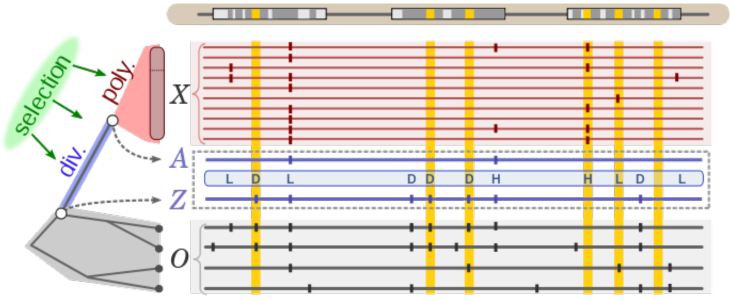
INSIGHT

Inference of Natural Selection from Interspersed Genomically coHerent elemenTs

version 1.1

User Manual



Contents

1. About INSIGHT

INSIGHT is a software package for inferring signatures of recent 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, positive selection, and neutral drift acting on the elements of interest.

INSIGHT-EM is the core software component of INSIGHT, which includes implementations for the two EM algorithms required for INSIGHT inference. The main EM algorithm produces maximum likelihood estimates of the model parameters that describe the effects of natural 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). Another EM algorithm is used to compute the ratio between neutral polymorphism parameters β_1 and β_3. The algorithms were designed and implemented by Ilan Gronau and Leonardo Arbiza at the Siepel lab in Cornell. The implementation is written in C, and should be compilable under all common platforms. More information on INSIGHT can be found in (Gronau *et al.*, 2013).

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:

<ig67@cornell.edu>

Cite: *Gronau I, Arbiza L, Mohammed I, Siepel A.* Inference of Natural Selection from Interspersed Genomic Elements Based on Polymorphism and Divergence. *Mol Biol Evol*. 2013. In press, doi: 10.1093/molbev/mst019

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_1.tar.gz`
3. Move to the unzipped directory ==> `cd INSIGHT/`
4. Compile INSIGHT ==> `make`
5. The INSIGHT binary (`INSIGHT-EM-v1.1`) will be in the bin/ subdirectory.
6. Post-install test run

   ==> `bin/INSIGHT-EM-v1.1 samples/thresholdedInput/miRNAs.f15.ins`

3. Package Contents

After extraction of the tar file, the INSIGHT directory will contain a Makefile, this user manual, and the following subdirectories (see README file in each directory for more details):

Directory	description
samples/	Samples directory (see samples/REAME for more detail)
scripts/	Scripts directory (see scripts/REAME for more detail)
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 (see src/README for more detail)

4. Compiling – GSL Dependencies

Compilation of INSIGHT-EM can be done by simple execution of the Makefile. Just enter 'make' in the root INSIGHT-EM 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 in the future 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-v1.1 -->  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-v1.1
```

INSIGHT-EM Compilation

5. Input File for INSIGHT-EM

Each line in the INSIGHT-EM input file should have one of the following configurations (spacers can be arbitrary white spaces)i:

- **samples <*N*>**
  N – number of chromosome samples used for population variation data (2 X number of individuals sampled). ***Input file should contain a single line of this format*.**

- **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. ***Input file should contain no more than line of this format. When running the beta1_3 EM, this line is not required*.**

- **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 <site*ID*> <*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, such that 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, 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.

```
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

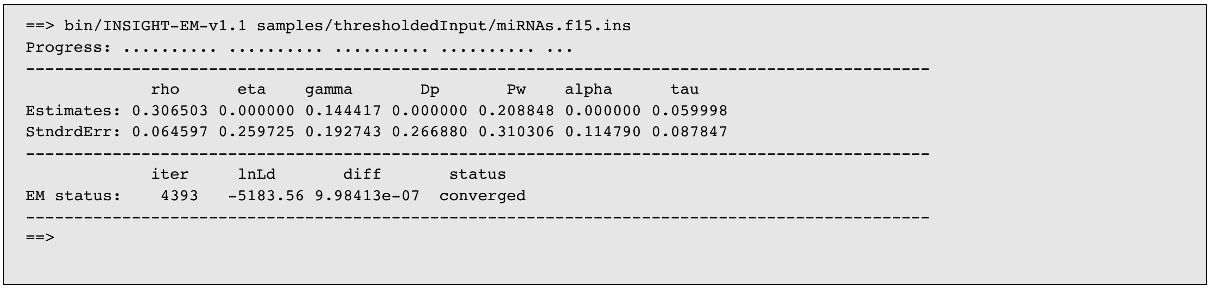
The above short input file contains sequence data for 21 nucleotide sites along the human genome (hg19) spanning two genomic blocks. The variation data considers 108 chromosome samples (the 54 unrelated samples in the Complete Genomics data set; see Gronau *et al.,* 2013). There are 19 monomorphic sites 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.

