EMBL-EBI

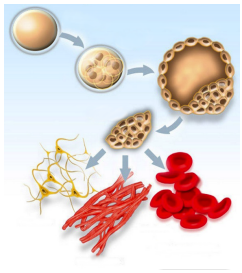
e!Ensembl

UNDERSTANDING GENE REGULATION – PART I

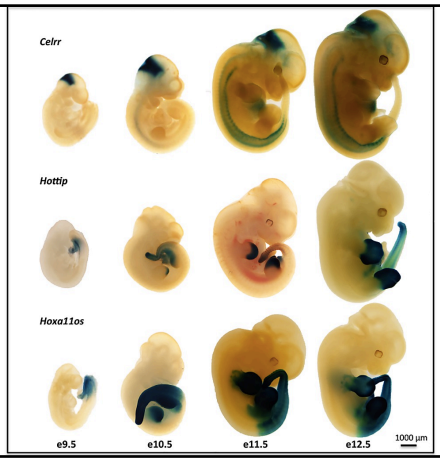
Myrto Kostadima

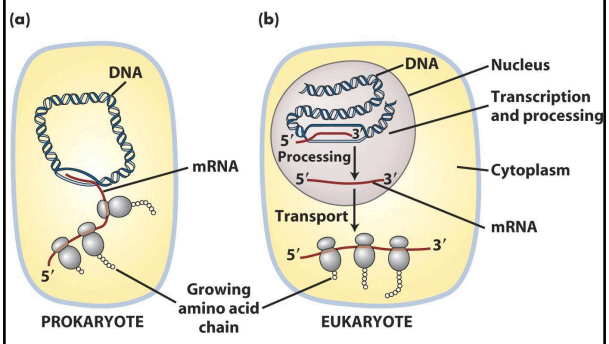
Ensembl Regulation Project Leader, EMBL-EBI

One Genome – Many cell types



- Essentially all cells of an organism share
  - the same genome
  - the same genes
- Hundreds of different cell types with a clearly distinct phenotype
- Difference?
  - Gene expression profiles





4

http://barleyworld.org/sites/default/files/figure-08-01_0.jpg

Gene regulation (almost sole) source of complexity

5

- Bacteria: 1,500-7,500 genes
- Drosophila: 14,000 genes
- C. elegans: 19,000
- Human: ?

2003

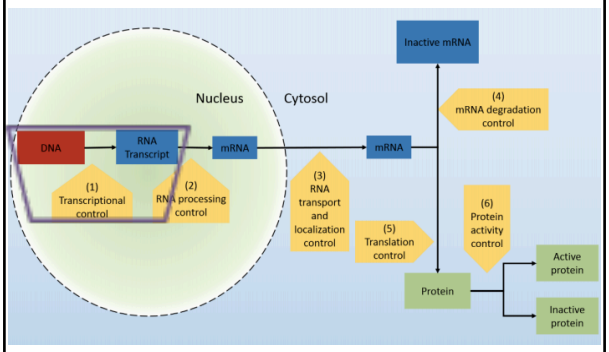
Completion of the Human Genome Project

Human genes: ~35,000

Human genes: > 100,000

Human genes: ~21,000

6

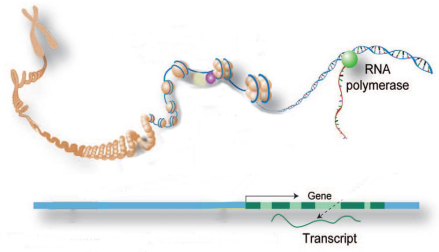


7

Alberts B et al., "Molecular Biology of the Cell", Sixth edition (2014)

From protein-coding DNA..

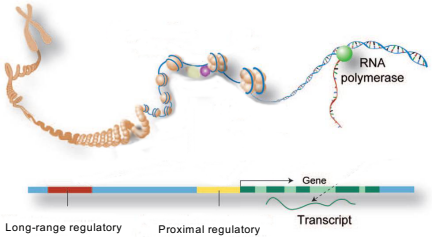
8



Adapted from "A user's guide to the encyclopedia of DNA elements (ENCODE).": PLoS Biol. 2011 Apr;9(4):e1001046

.. to regulatory DNA

9



Long-range regulatory elements (enhancers, repressors/silencers, insulators)

