Studying the transcriptome using RNA-seq

Cecilia Coimbra Klein
Computational Biology of RNA Processing, CRG
Departament de Genètica, IBUB, UB

Master in Omics Data Analysis
Jan. 2019

IBUB
Institut de Biomedicina
de la Universitat de Barcelona

UNIVERSITAT DE BARCELONA

CRG Centre for Genomic Regulation

UVIC UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL DE CATALUNYA

Master in Omics Data Analysis

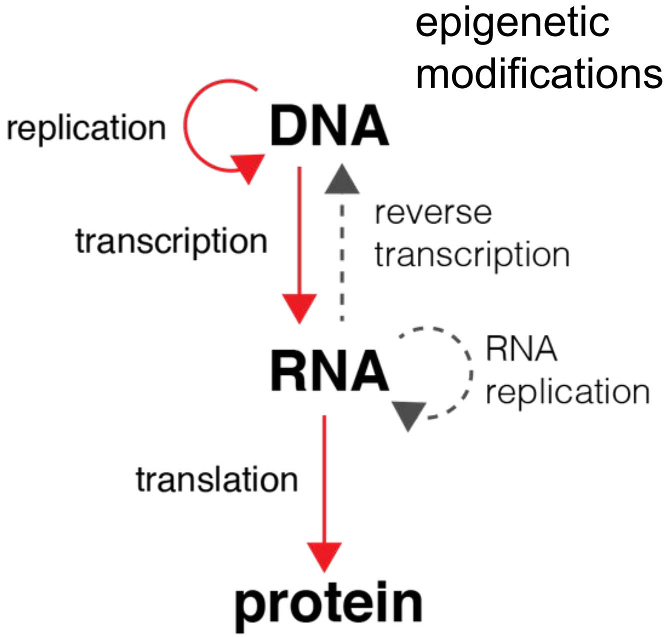
Outline

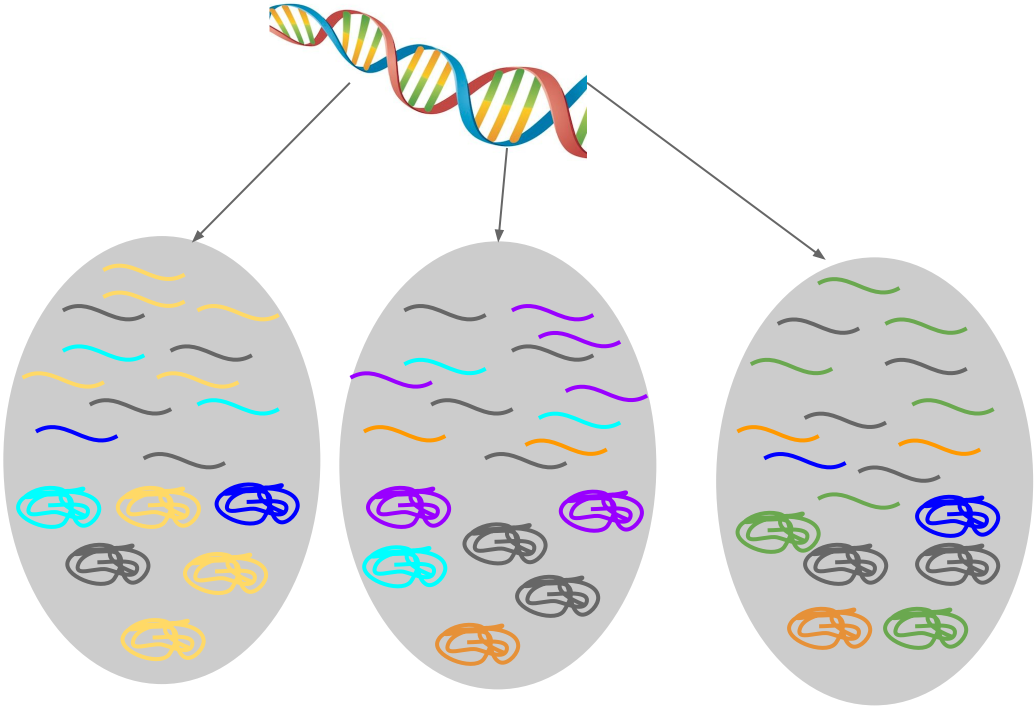
Outline

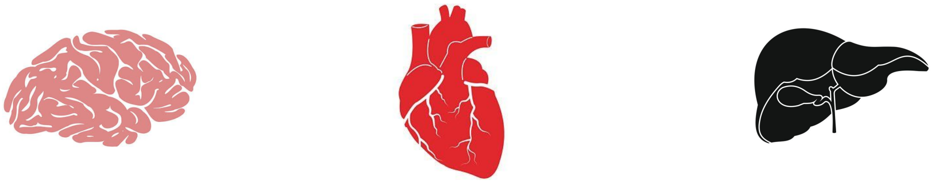
- Summary of the course
  - Day 1: RNA-seq introduction and processing
  - Day 2: RNA-seq analysis (clustering, differential gene expression, GO enrichment)
  - Day 3: RNA-seq analysis (splicing)
- MultiOmics
  - ChIP-seq (processing and data analysis)
  - ATAC-seq (visualization)
- Hands-on MultiOmics
  - ChIP-seq and ATAC-seq signal in the UCSC genome browser
  - promoter regions of differentially expressed genes
  - promoter regions of differentially spliced genes
  - omics portals
- Multiple-choice exercise

Day 1: RNA-seq introduction and processing

Molecular biology dogma



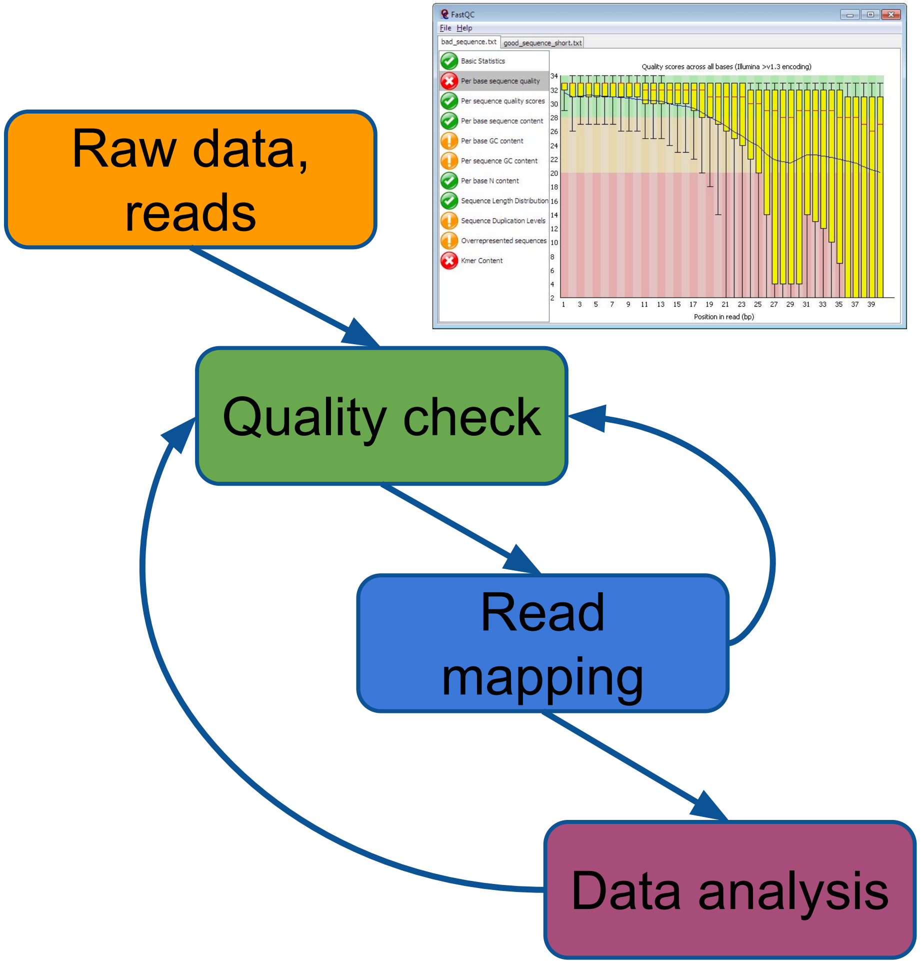




- The genome is identical in all cell types, however not all cell types have the same function. That’s why the transcriptome (and the epigenome) becomes also relevant.

https://sdtimes.com/23andme/sd-times-blog-using-dna-for-access-control/
https://www.pinterest.com/pin/702139398127827000/
https://www.etsy.com/market/human_heart_svg
https://www.vectorstock.com/royalty-free-vector/human-liver-black-icon-vector-7458173

Typical pipeline



Raw data, reads

Quality check

Read mapping

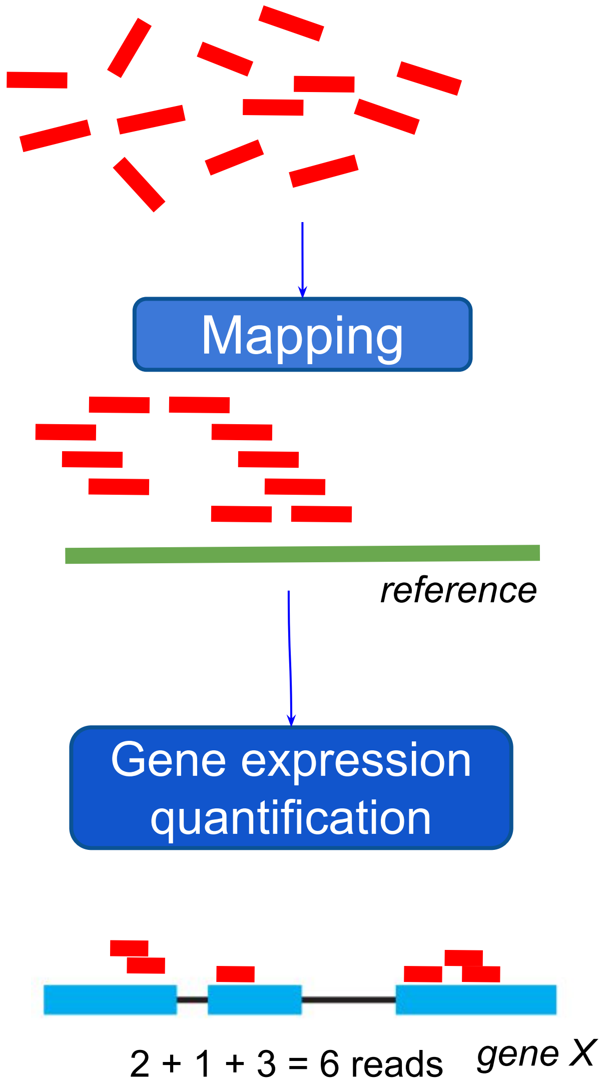
Data analysis

FastQC

File Help

bad_sequence.txt | good_sequence_short.txt

- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✗ Kmer Content



Mapping

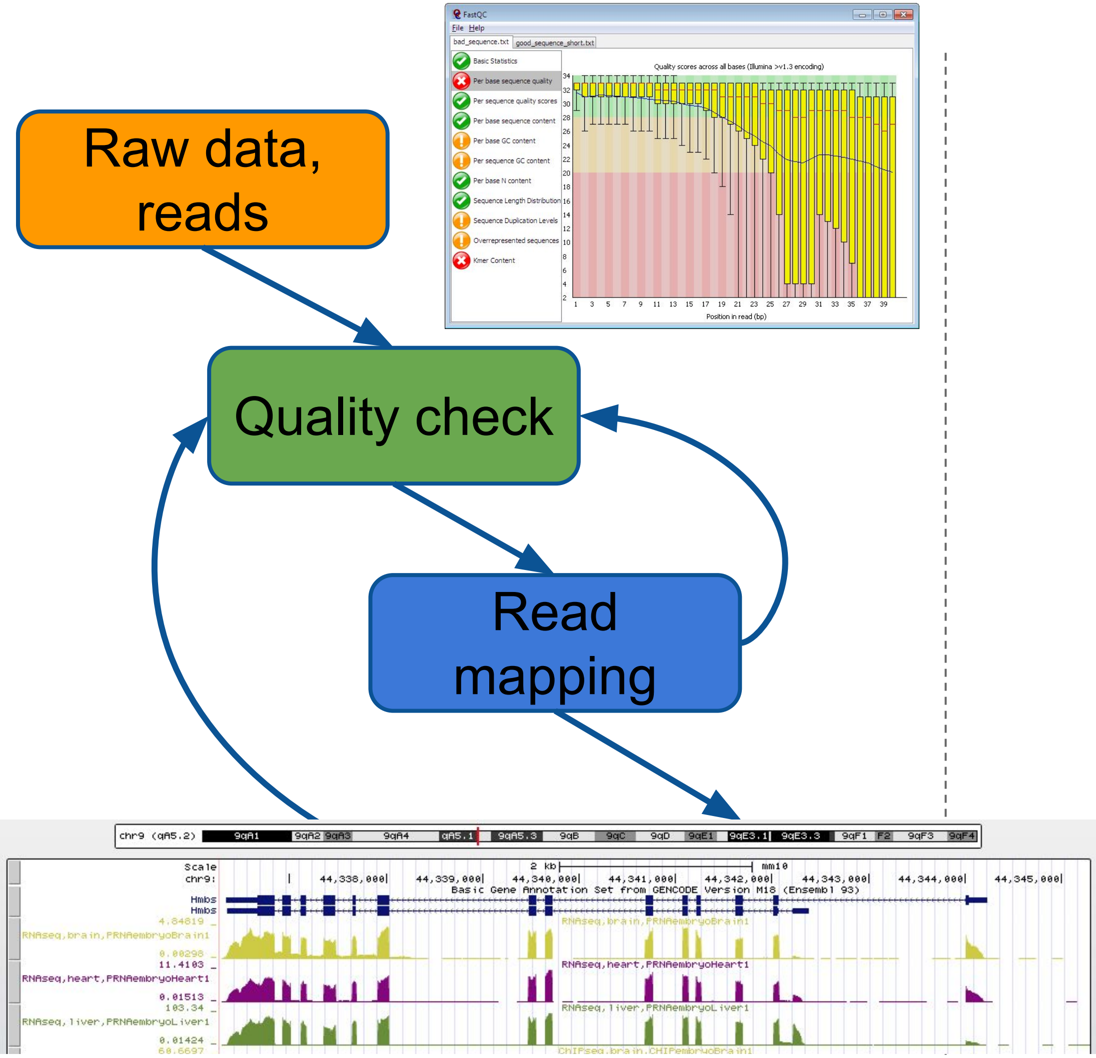
reference

Gene expression quantification

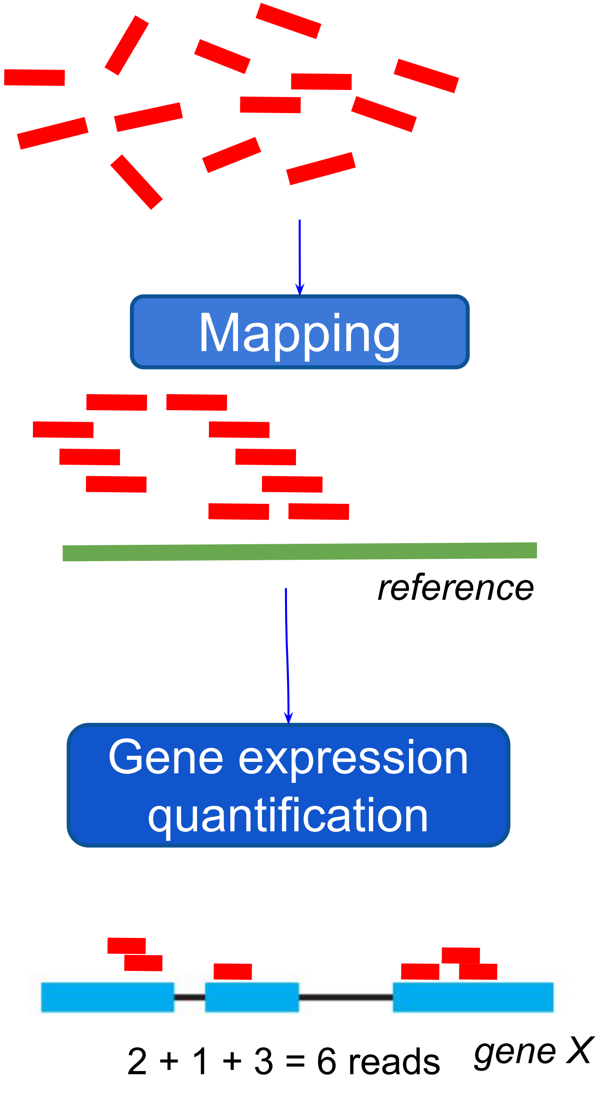
2 + 1 + 3 = 6 reads

gene X

Typical pipeline

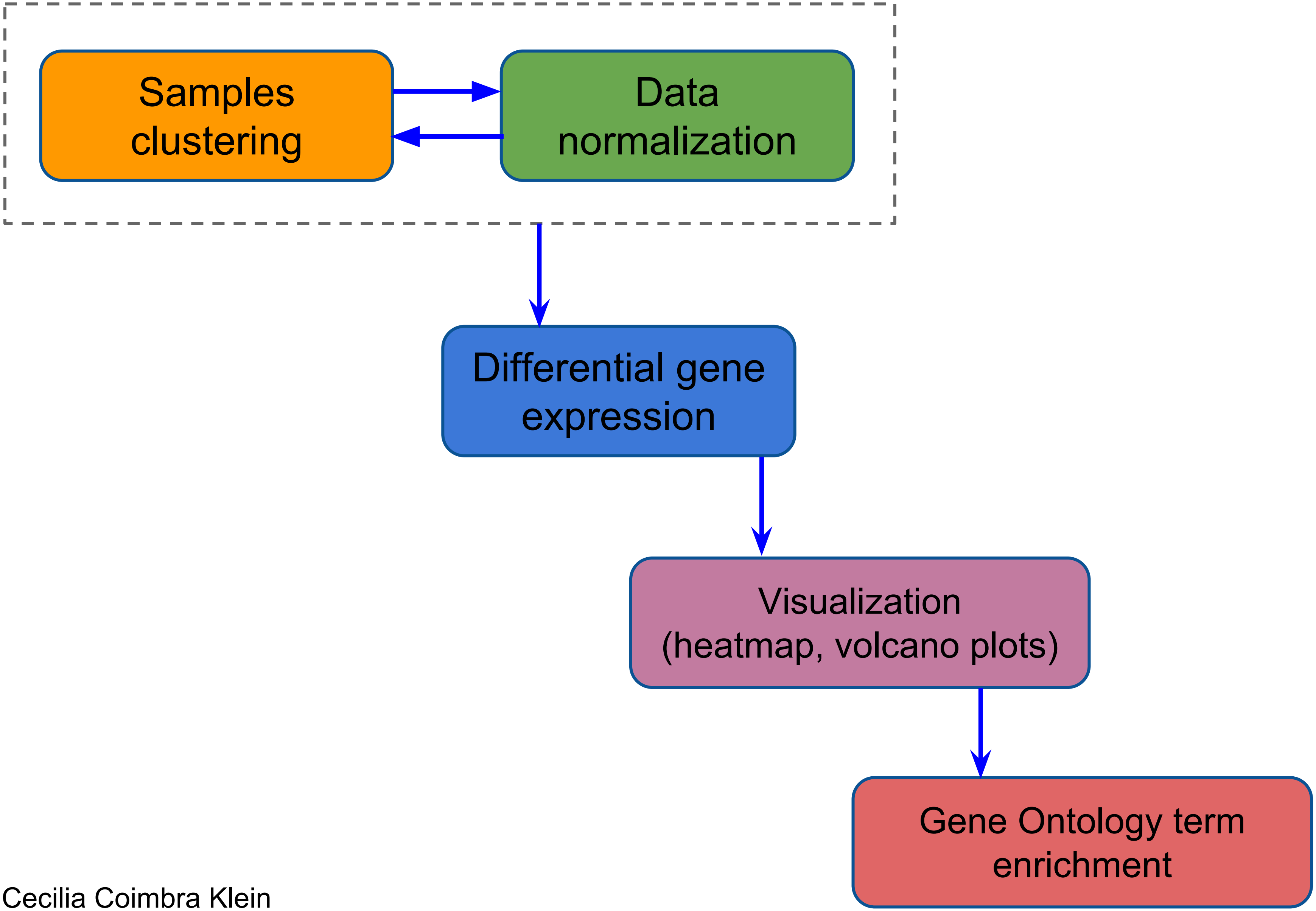


genome-euro.ucsc.edu



Day 2: RNA-seq analysis (clustering, differential gene expression, GO enrichment)

