samples <*N*>**
  N – number of chromosome samples used for population variation data (2 X number of individuals sampled)
- **beta <*beta1*> <*beta2*> <*beta3*>**
  beta1, beta2, beta3 – values for β_1, β_2, and β_3 neutral parameters of the model. The selection EM requires the user to supply these. The beta1_3 EM computes the relative contribution of β_1 and β_3 and thus does not require a 'beta' line in the input file.

Additionally, the input file consists a series of 'block' and 'site' lines in the following format:

- **block <*blockID*> theta <*theta*> lambda <*lambda*>**
  blockID – ID for genomic block (typically characterized by genomic coordinates)
  theta – value of block-specific neutral polymorphism rate θ_b associated with block.
  lambda – value of block-specific neutral divergence rate $\lambda_b t$ associated with block.
- **site <*siteID*> <*polyType*> <*majProb*> [<*minProb*>]**
  siteID – ID for site (genomic coordinate or element ID + position within element)
  polyType – **M** for monomorphic sites, **L** for polymorphic sites with low minor allele frequency, and **H** for polymorphic sites with high minor allele frequency.
  majProb – the prior probability that the deep ancestral state Z_i equals the observed major allele (or the only observed allele in case of an 'M' site).
  minProb – the prior probability that the deep ancestral state Z_i equals the observed minor allele (minProb is not specified for 'M' site).

Overall, the order of lines in the input file does not matter. So the 'samples' and 'beta' lines can appear anywhere in the file, and 'blocks' can be ordered arbitrarily. However, 'site' lines, which contain summaries of the sequence data in the analyzed elements, must be grouped according to their respective genomic blocks (the order of 'site' lines within a block does not matter). More formally, each 'site' line is associated with a genomic block defined by the 'block' line that is the closest to it among the 'block' lines that precede it.

The following short input file contains sequence data for 21 nucleotide sites along the human genome (hg19) spanning two genomic blocks. The variation data considers 108 chromosome samples (the 54 unrelated samples in the Complete Genomics data set; see Gronau et al., 2012). There are 19 monomorphic sites in this set, each of which is given with the prior probability that the deep ancestral state (Z_i) equals the observed allele in the population. There are two polymorphic sites with low minor allele frequencies. Those are given with two prior probabilities for the deep ancestral state: one corresponding to the major observed allele, and one corresponding to the minor observed allele. Note that for the two polymorphic sites, the file also provides information about the frequency of the major and minor alleles (107/1 for both sites). This information is not processed by the program, but can be used to relabel polymorphic sites as 'H' or 'L' according to different frequency thresholds.

```
samples 108
block   chr1:21602500-21607500      theta    0.000329247          lambda     0.00564974
site    chr1:21606850      M        0.999955
site    chr1:21606851      M        0.999956
site    chr1:21606852      M        0.999955
site    chr1:21606853      L        0.999955 4.00882e-05          107        1
site    chr1:21606854      M        0.999955
site    chr1:21606855      M        0.999955
site    chr1:21606856      M        0.999955
site    chr1:21606952      M        0.999955
site    chr1:21606953      M        0.999956
site    chr1:21606954      M        0.999955
site    chr1:21606955      M        0.999955
site    chr1:21606956      M        0.999955
site    chr1:21606957      M        0.999955
site    chr1:21606958      M        0.999956
block   chr1:21632500-21637500      theta    0.000538397          lambda     0.00263024
site    chr1:21634276      M        0.999969
site    chr1:21634277      M        0.999969
site    chr1:21634278      M        0.999969
site    chr1:21634279      L        0.999969 2.8055e-05           107        1
site    chr1:21634280      M        0.999969
site    chr1:21634281      M        0.999969
site    chr1:21634282      M        0.999969
beta    0.772809           0.205993 0.021198
```

An example of a short input file for INSIGHT-EM

6. Simple Running Example

Simple execution of INSIGHT-EM uses the following command line:

```
==> INSIGHT-EM-v1_0 inputFile.ins [optional flags]
```

For a complete list of all options, run INSIGHT-EM with the --help (-h) option (see section 6.1). Running INSIGHT-EM with default options on the sample input file corresponding to GATA2 binding sites (GATA2_TFBS-f15.ins) results in the following output:

```
==> bin/INSIGHT-EM-v1.0 samples/GATA2_TFBS-f15.ins
Progress: .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
          .......... .......... .......... .......... .......... .

----------------------------------------------------------
             rho      eta    gamma       Dp       Pw
Estimates: 0.244650 0.684396 0.434543 0.615027 0.509911
StndrdErr: 0.056520 0.161223 0.132705 0.242632 0.244464
----------------------------------------------------------
             iter     lnLd       diff       status
EM status:   15168   -15166.3 9.99751e-07  converged
----------------------------------------------------------
==>
```

Sample output for INSIGHT-EM

The output contains a progress indicator (each '.' indicates 100 EM iterations), followed by the EM results given in three separate lines.

