CHANCE MANUAL

AARON DIAZ, ABHINAV NELLORE, AND JUN S SONG

Abstract. **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 [**?**]. For the theoretical analysis and technical interpretation of the main statistical tests used by CHANCE, see [**?**]. For software downloads and the sample data referred to in this guide see: http://songlab.ucsf.edu/CHANCE.html and for source code, wiki, bugs or requests see https://github.com/songlab/chance/

1. Installation

1.1. **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

1.2. **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.

1.3. **If you are running Mac OSX.**

(1) Decompress the CHANCE archive
   (a) Unzip the file `CHANCE_MacOS.zip` by double clicking the `CHANCE_MacOS.zip` icon.
   (b) Open the folder `CHANCE_MacOS/`.
(2) Install MCR, the MATLAB Compiler Runtime:
   (a) Unzip `MCRInstaller.zip` by double clicking its icon
   (b) Double click `InstallForMacOSX`
   (c) Follow the on screen instructions, but keep track of the install location if you change the default.
(3) To start CHANCE:
   (a) Double click the chance icon
   (b) To start CHANCE from the command line:
      (i) Navigate to the `CHANCE_MacOS` folder
      (ii) 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.

1.4. **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:

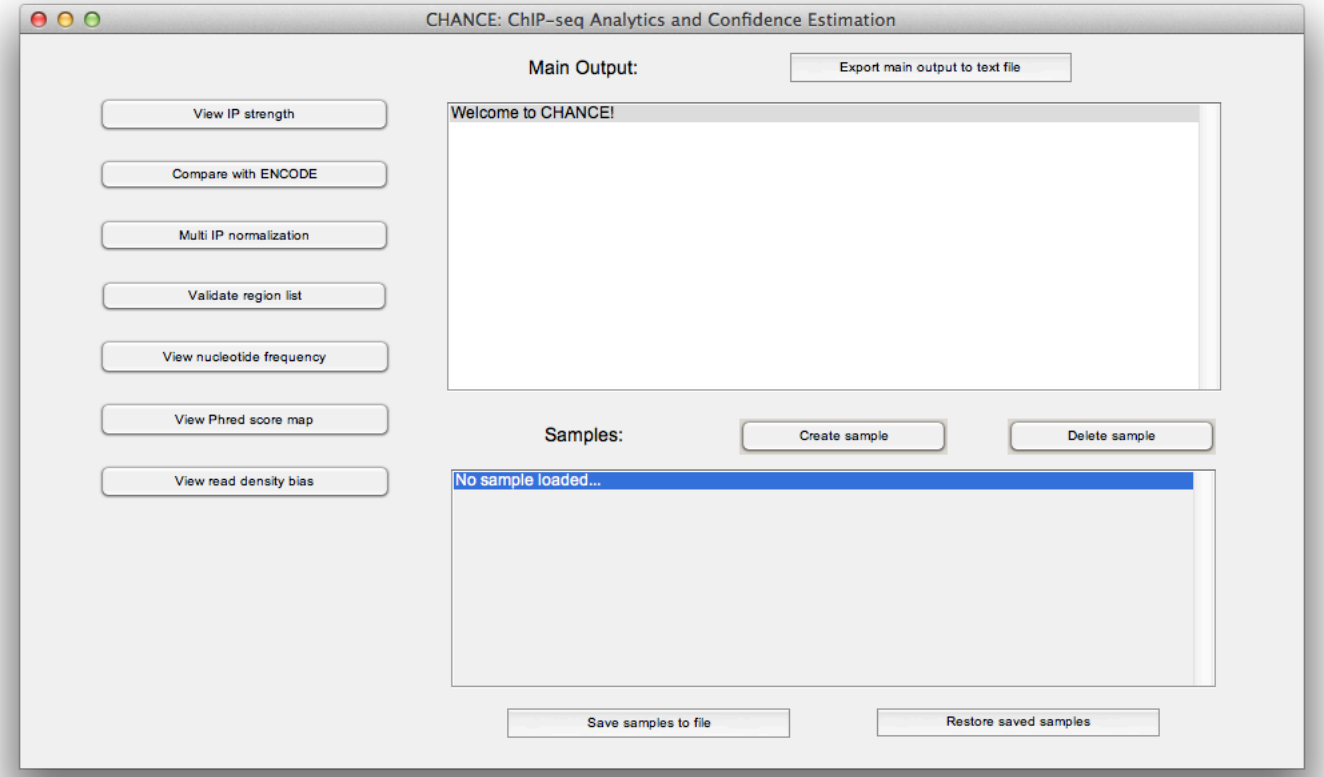

FIGURE 1.

```
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.

1.5. **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.

2. USING CHANCE

The main window of CHANCE, Figure **??**, 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 **??**.

2.1. **Adding samples to the workspace:**

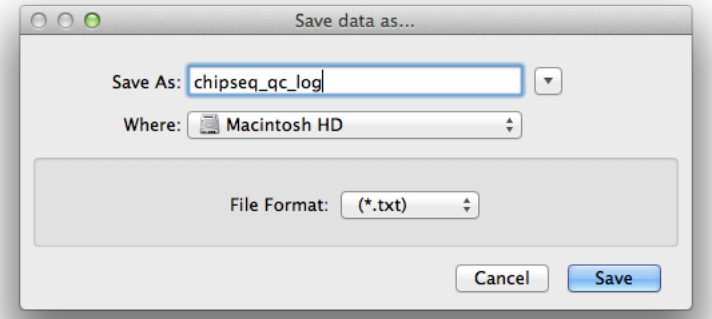

FIGURE 2.

2.1.1. *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.

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 **??**.

(5) When prompted enter a name for the sample, Figure **??** and **??**.

2.1.2. *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 **??**

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 **??**.

2.2. **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 **??**, an enrichment plot Figure **??**, and a second linearization plot Figure **??**.

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.

