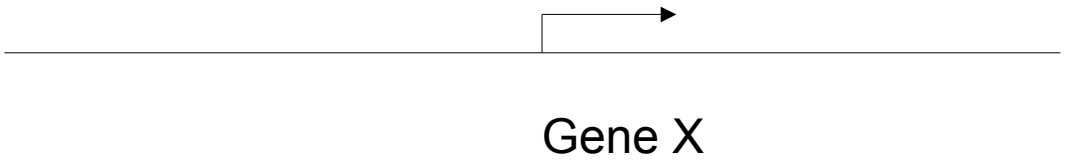
Integrative analysis of NGS data

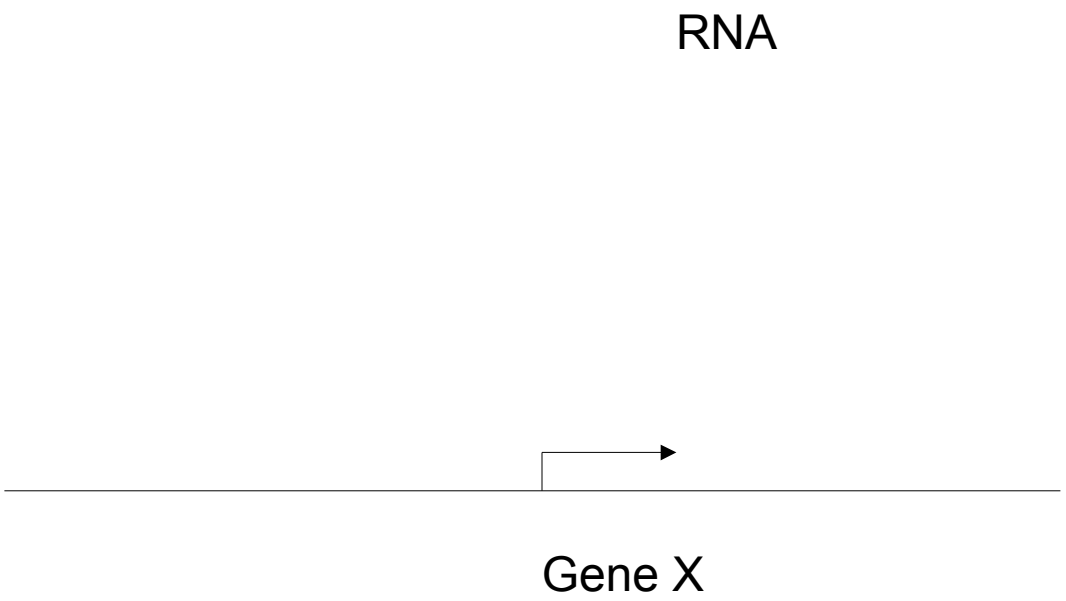
Alena van Bömmel (Alena.vanBoemmel@molgen.mpg.de R 3.3.8)
Wolfgang Kopp (kopp@molgen.mpg.de R 3.3.18)
Max Planck Institute for Molecular Genetics

