Genomics analysis

Yu LI (李煜)

Thursday, 31 October 2024

liyu95.com

liyu@cse.cuhk.edu.hk

Department of Computer Science and Engineering (CSE)

The Chinese University of Hong Kong (CUHK)

Gene enrichment analysis

- A biological pathway is a series of interactions among molecules in a cell that leads to a certain product or a change in a cell. Such a pathway can trigger the assembly of new molecules, such as a fat or protein. Pathways can also turn genes on and off, or spur a cell to move
  - KEGG pathway database
  - Each pathway contains a set of genes

- By experiments, researchers identified 213 genes associated with type-II diabetes

- Question: how to identify pathways related with type-II diabetes?

+ve / -ve correlation

What is Fisher’s exact test?

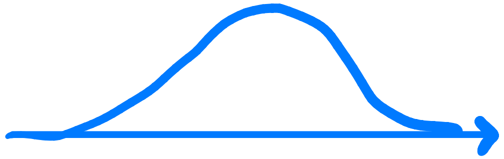
- ❖Fisher's exact test is a statistical significance test used in the analysis of contingency tables

- ❖Why is it called exact test?
  - ➢P-value can be calculated exactly from the table
  - ➢Recall t-test
  - ➢We calculate a t-value
  - ➢Based on a distribution, we get the p-value

normal distribution

	In gene set	Not in gene set	Total
In pathway	100 (a)	9000 (b)	9100
Not in pathway	113 (c)	11000 (d)	11113
Total	213	20000	20213

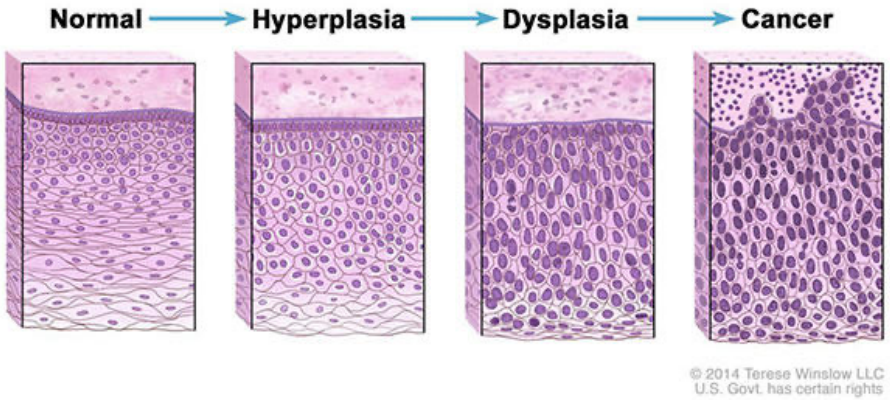
SD = 1

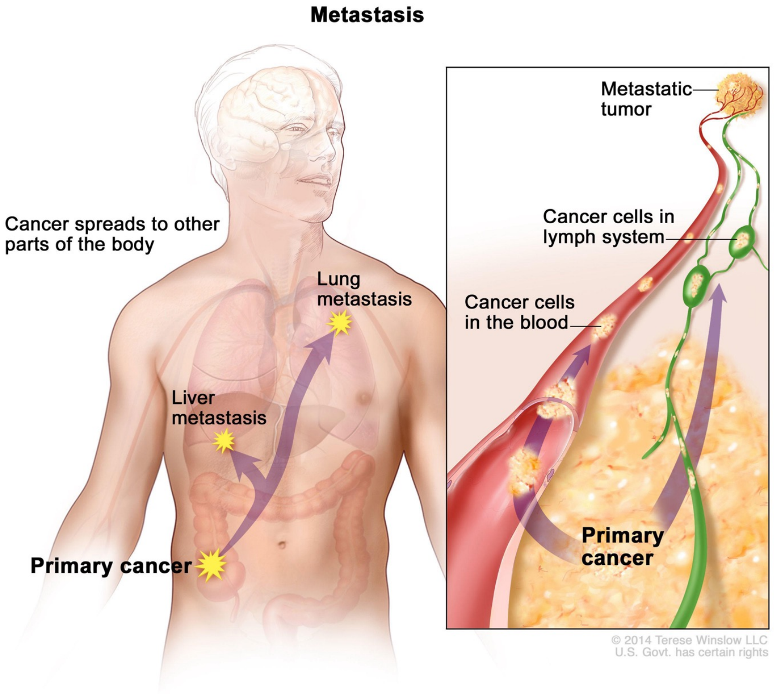


❖ $$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}$$

What is cancer?

- ❖Cancer is a disease in which some of the body’s cells grow uncontrollably and spread to other parts of the body




Metastasis
Cancer spreads to other parts of the body
Lung metastasis
Liver metastasis
Primary cancer
Metastatic tumor
Cancer cells in lymph system
Cancer cells in the blood
Primary cancer
© 2014 Terese Winslow LLC
U.S. Govt. has certain rights

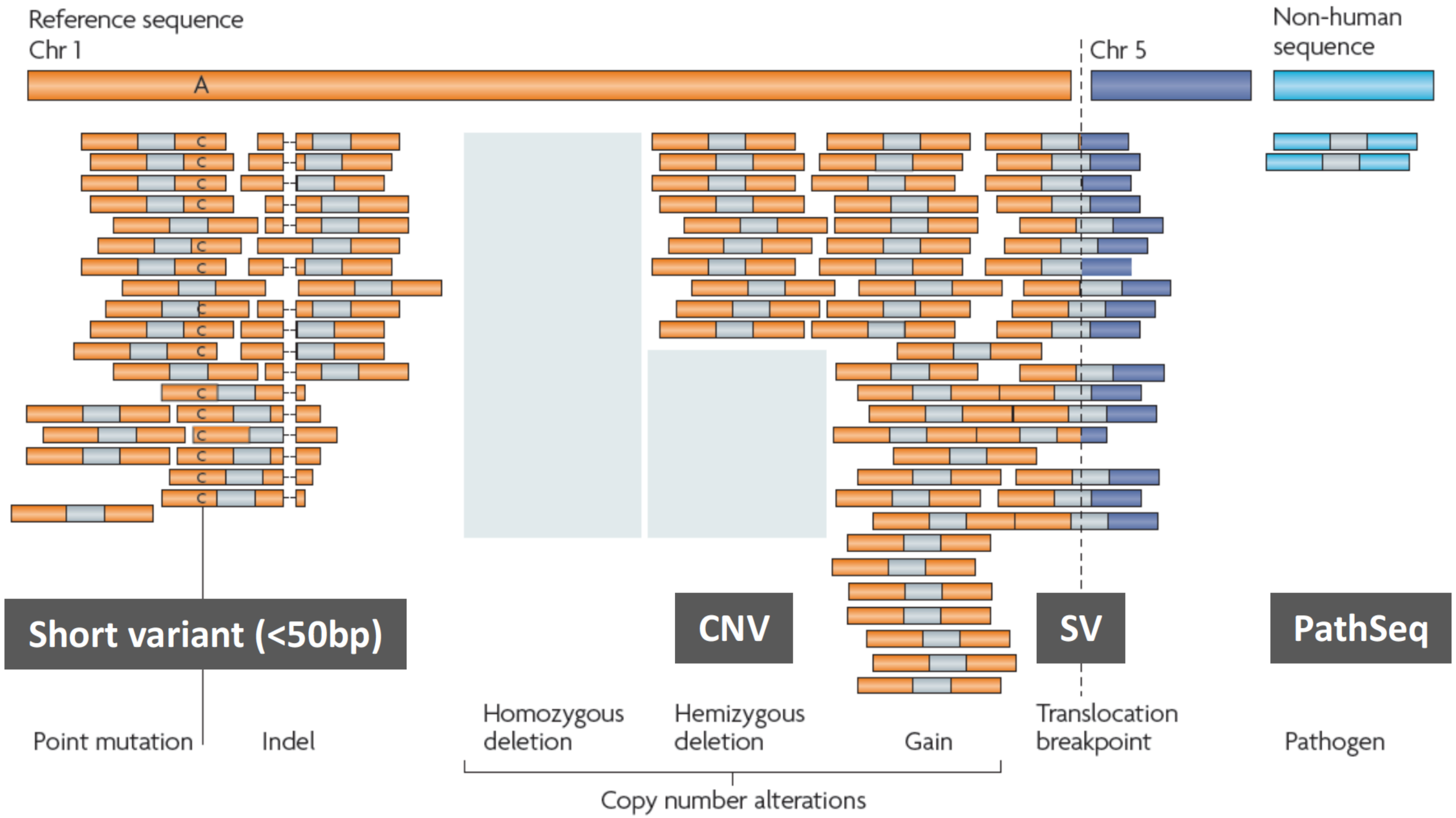
How do we study cancer?