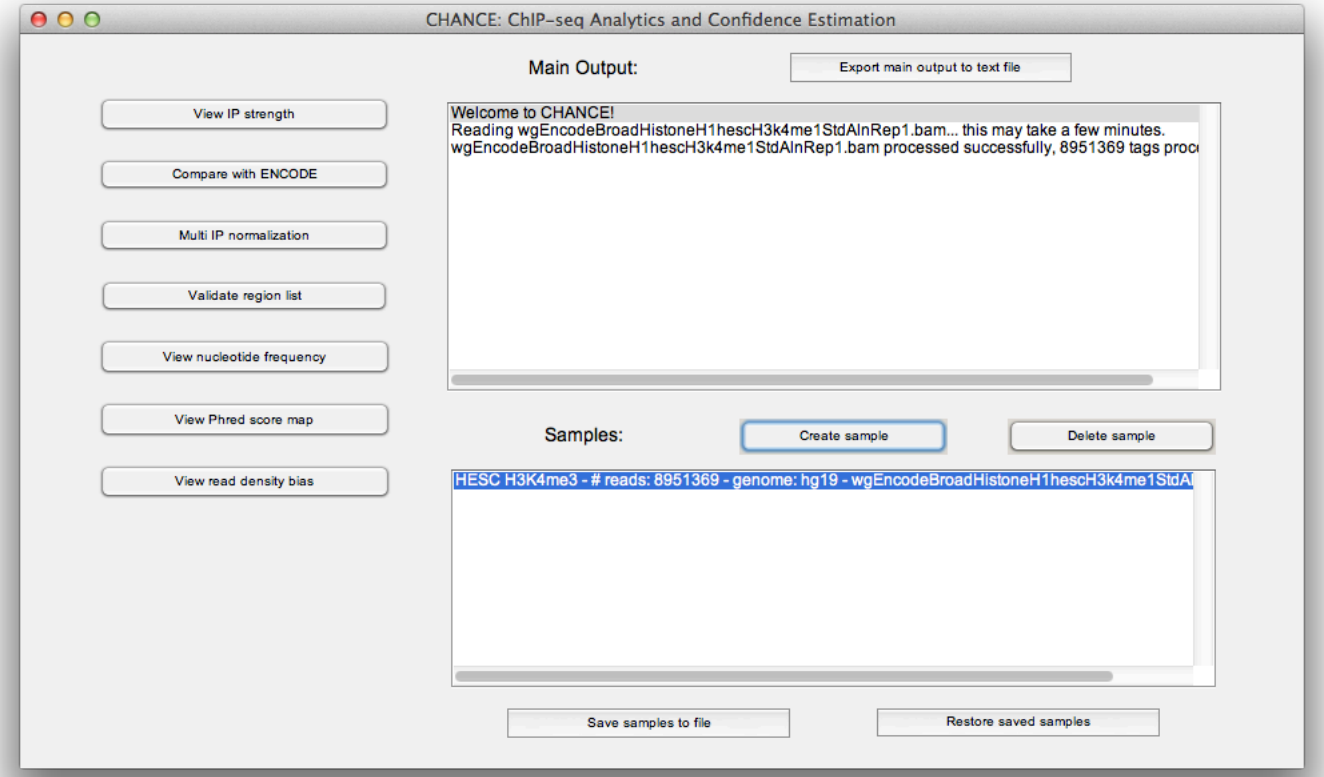

FIGURE 3.

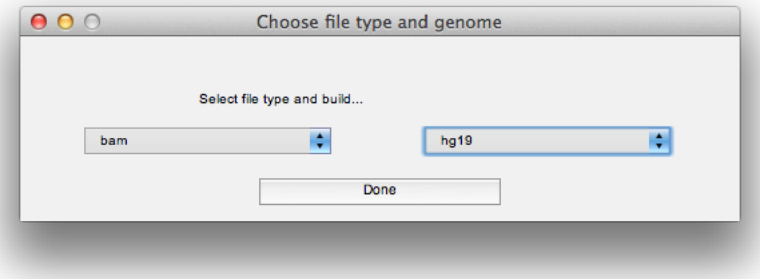

FIGURE 4.

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

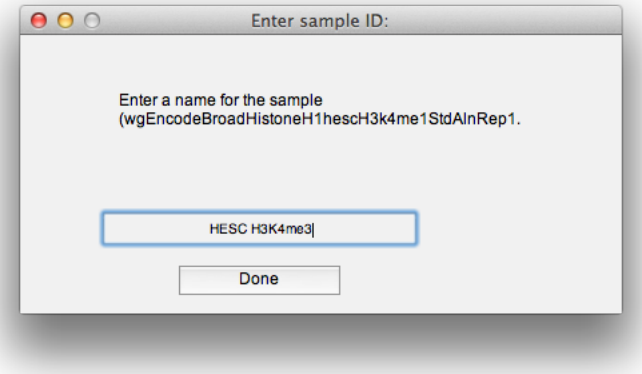

FIGURE 5.

