institutCurie

# Overview of NGS Applications and Analysis Workflows

Nicolas Servant (nicolas.servant@curie.fr)
DU Séquençage Haut Débit et Maladies Génétiques
Dijon, 16th of October 2013

# WHAT'S HAPPEN AFTER SEQUENCING ?

# NGS : **N**ext-**G**eneration **S**equencing

**Faster, Cheaper, Deeper**

Bringing analysis of sequence information to another level by generating millions of sequences, millions of samples, millions of genomes !

Human Genome : 30 Gbases

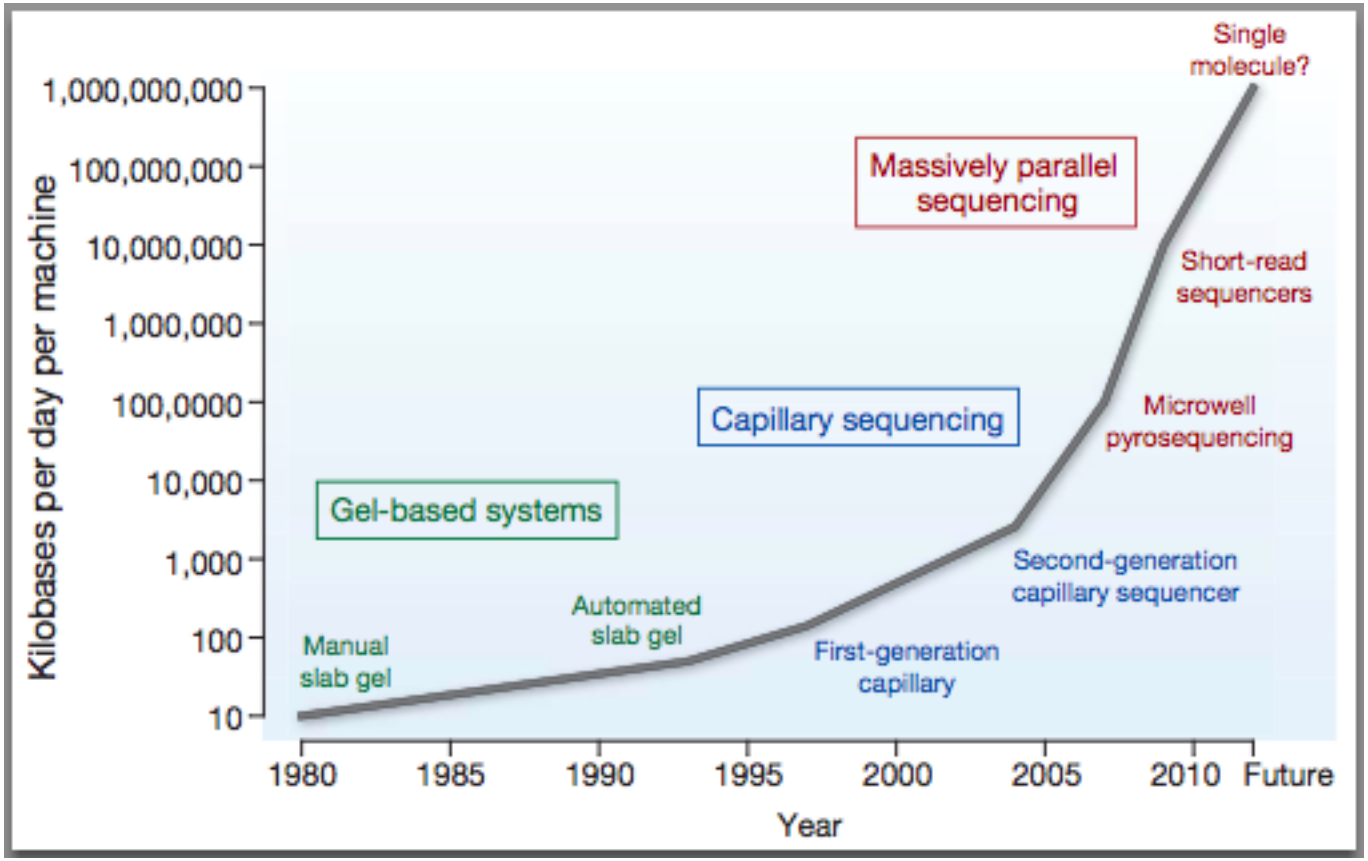
Sanger Sequencing : 1000 bases / run

NGS Sequencing : 100 Gbases / run

# Hi-seq, SOLiD, PGM, What does it mean ?

| Platform | Provider | Reads Number (M) | Max Reads Size (bp) | Throughput (Gb) | Time | Space |
|---|---|---|---|---|---|---|
| **GS Flex** | Roche | 1 | 700 | 0.7 | 8d | Base |
| **Hiseq 2000/2500 Normal mode** | Illumina | 3000 | 2x100 | 600 | 11d | Base |
| **Hiseq 2500 rapid mode** | Illumina | 600 | 2x150 | 120 | 40h | Base |
| **MiSeq** | Illumina | 15 | 2x250 | 8.5 | 40h | Base |
| **SOLiD** | LifeTech | 1400 | 75-35 | 150 | 25d | Color |
| **PGM 314** | Ion Torrent | 0.5 | 400 | >0.01 | 2-4h | Base |
| **PGM 316** | Ion Torrent | 2 | 400 | >0.1 | 2-4h | Base |
| **PGM 318** | Ion Torrent | 4 | 400 | >1 | 2-4h | Base |
| **Proton** | Ion Torrent | 60-80 M | 200 | >12 | 2-4h | Base |