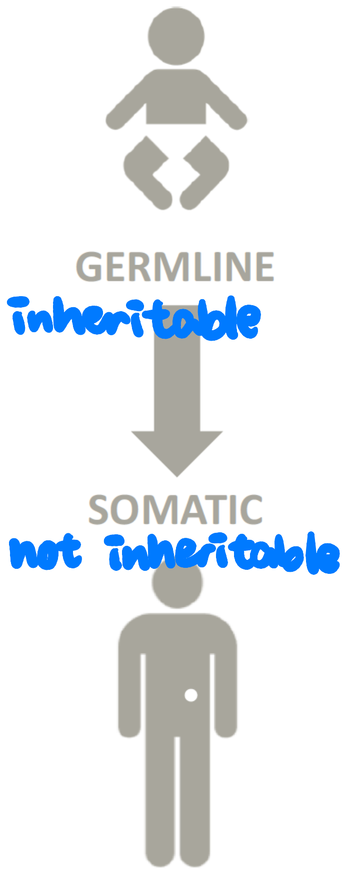
- Cancer is usually believed to be a genomic disease

- So, we will use genomics/multi-omics methods to study it

- Genome/Epigenome/Transcriptome/Proteome/Metabolome

Different types of genomic variants

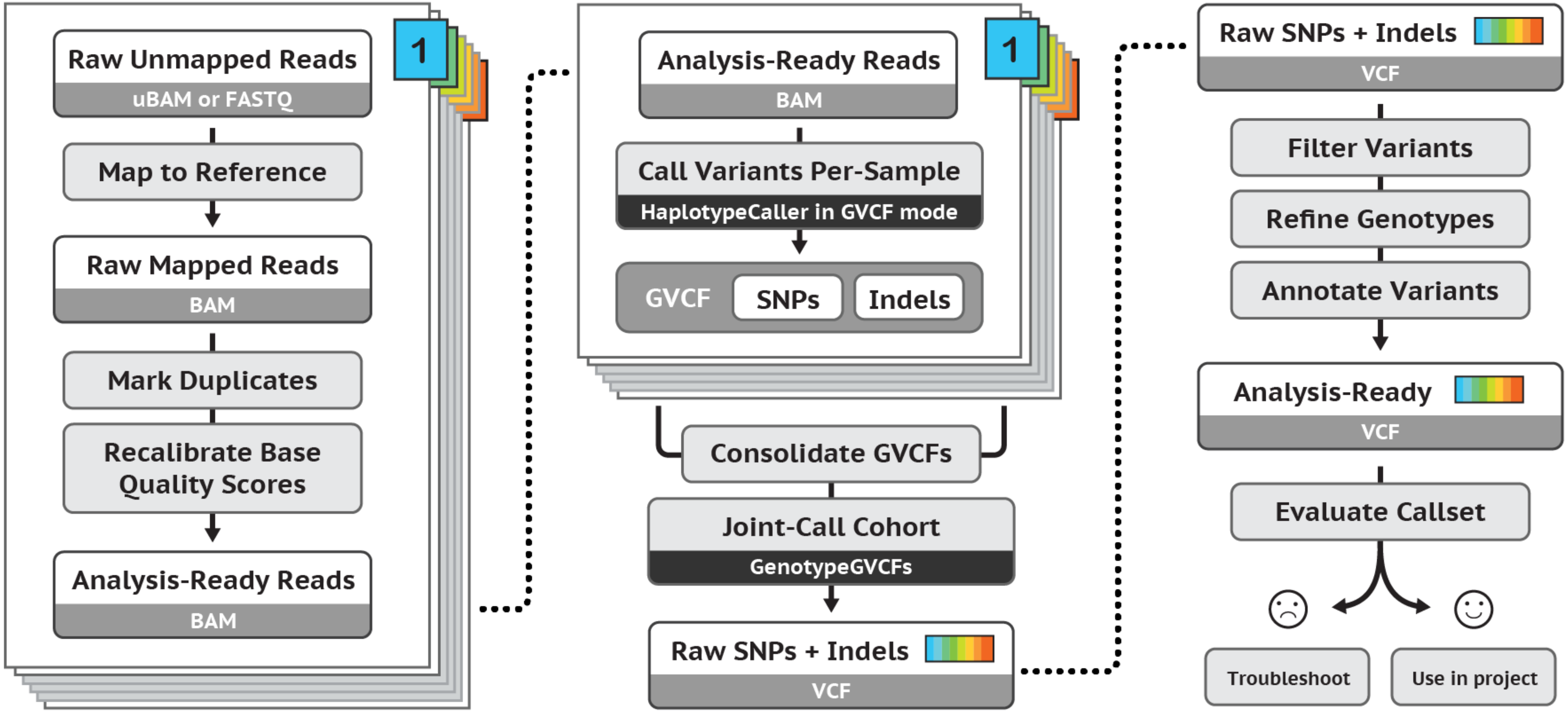




inheritable

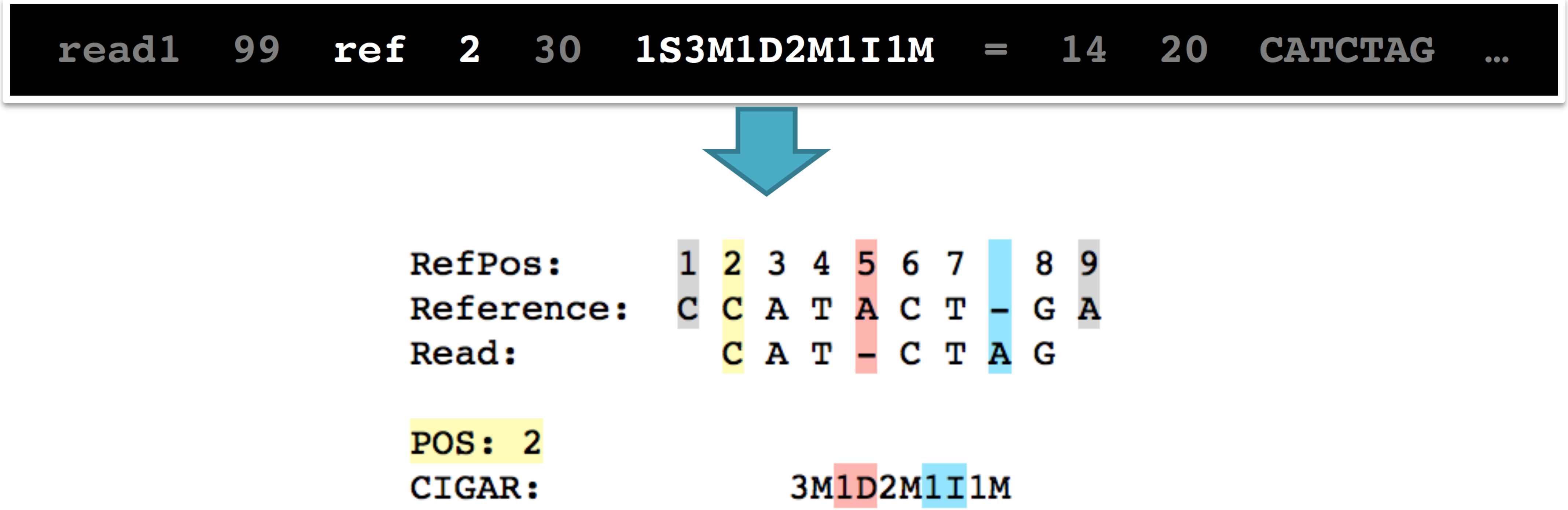
not inheritable

Variant calling in more detail



CIGAR summarizes alignment structure

CIGAR = Concise Idiosyncratic Gapped Alignment Report



```
read1   99   ref   2   30   1S3M1D2M1I1M   =   14   20   CATCTAG   …
```

```
RefPos:      1 2 3 4 5 6 7   8 9
Reference:   C C A T A C T - G A
Read:          C A T - C T A G

POS: 2
CIGAR:             3M1D2M1I1M
```

What you are expected to know from this part

- The reasons that we need to do the steps
  - For example, why we would like to remove the duplicates
- The ability to read the records in those files
  - Given an alignment, you should be able to convert it into a CIGAR string
  - Given a VCF record, you should know what has been changed

    ↘ mutation
- How different factors affect the quality of the mapping and the variant calling
  - Errors VS variants
  - Duplicates
  - Depth/coverage
  - Sequence quality

Mutation ≠ Cancer

Bonferroni correction

- Adjusted p-value = p-value/number of tests

- Suppose we have 1 million SNPs to test
  - Adjusted p-value = $\frac{0.05}{1{,}000{,}000}$
  - Adjusted p-value = $5 * 10^{-8}$

Decrease Type I error rates (FP)

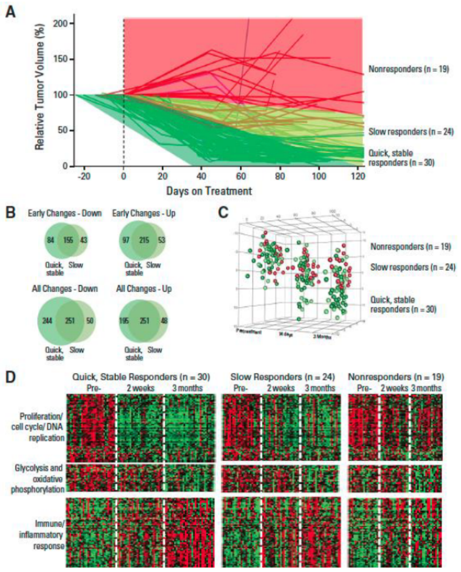
Increase Type II error rates (FN)

Today's agenda

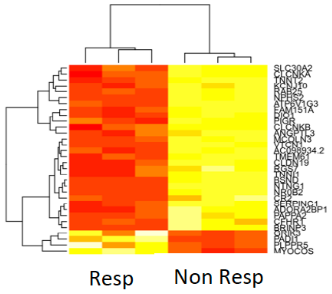
- RNA-seq
  - Gene fusion---structural variant
- Epigenome
  - Peak calling

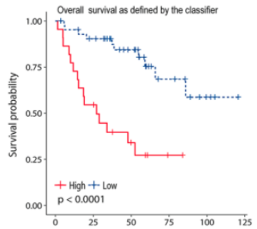
RNA-seq data analysis

Exploration → Choose informative genes → Choose and tune classification algorithm → Validation → Implementation



Whole genome: RNA-seq





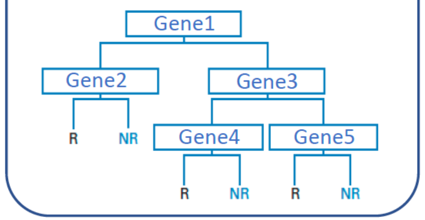
Differential Gene Expression

Multiple testing

(Bonferroni, FDR)

Linear scores = 3.5*Gene1 - 4.8*Gene2 + 5.0*Gene3 + 10.4*Gene4 ...

Decision trees



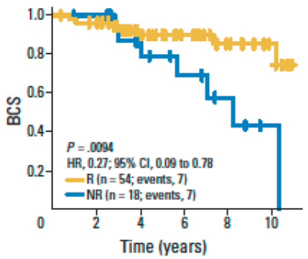
AI algorithms: SVM, RF, ANN, ...

Algorithm training

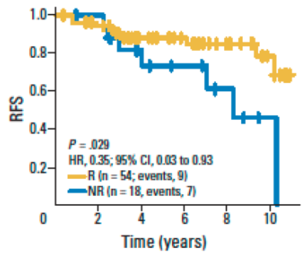
Overfitting

(Cross-Validation)

Training set



Validation set(s)



Independent dataset(s)

NICE

FDA

...

