ChIP-seq

Peak Calling

Peter N. Robinson

Institut für Medizinische Genetik und Humangenetik
Charité Universitätsmedizin Berlin

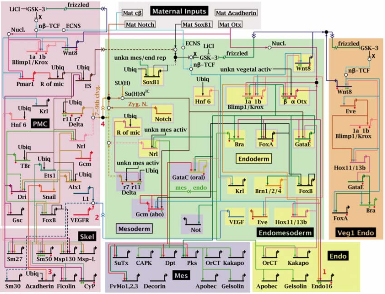
Genomics: Lecture #13

Outline

1. Gene Regulatory Networks
2. ChIP-Seq
3. XSET
4. FDR
5. MACS
6. ENCODE and the Irreproducible Discovery Rate (IDR)
7. The Big Picture

Gene Regulatory Networks

A genetic regulatory network (GRN) is a collection of genes which interact with each other indirectly (through their RNA and protein expression products) and with other substances in the cell, thereby governing the rates at which genes in the network are transcribed into mRNA, thereby mediating biological function.

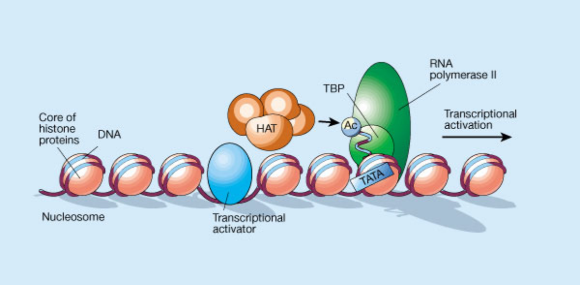


Ubiq=ubiquitous; Mat = maternal; activ = activator; rep = repressor;
unkn = unknown; Nucl. = nuclearization; χ = β-catenin source;
nβ-TCF = nuclearized b-β-catenin-Tcf1; ES = early signal;
ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

Copyright © 2001-2006 Hamid Bolouri and Eric Davidson

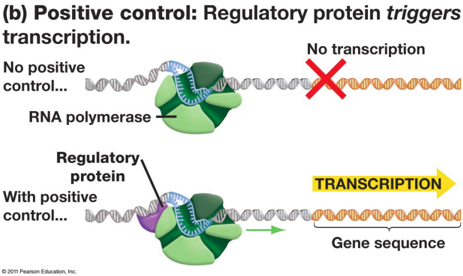
Gene Regulation

Genes are transcribed by RNA Polymerase II, but binding by more or less specific transcription factors is required to initialize this process



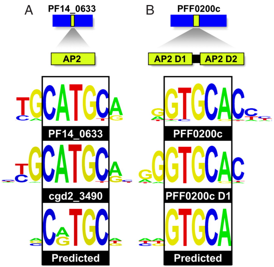
Gene Regulation

The following somewhat oversimplified cartoon illustrates the phenomenon of gene regulation by a specific regulatory protein (transcription factor), without which transcription does not occur.



Gene Regulation

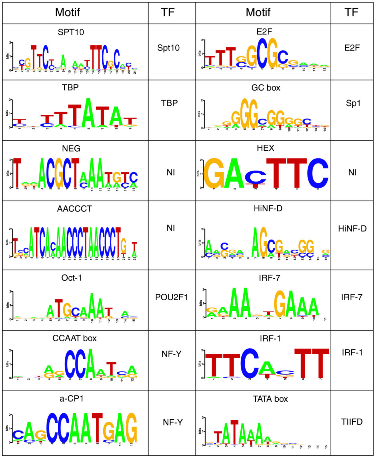
Proteins bind to DNA at more or less specific sequences, so called binding motifs. Genes that are regulated by a given transcription factor often have one or more DNA binding motifs for the protein within their promoter sequence or other regulatory sequences.



De Silva EK et al (2008) Specific DNA-binding by apicomplexan AP2 transcription factors. *PNAS* **105**:8393-8.

Gene Regulation

However, most DNA binding proteins do not have extremely specific binding motifs



Motif	TF	Motif	TF
SPT10	Spt10	E2F	E2F
TBP	TBP	GC box	Sp1
NEG	NI	HEX	NI
AACCCT	NI	HiNF-D	HiNF-D
Oct-1	POU2F1	IRF-7	IRF-7
CCAAT box	NF-Y	IRF-1	IRF-1
a-CP1	NF-Y	TATA box	TIIFD

Gene Regulation

- To understand gene regulation and gene regulatory networks, we want to know all of the sites in the genome to which transcription factors bind under different conditions[1]
- Because of the non-specificity of binding of transcription factors, a purely sequence-based approach to predicting transcription factor binding sites (**TFBS**) simply does not work well at all.
- Therefore, an experimental methodology has been developed that combines next-generation sequencing and chromatin immunoprecipitation.

[1] There are at least 1391 characterized transcription factors in the human genome- Vaquerizas JM et al (2009) A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**:252-63.

Outline

1. Gene Regulatory Networks
2. ChIP-Seq
3. XSET
4. FDR
5. MACS
6. ENCODE and the Irreproducible Discovery Rate (IDR)
7. The Big Picture

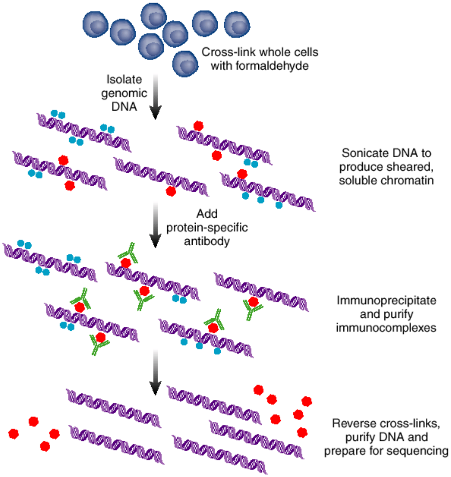
ChIP-Seq

Chromatin Immunoprecipitation following by next generation sequencing (**ChIP-seq**) is used to analyze protein interactions with DNA.