Biological background

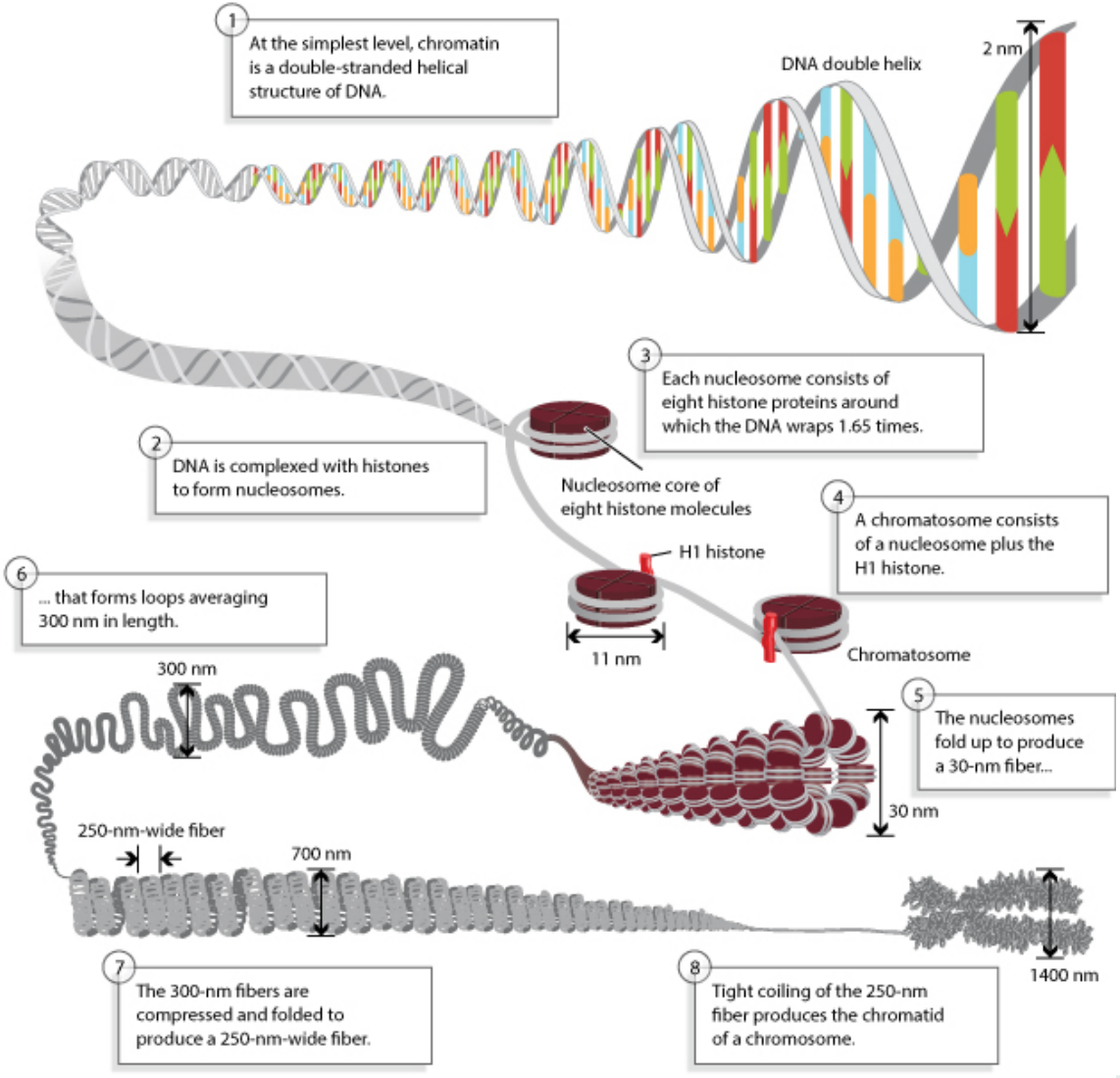
Gene expression



Gene expression

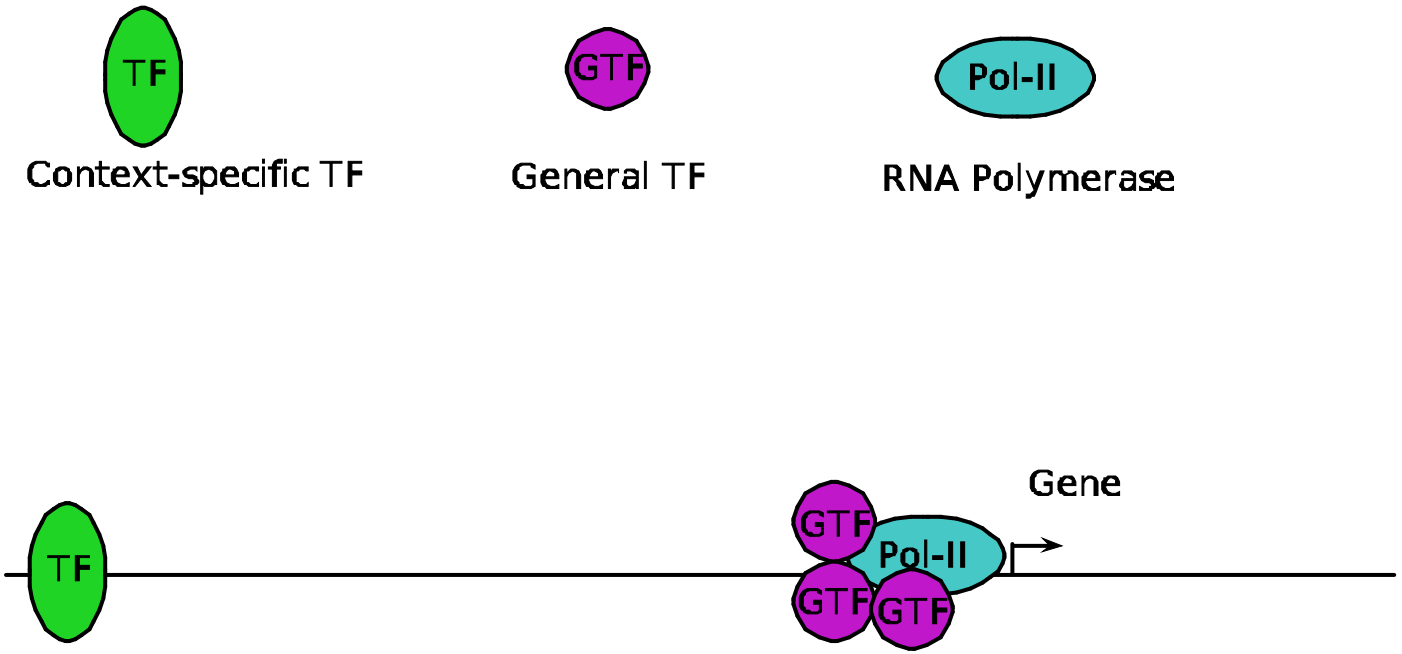


DNA

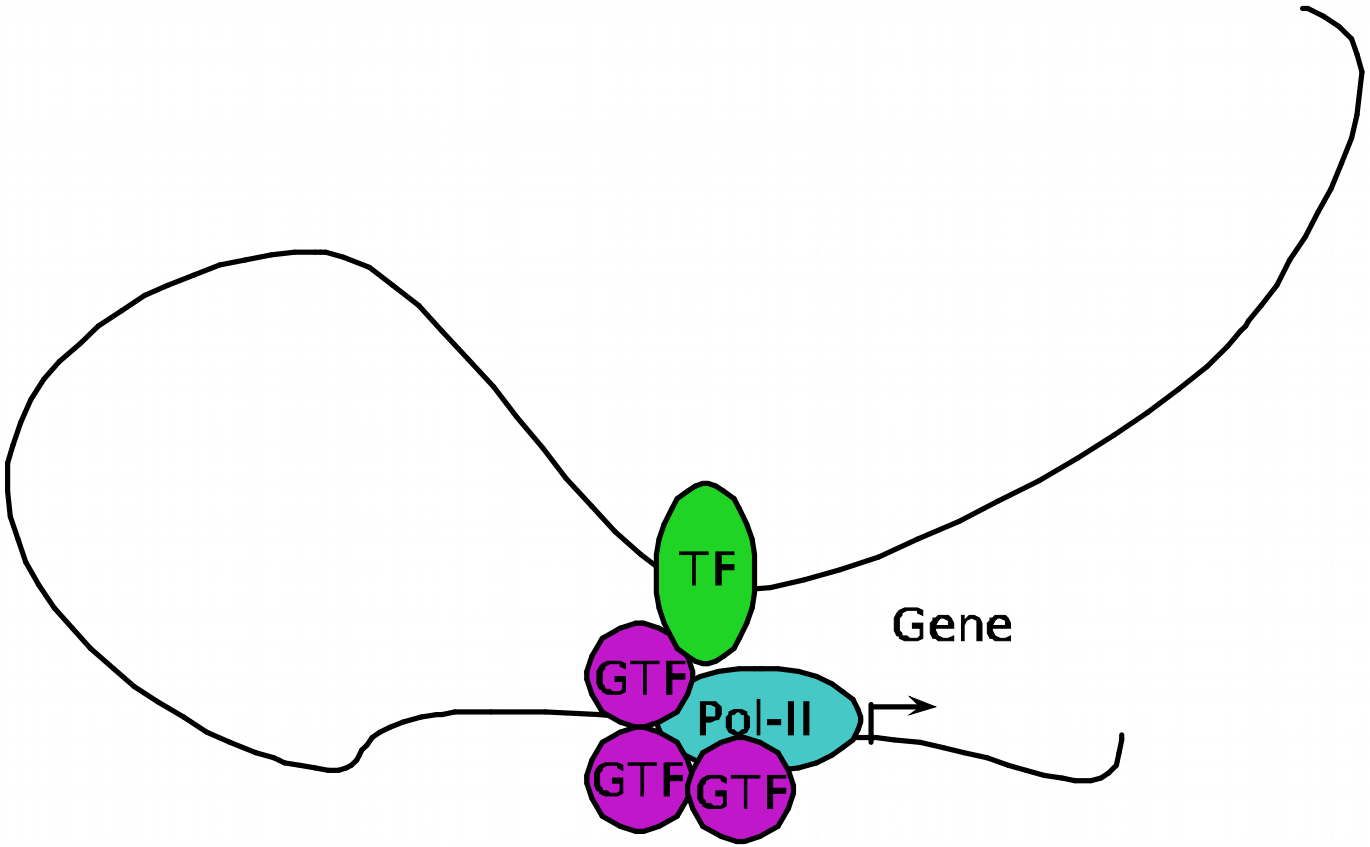

1
At the simplest level, chromatin is a double-stranded helical structure of DNA.
DNA double helix
2 nm
2
DNA is complexed with histones to form nucleosomes.
3
Each nucleosome consists of eight histone proteins around which the DNA wraps 1.65 times.
Nucleosome core of eight histone molecules
4
A chromatosome consists of a nucleosome plus the H1 histone.
H1 histone
11 nm
Chromatosome
5
The nucleosomes fold up to produce a 30-nm fiber...
30 nm
6
... that forms loops averaging 300 nm in length.
300 nm
250-nm-wide fiber
700 nm
7
The 300-nm fibers are compressed and folded to produce a 250-nm-wide fiber.
8
Tight coiling of the 250-nm fiber produces the chromatid of a chromosome.
1400 nm

