CS-E5875 High-Throughput Bioinformatics
ChIP-seq data analysis

Harri Lähdesmäki

Department of Computer Science
Aalto University

November 17, 2023

Contents

Transcriptional regulation

- Transcriptional regulation is largely controlled by protein-DNA interactions

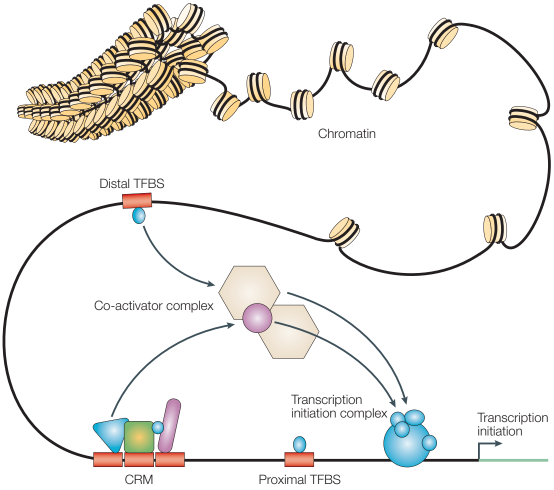


Figure from (Wasserman & Sandelin, 2004)

Transcriptional regulation

- Transcriptional regulation is largely controlled by protein-DNA interactions

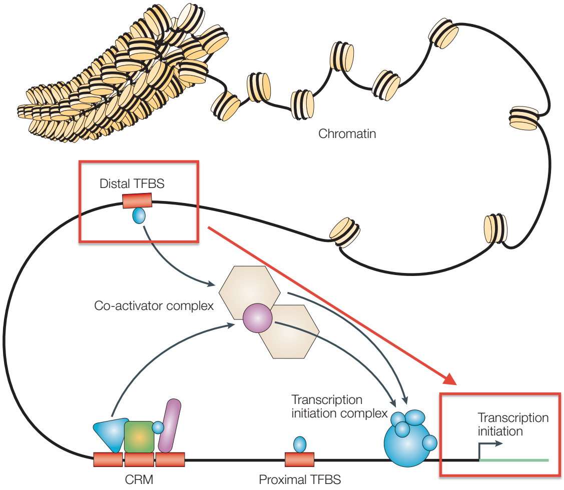


Figure from (Wasserman & Sandelin, 2004)

Protein-DNA binding

- A transcription factor (TF) is a protein that binds to DNA in a sequence specific manner
  - E.g. GATA2 protein preferentially recognizes and binds sequences ...[T/A]GATA[A/G]...
- TFs can:
  - Function alone or with other proteins
  - Recruit other co-factors to bind DNA
  - Activate or repress gene expression
  - ...

Protein-DNA binding

- Transcription factors contain DNA-binding domain(s) (DBDs) that encode their DNA-binding specificities

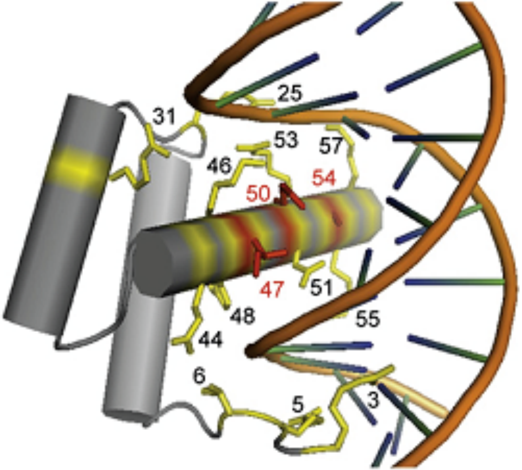


Figure from (Kissinger et al., 1990)

Modeling transcriptional regulation

- The goal
  - An accurate method to measure locations where a specific protein bind DNA
- Challenges
  - Human genome contains about 3 billion (3×10^9!) nucleotides
  - $\rightarrow$ Lots of putative binding sites
  - Human genome is physically about 2 meters long, packed in a cell nucleus with an average diameter in the range of micrometers
  - $\rightarrow$ Parts of the nucleus are densely packed and thus not available for TFs to interact