Three basic steps:

1. covalent cross-links between proteins and DNA are formed, typically by treating cells with formaldehyde
2. an antibody specific to the protein of interest is used to selectively coimmunoprecipitate the protein-bound DNA fragments that were covalently cross-linked.
3. the immunoprecipitated protein-DNA links are reversed and the recovered DNA is assayed to determine the sequences bound by that protein

ChIP-Seq: Workflow

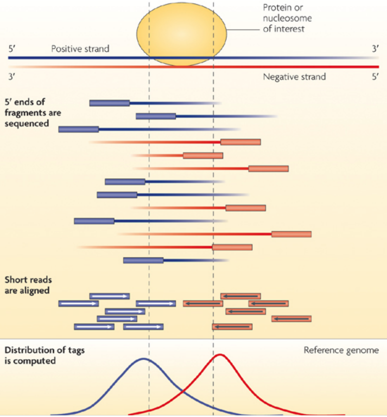

Cross-link whole cells with formaldehyde
Isolate genomic DNA
Sonicate DNA to produce sheared, soluble chromatin
Add protein-specific antibody
Immunoprecipitate and purify immunocomplexes
Reverse cross-links, purify DNA and prepare for sequencing

Elaine R Mardis ER (2007) ChIP-seq: welcome to the new frontier. *Nature Methods* **4**:613–614

ChIP-Seq: Workflow

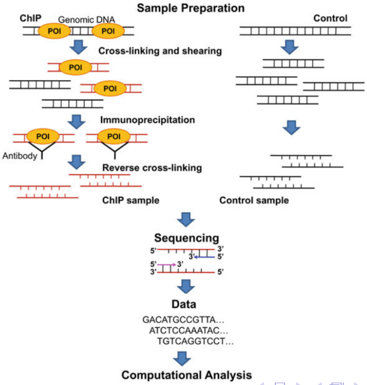
DNA fragments from a chromatin immunoprecipitation experiment are sequenced from the 5' end.

With ChIP-seq, the alignment of the reads to the genome results in two peaks (one on each strand) that flank the binding location of the protein or nucleosome of interest.


Protein or nucleosome of interest
5'
Positive strand
3'
3'
Negative strand
5'
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome

ChIP-Seq: Workflow

Most experimental protocols involve a control sample that is processed the same way as the test sample except that no specific antibody is used to enrich the bound protein. This serves to be able to calculate the background distribution.



Outline

1. Gene Regulatory Networks
2. ChIP-Seq
3. XSET
4. FDR
5. MACS
6. ENCODE and the Irreproducible Discovery Rate (IDR)
7. The Big Picture

XSET: A simple algorithm

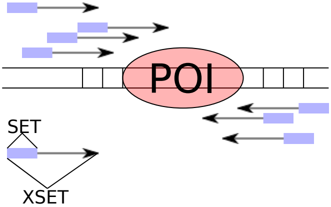
To set the stage, we will explain a simple algorithm from one of the very first ChIP-seq papers from 2007.

- The methodology involves a relatively simple scheme to calculate peak depth in ChIP-Seq experiments.

Robertson G et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing *Nature Methods* **4**:651–657.

XSET: A simple algorithm

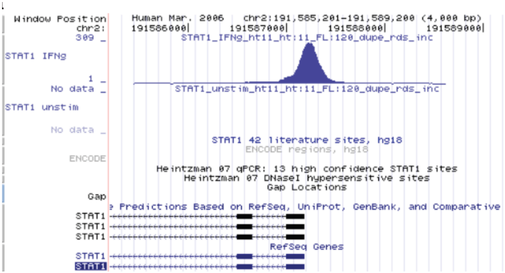
- We start with **single-end tags (SET)**, typically very short e.g., 36 bp. *Note fragments are sequenced from their 5' end in 5' to 3' direction only!*
- In a typical ChIP-Seq experiment, we will have 20 to 50 million reads that are mapped to the genome using "standard" methodologies
- The SETs are "computationally extended" in the 3' direction (e.g., 174-bp) into an **extended SET (XSET)**.
- XSET length is chosen to be the mean fragment length of the size selected DNA.



XSET: Overlap profiles

XSET overlap profiles are then calculated by counting the number of XSETs that are aligned to any given position of the genome.

- But how do we know whether any given peak is enriched? How do we know what is statistically significant?


Window Position
chr2:
309 _
STAT1 IFNg
1 _
No data _
STAT1 unstim
No data _
ENCODE
Gap
STAT1
STAT1
STAT1
STAT1
STAT1
Human Mar. 2006 chr2:191,585,201-191,589,200 (4,000 bp)
191586000|
191587000|
191588000|
191589000|
STAT1_IFNg_ht11_ht:11_FL:120_dupe_rds_inc
STAT1_unstim_ht11_ht:11_FL:120_dupe_rds_inc
STAT1 42 literature sites, hg18
ENCODE regions, hg18
Heintzman 07 qPCR: 13 high confidence STAT1 sites
Heintzman 07 DNaseI hypersensitive sites
Gap Locations
e Predictions Based on RefSeq, UniProt, GenBank, and Comparative
RefSeq Genes

Outline

1. Gene Regulatory Networks
2. ChIP-Seq
3. XSET
4. FDR
5. MACS
6. ENCODE and the Irreproducible Discovery Rate (IDR)
7. The Big Picture

False Discovery Rate

XSET employs the notion of False Discovery Rate (FDR) to estimate the significance of ChIP-Seq peaks. We will review the salient concepts.

The scenario:

- We want to simultaneously test m null hypotheses $H_1, \ldots, H_m$ at level α, giving p-values p_i
- Each hypothesis (in the current case) represents a candidate ChIP-Seq peak (transcription factor binding event), and the null hypothesis is that there is no true binding.
- m_0 of these hypotheses are truly null (no effect)

False Discovery Rate

Assume we are talking about a testing procedure based on *p*-values, and let us consider the rejection region Γ.