Analysis pipeline



Samples clustering

Data normalization

Differential gene expression

Visualization (heatmap, volcano plots)

Gene Ontology term enrichment

Analysis pipeline

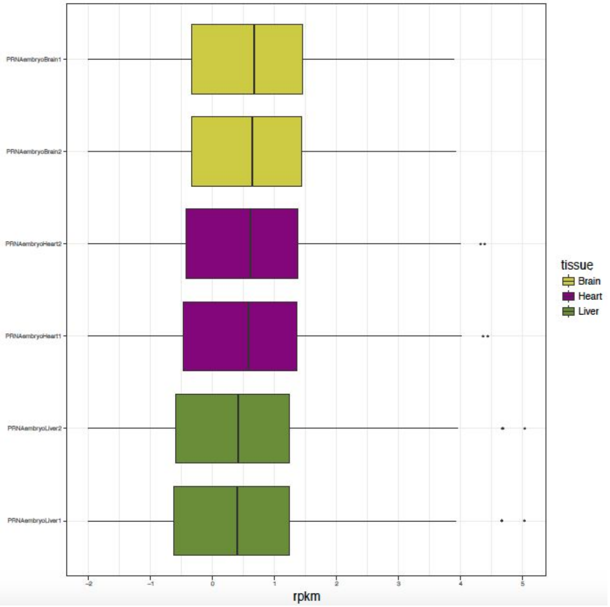
Samples clustering

Data normalization

Differential gene expression

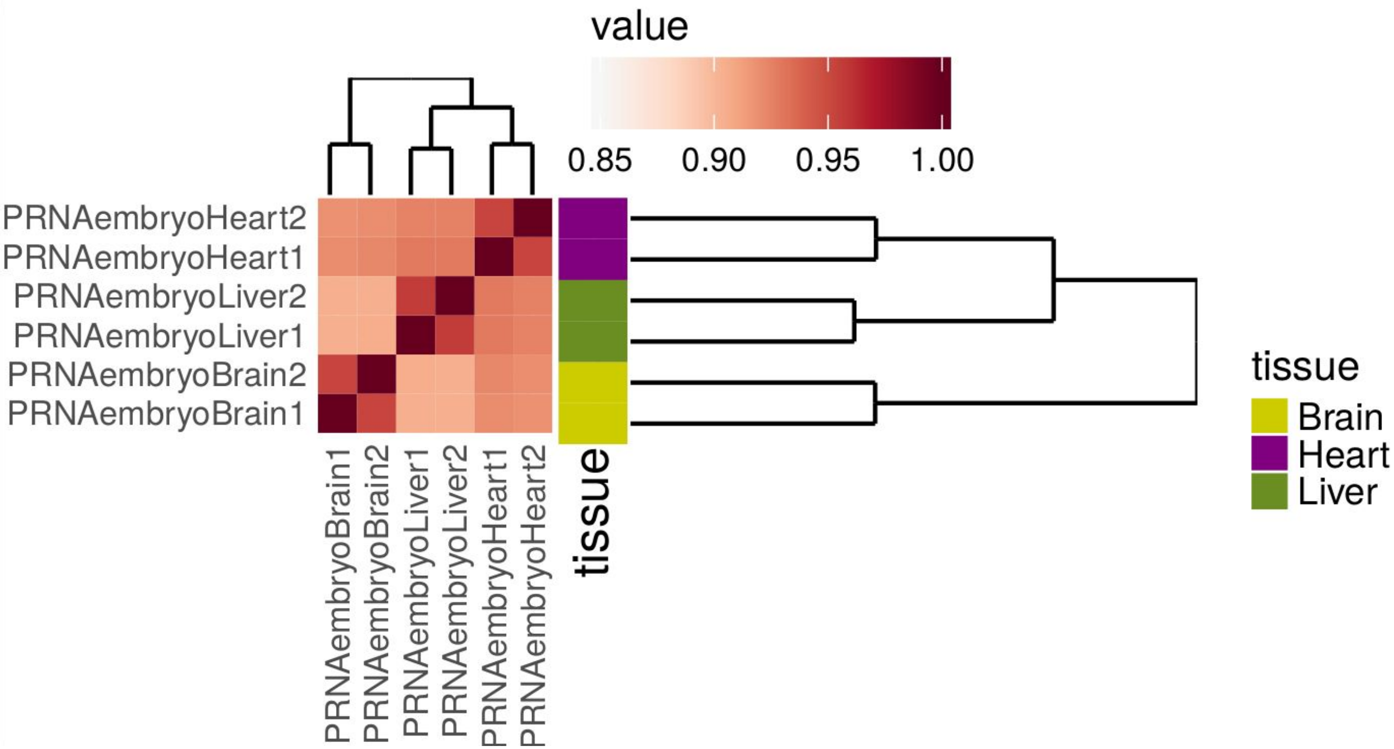
Visualization (heatmap, volcano plots)

Gene Ontology term enrichment



Cecilia Coimbra Klein

Samples clustering



Analysis pipeline

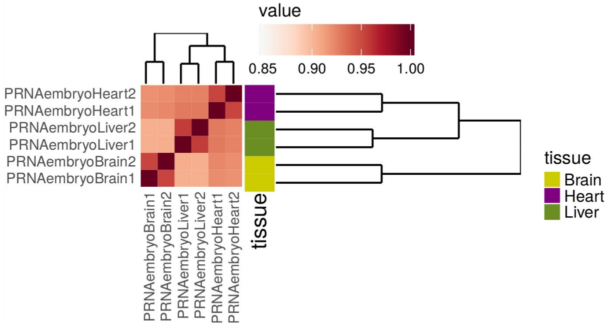
Samples clustering

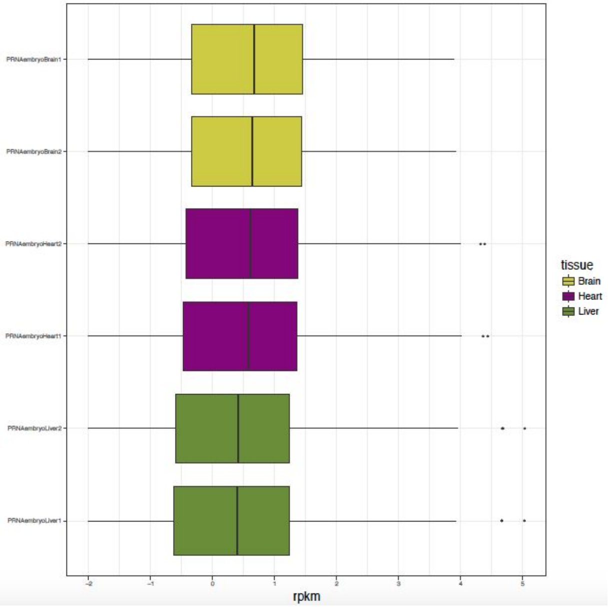
Data normalization

Differential gene expression

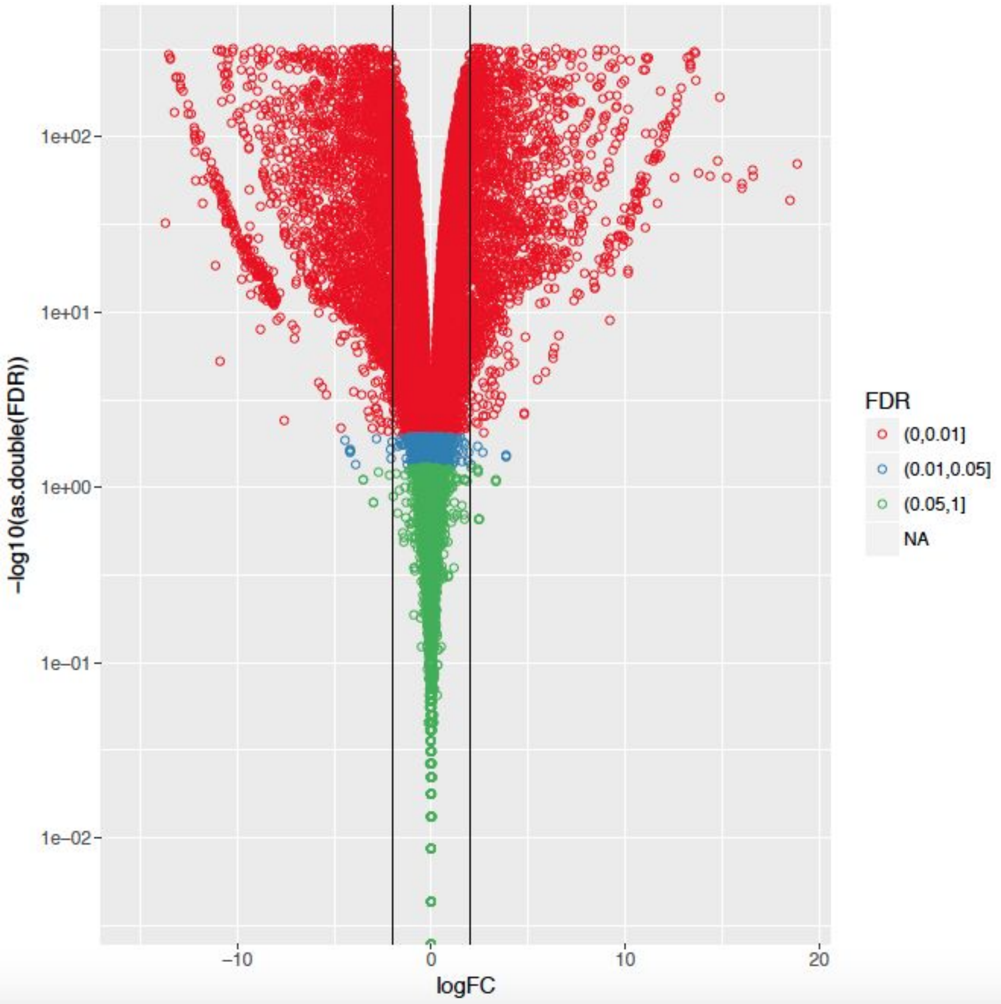
Visualization (heatmap, volcano plots)