# Today's sequencing machines produce a massive amount of data

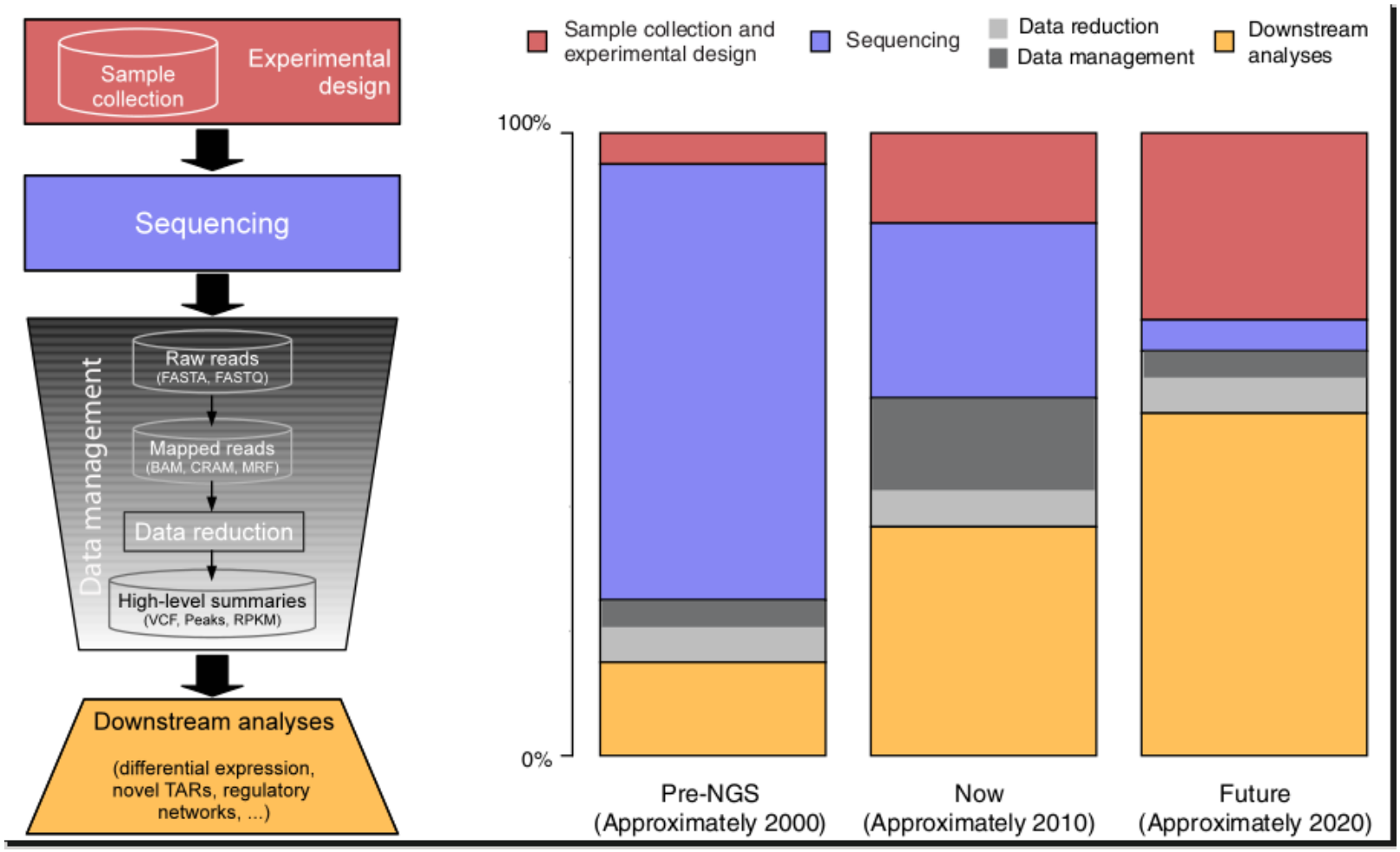



Improvements in the rate of DNA sequencing over the past 30 years and into the future

Stratton, MR. *et al. Nature* **458**, 719-724 (2009)