RT-PCR

NanoString

RNA-seq ...

Recall one question

- ❖What if there are two same mappings of the short reads to the genome sequence? how can we decide which section of the genome should it map to?

Genome	T	A	A	T	G	C	C	A	T	G	G	A	T	G
RNA-seq		C	C	A										
		2	3											

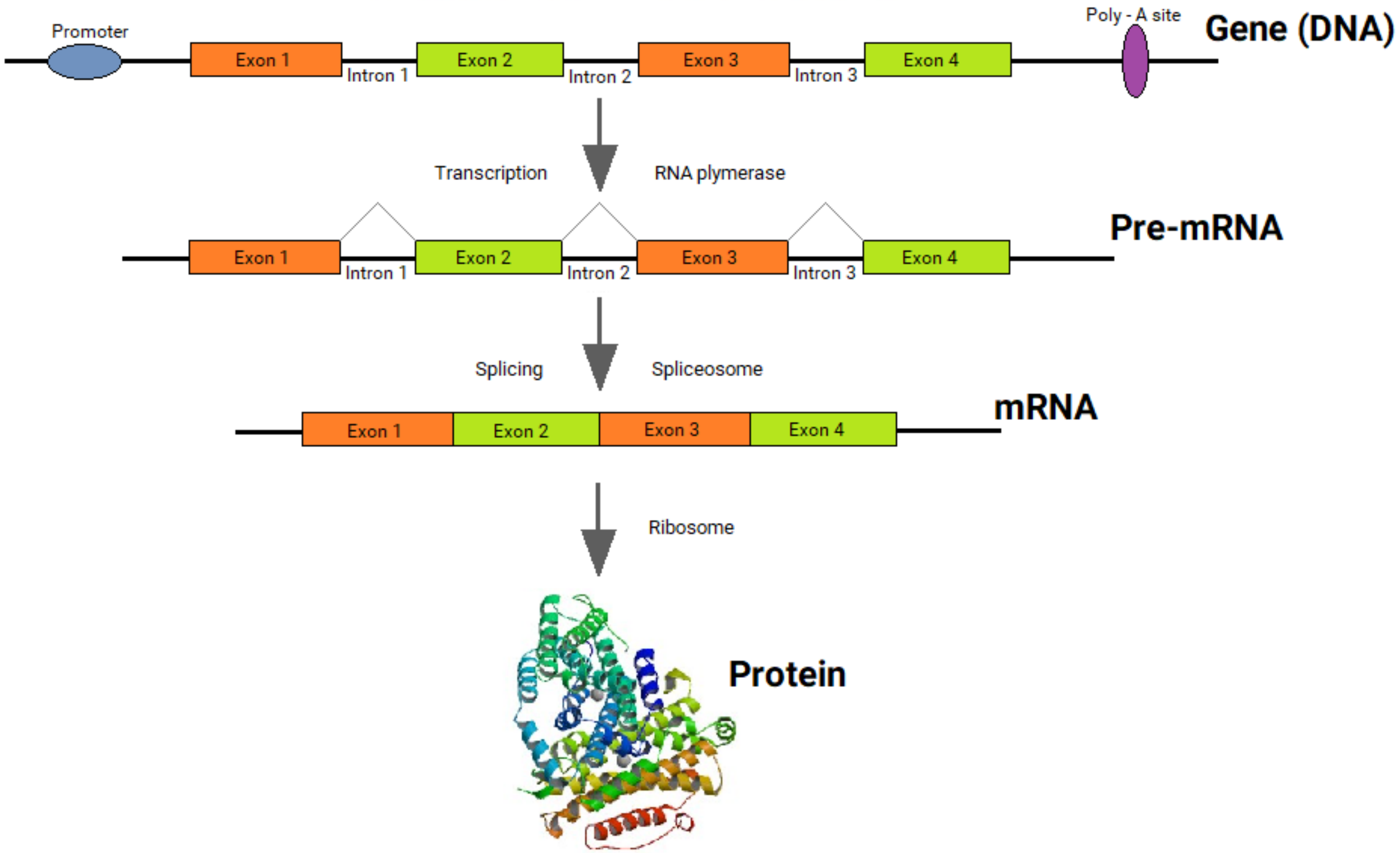
Long reads

Genome	T	A	A	T	G	C	C	A	T	G	G	C	C	A
RNA-seq		C	C	A										
		2	3											

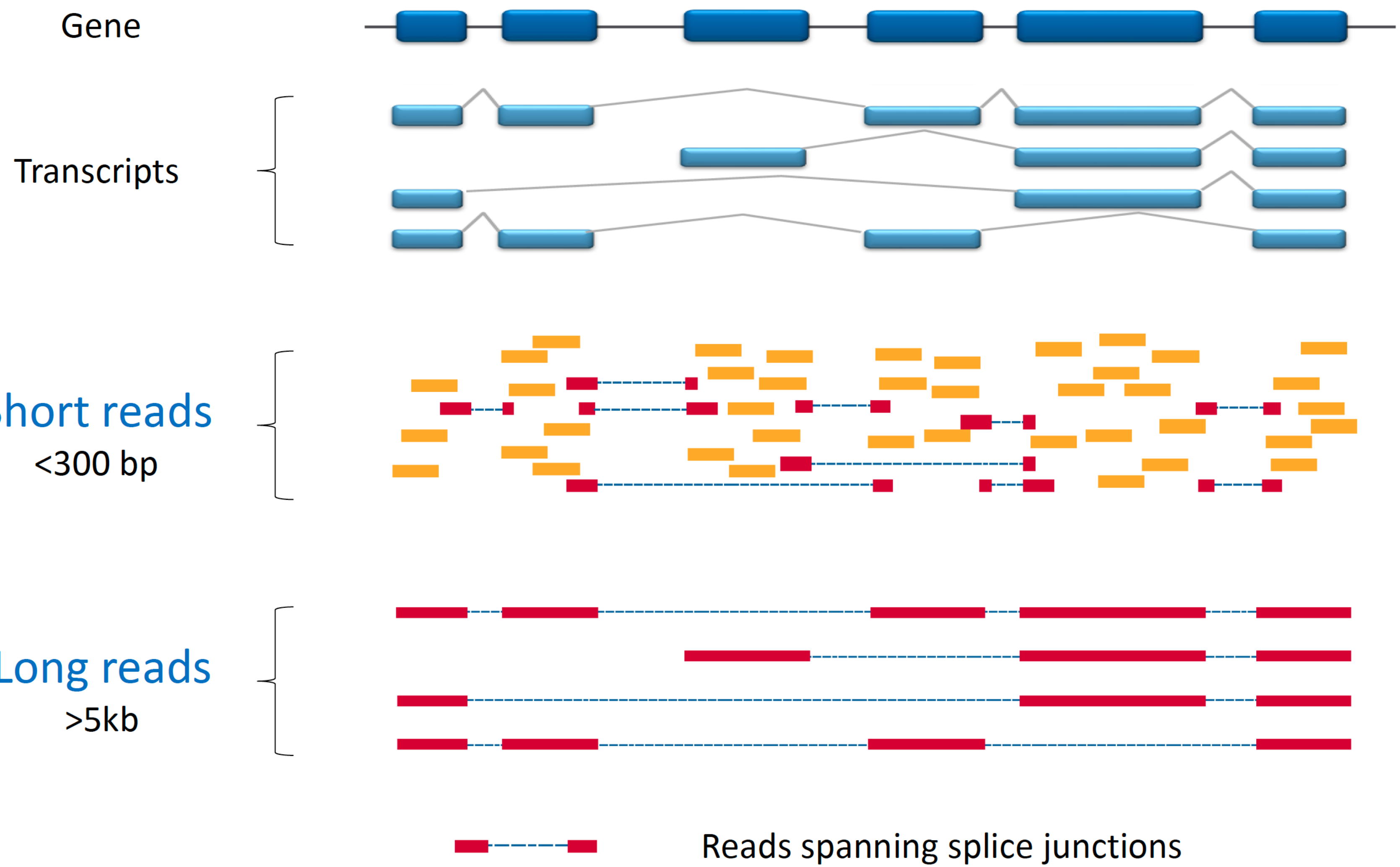
Statistical probability

Genomic context

Transcription, splicing and translation of a eukaryotic gene



Mapping spanning splice junctions

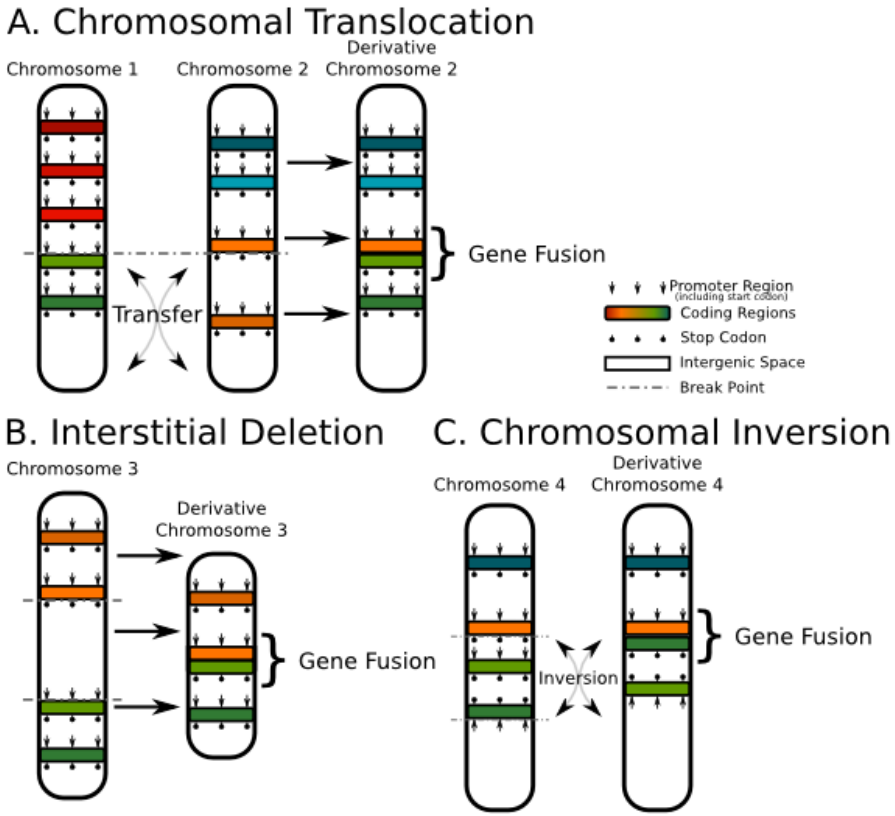


The mapping algorithm should be modified slightly. But it's helpful for identifying gene fusion.