6. A Simple Running Example

A simple execution of INSIGHT-EM uses the following command line:

```
==> INSIGHT-EM-v1.0 inputFile [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 (samples/thresholdedInput/GATA2-TFBS.f15.ins) results in the following output:



```
==> bin/INSIGHT-EM-v1.1 samples/thresholdedInput/miRNAs.f15.ins
Progress: .......... .......... .......... .......... ...
----------------------------------------------------------------------------------------------
             rho      eta    gamma       Dp       Pw    alpha      tau
Estimates: 0.306503 0.000000 0.144417 0.000000 0.208848 0.000000 0.059998
StndrdErr: 0.064597 0.259725 0.192743 0.266880 0.310306 0.114790 0.087847
----------------------------------------------------------------------------------------------
             iter     lnLd       diff       status
EM status:    4393   -5183.56 9.98413e-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.

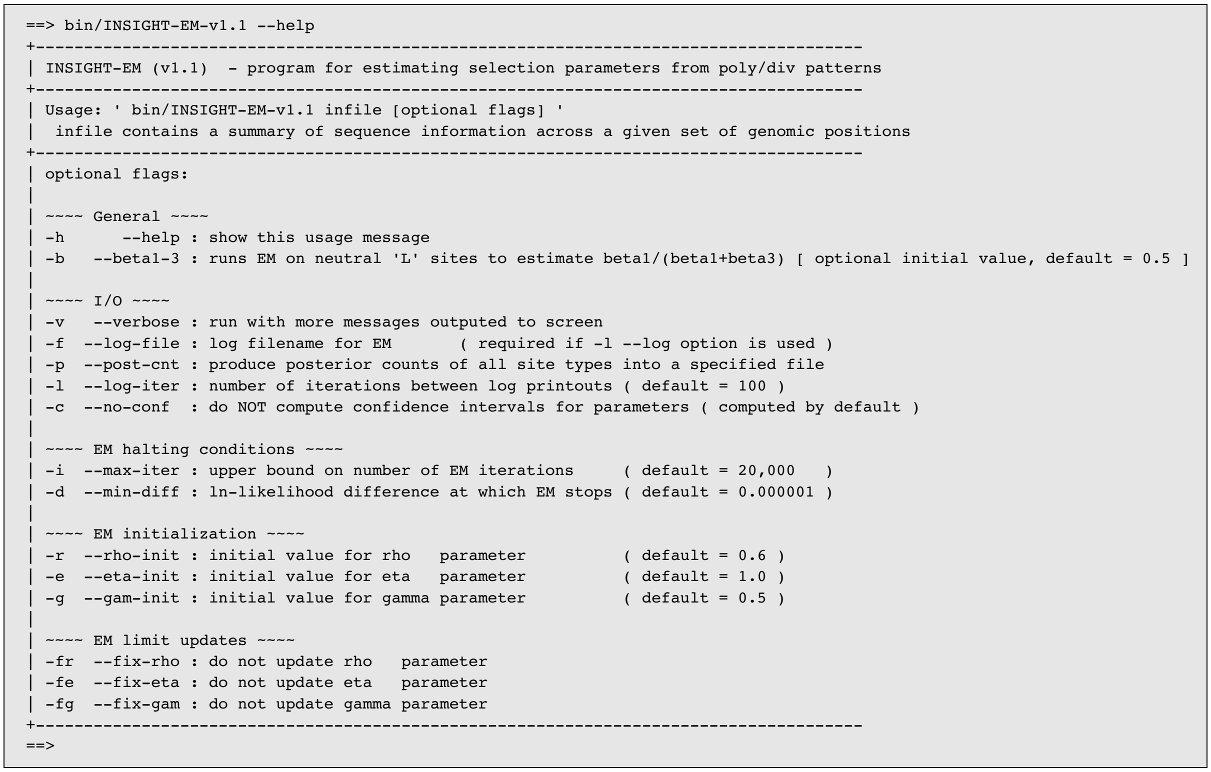
- The 'Estimates' line provides the maximum likelihood estimates produced by the EM algorithm for the three selection parameters (ρ,η, and γ) and the posterior expected values of the number of divergences under strong positive selection (E[D_p]) and the number of polymorphism under weak negative selection (E[P_w]), normalized per 1000 bp (kbp). We also provide versions of these posterior counts normalized by total (expected) number of divergences (α) and total number of polymorphisms (τ). In the above example, the data set is inferred to have 29% sites under selection (ρ), 0.81 adaptive divergences per kbp (E[D_p]), and 0.73 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 seven estimates, obtained using the *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=30.7%±6.5%.
- The 'EM status' line provides additional information on the progress of the EM algorithm.
  - number of EM iterations (4,393 in the above example)
  - the ln-likelihood associated with the final estimate (-5183.56 in the above 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:

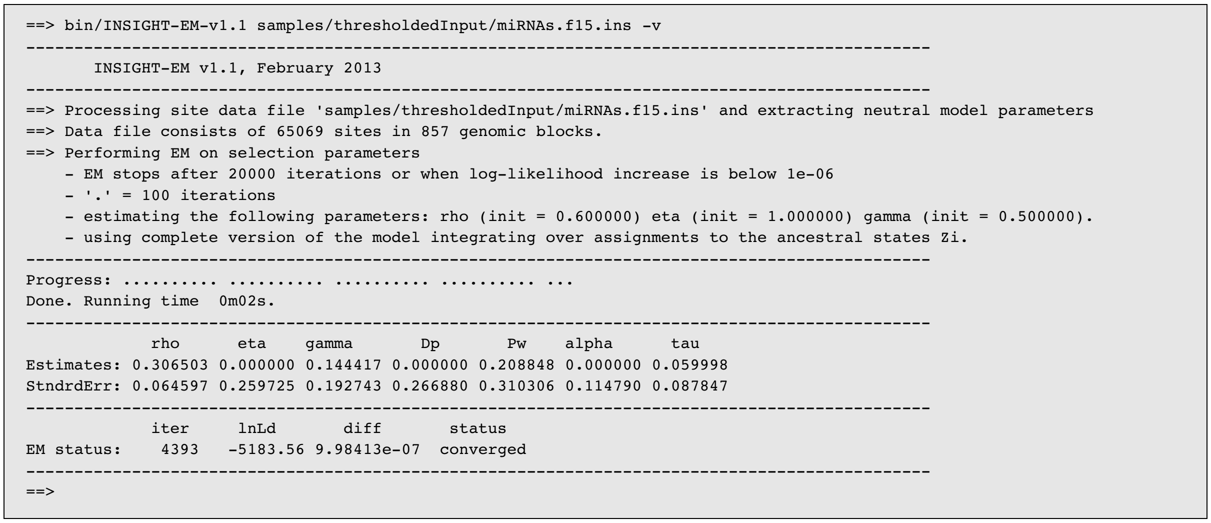


```
==> bin/INSIGHT-EM-v1.1 --help
+--------------------------------------------------------------------------------------
| INSIGHT-EM (v1.1)  - program for estimating selection parameters from poly/div patterns
+--------------------------------------------------------------------------------------
| Usage: ' bin/INSIGHT-EM-v1.1 infile [optional flags] '
|  infile contains a summary of sequence information across a given set of genomic positions
+--------------------------------------------------------------------------------------
| optional flags:
|
| ~~~~ General ~~~~
| -h      --help : show this usage message
| -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-v1.1 samples/thresholdedInput/miRNAs.f15.ins -v
----------------------------------------------------------------------------------------------
       INSIGHT-EM v1.1, February 2013
----------------------------------------------------------------------------------------------
==> Processing site data file 'samples/thresholdedInput/miRNAs.f15.ins' and extracting neutral model parameters
==> Data file consists of 65069 sites in 857 genomic blocks.
==> 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  0m02s.
----------------------------------------------------------------------------------------------
             rho      eta    gamma       Dp       Pw    alpha      tau
Estimates: 0.306503 0.000000 0.144417 0.000000 0.208848 0.000000 0.059998
StndrdErr: 0.064597 0.259725 0.192743 0.266880 0.310306 0.114790 0.087847
----------------------------------------------------------------------------------------------
             iter     lnLd       diff       status
EM status:    4393   -5183.56 9.98413e-07  converged
----------------------------------------------------------------------------------------------
==>
```

Verbose output of INSIGHT-EM

7.3 Altering the halting conditions

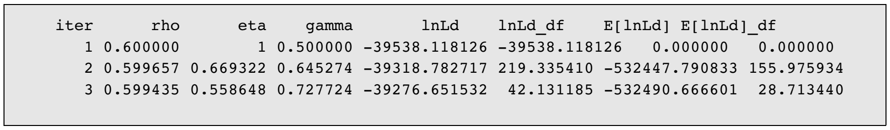
The EM algorithm halts if one of the following has occurred:

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

```
==> bin/INSIGHT-EM-v1.1 samples/thresholdedInput/miRNAs.f15.ins -v -r 0.1 -e 0.01 -g 10.0
---------------------------------------------------------------------------------------------
       INSIGHT-EM v1.1, February 2013
---------------------------------------------------------------------------------------------
==> Processing site data file 'samples/thresholdedInput/miRNAs.f15.ins' and extracting neutral model parameters
==> Data file consists of 65069 sites in 857 genomic blocks.
==> 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  0m02s.
---------------------------------------------------------------------------------------------
             rho      eta    gamma       Dp       Pw    alpha      tau
Estimates: 0.303322 0.000000 0.136035 0.000000 0.194683 0.000000 0.055916
StndrdErr: 0.064288 0.260753 0.195703 0.265155 0.309749 0.113528 0.087738
---------------------------------------------------------------------------------------------
             iter     lnLd       diff       status
EM status:    4909   -5183.56 9.99662e-07  converged
---------------------------------------------------------------------------------------------
==>
```

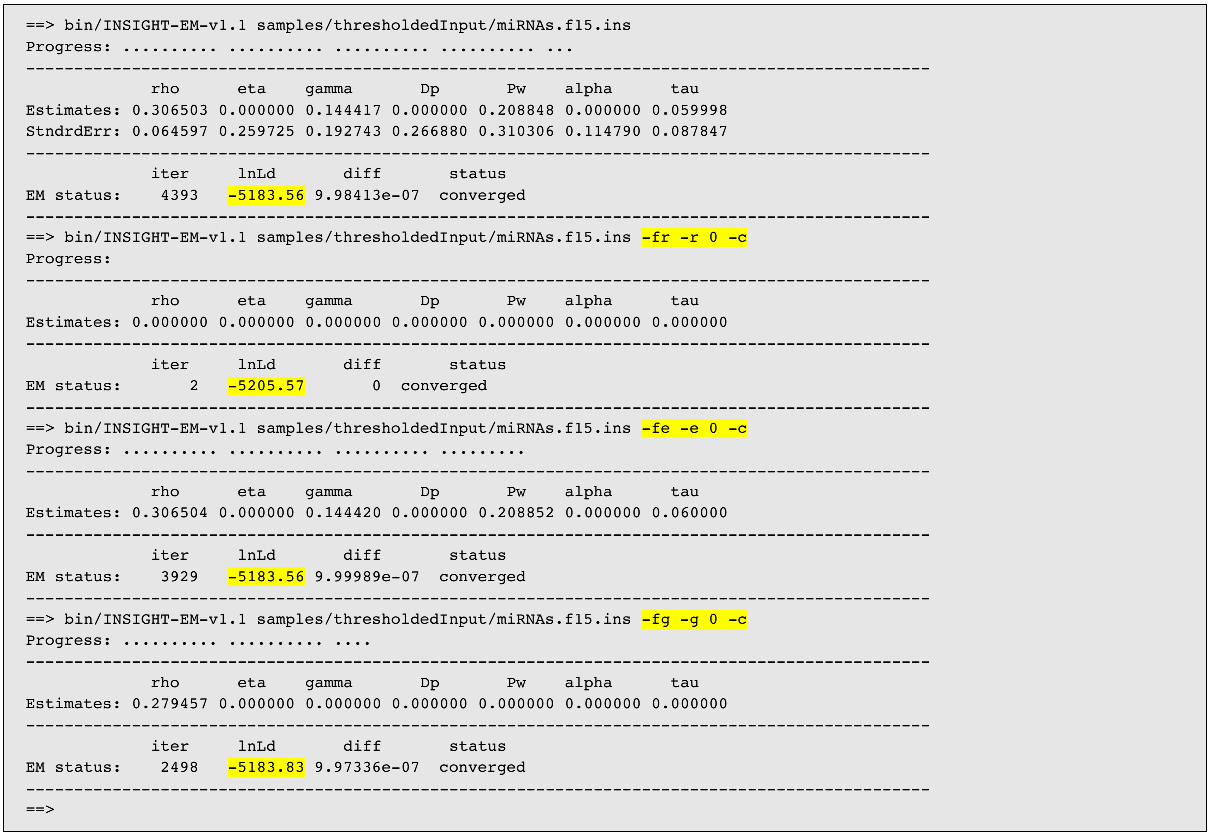
Alternative starting point for INSIGHT-EM