Gene regulation by TFs

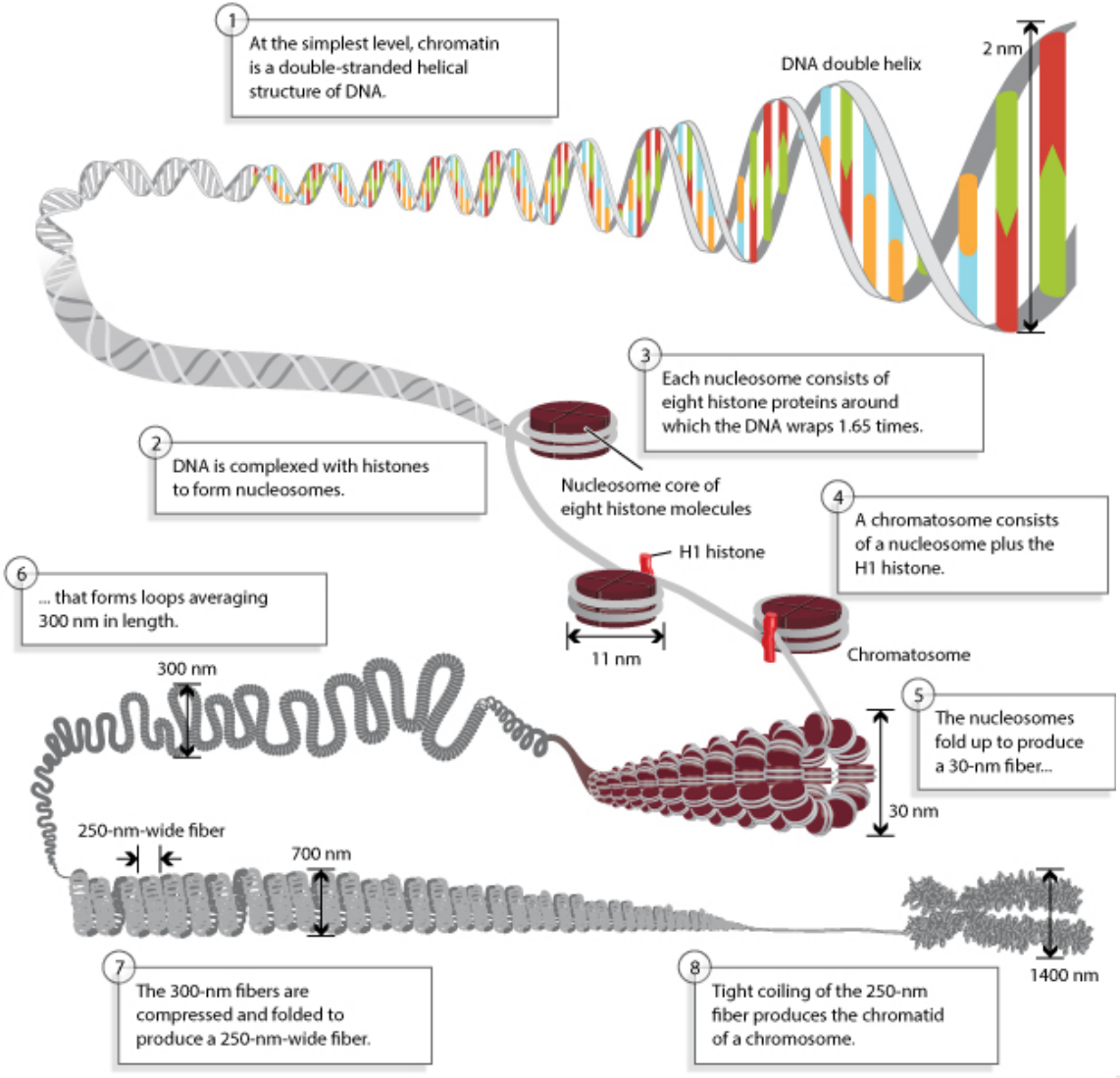
Gene regulation by TFs



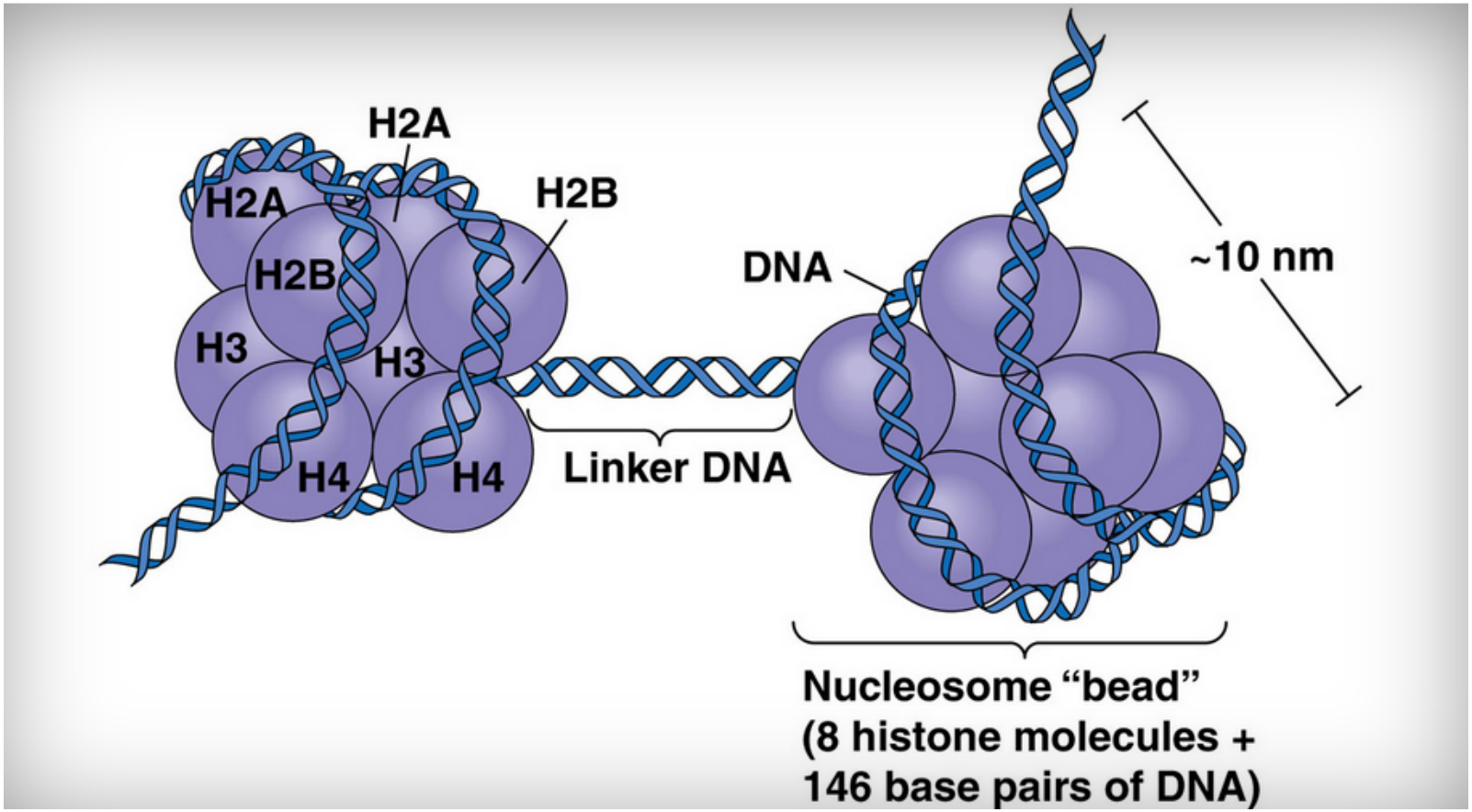
Gene regulation by TFs



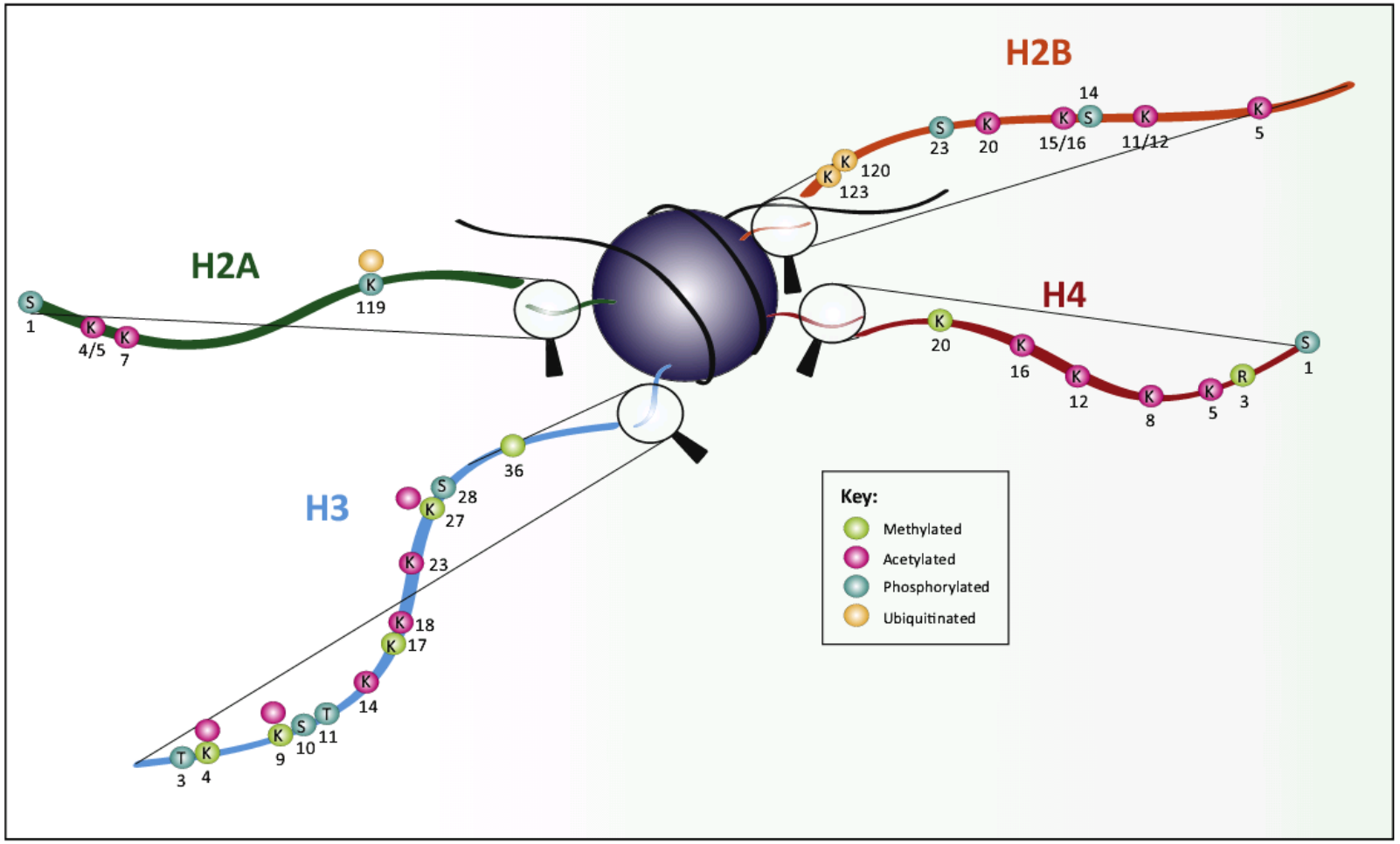
DNA packaging



Nucleosome and histones



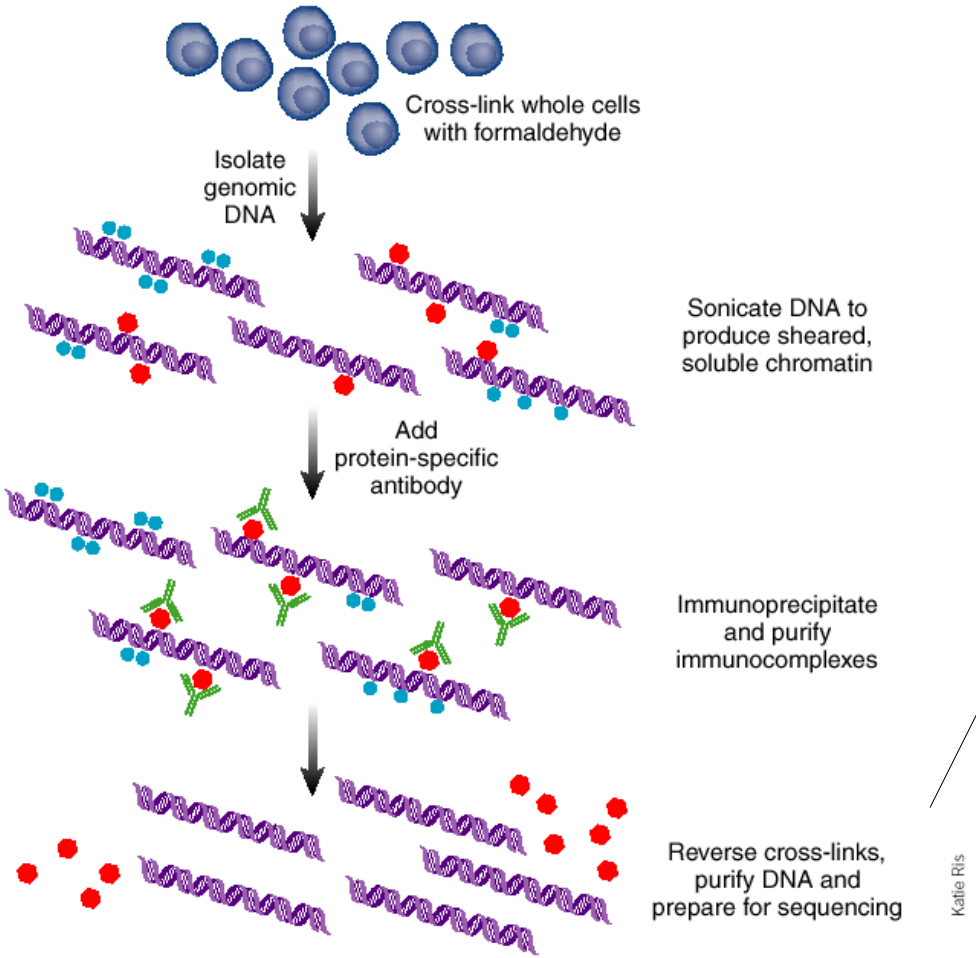
