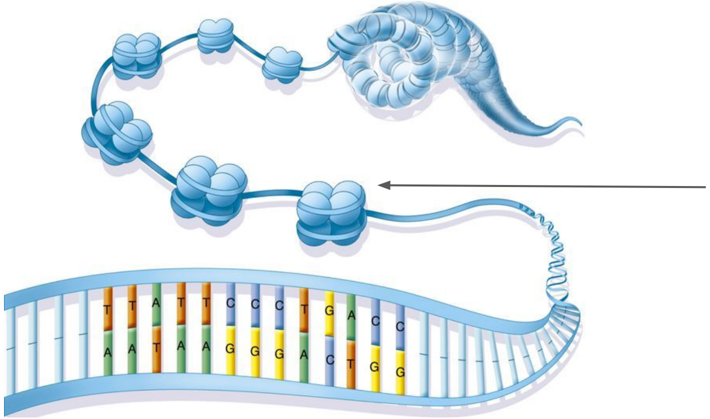
Recent Advances in ChIP-seq analysis: from quality management to whole-genome annotation

Nakato and Shirahige, 2017

Andrew Liu and Kevin Xie

ChIP-seq - Background

- Chromatin is the material chromosomes are made of
  - Consists of protein, RNA, and DNA
- Immunoprecipitation extracts specific protein antigens by utilizing a specific antibody associated with that antigen
- Chromatin Immunoprecipitation followed by sequencing analysis (ChIP-seq) can detect protein and DNA binding and histone-modification sites across an entire genome.