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 (η>0) or weak negative selection (γ>0), we suggest comparing to a χ^2 distribution with one degree of freedom, and for testing significant evidence for overall selection (ρ>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-v1.1 samples/thresholdedInput/miRNAs.f15.ins
Progress: .......... .......... .......... .......... ...
--------------------------------------------------------------------------------------------
             rho      eta    gamma       Dp       Pw    alpha      tau
Estimates: 0.306503 0.000000 0.144417 0.000000 0.208848 0.000000 0.059998
StndrdErr: 0.064597 0.259725 0.192743 0.266880 0.310306 0.114790 0.087847
--------------------------------------------------------------------------------------------
             iter     lnLd       diff       status
EM status:    4393   -5183.56 9.98413e-07  converged
--------------------------------------------------------------------------------------------
==> bin/INSIGHT-EM-v1.1 samples/thresholdedInput/miRNAs.f15.ins -fr -r 0 -c
Progress:
--------------------------------------------------------------------------------------------
             rho      eta    gamma       Dp       Pw    alpha      tau
Estimates: 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
--------------------------------------------------------------------------------------------
             iter     lnLd       diff       status
EM status:       2   -5205.57       0  converged
--------------------------------------------------------------------------------------------
==> bin/INSIGHT-EM-v1.1 samples/thresholdedInput/miRNAs.f15.ins -fe -e 0 -c
Progress: .......... .......... .......... .........
--------------------------------------------------------------------------------------------
             rho      eta    gamma       Dp       Pw    alpha      tau
Estimates: 0.306504 0.000000 0.144420 0.000000 0.208852 0.000000 0.060000
--------------------------------------------------------------------------------------------
             iter     lnLd       diff       status
EM status:    3929   -5183.56 9.99989e-07  converged
--------------------------------------------------------------------------------------------
==> bin/INSIGHT-EM-v1.1 samples/thresholdedInput/miRNAs.f15.ins -fg -g 0 -c
Progress: .......... .......... ....
--------------------------------------------------------------------------------------------
             rho      eta    gamma       Dp       Pw    alpha      tau
Estimates: 0.279457 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000
--------------------------------------------------------------------------------------------
             iter     lnLd       diff       status
EM status:    2498   -5183.83 9.97336e-07  converged
--------------------------------------------------------------------------------------------
==>
```

Likelihood ratio tests using INSIGHT-EM

These runs derive a test statistic of 2*(5205.57 - 5183.56) = 44.02 for the hypothesis that $\rho>0$, which implies an approximate P-value smaller than 0.00001 according to above table. Similarly, test statistics of 2*(5183.56 - 5183.56) = 0 and 2*(5183.83 - 5183.56) = 0.54 are associated with the hypotheses $\eta>0$ and $\gamma>0$ (resp.).

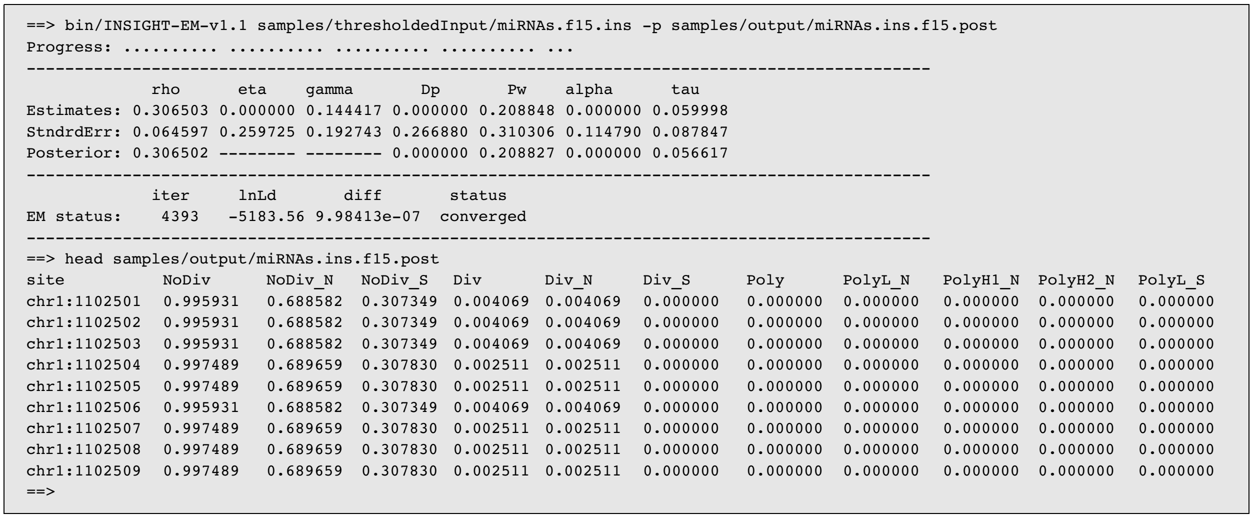
7.7 Approximate standard errors

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 (ρ=0 or ρ=1 or η=0 or γ=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 probabilities

It is possible to instruct INSIGHT-EM to output posterior probabilities for configurations of the hidden variables (A_i, Z_i, and S_i) at all sites. This mode is invoked by the **--post-cnt (-p)** option, followed by a file name to which to output the posterior distributions. First, the posterior probabilities allow us to produce estimates of the selection measures, ρ, E[D_p], and E[P_w], based on posterior counts. These should be very close to the direct estimates (psoterior-based estimates of E[D_p] and E[P_w] are typically slightly smaller than their direct estimates). The posterior probability table written into the output file contains 12 columns:

Column	**description**
site	id of the site, as given in the EM input file
NoDiv	Probability site is monomorphic non-divergent
NoDiv_N	Probability site is monomorphic non-divergent and neutral
NoDiv_S	Probability site is monomorphic non-divergent and under selection (NoDiv_N + NoDiv_S = NoDiv)
Div	Probability site is monomorphic divergent (NoDiv + Div = 1 for M sites)
Div_N	Probability site is monomorphic divergent and neutral
Div_S	Probability site is monomorphic divergent and under selection (Div_N + Div_S = Div)
Poly	Probability site is polymorphic (Poly = 1 for L and H sites)
PolyL_N	Probability site is polymorphic, with low derived allele frequency (<f) and neutral
PolyH1_N	Probability site is polymorphic, with intermediate derived allele frequency (in [f,1-f]) and neutral
PolyH2_N	Probability site is polymorphic, with high derived allele frequency (>1-f) and neutral
PolyL_S	Probability site is polymorphic, with low derived allele frequency (<f) and under selection (PolyL_N + PolyH2_N + PolyL_S = 1 for L sites and PolyH1_N = 1 for H sites)