Modeling transcriptional regulation

- The goal
  - An accurate method to measure locations where a specific protein bind DNA
- Challenges
  - Human genome contains about 3 billion (3×10^9!) nucleotides
  - $\rightarrow$ Lots of putative binding sites
  - Human genome is physically about 2 meters long, packed in a cell nucleus with an average diameter in the range of micrometers
  - $\rightarrow$ Parts of the nucleus are densely packed and thus not available for TFs to interact
- Protein-DNA binding can be studied using e.g.
  - Biophysics: all atom-level modeling
  - Probabilistic models for biological sequences
  - Biological experiments + statistical analysis:
    - ChIP-seq, protein binding microarray, high-throughput SELEX, chromatin accessibility

Contents

ChIP-seq

- For any given condition, how do we find the genomic locations where DNA binding proteins bind?
- The current state-of-the-art method: chromatin immunoprecipitation followed by sequencing (ChIP-seq)
- ChIP-seq can identify genomic binding locations for a single DNA binding protein at a time
- The basic principle:
  1. Use a specific antibody to label a protein of interest
  2. Fragment the DNA (with proteins still binding the DNA)
  3. ChIP step enriches for those proteins that are bound/labeled by the antibody
  4. Extract DNA fragments from the enriched proteins
  5. These DNA fragments are then sequenced

ChIP-seq protocol

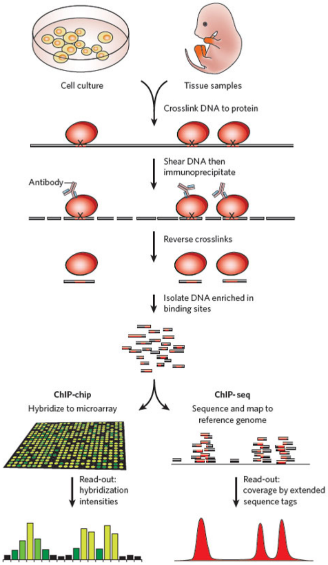


Figure from (Visel et al., 2009)

ChIP-seq steps:

- Crosslink DNA-binding proteins with DNA in vivo
- Shear the chromatin into small fragments (e.g. 200bp-1000bp) amenable for sequencing (sonication)
- Immunoprecipitate the DNA-protein complex with a specific antibody
- Reverse the crosslinks
- Assay enriched DNA to determine the sequences bound by the protein of interest

ChIP-seq protocol again

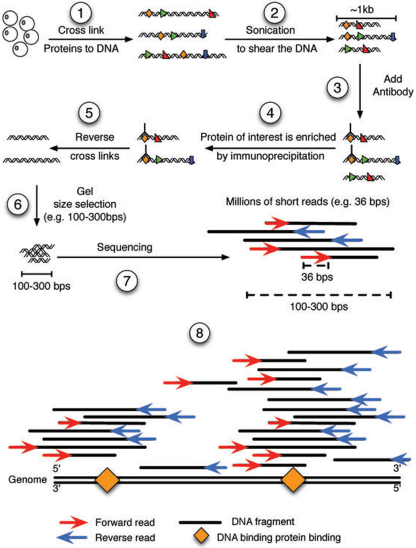


Figure from (Zhang et al., 2011)

Strand specificity and read density visualization

- A "data view" of protein-DNA binding

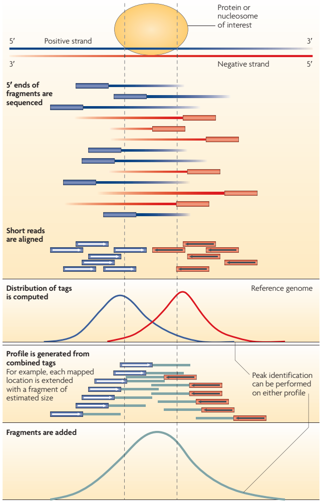


Figure from (Park, 2009)