- The 'Estimates' line provides the maximum likelihood estimates produced by the EM algorithm for the three selection parameters (ρ,η, and γ) and the posterior expected values of the number of divergences under strong positive selection ($E[D_p]$) and the number of polymorphism under weak negative selection ($E[P_w]$), normalized per 1000 bp (kbp). In the above example, the data set is inferred to have 24% sites under selection (ρ), 0.62 adaptive divergences per kbp ($E[D_p]$), and 0.51 weakly deleterious polymorphisms per kbp ($E[P_w]$). The selection parameters η and γ describe the relative divergence and polymorphism rates for site under selection (compared to the local neutral rates), and their estimated values do not have a straightforward interpretation. The expected counts $E[D_p]$ and $E[P_w]$ encapsulate these estimates in measures that are more easy to interpret.

- The 'StndrdErr' line provides approximate standard errors for the five estimates, obtained using the so-called *curvature method*, which uses the curvature (second derivative) of the log-likelihood function at the point of estimation to assess confidence in the estimates. In the above example, the fraction of sites under selection is estimated as ρ=24%±6%.
- The 'EM status' line provides additional information on the progress of the EM algorithm.
  - number of EM iterations (15,168 in the above example)
  - the ln-likelihood associated with the final estimate (-15166.3 in the following example)
  - the ln-likelihood difference between the last two EM iterations.
  - The final status of the EM:
    - 'converged' – EM reached successful convergence (ln-likelihood difference is below threshold defined for halting)
    - 'timeout' – EM reached maximum number of alloted iterations
    - 'overshoot' – EM stopped due to decrease in ln-likelihood (might indicate rounding errors or sub-optimal solution provided by optimization procedure)
    - 'zero-likelihood' – EM converged to a solution with zero likelihood (when parameters are constrained, or falsely initialized)
    - 'error' – EM encountered some internal error (will be accompanied with an error message)

7. Other Running Modes and Options

7.1 Full list of options

A full list of all flags and options is given when running INSIGHT-EM with the **--help (-h)** flag:

```
==> bin/INSIGHT-EM-v1.0 --help
+-------------------------------------------------------------------------------------
| INSIGHT-EM (v1.0)  - program for estimating selection parameters from poly/div patterns
+-------------------------------------------------------------------------------------
| Usage: ' bin/INSIGHT-EM-v1.0 infile [optional flags] '
|  infile contains a summary of sequence information across a given set of genomic positions
+-------------------------------------------------------------------------------------
| optional flags:
|
| -h      --help : show this usage message
| -s    --simple : reverts to simpler version of the model conditioning on known ancestral states
| -b   --beta1-3 : runs EM on neutral 'L' sites to estimate beta1/(beta1+beta3) [ optional initial value, default = 0.5 ]
|
| ~~~~ I/O ~~~~
| -v   --verbose : run with more messages outputed to screen
| -f  --log-file : log filename for EM       ( required if -l --log option is used )
| -p  --post-cnt : produce posterior counts of all site types into a specified file
| -l  --log-iter : number of iterations between log printouts ( default = 100 )
| -c  --no-conf  : do NOT compute confidence intervals for parameters ( computed by default )
|
| ~~~~ EM halting conditions ~~~~
| -i  --max-iter : upper bound on number of EM iterations     ( default = 20,000   )
| -d  --min-diff : ln-likelihood difference at which EM stops ( default = 0.000001 )
|
| ~~~~ EM initialization ~~~~
| -r  --rho-init : initial value for rho   parameter          ( default = 0.6 )
| -e  --eta-init : initial value for eta   parameter          ( default = 1.0 )
| -g  --gam-init : initial value for gamma parameter          ( default = 0.5 )
|
| ~~~~ EM limit updates ~~~~
| -fr  --fix-rho : do not update rho   parameter
| -fe  --fix-eta : do not update eta   parameter
| -fg  --fix-gam : do not update gamma parameter
+-------------------------------------------------------------------------------------
==>
```