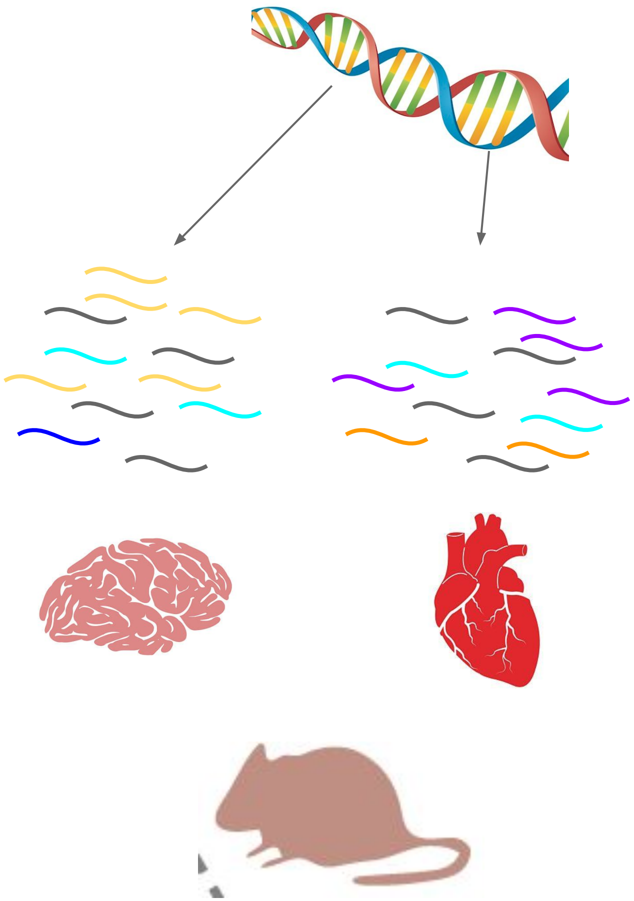
Gene Ontology term enrichment



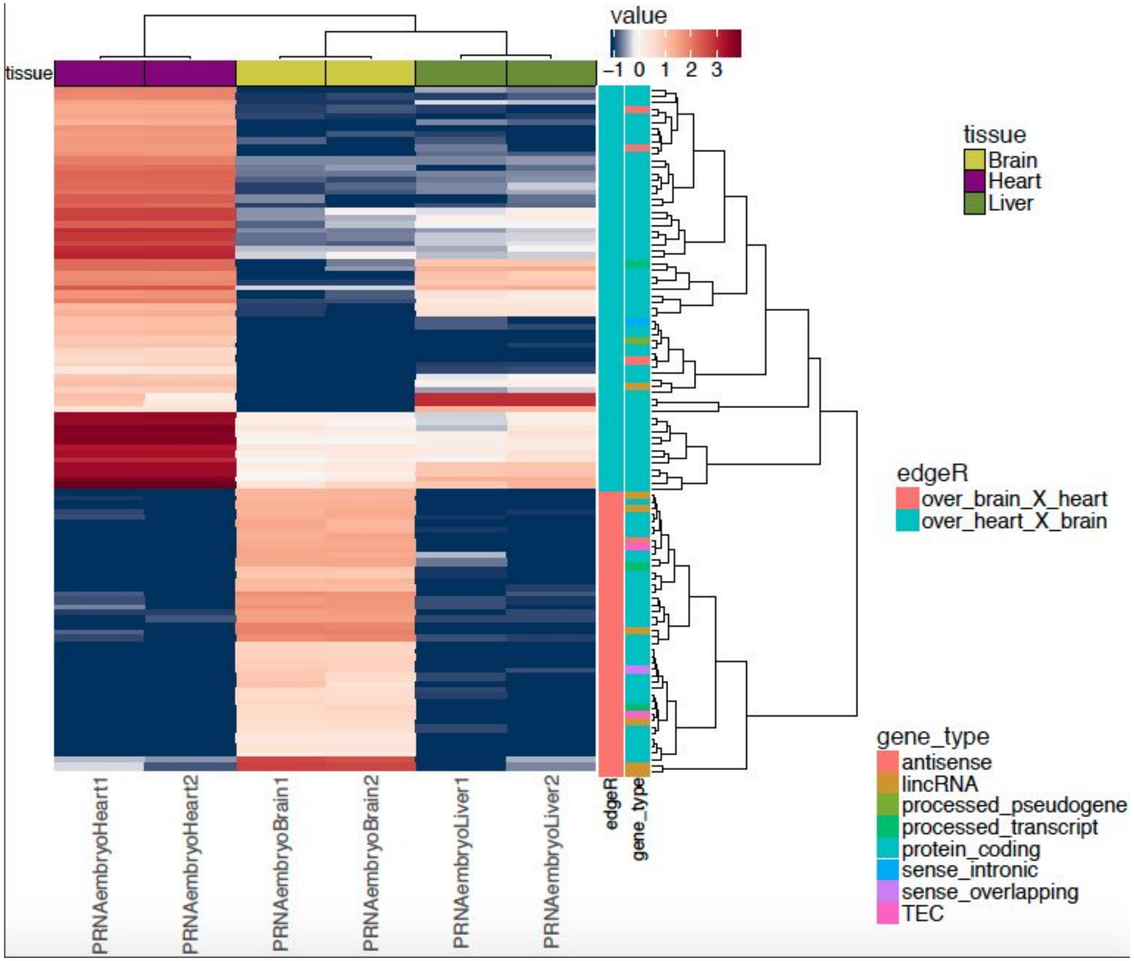


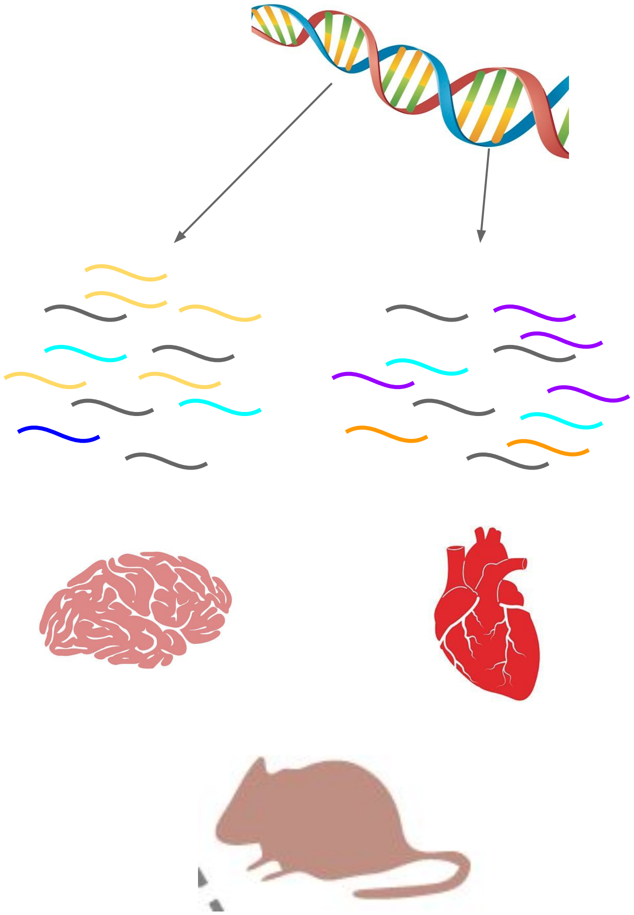
Differential Gene Expression





Differential Gene Expression

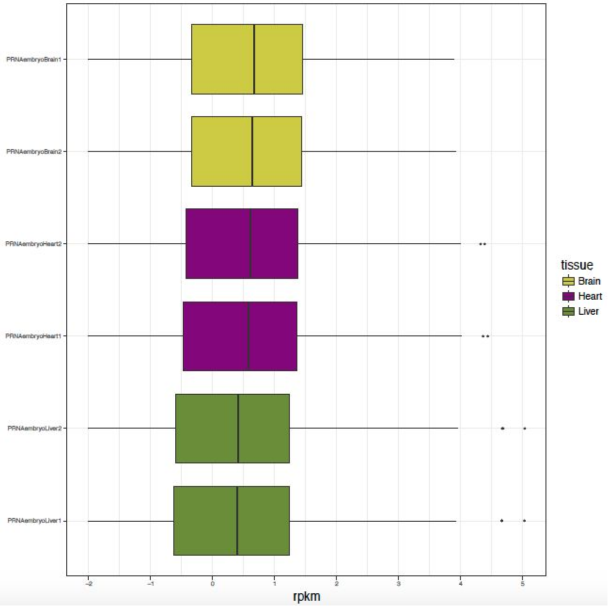


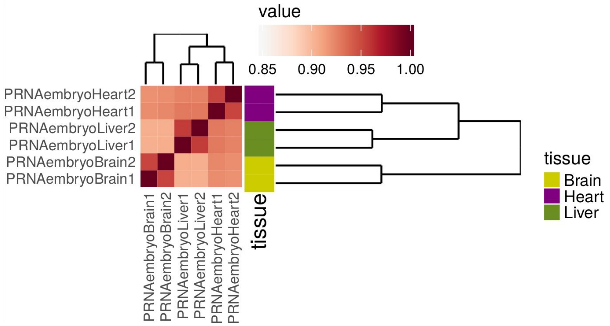


Analysis pipeline

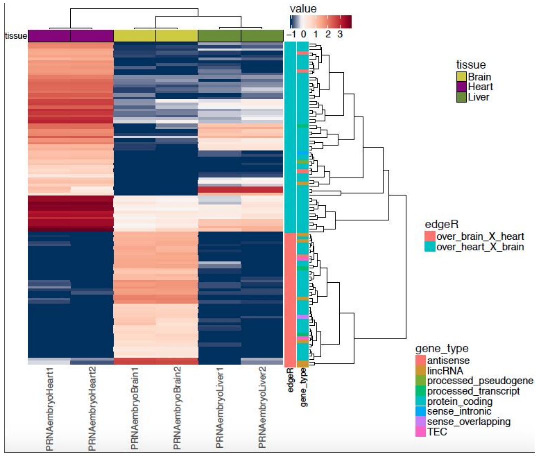
Samples clustering

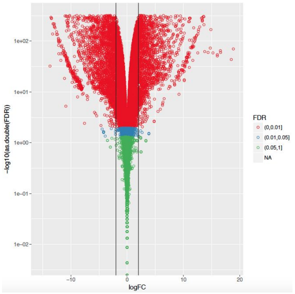
Data normalization





Differential gene expression



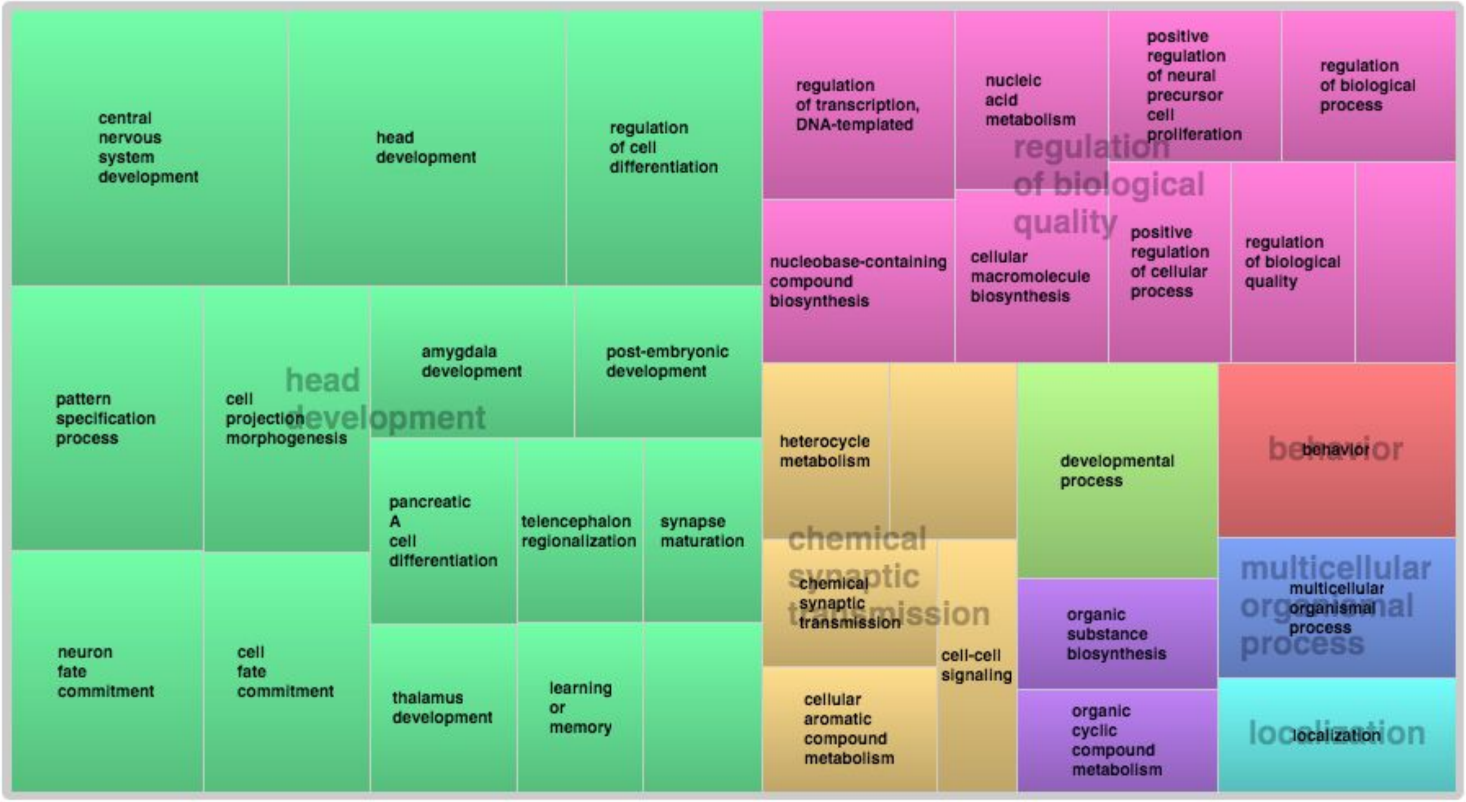


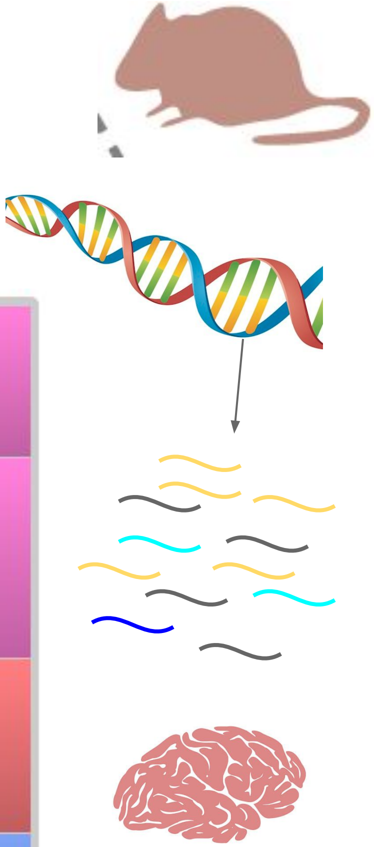
Visualization (heatmap, volcano plots)

Gene Ontology term enrichment

Cecilia Coimbra Klein

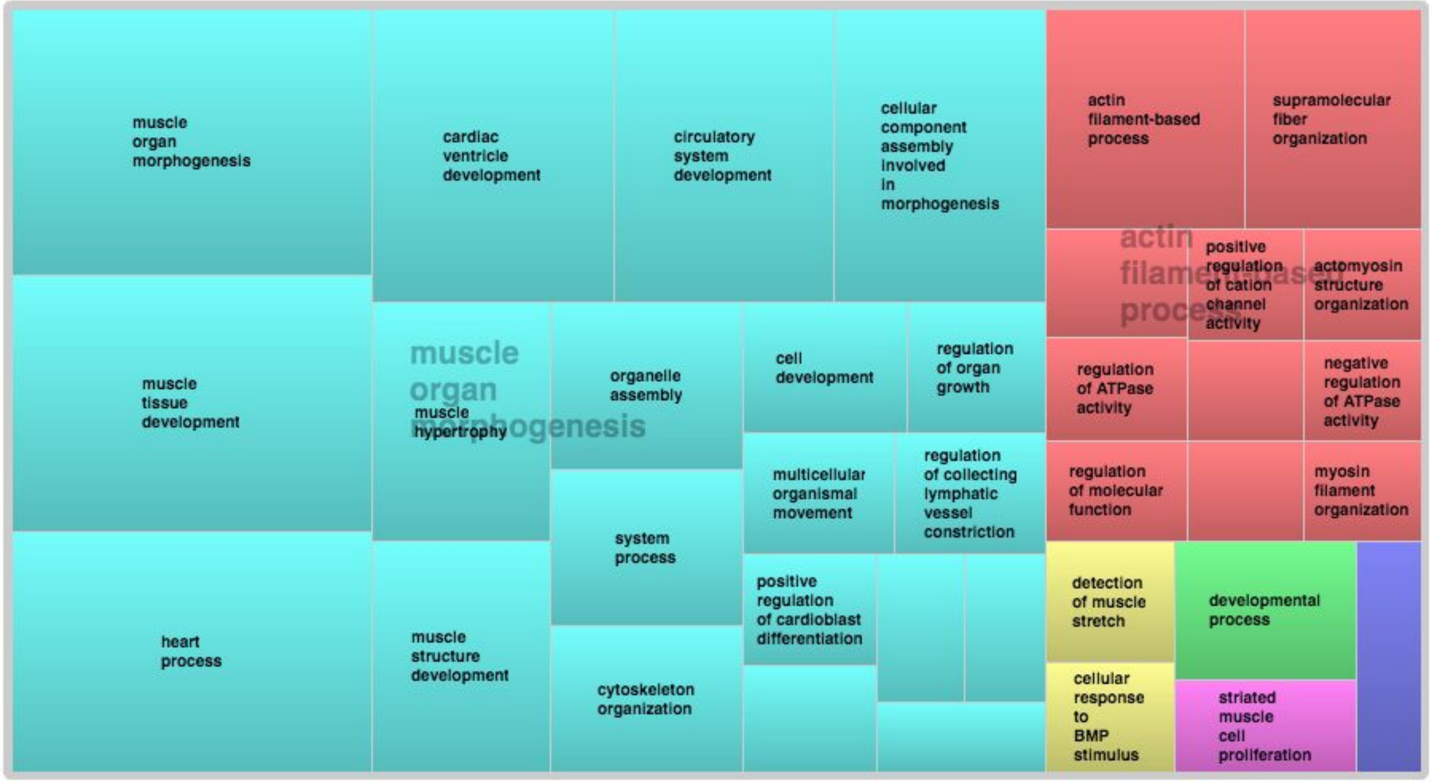
GO term enrichment

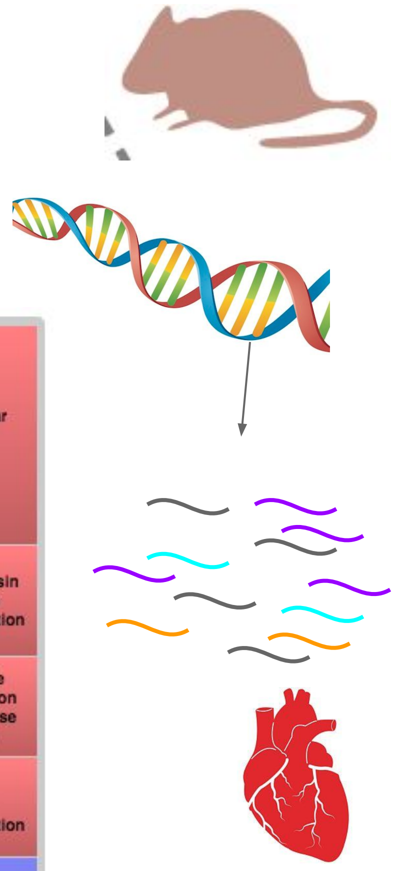




Cecilia Coimbra Klein

GO term enrichment

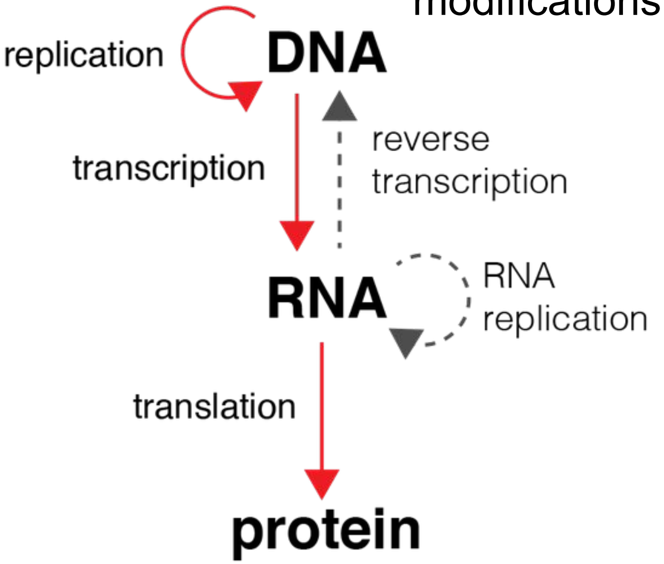


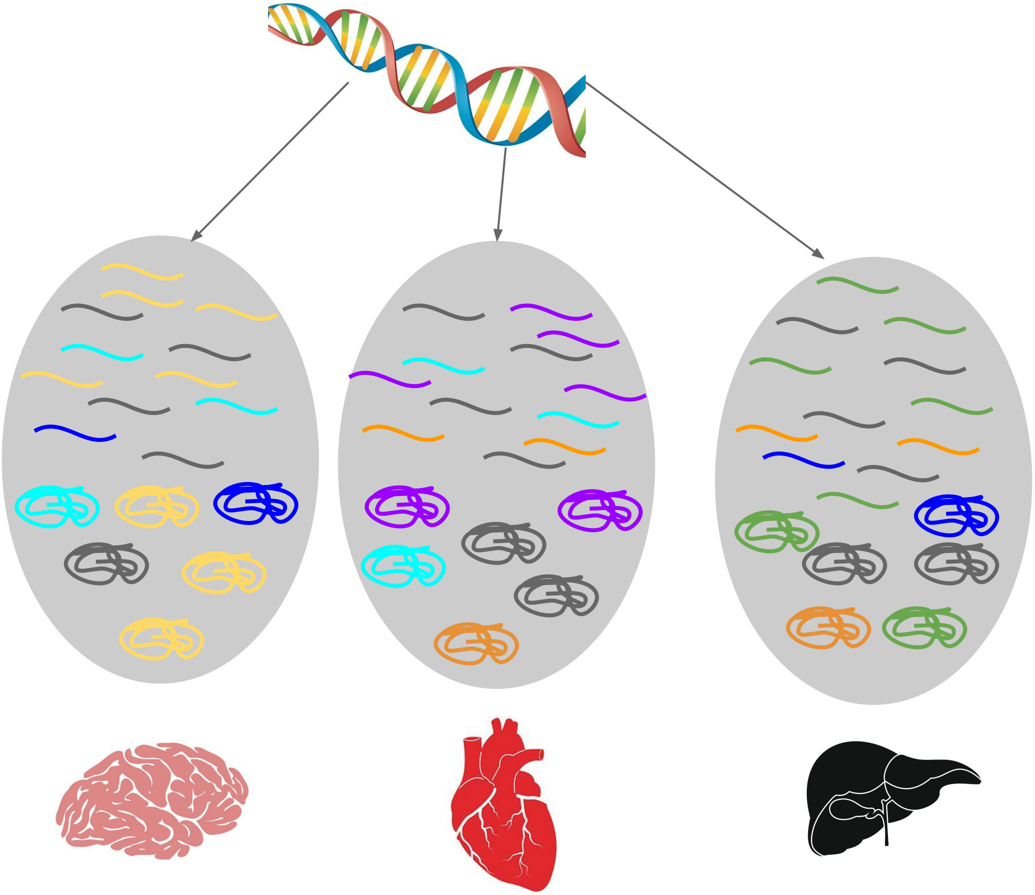


Day 3: RNA-seq analysis (splicing)

Molecular biology dogma

epigenetic
modifications

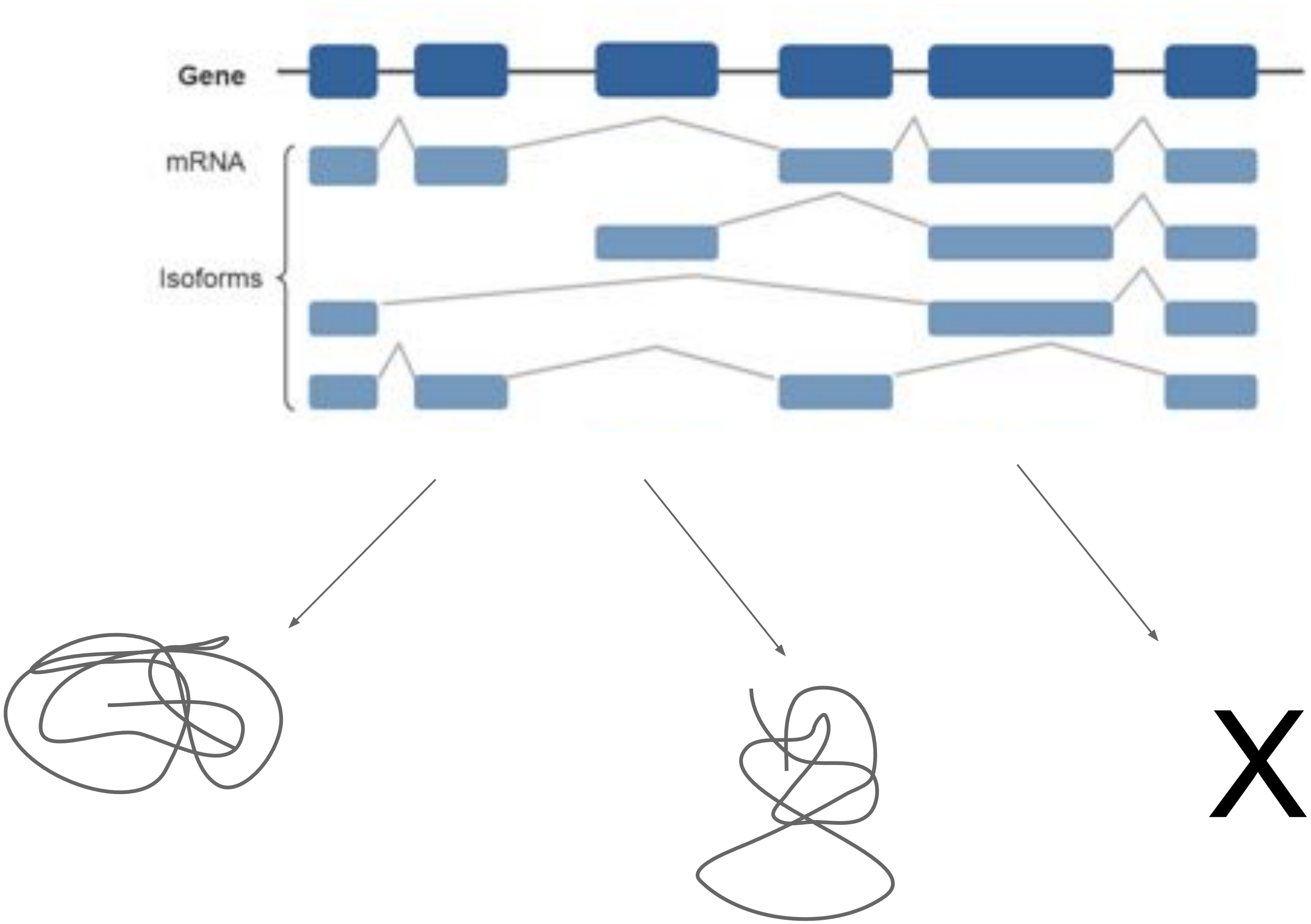




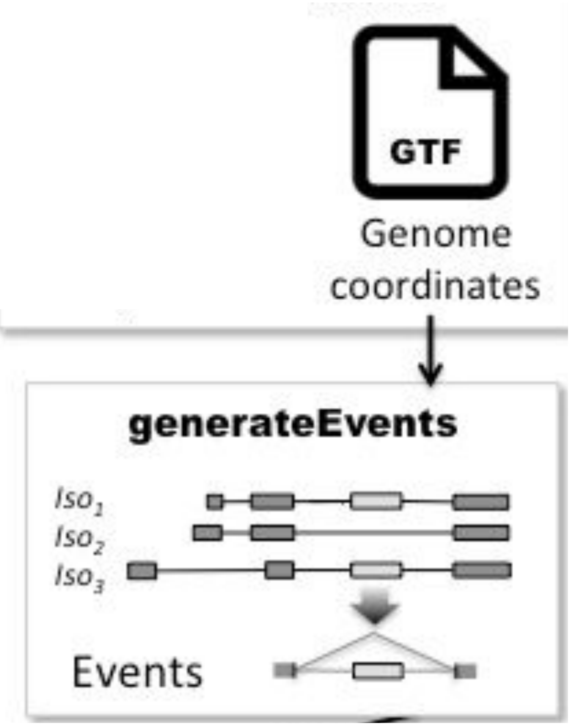
- The genome is identical in all cell types, however not all cell types have the same function. That’s why the transcriptome (and the epigenome) becomes also relevant.

