CHANCE user guide

Aaron Diaz DiazA2@humgen.ucsf.edu Abhinav Nellore anellore@gmail.com Jun S Song jssong@humgen.ucsf.edu

> **Ch**IP-seq **An**alytics and **C**onfidence **E**stimation (CHANCE) is a software for assessing the quality of ChIP-seq experiments and providing feedback for the optimization of ChIP and library generation protocols. This document is a brief guide to using the software and interpreting its results. If you find this software useful please cite A. Diaz, A. Nellore, J. S. Song, "CHANCE: comprehensive software for quality control and validation of ChIP-seq data". Genome Biology. Vol. 13 Issue 10, October 2012. For the theoretical analysis and technical interpretation of the main statistical tests used by CHANCE, see [Diaz et al. Statistical Applications in Genetics and Molecular Biology.11(3) March 2012] (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3342857/). For software downloads and the sample data referred to in this guide see: [http://songlab.ucsf.edu/CHANCE.html] (http://songlab.ucsf.edu/CHANCE.html) and for source code, wiki, bugs or requests see https://github.com/songlab/chance/

Installation

Installing executables

CHANCE runs under most 64bit Mac OSX, Windows 7, and Linux distributions. Start by downloading the appropriate installation package from: http://songlab.ucsf.edu/CHANCE.html or from http://github.com/songlab/chance/downloads

Installing CHANCE development tools

CHANCE is released under the GNU General Public License: http://www.gnu.org/licenses/. The CHANCE MATLAB toolbox, source code, development toolkit, and sample demos can be obtained from https://github.com/songlab/chance.

If you are running Mac OSX

1. Decompress the CHANCE archive
    1. Unzip the file `CHANCE_MacOS.zip` by double clicking the `CHANCE_MacOS.zip` icon.
    2. Open the folder `CHANCE_MacOS/`.
2. Install MCR, the MATLAB Compiler Runtime:
    1. Unzip `MCRInstaller.zip` by double clicking its icon
    2. Double click `InstallForMacOSX`
    3. Follow the on screen instructions, but keep track of the install location if you change the default.
3. To start CHANCE:
    1. Double click the chance icon
    2. To start CHANCE from the command line:
        1. Navigate to the `CHANCE_MacOS` folder
        2. Execute `./run_chance.sh path_to_mcr`, where `path_to_mcr` is the path to the MCR you installed. The default path is `/Applications/MATLAB/MATLAB_Compiler_Runtime/v717/`
4. NOTES: Drag or copy the CHANCE icon (`chance.app`) to the `/Applications` folder or any other location if you like. If you want to start chance from the command line the shell script `run_chance.sh` needs to be in the same folder as `chance.app`. The first time you start CHANCE it will take a little longer than usual to start since CHANCE needs to configure the MCR.

If you are running 64bit Linux:

1. Navigate to where you downloaded `CHANCE_Linux.zip`
2. Decompress the CHANCE archive

   ```
   unzip CHANCE_Linux.zip
   cd chance_linux
   ```

3. Install MCR, the MATLAB Compiler Runtime:

   ```
   unzip MCRInstaller.zip
   sudo ./install
   ```

   Follow the on screen instructions, keep track of the install location if you change the default
4. To start CHANCE: `./run_chance.sh path_to_mcr` where `path_to_mcr` is the path to the MCR you installed,the default is `/usr/local/MATLAB/MATLAB_Compiler_Runtime/v717/`
5. The first time you start CHANCE it will take a little longer than usual to start since CHANCE needs to configure the MCR.

If you are running 64bit Windows 7

1. Double click the installer executable `CHANCE_Windows.exe`
2. To start CHANCE: double click `chance.exe`.
3. The first time you start CHANCE it will take a little longer than usual to start, since CHANCE needs to configure the MCR.

Using CHANCE

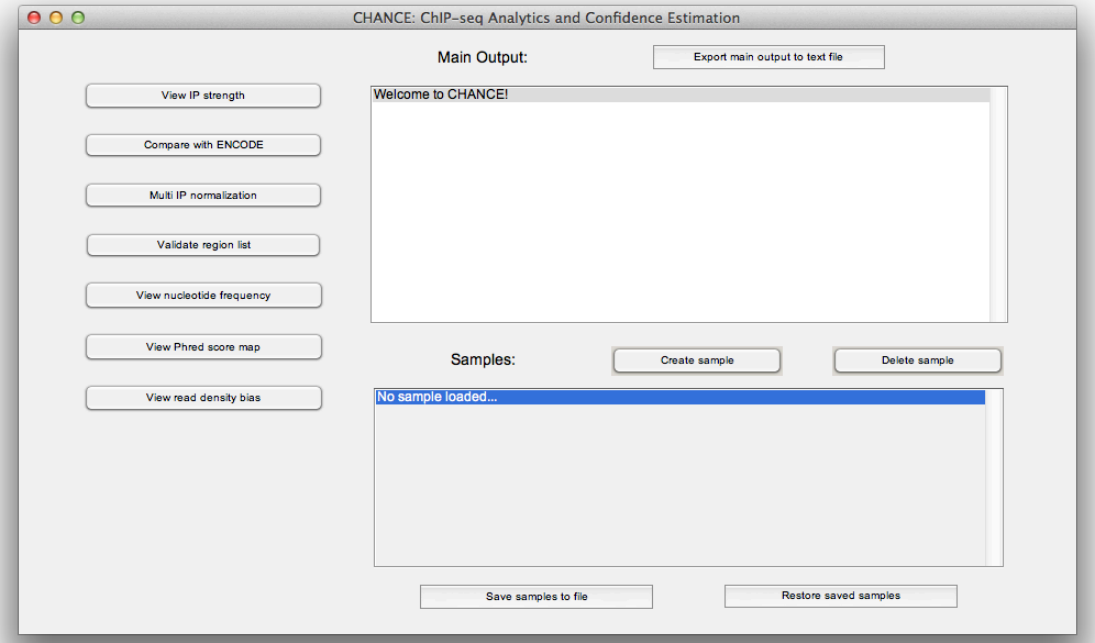

Figure 1

The main window of CHANCE, Figure main, has three parts: at the top right is an output window that logs the session, at the bottom right is a list of samples in your workspace, CHANCE's quality controls can be accessed on the left. The session log can be exported to a text file by clicking "Export main output to text file", Figure 2.

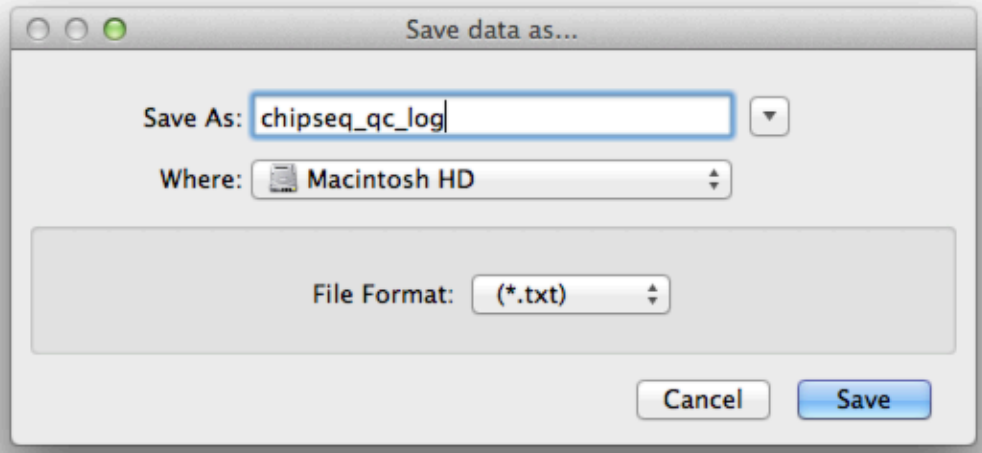

Figure 2

Adding samples to the workspace:

Creating new samples from mapped reads.

CHANCE works with reads mapped to a reference genome from IP and control (Input) samples. It can import reads in `BED`, `tagAlign` , `SAM`, and `BAM` formats, as well as `BOWTIE` output. CHANCE works with reads mapped to mm9, hg18, hg19, and tair10.

*NOTE: CHANCE expects chromosome identifiers to match the standard identifiers for a given build. For a list of valid chromosome identifiers, please see Appendix 1.

For example to generate a sample from the provided Broad Institute H1 HESC H3K4me3 `BAM` file \ `wgEncodeBroadHistoneH1hescH3k4me1StdAlnRep1.bam`:

1. Click the “Create Sample” button in the middle of the CHANCE window.
2. Navigate to the `sample_data/` folder
3. Select `wgEncodeBroadHistoneH1hescH3k4me1StdAlnRep1.bam` and click Open
4. When prompted enter the type of the file to be read (bam) and the build, Figure 4.