Usage options for INSIGHT-EM

7.2 Verbose output

the **--verbose (-v)** flag produces a more verbose output trace, containing information on the EM halting conditions, logging options, initial parameter values, and running time (indicated before the three EM result lines).

```
==> bin/INSIGHT-EM-v1.0 samples/GATA2_TFBS-f15.ins -v
----------------------------------------------------------------------------------------------
       INSIGHT-EM v1.0, September 2012
----------------------------------------------------------------------------------------------
==> Processing site data file 'samples/GATA2_TFBS-f15.ins' and extracting neutral model parameters
==> Performing EM on selection parameters
    - EM stops after 20000 iterations or when log-likelihood increase is below 1e-06
    - '.' = 100 iterations
    - estimating the following parameters: rho (init = 0.600000) eta (init = 1.000000) gamma (init = 0.500000).
    - using complete version of the model integrating over assignments to the ancestral states Zi.
----------------------------------------------------------------------------------------------
Progress: .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
          .......... .......... .......... .......... .......... .
Done. Running time  1m40s.
----------------------------------------------------------------------------------------------


----------------------------------------------------------
             rho      eta    gamma       Dp       Pw
Estimates: 0.244650 0.684396 0.434543 0.615027 0.509911
StndrdDev: 0.056520 0.161223 0.132705 0.242632 0.244464
----------------------------------------------------------
             iter     lnLd       diff       status
EM status:   15168   -15166.3 9.99751e-07  converged
----------------------------------------------------------
==>
```

Verbose output of INSIGHT-EM

7.3 Altering the halting conditions

The EM algorithm halts if one of the following has occurred:

1. The maximum number of iterations has been exceeded. The default maximum is 20,000 iterations, and it can be modified using the **--max-iter (-i) <*maxIter*>** option.
2. The difference between the ln-likelihood of the last two EM iterations went below the ln-likelihood threshold. The default threshold is 0.000001, and it can be modified using the **--min-diff (-d) <*diff*>** option. **Note:** this holds as long as the ln-likelihood increases.
3. The ln-likelihood decreased from the previous iteration. This happens only when the optimization procedure fails, or due to precision issues. It will be indicated by an 'overshoot' status in the 'EM status' line.
4. The EM encountered some error. An 'error' status will be given in such a case.

The maximum number of iterations is a “safety” feature ensuring that the EM does not run indefinitely, but you eventually want to let the EM converge for each data set (by sufficiently increasing the maximum number of iterations). It is also good practice to adjust the ln-likelihood threshold to ensure the EM did not converge to a very wide plateau (this should also be indicated by large standard errors for the parameter estimates).

7.4 Logging EM progress

A trace of the EM algorithm can be written into a file, for diagnostic purposes, by specifying a log file using the **--log-file (-f) <file-name>** option . Snapshots of the EM algorithm are logged every *logIter* iterations, where *logIter* is 100 by default and can be set using the **--log-iter (-l) <*logIter*>** option. Note: the --log-iter option can be used only together with the --log-file option. Each line of the log file contains the iteration index, the current values of three selection parameters (ρ, η, and γ), the current ln-likelihood, the ln-likelihood difference from the previous iteration, and the same with the expected ln-likelihood (which is the measure being maximized in each iteration of the EM). Below is an example of the first three lines in a log file, where *logIter* is set to 1. Note that while the ln-likelihood consistently improves, the expected ln-likelihood is a different function in every iteration (it depends on the current parameter values). The improvement in the expected ln-likelihood uses the current function and the two sets of parameters: the current one and the updated one (used in the next iteration). Tracking the improvements in the expected ln-likelihood can be used to diagnose errors and slow convergence.

```
iter      rho      eta    gamma       lnLd    lnLd_df    E[lnLd] E[lnLd]_df
   1 0.600000        1 0.500000 -39538.118126 -39538.118126    0.000000   0.000000
   2 0.599657 0.669322 0.645274 -39318.782717 219.335410 -532447.790833 155.975934
   3 0.599435 0.558648 0.727724 -39276.651532  42.131185 -532490.666601  28.713440
```

Log file example

7.5 Initial parameter values

Since the EM algorithm is an iterative update process for finding the maximum likelihood estimates, it requires indicating a starting point for that search. The default starting point for INSIGHT-EM is defined as (ρ=0.6 ; η=1.0 ; γ=0.5). An alternative starting point can be given by the user through the **--rho-init (-r) <*rhoInit*>**, **--eta-init (-e) <*etaInit*>**, or **--gam-init (-g) <*gammaInit*>** options. **Note:** choosing a starting point at the boundary of the parameter space (ρ=0 or ρ=1 or η=0 or γ=0) will restrict the search to that boundary, so INSIGHT-EM allows specifying a starting point at the boundary only if the relevant parameter is explicitly indicated to be fixed in the EM algorithm (see next section).