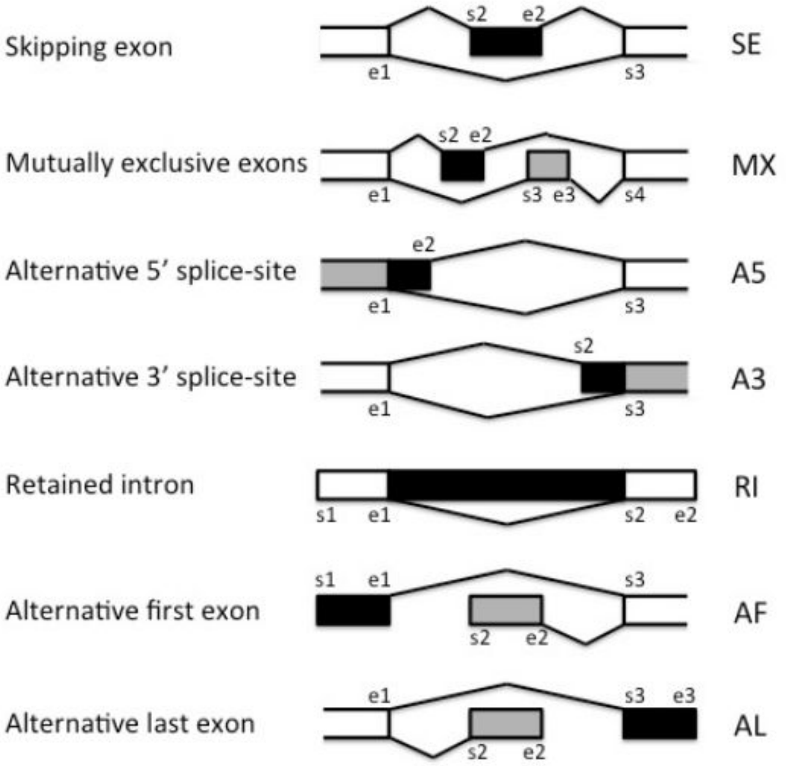
https://sdtimes.com/23andme/sd-times-blog-using-dna-for-access-control/
https://www.pinterest.com/pin/702139398127827000/
https://www.etsy.com/market/human_heart_svg
https://www.vectorstock.com/royalty-free-vector/human-liver-black-icon-vector-7458173

Isoform usage



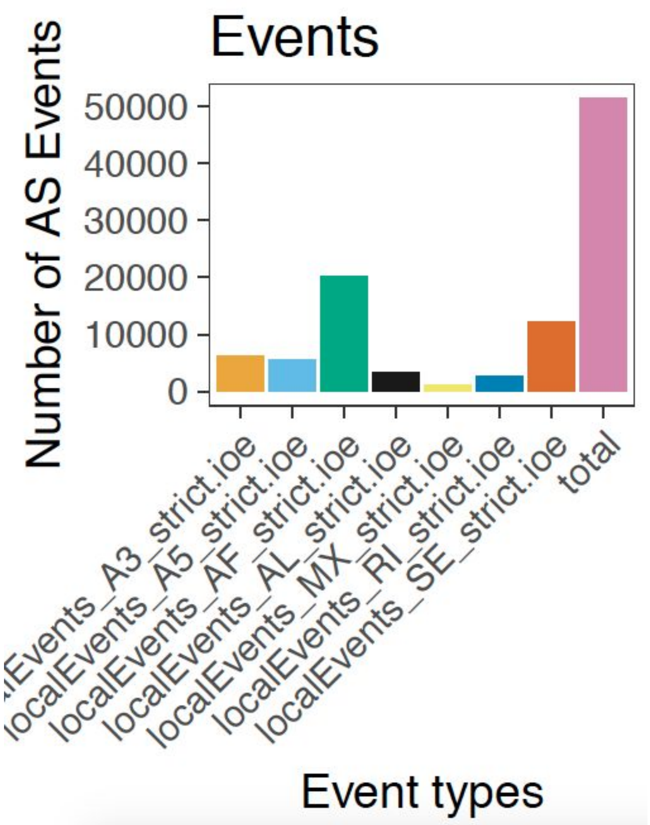
SUPPA: generate events based on gene annotation

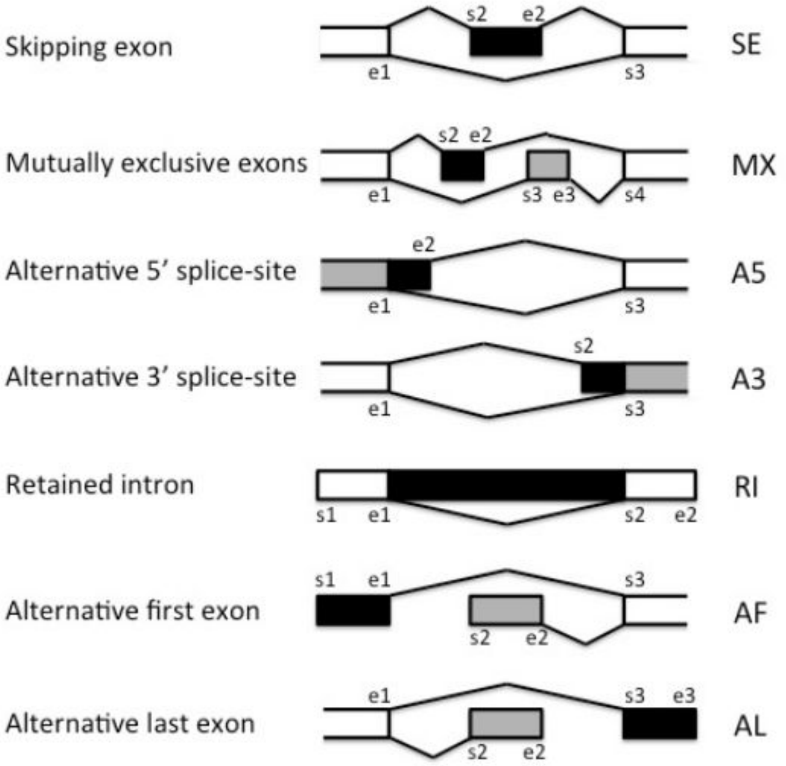




https://bitbucket.org/regulatorygenomicsupf/suppa

SUPPA: generate events based on gene annotation



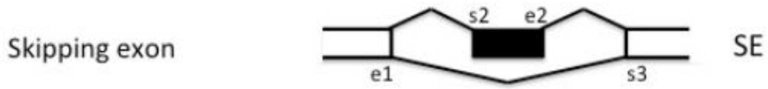


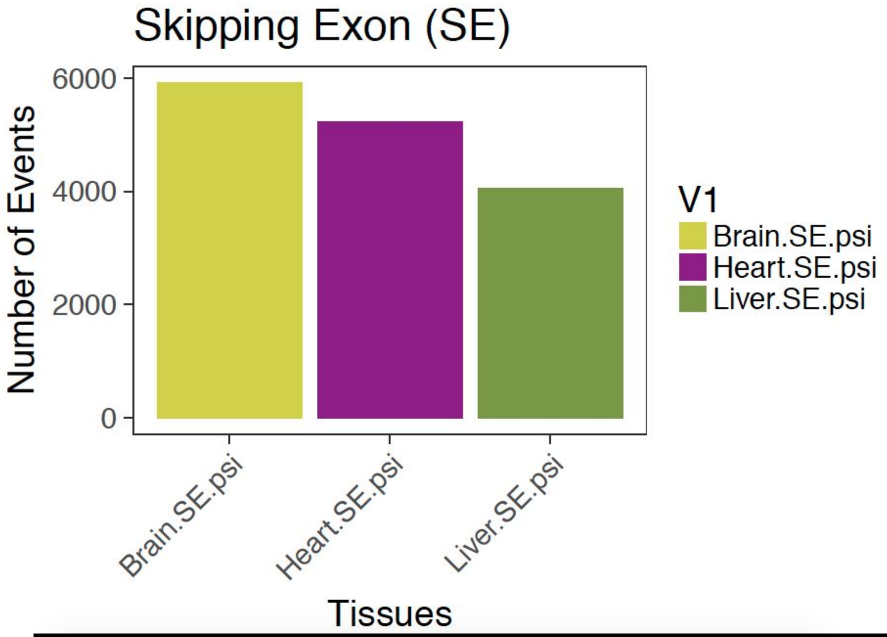
```
# number of alternative events with PSI values
ggbarplot.R -i input.tsv -o number_of_events.pdf --title "Events"
--y_title "Number of AS Events" --x_title "Event types" --palette_fill
/tutorial/palettes/cbbPalette.8.txt  -f 1
```

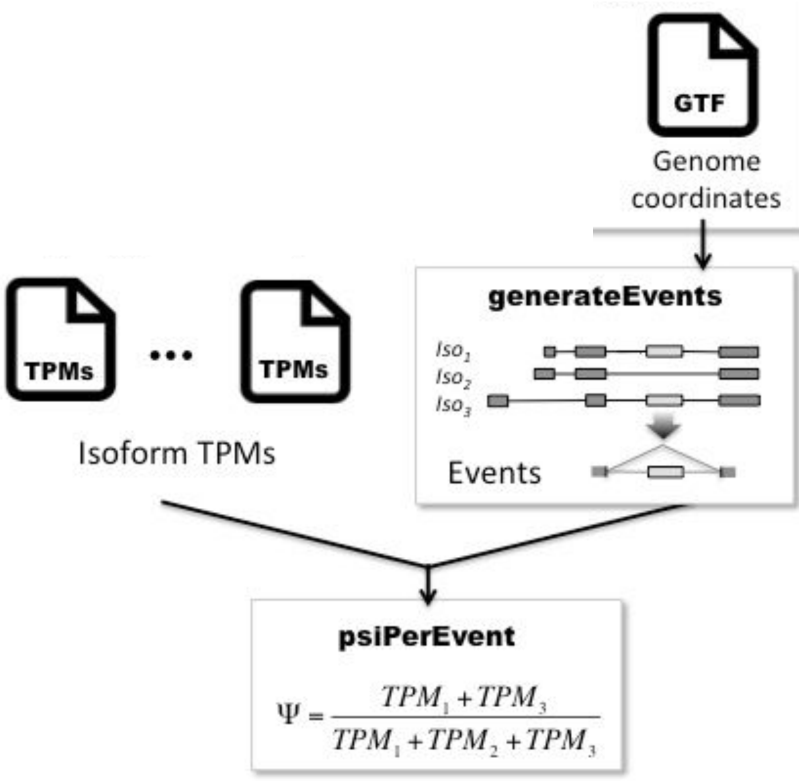
Cecilia Coimbra Klein

https://bitbucket.org/regulatorygenomicsupf/suppa

SUPPA: Quantify event inclusion levels (PSIs)



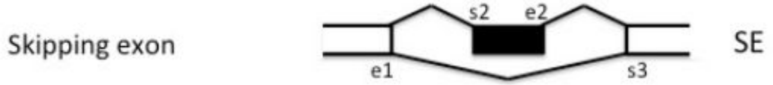


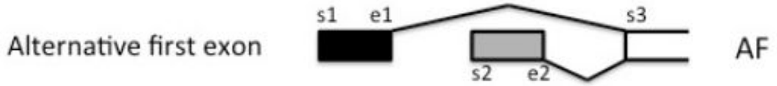


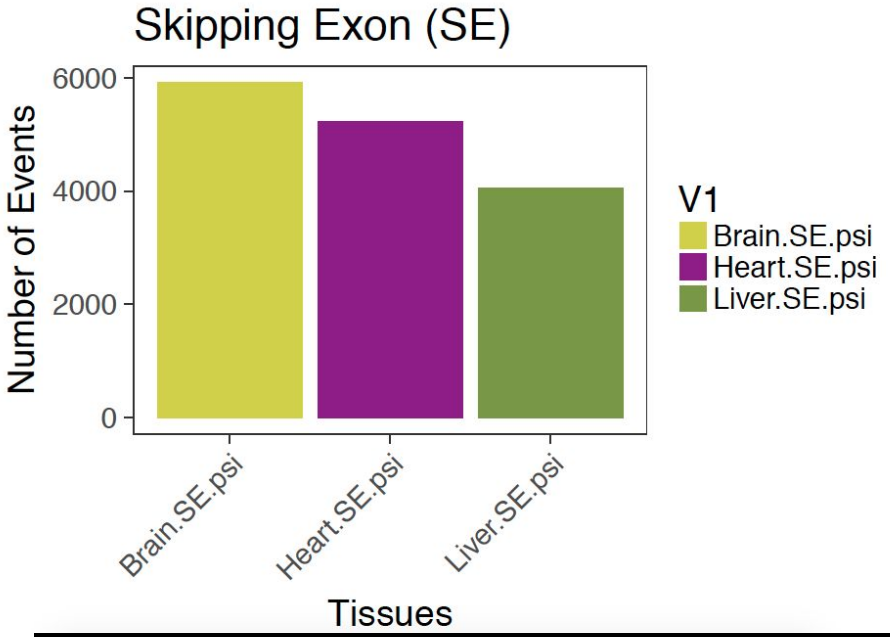
```
# number of alternative single exon skipping
ls *.SE.psi|while read f;do echo -e $f"\t"$(grep -v nan $f |wc -l);done |
ggbarplot.R -i stdin -o number_of_SE.pdf -f 1 --palette_fill
/tutorial/palettes/palTissue.txt  --title "Skipping Exon (SE)" --y_title
"Number of Events" --x_title "Tissues"
```

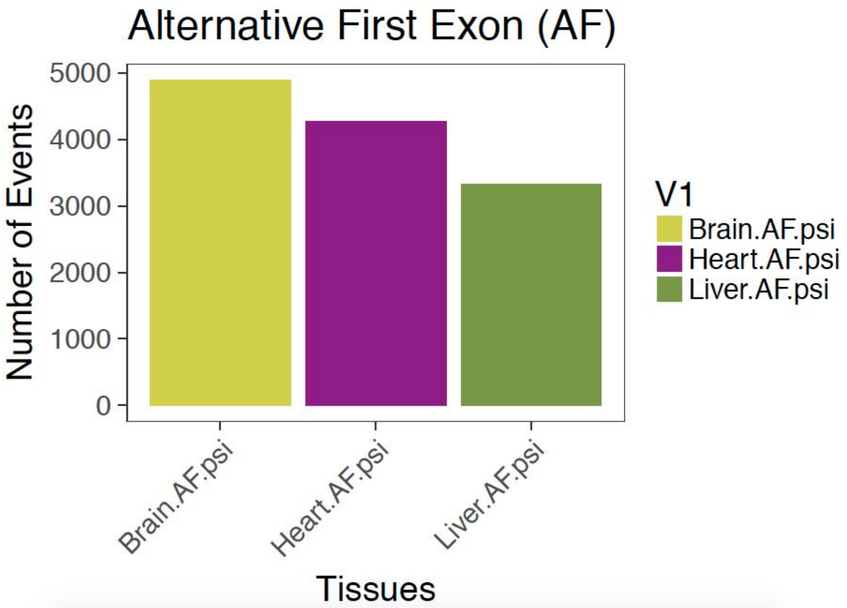
https://bitbucket.org/regulatorygenomicsupf/suppa

SUPPA: Quantify event inclusion levels (PSIs)







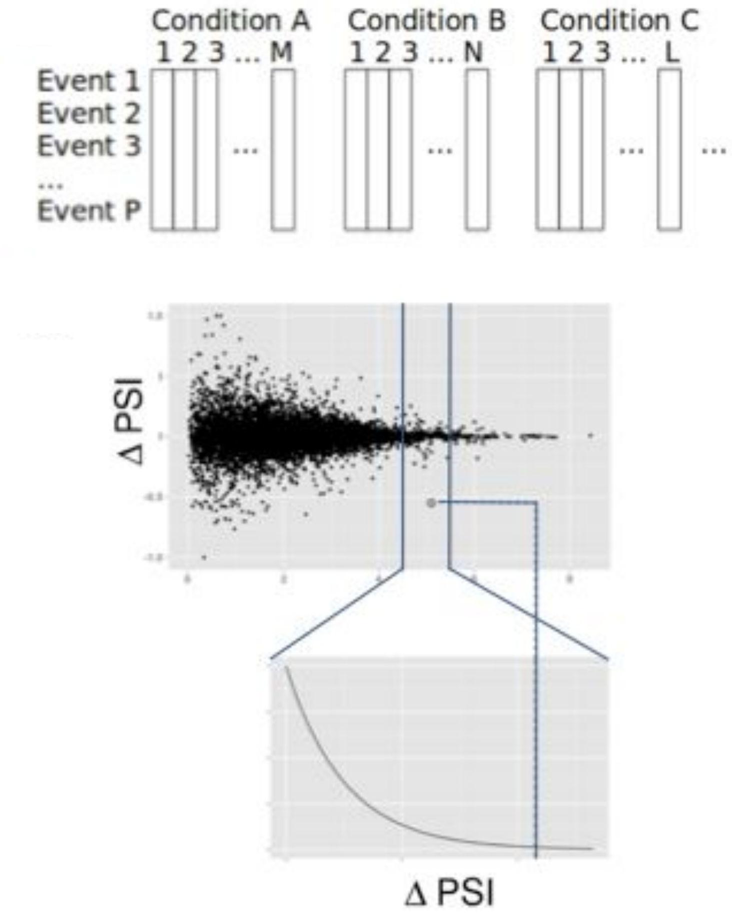


```
# number of alternative first exons
ls *.AF.psi|while read f;do echo -e $f"\t"$(grep -v nan $f |wc -l);done |
ggbarplot.R -i stdin -o number_of_AF.pdf -f 1 --palette_fill
/tutorial/palettes/palTissue.txt  --title "Alternative First Exon (AF)"
--y_title "Number of Events" --x_title "Tissues"
```