5. When prompted enter a name for the sample, Figure 5.

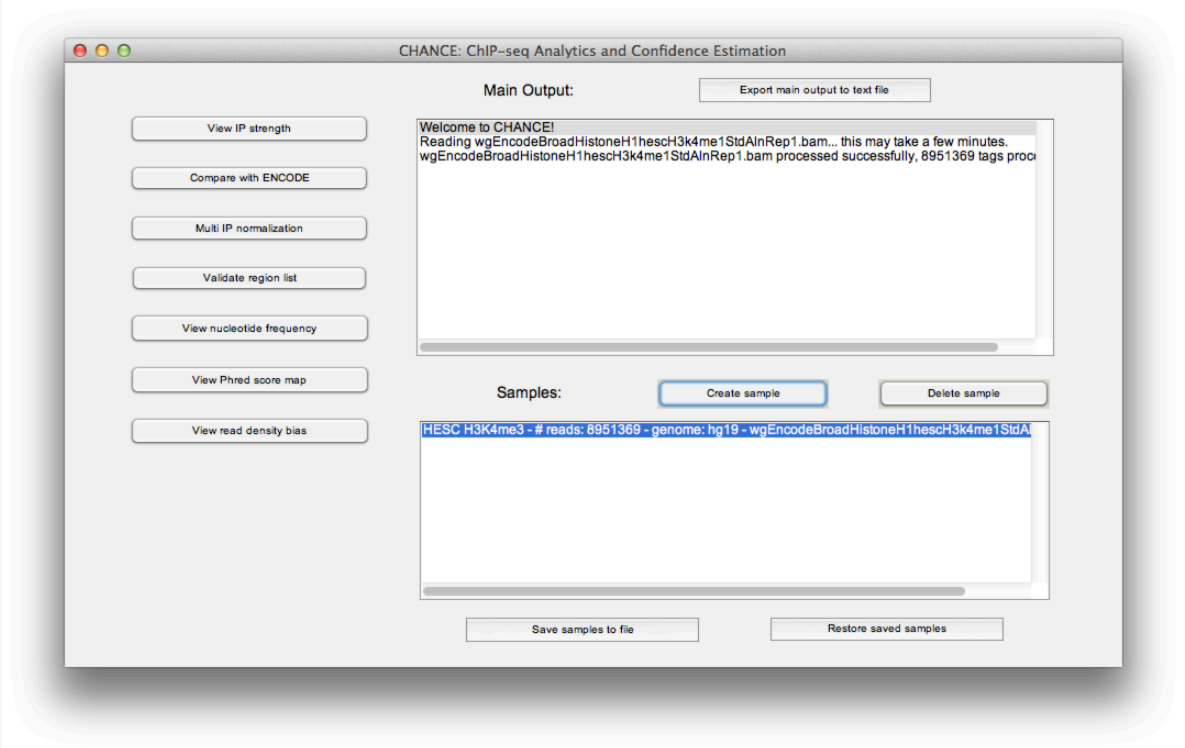

Figure 3

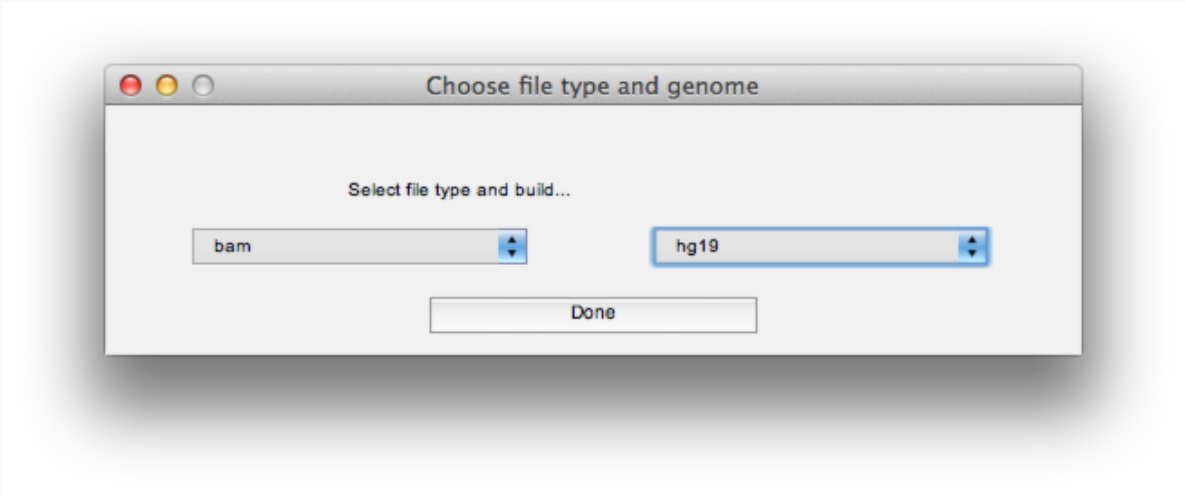

Figure 4

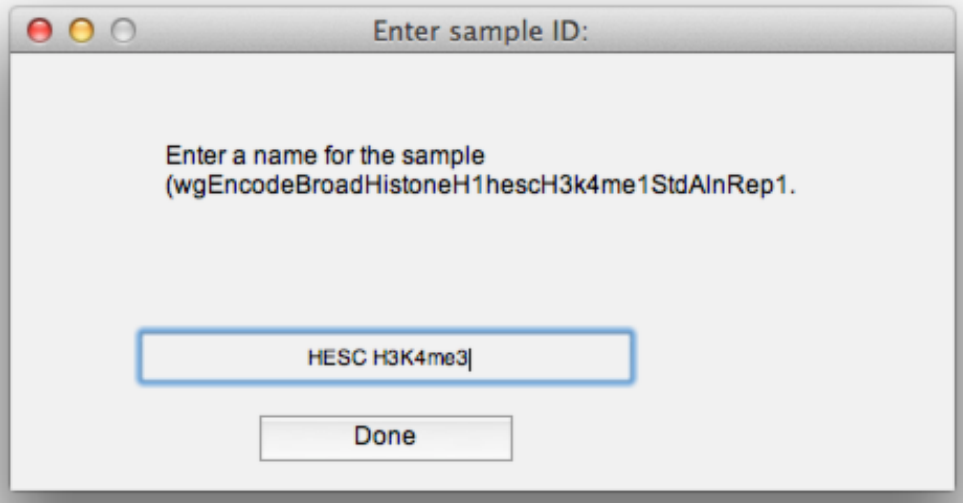

Figure 5

Restoring samples from a previous session

To reload samples from an old session, click “Restore saved samples”. For example:

1. Click the “Restore saved samples” button
2. Navigate to the `sample_data/` directory
3. Select `broad_data.mat` and click open (This contains H3K4me1,2,3, H3K36me3, H3K27me3, and Input samples in HESC, from Broad via ENCODE.), Figure 6

Samples can be deleted and all the samples in the current workspace can be written to file using the “Delete sample” and “Save samples to file”, Figure 7.

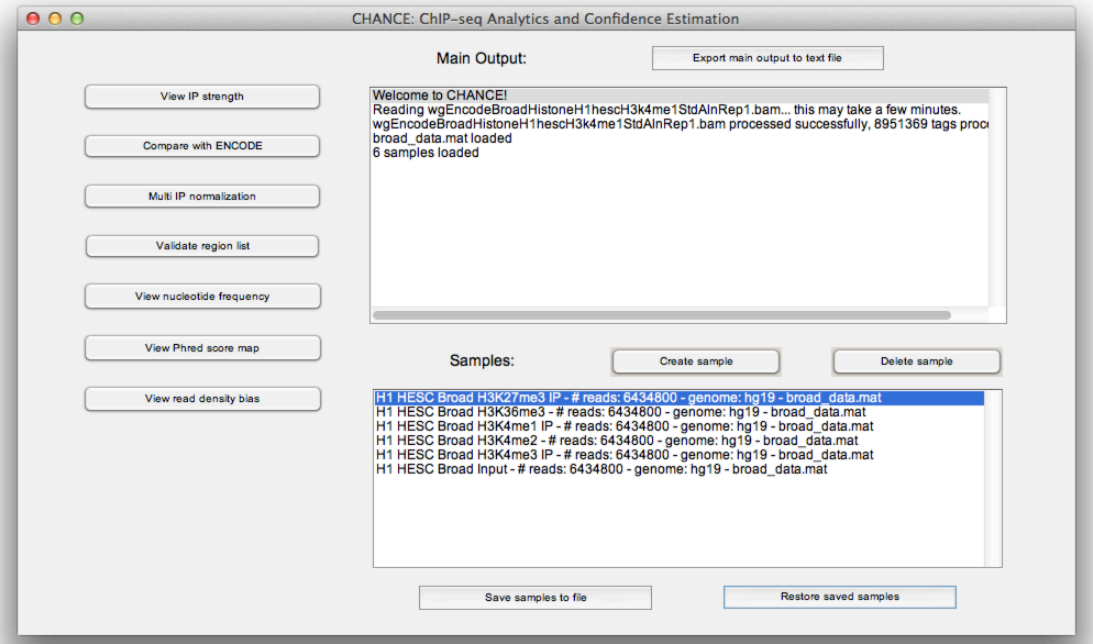

Figure 6

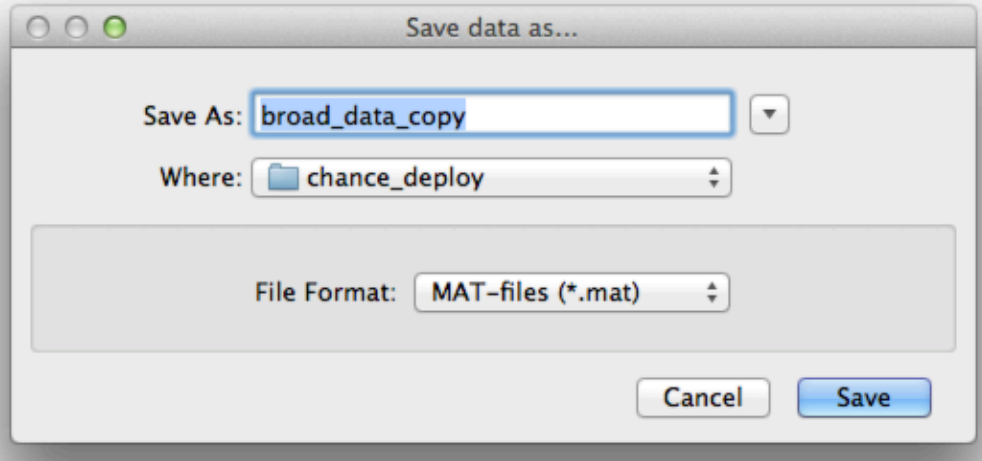

Figure 7

Checking the strength of enrichment in the IP

One of the primary uses of CHANCE is to check the strength of the IP. To begin, select a sample to be used as a primary sample. If you are doing comparison of ChIP to Input control then select the the IP as your primary sample. To compare to arbitrary samples select the sample you want to use as treatment. Then click “View IP strength,” and select the Input or control sample. For example:

1. Select an IP sample (H1 HESC Broad H3K4me3 IP) in the “Samples:” window.
2. Click the “View IP strength” button.
3. Select a matching Input sample (H1 HESC Broad Input) in the drop down dialog box.
4. CHANCE will now spawn 3 windows: a summary statement Figure 8, an enrichment plot Figure 9, and a second linearization plot Figure 10.