Contents

Identification of binding sites from ChIP-seq data

- First steps in ChIP-seq data analysis:
  - Quality control, and short read alignment
- Quantify read coverage (also called read density), which refers to "pile-up" of aligned reads along genome (see previous lectures)
- Given read coverages/densities on both strands along genome, the actual data analysis task involves identification of the protein binding sites
- Given the above information about the experimental steps, we should expect to see two "signal peaks" on opposite DNA strands within a proper distance

  $\rightarrow$ This analysis is often called "peak detection"

Identification of binding sites from ChIP-seq data

- But how much signal (how many reads) in a putative genomic region is considered enough to call a protein-DNA interaction site?
- What affects the signal strength?
  1. Protein binding in the first place
  2. Sequencing depth (i.e., total number of sequencing reads)
  3. Chromatin accessibility
  4. Fragmentation efficiency
  5. Mappability (i.e., uniqueness) of a local genomic region
- All these aspects affect binding locally, i.e., not uniformly along the whole genome

ChIP-seq controls

- The best way to assess significance of a signal at putative binding sites is to use a control for ChIP-seq
  - Input-DNA: sequencing data of the (fragmented) genomic DNA from the same sample without any antibody/immunoprecipitation
  - ChIP-seq experiment with an unspecific antibody which does not detect any specific protein
- ChIP-seq controls can be used to account for many of the biases (e.g. biases 3–5 listed on the previous page) which affect the signal strength
- Input-DNA is currently considered to be the best control

Detecting binding sites from ChIP-seq data

- Early methods used a single cut-off for signal strength or a log-fold enrichment

$$\text{score} = \log \frac{\#\ \text{ChIP-seq reads in a genomic region}}{\#\ \text{Input DNA reads in a genomic region}}$$

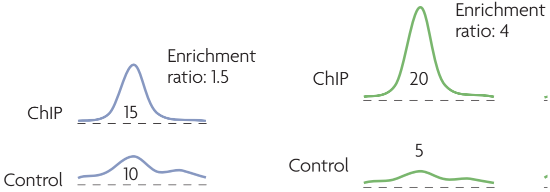


Figure from (Park, 2009)

- Current state-of-the-art methods are probabilistic

Model-based Analysis of ChIP-Seq (MACS)

- A commonly used method for detecting TF binding sites from ChIP-seq data: MACS (Zhang et al, 2008)
- Analyzes each biological sample separately
- Note: here words "sequencing read" and "tag" are used interchangably

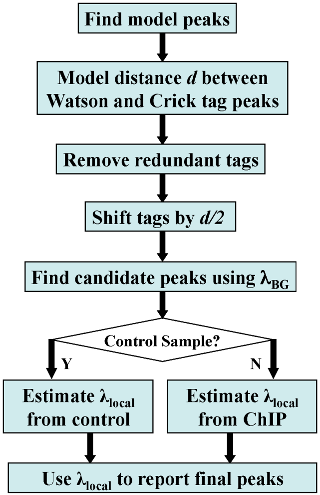


Figure from (Zhang et al., 2008)

Model-based Analysis of ChIP-Seq (MACS)

Find model peaks:

- Define two parameters, $\mathrm{mfold}_{\mathrm{low}}$ and $\mathrm{mfold}_{\mathrm{high}}$, to find genomic regions with high confidence fold-enrichment
- $\mathrm{bandwidth} =$ assumed sonicated DNA fragment size
- MACS slides $2 \times \mathrm{bandwidth}$ window across the genome to find genomic regions that satisfy:

$$\mathrm{mfold}_{\mathrm{low}} \leq \text{exp(score)} \leq \mathrm{mfold}_{\mathrm{high}}$$

- The first inequality identifies high confidence binding sites
- The second inequality filters out putative artefacts, such as PCR duplicates

Model-based Analysis of ChIP-Seq (MACS)

Model the shift size of ChIP-seq tags

- Take 1000 high confidence genomic regions (randomly) from the previous step
- Separate sequencing reads that are aligned to Watson and Crick
- Align the reads by the mid point between their Watson and Crick tag centers
- Find d: distance between the modes of the Watson and Crick peaks in the alignment