```
==> bin/INSIGHT-EM-v1.1 samples/thresholdedInput/miRNAs.f15.ins -p samples/output/miRNAs.ins.f15.post
Progress: .......... .......... .......... .......... ...
-----------------------------------------------------------------------------------------------
             rho      eta    gamma       Dp       Pw    alpha      tau
Estimates: 0.306503 0.000000 0.144417 0.000000 0.208848 0.000000 0.059998
StndrdErr: 0.064597 0.259725 0.192743 0.266880 0.310306 0.114790 0.087847
Posterior: 0.306502 -------- -------- 0.000000 0.208827 0.000000 0.056617
-----------------------------------------------------------------------------------------------
             iter     lnLd       diff       status
EM status:    4393   -5183.56 9.98413e-07  converged
-----------------------------------------------------------------------------------------------
==> head samples/output/miRNAs.ins.f15.post
site           NoDiv      NoDiv_N   NoDiv_S   Div       Div_N     Div_S      Poly      PolyL_N    PolyH1_N  PolyH2_N   PolyL_S
chr1:1102501   0.995931   0.688582  0.307349  0.004069  0.004069  0.000000   0.000000  0.000000   0.000000  0.000000   0.000000
chr1:1102502   0.995931   0.688582  0.307349  0.004069  0.004069  0.000000   0.000000  0.000000   0.000000  0.000000   0.000000
chr1:1102503   0.995931   0.688582  0.307349  0.004069  0.004069  0.000000   0.000000  0.000000   0.000000  0.000000   0.000000
chr1:1102504   0.997489   0.689659  0.307830  0.002511  0.002511  0.000000   0.000000  0.000000   0.000000  0.000000   0.000000
chr1:1102505   0.997489   0.689659  0.307830  0.002511  0.002511  0.000000   0.000000  0.000000   0.000000  0.000000   0.000000
chr1:1102506   0.995931   0.688582  0.307349  0.004069  0.004069  0.000000   0.000000  0.000000   0.000000  0.000000   0.000000
chr1:1102507   0.997489   0.689659  0.307830  0.002511  0.002511  0.000000   0.000000  0.000000   0.000000  0.000000   0.000000
chr1:1102508   0.997489   0.689659  0.307830  0.002511  0.002511  0.000000   0.000000  0.000000   0.000000  0.000000   0.000000
chr1:1102509   0.997489   0.689659  0.307830  0.002511  0.002511  0.000000   0.000000  0.000000   0.000000  0.000000   0.000000
==>
```

Posterior probabilities

7.9 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.

```
==> bin/INSIGHT-EM-v1.1 -b samples/thresholdedInput/miRNAs.flankPoly.forBetas.f15.ins -v
----------------------------------------------------------------------------------------------
       INSIGHT-EM v1.1, February 2013
----------------------------------------------------------------------------------------------
==> Estimating the ratio beta1/(beta1+beta3) using neutral sites with low MAF.
----------------------------------------------------------------------------------------------
==> Processing site data file 'samples/thresholdedInput/miRNAs.flankPoly.forBetas.f15.ins' and lambda parameters
==> Performing EM on ratio beta1/(beta1+beta3) [ initial value = 0.5 ]
    - EM stops after 20000 iterations or when log-likelihood increase is below 1e-06
    - '.' = 100 iterations
----------------------------------------------------------------------------------------------
Progress:
Done. Running time  0m00s.