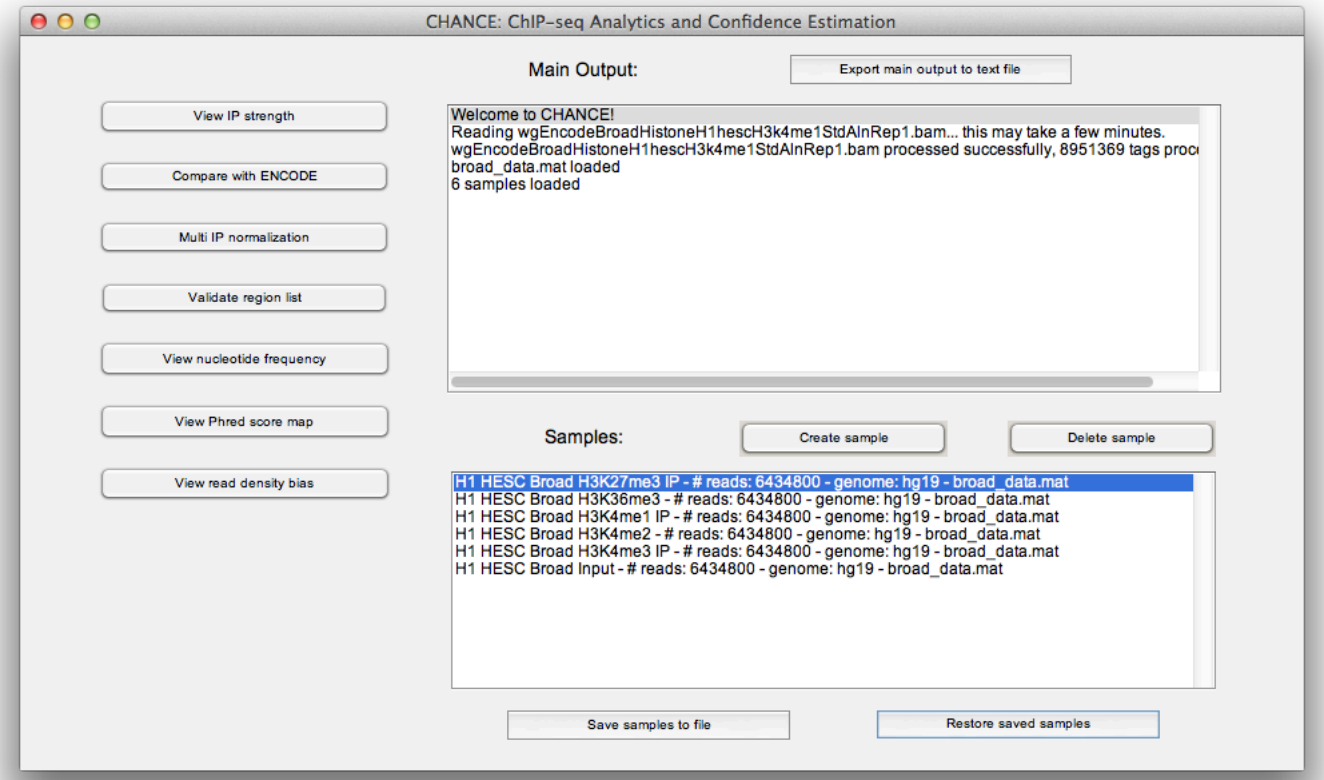

FIGURE 6.

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

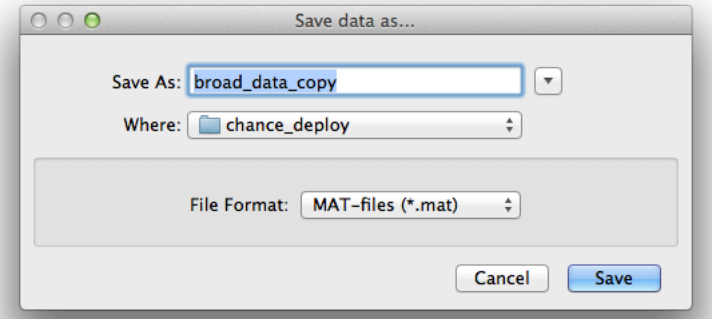

FIGURE 7.

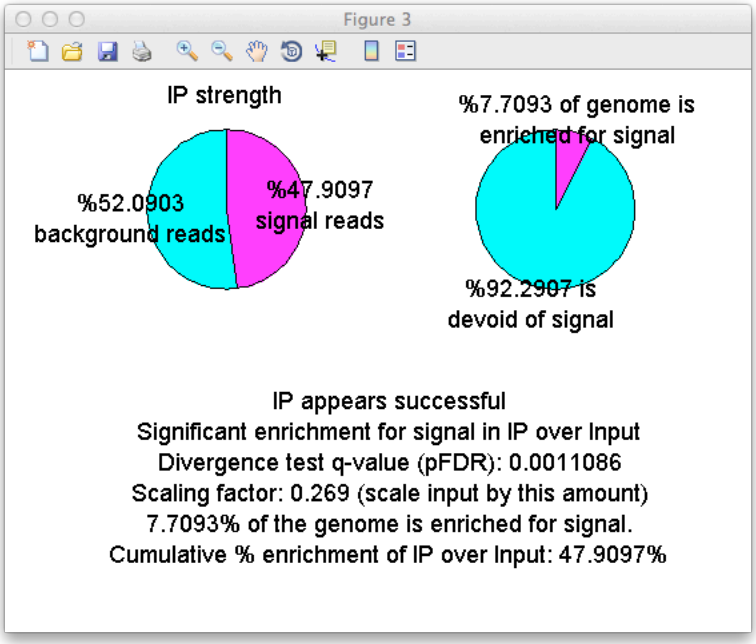

FIGURE 8.

implies a large number of regions with zero coverage, deflating the cumulative tag density. For example, Figure **??** 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, then the percentage enrichment and percent of the genome enriched estimations are not meaningful and CHANCE does not report them.

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.

2.3. **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:

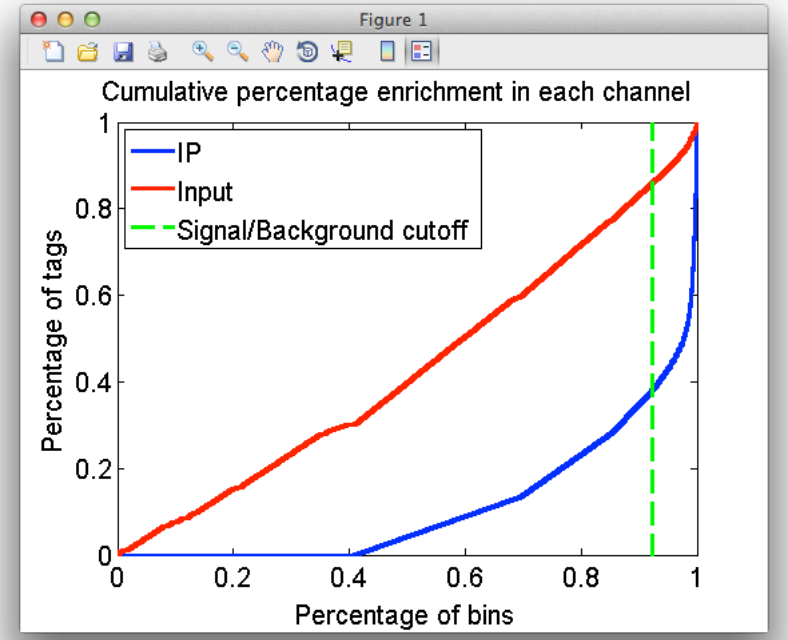

FIGURE 9.