The scenario:

- Let R be the number of rejections (p-value lower than significance threshold)
- Let V be the number of rejections of truly null hypotheses (false positive rejections)
- Intuitively, we would like to define $\mathrm{FDR} = \frac{V}{R}$, i.e., the proportion of false positive rejections amongst all rejections.
- We will not go into this topic in detail here[2]

[2] See especially various writings by Storey for more about FDR.

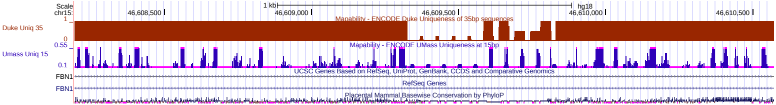
XSET: FDR

XSET uses an empirical procedure to provide an estimate of the FDR based on the characteristics of the data

- Randomly place the same number of reads as in the real data onto the genome
- Each random read is defined to have the XSET length
- Calculate the random expectation for the probability of observing peaks with a particular height, taking **mapability** into account

Mappability

Not all reads can be mapped uniquely to the genome. **Mappability** of a sequence of length *n* relates to the uniqueness (or not) of a sequence of length *n* that starts at a particular position of the genome. If there is another identical sequence somewhere else, then the *n*-mer sequence is not mappable.


Scale
chr15:
1 kb
hg18
46,608,500
46,609,000
46,609,500
46,610,000
46,610,500
Mapability - ENCODE Duke Uniqueness of 35bp sequences
Duke Uniq 35
1
0
Mapability - ENCODE UMass Uniqueness at 15bp
Umass Uniq 15
0.55
0.1
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
FBN1
RefSeq Genes
FBN1
Placental Mammal Basewise Conservation by PhyloP

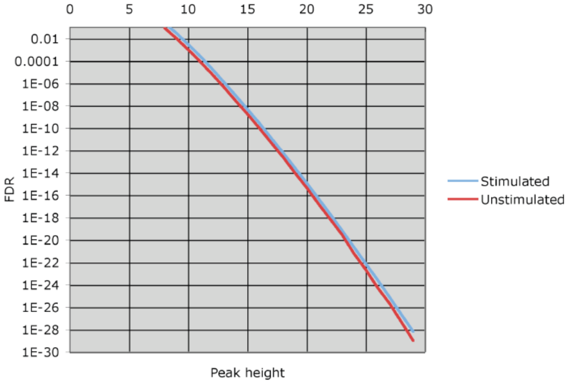
Mappability: The uniqueness of a stretch of DNA sequence compared with a whole-genome sequence. Short sequence reads can be confidently mapped to unique sequence, but less confidently mapped to sequence that occurs multiple times in a genome. Mappability increases substantially with read length

XSET: FDR

- It is easy to show that 27-bp reads can be mapped uniquely to $\sim$ 90% of the human genome
- Therefore, the background simulations for XSET for reads of 27bp uses a mappable genome length that was 90% of 3.08 Gb.
- For a given peak height, one can estimate the FDR as the number of peaks found in the randomized data (these are by definition false positive) to the number of peaks that were actually observed (these are presumably not all true positives, but seem a reasonable estimate thereof)

XSET: FDR

Relationship between the peak height threshold (number of XSETs that are aligned across a peak) and the estimated FDR



XSET: FDR

For each profile, we chose a threshold peak height as the smallest height that was equivalent to FDR < 0.001 for peaks of that height. All peaks of at least this height were retained in the profile.

- For the random data we can calculate a global coverage level as

$$\lambda = \frac{\ell \times N}{G^*}$$

- Here, ℓ is the length of the XSETs (174bp in our example), N is the number of XSETs in the ChIP-Seq experiment, and G^* is the mapability-adjusted genome size (for 27bp reads, 0.9×3.08 Gb)

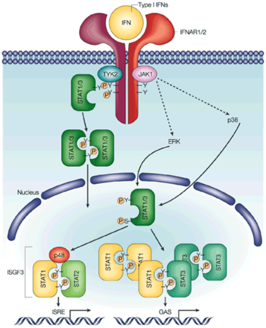
XSET: FDR

- Given a λ value calculated as above, the probability of observing a peak with a height of at least h is given by a sum of Poisson probabilities as:

$$P(H \geq h) = \sum_{k=h}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = 1 - \sum_{k=0}^{h-1} \frac{e^{-\lambda}\lambda^k}{k!} \tag{1}$$

Stat1 and Interferon

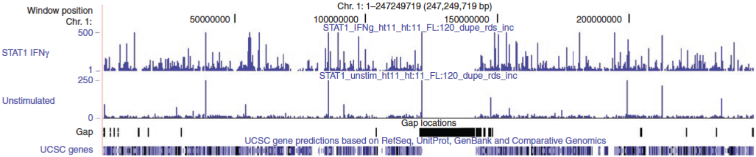
Let us now look at a typical ChIP-Seq experiment. Stat1 is a transcription factor that can be activated by stimulation of cells by interferon-γ. Thus, by performing one experiment before and one after interferon-γ stimulation, comparison of the peaks indicates the biological effect due to the stimulation.



Nature Reviews | Immunology

Stat1 and Interferon

FDR-thresholded XSET profiles and peaks (the significance threshold was estimated at $\lambda = 11$). Stimulated and unstimulated FDR-thresholded XSET profiles for the 247 Mb chromosome 1

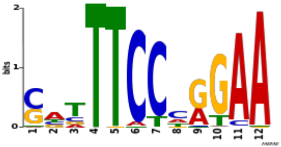


Stat1 and Interferon

The set of peaks and their location then form the basis for biological interpretation of the actions of the transcription factor being investigated.

Parameter	stimulated	unstimulated
peak height at FDR threshold	11	11
Number of peaks	41,582	11,004
Average height	29.2	21.0
Median height	16	13

STAT1 motif inferred from sequences at peaks



Outline

1. Gene Regulatory Networks
2. ChIP-Seq
3. XSET
4. FDR
5. **MACS**
6. ENCODE and the Irreproducible Discovery Rate (IDR)
7. The Big Picture

MACS

We will now present **Model-based Analysis of ChIP-Seq data (MACS)**, which has been one of the most commonly used peak finders. MACS introduced a more sophisticated way of modeling the fragment size.