Map part of sequence

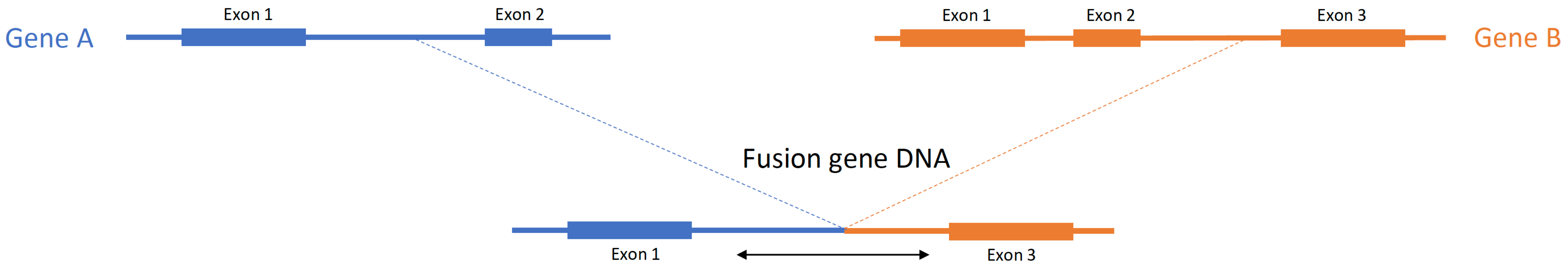
What is gene fusion?

- ❖The first fusion gene was described in cancer cells in the early 1980s
- ❖Novel gene formed by fusion of two distinct wild type genes
- ❖In cancer: produced by somatic genome rearrangements

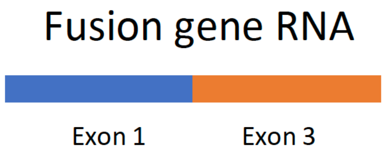


Gene fusion is a specific kind of structural variant related to cancer

RNA-seq for gene fusion detection

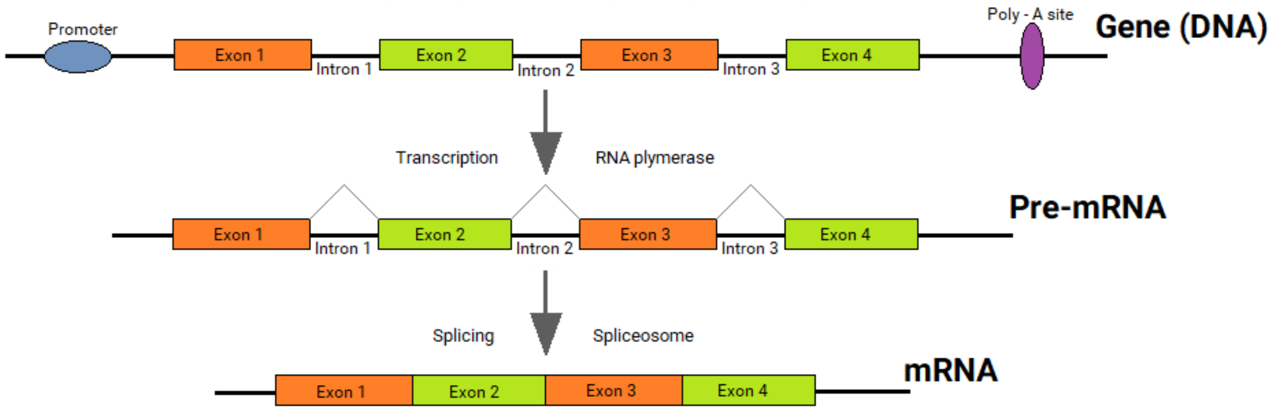


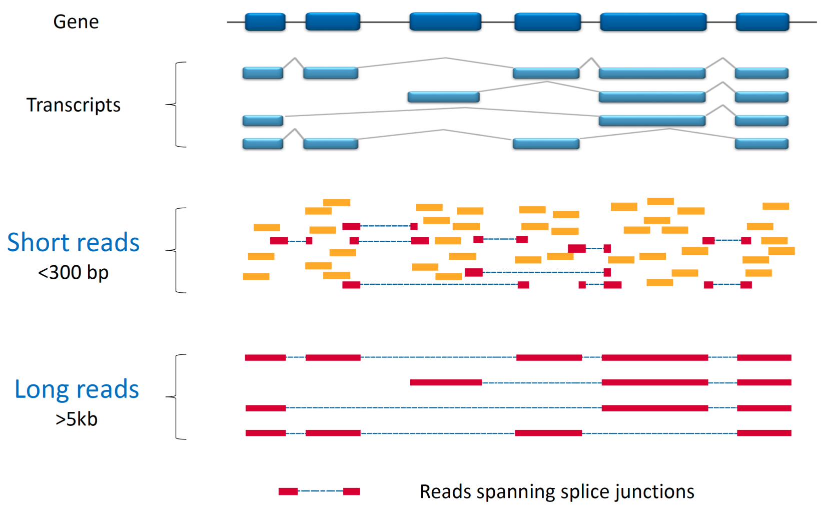
Break-points are in introns
We need whole genome sequencing
Whole exome sequencing is not enough

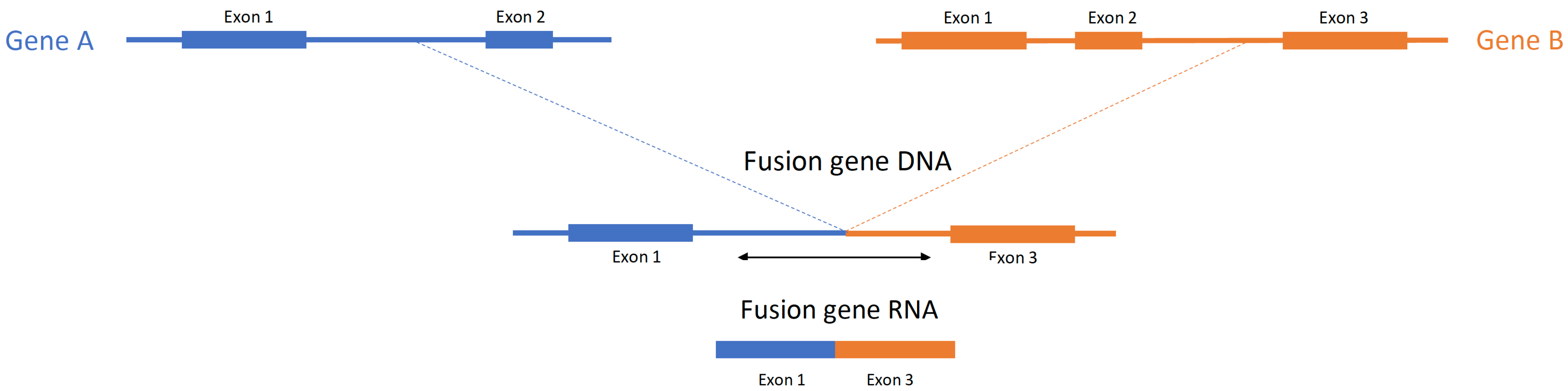


Detecting fusion in RNA-seq requires much less sequencing than WGS, especially with long reads

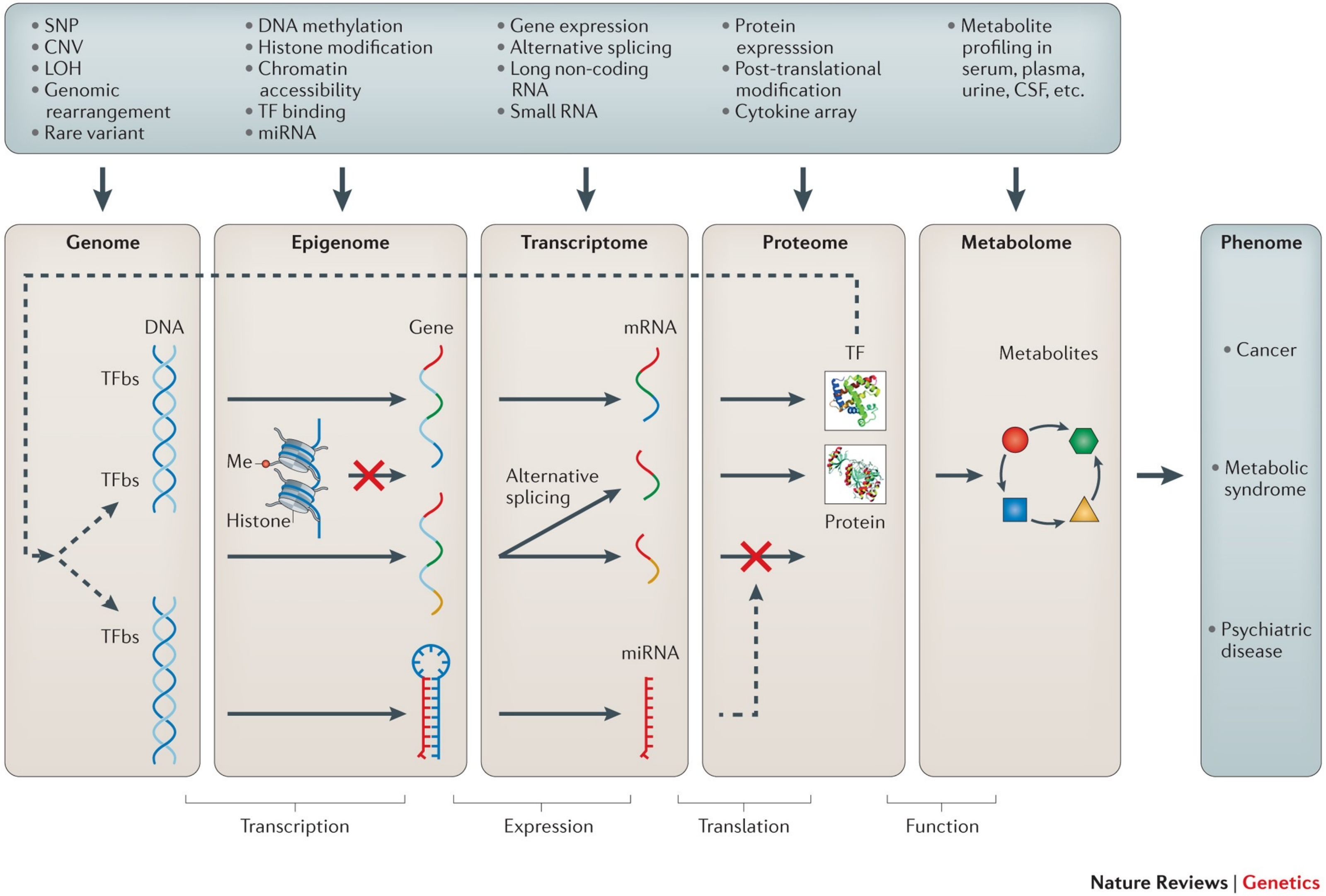
Why can it be detected by RNA-seq?