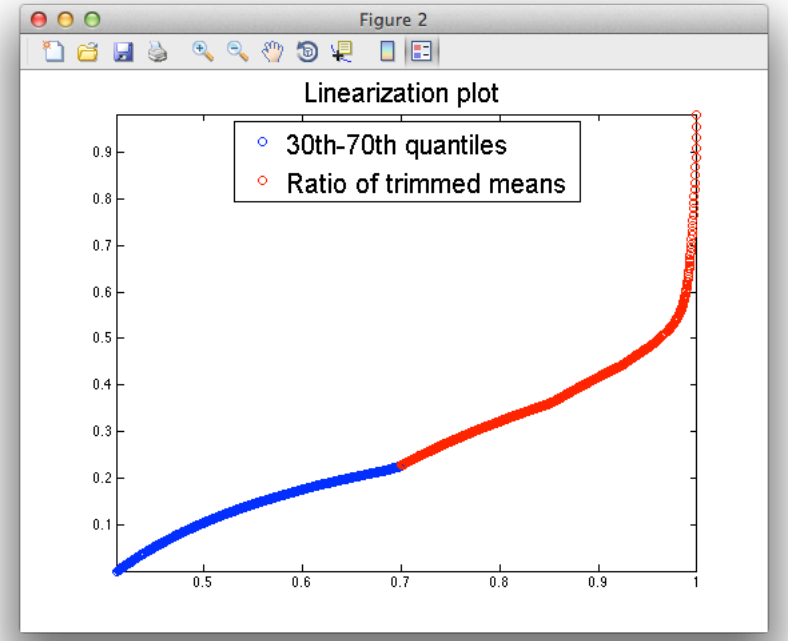

FIGURE 10.

- 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.

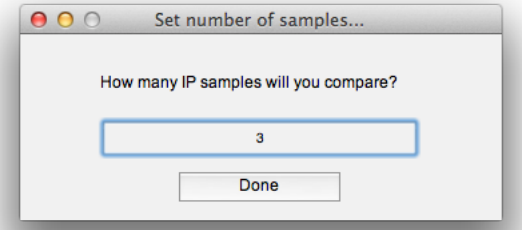

FIGURE 11.

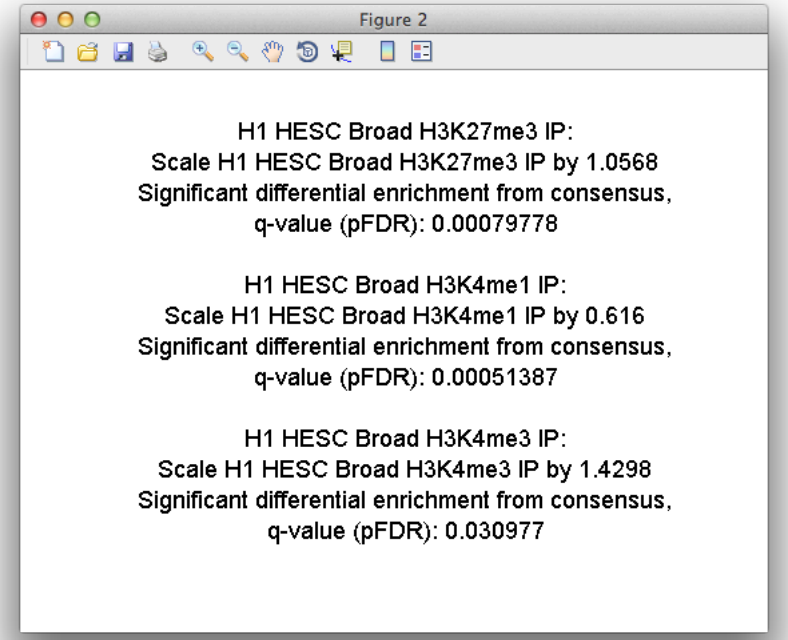

FIGURE 12.

To get started:

(1) Click the "Multi IP normalization" button.

(2) Enter the number of experiments to compare in the dialog box, Figure **??**.

(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 **??**, a differential enrichment comparison matrix Figure **??**,and an enrichment plot Figure **??**.

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 described in [**?**, **?**, **?**].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. 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 **??**.