----------------------------------------------------------
           beta1
Estimates: 0.958698
----------------------------------------------------------------------------------------------
             iter     lnLd       diff       status
EM status:       6   -893.597 3.1966e-07  converged
----------------------------------------------------------------------------------------------
==>
```

Estimating the ratio between beta1 and beta3

8. Additional Scripts.

We provide several scripts that simplify usage of INSIGHT-EM. Mostly, these scripts allow the user to apply an L/H frequency threshold to a given input file (in which all polymorphic sites are indicated as 'P', and are given with allele counts), and run a complete analysis, including estimation of betas, selection parameters, and LRTs. **All scripts should be run in the root INSIGHT-EM directory** (alternatively, set the scriptDir variable in these scripts appropriately). File samples/README contains more details. Here we highlight a few main examples.

The script runINSIGHT-EM.sh is the main script that performs three steps, that are executed using the following three scripts.

The script getBetaEstimates.sh computes estimates for β_1, β_2, and β_3, given an EM input file for the flanking neutral polymorphic sites, and an L/H frequency cutoff. In the example below, we estimate beta parameters for the sample miRNAs using an L/H frequency cutoff of 15%. The log of INSIGHT-EM (for estimation of beta1/beta3) is written into the log file specified (samples/output/miRNAs.ins.log). The estimated betas in this example, β_1=0.767 β_2=0.200 and β_3=0.033, should be used when estimating selection on the same data when using a frequency cutoff of 15%.

```
==> bash scripts/getBetaEstimates.sh samples/baseInput/miRNAs.flankPoly.forBetas.ins 15 samples/output/miRNAs.ins.log
=================================================================
  Estimating beta1, beta2, and beta3 for freq threshold 15 on file samples/baseInput/miRNAs.flankPoly.forBetas.ins
=================================================================
1.  applying frequency threshold to input file
2.  estimating beta2 from L and H counts in flankPoly file
    beta2 = 0.199733
3.  estimating beta1 and beta3 by applying INSIGHT-EM to flank input file
    Invoking  ./bin/INSIGHT-EM-v1.1 -v -i 100000 -b samples/baseInput/miRNAs.flankPoly.forBetas.ins.f15
    INSIGHT-EM converged.
4.  getting beta line
    Beta line: beta 0.767214 0.199733 0.0330526
Done.
==>
```

Running script `getBetaEstimates.sh`

The script `getSelectionEstimates.sh` computes estimates for the selection parameters, including LRT statistics, given an L/H frequency cutoff and the appropriate beta parameters. In the example below, we plug in the beta parameters estimated above to estimate selection parameters and LRTs in the sample miRNAs using an L/H frequency cutoff of 15%. The logs of INSIGHT-EM (for all runs involved) are appended to the previous log file (samples/output/miRNAs.ins.log).