People study cancer at multiple levels



- ➢ Genetic variants
  - Genome
  - Gene fusion (RNA-seq)
- ➢ Abnormal gene expression
  - Genome (genetic information)
  - Epigenome (environment)
  - Transcriptome (direct measurement)

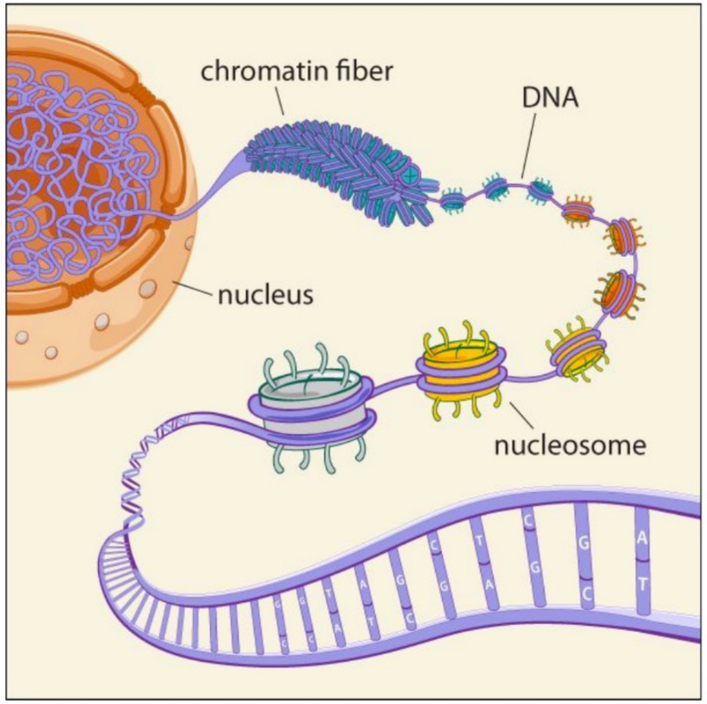
Ritchie et al., 2015, Nature Review Genetics

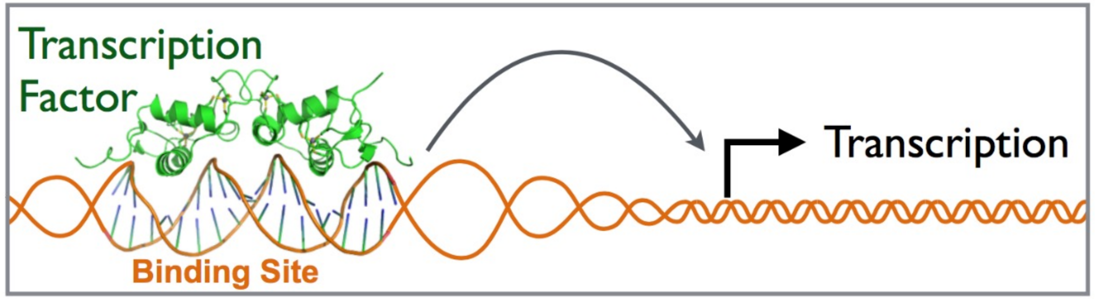
Today’s agenda

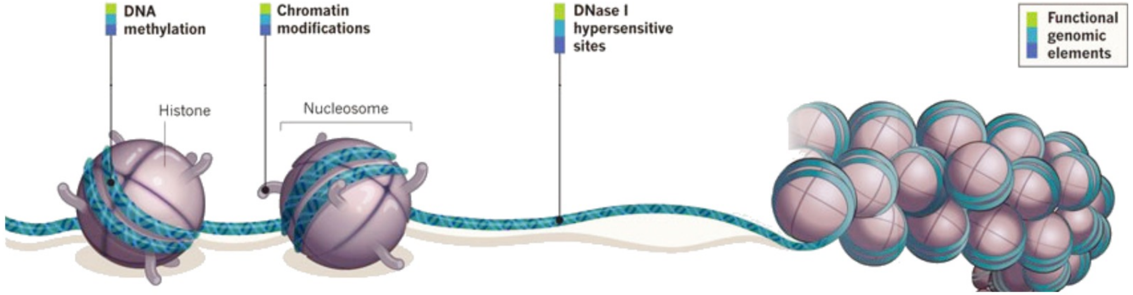
- RNA-seq
  - Gene fusion---structural variant