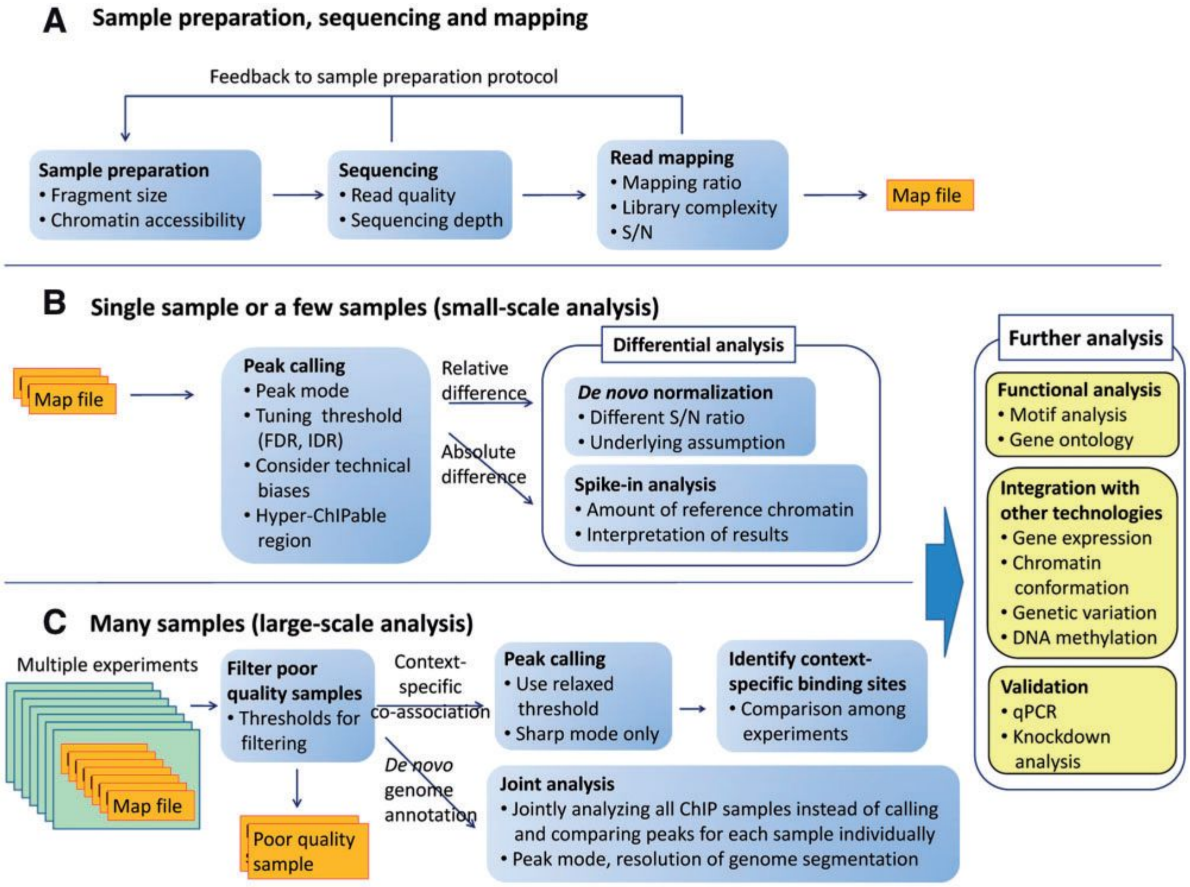


Histone (proteins that help package DNA)

ChIP-seq

- ChIP-seq was developed to understand the cooperation and interactions that occur in genomic function for different organisms using next generation sequencing (NGS)
- ChIP-seq is a mainstream method in genomics and epigenomics
    - Disease-associated transcriptional regulation
    - Tissue-specificity of epigenetic regulation
    - Chromatin organization

ChIP Protocols



A Sample preparation, sequencing and mapping

Feedback to sample preparation protocol

Sample preparation
- Fragment size
- Chromatin accessibility

Sequencing
- Read quality
- Sequencing depth

Read mapping
- Mapping ratio
- Library complexity
- S/N

Map file

B Single sample or a few samples (small-scale analysis)

Map file

Peak calling
- Peak mode
- Tuning threshold (FDR, IDR)
- Consider technical biases
- Hyper-ChIPable region

Relative difference

Absolute difference

Differential analysis

***De novo* normalization**
- Different S/N ratio
- Underlying assumption

Spike-in analysis
- Amount of reference chromatin
- Interpretation of results

C Many samples (large-scale analysis)

Multiple experiments

Map file

Filter poor quality samples
- Thresholds for filtering

Poor quality sample

Context-specific co-association

De novo genome annotation

Peak calling
- Use relaxed threshold
- Sharp mode only

Identify context-specific binding sites
- Comparison among experiments

Joint analysis
- Jointly analyzing all ChIP samples instead of calling and comparing peaks for each sample individually
- Peak mode, resolution of genome segmentation

Further analysis

Functional analysis
- Motif analysis
- Gene ontology

Integration with other technologies
- Gene expression
- Chromatin conformation
- Genetic variation
- DNA methylation

Validation
- qPCR
- Knockdown analysis

ChIP Protocol

- Cross-linked chromatin is sonicated (fragmented through vibrations).
    - Purified with and without immunoprecipitation (ChIP-seq and corresponding input DNA fragments)
- DNA fragments are sequenced as reads
    - These reads are then mapped onto the reference genome
- Enriched ChIP reads compared to input reads are detected as peaks.
    - Other genomic regions are known as nonspecific background.
- Peaks represent candidates of targeted protein/DNA-binding and histone modification sites
    - These targets can help identify associated functional annotations (what genes do)

ChIP-seq

- The results from ChIP-seq can also be integrated with other genomic assays
  - Gene Expression
  - DNA Methylation
  - Chromatin Conformation
- These results help us understand mechanisms for genomic functions from multiple aspects.