Proximal regulatory elements (promoters, transcription factor binding sites)

Adapted from "A user's guide to the encyclopedia of DNA elements (ENCODE).": PLoS Biol. 2011 Apr;9(4):e1001046

Elements of gene regulation

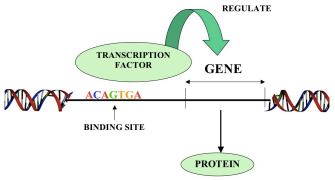
10

- Transcription factors
  - Promoters, enhancers, insulators..
- RNA Pol II machinery
- Epigenetics
  - DNA methylation
  - Nucleosome positioning/Open chromatin
  - Histone modification
  - Euchromatin/heterochromatin
- 3D Chromatin organisation

Transcription factors (TFs)

11

- TFs are a subset of genes that encode proteins that bind to sequences to regulate transcription



  - ~1, 000 human genes are TFs (depending on definition)

Vaquerizas JM et al., Nat Rev Genet. 2009 Apr;10(4):252-63.

Prokaryotes vs Eukaryotes

12

- simple process
- genes organised in operons
- one TF per operon

- complex process
- many TFs per gene
- many regulatory elements per gene

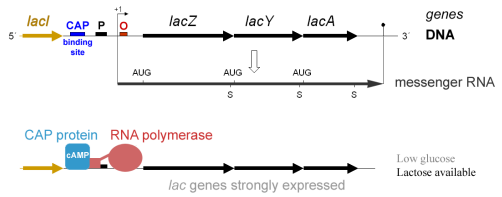
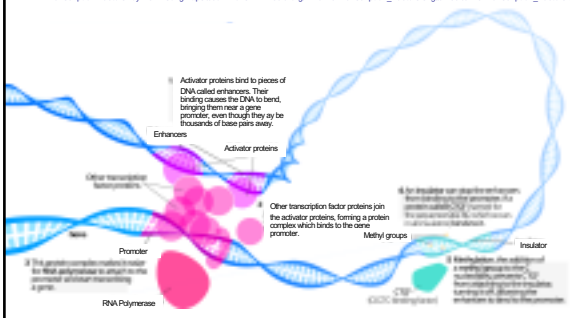


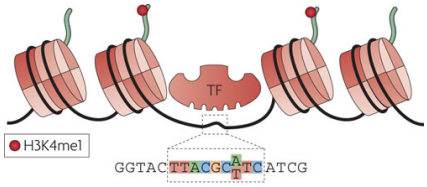
Image adapted from: https://commons.wikimedia.org/wiki/File:Lac_operon-2010-21-01.png#/media/File:Lac_operon-2010-21-01.png

TFs and regulatory regions

"Transcription Factors" by Kelvinsong https://commons.wikimedia.org/wiki/File:Transcription_Factors.svg#/media/File:Transcription_Factors.svg



TF binding site and motif



Nature Reviews | Genetics

Position Weight Matrices

15

GAGGTAAAC
TCCGTAAGT
CAGGTTGGA
ACAGTCAGT
TAGGTCATT
TAGGTACTG
ATGGTAACT
CAGGTATAC
TGTGTGAGT
AAGGTAAGT

Consensus sequence

TAGGTAAGT

- A PWM is a commonly used representation of motifs in biological sequences.
- PWMs are often derived from a set of aligned sequences that are thought to be functionally related.
- They have become an important part of many software tools for computational motif discovery.

Position Weight Matrices

16

```
POS 1|2|3|4|5|6|7|8|9
 1. G|A|G|G|T|A|A|A|C
 2. T|C|C|G|T|A|A|G|T
 3. C|A|G|G|T|T|G|G|A
 4. A|C|A|G|T|C|A|G|T
 5. T|A|G|G|T|C|A|T|T
 6. T|A|G|G|T|A|C|T|G
 7. A|T|G|G|T|A|A|C|T
 8. C|A|G|G|T|A|T|A|C
 9. T|G|T|G|T|G|A|G|T
10. A|A|G|G|T|A|A|G|T
```

```
POS 1 2 3  4  5 6 7 8 9
A [ -  - - -  - - -  - - ]
C [ -  - - -  - - -  - - ]
G [ -  - - -  - - -  - - ]
T [ -  - - -  - - -  - - ]
```