- Epigenome
  - Peak calling

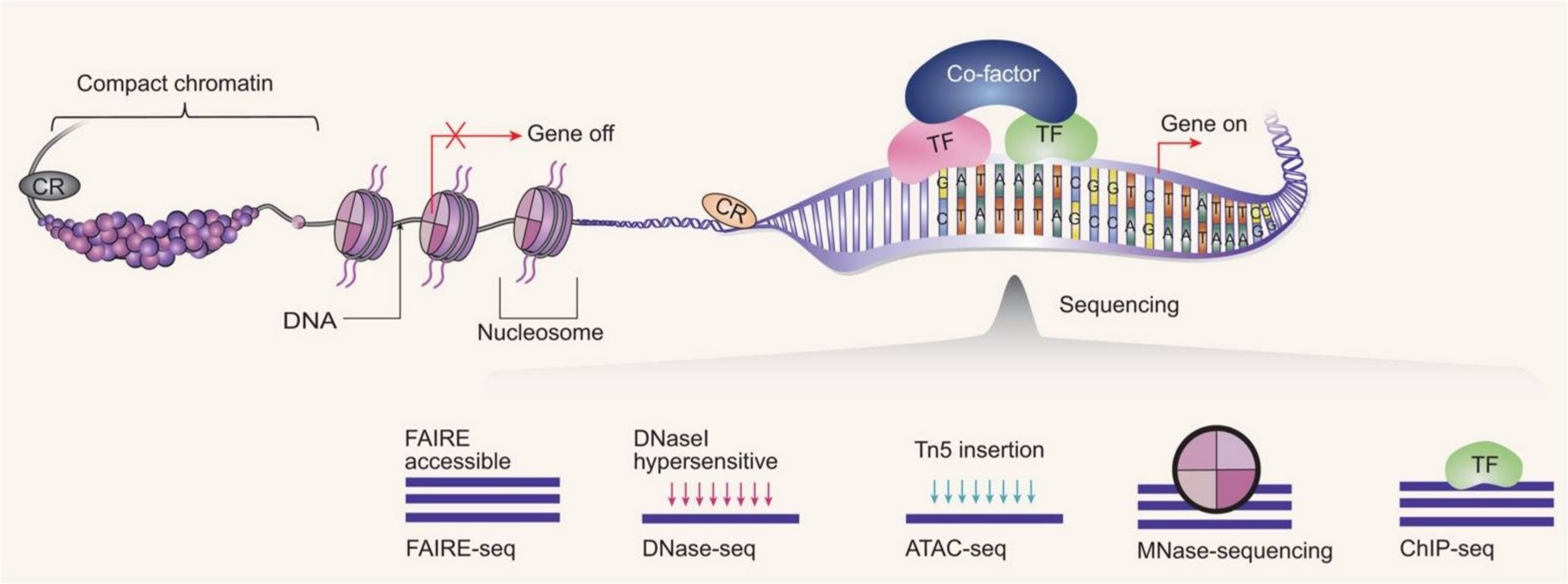
Epigenomics





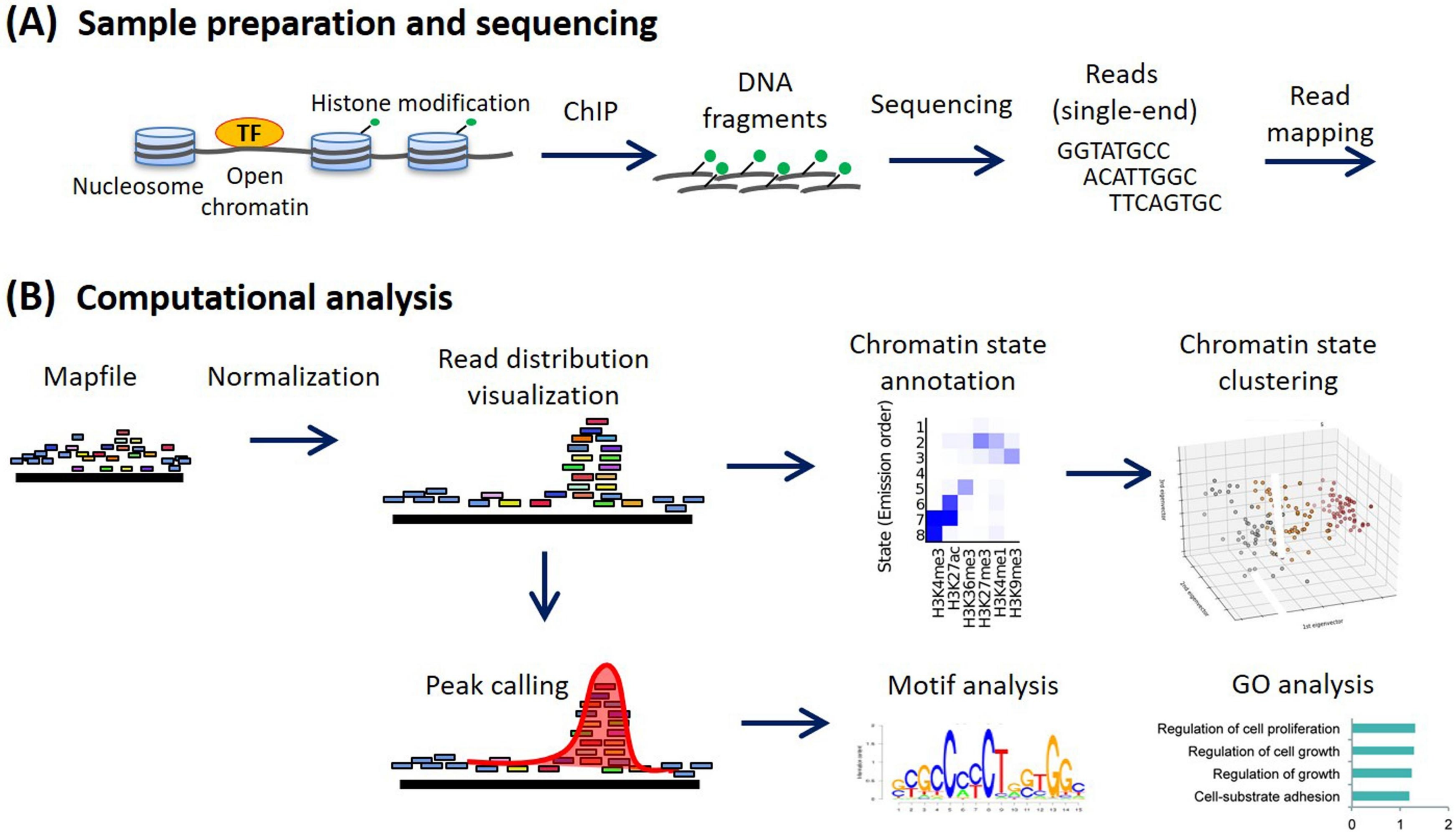


Sequencing protocols



Wang et. al., 2016, Natl Sci Rev

The overall data analytics pipeline for epigenomics



(A) Sample preparation and sequencing

Nucleosome → TF → Open chromatin → Histone modification → ChIP → DNA fragments → Sequencing → Reads (single-end) → Read mapping

GGTATGCC
ACATTGGC
TTCAGTGC

(B) Computational analysis

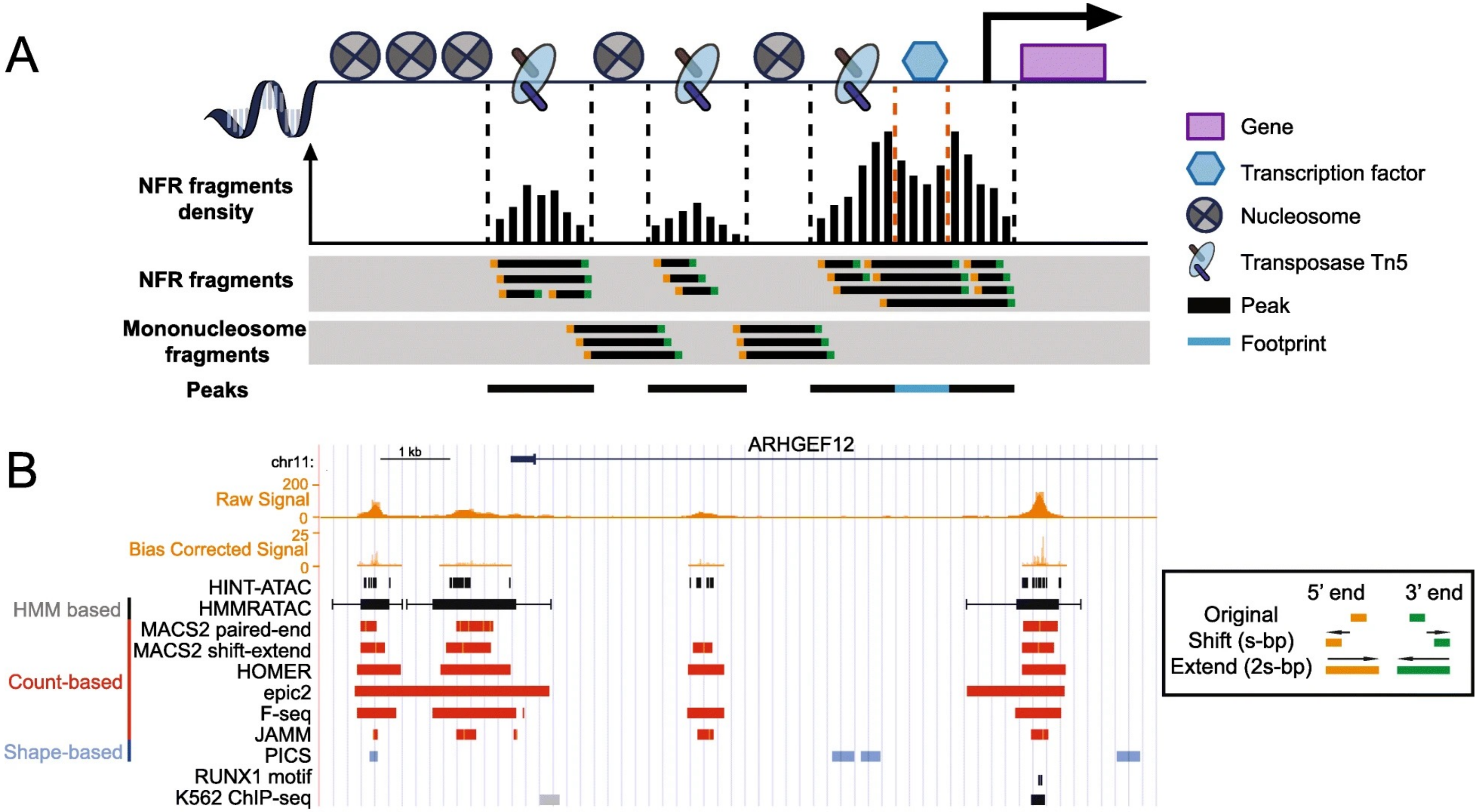
Mapfile → Normalization → Read distribution visualization → Chromatin state annotation → Chromatin state clustering