- Clearly, the estimation of the fragment size is critical to the performance of an algorithm such as XSET: The larger the fragment size, the higher the average coverage of the genome is, which has a direct influence on the calculation of the estimated significance threshold

Zhang Y (2008) Model-based Analysis of ChIP-Seq (MACS) *Genome Biology* **9**:R137

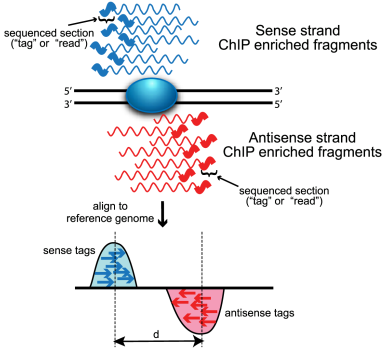
MACS Bimodal enrichment

Since ChIP-DNA fragments are equally likely to be sequenced from both ends, the tag density around a true binding site should show a bimodal enrichment pattern

- Watson strand tags enriched upstream of binding and Crick strand tags enriched downstream.
- Tags are often shifted/extended towards the 3' direction to better represent the precise protein-DNA interaction site (as with XSETs). The size of the shift is, however, often unknown to the experimenter.

MACS Bimodal enrichment

The 5’ to 3’ sequencing requirement and short read length produce stranded bias in tag distribution.



The separation between peaks (d) corresponds to the average sequenced fragment length.

Wilbanks EG (2010) Evaluation of Algorithm Performance in ChIP-Seq Peak Detection *PLoS ONE* **5**:e11471.

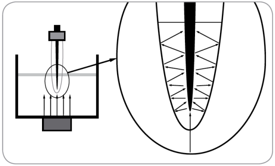
ChIP-seq Fragment Length

- Why does the separation between peaks (d) correspond to the average sequenced fragment length?

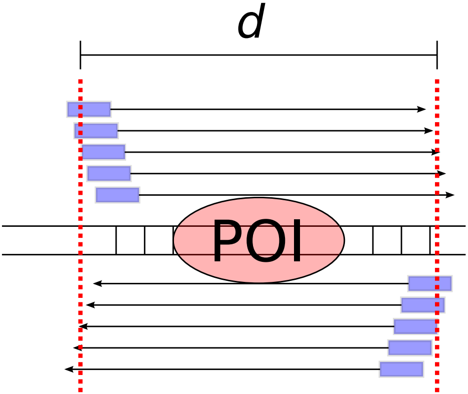
Recall: Library Prep: Fragmentation

- Most Illumina protocols require that DNA is fragmented to less than 800 nt.
- Ideally, fragments have uniform size
- Sonication uses ultrasound waves in solution to shear DNA.
- Ultrasound waves pass through the sample, expanding and contracting liquid, creating “bubbles” in a process called *cavitation*.
- Bubbles ⇒ focused shearing forces ⇒ fragment the DNA

- Sketch of sonication in “Eppi”
- Source: Bioruptor (http://www.diagenode.com/)



ChIP-seq Fragment Length



The blue box shows the region of the fragment that actually is sequenced (often 36bp). The entire fragment is longer, with the exact size depending on the experimental fragmentation protocol. On average, the protein of interest (POI) is located in the middle of the fragment, so that the average distance between reads corresponds to the average fragment length

MACS: Estimation of fragment size

Given a sonication size (*bandwidth*) and a high-confidence fold-enrichment (*mfold*), MACS slides windows of length $2 \times \mathrm{bandwidth}$ across the genome to find regions with tags more than mfold enriched relative to a random tag genome distribution

- bandwidth and mfold are user parameters
- **mfold** specifies an interval of high-confidence enrichment ratio against the background on which to build the model. The default value 10, 30 means that a model will be built on the basis of regions having read counts that are 10- to 30-fold of the background.
- **bandwidth**, which is half of the sliding window size used in the model-building step, is set according to the length of the fragments expected experimentally from the sonication procedure

MACS: Shift size

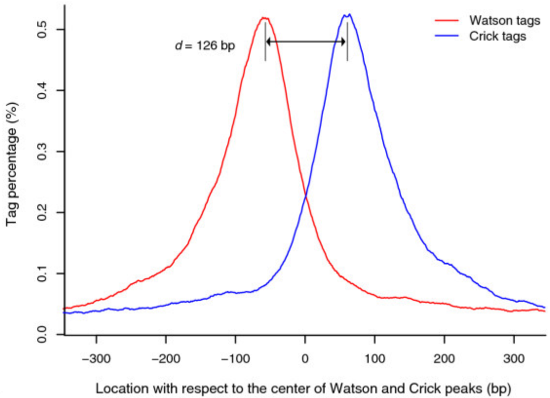
Algorithm 1 Estimate Fragment Size

```
Slide a window of $2 \times \text{bandwidth}$[3] across genome
Identify regions of moderate enrichment (mfold: 10-30 fold)
for each peak $i$ of 1000 randomly chosen enriched regions do
    separate reads into + and - strand
    Calculate mode of + and - summit
    $d_i \leftarrow |\text{mode}_+ - \text{mode}_-|$
end for
$d \leftarrow \text{average}_i(d_i)$
```

- Thus, the distance between bimodal summits is assumed to be the the estimated DNA fragment size *d*

[3] roughly twice the size of the sheared chromatinacross the genome

MACS: Shift size



MACS: Shift size

Once d has been estimated, all reads are shifted by $d/2$ to their 3' end, i.e., towards the center of the overall peak.

- A statistical test is then used to determine significant peaks
- A dynamic λ_{local} is defined to capture local biases in the genome.

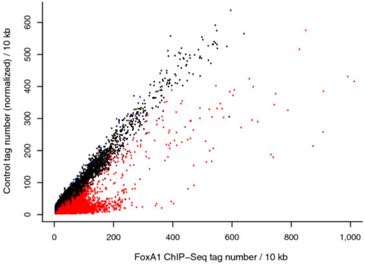
ChIP-Seq: Background bias

Similar to the situation with read-depth analysis in genome sequencing, local characteristics of the genome can lead to a bias in the number of reads being mapped.