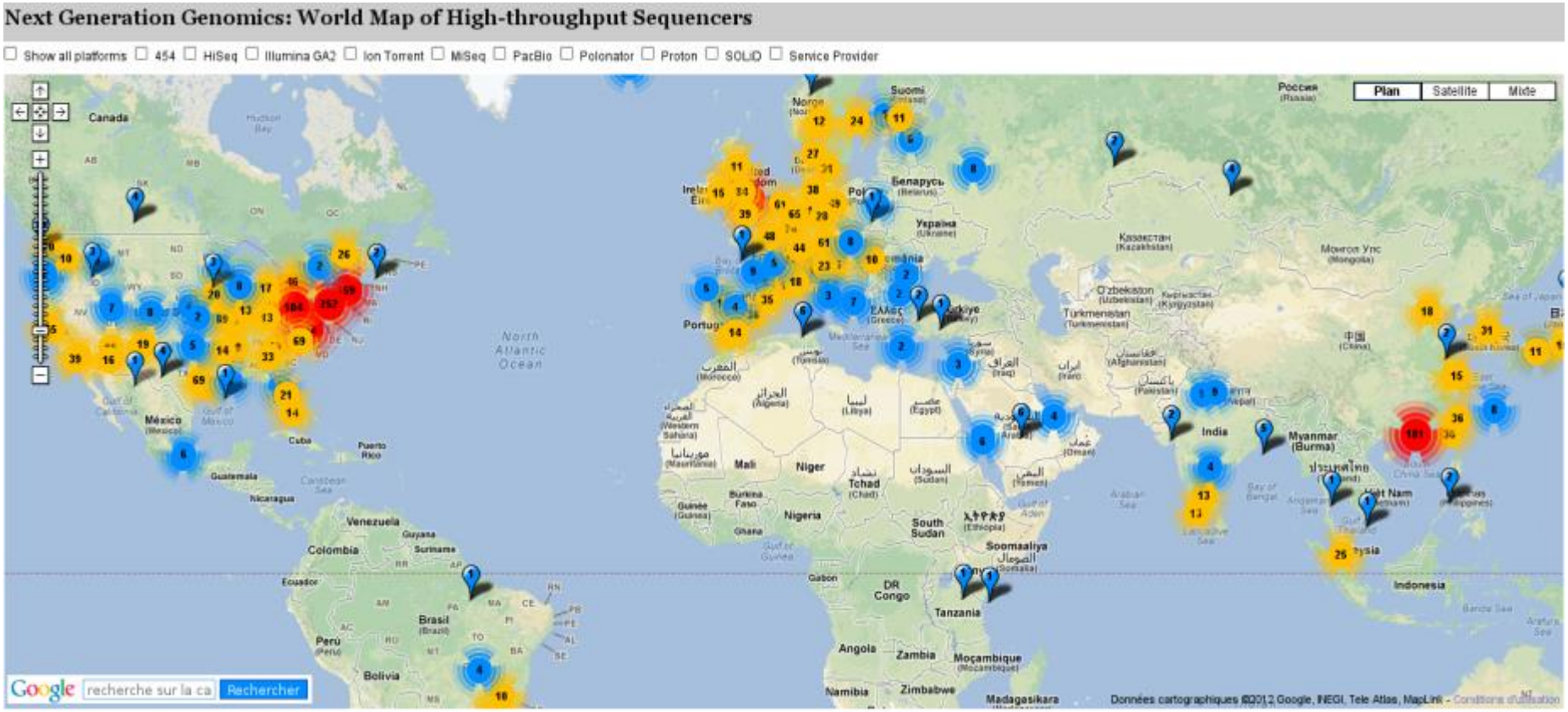
From R. Daveau

# Evolution of sequencing cost




Sboner et al. 2011. Genome Biology, **12**:125

# Sequencing, sequencing, sequencing




http://omicsmaps.com/

# NGS and (Bio)Informatics

« The rule of thumb in the genomics community is that every dollar spent on sequencing hardware must be matched by a comparable investment in informatics »

http://www.the-scientist.com/?articles.view/articleNo/30731/title/Sequence-Analysis-101/

# How does next-gen sequencing work ?

DNA fragmentation

Size selection and clonal amplification

Massive parallel sequencing

ACCGTTTGCCG… $10^6$-$10^9$ short reads

# Different libraries for different applications

## Single-end sequencing

Length of end sequences: depends on the platform

- **ChIP-seq, sRNA-seq**

## Paired-end sequencing

Sequence both ends of DNA fragment

Insert size: <800nt

Length of end sequences: depends on the platform

- **RNA-seq, Exome-seq, DNA-seq**

# Different libraries for different applications

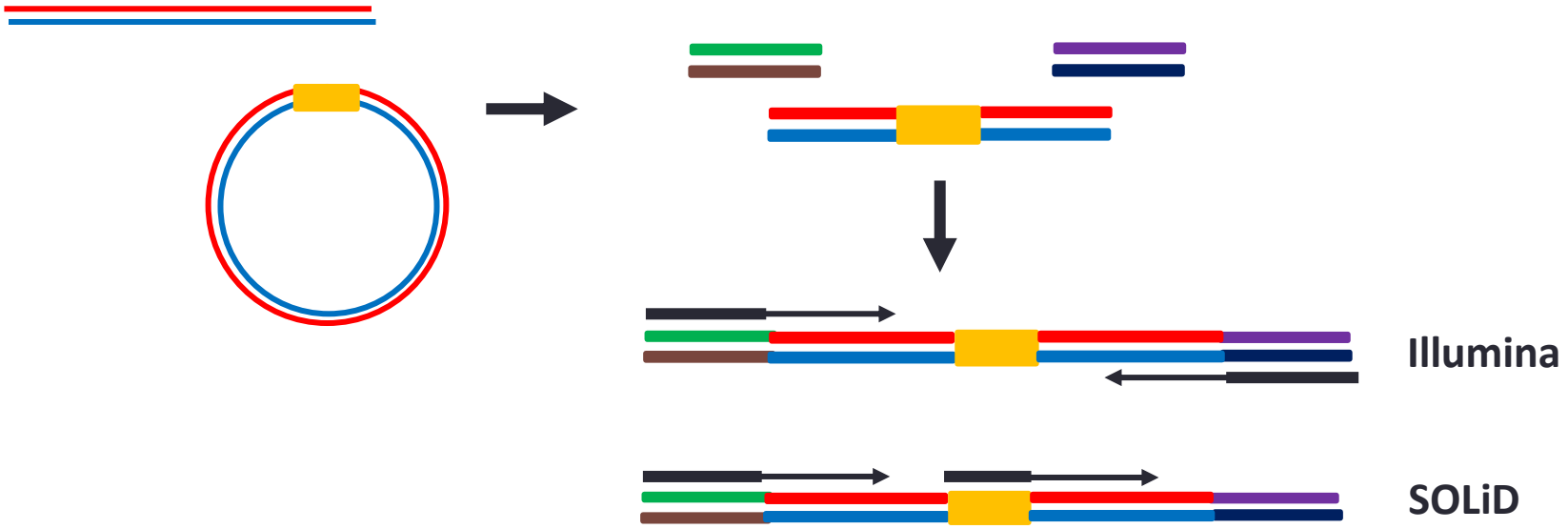
## Mate pair sequencing

Sequence both ends of DNA fragment

Insert size: >500nt

Length of end sequences: depends on the platform

- **Detection of large structural variations**



# FASTQ Format : raw unaligned reads

- Extension from traditional FASTA format
- Each block has 4 elements (in 4 lines):
  1. Sequence name (read name, group, etc…)
  2. Sequence
  3. + (optional: sequence name again)
  4. Associated quality scores (phred-scaled)
- Example record:

```
@FCD19MJACXX:2:1101:1735:1993#GTTCGACA/1
NGAGGCTGAGGCGGGCAGAGGTCAGGAGATCGAGACCATC
+
BP\ccccc]ceecheheeZbe_cZbd_dbbdd\aXab_`b
```