Histone modifications



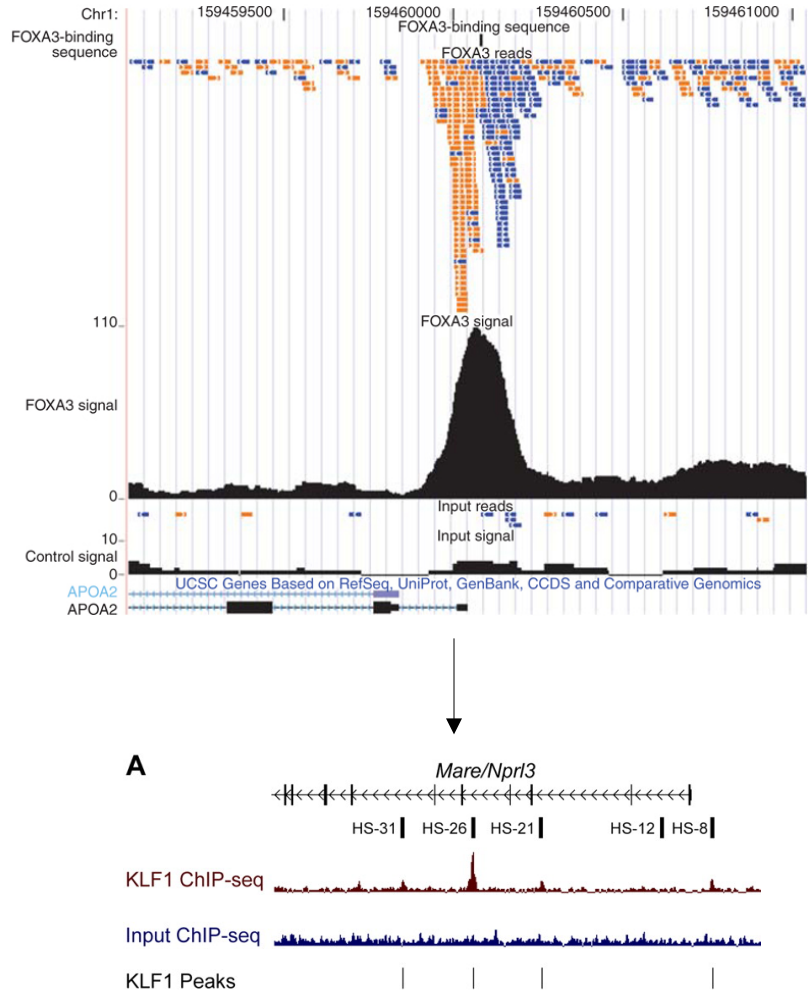
Lawrence et al., Trends in Genetics 2016

Experimental assays

ChIP-seq



Map reads to the genome



ChIP-seq (2)

- Pros:
  - **Direct measure of genome-wide protein-DNA interaction(*)**
- Cons:
  - Don't know whether binding causes changes in gene expression
  - Need an antibody against your protein of interest
  - Expensive

Sequencing data

- **raw data=reads** usually very large file (few GB)
- **format** fastq (ENCODE) or SRA (Sequence Read Archive of NCBI)