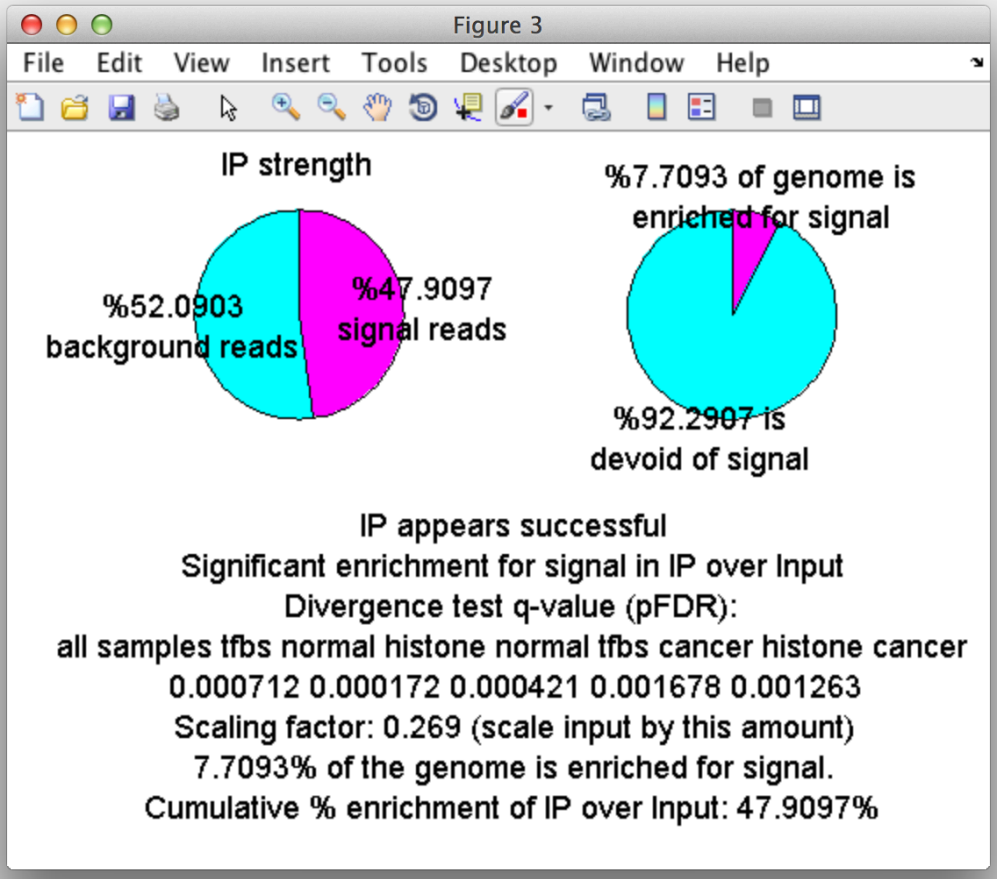
Figure 3
File Edit View Insert Tools Desktop Window Help
IP strength
%52.0903
background reads
%47.9097
signal reads
%7.7093 of genome is
enriched for signal
%92.2907 is
devoid of signal
IP appears successful
Significant enrichment for signal in IP over Input
Divergence test q-value (pFDR):
all samples tfbs normal histone normal tfbs cancer histone cancer
0.000712 0.000172 0.000421 0.001678 0.001263
Scaling factor: 0.269 (scale input by this amount)
7.7093% of the genome is enriched for signal.
Cumulative % enrichment of IP over Input: 47.9097%

Figure 8

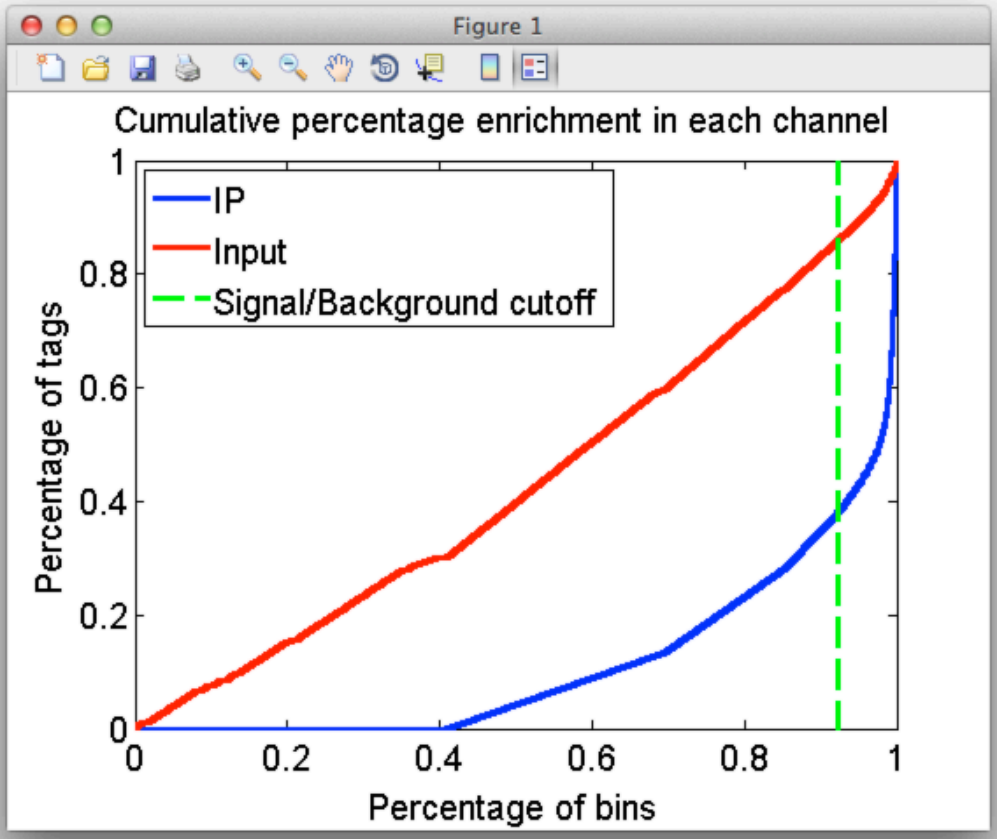
Figure 1
Cumulative percentage enrichment in each channel
IP
Input
Signal/Background cutoff
Percentage of tags
Percentage of bins
0
0.2
0.4
0.6
0.8
1

Figure 9

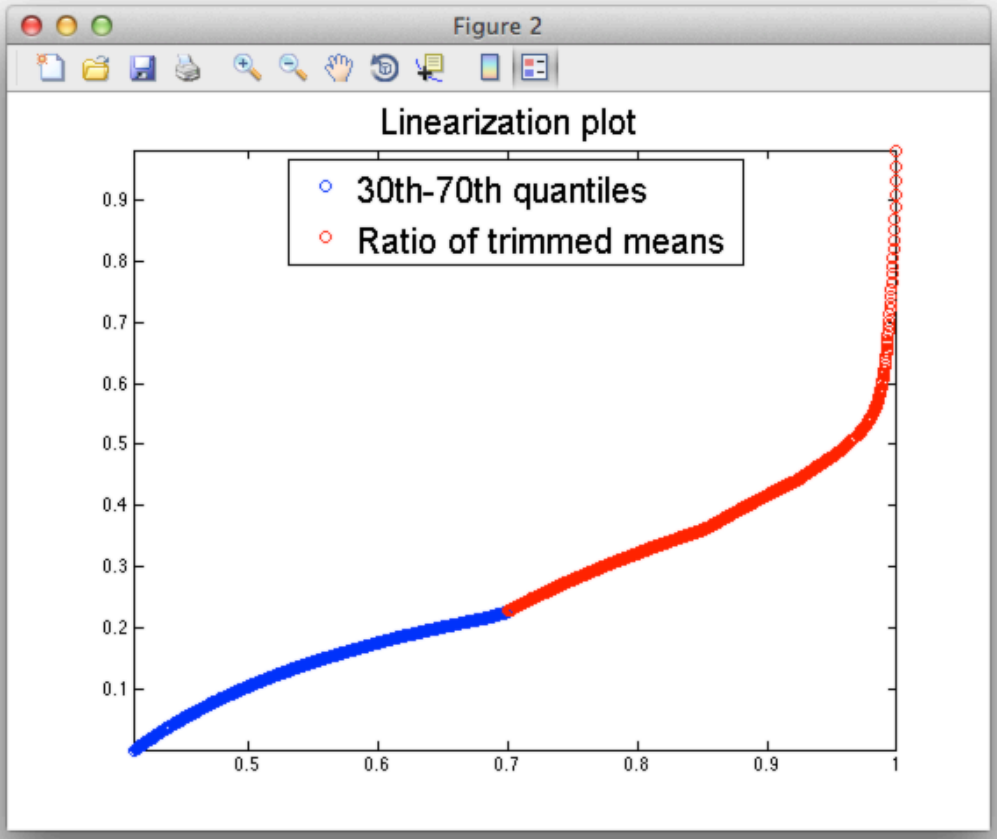

Figure 10

Begin your analysis with the summary statement which will give you an estimate of the percentage of the IP reads which map DNA fragments pulled down by the antibody used for the ChIP. In addition to the size of this *signal component* within the IP CHANCE reports the fraction of the genome these signal reads cover, as well as the statistical significance of the genome wide percentage enrichment relative to control in the form of a q-value (positive false discovery rate). CHANCE has been trained on CHIP-seq experiments from the ENCODE repository by making over 10,000 Input to IP and Input to replicate Input comparisons. The q-value reported gives then the fraction of comparisons between Input sample techinical replicates that report an enrichment for signal in one sample compared to another equal to the user provided sample or greater.