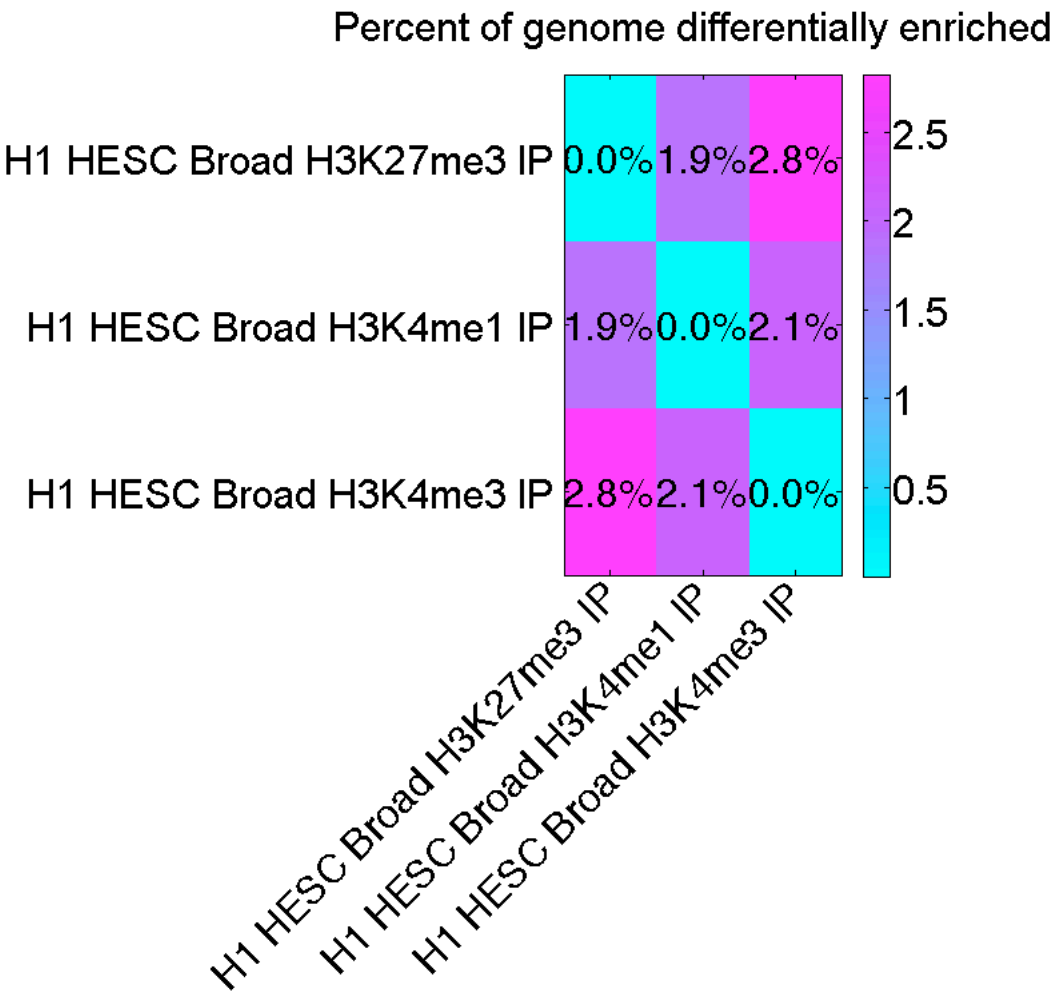

FIGURE 13.

(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.

2.4. **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 **??**.

(4) Select the transcription factor or epigenetic mark type from the drop down dialog box.

(5) CHANCE will now spawn a plot window:

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

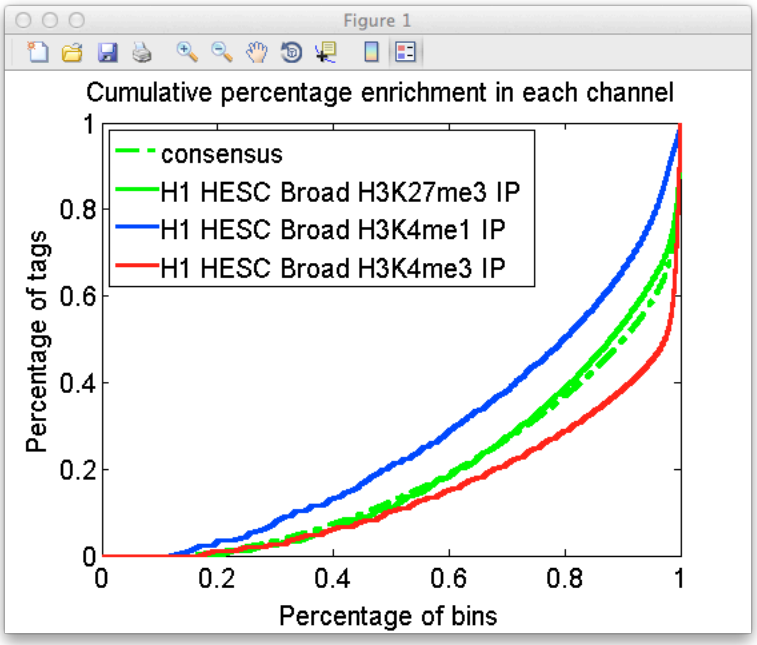

FIGURE 14.

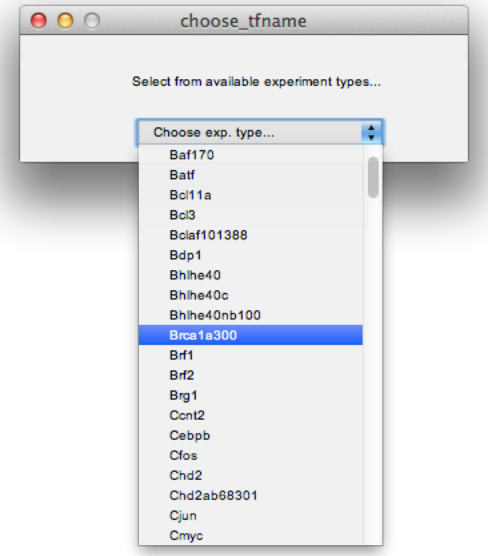

FIGURE 15.

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.

2.5. **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.

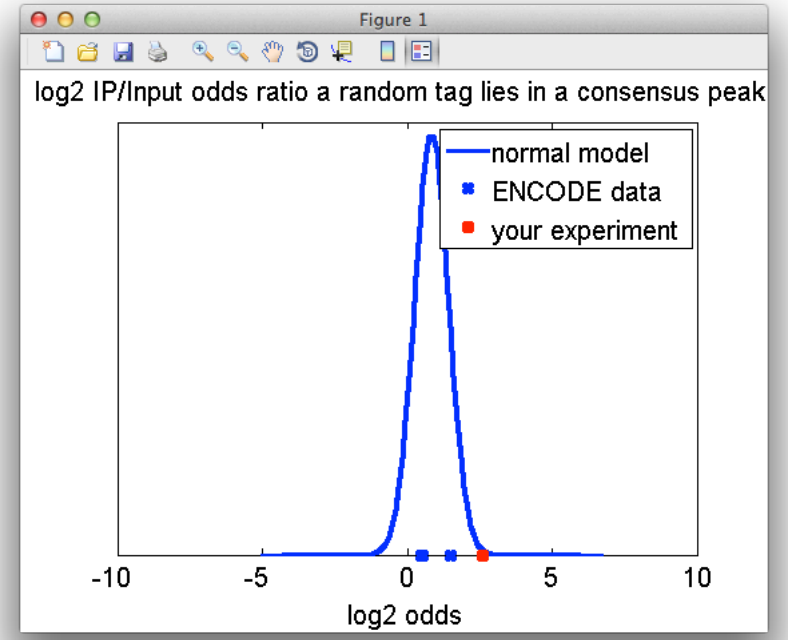

FIGURE 16.