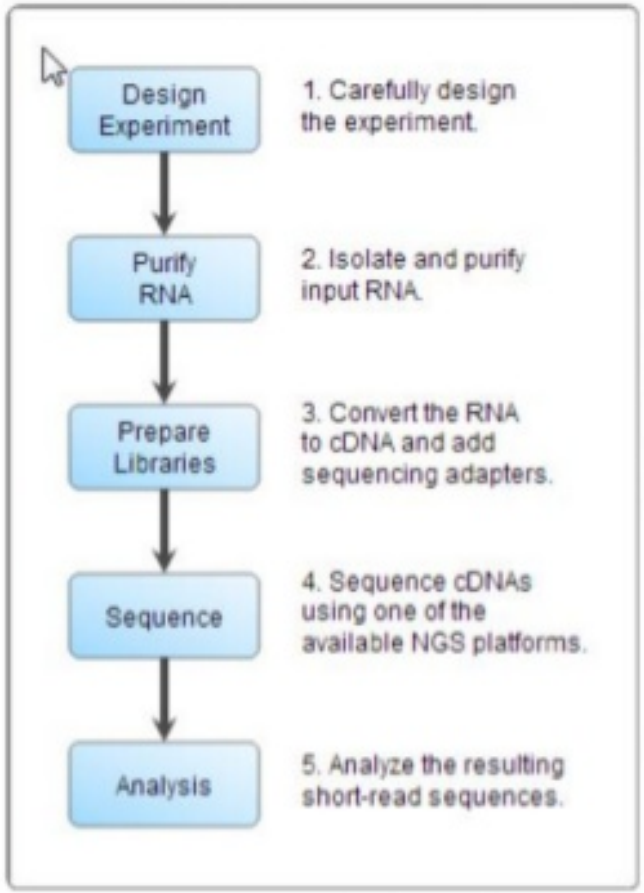
Analysis

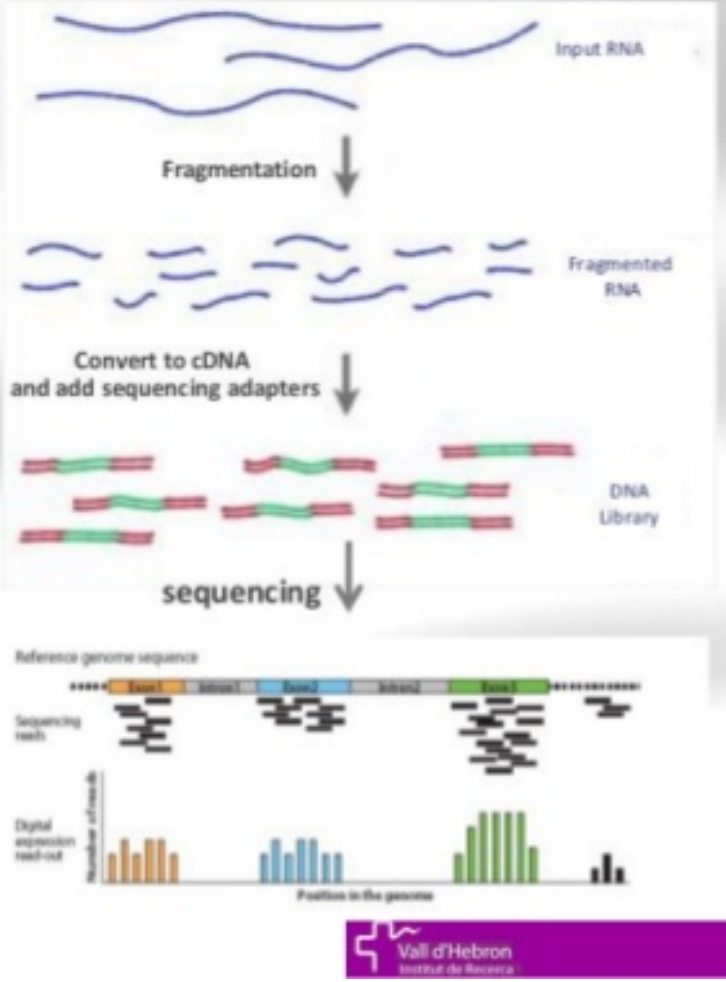
1) **Quality control** with **fastqc, ...**
2) **Mapping** of the reads to the reference genome (**bwa** or **Bowtie**)
3) **Visualizing** the genomic regions (deepTools, IGV)
4) **Peak calling** (**MACS2**)

```
@HS1:161:D0THYACXX:8:1101:1217:2135 1:N:0:CAGATC
CAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGGGGCAACGCCAATC
+
@@@DDDDAFDDFFEFIA)66B=CF;F1=257?A76?##############
@HS1:161:D0THYACXX:8:1101:1483:2089 1:N:0:CAGATC
NTCCATCGACAAGCTGACCAATACTGACCTTAGCTTCGGTCCGTTCAAAG
+
#1:BDA?D>?FHFGCHIGIIIFHGEHHGGCG>?DCEG@F0?FG(8BFDH@
@HS1:161:D0THYACXX:8:1101:1333:2090 1:N:0:CAGATC
NGCATTATTTTCTTTTATTACATTGGTTTATTTGCCATTTTGTTTTAATT
+
#1=DDDBDFHDBDFIIIIGGHIHIIIEGEGIIGGHGGIIIIIIIIIIHGH
@HS1:161:D0THYACXX:8:1101:1361:2104 1:N:0:CAGATC
NCCTGGTATTTTAGCACGGGAAGACCCTGATCGTGGTACATCGCCAGCAC
+
#1:1AD=::CFF?F:EE<EFFE?BEFG)?D<F108@8BDCA?DDIF>FEI
@HS1:161:D0THYACXX:8:1101:1446:2129 1:N:0:CAGATC
TAATGGGATAGGTCACGTTGGTGTAGATGGGCGCATCGTAACCGTGCATC
+
CCCFFFFFHHHHHJJJJJJJJHIJIJIIJJJJJJJJJJIGIJIJHIJJJJ
@HS1:161:D0THYACXX:8:1101:1383:2186 1:N:0:CAGATC
CGGATCCATGTCTGACCTTGTCCTTCTGTTCTTGAGAATTGAGAGCATCT
+
@@@BBDDDFB:AA<4EGEH<C3FCF?C<CH>CH9:E9?CHECDBHBDGHF
@HS1:161:D0THYACXX:8:1101:1568:2088 1:N:0:CAGATC
NTACTGACAACCTGAAAGCAATTGACGCCTGTATTACTCGTCTGCGCCTG
+
#11=AB;ADFBFFBAFHB?FG@H@A?@?G@C::?D?CF@FBDB??@?F:8
@HS1:161:D0THYACXX:8:1101:1604:2110 1:N:0:CAGATC
NATGATATGCAATCAACTTCTTATTTATACCTAATAACGAACTGGGATCA
+
#1BDDFFFHHHHHJJJJJJIJJIJJJGJIJJJIIIJJIJIJJJJJJJJII
@HS1:161:D0THYACXX:8:1101:1565:2143 1:N:0:CAGATC
GACTCACATTACCTTAGGAGACCTTGATTTCAGCAACAAATCATGTACCA
+
@@@DDFDDFHDHHGIGFHGFGEHGIHEFHIHGHGGIGIJIEFHEIIIIIJ
@HS1:161:D0THYACXX:8:1101:1691:2168 1:N:0:CAGATC
GACCATCAGTGTTCCGTTATGGTGGTTGGTTCCCCCAGCCCTGTGTTGGG
+
CCCFFFFFHHHFFGIIIIGIII?EHFHGGDFEFHGIGHGGH@D?@BGACG
```

Example of fastq data file

RNA-seq analysis workflow





RNA-seq data

- **raw data=reads** usually very large file (few GB)
- **format** fastq (ENCODE) or SRA (Sequence Read Archive of NCBI)

Analysis

1) **Quality control** with **fastqc**
2) **Mapping** of the reads to the reference genome (**tophat2**)
3) **Visualizing** the genomic regions (IGV)
4) **Gene expression levels** (in FPKM using Cufflinks)

```
@HS1:161:D0THYACXX:8:1101:1217:2135 1:N:0:CAGATC
CAATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGGGGCAACGCCAATC
+
@@@DDDDAFDDFFEFIA)66B=CF;F1=257?A76?##############
@HS1:161:D0THYACXX:8:1101:1483:2089 1:N:0:CAGATC
NTCCATCGACAAGCTGACCAATACTGACCTTAGCTTCGGTCCGTTCAAAG
+
#1:BDA?D>?FHFGCHIGIIIFHGEHHGGCG>?DCEG@F0?FG(8BFDH@
@HS1:161:D0THYACXX:8:1101:1333:2090 1:N:0:CAGATC
NGCATTATTTTCTTTTATTACATTGGTTTATTTGCCATTTTGTTTTAATT
+
#1=DDDBDFHDBDFIIIIGGHIHIIIEGEGIIGGHGGIIIIIIIIIIHGH
@HS1:161:D0THYACXX:8:1101:1361:2104 1:N:0:CAGATC
NCCTGGTATTTTAGCACGGGAAGACCCTGATCGTGGTACATCGCCAGCAC
+
#1:1AD=::CFF?F:EE<EFFE?BEFG)?D<F108@8BDCA?DDIF>FEI
@HS1:161:D0THYACXX:8:1101:1446:2129 1:N:0:CAGATC
TAATGGGATAGGTCACGTTGGTGTAGATGGGCGCATCGTAACCGTGCATC
+
CCCFFFFFHHHHHJJJJJJJJHIJIJIIJJJJJJJJJJIGIJIJHIJJJJ
@HS1:161:D0THYACXX:8:1101:1383:2186 1:N:0:CAGATC
CGGATCCATGTCTGACCTTGTCCTTCTGTTCTTGAGAATTGAGAGCATCT
+
@@@BBDDDFB:AA<4EGEH<C3FCF?C<CH>CH9:E9?CHECDBHBDGHF
@HS1:161:D0THYACXX:8:1101:1568:2088 1:N:0:CAGATC
NTACTGACAACCTGAAAGCAATTGACGCCTGTATTACTCGTCTGCGCCTG
+
#11=AB;ADFBFFBAFHB?FG@H@A?@?G@C::?D?CF@FBDB??@?F:8
@HS1:161:D0THYACXX:8:1101:1604:2110 1:N:0:CAGATC
NATGATATGCAATCAACTTCTTATTTATACCTAATAACGAACTGGGATCA
+
#1BDDFFFHHHHHJJJJJJIJJIJJJGJIJJJIIIJJIJIJJJJJJJJII
@HS1:161:D0THYACXX:8:1101:1565:2143 1:N:0:CAGATC
GACTCACATTACCTTAGGAGACCTTGATTTCAGCAACAAATCATGTACCA
+
@@@DDFDDFHDHHGIGFHGFGEHGIHEFHIHGHGGIGIJIEFHEIIIIIJ
@HS1:161:D0THYACXX:8:1101:1691:2168 1:N:0:CAGATC
GACCATCAGTGTTCCGTTATGGTGGTTGGTTCCCCCAGCCCTGTGTTGGG
+
CCCFFFFFHHHFFGIIIIGIII?EHFHGGDFEFHGIGHGGH@D?@BGACG
```