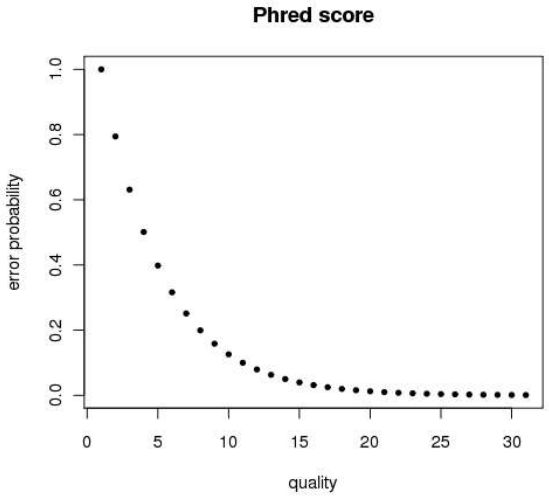
# Sequence quality encoding

- The base calling (A, T, G or C) is performed based on Phred Scores.

| Phred Quality Score | Probability of Incorrect Base Call | Base Call Accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |

Error rate

-> 1%
-> 0.1%
-> 0.01%

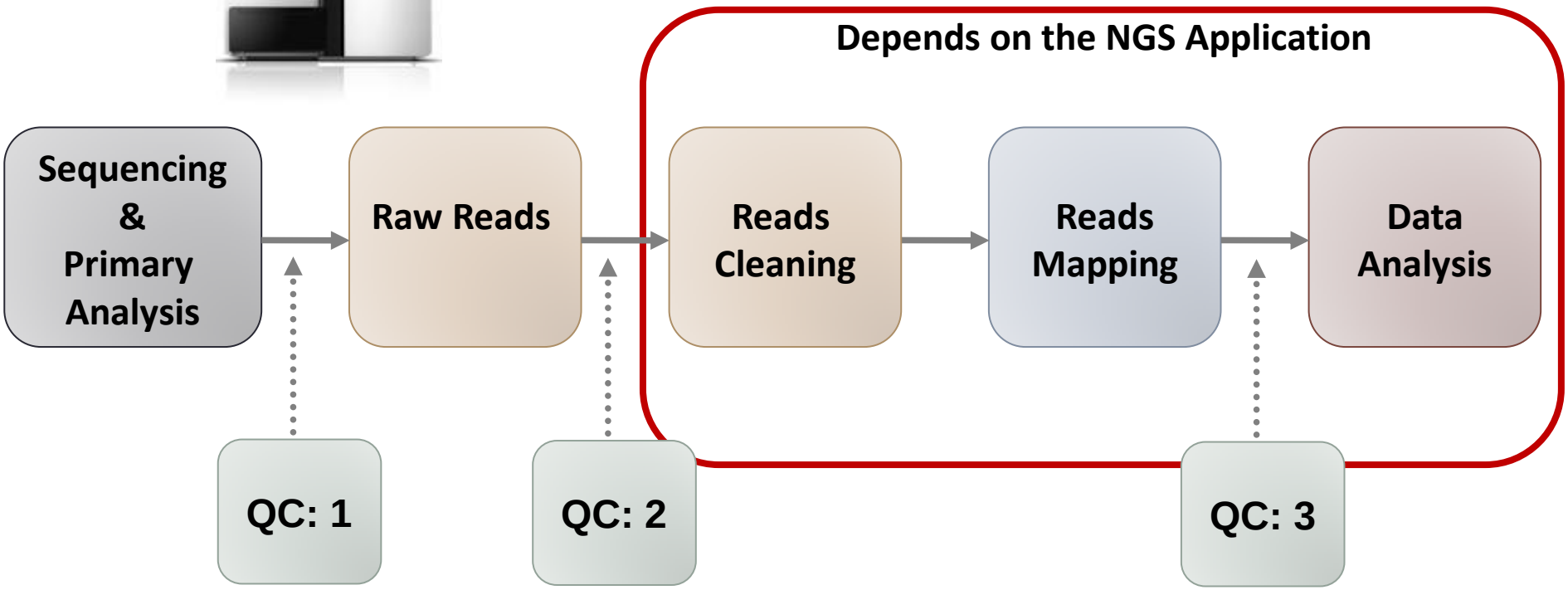



- Phred scores provide log(10)-transformed error probability values:

If p is probability that the base call is wrong the Phred score is

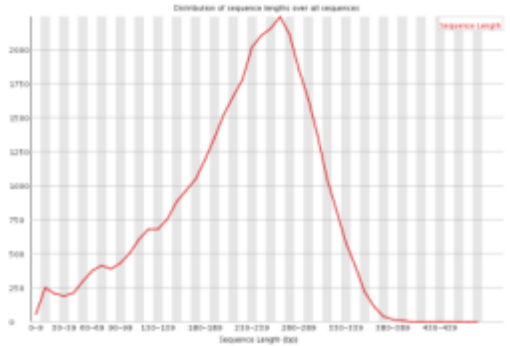
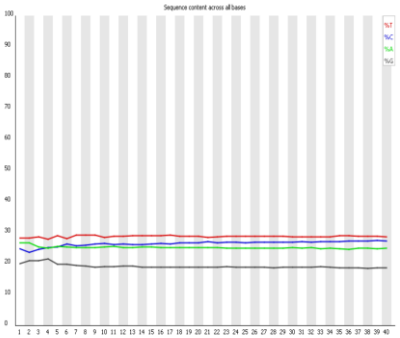
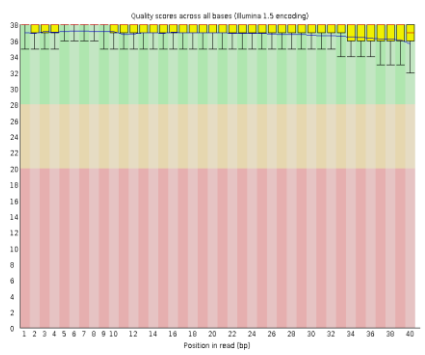
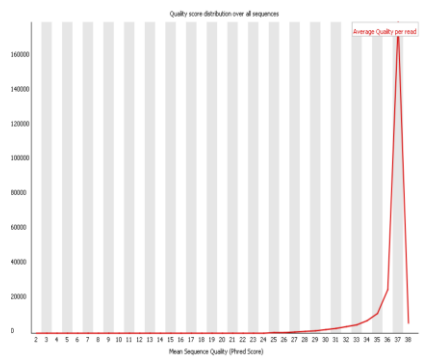
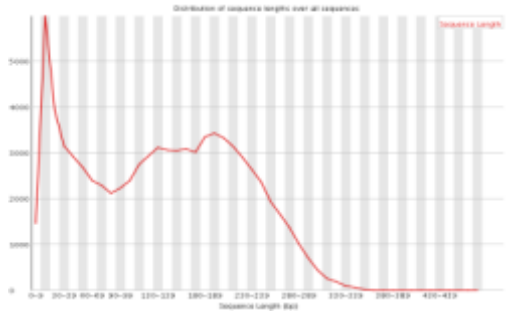
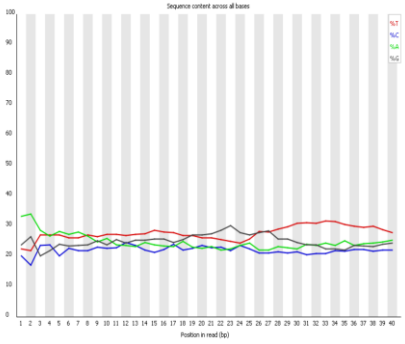
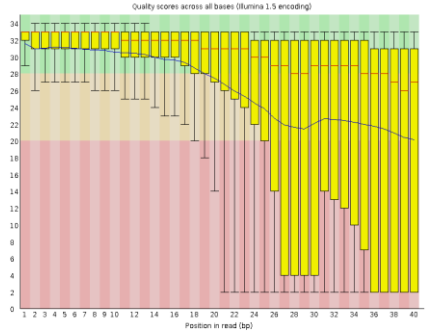
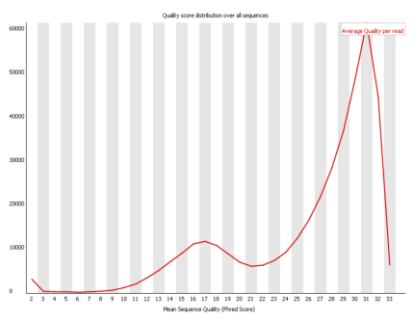
$$Q = -10 \log_{10} P \quad \Longleftrightarrow \quad P = 10^{\frac{-Q}{10}}$$

# Standard Workflow for NGS Analysis



# Quality Control on Raw Data

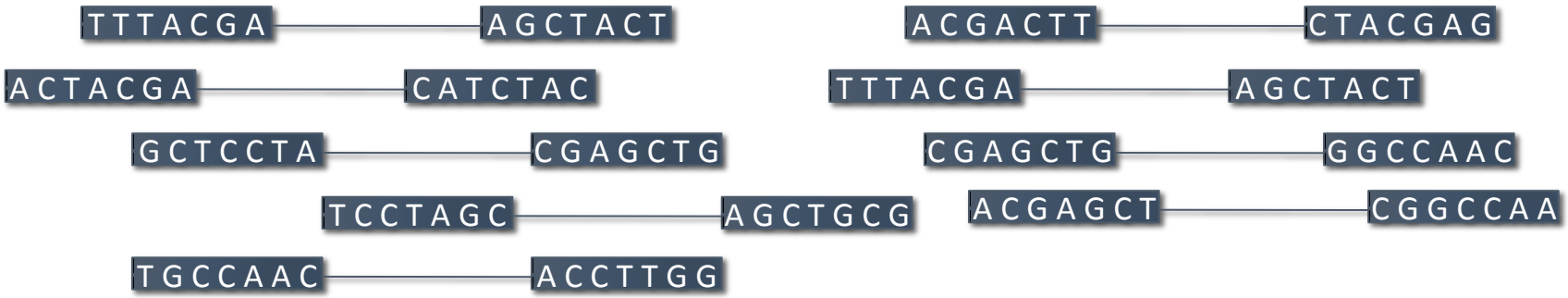
- Read length distribution
- Sequence content per base and % of GC
- Quality score per base and over the reads
- Overrepresented sequences
- Duplicated reads

S. Legras – 11/12/13

# Alignment on a reference genome

The reference genome is a **known** sequence, supposed to be **as close as possible** to the input genome, and which is used as an **anchor** to organize the single reads information.