Note that in addition to the "divergence test q-value", CHANCE also reports several additional numbers bellow. Each of these numbers is also a q-value, corresponding to a different categorey

of data. ChIP-seq experiments from cancer samples can sometimes have amplified copy number of backgound regions. This can lead to an oversampling of these amplified regions and an increase in the number of reads mapping to amplified loci. This increase in tag density can read like signal in some extreme cases. Similarly, ChIP enrichment profiles for histone modification marks are generally shallower and more difuse than profiles of transcription factors which tend to be more punctate. For these reasons there will tend to be higher false discovery rates when analyzing cancer samples or samples with a ChIP for a histone mark. CHANCE reports secondary false discovery rates which the user can compare to their own sample. The abbreviations are:

1. TN: transcription factor ChIP - normal cells
2. HN: histone mark ChIP - normal cells
3. TC: transcription factor ChIP - cancer cells
4. HC: histone mark ChIP - cancer cells

CHANCE uses Signal Extraction Scaling (SES) to estimate the portion of the genome where the IP channel is distributed in the same way as the Input (in terms of cumulitive percentage tag density). On this *background* component of the genome the behavior of the cumulative percentage tag densities in the IP and Input channels can be visually inspected in the enrichment and linearization plots. On both plots the x-axis represents percentage of the genome as a function of increasing tag density (percentage of the genome with a given number of tags per kbp or less). In other words as we move from left to right we consider regions of the genome with larger and larger tag counts. An ideal linearization plot will initially look like a flat line since when confined to background regions of relatively low tag count the IP distribution behaves just like the control sample, allocating percentages of its total tag count nearly uniformly to the quantiles of tag count. In regions of tag density sufficiently large the percentage of tags found in regions of differential tag count will be much larger for the IP sample since ChIP enriched regions with greater tag density occur with greater genome wide frequency. This will cause the linearization plot to vier up. On the other hand the ideal enrichment plot will show an increasing difference between Input and IP cumulative tag density since at each quantile of tag count regions with tag density of that count or less occupy a larger percentage of the genome for the Input sample compared to the IP. The point of maximal distance therefore indicates the largest bimodal partition of low count background regions vs. high tag count signal regions achievable by triaging regions by tag density. Note that both curves on the enrichment plot must begin at 0 and end at 1.

Things to look for:

- If the difference between IP and Input tag density is very shallow across the whole graph or if the linearization plot looks completely like a straight line across the whole graph then that indicates a weak IP. Since this occurs when the IP tag density matches the Input channel, ie.

the IP and Input are barely distinguishable, in cumulative percentage distribution.

- In the enrichment plot: if either the IP or Input curves are zero as you move from left to right in the along the x-axis for a large portion, then this indicates a lack of sequencing depth in the corresponding sample since that implies a large number of regions with zero coverage, deflating the cumulative tag density. For example, Figure 7 demonstrates this type of zero-inflation in the IP channel. CHANCE will auto-detect when a channel is severely zero-inflated and issue a warning.
- Another phenomenon which frequently occurs is when there are a large number repeat reads mapping to the same location, often induced by PCR amplification bias during library creation. You can see this in the enrichment plot by an Input curve which is depressed for most x and only rises sharply near x=1. If you encounter this you can often increase the statistical power of your data by a bioinformatic de-duplication of your reads.
- If the q-value is greater than 0.05 in all categories, then the percentage enrichment and percent of the genome enriched estimations are not meaningful and CHANCE does not report them.
- When checking false discovery rates for a sample, one should expect to see slightly higher false discovery rates in cancer samples and in histone mark samples.

Note that CHANCE also reports the factor by which to scale the Input channel when comparing the enrichment profile to the IP, for example when viewing the enrichment profile in a genome browser such as, IGB, IGV, or the ENCODE Genome Browser.

Multi-IP normalization

Multi-IP normalization can be used when performing differential analysis of more than two samples, for example when looking for co-localization of multiple transcription factors. The purpose of this module is to:

- Compute scaling factors for comparing multiple enrichment profiles in a genome browser such as IGB, IGV, or the UCSC Genome browser, or for use in downstream differential analysis.
- Identify batch effects in replicate samples.
- Estimate the percentage of the genome differentially enriched for between two samples.

To get started:

1. Click the “Multi IP normalization” button.
2. Enter the number of experiments to compare in the dialog box, Figure
    1.
3. Select, one by one, the experiments to include in the drop down dialog box.
4. CHANCE will now spawn 3 windows: a summary statement Figure 12, a differential enrichment comparison matrix Figure 13, and an enrichment plot Figure 14.

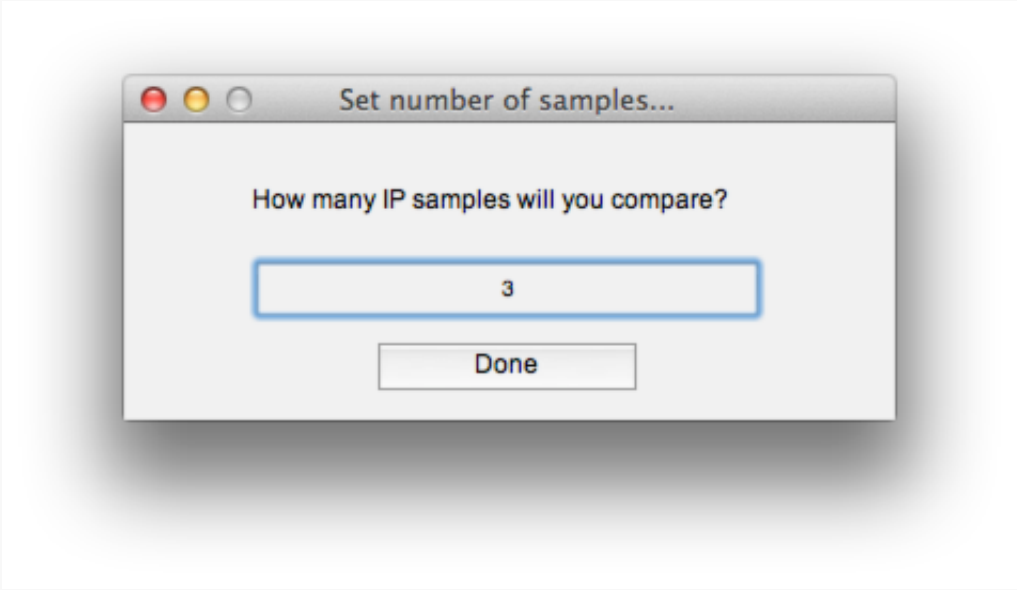

Figure 11

For multiple IP differential analysis, CHANCE first normalizes each sample to the mean read depth over all samples considered. CHANCE then forms a consensus sample using a multi-channel signal combiner. This has the effect of determining a consensus whose background component will be the largest possible subset of the genome of mutual background for all original samples. Lastly, SES is used to determine differential enrichment of each sample from the consensus, as well as the pairwise differential comparisons between samples.

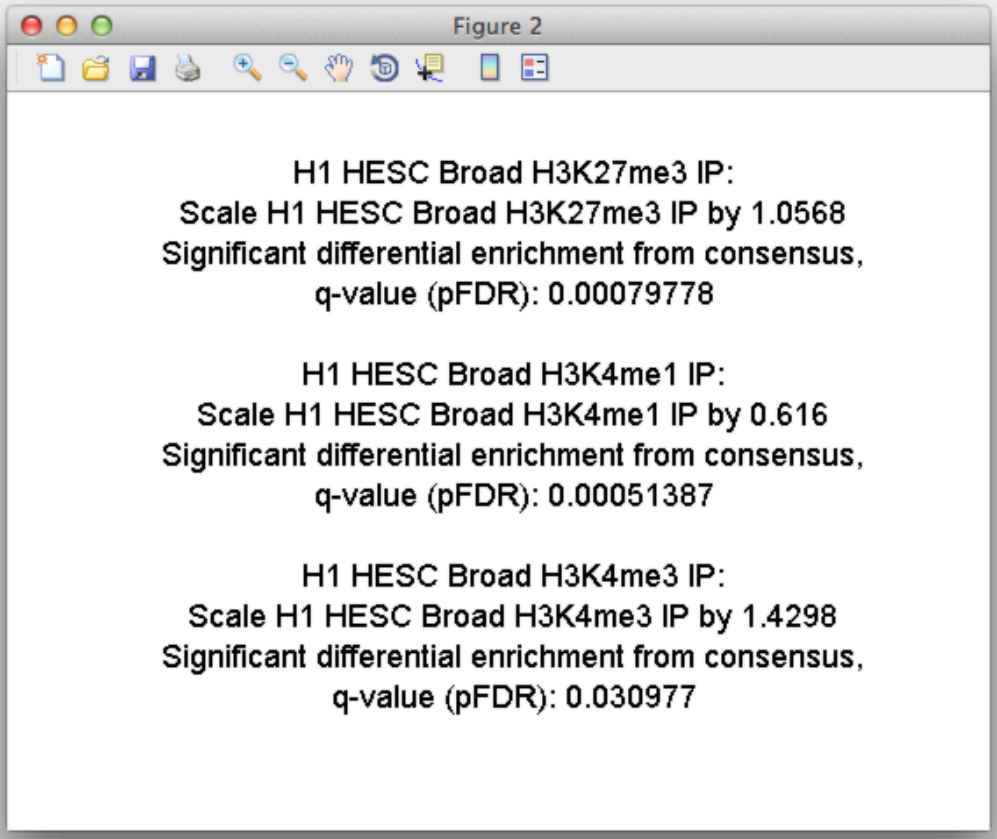
Figure 2
H1 HESC Broad H3K27me3 IP:
Scale H1 HESC Broad H3K27me3 IP by 1.0568
Significant differential enrichment from consensus,
q-value (pFDR): 0.00079778
H1 HESC Broad H3K4me1 IP:
Scale H1 HESC Broad H3K4me1 IP by 0.616
Significant differential enrichment from consensus,
q-value (pFDR): 0.00051387
H1 HESC Broad H3K4me3 IP:
Scale H1 HESC Broad H3K4me3 IP by 1.4298
Significant differential enrichment from consensus,
q-value (pFDR): 0.030977