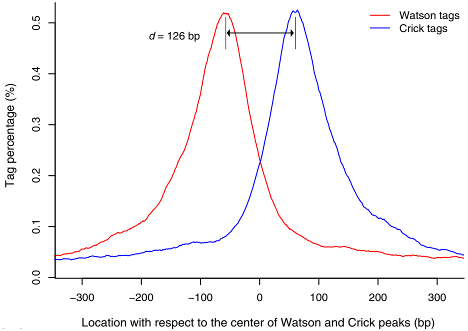


Figure from (Zhang et al., 2008)

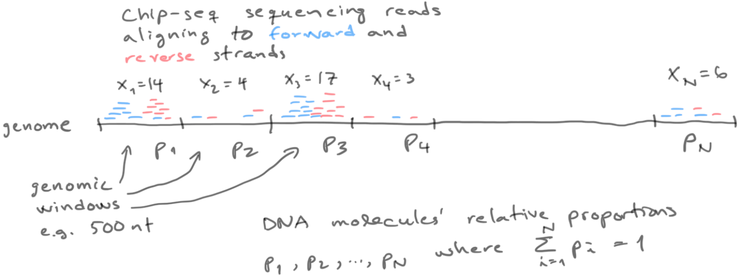
Model-based Analysis of ChIP-Seq (MACS)

- Shift all reads by $d/2$ toward the 3' ends: the most likely protein-DNA interaction sites
- An alternative strategy could be to extend all aligned sequencing reads to length d
- Remove redundant tags:
  - Sometimes the same read can be sequenced repeatedly, more than expected from a random genome-wide tag distribution
  - Such reads might arise from biases during ChIP-DNA amplification and sequencing library preparation (PCR duplicates)
  - These are likely to add noise to the final peak calls
  - MACS removes duplicate reads in excess of what is warranted by the sequencing depth (binomial distribution p-value $< 10^{-5}$)
  - For example, for the 3.9 million ChIP-seq reads, MACS allows each genomic position to contain no more than one tag and removes all the redundancies

Model-based Analysis of ChIP-Seq (MACS)

Identifying the most likely binding sites

- Counting process is exactly analogous to that of RNA-seq counting process
- Assume: reads are sampled independently from a population with fixed probabilities $(p_1, \ldots, p_N)$ for all N genomic locations ($\sum_{i=1}^{N} p_i = 1$)
- Then, the read counts $x_1, x_2, \ldots, x_N$ across the genomic locations/windows follow the multinomial distribution (total number of reads is $\sum_{i=1}^{N} x_i = n$)
- For a single genomic location i, the read count x_i follows the binomial distribution with $p = p_i$, which can be approximated by the Poisson distribution



Binomial and Poisson distributions

- Recall the definition of the binomial distribution (of a random variable X)

$$\mathrm{Binomial}(k; p, n) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Consider the mean of the binomial $E(X) = \sum_{x=0}^{n} x \cdot P(X = x) = np$ and denote the mean by λ

$$\lambda = np \Leftrightarrow p = \frac{\lambda}{n}$$

- Substitute $p = \frac{\lambda}{n}$ into the binomial distribution and take limit $n \to \infty$

Binomial and Poisson distributions

- We have

$$
\begin{aligned}
\lim_{n\to\infty} P(X=k) &= \lim_{n\to\infty} \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \left(\frac{\lambda^k}{k!}\right) \lim_{n\to\infty} \frac{n!}{(n-k)!}\left(\frac{1}{n^k}\right)\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} \\
&= \left(\frac{\lambda^k}{k!}\right) \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} \\
&= \left(\frac{\lambda^k}{k!}\right) \lim_{n\to\infty} \underbrace{\left(\frac{n^k+O(n^{k-1})}{n^k}\right)}_{\to 1} \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{e^{-\lambda *}} \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{\to 1} \\
&= \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}
$$

---

*Because $\lim_{x\to\infty}\left(1+\frac{1}{x}\right)^x = e$

Binomial and Poisson distributions

- Poisson approximation to binomial distribution is accurate when n is large and p is small
- Poisson approximation is convenient because is has only a single parameter λ