Position Weight Matrices

17

```
POS 1|2|3|4|5|6|7|8|9
 1. G|A|G|G|T|A|A|A|C
 2. T|C|C|G|T|A|A|G|T
 3. C|A|G|G|T|T|G|G|A
 4. A|C|A|G|T|C|A|G|T
 5. T|A|G|G|T|C|A|T|T
 6. T|A|G|G|T|A|C|T|G
 7. A|T|G|G|T|A|A|C|T
 8. C|A|G|G|T|A|T|A|C
 9. T|G|T|G|T|G|A|G|T
10. A|A|G|G|T|A|A|G|T
```

```
POS 1 2 3  4  5 6 7 8 9
A [ 3  6 1  0  0 6  7 2 1 ]
C [ 2  2 1  0  0 2  1 1 2 ]
G [ 1  1 7 10  0 1  1 5 1 ]
T [ 4  1 1  0 10 1  1 2 6 ]
```

Position Weight Matrices

18

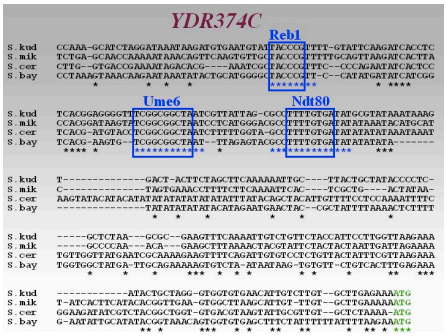
```
POS 1|2|3|4|5|6|7|8|9
 1. G|A|G|G|T|A|A|A|C
 2. T|C|C|G|T|A|A|G|T
 3. C|A|G|G|T|T|G|G|A
 4. A|C|A|G|T|C|A|G|T
 5. T|A|G|G|T|C|A|T|T
 6. T|A|G|G|T|A|C|T|G
 7. A|T|G|G|T|A|A|C|T
 8. C|A|G|G|T|A|T|A|C
 9. T|G|T|G|T|G|A|G|T
10. A|A|G|G|T|A|A|G|T
```

```
POS 1 2 3  4  5 6 7 8 9
A [ 3  6 1  0  0 6  7 2 1 ]
C [ 2  2 1  0  0 2  1 1 2 ]
G [ 1  1 7 10  0 1  1 5 1 ]
T [ 4  1 1  0 10 1  1 2 6 ]

A [ .3 .6 .1  0  0 .6 .7 .2 .1 ]
C [ .2 .2 .1  0  0 .2 .1 .1 .2 ]
G [ .1 .1 .7  1  0 .1 .1 .5 .1 ]
T [ .4 .1 .1  0  1 .1 .1 .2 .6 ]
```

Phylogenetic Footprinting

19



Systematic Evolution of Ligands by Exponential Enrichment (SELEX)

20

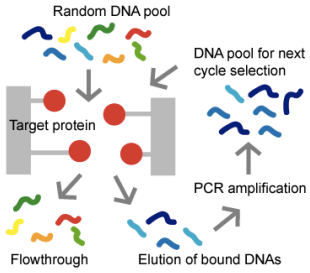
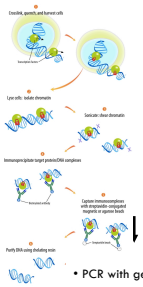


Image from: http://altair.sci.hokudai.ac.jp/g6/Projects/Selex-e.html

Chromatin ImmunoPrecipitation (ChIP)

- Transcription factors
- RNA Polymerase II

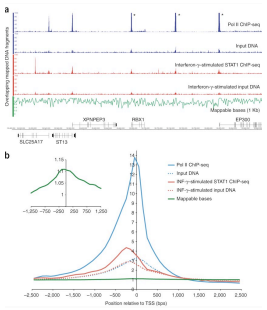


- PCR with gene specific primers
- Hybridisation on microarrays
- Sequencing

Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq)

- One of the early applications of high-throughput sequencing
- First studies published in 2007
  - Johnson et al (Science) - NRSF
  - Barski et al (Cell) - histone methylation
  - Robertson et al (Nature Methods) - STAT1
  - Mikkelsen et al (Nature) - histone modification
- Thousands of publications currently in PubMed