```
==> bash scripts/getSelectionEstimates.sh samples/baseInput/miRNAs.ins 15 0.767214 0.199733 0.0330526 \
          samples/output/miRNAs.ins.log
=================================================================
  Estimating selection parameters for freq threshold 15 on file samples/baseInput/miRNAs.ins
=================================================================
1.  applying frequency threshold to input file and adding beta line
2.  estimating selection parameters
    Invoking  ./bin/INSIGHT-EM-v1.1 -v -i 100000 samples/baseInput/miRNAs.ins.f15
    INSIGHT-EM converged.
3.  restricting rho=0 for LRT for selection
    Invoking  ./bin/INSIGHT-EM-v1.1 -v -i 100000 -fr -r 0.0 -c samples/baseInput/miRNAs.ins.f15
    INSIGHT-EM converged.
4.  restricting eta=0 for LRT for positive selection
    Invoking  ./bin/INSIGHT-EM-v1.1 -v -i 100000 -fe -e 0.0 -c samples/baseInput/miRNAs.ins.f15
    INSIGHT-EM converged.
5.  restricting gamma=0 for LRT for selection
    Invoking  ./bin/INSIGHT-EM-v1.1 -v -i 100000 -fg -g 0.0 -c samples/baseInput/miRNAs.ins.f15
    INSIGHT-EM converged.
Done.
==>
```

Running script `getSelectionEstimates.sh`

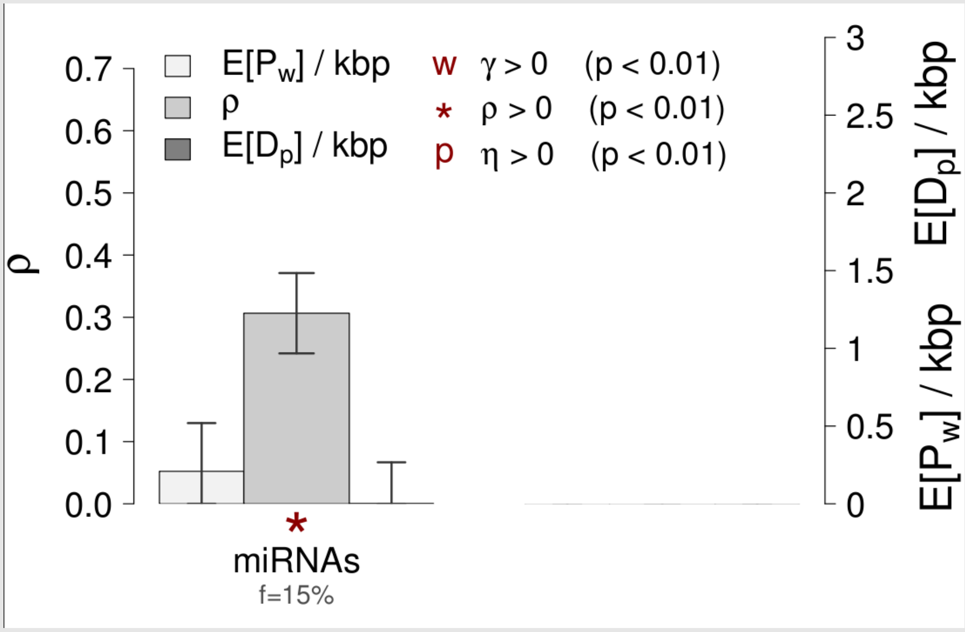
The script `processEMresults.sh` parses the log file produced by `getSelectionEstimates.sh`, and writes to a file a line containing estimates of all parameters of interest, standard errors, and LRT statistics associated with the three tests for selection (general, positive, and weak negative). By specifying '--header' instead of a log file, this script prints out the column names. The last column indicates the status of each of the four runs of the EM used in this analysis.

```
==> bash scripts/processEMresults.sh --header NONE samples/output/miRNAs.insight.results.txt
==> bash scripts/processEMresults.sh samples/output/miRNAs.ins.log miRNAs samples/output/miRNAs.insight.results.txt
==> cat samples/output/miRNAs.insight.results.txtmiRNAs /dev/stdout
dataID thres rho      rho_stderr  E[A]      E[A]_stderr E[W]     E[W]_stderr alpha      alpha_stderr tau      tau_stderr
             eta      eta_stderr  gamma     gamma_stderr lnLd     LRT[rho>0]  LRT[eta>0] LRT[gamma>0] em_status
miRNAs f=15% 0.306503 0.064597    0.000000  0.266880    0.208848 0.310306    0.000000   0.114790     0.059998 0.087847
             0.000000 0.259725    0.144417  0.192743    -5183.56 44.02       0          0.54         main-converged-rho0
-converged-eta0-converged-gam0-converged
==>
```

Running script `processEMresults .sh`

The R script `plotResults-INSIGHT.r` provides a graphical summary of INSIGHT results. Provided with a text table of INSIGHT results generated using `processEMresults.sh` and the name of a PDF file for the plot, this scripts generates bar charts with results of the analysis. Each line in the results table is summarized by three bars for $E[P_w]$, ρ, and $E[D_p]$ (in that order). The left axis provides the scale for ρ, and the right axis provides the scale of the expected counts $E[P_w]$ and $E[D_p]$. Below each bar corresponding to a value that is significantly greater than zero there is a significance indicator in red. If the results table contains multiple lines, then results are given for all lines in a single plot (organized in triplets of bars in sequence).

```
==> R  --vanilla < ./scripts/plotResults-INSIGHT.r —args \
       samples/output/miRNAs.insight.results.txt samples/output/miRNAs.insight.plot.pdf &> /dev/null
==> evince samples/output/miRNAs.insight.plot.pdf &
```



Graphical Presentation of INSIGHT Results Using `plotResults-INSIGHT.r`

9. Additional Samples.

Additional samples are available for download on the INSIGHT website http://compgen.bscb.cornell.edu/INSIGHT/.

10. Errors and Bugs.

INSIGHT-EM has been extensively tested on numerous simulated and real genomic data sets. However, it is a farily young program, and as such, is not free of bugs. 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>.