Read distribution visualization → Peak calling → Motif analysis

GO analysis

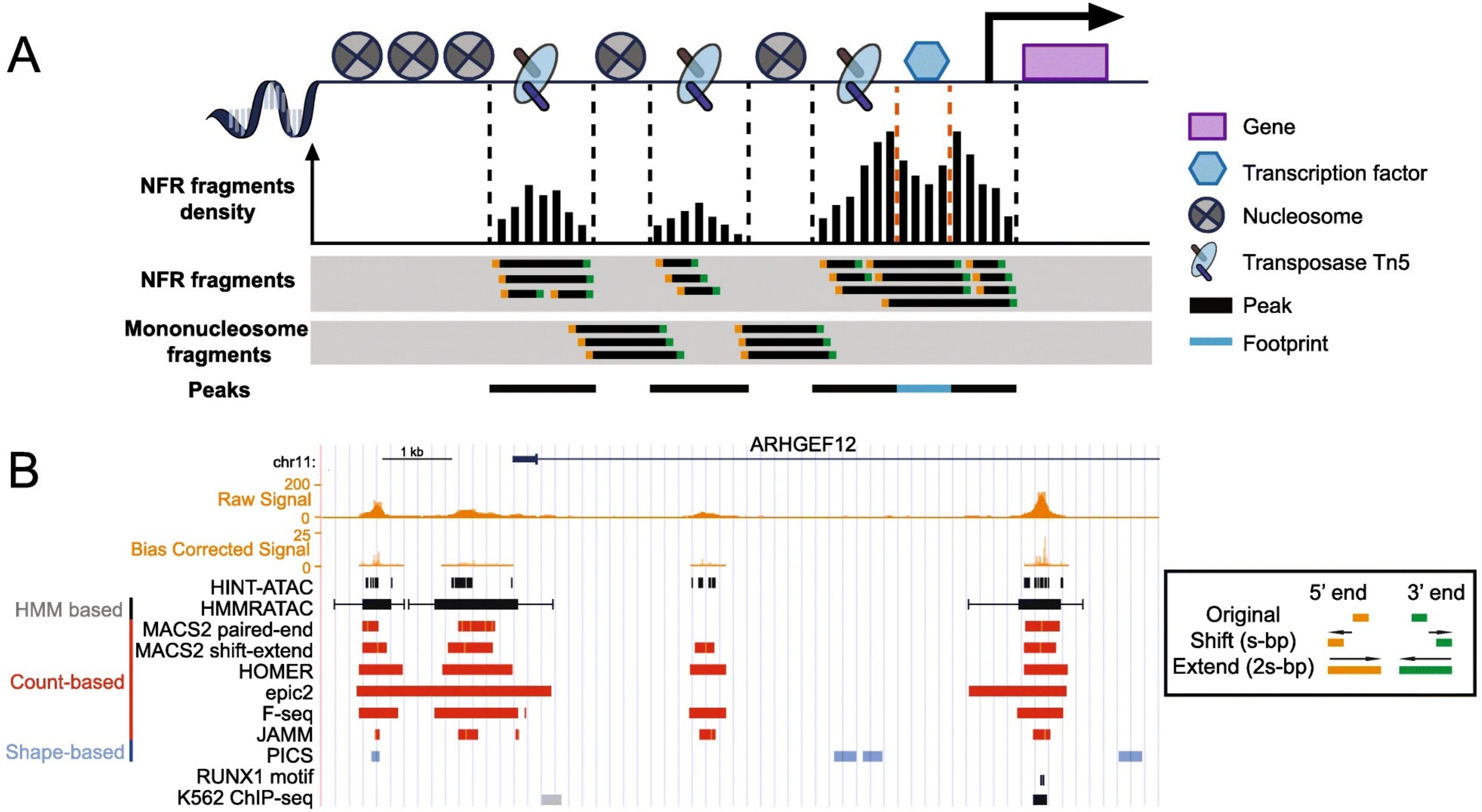
Nakato et. al., 2021, Methods

Peak calling



Yan et. al., 2020, Genome Biology

Peak calling

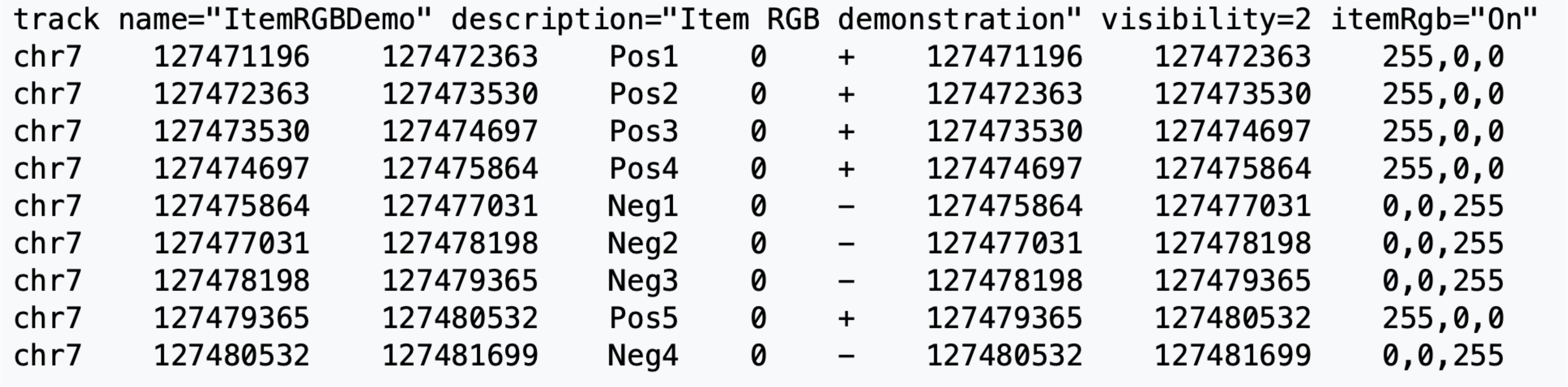


Statistical testing: Peak shape VS random background

Yan et. al., 2020, Genome Biology

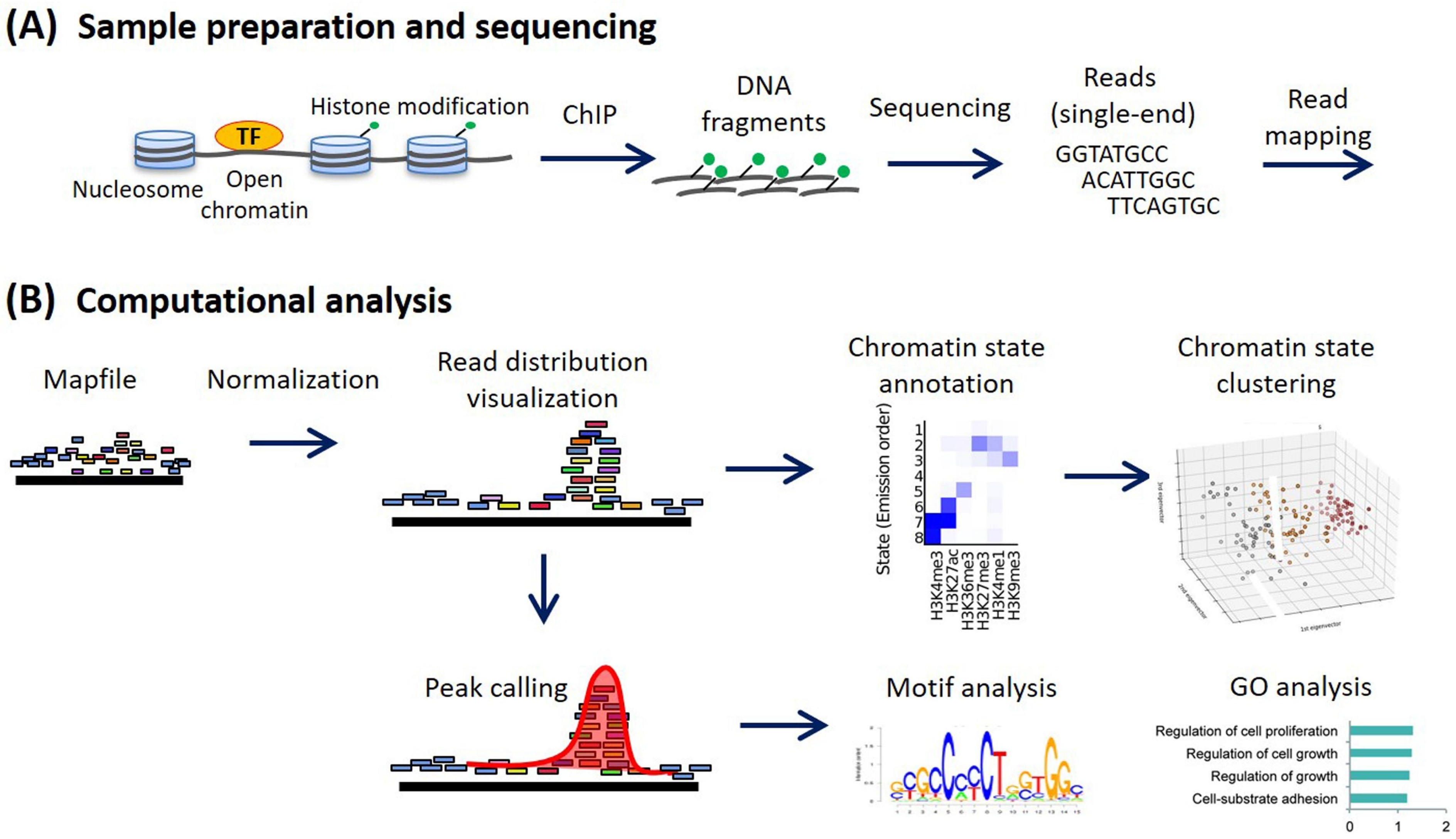
Peak calling output-BED file

- Browser Extensible Data (BED) format
  - Chromosome
  - Start
  - End
  - Label
  - …



```
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7    127471196    127472363    Pos1    0    +    127471196    127472363    255,0,0
chr7    127472363    127473530    Pos2    0    +    127472363    127473530    255,0,0
chr7    127473530    127474697    Pos3    0    +    127473530    127474697    255,0,0
chr7    127474697    127475864    Pos4    0    +    127474697    127475864    255,0,0
chr7    127475864    127477031    Neg1    0    -    127475864    127477031    0,0,255
chr7    127477031    127478198    Neg2    0    -    127477031    127478198    0,0,255
chr7    127478198    127479365    Neg3    0    -    127478198    127479365    0,0,255
chr7    127479365    127480532    Pos5    0    +    127479365    127480532    255,0,0
chr7    127480532    127481699    Neg4    0    -    127480532    127481699    0,0,255
```

The overall data analytics pipeline for epigenomics



(A) Sample preparation and sequencing

Histone modification

TF

Nucleosome

Open chromatin

ChIP

DNA fragments

Sequencing

Reads (single-end)

GGTATGCC
ACATTGGC
TTCAGTGC

Read mapping

(B) Computational analysis

Mapfile

Normalization

Read distribution visualization

Peak calling

Chromatin state annotation

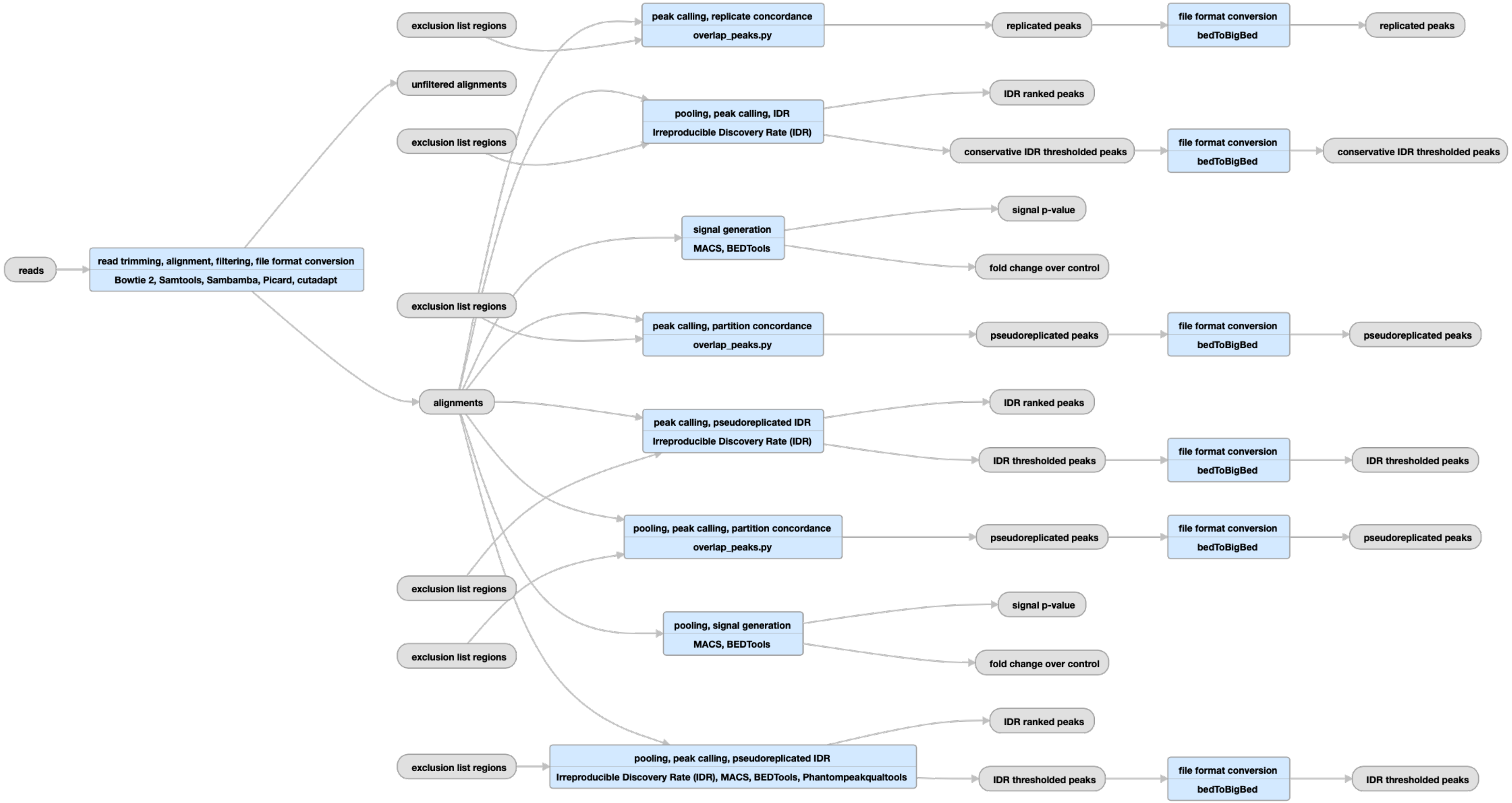
Chromatin state clustering

Motif analysis

GO analysis

Nakato et. al., 2021, Methods

The entire detailed pipeline (ATAC-seq as an example)