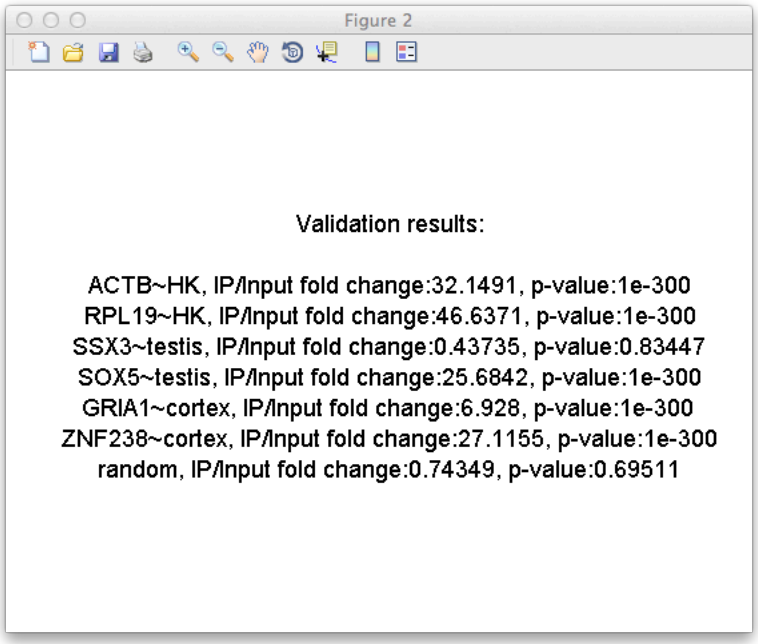

FIGURE 17.

(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, Figures **??** and **??**.

Things to look for:

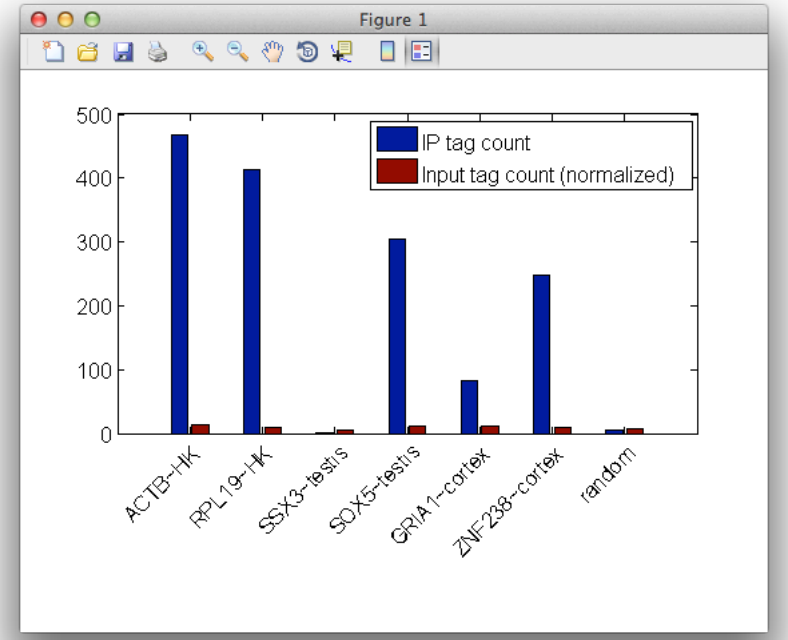

FIGURE 18.

(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.

2.6. **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 **??** and **??**.

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 **??**. 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

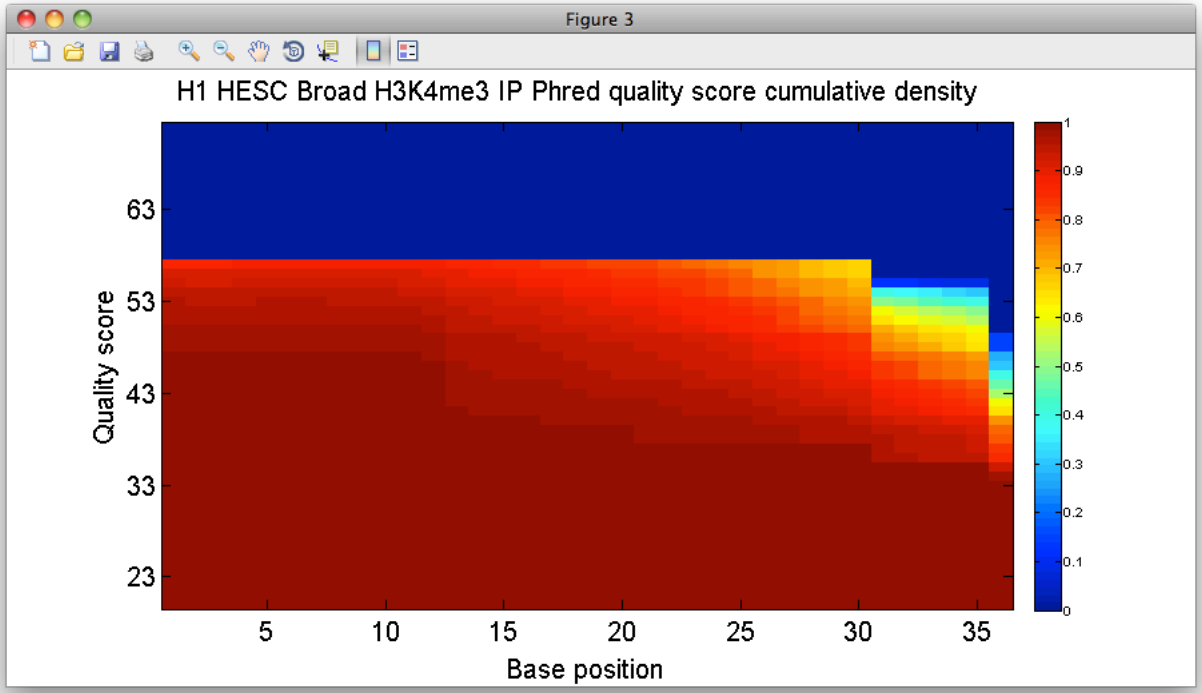

FIGURE 19.