ChIP-seq Peaks

- Peaks represent candidates of targeted protein/DNA-binding sites and consists of three modes:
- Sharp Mode
    - Located at specific positions in the genome
    - Associated with transcription factors (TFs) and localized chromatin markers
    - Majority of peak-calling algorithms are designed for this mode
- Broad Mode
    - Associated with large genomic domains
    - Some proteins and histone modifications
- Mixed mode
    - Involved with both sharp and broad modes
    - RNA polymerase II and transcription elongation factors
- Different strategies for peak-calling are required for each shape

ChIP-seq Analysis

- Recent advances in technology allow us to analyze hundreds of ChIP samples simultaneously.
  - Large-scale analyses observe high-dimensional interrelationship between regulatory elements
- These large-scale analyses are sensitive to sample quality
  - Multilateral quality assessments during the computational procedures are essential
- Unfortunately, there is not one workflow that is appropriate or optimal for all conditions.
  - To obtain unbiased and reasonable data, the protocol design is important

Sample Preparation and Sequencing

- Once cross-linked chromatin is fragmented, the DNA fragments (150-500 bp) are sequenced as reads (36-100 bp)
    - Single-end reads are used in general
    - Paired-end reads improve the library complexity and increase mapping efficiency
        - Helpful for repetitive regions
- Chromatin accessibility during fragmentation is not uniform across the genome.
    - DNA is most easily fragmented and therefore preferentially represented in the sample (false positives)
    - Tightly packed regions like heterochromatin are not as easily accessible (confounding weak enrichment of true binding sites for heterochromatin markers)
    - This is why longer DNA fragments are preferred although it may not be efficient

Read Mapping

- Sequenced reads are mapped onto the genome using mapping tools
    - Most ChIP-seq experiments do not require gapped alignments that consider indels because the sequenced reads do not contain them.
        - Exception is cross-specifices analysis, which map onto other species’ genome
- Inclusion of multiple mapped reads
    - Reads mapped to multiple loci on the reference genome
    - Allowing for multiple mapped reads increases number of usable reads and sensitivity of peak detection
        - False positives may increase
    - In general, uniquely mapped reads are sufficient for analyze transcription factors

Read Mapping - Mappability

- Most ChIP-seq studies utilize uniquely mapped reads
- Mappability of a reference genome depends on the read length, type, and mapping tool and parameters.
- In general, calculating the genome-wide mappability for each genome is time consuming
  - Difficult to calculate the mappability of paired and gapped alignment data
- Most practical to use mappability data publicly available for similar parameter sets.
- Low-mappable regions may require multiple mapped reads or use paired-end reads

Read Mapping - Library Complexity

- Library Complexity is measured by non-redundant fraction (NRF)
  - Non-redundant reads/Total Number of mapped reads (N-nonred/N-all)
  - Non-redundant reads - reads mapped on the same genomic positions T times or less
    - Where T (usually 1, since expected number of mapped reads per base pair is much less than 1) is a set threshold for redundant reads
    - When expected mapped reads per base pair > 1, enriched regions for high signal-to-noise ratio (S/N), T is relaxed
    - When observing highly repetitive regions, filtering redundant reads should be omitted.
- Since NRF depends on total number of mapped reads, read sampling is necessary when comparing NRF scores across samples

Read Mapping - Sequencing Depth

- Number of peaks increase with sequencing depth
- Since weak protein binding may have important subfunctions, it is important to capture all functional sites.
- Sufficient depth depends on Signal-to-Noise ratio (S/N)
    - Saturation analysis can be used
        - However saturation thresholds have not been extensively defined for most histone modifications.
        - Agreeable depth is determined empirically
- For human samples, the Encyclopedia of DNA Elements (ENCODE) consortium suggests at least 2 biological replicates with 10 million uniquely mapped reads
- Others have suggested up to 60 million reads may be necessary for broad histone markers

Signal-to-Noise Ratio (S/N)

- Evaluated by the number and strength of peaks for each ChIP sample.
- ENCODE proposed two metrics
  - Fraction of reads in peaks (FRiP)
    - Number of reads falling within peak regions/Number of non-redundant
  - Cross-correlation profiles (CCP)
    - Plots Pearson cross-correlations between mapped read densities of positive and negative strands (y-axis) with shifting one strand (x-axis)
    - Samples with large and small S/N's typically have high CCs at shift points corresponding to
      - Fragment Length (C-frag)
      - Read Length (C-read)