Example of fastq data file

Tasks

Tasks

- Analysis of TF binding across the genome (TAF1, JUND)
- Analysis of histone modifications across the genome (H3K4me3, H3K4me1, H3K27ac)
- Cell-types: K562, GM12878 and H1-hESC (one per group)

- From the ENCODE project (see papers)
- genome.ucsc.edu/ENCODE or
- https://www.encodeproject.org/

Group

- Each group should work in a different cell-type
- Group 1: K562
- Group 2: GM12878
- Group 3: H1-hESC

Literature survey

What is TAF1, H3K4me3, H3K4me1, H3K27ac and JUND?

- Where does one find those marks or proteins in the genome?
- Do they bind to promoters and/or enhancers?
- What are their roles in gene regulation?
- Are there known motifs associated with the TFs (e.g. Jaspar)?
- What is the role of high and low CpG promoters?
- Where can you find the dataset? Specify the exact source and name of the file/experiment (including RNA-seq for your cell line).

- Find publications that address those points
- Use Google and/or scholar.google.com
- **Until next Monday**

Preliminary analysis steps (ChIP-seq)

- Download ChIP-seq raw reads (fastq/fq) for TAF1, JUND, H3K4me1, H3K4me3 and H3K27ac
- Also, download corresponding Input (control) experiments
- Align the ChIP-seq reads to hg19 with bowtie2
- Check the ChIP-seq quality
  - Using fastqc and phantompeakqualtools (only for ChIP-seq. Hint: Is NSC and RSC acceptable?)
  - Is the quality sufficient? Why or why not?
- Call peaks for all experiments with macs2

Preliminary analysis steps (RNA-seq)

- Download RNA-seq reads (fastq)
- Align the RNA-seq reads to hg19 with tophat2
- If paired-end, there must be two fastq files
- Check the RNA-seq quality
  - Using fastqc
  - Is the quality sufficient? Why or why not?
- Compute FPKM expression values with cufflinks

Genomic features and overlap analysis

Do the peaks overlap (for different marks and proteins)?

Bedtools or R/Bioconductor: Genomic Ranges

Draw a Venn-diagram

Share the peak regions with the other groups

What is the overlap with the other groups?

- Which genomic features do they overlap with?

Intergenic, gene body, promoters, exons, introns, etc.

Generate a heatmap centered at the peak summit (with deepTools)

Generate a profile aligned at the TSS (with deepTools)

Interpret the results

Sequence analysis

- Extract the sequences from the peak regions
  Using R/Bioconductor or bedtools
- Analyse motifs in the sequences
  Using MEME-ChIP
  Which motifs do you find? Interpret the results
- Do the TAF1 peaks overlap with promoters? Are these high or low CpG promoters? (Hint: analyse dinucleotide frequency)

Gene expression analysis

- How do the peaks explain gene expression levels?
  Correlation or linear regression
  How well does the H3K4me3 level at a promoter explain gene expression?
  How well does TAF1 level at promoters predict gene expression?
  How well does JUND predict gene expression
  How well does H3K27ac and H3K4me

Schedule

- 13.03. Introduction lecture
- 20.03. Presentation of the detailed plan of each group (Literature survey, data file information, schedule) 10:15am, 11:00am, 11:45am
- every Monday 10:15am, 11:00am, 11:45am progress meetings
- 27.04. Final report deadline
- 03.05. Discussion of final reports
- 08.05. Final presentations

Bioinformatics resources

READ THE MANUALS!

- Bowtie2 and bwa (to align ChIP-seq reads)
- Tophat2 (to align RNA-seq reads)
- Samtools (to convert SAM files to BAM files)
- Cufflinks (to determine gene expression levels)
- Bedtools (to analyse genomic regions – e.g. overlap, distance, extracting DNA sequences for some regions, find closest gene, ...)
- Fastqc (to analyse the ChIP-seq/RNA-seq quality)
- Phantompeakqualtools (to analyse ChIP-seq quality – Cross-correlation plot, etc.)
- DeepTools (to plot average profiles and heatmaps)
- MEME-ChIP (to discovery motifs)
- Bioconductor www.bioconductor.org/

Useful resources

- JASPAR
- IGV
- Genome.ucsc.edu/ENCODE and www.encodeproject.org
- Google and scholar.google.com
- http://hgdownload.cse.ucsc.edu/downloads.html
- https://www.gencodegenes.org/ (Gene annotations, Hint: hg19 corresponds to GRCh37)

Useful resources

ENCODE papers (An intergated encyclopedia of DNA elements in the human genome, etc.)

Bailey et al Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. PLoS Comput Biol (2013). (This explains some quality aspects of ChIP-seq data)

Saxonov et al A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters (2006).

Any papers that explain TAF1, JUND, H3K4me4, H3K4me1, K3K27ac

Any papers that explain the methods

Office hours

Alena: Monday and Tuesday at 1:30 pm

Wolfgang: Thursday and Friday at 9:30 am