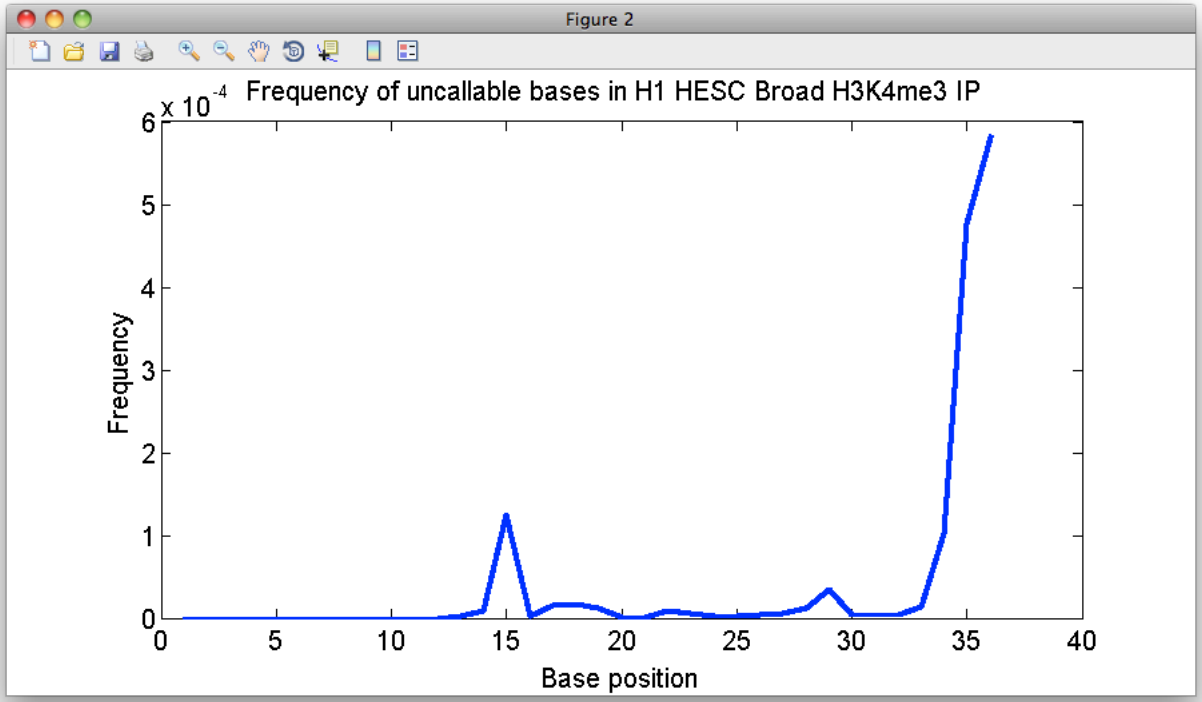

FIGURE 20.

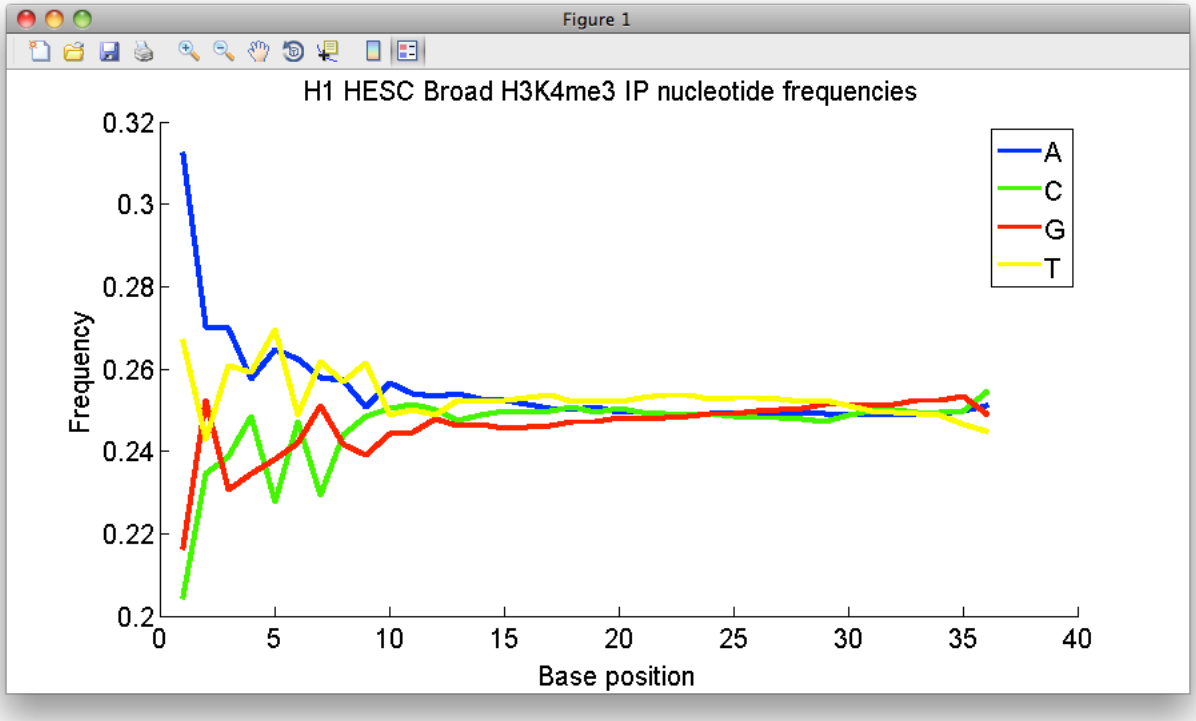

FIGURE 21.