Figure 12

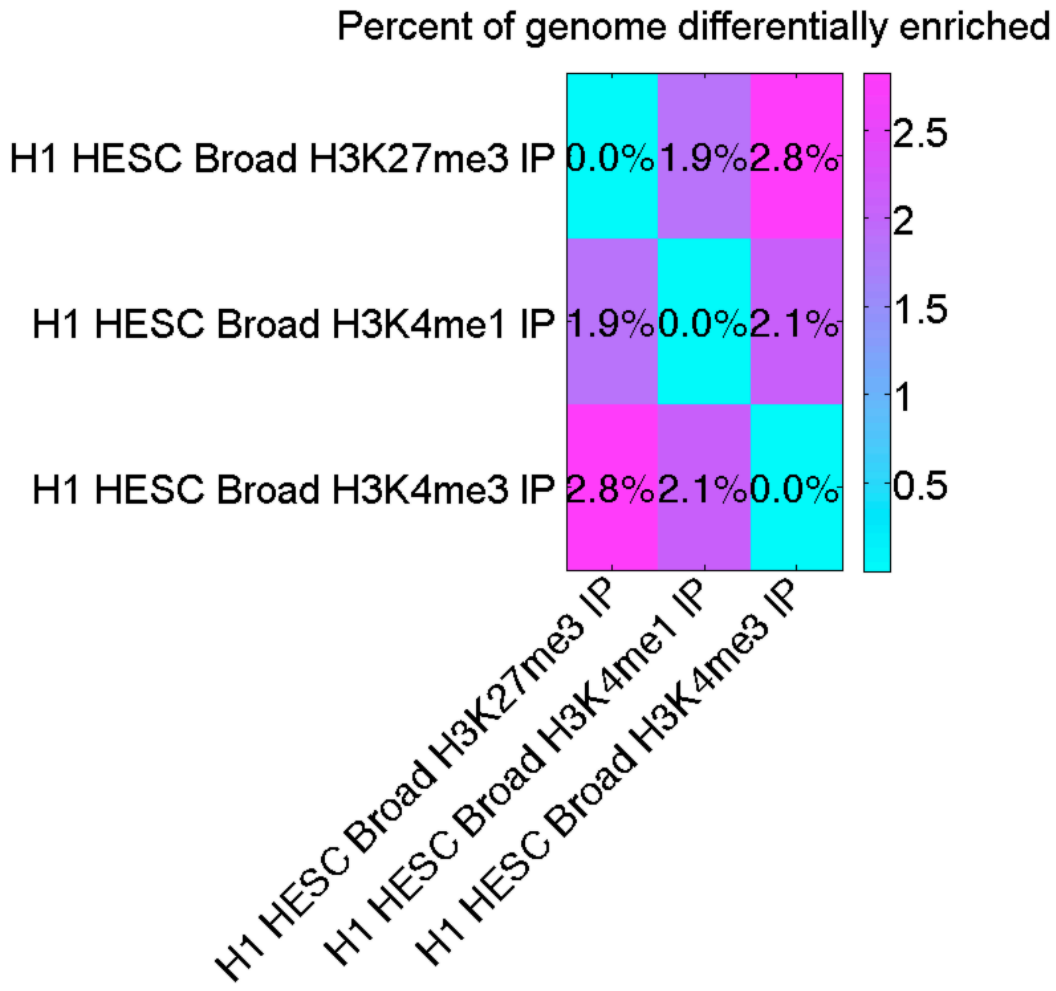
Percent of genome differentially enriched
H1 HESC Broad H3K27me3 IP
H1 HESC Broad H3K4me1 IP
H1 HESC Broad H3K4me3 IP
0.0%
1.9%
2.8%
1.9%
0.0%
2.1%
2.8%
2.1%
0.0%
2.5
2
1.5
1
0.5
H1 HESC Broad H3K27me3 IP
H1 HESC Broad H3K4me1 IP
H1 HESC Broad H3K4me3 IP

Figure 13

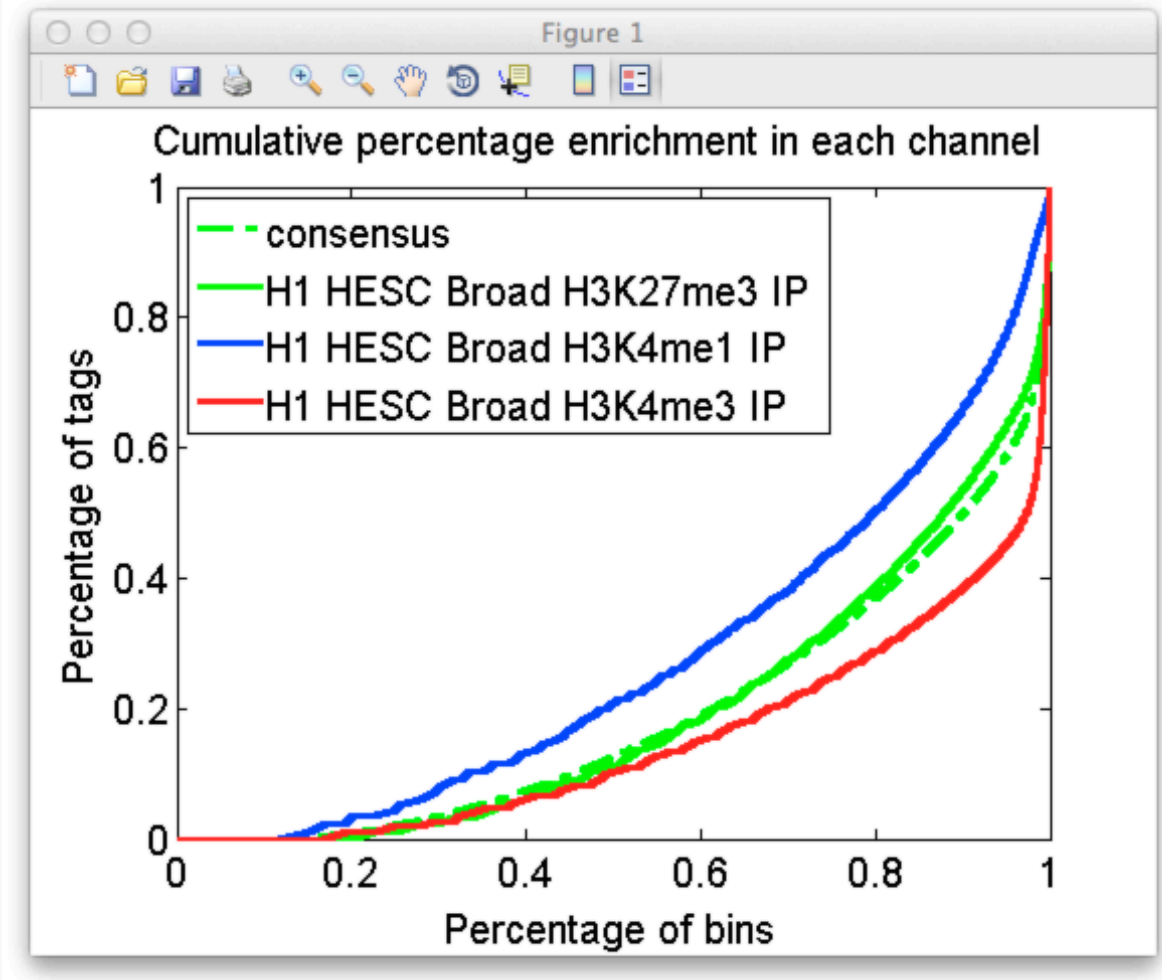

Figure 14

Things to look for:

1. The summary statement gives scaling factors to use when viewing the enrichment profiles of the samples collectively, say on IGB, IGV, or the UCSC Genome Browser, Figure 11.
2. Technical replicates should show no differential enrichment in the matrix. If CHANCE detects significant differential enrichment in replicates this may indicate a batch effect.

Compare with ENCODE

This feature compares your ChIP-seq experiment with similar experiments from the ENCODE

repository. This comparison checks the fold change in IP/Input read count in peak regions defined as the union of all ENCODE peaks for your experiment type in your organism/build. For example:

1. Select an IP sample (H1 HESC Broad H3K4me3 IP) in the “Samples:” window.
2. Click the “Compare with ENCODE” button.
3. Select a matching Input sample (H1 HESC Broad Input) in the drop down dialog box, Figure 15.
4. Select the transcription factor or epigenetic mark type from the drop down dialog box.
5. CHANCE will now spawn a plot window:

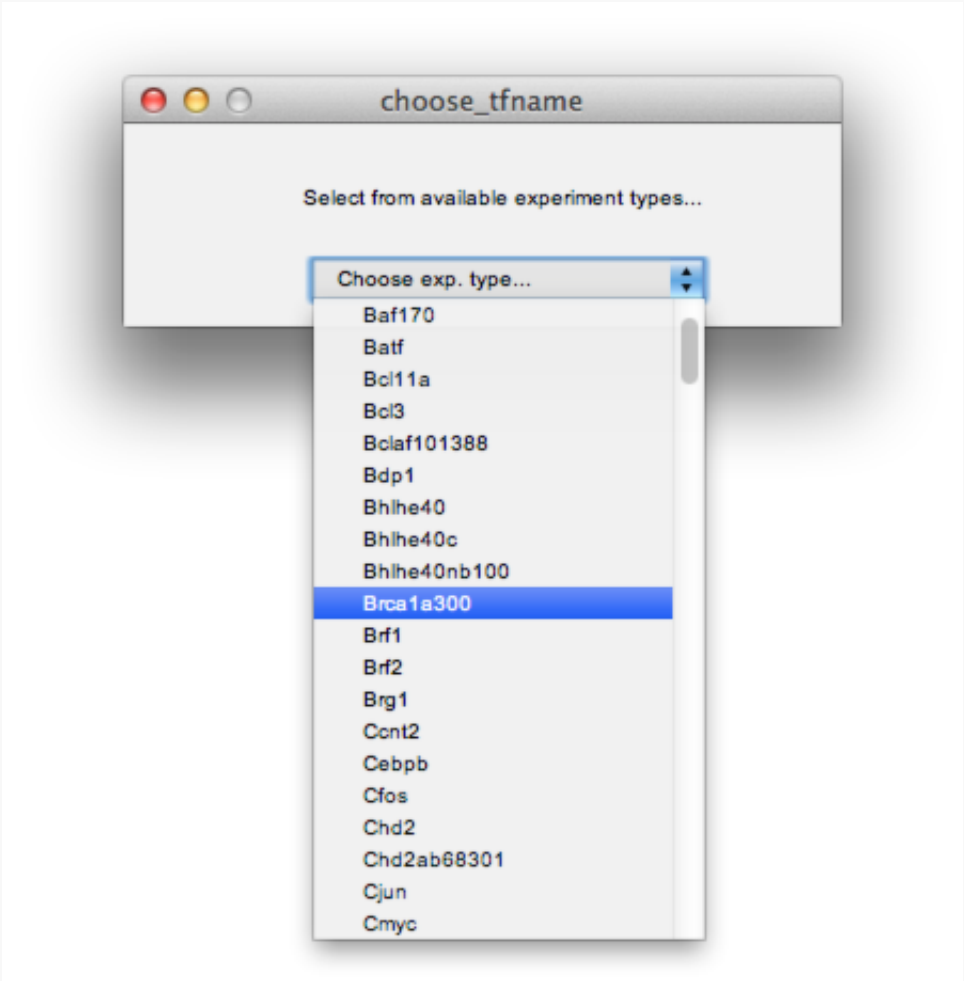

Figure 15

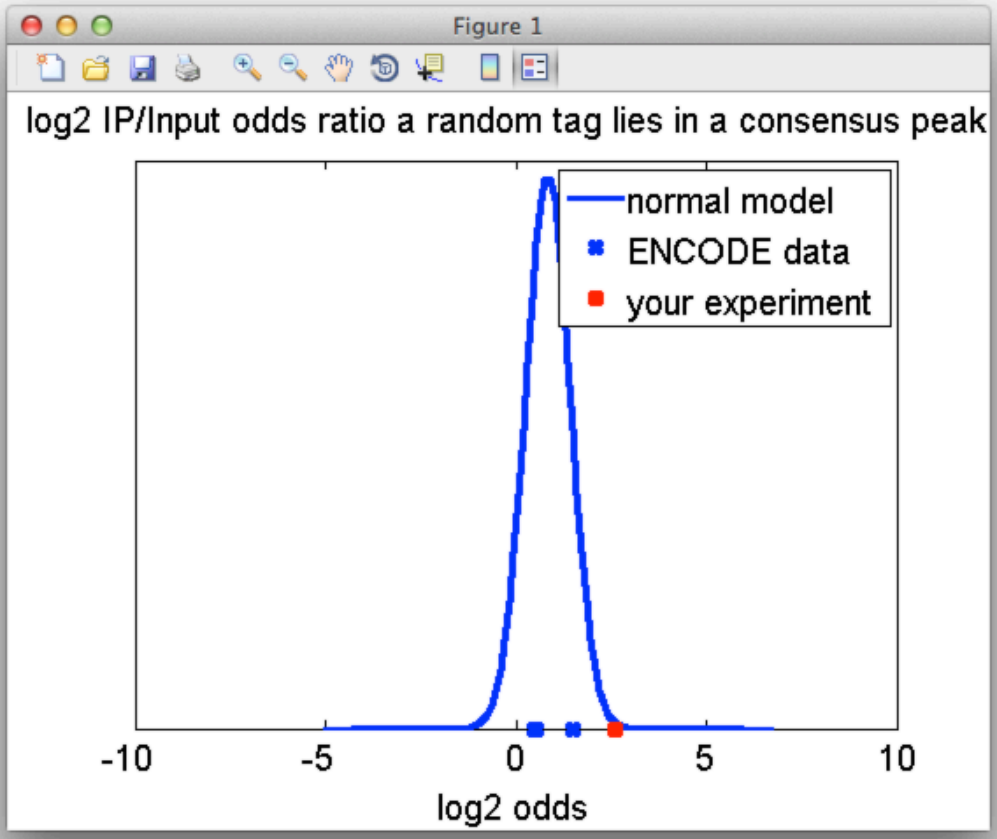

Figure 16

The blue circles denote IP samples from ENCODE, and the red star is your sample. The blue bell curve is a probability model fitted to all available data. When the red star lies among the blue circles, close to the center of the blue bell curve, your experiment resembles the experiments from the ENCODE repository, in the sense that there is a similar odds of finding enrichment in your sample as in other ENCODE samples in the union of all ENCODE peaks. The probability reported in the main window can be interpreted as the fraction of experiments from ENCODE which have less IP enrichment over Input the union set than your experiment. Note that disagreement with ENCODE in this fashion is not definitive of a failed experiment. Transcription factor binding and epigenetic state can be highly dynamic and cell type specific and different antibodies may have been used for the same protein of interest. None the less this test can give you a sense as to what extent your data constitutes a statistical outlier when compared to others.

Validate region list

Often one spot validates ChIP enrichment by PCR amplification of positive control regions. It is natural to ask if the enrichment detected by PCR is present in the sequencing data. CHANCE allows the user to spot check an arbitrary set of regions. To get started:

1. First construct a tab delimited text file with one line per control region. Each line should be tab separated with the following format: 'chrom start stop ID' Where `chrom` is a chromosome identifier matching one of the chomosome identifiers in the original file of reads from which the sample was generated, for example: `chr1` or `chrY`. `start` and `stop` are integers giving the genomic coordinates where the defined region starts and stops. `ID` is any user string which will identify the region, for example the name of a gene. In the provided sample data the file `gene_promoter_list.txt` is an example of such a file.
2. Select an IP sample in the “Samples:” window.
3. Click the “Validate region list” button.
4. Select a matching Input sample in the drop down dialog box.
5. CHANCE will now spawn a plot window.

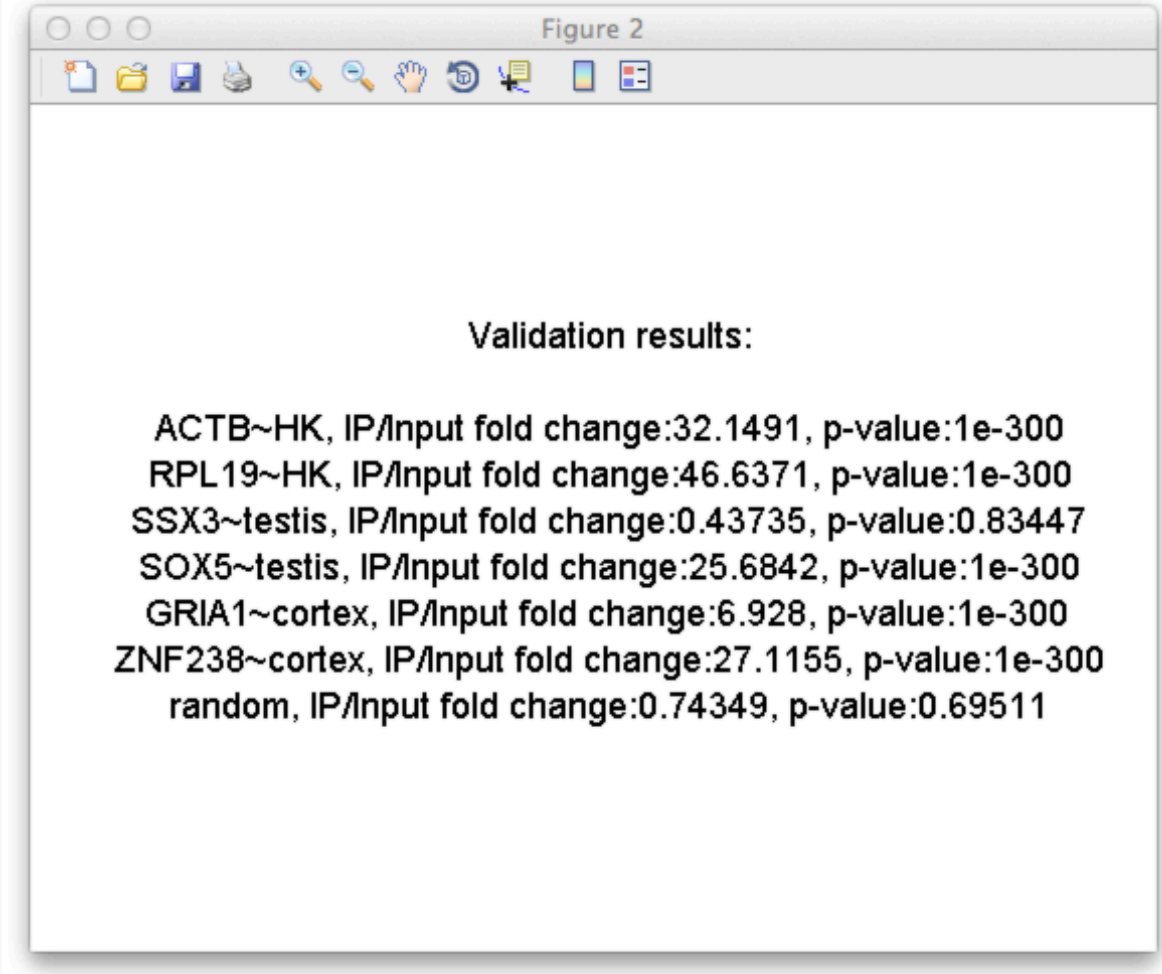
Figure 2
Validation results:
ACTB~HK, IP/Input fold change:32.1491, p-value:1e-300
RPL19~HK, IP/Input fold change:46.6371, p-value:1e-300
SSX3~testis, IP/Input fold change:0.43735, p-value:0.83447
SOX5~testis, IP/Input fold change:25.6842, p-value:1e-300
GRIA1~cortex, IP/Input fold change:6.928, p-value:1e-300
ZNF238~cortex, IP/Input fold change:27.1155, p-value:1e-300
random, IP/Input fold change:0.74349, p-value:0.69511