reads

read trimming, alignment, filtering, file format conversion
Bowtie 2, Samtools, Sambamba, Picard, cutadapt

unfiltered alignments

alignments

exclusion list regions

peak calling, replicate concordance
overlap_peaks.py

replicated peaks

file format conversion
bedToBigBed

replicated peaks

pooling, peak calling, IDR
Irreproducible Discovery Rate (IDR)

IDR ranked peaks

conservative IDR thresholded peaks

file format conversion
bedToBigBed

conservative IDR thresholded peaks

signal generation
MACS, BEDTools

signal p-value

fold change over control

peak calling, partition concordance
overlap_peaks.py

pseudoreplicated peaks

file format conversion
bedToBigBed

pseudoreplicated peaks

peak calling, pseudoreplicated IDR
Irreproducible Discovery Rate (IDR)

IDR ranked peaks

IDR thresholded peaks

file format conversion
bedToBigBed

IDR thresholded peaks

pooling, peak calling, partition concordance
overlap_peaks.py

pseudoreplicated peaks

file format conversion
bedToBigBed

pseudoreplicated peaks

pooling, signal generation
MACS, BEDTools

signal p-value

fold change over control

pooling, peak calling, pseudoreplicated IDR
Irreproducible Discovery Rate (IDR), MACS, BEDTools, Phantompeakqualtools

IDR ranked peaks

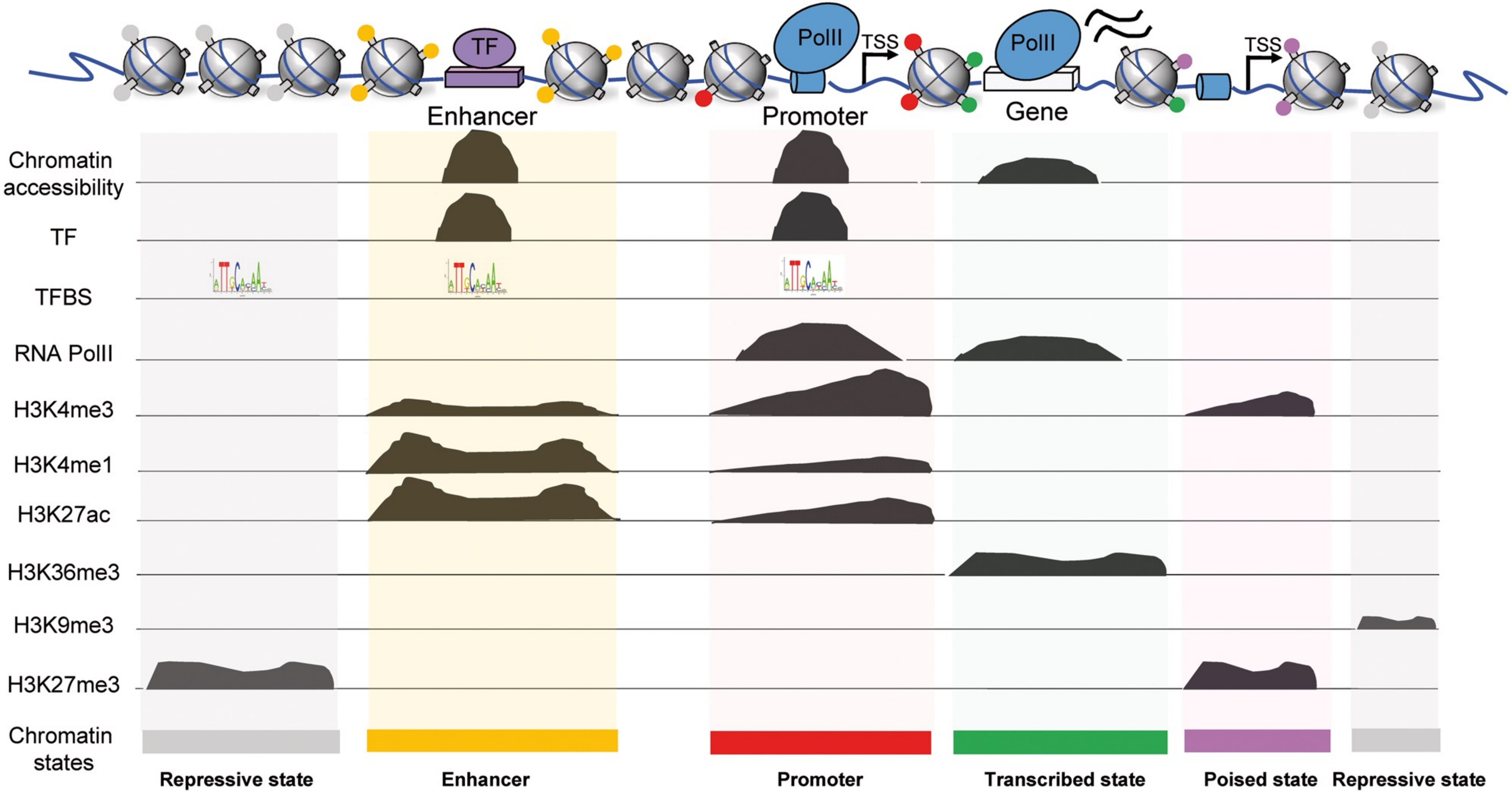
IDR thresholded peaks

file format conversion
bedToBigBed

IDR thresholded peaks

https://www.encodeproject.org/pipelines/ENCPL787FUN/

Histone marks and chromatin accessibility



To make you awake

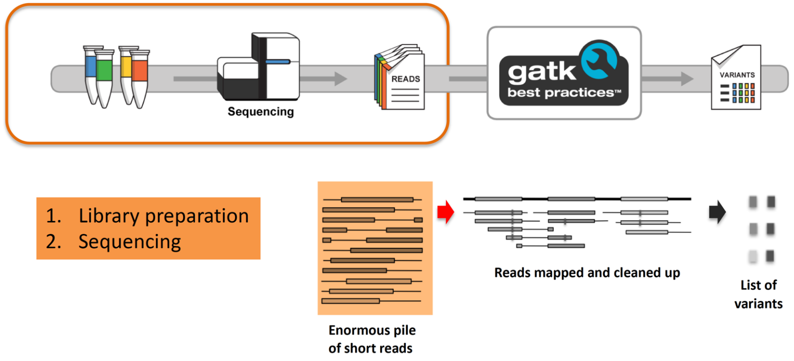
https://ureply.mobi/teacher

Take-home message

- Variant calling pipeline
  - Reasons for the steps
  - File interpretation
  - Factors affect variant calling
- GWAS
  - P-value correction
- Gene fusion
  - Definition
  - RNA-seq can detect it
- Epigenomics
  - Gene expression regulation: structure and environment
  - Data analytics pipeline

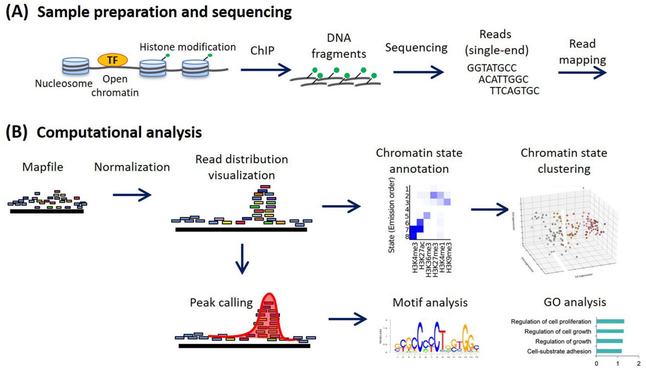
Potential projects-4,5,6

❖4. Genetic variant calling pipeline

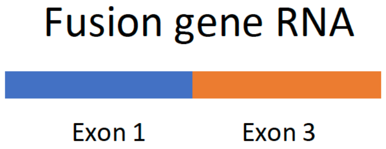


1. Library preparation
2. Sequencing

❖5. Epigenetic data processing pipeline



❖6. Gene fusion detection pipeline



Resources

- https://www.ebi.ac.uk/training/materials/cancer-genomics-materials/

- GATK workshop slides: https://drive.google.com/drive/folders/1y7q0gJ-ohNDhKG85UTRTwW1Jkq4HJ5M3

- GATK workshop video: https://www.youtube.com/watch?v=sM9cQPWwvn4

- GWAS workshop: https://www.youtube.com/watch?v=xw419NKqMqw

- Epigenetics: https://www.youtube.com/watch?v=IAu44BkOaSs

- https://www.encodeproject.org/atac-seq/

Post-lecture survey

- https://forms.gle/dRgK23XzEfhThDed8

Next time

- Single-cell RNA-seq

Thank you!

Yu LI (李煜)

liyu95.com

liyu@cse.cuhk.edu.hk

The Artificial Intelligence in Healthcare (AIH) Group

Department of Computer Science and Engineering (CSE)

The Chinese University of Hong Kong (CUHK)