The following example demonstrates an INSIGHT-EM execution with an alternative starting point (see example in section 7.2 for comparison).

```
==> bin/INSIGHT-EM-v1.0 samples/GATA2_TFBS-f15.ins -v -r 0.1 -e 0.01 -g 10.0
----------------------------------------------------------------------------------------------
       INSIGHT-EM v1.0, September 2012
----------------------------------------------------------------------------------------------
==> Processing site data file 'samples/GATA2_TFBS-f15.ins' and extracting neutral model parameters
==> Performing EM on selection parameters
    - EM stops after 20000 iterations or when log-likelihood increase is below 1e-06
    - '.' = 100 iterations
    - estimating the following parameters: rho (init = 0.100000) eta (init = 0.010000) gamma (init = 10.000000).
    - using complete version of the model integrating over assignments to the ancestral states Zi.
----------------------------------------------------------------------------------------------
Progress: .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
          .......... .......... .......... .......... .......... .......... .......... ....
Done. Running time  2m10s.
----------------------------------------------------------------------------------------------

----------------------------------------------------------
             rho      eta    gamma       Dp       Pw
Estimates: 0.236245 0.660969 0.422455 0.573568 0.478738
StndrdErr: 0.057682 0.171142 0.140193 0.248119 0.248169
----------------------------------------------------------
             iter     lnLd       diff       status
EM status:   17446   -15166.3 9.99647e-07  converged
----------------------------------------------------------
==>
```

Alternative starting point for INSIGHT-EM

7.6 Fixing parameters and likelihood ratio tests (LRTs)

It is possible to instruct the EM algorithm to keep one or more of the parameters fixed at their initial value using the **--fix-rho (-fr)**, **--fix-eta (-fe)**, or **--fix-gam (-fg)** flag. This enables INSIGHT-EM to find maximum likelihood estimates (MLEs) within subspaces of the entire parameter space. Comparing the ln-likelihood of the restricted MLE with that of the general MLE enables hypothesis testing through a likelihood ration test (LRT). Twice the ln-likelihood difference is treated as a test statistic and compared to the appropriate χ^2 distribution. For testing significant evidence for positive selection ($\eta>0$) or weak negative selection ($\gamma>0$), we suggest comparing to a χ^2 distribution with one degree of freedom, and for testing significant evidence for overall selection ($\rho>0$), we suggest comparing to a χ^2 distribution with three degrees of freedom.

The following table provides ln-likelihood differences for various LRT significance thresholds.

P-value	**0.05**	**0.01**	**0.005**	**0.001**	**0.0005**	**0.0001**	**0.00005**	**0.00001**
$\chi^2_{df=3}$	3.91	5.67	6.42	8.13	8.87	10.55	11.28	12.95
$\chi^2_{df=1}$	1.92	3.32	3.94	5.41	6.06	7.57	8.22	9.76