S/N Cross Correlation Profiles

- After measuring the two lengths (C-frag and C-read) and a minimum cross correlation determined (C-min), two quantitative measures are scored:
    - Normalized strand coefficient (NSC)
        - NSC = C-frag/C-min
    - Relative strand correlation (RSC)
        - RSC = (C-frag - C-min)/(C-read - C-min)
- ENCODE recommends NSC > 1.05 and an RSC > 0.8 for typical TF (sharp mode)
- Using FRiP and CCP (S/N metrics), there are still many data sets that were of insufficient quality, which means that published data may need to be re-evaluated.

Visualization of ChIP - Analysis

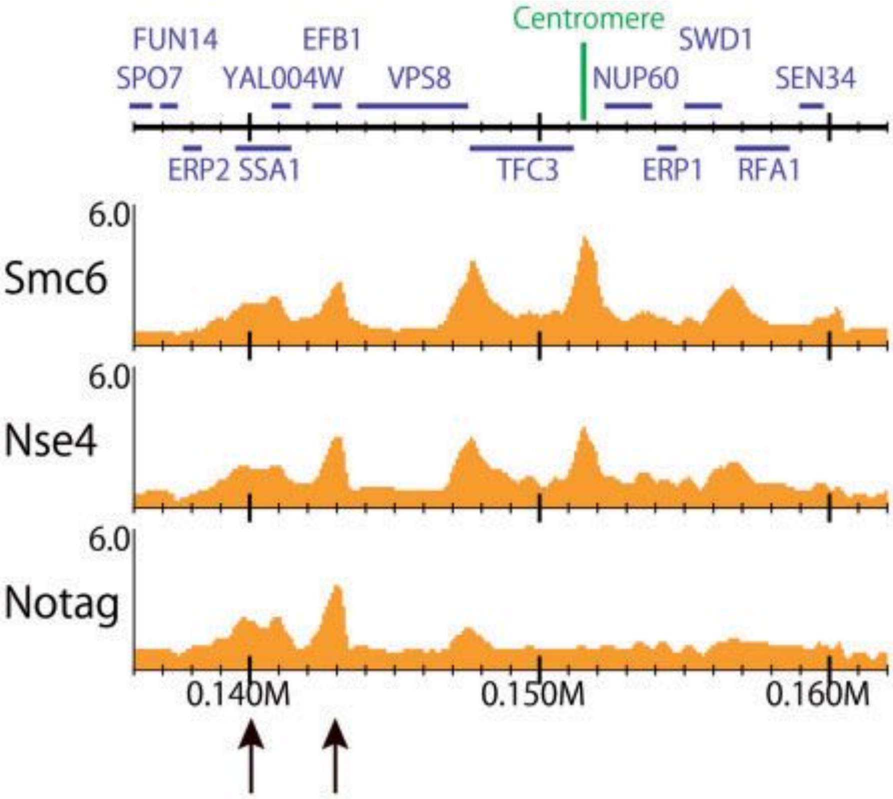
Peak Calling History

- Peaks detection uses a corresponding input sample to estimate the background distribution at a genomic locus
- Early programs adopted the Poisson Model, which assumes that the background reads are uniformly distributed along the genome.
    - However, it has been found that there is a greater variation in the read distribution than allowed.
- Negative binomial model was adopted to approximate the the overdispersion
    - Extended to zero-inflated negative binomial model
        - Accounts for lack of sequencing depth and low-mappable regions
- Local Poisson Model estimates parameter for each local genomic position
- Expectation-Maximization algorithm to predict protein-binding events
- Multiple hypothesis correction is performed to calculate false discovery rates (FDR) using Benjamini-Hochberg procedure

Peak Calling - Broad Peaks

- Detection of sharp peaks is much easier than detection of broad or mixed peaks.
- Proteins that are expected to be distributed within genic region, using aggregation plots around active genes and differential gene expression analyses is useful.
  - For proteins not expected to be distributed within genic region, genome-wide visualization and comparison to genome-wide maps is preferred
- Travelling ratio (TR) is proposed for Polymerase II (Pol II)
  - TRi = d-pp/d-gene
    - D-pp is the density in the promoter-proximal region
    - D-gene is the density in the gene body of gene i .
  - This score indicates whether the promoter-proximal Pol II is stalled at the gene
- Mutational Significance in Cancer (MUSIC) discriminates between two binding modes, and between stalled and elongating forms of Pol II.