- chromatin state (e.g. euchromatin fragments easier than silenced chromatin)
- GC content
- Therefore, ChIP-Seq experiments often include a control sample, consisting of the he input material of the ChIP processed with an unspecific immunoprecipitation with "generic" (i.e., mixed) IgG

ChIP-Seq: Background bias

Similar to the situation with read-depth analysis in genome sequencing, local characteristics of the genome can lead to a bias in the number of reads being mapped.



The tag count in ChIP versus control in 10 kb windows across the genome. Each dot represents a 10 kb window; red dots are windows containing ChIP peaks and black dots are windows containing control peaks

MACS: Peak calling

Because of these biases, instead of using a uniform λ_{BG} estimated from the whole genome, MACS uses a dynamic parameter, λ_{local} , defined for each candidate peak as:

$$\lambda_{\text{local}} = \max(\lambda_{BG}, \lambda_{1k}, \lambda_{5k}, \lambda_{10k}) \tag{2}$$

- λ_{BG} is calculated over the entire genome, and $\lambda_{1k}, \lambda_{5k}, \lambda_{10k}$ are calculated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control sample.

MACS: Peak calling

λ_{local} reduces the influence of local biases, and is robust against occasional low tag counts at small local regions. MACS uses λ_{local} to calculate the p-value of each candidate peak.

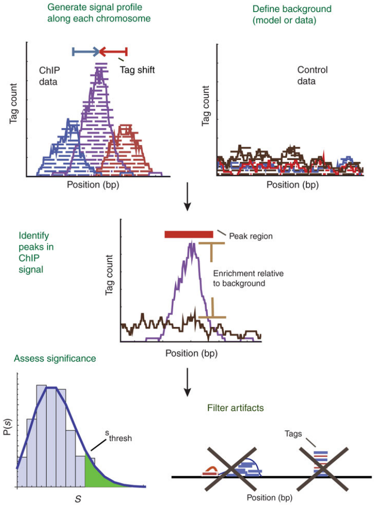
- Candidate peaks with p-values below a user-defined threshold p-value (default 10^{-5}) are called (Poisson distribution)
- The ratio between the ChIP-Seq tag count and λ_{local} is reported as the fold_enrichment.

ChIP-Seq: Artifacts

It may also be useful to filter out certain classes of peaks that are likely to be artifacts

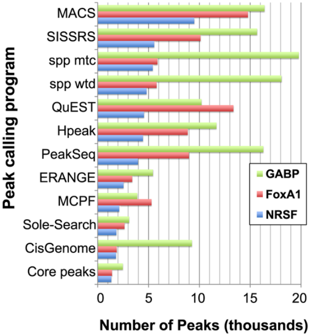
- Peaks with many reads starting from the same position
- Peaks with reads mainly from only one strand

Pepke S et al. (2009) Computation for ChIP-seq and RNA-seq studies *Nature Methods* **6**:S22–S32



ChIP-Seq: An unsolved problem

ChIP-Seq programs report different numbers of peaks, when run with their default or recommended settings on the same dataset.



Wilbanks EG (2010) Evaluation of Algorithm Performance in ChIP-Seq Peak Detection *PLoS ONE* **5**:e11471.

Outline

1. Gene Regulatory Networks
2. ChIP-Seq
3. XSET
4. FDR
5. MACS
6. ENCODE and the Irreproducible Discovery Rate (IDR)
7. The Big Picture

ChIP-Seq: Quality Control

In real life, there are innumerable ways that experiments can go wrong, and an essential part of bioinformatics is quality control of genomics data.

Essential Q/C parameters

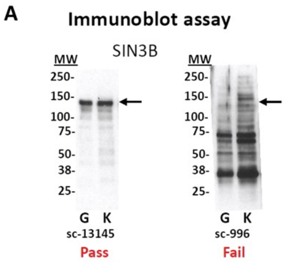
- Biological reproducibility
- Enrichment factor of immunoprecipitation
- Size and uniformity of fragmentation
- Library size and read count
- PHRED quality profile of reads
- weird stuff that nobody understands . . .

ChIP-Seq: Quality Control

We will discuss a few bioinformatic Q/C measures from

Landt SG et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**:1813-31.

ChIP-Seq: Why do we need Q/C?



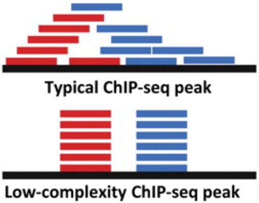
- Lanes contain nuclear extract from GM12878 cells (G) and K562 cells (K). Arrows indicate band of expected size of 133 kDa for transcription factor SIN3B.
- The primary reactive band should contain at least 50% of the signal and ideally correspond to the expected size of the protein
- A number of other wetlab Q/C measures are discussed in the paper

ChIP-Seq: Experimental Planning

A practical goal is to maximize site discovery by optimizing immunoprecipitation and sequencing deeply, within reasonable expense constraints.

- Different TFs and enhancer sequences have different binding affinities, so it is not possible to provide a one-size for all recommendation for sequencing depth, but for mammals, each replicate should generally have at least 10 million mappable reads.
- Library complexity: are there a lot of duplicate reads? Obviously, the deeper one sequences, the more likely it is to obtain duplicate reads, but an elevated number of duplicates (i.e., low library complexity) can indicate that too little DNA was isolated by immunoprecipitation or that there were problems with library construction

ChIP-Seq: Library complexity



NRF: Nonredundant fraction

A useful complexity metric is the fraction of nonredundant mapped reads in a data set (nonredundant fraction or NRF), which we define as the ratio between the number of positions in the genome that uniquely mappable reads map to and the total number of uniquely mappable reads.