Cecilia Coimbra Klein

https://bitbucket.org/regulatorygenomicsupf/suppa

SUPPA: compare conditions

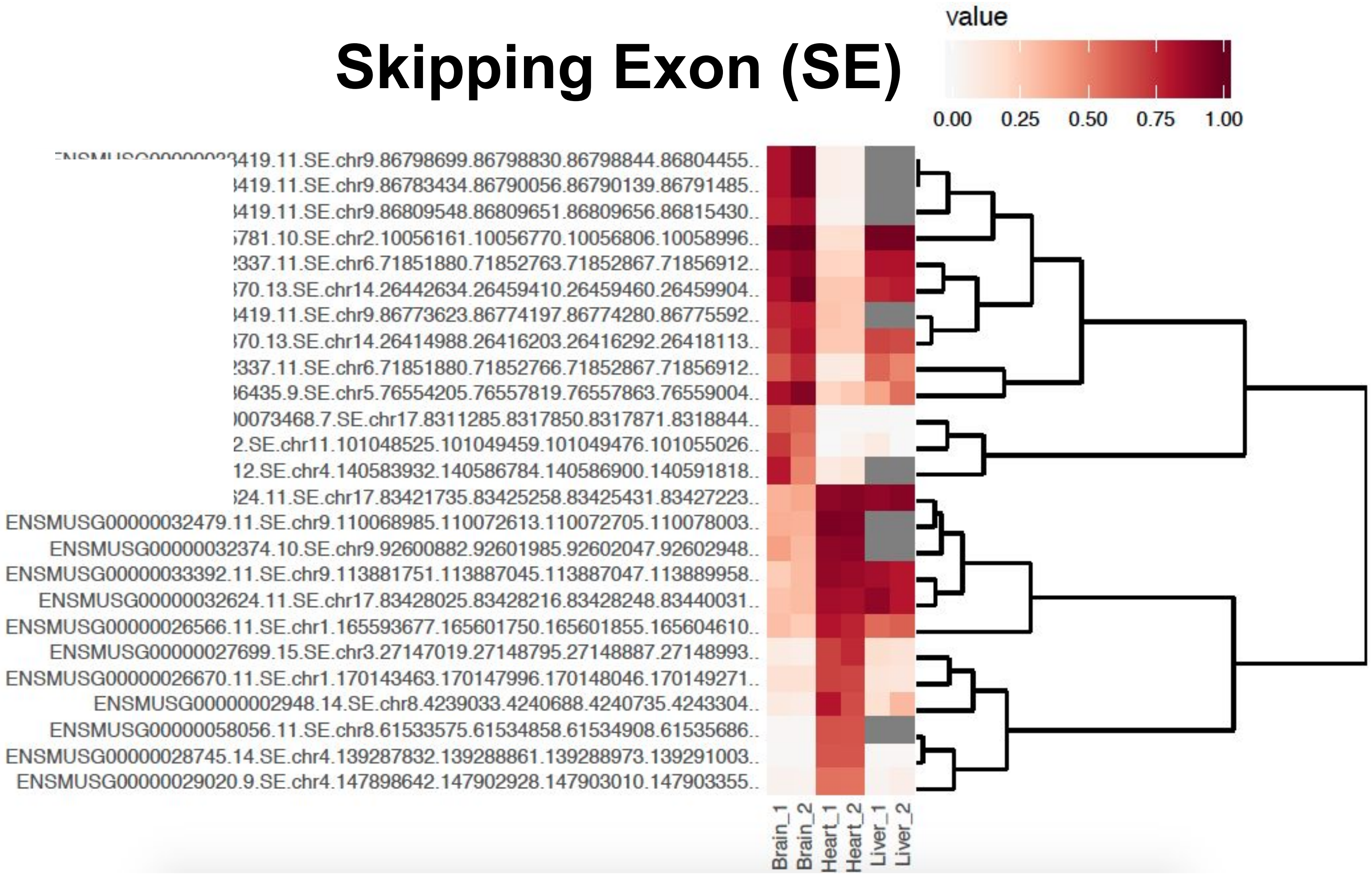


- SUPPA calculates the magnitude of splicing change (ΔPSI) and their significance across multiple biological conditions, using two or more replicates per condition.

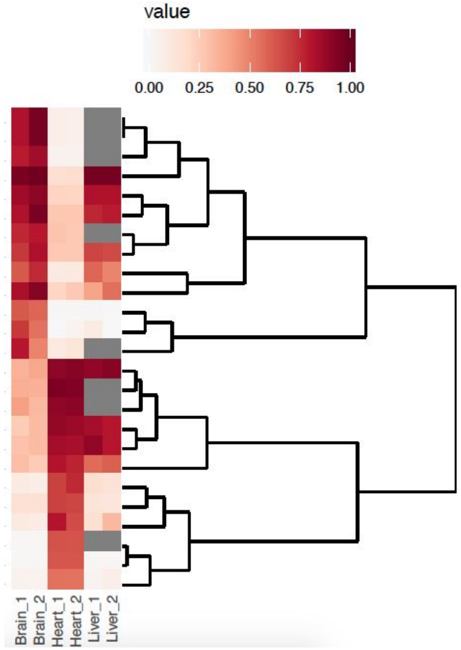
- Statistical significance is calculated by comparing the observed ΔPSI between conditions with the distribution of the ΔPSI between replicates as a function of the gene expression (measured as the expression of the transcripts defining the events).

https://bitbucket.org/regulatorygenomicsupf/suppa

Skipping Exon (SE)



Skipping Exon (SE)

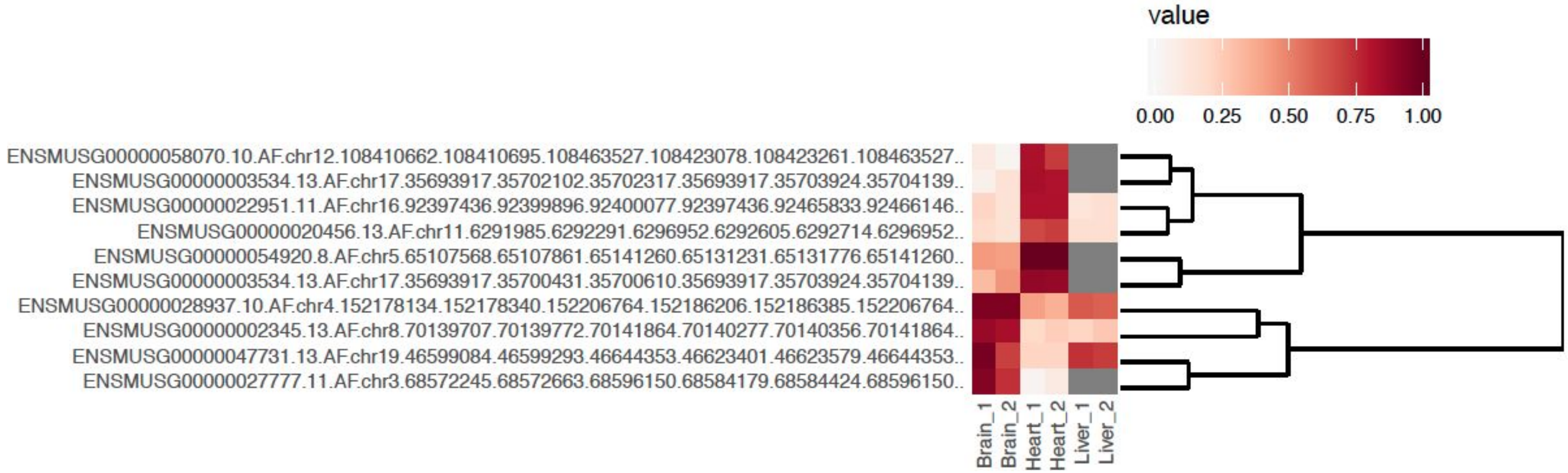


- Select top events from pairwise comparison of Brain and heart
- p-value < 0.05
- ΔPSI > 0.5
  - PSI = 1 inclusion of exon
  - PSI = 0 exclusion of exon

```
# prepare input for heatmap
event=SE; awk 'BEGIN{FS=OFS="\t"}NR>1 && $2!="nan" && ($2>0.5 || $2<-0.5)
&& $3<0.05{print}' DS.${event}.dpsi|cut -f1 > top-examples-SE.txt
selectMatrixRows.sh top-examples-SE.txt DS.SE.psivec >
matrix.top-examples-SE.tsv

# heatmap SE
ggheatmap.R -i matrix.top-examples-SE.tsv -o heatmap_top-examples-SE.pdf
--matrix_palette /tutorial/palettes/palSequential.txt --row_dendro
--matrix_fill_limits "0,1" -B 8
```

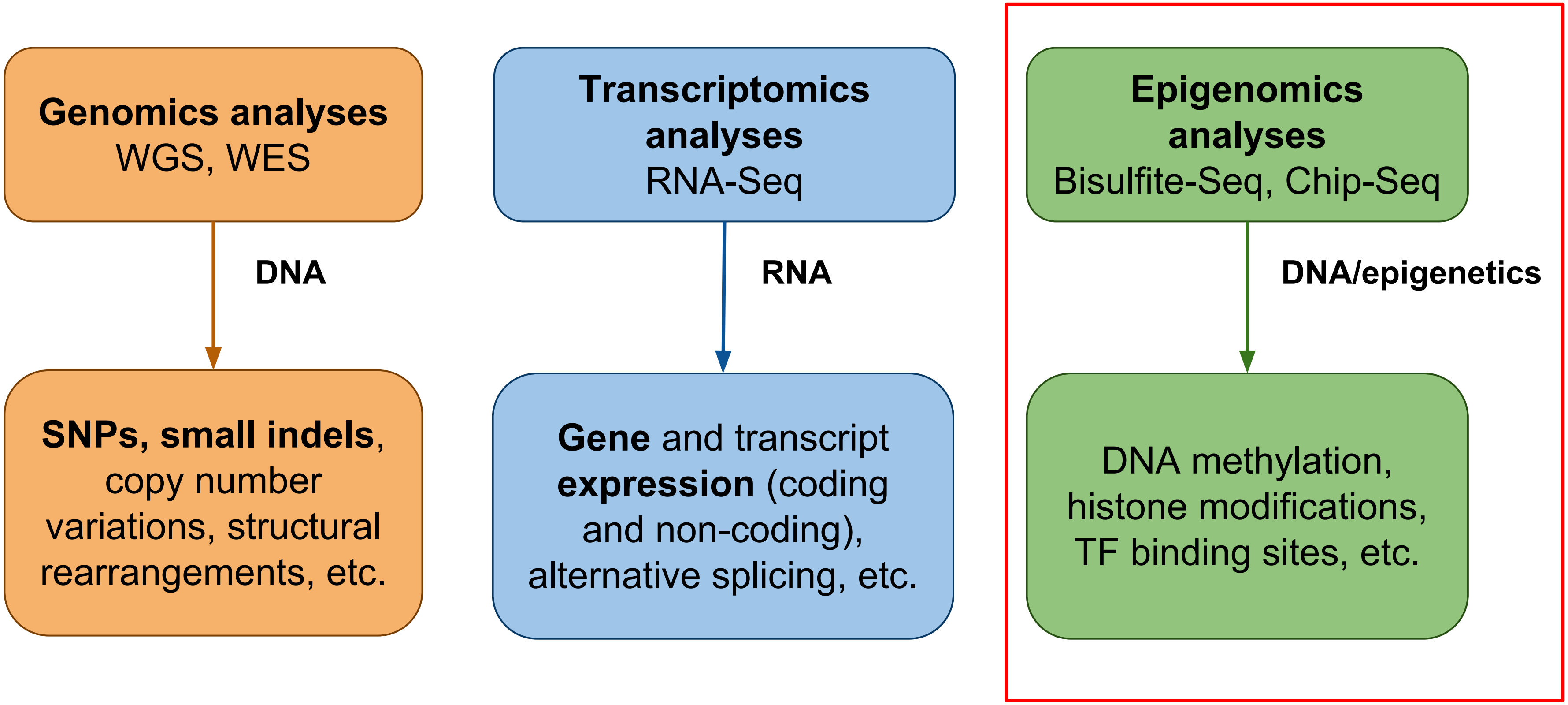
Alternative First exon (AF)



```
# prepare input for heatmap
event=AF; awk 'BEGIN{FS=OFS="\t"}NR>1 && $2!="nan" && ($2>0.5 || $2<-0.5)
&& $3<0.05{print}' DS.${event}.dpsi|cut -f1 > top-examples-AF.txt
selectMatrixRows.sh top-examples-AF.txt DS.AF.psivec >
matrix.top-examples-AF.tsv

# heatmap alternative first exons top examples
ggheatmap.R -i matrix.top-examples-AF.tsv -o heatmap_top-examples-AF.pdf
--matrix_palette /tutorial/palettes/palSequential.txt --row_dendro
--matrix_fill_limits "0,1" -B 8
```

Which *-Seq do I need?



Genomics analyses
WGS, WES

DNA

SNPs, small indels, copy number variations, structural rearrangements, etc.

Transcriptomics analyses
RNA-Seq

RNA

Gene and transcript **expression** (coding and non-coding), alternative splicing, etc.

Epigenomics analyses
Bisulfite-Seq, Chip-Seq

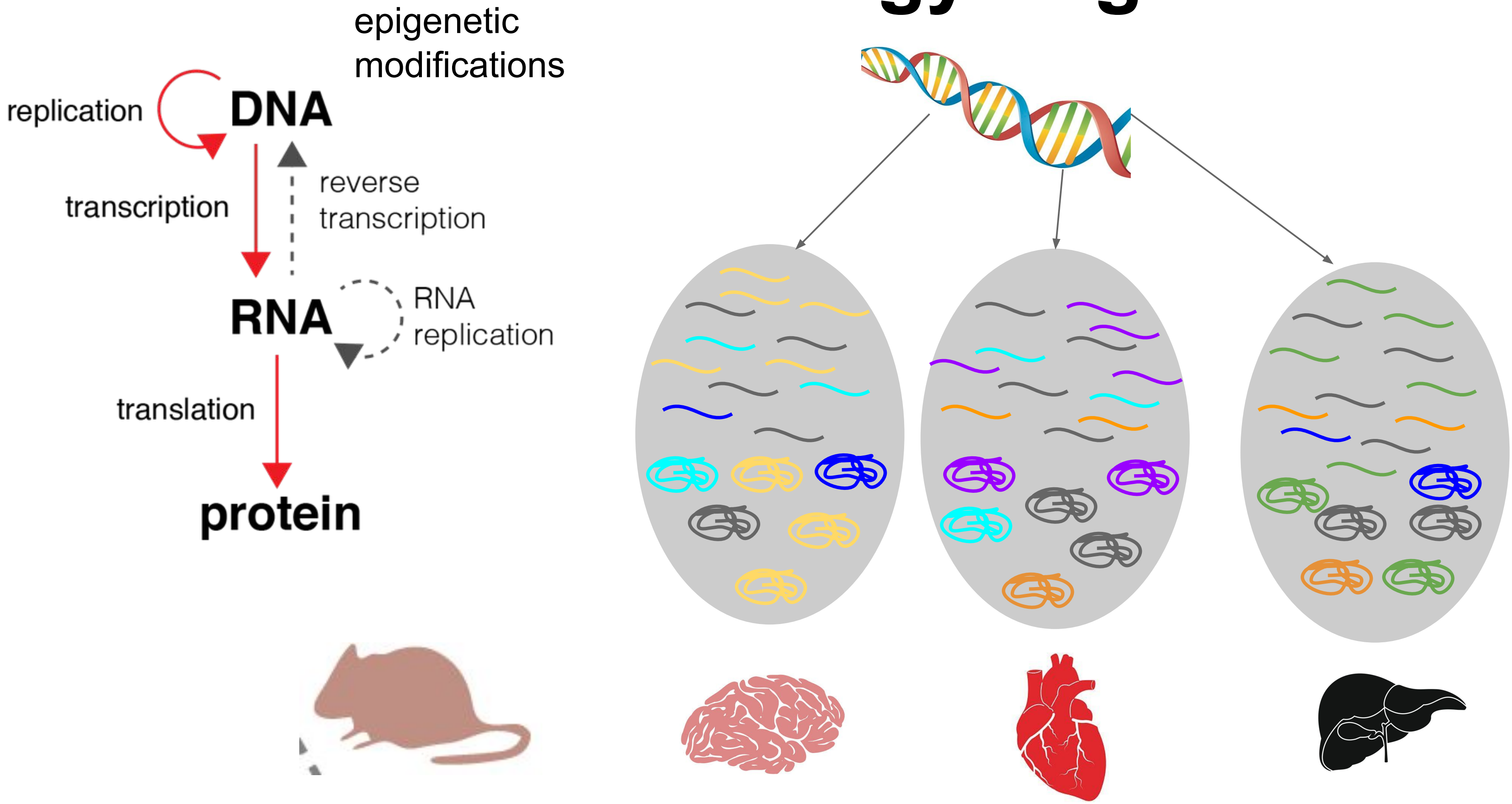
DNA/epigenetics

DNA methylation, histone modifications, TF binding sites, etc.

- Learn more about your favourite *-Seq here!
- Note that we are always talking about *re-sequencing*, which is something different from *de novo sequencing* (what is done for a new genome assembly)

Basic concepts

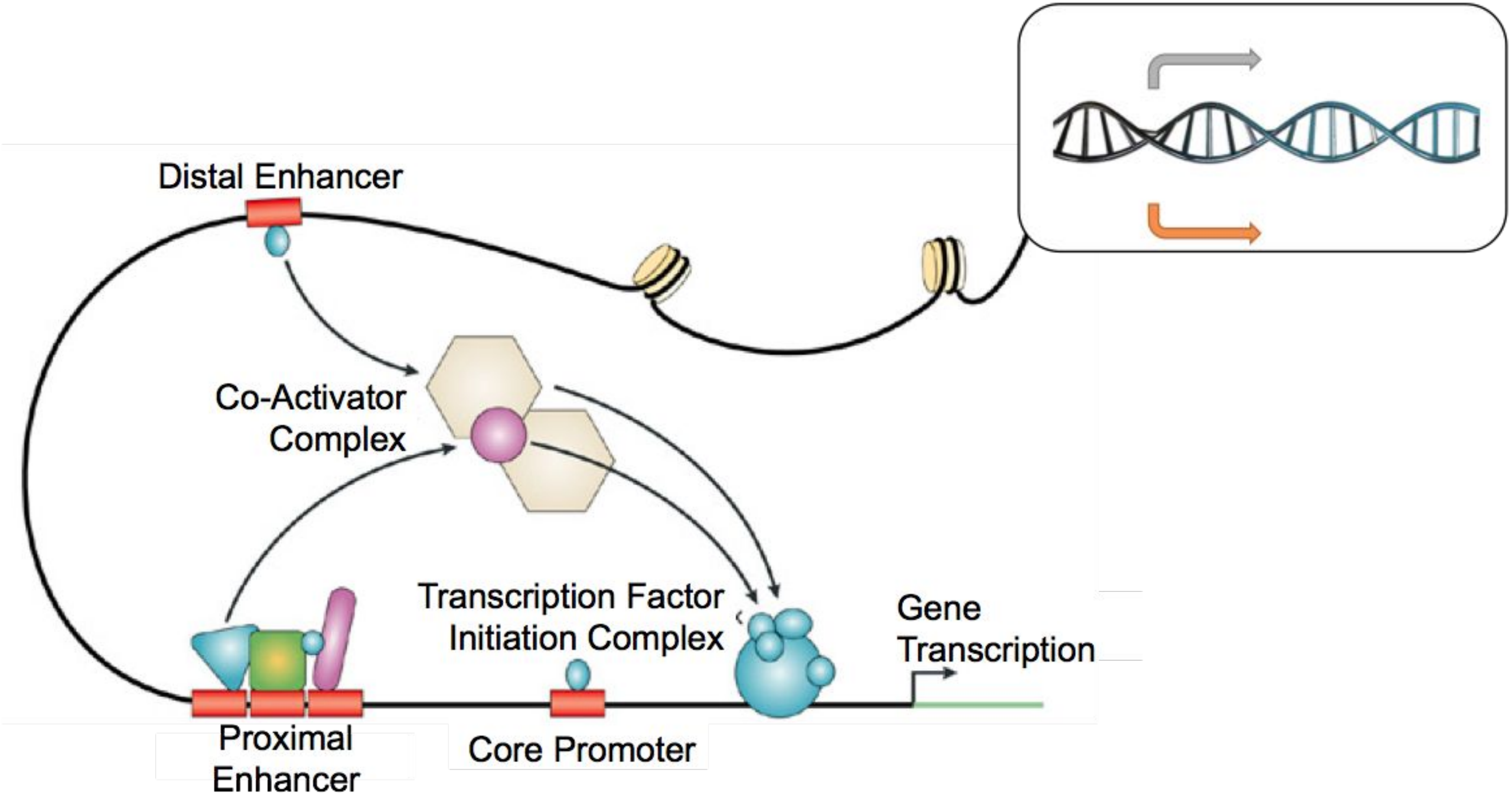
Molecular biology dogma



- The genome is identical in all cell types, however not all cell types have the same function. That’s why the transcriptome (and the epigenome) becomes also relevant.