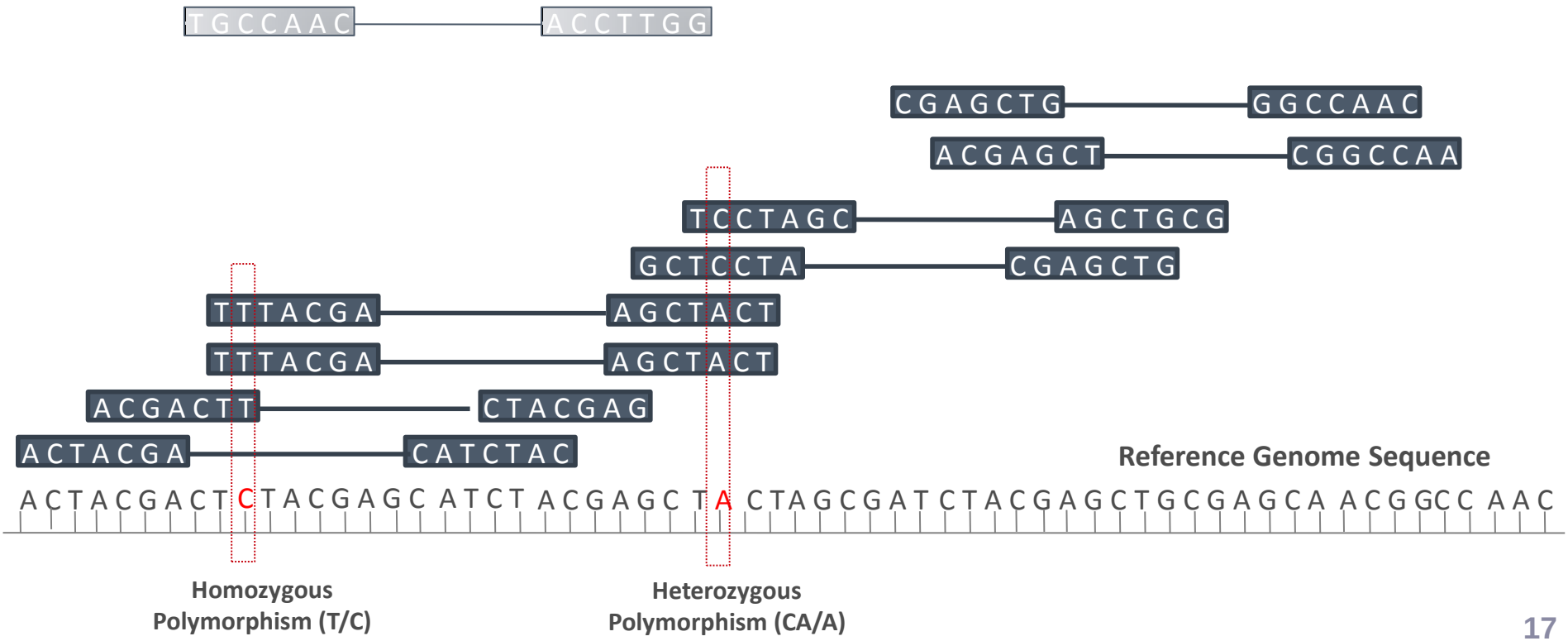



**Reference Genome Sequence**

ACTACGACTCTACGAGC ATCT ACGAGCTA CTAGCGATCTACGAGCTGCGAGCA ACGGCC AAC

# Alignment on a reference genome

The reference genome is a **known** sequence, supposed to be **as close as possible** to the input genome, and which is used as an **anchor** to organize the single reads information.



# Alignment on a reference genome
## Challenges

New alignment algorithms must address the requirements and characterics of NGS reads

- Millions of reads per run (30x of genome coverage)
- Reads of different size (35bp - 200bp)
- Different types of reads (single-end, paired-end, mate-pair, etc.)
- Base-calling quality factors
- Sequencing errors ( ~ 1%)
- Repetitive regions
- Sequencing organism vs. reference genome
- Must adjust to evolving sequencing technologies and data formats

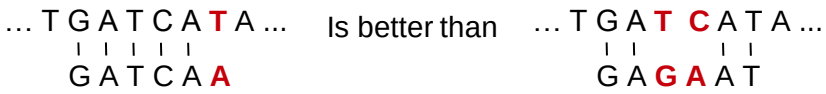
# Finding the best alignment

## Rational

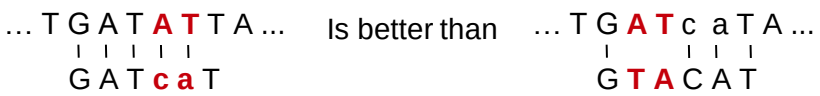
Given a reference and a set of reads, report at least one “good” local alignment for each read if one exists

What is “good”? For now, we concentrate on:

– Fewer mismatches is better




– Failing to align a low-quality base is better than failing to align a high-quality base




Based on a scoring system, i.e. score for a match (1), MM penalty (3), gap open penalty (5), gap extension penalty (2). The best alignment is the one with the highest score.

# Alignment on a reference genome

## Key points

- The alignment is a crutial step of the NGS analysis
- The alignment has to be define according to the biological application
- The reference genome has to be carefully chose
- The mappability of the region of interest has to be takken into account (primer design)
- The scoring method has to be chose accordingly to the sequencing error rate and the quality of the raw reads
- The alignment parameters have to be set properly

# The BAM format


- Allows to represent the data of any sequencer
- Analyses can then be conducted whichever the particular sequencer used
- Can contain data from a single or from several samples
- Example record (truncated):

Read identifier
Mapped locus
Read sequence
Sequence quality
Metadata

```
@FCD19MJACXX:2:1101:1735:1993#GTTCGACA/1 \
chr1    14410                    20M6I74M \
CTCAGTTCTTTATTGATTGGTGGGCCGTTTTCTCTGGAAT \
??;112=,2CF8?F9E42++33+<C8?1??GGC9:**:<0 \
AS:i:-9 XN:i:0  XM:i:3  XO:i:0  XG:i:0 ...
```