Why we need a control sample



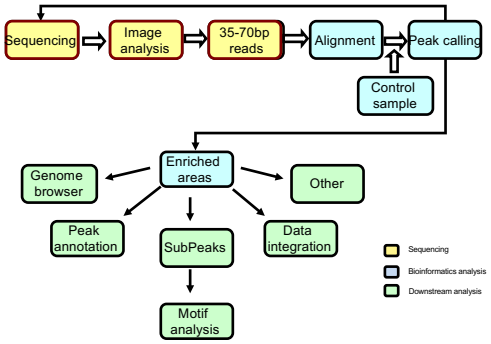
- Open chromatin regions are fragmented more easily than closed regions.
- Repetitive sequences might seem to be enriched (inaccurate repeats copy number in the assembled genome).
- Uneven distribution of sequence tags across the genome.
- A ChIP-seq peak should be compared with the same region in a matched control.

Rozowsky J et al., Nat Biotechnol. 2009 Jan;27(1):66-75.

Control types in ChIP-seq

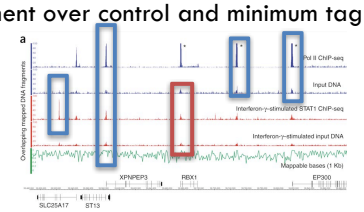
- Input DNA
- Mock IP - DNA obtained from IP without antibody
  - Very little material can be pulled down leading to inconsistent results of multiple mock Ips.
- Nonspecific IP - using an antibody against a protein that is not known to be involved in DNA binding
- Sequencing a control can be avoided when looking at:
  - time points
  - differential binding pattern between conditions

Analysis - Overview



Peak Calling

Regions are scored by the number of tags in a window of a given size. Then assessed by enrichment over control and minimum tag density.



Rozowsky J et al., Nat Biotechnol. 2009 Jan;27(1):66-75.

Peak Calling - Challenges

- Adjust for sequence mappability - regions that contain repetitive elements have different expected tag count
- Different ChIP-seq applications produce different type of peaks. Most current tools have been designed to detect sharp peaks (TF binding)
- Alternative tools exist for broader peaks (histone modifications that mark domains - transcribed or repressed), e.g. SICER

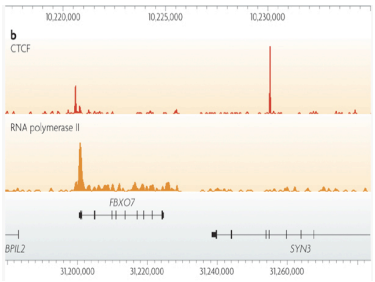
MACS - Peak detection

- Remove duplicate tags (in excess of what can be expected by chance)
- Slide window across the genome to find candidate peaks with a significant tag enrichment (Poisson distribution, global background, default p-value 10e-5)
- Merge overlapping peaks, and extend each tag d bases from its center
- Also looks at local background levels and eliminates peaks that are not significant with respect to local background
- Uses the control sample to eliminates peaks that are also present there

Analysis downstream to peak calling

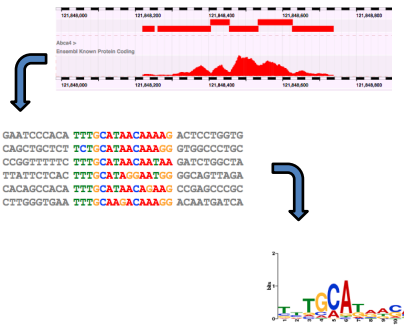
- Visualisation - genome browser: Ensembl, UCSC, IGV
- Peak Annotation - finding interesting features surrounding peak regions
- Correlation with expression data
- Discovery of binding sequence motifs
- Gene Ontology analysis on genes bound by the same transcription factor
- Correlation with SNP data to find allele-specific binding

Visualisation in a genome browser



Park J, Nature Reviews Genetics, 2009

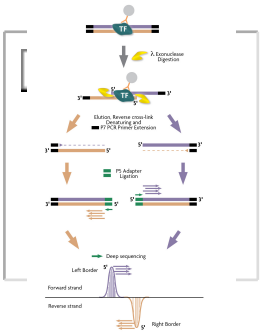
Motif Analysis



GAATCCCACA TTTGCATAACAAAAG ACTCCTGGTG
CAGCTGCTCT TCTGCATAACAAAGG GTGGCCCTGC
CCGGTTTTTC TTTGCATAACAATAA GATCTGGCTA
TTATTCTCAC TTTGCATAGGAATGG GGCAGTTAGA
CACAGCCACA TTTGCATAACAGAAG CCGAGCCCGC
CTTGGGTGAA TTTGCAAGACAAAGG ACAATGATCA