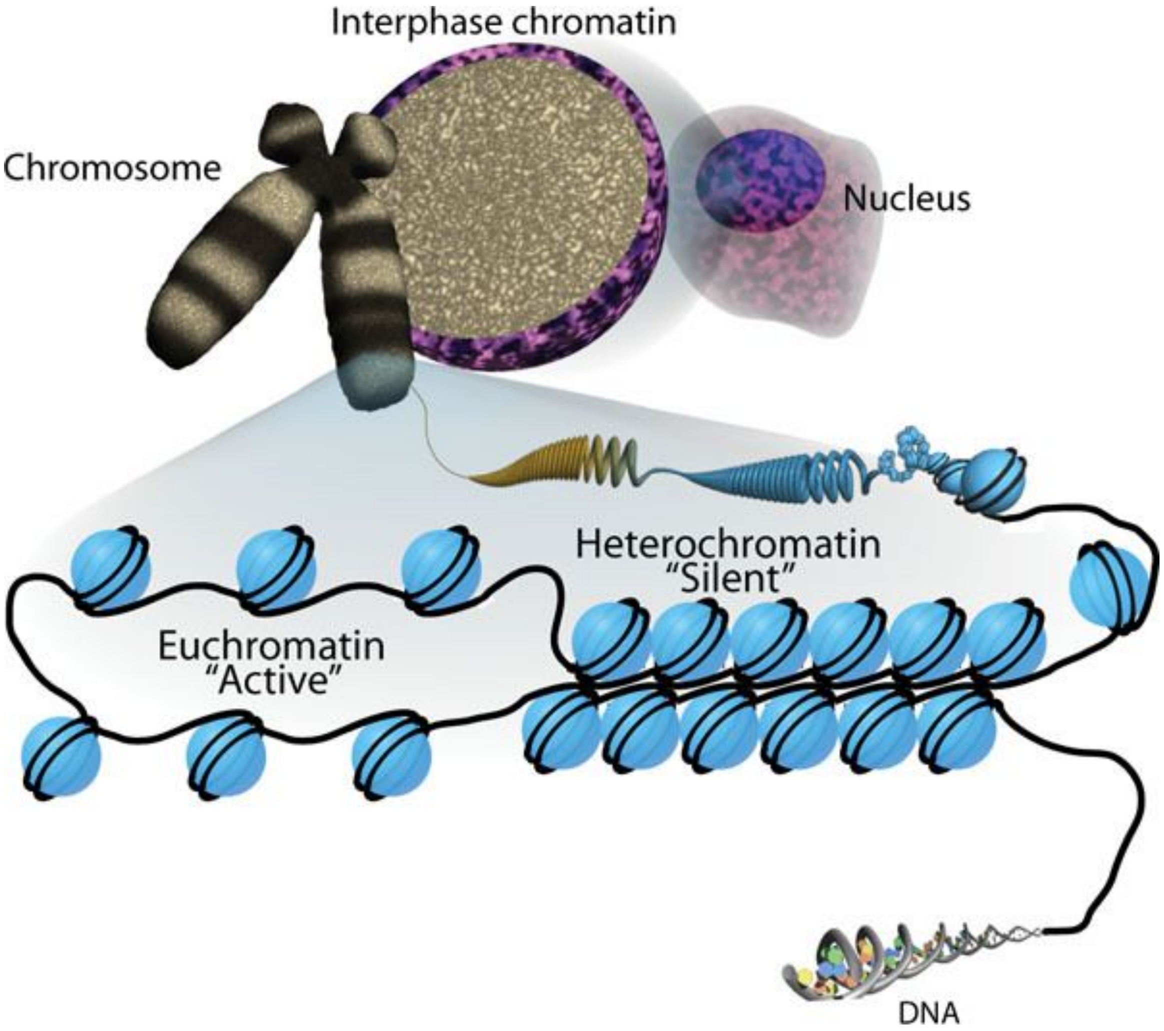
https://sdtimes.com/23andme/sd-times-blog-using-dna-for-access-control/
https://www.pinterest.com/pin/702139398127827000/
https://www.etsy.com/market/human_heart_svg
https://www.vectorstock.com/royalty-free-vector/human-liver-black-icon-vector-7458173

Dynamics of gene regulation



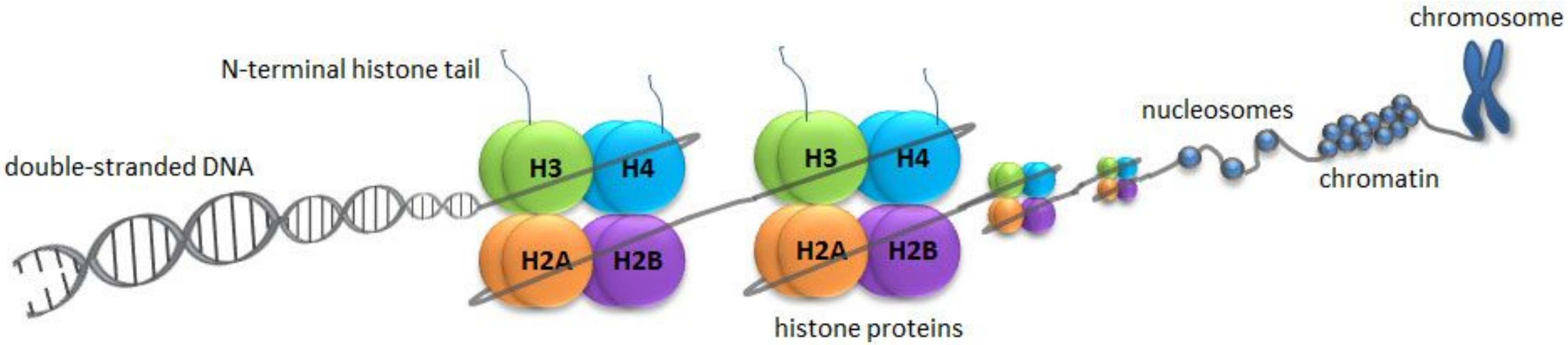
Cecilia Coimbra Klein

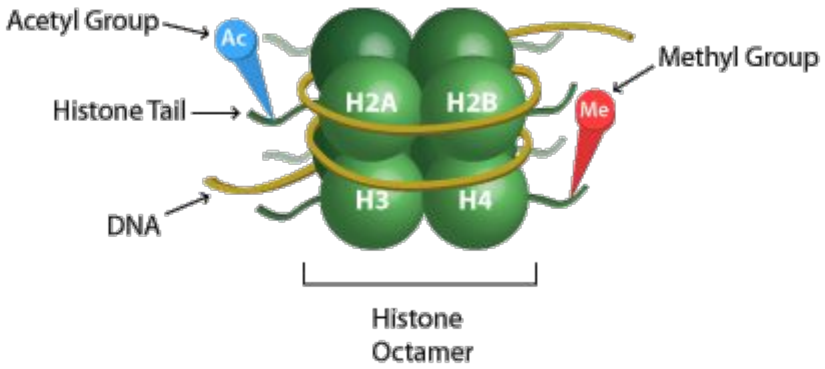
Chromatin organization



The chromatin signature of pluripotent cells, StemBook, NCBI
https://www.ncbi.nlm.nih.gov/books/NBK27041/figure/thechromatinsignature.F1/

Histone modifications

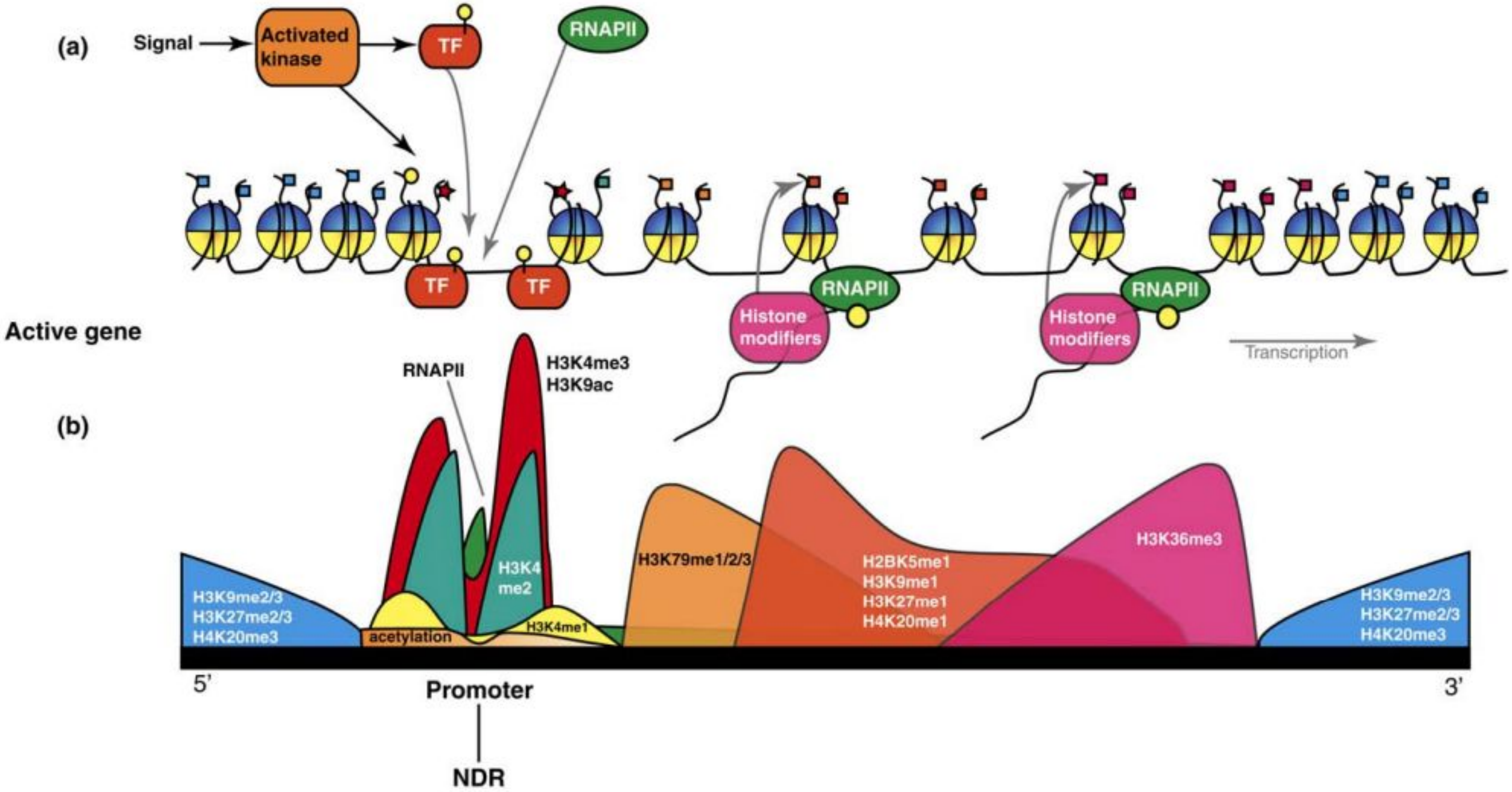




- methylation, acetylation and phosphorylation
- involve covalent post-translational modifications mostly to the residues at the positively charged N-terminal tails of histones.

https://www.whatisepigenetics.com/histone-modifications/
https://www.epigentek.com/catalog/advanced-epigenetic-overview-of-histone-modifications-n-5.html?currency=es

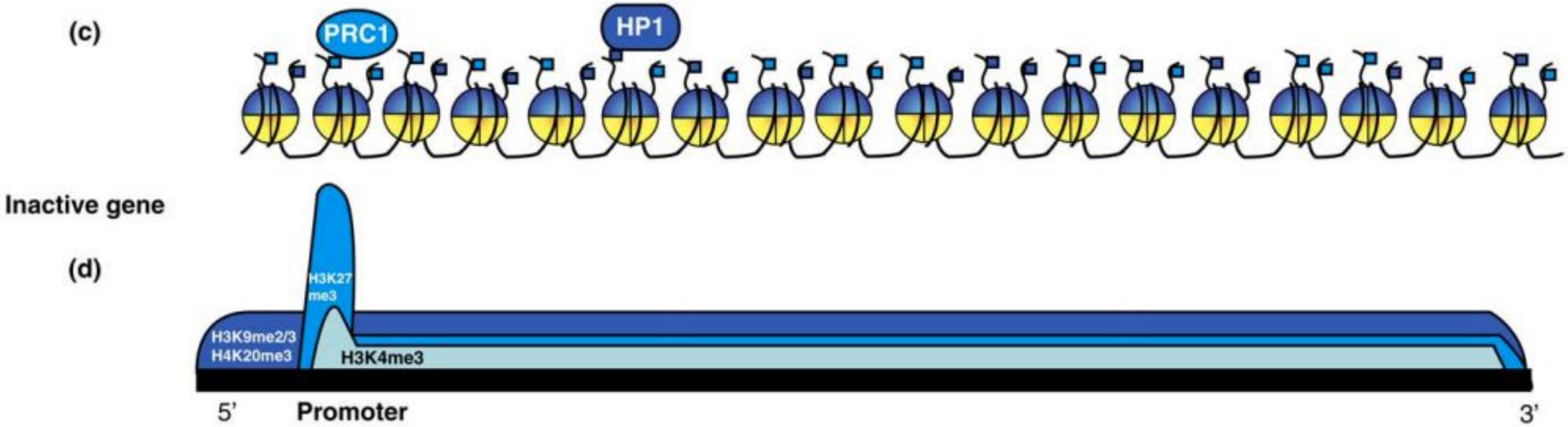
Histone modifications



Barth & Imhof (2010) Trends in Biochemical Sciences

Cecilia Coimbra Klein

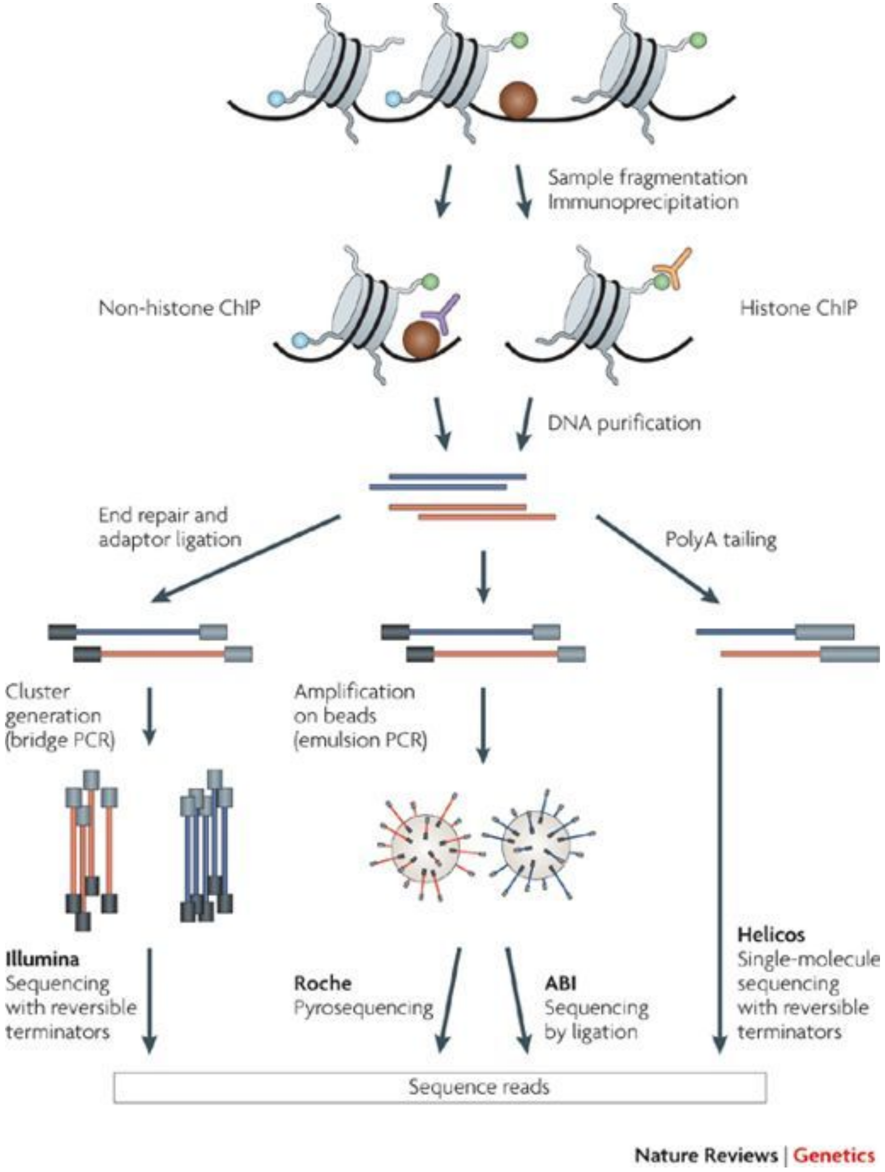
Histone modifications



Barth & Imhof (2010) Trends in Biochemical Sciences

Cecilia Coimbra Klein

ChIP-seq: Chromatin ImmunoPrecipitation

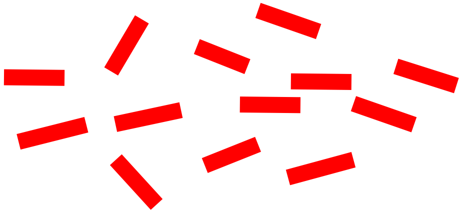

Sample fragmentation
Immunoprecipitation
Non-histone ChIP
Histone ChIP
DNA purification
End repair and adaptor ligation
PolyA tailing
Cluster generation (bridge PCR)
Amplification on beads (emulsion PCR)
Illumina
Sequencing with reversible terminators
Roche
Pyrosequencing
ABI
Sequencing by ligation
Helicos
Single-molecule sequencing with reversible terminators
Sequence reads
Nature Reviews | Genetics