ChIP/input Enrichment Distribution



- S. cerevisiae (yeast)
- Black arrows indicate false positives
- The reads from the ChIP-seq data are mapped onto the genome

Reliability

- Lack of true binding sites have made development of computation methods to evaluate peaks limited.
- Motif-based evaluations are not applicable for histone modifications since they do not have sequence specificity
- Proteins that do have motifs can have many tissue-specific binding sites recruited by other factors
- An alternative approach is to focus on reproducibility
  - Irreproducible discovery rate (IDR) assess the rank consistency of common peaks between two replicates
  - Can be used as a threshold robust for the technical variance

Low-quality samples

- Different antibodies can often produce different peak distributions
- Difficult to ascertain the sources of bias in a sample preparation
- Low-quality samples often has other problems (strong GC bias)
- Therefore, since there are many problems that can arise during the preparation process, it is important to fine tune each experiment to produce high quality samples.
- One thing that should be done is to limit the genomic regions to be investigated and then validating them using other biological experiments

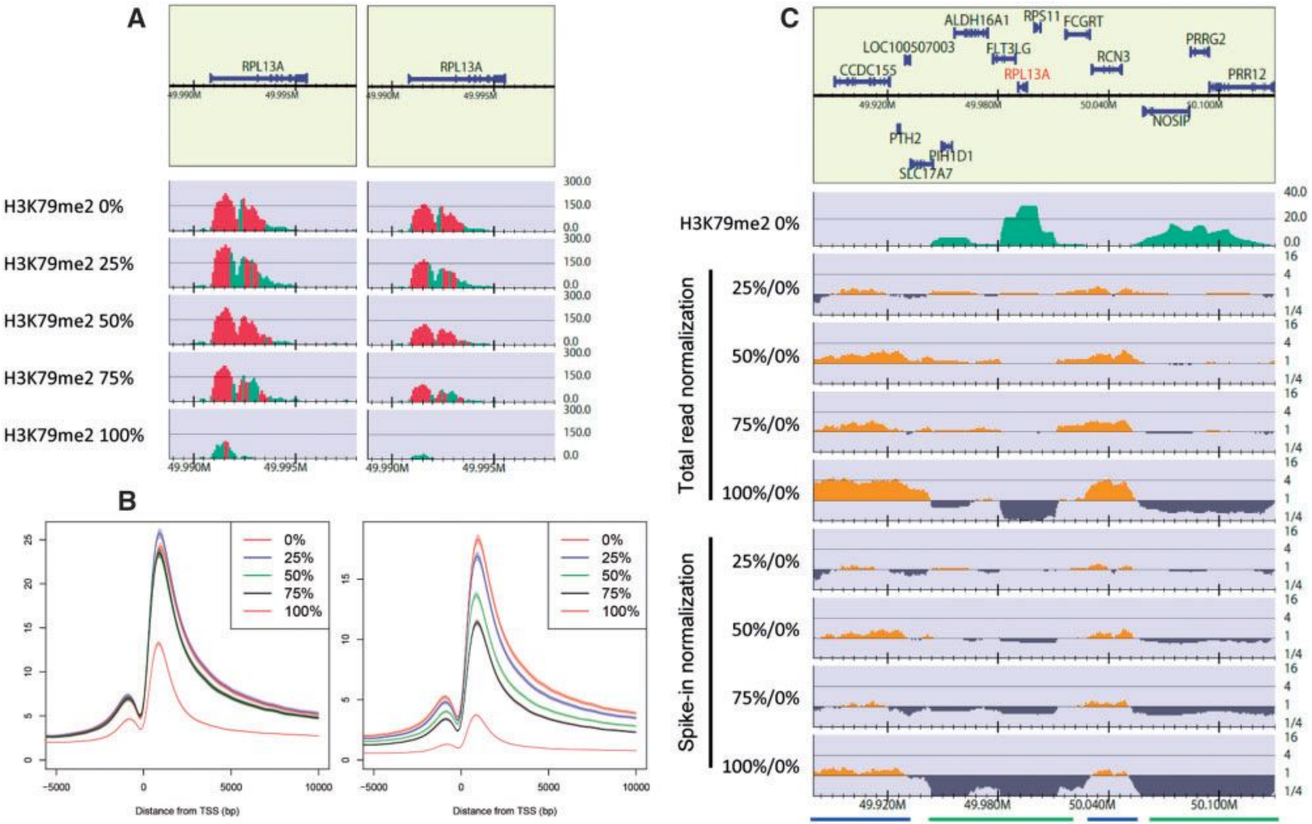
Normalization for differential analysis

- Simplest approach is to scale reads using total read number (N-nonred)
- Nonlinear methods using locally weighted regression (LOESS)
  - Remove effects of bias and systematic errors
  - However, this involves the assumption that means and variance are equal which is not always the case.
- Methods for differential gene expression can be used to compare more than two groups which do not consider S/N
- In contrast, MAnorm and ChIPcomp are designed to consider different S/N
  - MAnorm scales the reads of peaks common to two samples using a robust linear regression
  - ChIPcomp performs quantitative comparison of multiple ChIP samples
    - Measures genomic background using control data and considers multiple-factor experimental designs.

Absolute-level difference (spike-in analysis)

- If genome-wide peak distribution changes drastically, then assumptions used for the previous methods do not hold.
- Previous methods also analyze different in protein binding among samples.
- Spike-in analysis adds the same quantities of chromatin DNA to all samples.