Model-based Analysis of ChIP-Seq (MACS)

- Let x_i denote the number of sequencing reads in the ith position / window in a genome
- Each genomic window is analyzed independently

$$x_i \sim \mathrm{Poisson}(\cdot|\lambda_{\mathrm{BG}}) = \frac{\lambda_{\mathrm{BG}}^{x_i}}{x_i!} e^{-\lambda_{\mathrm{BG}}}, \quad x_i = 0, 1, 2, \ldots$$

where λ_{BG} is the rate of observing reads in the control sample along the whole genome

Model-based Analysis of ChIP-Seq (MACS)

- Let x_i denote the number of sequencing reads in the ith position / window in a genome
- Each genomic window is analyzed independently

$$x_i \sim \mathrm{Poisson}(\cdot|\lambda_{\mathrm{BG}}) = \frac{\lambda_{\mathrm{BG}}^{x_i}}{x_i!} e^{-\lambda_{\mathrm{BG}}}, \quad x_i = 0, 1, 2, \ldots$$

  where λ_{BG} is the rate of observing reads in the control sample along the whole genome
- MACS linearly scales the total number of sequencing reads in the control experiment N_{control} and in the ChIP experiment N_{ChIP}, i.e.,

$$\lambda_{\mathrm{BG}} := N_{\text{ChIP}} / N_{\text{control}} \cdot \lambda_{\mathrm{BG}}$$

Model-based Analysis of ChIP-Seq (MACS)

- Let x_i denote the number of sequencing reads in the ith position / window in a genome
- Each genomic window is analyzed independently

$$x_i \sim \mathrm{Poisson}(\cdot|\lambda_{\mathrm{BG}}) = \frac{\lambda_{\mathrm{BG}}^{x_i}}{x_i!} e^{-\lambda_{\mathrm{BG}}}, \quad x_i = 0, 1, 2, \ldots$$

  where λ_{BG} is the rate of observing reads in the control sample along the whole genome
- MACS linearly scales the total number of sequencing reads in the control experiment N_{control} and in the ChIP experiment N_{ChIP}, i.e.,

$$\lambda_{\mathrm{BG}} := N_{\mathsf{ChIP}} / N_{\mathsf{control}} \cdot \lambda_{\mathrm{BG}}$$

- Because ChIP-seq data has several bias sources which vary across the genome, it is better to model the data using a "local" or "dynamic" Poisson

$$\lambda_{\mathrm{local}}^{(i)} = \max(\lambda_{\mathrm{BG}}, [\lambda_{\mathrm{1K}}^{(i)}], \lambda_{\mathrm{5K}}^{(i)}, \lambda_{\mathrm{10K}}^{(i)}),$$

  where $\lambda_{X\mathrm{K}}^{(i)}$ is estimated from the control sample (e.g. input-DNA) using the window of size $X\mathrm{K}$ centered at the ith position ($[\cdot]$ denotes an optional input argument)

Model-based Analysis of ChIP-Seq (MACS)

- Assessing statistical significance of x_i reads (in a genomic region i) using hypothesis testing
  - H_0: the ith location is not a binding site
  - H_1: the ith location is a binding site
- The p-value is the probability of observing x_i many reads or more, assuming the null hypothesis is true:

$$p - \mathrm{value} = \sum_{k=x_i}^{\infty} \mathrm{Poisson}(k|\lambda_{\mathrm{local}}^{(i)})$$

Model-based Analysis of ChIP-Seq (MACS)

- Assessing statistical significance of x_i reads (in a genomic region i) using hypothesis testing
  - H_0: the ith location is not a binding site
  - H_1: the ith location is a binding site
- The p-value is the probability of observing x_i many reads or more, assuming the null hypothesis is true:

$$p - \text{value} = \sum_{k=x_i}^{\infty} \text{Poisson}(k|\lambda_{\text{local}}^{(i)})$$

- For genomic regions for which the null hypothesis is rejected:
  - The location with the highest pileup of aligned sequencing reads (shifted by $d/2$) is used as an estimate of the nucleotide-level binding location: called summit
- The ratio between the ChIP-seq read count x_i and $\lambda_{\text{local}}^{(i)}$ is reported as the fold_enrichment