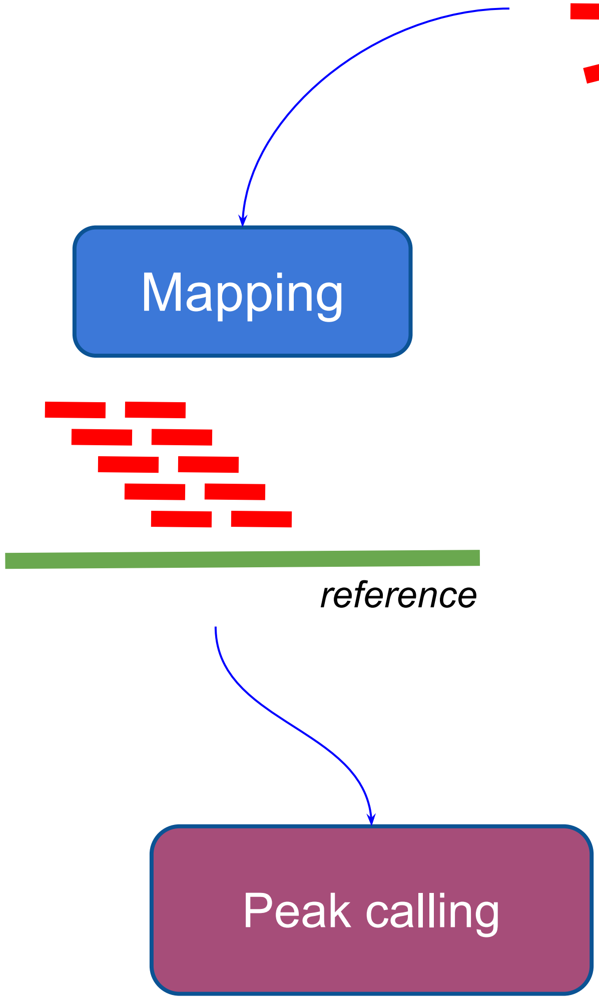
Park (2009) doi: 10.1038/nrg2641

Cecilia Coimbra Klein

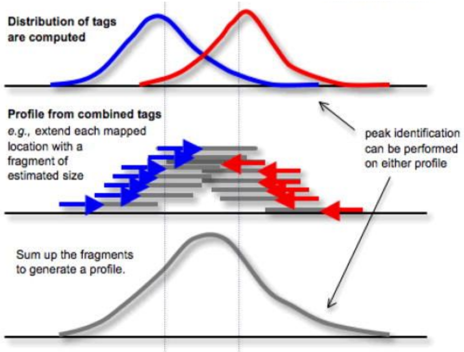
Processing

Mapping



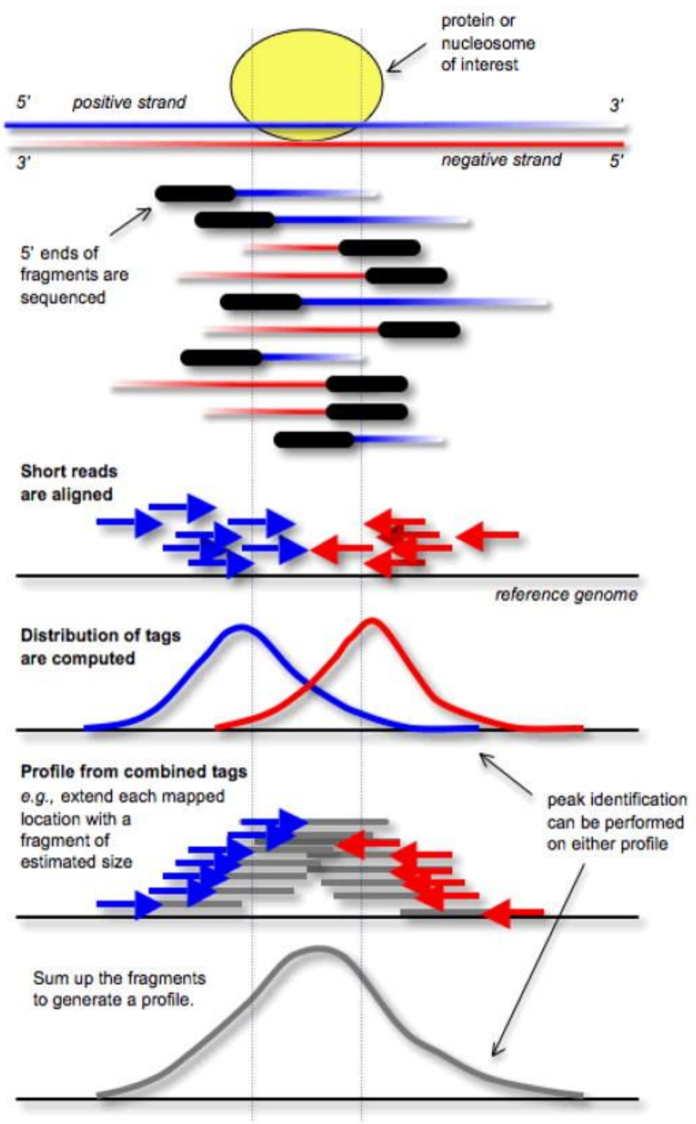


Find a correspondence between the query sequences (ChIP-seq reads) and our prior knowledge (reference genome sequence).


Distribution of tags are computed
Profile from combined tags
e.g., extend each mapped location with a fragment of estimated size
peak identification can be performed on either profile
Sum up the fragments to generate a profile.

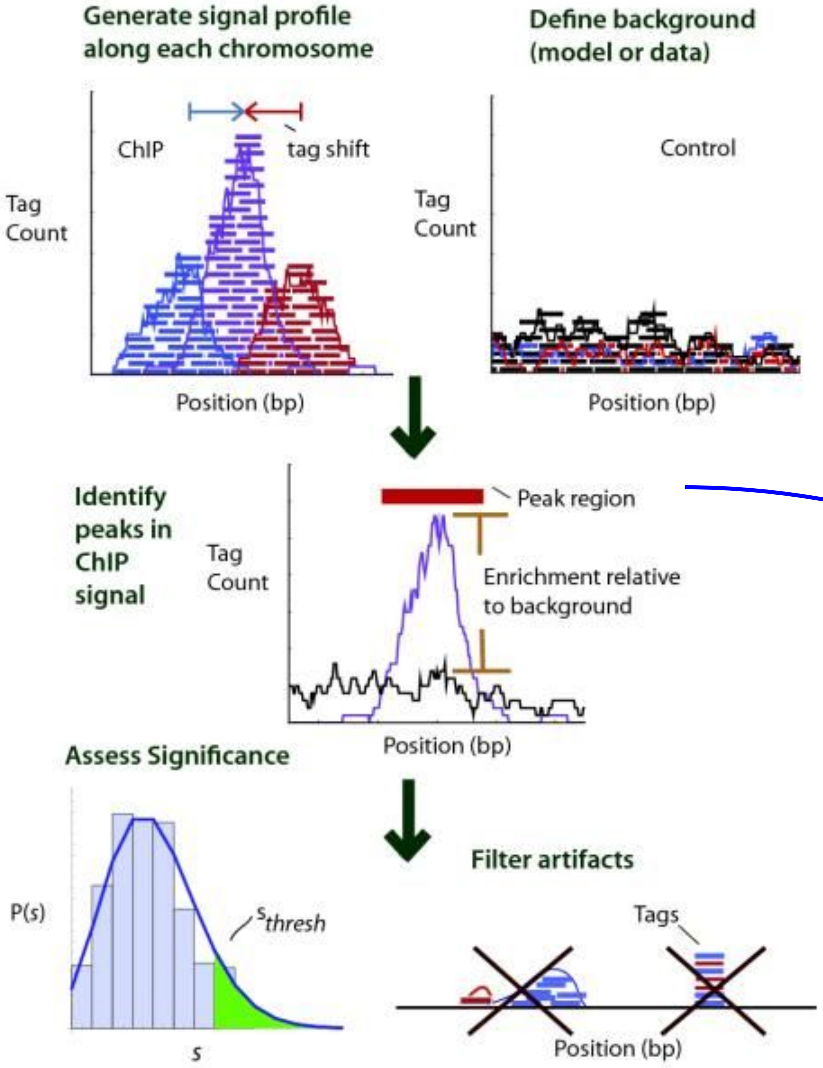
Cecilia Coimbra Klein

Peak calling


protein or nucleosome of interest
5'
positive strand
3'
3'
negative strand
5'
5' ends of fragments are sequenced
Short reads are aligned
reference genome
Distribution of tags are computed
Profile from combined tags
e.g., extend each mapped location with a fragment of estimated size
peak identification can be performed on either profile
Sum up the fragments to generate a profile.

Park (2009) doi: 10.1038/nrg2641

Peak calling: MACS2



Signal profile (bigWig)

Peak regions (BED)

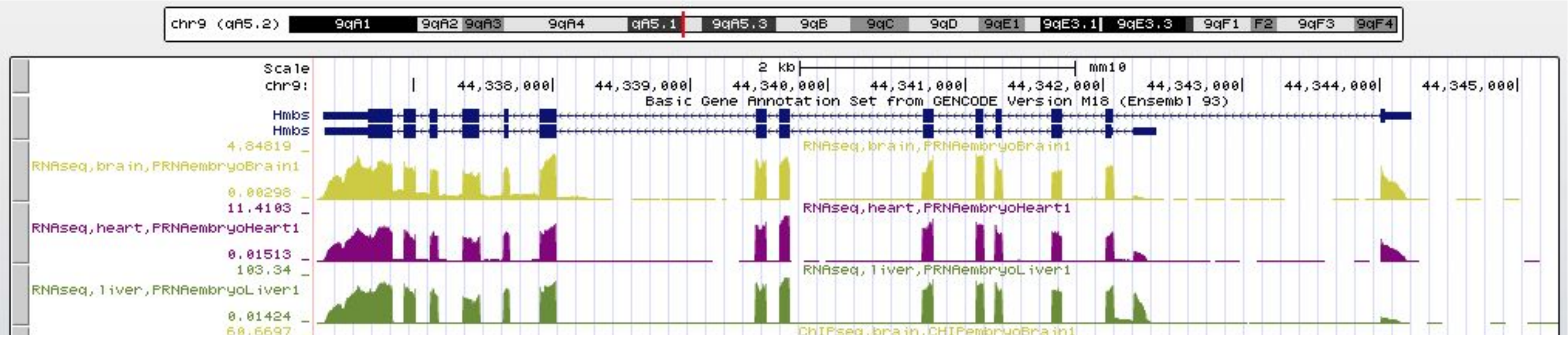
Pepke et al. (2009) doi: 10.1038/nmeth.1371

Cecilia Coimbra Klein

Visualization

RNA-seq signal

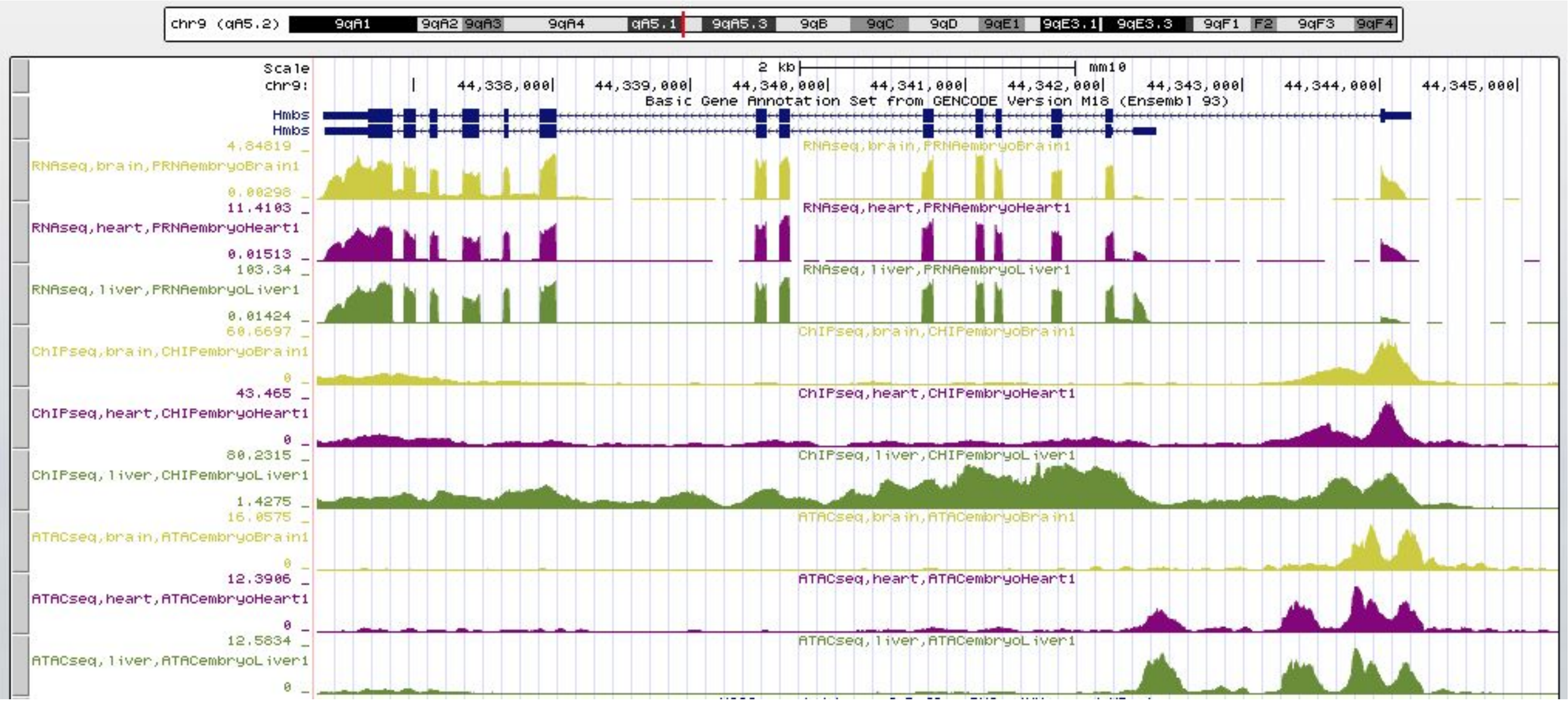
genome-euro.ucsc.edu



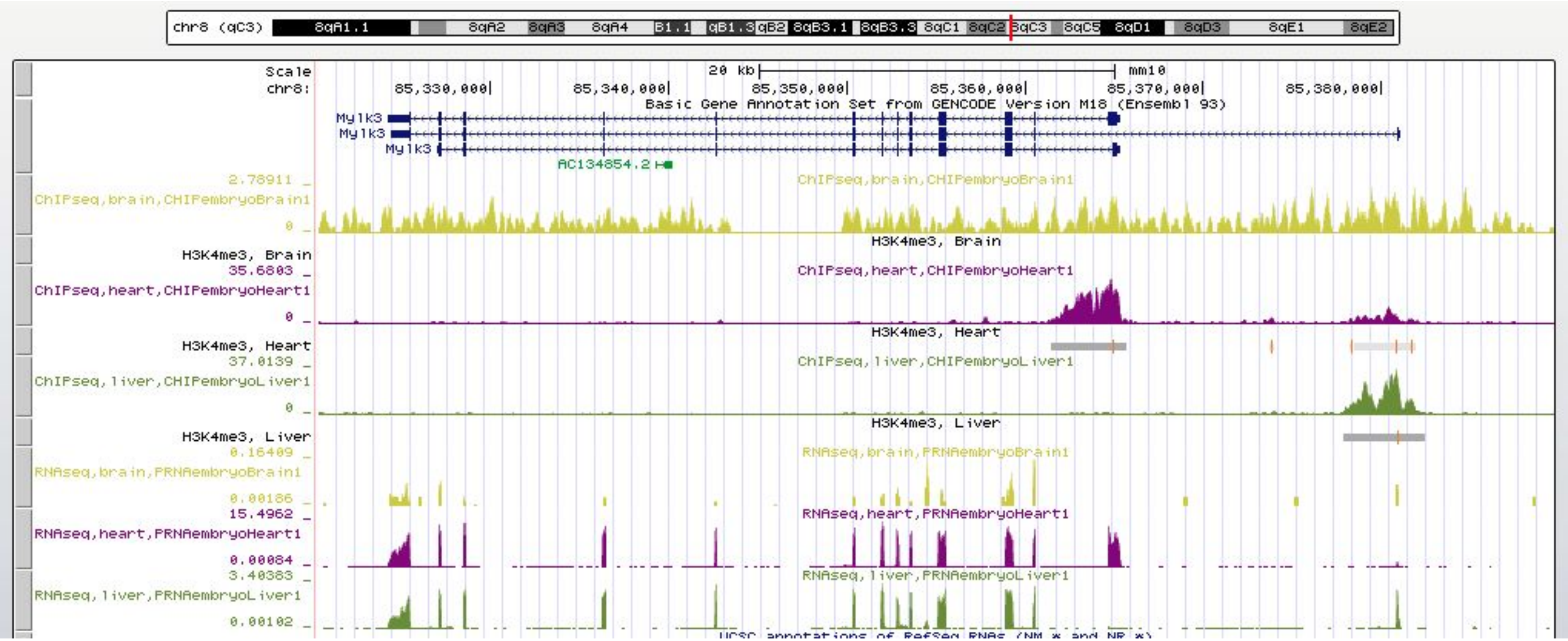
- expected read depth at each position in the genome
- can be normalized (e.g. RPM, reads per million reads)

UCSC: signal files

genome-euro.ucsc.edu


chr9 (qA5.2)
9qA1
9qA2
9qA3
9qA4
qA5.1
9qA5.3
9qB
9qC
9qD
9qE1
9qE3.1
9qE3.3
9qF1
F2
9qF3
9qF4
Scale
chr9:
2 kb
mm10
44,338,000
44,339,000
44,340,000
44,341,000
44,342,000
44,343,000
44,344,000
44,345,000
Basic Gene Annotation Set from GENCODE Version M18 (Ensembl 93)
Hmbs
Hmbs
4.84819
RNAseq,brain,PRNAembryoBrain1
0.00298
11.4103
RNAseq,heart,PRNAembryoHeart1
0.01513
103.34
RNAseq,liver,PRNAembryoLiver1
0.01424
60.6697
ChIPseq,brain,CHIPembryoBrain1
0
43.465
ChIPseq,heart,CHIPembryoHeart1
0
80.2315
ChIPseq,liver,CHIPembryoLiver1
1.4275
16.0575
ATACseq,brain,ATACembryoBrain1
0
12.3906
ATACseq,heart,ATACembryoHeart1
0
12.5834
ATACseq,liver,ATACembryoLiver1
0