Figure 17

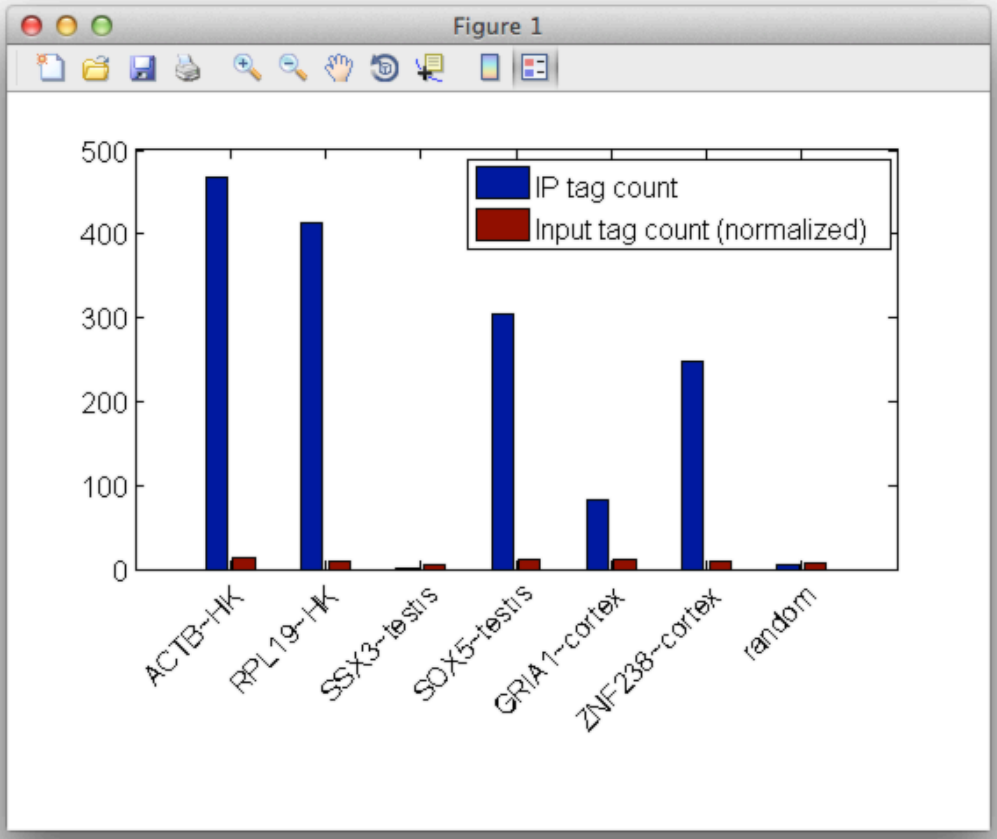

Figure 18

Things to look for:

1. The resulting bar graph gives the tag count in each channel in the defined region as well as a p-value indicating the statistical significance of the fold change in read count over the defined region.

2. Although the Poisson null model is widely used the Poisson null model is often artificially low due to local bias in read density . The Poisson model is used here as a simple quick check of enrichment and this module should not be used for peak calling.

3. Often comparison to a negative control can be informative. If you have ChIPed for a transcription factor for example, choosing a region which is known *not* to interact with your

protein of interest or a random region as a negative control is a good idea.

View nucleotide frequency and View Phred score map

These features help you detect read quality and content biases in your data. CHANCE constructs two plots from the sample data: the first is of nucleotide frequency vs. base position, and the second is of frequency of uncallable bases vs. base position. For example:

1. Select a sample (H1 HESC Broad H3K4me3 IP) in the “Samples:” window.
2. Click the “View nucleotide frequency” button.
3. CHANCE will now spawn a 2 plot windows: the frequencies of A,C,G,T, and the frequencies of uncallable bases, Figures 21 and 22.

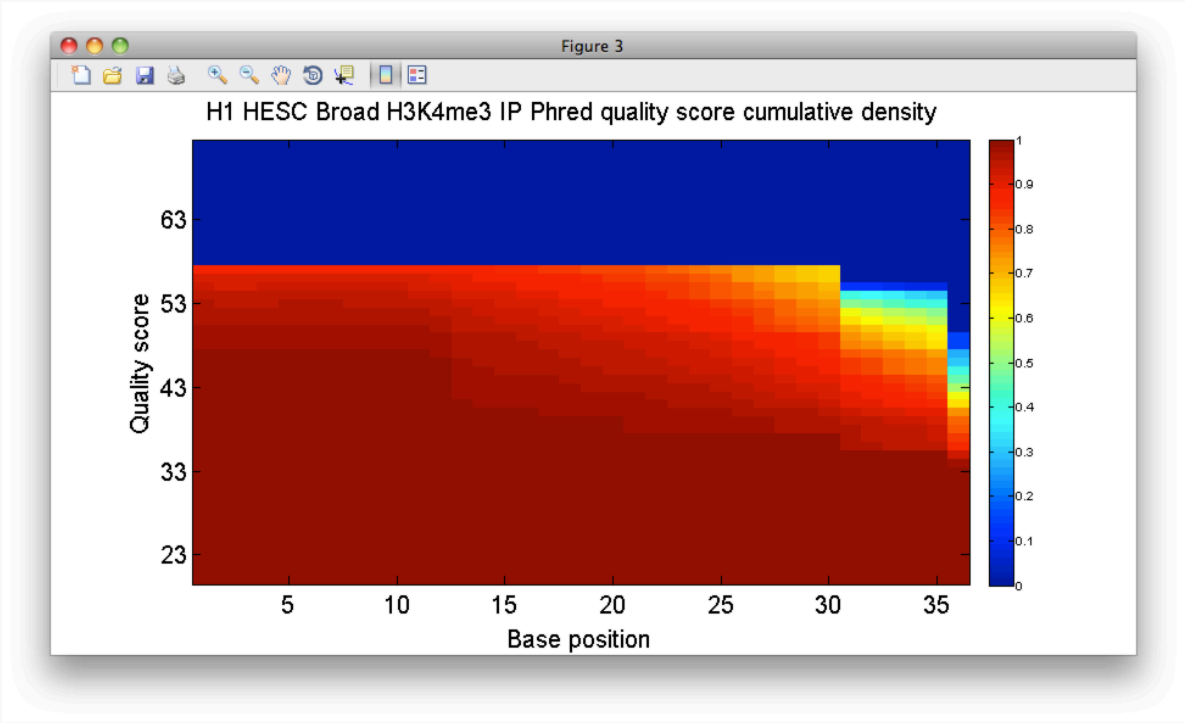

Figure 19

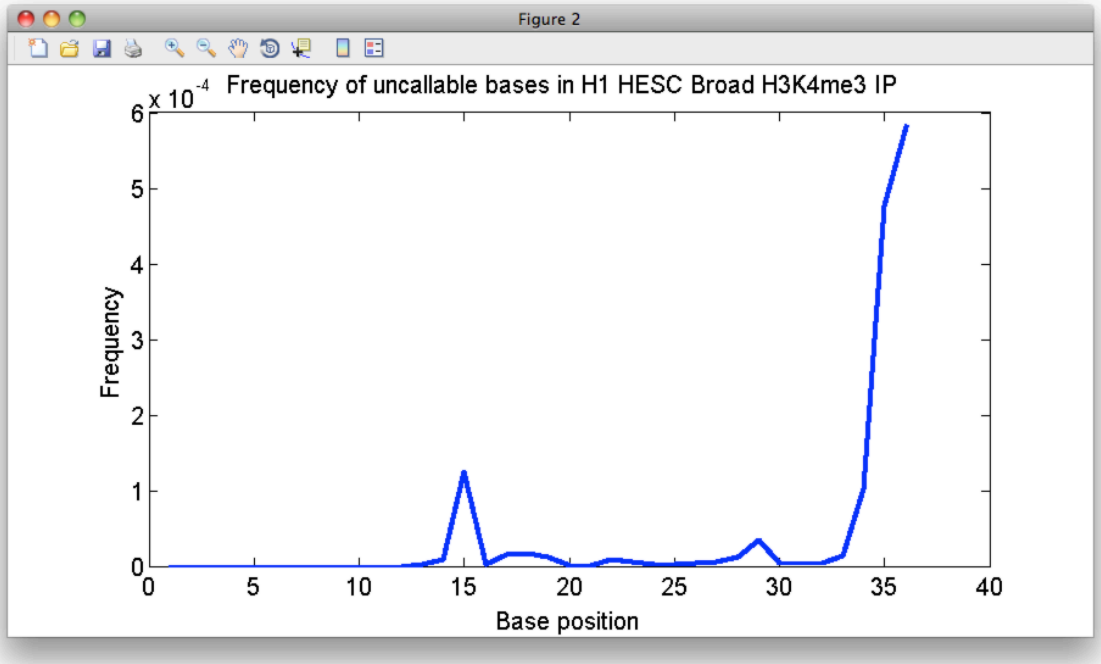
Figure 2
Frequency of uncallable bases in H1 HESC Broad H3K4me3 IP
x 10^-4
Frequency
Base position
0
1
2
3
4
5
6
0
5
10
15
20
25
30
35
40