ChIP-exo

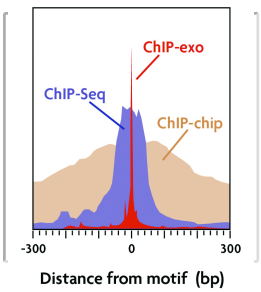
32



- ChIP-exo is a modification of the ChIP-seq protocol, improving the resolution of binding sites from hundreds of base pairs to almost one base pair.
- It employs the use of exonucleases to degrade strands of the protein-bound DNA in the 5'-3' direction to within a small number of nucleotides of the protein binding site.

ChIP-exo

33

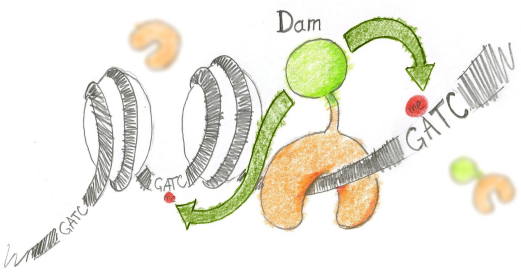


- ChIP-exo is a modification of the ChIP-seq protocol, improving the resolution of binding sites from hundreds of base pairs to almost one base pair.
- It employs the use of exonucleases to degrade strands of the protein-bound DNA in the 5'-3' direction to within a small number of nucleotides of the protein binding site.

Image adapted from: https://www.activemotif.com/images/products/ChIP-exo_Peak_image.jpg

DNA adenine methyltransferase identification

34

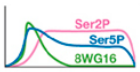


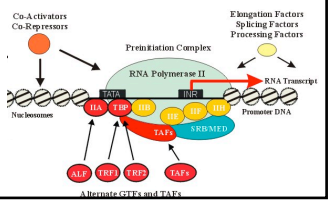
Van Steensel B et al., Nature Biotechnology 18, 424 - 428 (2000)
"DamID concept" by Guillaume Filion - Own work. Licensed under Public Domain via Commons
https://commons.wikimedia.org/wiki/File:DamID_concept.jpg#/media/File:DamID_concept.jpg

RNA Pol II and the transcriptional machinery

35

- Pol II does not act alone
- Needs a Pre-initiation complex (TFs, mediator proteins)
- Pol II is chemically modified in different phases of transcription

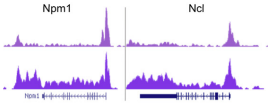




RNA Pol II pausing

36

- RNA Pol II modification & elongation factors needed for full transcripts
- RNA Pol II Pre-initiation complex can be present, but no (or low) transcription: *RNA Pol II pausing*



- RNA Pol II ChIP: Compute Pausing Index/Travelling ratio (signal at TSS vs gene body)

Further reading

37

REPORT

Suboptimization of developmental enhancers

Emma K. Farley[1,2,*], Katrina M. Olson[1,2], Wei Zhang[3], Alexander J. Brandt[4], Daniel S. Rokhsar[1], Michael S. Levine[1,2,*]

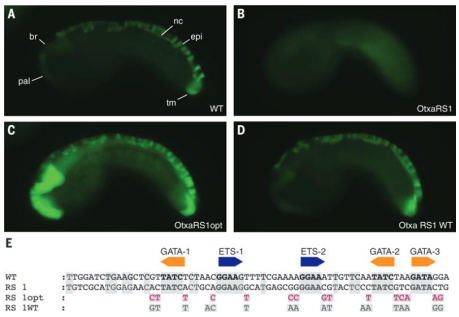
+ Author Affiliations

↩*Corresponding author. E-mail: msl2@princeton.edu (M.S.L.); ekfarley@princeton.edu (E.K.F.)

Science 16 Oct 2015:
Vol. 350, Issue 6258, pp. 325-328
DOI: 10.1126/science.aac6948

Further reading

38



39

Gene Regulation Practical

Prerequisites for next week's practical:

1. Install R packages:
    1. rtracklayer
    2. ggplo2
2. Add the following to your UNIX $PATH:
    1. /local/data/genome_informatics/programs/bedtools2