Multiple correction in MACS

- For a ChIP-seq experiment with controls, MACS empirically estimates the false discovery rate (FDR)
- At each *p*-value, MACS uses the same parameters to find
  - ChIP-seq peaks over control, and
  - Control peaks over ChIP-seq (i.e., an analysis using swapped samples)
- The empirical FDR is defined as

$$\text{empirical FDR} = \frac{\#\text{control peaks}}{\#\text{ChIP peaks}}$$

ChIP-seq peak: Illustration

- An illustration of a strong TF binding site

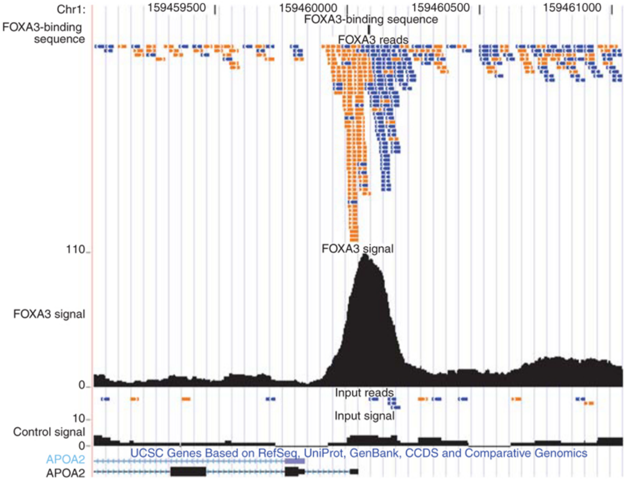


Figure from http://www.nature.com/nmeth/journal/v6/n4/images/nmeth.f.247-F2.jpg

Summary

- ChIP-seq is a powerful way to detect TF binding sites
- ChIP-seq method is limited in that
  - Only a subset of all TFs have a chip-grade antibody
  - None of the antibodies are perfect
  - A single experiment will profile a single protein
- ChP-seq can be applied to profile practically any protein / protein complex / molecule that interacts with DNA, assuming an antibody exists:
  - DNA methylation
  - RNA polymerase
  - Histone proteins / nucleosomes
  - Post-translationally modified histone proteins
  - ...

Contents

ENCODE project

- The ENCODE Project: ENCyclopedia Of DNA Elements
- Identify all functional elements in the human and mouse genomes
- Large amounts of functional and epigenetic data from several number of cell types/lines

ENCODE project

- Large amounts of functional and epigenetic data from several number of cell types/lines

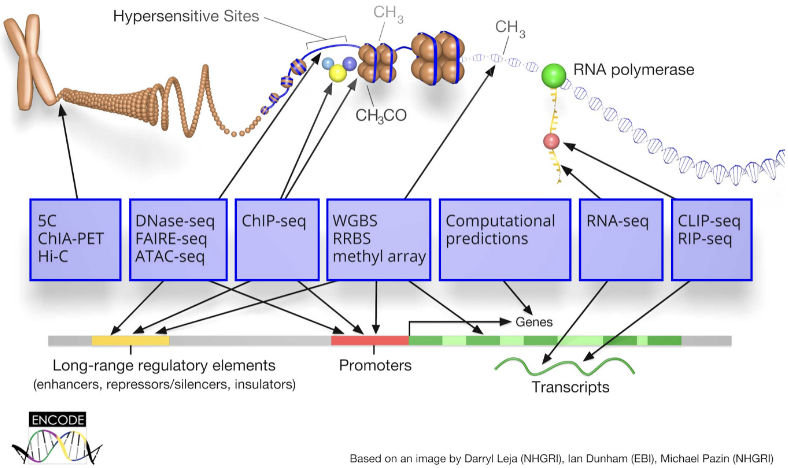


Figure from `https://www.encodeproject.org`

ENCODE project

- Large amounts of functional and epigenetic data from several number of cell types/lines