In order to assess the significance of all types of selection for the GATA2_TFBS-f15.ins data set, execute the four runs of INSIGHT-EM, as follows:

```
==> bin/INSIGHT-EM-v1.0 samples/GATA2_TFBS-f15.ins
Progress: .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
          .......... .......... .......... .......... .......... .
----------------------------------------------------------
             rho      eta    gamma       Dp       Pw
Estimates: 0.244650 0.684396 0.434543 0.615027 0.509911
StndrdErr: 0.056520 0.161223 0.132705 0.242632 0.244464
----------------------------------------------------------
             iter     lnLd       diff       status
EM status:   15168   -15166.3 9.99751e-07  converged
----------------------------------------------------------
==> bin/INSIGHT-EM-v1.0 samples/GATA2_TFBS-f15.ins -fr -r 0
Progress:
----------------------------------------------------------
             rho      eta    gamma       Dp       Pw
Estimates: 0.000000 0.000000 0.000000 0.000000 0.000000
StndrdErr: 0.000000      nan      nan      nan      nan
----------------------------------------------------------
             iter     lnLd       diff       status
EM status:       2   -15180.5       0   converged
----------------------------------------------------------
==> bin/INSIGHT-EM-v1.0 samples/GATA2_TFBS-f15.ins -fe -e 0
Progress: .......... .......... .......... .......... .......... .......... .......... ..........
----------------------------------------------------------
             rho      eta    gamma       Dp       Pw
Estimates: 0.119566 0.000000 0.080688 0.000000 0.046392
StndrdErr: 0.030366 0.000000 0.280502      nan 0.168918
----------------------------------------------------------
             iter     lnLd       diff       status
EM status:    8069   -15168.9 9.99231e-07  converged
----------------------------------------------------------
==> bin/INSIGHT-EM-v1.0 samples/GATA2_TFBS-f15.ins -fg -g 0
Progress: .......... .......... .......... .......... .......... .......... .......... ...
----------------------------------------------------------
             rho      eta    gamma       Dp       Pw
Estimates: 0.139022 0.440733 0.000000 0.225060 0.000000
StndrdErr: 0.029027 0.280096 0.000000 0.167721      nan
----------------------------------------------------------
             iter     lnLd       diff       status
EM status:    7309   -15168.1 9.99464e-07  converged
----------------------------------------------------------
```

Likelihood ratio tests using INSIGHT-EM

These runs derive a test statistic of 15180.5 - 15166.3 = 14.2 for the hypothesis that $\rho>0$, which implies an approximate P-value smaller than 0.00001 according to above table. Similarly, test statistics of 15168.9 - 15166.3 = 2.6 and 15168.1 - 15166.3 = 1.8 are associated with the hypotheses $\eta>0$ and $\gamma>0$ (resp.), which imply an approximate P-value between 0.05 and 0.01 for positive selection and an approximate P-value greater than 0.05 for weak negative selection.

7.7 Approximate standard errors

By default, INSIGHT-EM computes approximate standard errors for the selection parameters ρ, η, and γ (as well as $E[D_p]$ and $E[P_w]$). This option can be turned off using the **--no-conf (-c)** flag. **Note:** the approximate standard errors will typically be less accurate near the boundaries of the parameter space ($\rho=0$ or $\rho=1$ or $\eta=0$ or $\gamma=0$). This might lead to errors in the computation of the variance/covariance matrix, in particular negative diagonal elements, which result in an error message.

7.8 Posterior expected counts

It is possible to instruct INSIGHT-EM to output posterior expected counts for all configurations of the hidden variables (A_i, Z_i, and S_i). **Note:** this mode (invoked by the **--post-cnt (-p)** option) is in trial stages, and will be operational in the next subversion. **Please do not use !**

7.9 Conditioning on known ancestral states

It is possible to run INSIGHT-EM under the assumption of known deep ancestral states (Z_i), rather than considering uncertainty in ancestral specification. This implies a slightly simpler model that can make use of a simpler format of the input file, where for each genomic block, the number of sites in each type (monomorphic nondivergent, monomorphic divergent, polymorphic low and polymorphic high) is indicated in a single line. **Note: this mode (invoked by the --simple (-s) option) is depracacted, and will likely not be operational in future subversions. Please do not use !**

7.10 Running EM to estimate proportion between β_1 and β_3.

INSIGHT-EM has a mode of operation, **--beta1-3 (-b)**, that instructs it to estimate the proportion between β_1 and β_3. This mode receives an optional argument ***<initB1>***, which is the initial proportion. The output of this procedure is the maximum likelihood estimate of $\beta_1/(\beta_1 + \beta_3)$. Note that the value of β_2 is separately estimated according to the observed number of polymorphic sites with high minor allele frequency. This procedure receives as input a file with similar structure as that of the standard input file (see section 5). The file should contain all 'L' polymorphic sites within flanking sites belonging to genomic blocks that contain elements of interest. The EM procedure uses information from these sites, together with the pre-estimated neural divergence rates ($\lambda_b t$) associated with the relevant genomic blocks. Any non-'L' site in the input file is ignored by this procedure, as well as the neutral polymorphism rates (θ_b). Additionally, a 'beta' line is not required in this mode.

8. Errors and Bugs.

MORE INFORMATION TO BE ADDED

In the meanwhile, if you encounter an error, please send the exact error message together with the input file that generated the error to Ilan Gronau <ig67@cornell.edu>.