From R. Daveau

# Quality Control on Aligned Data

**In pratice, how to validate my alignment ?**

Be aware of the mapping strategy used

Look at simple descriptive statistics

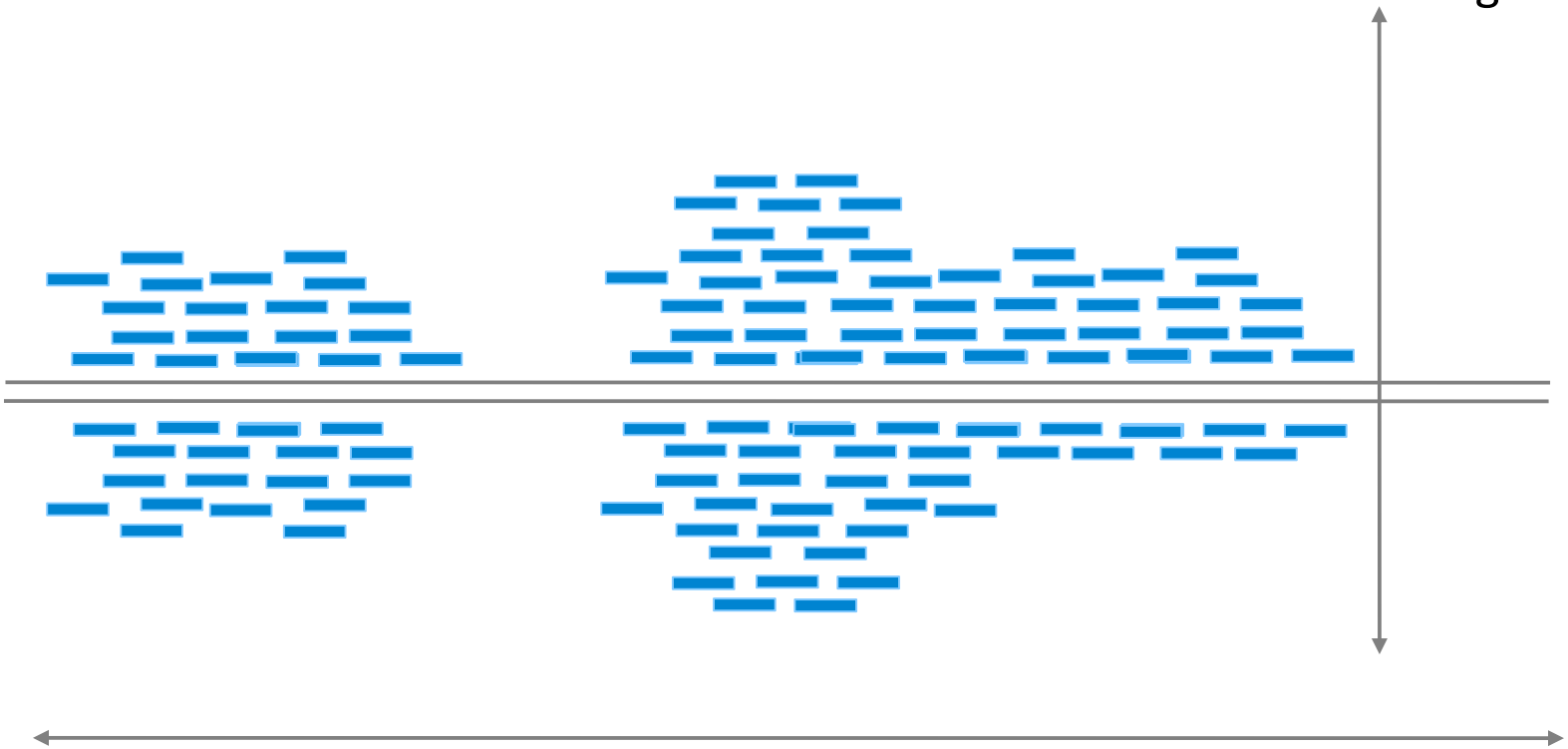
- Number of aligned reads
- Coverage/Depth
- Mapping quality
- Number of normal/abnormal pairs for paired-end data
- Strand bias
- …

# Mapping Statistics

## Coverage and Depth

**Depth of coverage**

= mean number of reads covering a base (X)