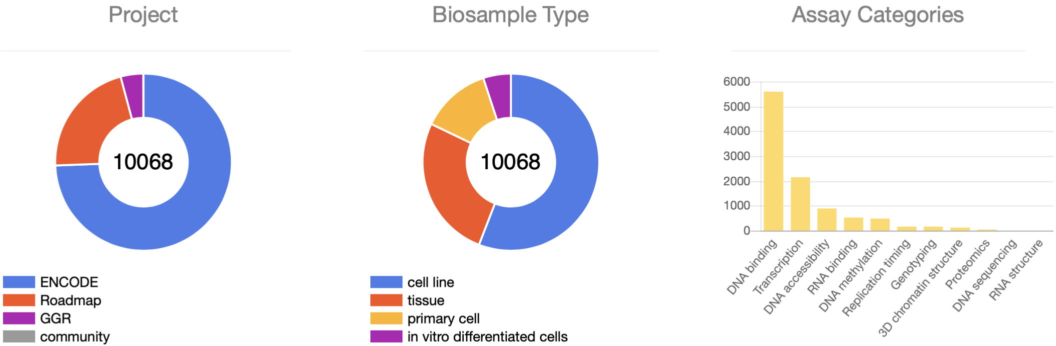


Figure from https://www.encodeproject.org

ENCODE project

Understand non-coding disease associated variants

- Co-localization of SNPs in protein-DNA interaction sites
- Can e.g. increase/decrease the strength of interaction and thereby affect e.g. gene transcription

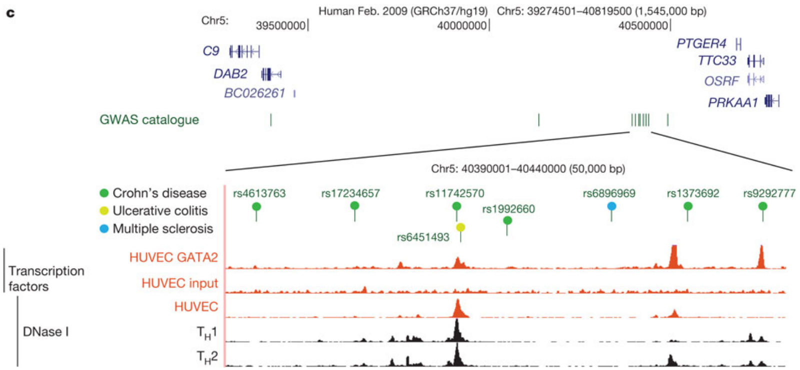
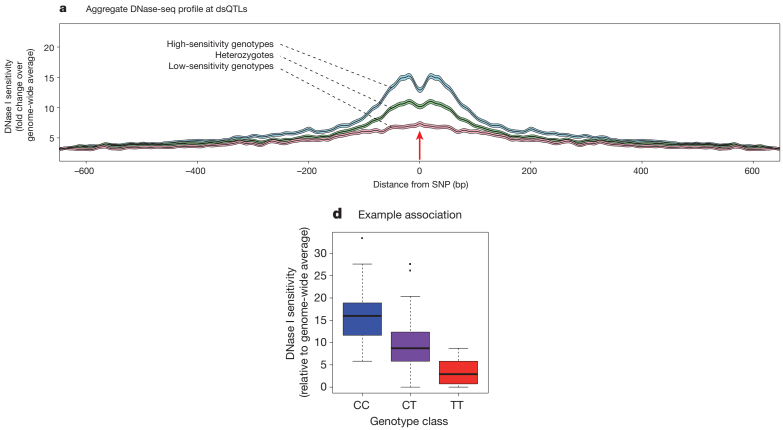


Figure from (The ENCODE Project Consortium, 2012)