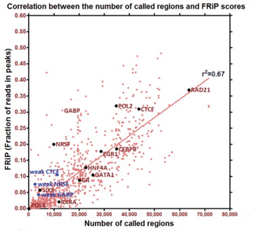
$$\mathrm{NRF} = \frac{\#\text{unique start positions of uniquely mappable reads}}{\#\text{uniquely mappable reads}} \tag{3}$$

- Note that NRF decreases with sequencing depth,
- ENCODE recommends target of $\mathrm{NRF} \geq 0.8$ for 10 million uniquely mapped reads

Measuring global ChIP enrichment (FRiP)

Typically, a minority of reads in ChIP-seq experiments occur in significantly enriched genomic regions (i.e., peaks); the remainder of the read represents background. The fraction of reads falling within peak regions is therefore a useful and simple first-cut metric for the success of the immunoprecipitation, and is called **FRiP** (fraction of reads in peaks).

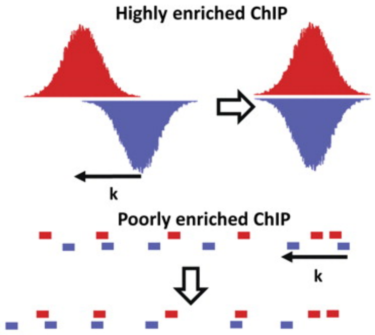
- Most (787 of 1052) ENCODE data sets have a FRiP enrichment of 1% or more when peaks are called using MACS with default parameters.
- There is a rough correlation with the number of peaks called



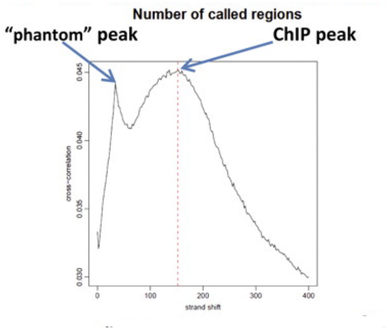
Cross-correlation analysis

High-quality ChIP-seq experiment produces significant clustering of enriched DNA sequence tags at locations bound by the protein of interest, and that the sequence tag density accumulates on forward and reverse strands centered around the binding site.

- "true signal" sequence tags are positioned at a distance k from the binding site center that depends on the fragment size distribution
- A control experiment lacks this pattern of shifted stranded tag densities

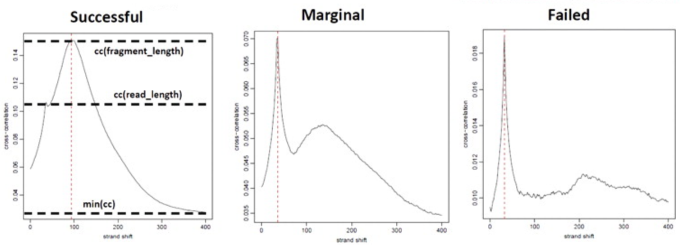


Cross-correlation analysis



- Reads are shifted in the direction of the strand they map to by an increasing number of base pairs and the Pearson correlation between the per-position read count vectors for each strand is calculated.
- This typically produces two peaks when cross-correlation is plotted against the shift value: a peak of enrichment corresponding to the predominant fragment length and a peak corresponding to the read length ("phantom" peak)

Cross-correlation analysis

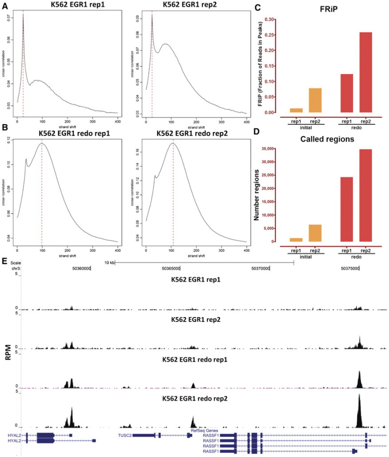


$$NSC = \frac{cc(fragment\ length)}{min(cc)} \qquad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

- The normalized ratio between the fragment-length cross-correlation peak and the background cross-correlation (normalized strand coefficient, NSC) and the ratio between the fragment-length peak and the read-length peak (relative strand correlation, RSC), are strong metrics for assessing signal-to-noise ratios in a ChIP-seq experiment.
- ENCODE cutoff: NSC values < 1.05 and RSC values < 0.8
  (repeat / reject experiments with these NSC/RSC values)

ChIP-seq: Biological replicates

Stuff happens: Sometimes the wetlab experiment simply doesn't work. Bioinformatics analysis needs to recognize this and warn the experimentalists: Garbage in garbage out!



Consistency of replicates: Analysis using IDR

Definition (IDR)

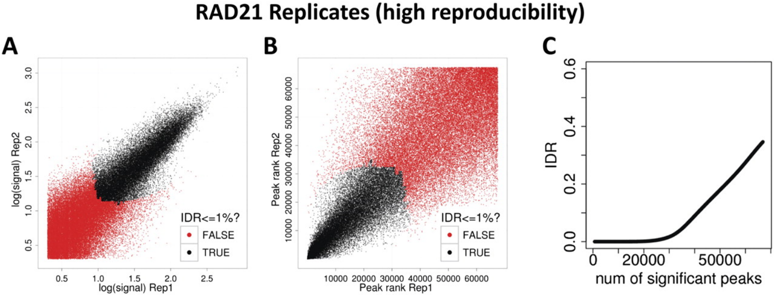
The irreproducible discovery rate (IDR) is a unified approach to measure the reproducibility of findings identified from replicate high-throughput experiments

- The scenario: We have two ChIP-seq experiments and have called peaks for each separately of them using MACS or some other tool
- Thus, each peak in each experiment has been assigned a *p*-value

Consistency of replicates: Analysis using IDR

- Each list of peaks is ranked according to p-value
- The IDR method then fits the bivariate rank distributions over the replicates in order to separate signal from noise based on a defined confidence of rank consistency and reproducibility of identifications
- We will not cover the details of the method, which was presented in Li Q (2011) Measuring reproducibility of high-throughput experiments. *Ann Appl Stat* **5**:1752–1779.