Figure 20

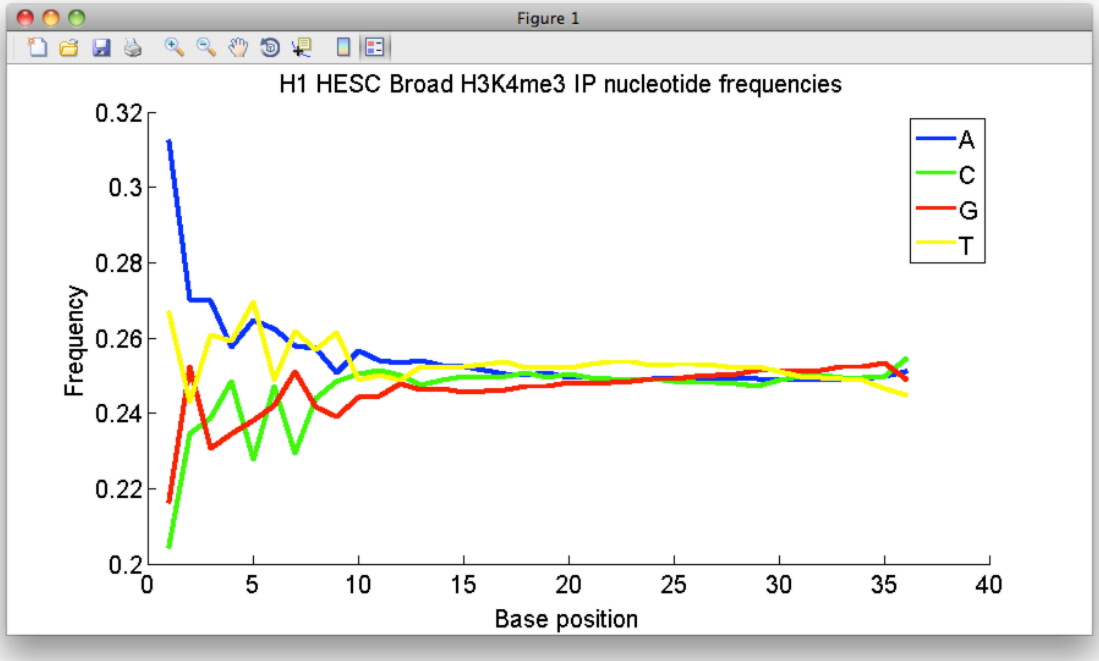
Figure 1
H1 HESC Broad H3K4me3 IP nucleotide frequencies
A
C
G
T
Frequency
Base position
0.2
0.22
0.24
0.26
0.28
0.3
0.32
0
5
10
15
20
25
30
35
40

> Figure 21

Now click on “View Phred score map,” and choose a score offset according to the equipment you used for your experiment. A heat map of quality scores vs. base positions is displayed, Figure 19. Consider a point (b, Q) on this map, for b a base position and Q a quality score. The color at that point indicates the proportion of calls at base position b that have quality scores greater than or equal to Q.

Things to look for:

1. Generally, you will see a type of ridge in the heat map which you want to be high for the bulk of the read length, representing a small fraction of reads with low quality base calls through out the read. Realistically, you will see a dip in quality, becoming gradually more pronounced as you move from the first to last base called.
2. To uncover quality biases, look for dips in quality cores on the heat map. The bias is often accompanied by an increase in the frequency of uncallable bases and an abrupt change in nucleotide frequency with respect to base position

---

View read density bias

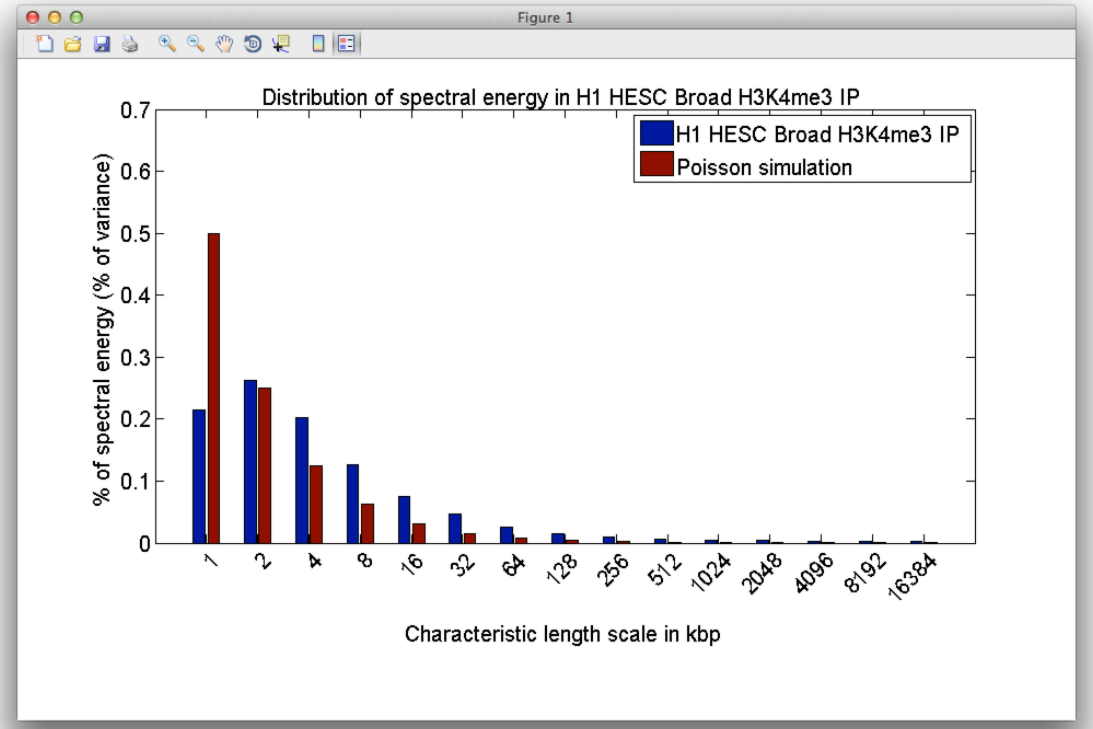

Figure 22

Bias in Input read density will decrease the power of a statistical test to detect enrichment. Fewer real peaks will be detected at a given false discovery rate (FDR) threshold. CHANCE detects bias in read density by using a signal processing technique known as spectral analysis. To get started click a sample from the list in the lower right of CHANCE's main window, then click "View read density bias." A plot of the distribution of spectral energy in the sample (blue) is displayed alongside that for an idealized Poisson simulation (red) based on the sample, Figure 20. The Poisson simulation represents an ideal version of the user's data which is unbiased but has been sequenced to the same depth of coverage as the user's data. On the x-axis of the plot, we have a set of length scales, from 1kbp to 16.384Mbp. On the y-axis, we have the percentage of variance in read density which is observed in the user's data at each length scale. If the chromatin sonication or digestion process were unbiased , if the library preparation, sequencing, and mapping were all done without bias or error then the break points introduced in chromatin would be uniformly distributed genome-wide, and the number of reads mapping to a particular region would be approximately Poisson-distributed with a mean constant throughout the genome. This expected trend would appear in the spectral analysis plots as a spectral energy distribution which was highest at 1kbp, indicating a read density profile composed of primarily of high frequency

fluctuations about a global mean. The spectral energy distribution would then rapidly drop down as we increase the length scale along the x-axis. If there is minimal read-density bias in the data, the Poisson simulation results should agree roughly with the sample results.

Appendix 1: valid chromosome identifiers

hg19:

chr1 chr10 chr11 chr11_gl000202_random chr12 chr13 chr14 chr15 chr16 chr17 chr17_ctg5_hap1 chr17_gl000203_random chr17_gl000204_random chr17_gl000205_random chr17_gl000206_random chr18 chr18_gl000207_random chr19 chr19_gl000208_random chr19_gl000209_random chr1_gl000191_random chr1_gl000192_random chr2 chr20 chr21 chr21_gl000210_random chr22 chr3 chr4 chr4_ctg9_hap1 chr4_gl000193_random chr4_gl000194_random chr5 chr6 chr6_apd_hap1 chr6_cox_hap2 chr6_dbb_hap3 chr6_mann_hap4 chr6_mcf_hap5 chr6_qbl_hap6 chr6_ssto_hap7 chr7 chr7_gl000195_random chr8 chr8_gl000196_random chr8_gl000197_random chr9 chr9_gl000198_random chr9_gl000199_random chr9_gl000200_random chr9_gl000201_random chrM chrUn_gl000211 chrUn_gl000212 chrUn_gl000213 chrUn_gl000214 chrUn_gl000215 chrUn_gl000216 chrUn_gl000217 chrUn_gl000218 chrUn_gl000219 chrUn_gl000220 chrUn_gl000221 chrUn_gl000222 chrUn_gl000223 chrUn_gl000224 chrUn_gl000225 chrUn_gl000226 chrUn_gl000227 chrUn_gl000228 chrUn_gl000229 chrUn_gl000230 chrUn_gl000231 chrUn_gl000232 chrUn_gl000233 chrUn_gl000234 chrUn_gl000235 chrUn_gl000236 chrUn_gl000237 chrUn_gl000238 chrUn_gl000239 chrUn_gl000240 chrUn_gl000241 chrUn_gl000242 chrUn_gl000243 chrUn_gl000244 chrUn_gl000245 chrUn_gl000246 chrUn_gl000247 chrUn_gl000248 chrUn_gl000249 chrX chrY

hg18:

chr1 chr10 chr10_random chr11 chr11_random chr12 chr13 chr13_random chr14 chr15 chr15_random chr16 chr16_random chr17 chr17_random chr18 chr18_random chr19 chr19_random chr1_random chr2 chr20 chr21 chr21_random chr22 chr22_h2_hap1 chr22_random chr2_random chr3 chr3_random chr4 chr4_random chr5 chr5_h2_hap1 chr5_random chr6 chr6_cox_hap1 chr6_qbl_hap2 chr6_random chr7 chr7_random chr8 chr8_random chr9 chr9_random chrM chrX chrX_random chrY

mm9

chr1 chr10 chr11 chr12 chr13 chr13_random chr14 chr15 chr16 chr16_random chr17 chr17_random chr18 chr19 chr1_random chr2 chr3 chr3_random chr4 chr4_random chr5 chr5_random chr6 chr7 chr7_random chr8 chr8_random chr9 chr9_random chrM chrUn_random chrX chrX_random chrY chrY_random

tair10:

chr1 chr2 chr3 chr4 chr5