Applications

Understand non-coding disease associated variants

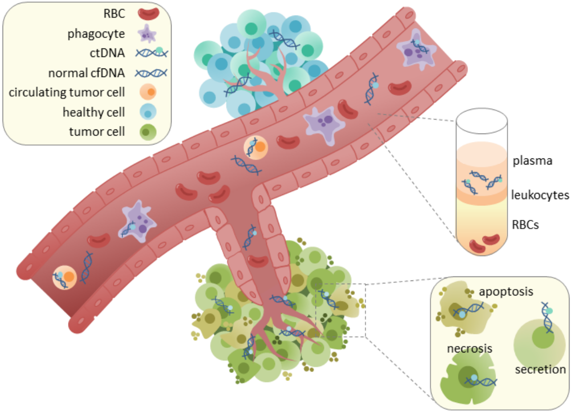
- Quantify how SNPs affect chromatin accessibility (and thus TF binding and gene transcription)



Figures from (Degner et al, 2012)

Circulating free/tumor DNA

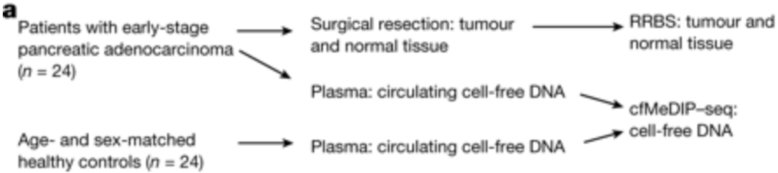
- Circulating free DNA (cfDNA) are degraded DNA fragments released to the blood plasma
- Circulating tumor DNA (ctDNA) are tumor-derived DNA fragments in the blood plasma
- Somatic mutations or epigenetic modifications in these cfDNA fragments can provide a highly accurate and sensitive non-invasive cancer diagnostics



Figures from `https://en.wikipedia.org/wiki/Circulating_tumor_DNA`

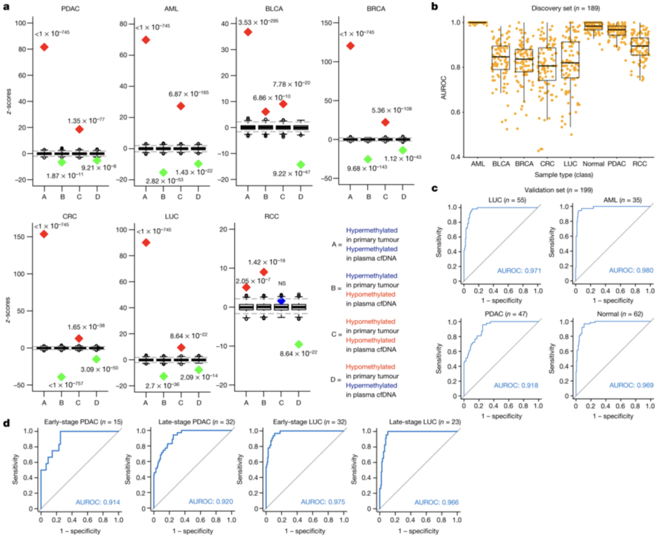
Circulating free/tumor DNA

- ChIP-seq based quantification of DNA methylation shows great potential in cancer diagnostics



Figures from (Shen et al., 2018)

Circulating free/tumor DNA



Figures from (Shen et al., 2018)

References

- Jacob F. Degner, DNase I sensitivity QTLs are a major determinant of human expression variation, Nature, 390, 482.
- The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489, 57-74, 2012.
- Metzker ML (2010) Sequencing technologies - the next generation, *Nat Rev Genet.* 11(1):31-46.
- Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology, *Nat Rev Genet.* 10(10):669-80.
- Shu Yi Shen, et al., (2018) Sensitive tumour detection and classification using plasma cell-free DNA methylomes, *Nature*, 563:579–583.
- Axel Visel, Edward M. Rubin & Len A. Pennacchio (2009) Genomic views of distant-acting enhancers," *Nature* 461, 199-205.
- Zhang Y et al. (2008), Model-based analysis of ChIP-Seq (MACS), *Genome Biol.* 9(9):R137.
- Zhang X et al. (2011) PICS: probabilistic inference for ChIP-seq, *Biometrics*, 67(1): 151-163.
- Wyeth W. Wasserman & Albin Sandelin, Applied bioinformatics for the identification of regulatory elements, Nature Reviews Genetics 5, 276-287, 2004.