UCSC

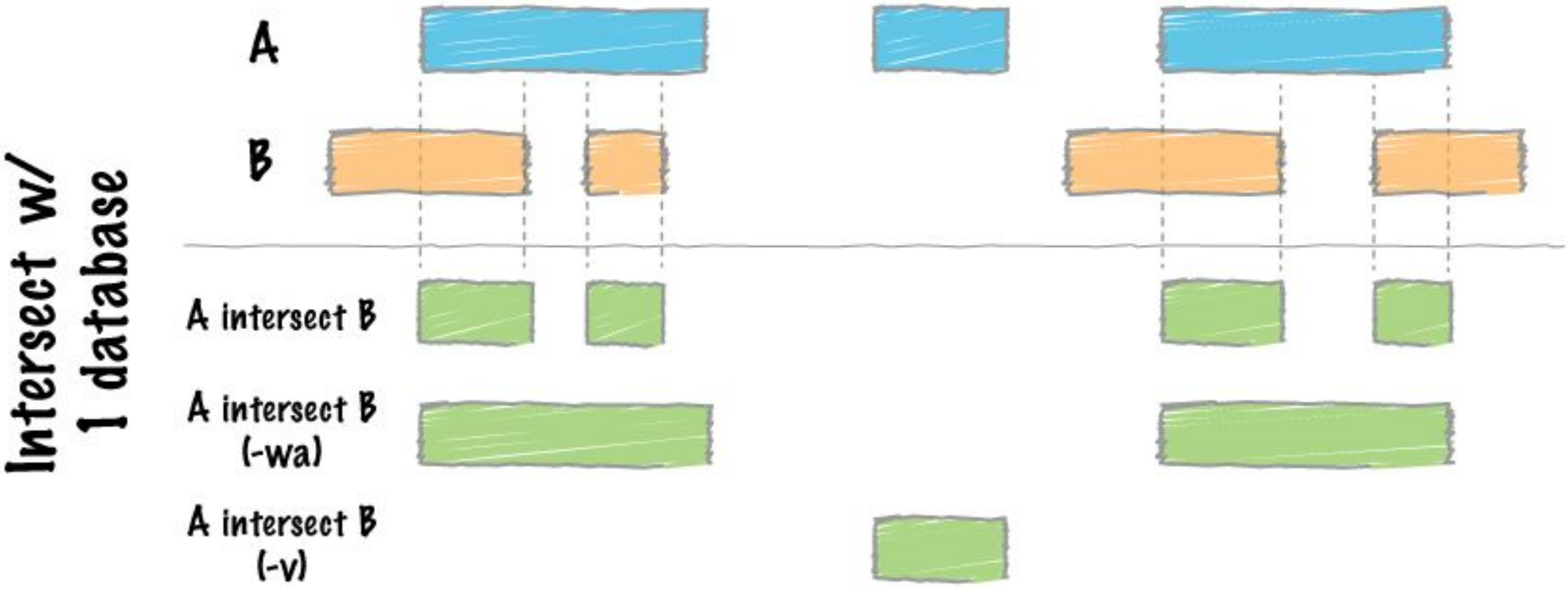


chr8 (qC3)
8qA1.1
8qA2
8qA3
8qA4
B1.1
qB1.3
qB2
8qB3.1
8qB3.3
8qC1
8qC2
8qC3
8qC5
8qD1
8qD3
8qE1
8qE2
Scale
chr8:
20 kb
mm10
85,330,000
85,340,000
85,350,000
85,360,000
85,370,000
85,380,000
Basic Gene Annotation Set from GENCODE Version M18 (Ensembl 93)
Mylk3
Mylk3
Mylk3
AC134854.2
2.78911 _
ChIPseq,brain,CHIPembryoBrain1
0 _
ChIPseq,brain,CHIPembryoBrain1
H3K4me3, Brain
H3K4me3, Brain
35.6803 _
ChIPseq,heart,CHIPembryoHeart1
0 _
ChIPseq,heart,CHIPembryoHeart1
H3K4me3, Heart
H3K4me3, Heart
37.0139 _
ChIPseq,liver,CHIPembryoLiver1
0 _
ChIPseq,liver,CHIPembryoLiver1
H3K4me3, Liver
H3K4me3, Liver
0.16409 _
RNAseq,brain,PRNAembryoBrain1
0.00186 _
RNAseq,brain,PRNAembryoBrain1
15.4962 _
RNAseq,heart,PRNAembryoHeart1
0.00084 _
RNAseq,heart,PRNAembryoHeart1
3.40383 _
RNAseq,liver,PRNAembryoLiver1
0.00102 _
RNAseq,liver,PRNAembryoLiver1
UCSC annotations of RefSeq RNAs (NM_* and NR_*)

Cecilia Coimbra Klein

Analysis

BEDTools intersect

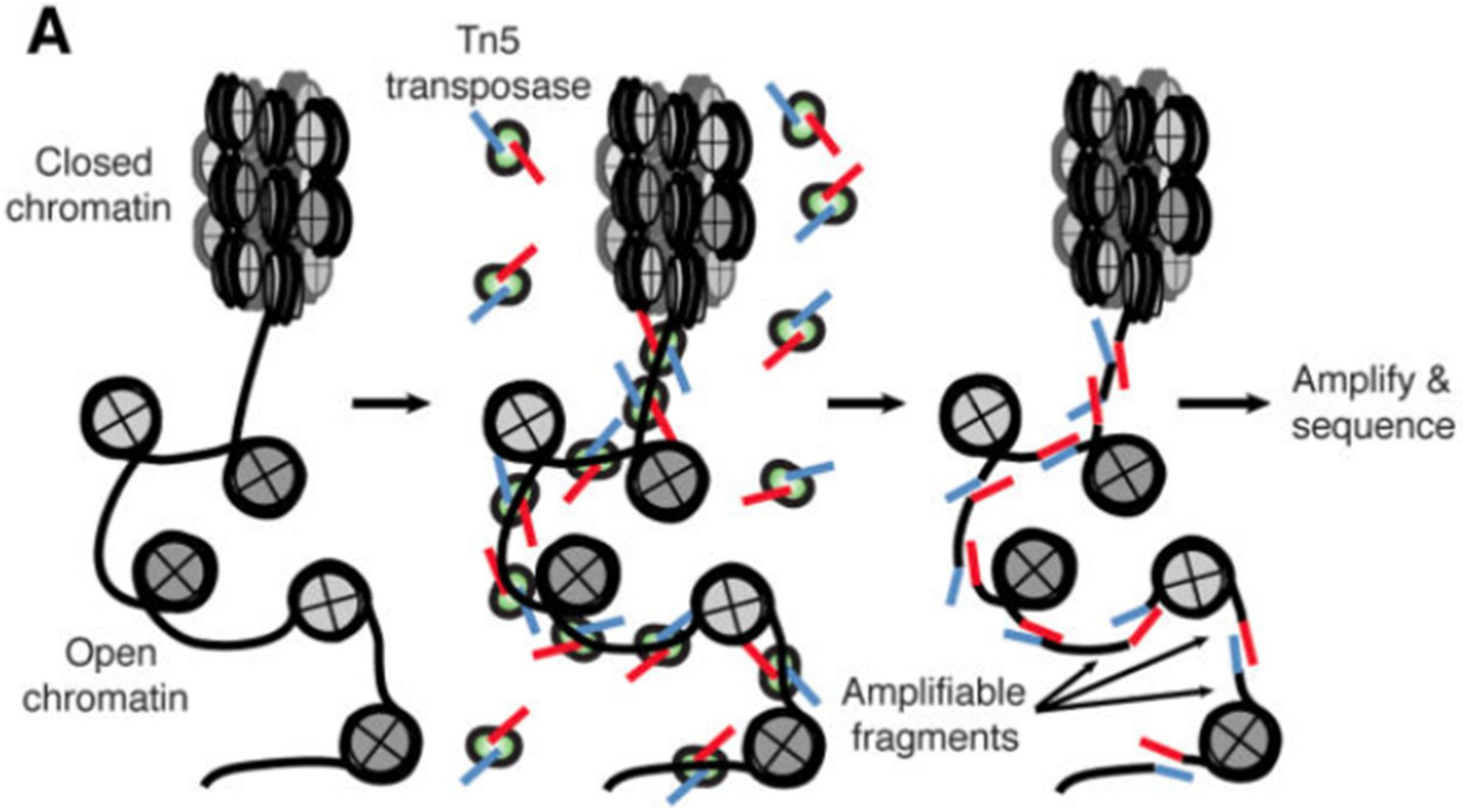


https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html

Cecilia Coimbra Klein

ATAC-seq

Open chromatin



Buenrostro *et al.* (2015) doi:10.1002/0471142727.mb2129s109

Hands-on

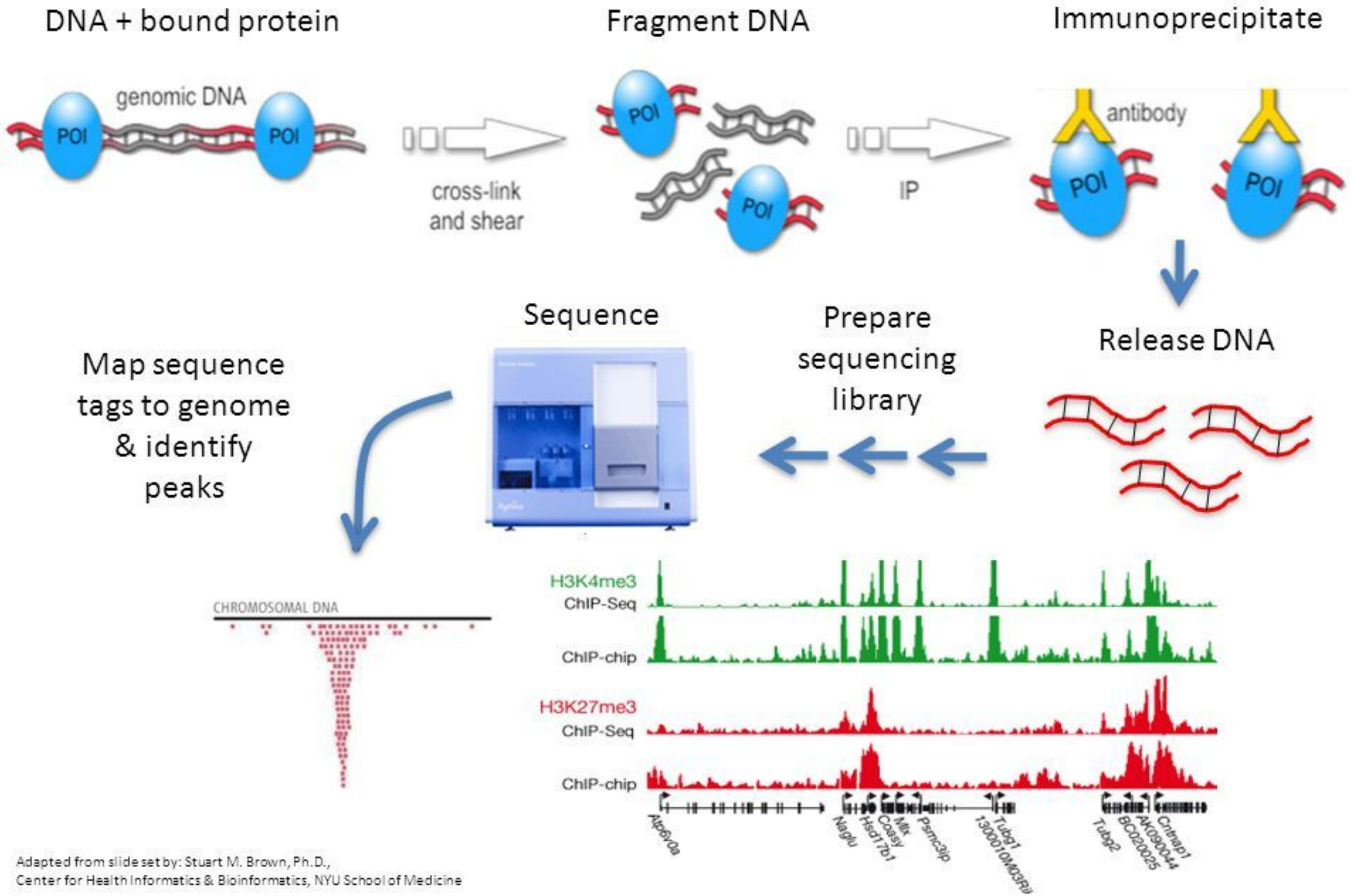
Setup environment 1

Integrative data analysis 7

https://public_docs.crg.es/rguigo/Data/cklein/courses/UVIC/handsOn/

Additional slides

ChIP-seq: Chromatin ImmunoPrecipitation



DNA + bound protein

Fragment DNA

Immunoprecipitate

Release DNA

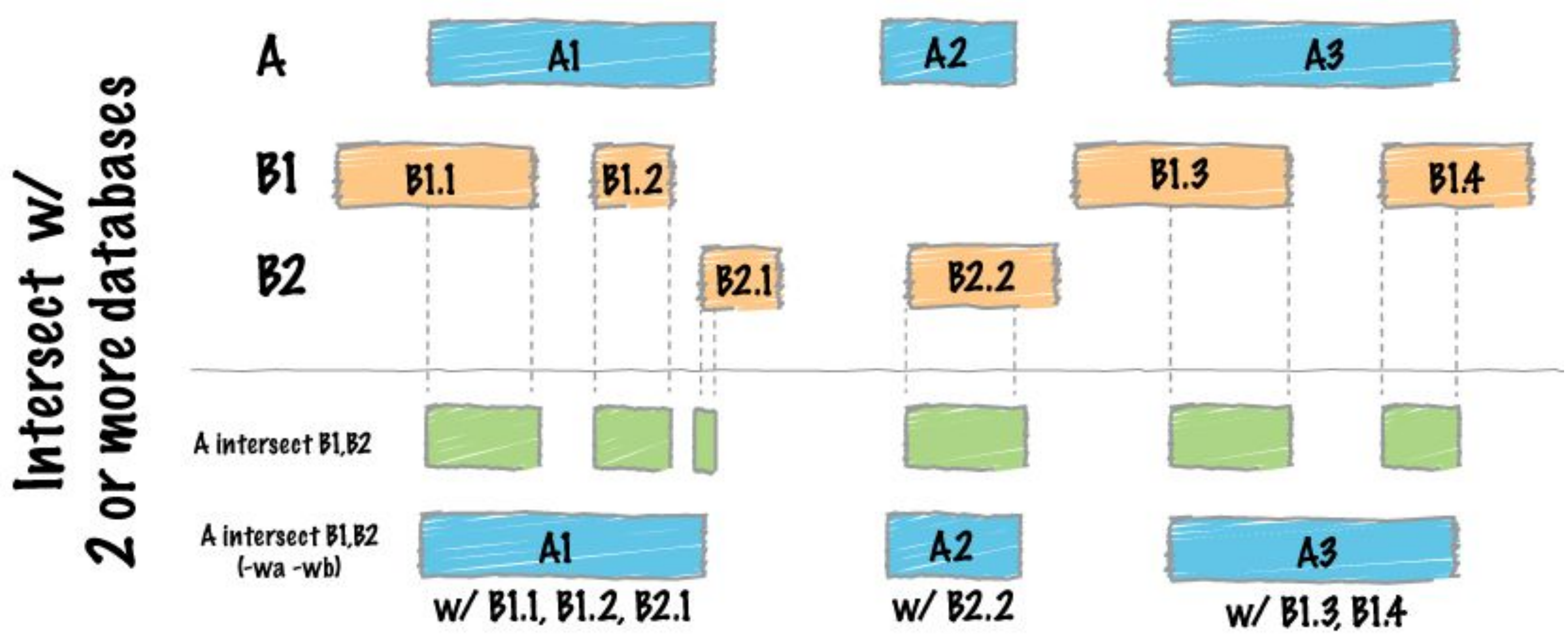
Prepare sequencing library

Sequence

Map sequence tags to genome & identify peaks

Adapted from slide set by: Stuart M. Brown, Ph.D.,
Center for Health Informatics & Bioinformatics, NYU School of Medicine

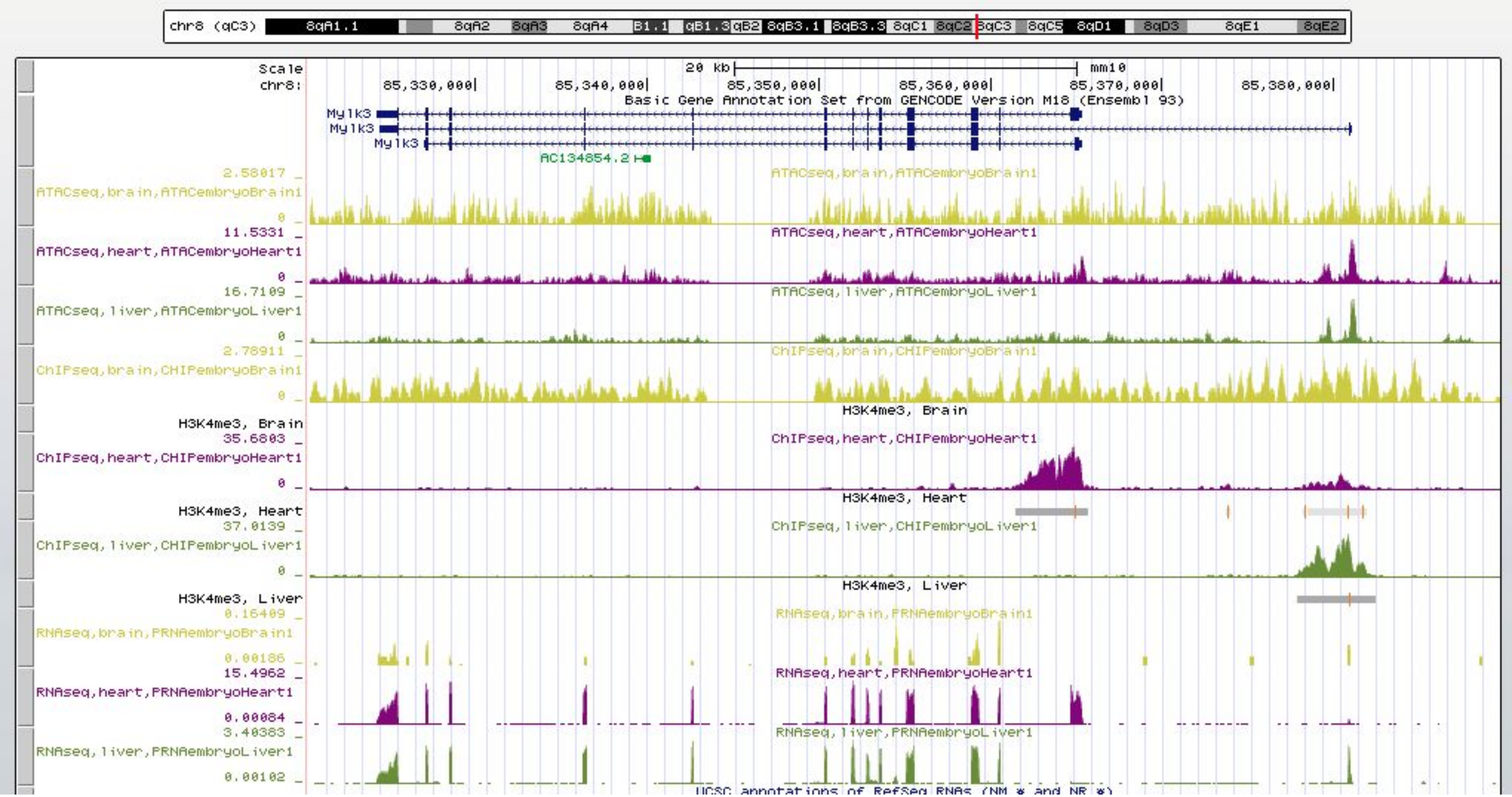
BEDTools intersect



https://bedtools.readthedocs.io/en/latest/content/tools/intersect.html

Cecilia Coimbra Klein

UCSC


chr8 (qC3)
8qA1.1
8qA2
8qA3
8qA4
B1.1
qB1.3
qB2
8qB3.1
8qB3.3
8qC1
8qC2
8qC3
8qC5
8qD1
8qD3
8qE1
8qE2
Scale
chr8:
20 kb
mm10
85,330,000
85,340,000
85,350,000
85,360,000
85,370,000
85,380,000
Basic Gene Annotation Set from GENCODE Version M18 (Ensembl 93)
Mylk3
Mylk3
Mylk3
AC134854.2
2.58017
ATACseq,brain,ATACembryoBrain1
0
11.5331
ATACseq,heart,ATACembryoHeart1
0
16.7109
ATACseq,liver,ATACembryoLiver1
0
2.78911
ChIPseq,brain,CHIPembryoBrain1
0
H3K4me3, Brain
35.6803
ChIPseq,heart,CHIPembryoHeart1
0
H3K4me3, Heart
37.0139
ChIPseq,liver,CHIPembryoLiver1
0
H3K4me3, Liver
0.16409
RNAseq,brain,PRNAembryoBrain1
0.00186
15.4962
RNAseq,heart,PRNAembryoHeart1
0.00084
3.40383
RNAseq,liver,PRNAembryoLiver1
0.00102
UCSC annotations of RefSeq RNAs (NM_* and NR_*)