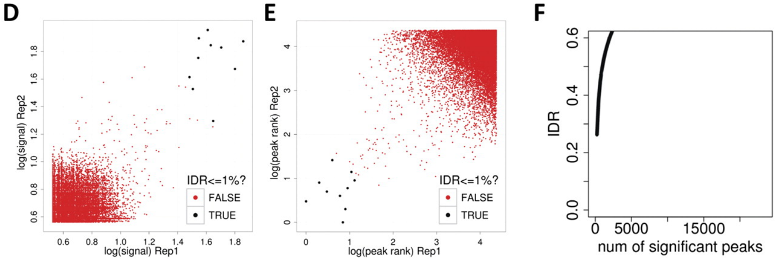
IDR: Good Quality



- Scatter plots of signal scores of peaks that overlap in each pair of replicates.
- Note that low ranks correspond to high signal and vice versa.
- **Black** data points represent pairs of peaks that pass an IDR threshold of 1%, whereas the red data points represent pairs of peaks that do not pass the IDR threshold of 1%.
- The RAD21 replicates show high reproducibility with ∼30,000 peaks passing an IDR threshold of 1%

IDR: Good Quality

SPT20 Replicates (low reproducibility)



- The SPT20 replicates show poor reproducibility with only six peaks passing the 1% IDR threshold

Outline

1. Gene Regulatory Networks
2. ChIP-Seq
3. XSET
4. FDR
5. MACS
6. ENCODE and the Irreproducible Discovery Rate (IDR)
7. The Big Picture

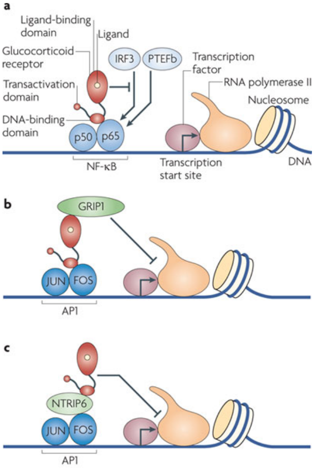
The Big Picture: Using ChIP-seq to answer biological questions

Transcription in eukaryotes involves interactions between multi-protein complexes and chromosomal DNA to coordinately regulate gene expression in a stimulus-specific, temporal, and tissue-specific fashion

- ChIP-seq is one of the most important genomics methodologies to investigate gene regulation
- We will present a bird's eye view of a nice paper on the subject: Stender JD et al (2010) Genome-wide analysis of estrogen receptor alpha DNA binding and tethering mechanisms identifies Runx1 as a novel tethering factor in receptor-mediated transcriptional activation. *Mol Cell Biol* **30**:3943-55.

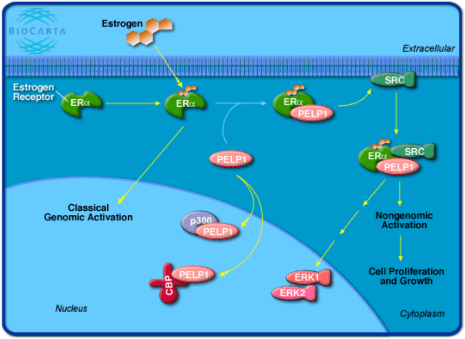
Multiprotein complexes are important for regulation

- Transcription factors have the ability to regulate gene expression by binding directly to DNA at sequence-specific response elements or by tethering to other response elements through protein-protein interactions with other DNA-bound factors
- The combinatorial usage of these response elements drives the regulation of target genes and ultimately determines stimulus and tissue specificity.



Nature Reviews | Immunology

Estrogen Receptor



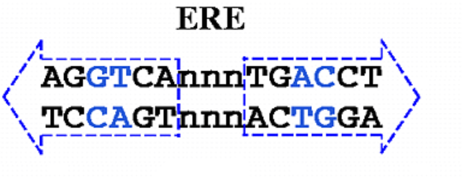
- Estrogen receptor alpha (ERα), a member of the nuclear hormone receptor family, is a ligand-activated transcription factor that controls the expression of hundreds of genes
- Two regulatory mechanisms

  – Direct binding to DNA at estrogen response elements (EREs) through its zinc finger-containing DNA binding domain

  – Protein-protein interactions with other direct DNA binding transcription factors,

Estrogen Receptor Element: ERE

The Estrogen Receptor Element (ERE) is a DNA motif to which the estrogen receptor α (ERα) can bind.

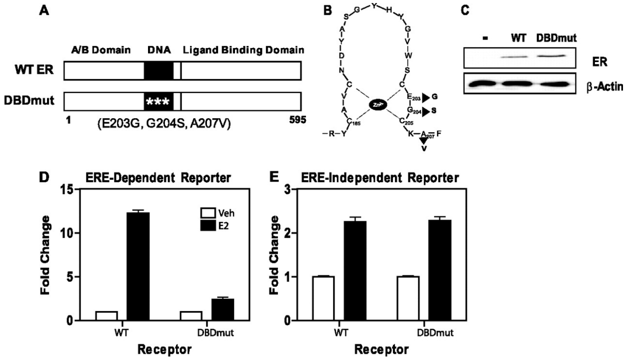


Estrogen Receptor

- The authors Stender et al. examine the genome-wide chromatin localization of a mutant nuclear hormone receptor, one in which point mutations in the DNA binding domain disable the receptor's ability to bind to its palindromic DNA response element.
- Thus, they have a molecular system to distinguish between direct DNA-binding and protein-protein interactions with indirect DNA binding

Mutant Estrogen Receptor Construct

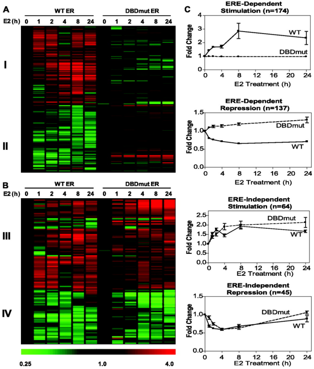
The estrogen receptor DNA binding domain mutant selectively activates ERE binding-independent estrogen signaling.