    - This acts as a control for read normalization
    - Can detect global differences that cannot be identified by de novo normalization methods
- However, its applicability and limitations are still not clear.
    - Balancing the additional chromatin to chromatin of interest
    - Occasionally observe decrease in read density in both peak regions and in background.
        - It can be difficult to determine reason for decrease
- Overall it is important to replicate experiment using real data and simulated data.

Spike-in Analysis of H3K79me2



- EPZ5676-treated Jurkat cells (inhibits histone modification)
- Before spike-in and after spike-in analysis

Integrative analysis for a de novo genome annotation

- Producing many ChIP-seq data sets is possible at reasonable costs
- Comparison and Integration are not trivial.
- Suppose we have four proteins obtained under two conditions with knockdown effects of interest
  - Investigate differences among four proteins under two conditions and between wild-type and knockdown cells.
  - Thus it can be difficult to integrate all of the results since the results depend on the peak-calling result of each individual sample

Joint Analysis

- Use unsupervised machine learning methods to annotate a whole-genome sequence
- Receive all ChIP sample data and analyze them simultaneously instead of calling peaks and comparing the samples individually
- ChromHMM and Segway were developed to identify specific combination pattern of histone modifications
  - These methods can detect large-scale variations of histone marks across the genome
- ChromHMM
  - Models binary vectors for each 200-bp bin converted from raw read counts using sample-specific threshold as an independent Bernoulli random variable.
- Segway
  - Transforms the counts into real values and uses a dynamic Bayesian network at a 1-bp resolution.

Limitations and Challenges

- Requires large amounts of starting material (~10e5 cells)
- No single cell analysis
- Methods are being developed, but cannot translate large-scale analysis to single cell analysis
- Classify direct and indirect binding sites
- Capture temporary and non-site-specific TF binding
- Investigation of highly repetitive regions
- Need to integrate with other analyses
  - Human genetic variation
  - Genome editing
  - De novo Assembly

Questions?