2.7. **View read density bias.** 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 **??**. 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.

REFERENCES

[1] Diaz A, Nellore A, S SJ: **CHANCE - CHip-seq ANalytics and Confidence Estimation: comprehensive software for quality control and validation of ChIP-seq data**. *Genome Biology* submitted.
[2] Diaz A, Park K, Lim DA, Song JS: **Normalization, bias correction, and peak calling for ChIP-seq**. *Statistical Applications in Genetics and Molecular Biology* 2012, **11**(3), [http://http://www.ncbi.nlm.nih.gov/pubmed/22499706].
[3] **BED/tagAlign file format** [http://genome.ucsc.edu/FAQ/FAQformat].
[4] **SAM/BAM file format** [http://samtools.sourceforge.net/].
[5] Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome biology* 2009, **10**(3):R25, [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2690996&tool=pmcentrez&rendertype=abstract].
[6] Cover TM, Thomas JA: *Elements of Information Theory*. New York: John Wiley and Sons 2006.
[7] Cheung K, Vilnrotter V: **Channel Capacity of an Array System for Gaussian Channels With Applications to Combining and Noise Cancellation**. Tech. rep., NASA Jet Propulsion Laboratory, Communications Systems and Research Section 1996.
[8] Guo D: **Gaussian Channels : Information , Estimation and Multiuser Detection**. *PhD thesis*, Princeton University 2004.

E-mail address: DiazA2@humgen.ucsf.edu,anellore@gmail.com,jssong@humgen.ucsf.edu

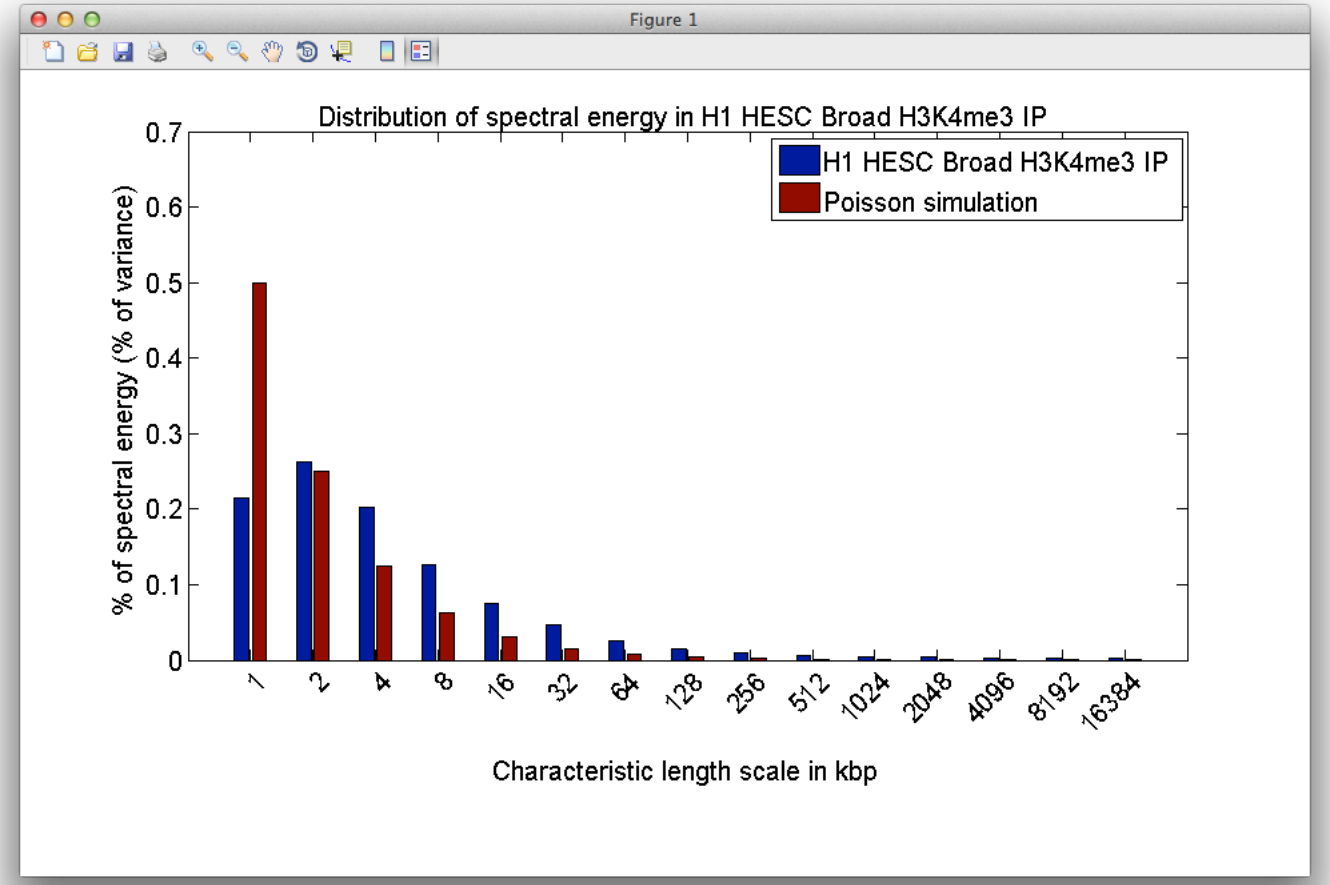
Figure 1
Distribution of spectral energy in H1 HESC Broad H3K4me3 IP
H1 HESC Broad H3K4me3 IP
Poisson simulation
% of spectral energy (% of variance)
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
4
8
16
32
64
128
256
512
1024
2048
4096
8192
16384
Characteristic length scale in kbp

FIGURE 22.