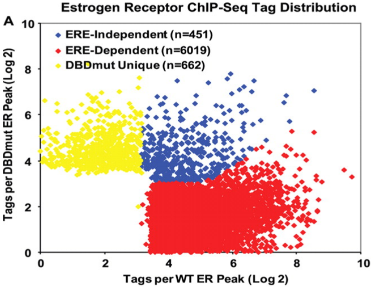
WT vs. mutant ER: Effects on gene expression

MDA-MB-231 cells stably expressing either the WT ER or DBDmut ER. Upregulated genes are shown in red, and genes downregulated by E2 are shown in green

Hierarchical clustering of these E2-regulated genes using the microarray expression data from the WT or DBDmut ER-expressing cells segregated the E2-regulated genes into two major classes: (i) genes that were regulated only by the WT ER (Fig. 2A) and (ii) genes that were regulated by both the WT ER and DBDmut ER (Fig. 2B).

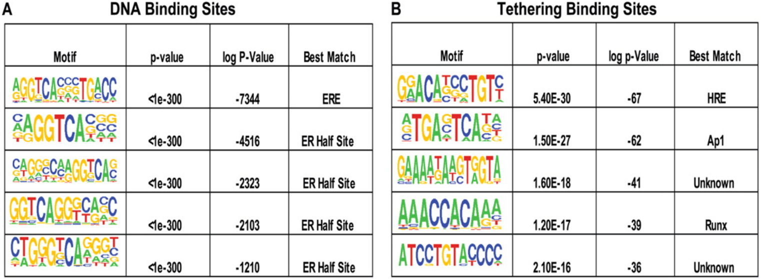


WT vs. mutant ER: ChIP-seq



- Peaks preferential for WT ER recruitment (n = 6,019) are denoted in red, while peaks common for both WT ER and DBDmut ER (n = 451) are blue. Peaks unique for the DBDmut ER (n = 662) are shown in yellow.
- The DBDmut colocalized to only 451 (7%) (blue dots) of the 6470 WT binding peaks (red dots plus blue dots), which indicates that the majority of ER recruitment to ER binding sites requires a fully functional DNA binding domain

WT vs. mutant ER: DNA binding motifs



A DNA Binding Sites

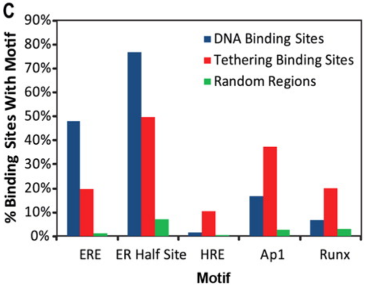
Motif	p-value	log P-Value	Best Match
AGGTCACCCTGACC	<1e-300	-7344	ERE
CAGGTCACGG	<1e-300	-4516	ER Half Site
CAGGGCAAGGTCAG	<1e-300	-2323	ER Half Site
GGTCAGGGCACC	<1e-300	-2103	ER Half Site
CTGGGTCAGGGT	<1e-300	-1210	ER Half Site

B Tethering Binding Sites

Motif	p-value	log p-Value	Best Match
GGACATCCTGTC	5.40E-30	-67	HRE
ATGAGTCATA	1.50E-27	-62	Ap1
GAAATAAGTGGTA	1.60E-18	-41	Unknown
AAACCACAAA	1.20E-17	-39	Runx
ATCCTGTACCCC	2.10E-16	-36	Unknown

- The DNA sequences corresponding to direct ER binding sites were searched for enriched motif sequences
- The ERE was the most enriched motif for WT ER (as expected)
- The tethered binding sites were investigated while using direct WT ER binding sites as a background set. In contrast to direct binding sites, the most enriched motifs for the tethering sites included Ap1, Runx, and HRE

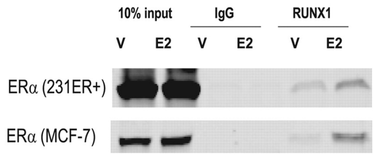
WT vs. mutant ER: DNA binding motifs



- The Ap1 motif was present in 37% of the binding sites of the DBDmut ER, while being present in only 16% of the WT ER DNA binding sites
- In addition, the Runx motif was present in 20% of DBDmut sites, while only 7% of the WT ER binding sites contained a Runx motif.
- **These data suggest that members of the Ap1 and the Runx families may be potential candidate tethering factors involved in mediating ERα-dependent gene regulation.**

Runx1 is a cofactor of ER

The observation that the Runx motif was specifically enriched in a subset of ER tethering sites suggested the possibility that Runx1 might bind to and serve as a tethering protein for ERα at distinct chromosomal locations.



- Cells were treated with vehicle (i.e., negative control) or 10 nM E2 for 45 min prior to immunoprecipitation with Runx1 antibody or IgG followed by Western immunoblotting for ERα.
- The fact that immunoprecipitation by a Runx1 antibody shows a signal, but that with IgG (also a negative control) does not indicates a binding interaction between Runx1 and ERα.

Elegant Genomics!

The paper I just presented has all of the hallmarks of an **elegant** genomics experiment!

- The experiment begins with an hypothesis
- The experimental design involves a global search or investigation[4]
- The experiment involves several interventions that allow genomic scale effects to be evaluated with at least some degree of specificity (ERα wildtype vs. mutant construct, stimulation by oestrogen vs. vehicle)
- Comprehensive and integrated bioinformatics analysis that is informed by the biological question
- The results of bioinformatics analysis lead to a targeted molecular experiment that validated the results of the bioinformatic analysis

[4] otherwise it wouldn't really be genomics . . .

Finally

- Email: peter.robinson@charite.de
- Office hours by appointment

Further reading

- Park PJ. ChIP-seq: advantages and challenges of a maturing technology (2009) *Nat Rev Genet* **10**:669-80. ⇒ excellent review
- Ibrahim DM, Hansen P, Rödelsperger C, Stiege AC, Doelken SC, Horn D, Jäger M, Janetzki C, Krawitz P, Leschik G, Wagner F, Scheuer T, Schmidt-von Kegler M, Seemann P, Timmermann B, Robinson PN, Mundlos S, Hecht J (2013) Distinct global shifts in genomic binding profiles of limb malformation-associated HOXD13 mutations. *Genome Res* **23**:2091-102. ⇒ We use ChIP-seq to investigate the pathogenesis of mutations in the transcription factor HOXD13