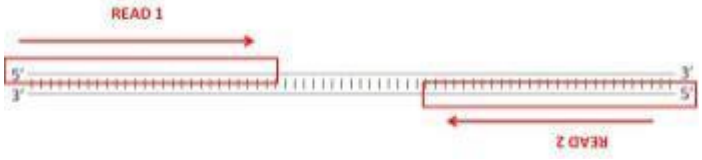


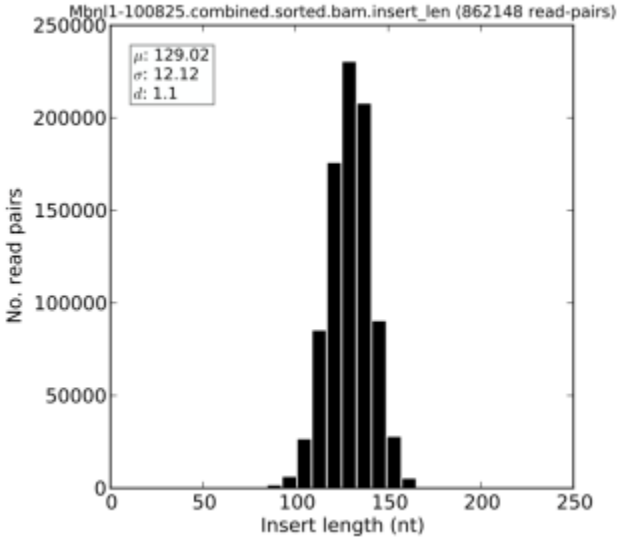
**Coverage**

= part of the reference with at least one read

# Paired-end mapping

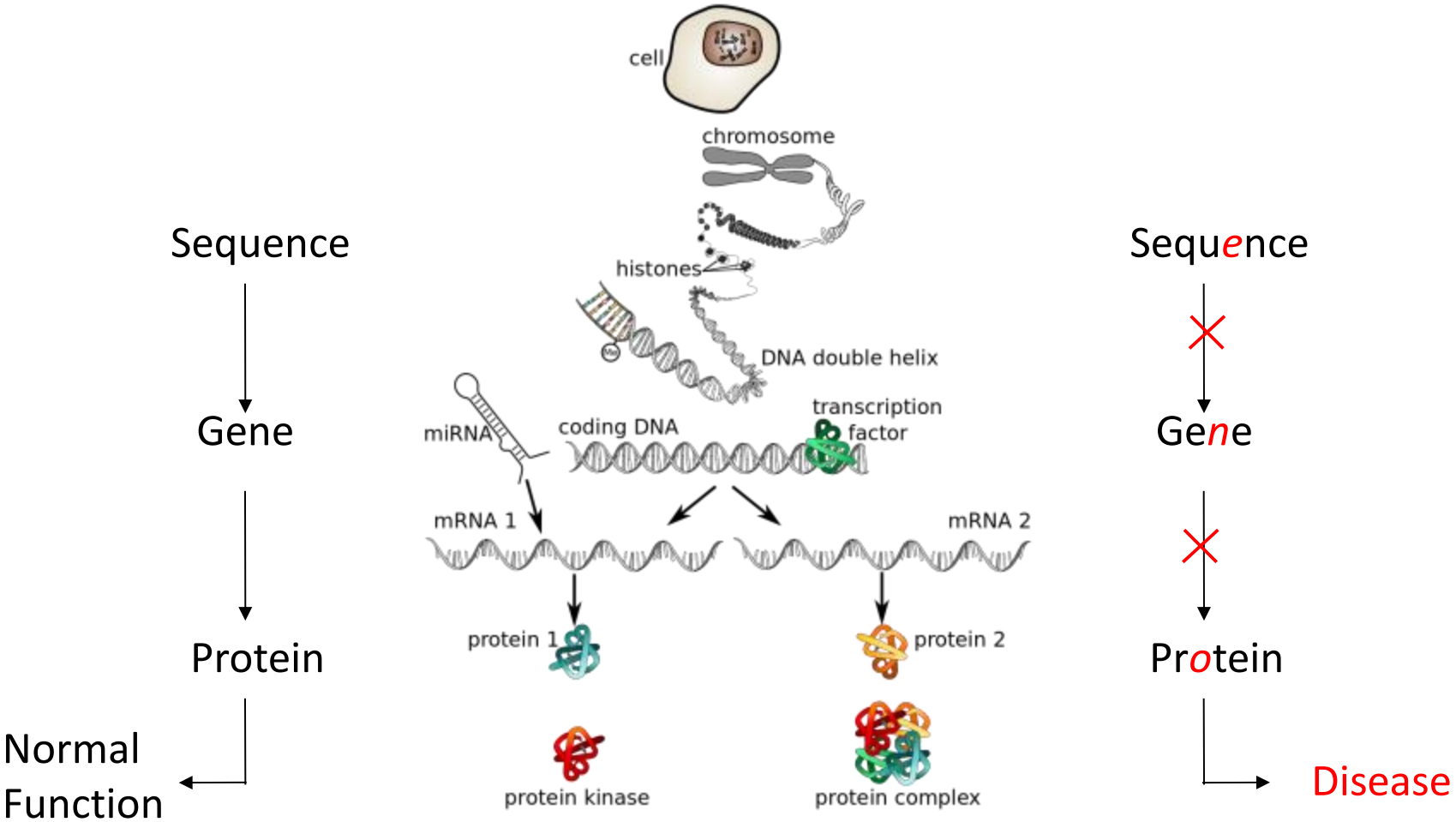
- Insert-size checking



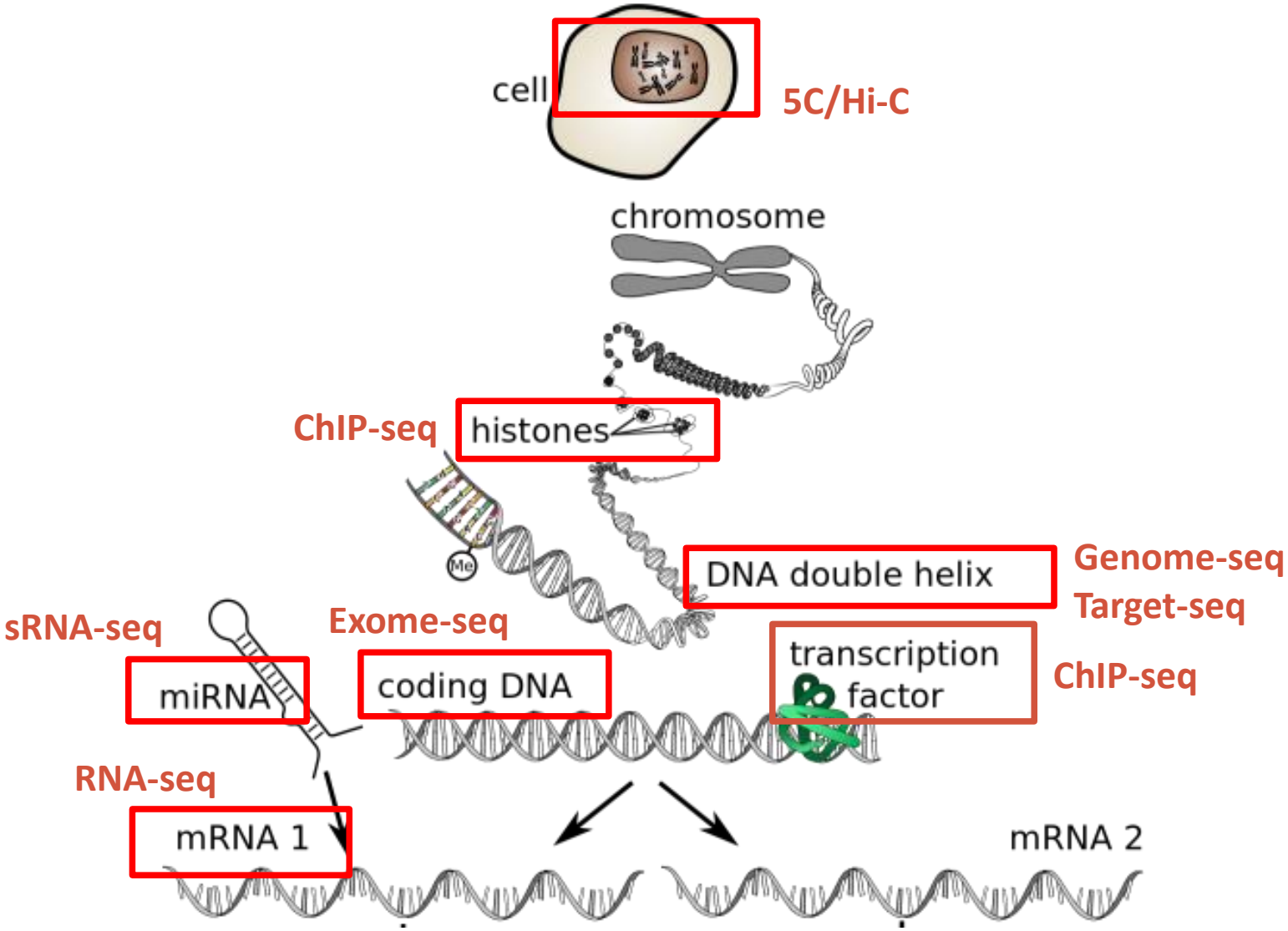




- % of **"All Good"**= both reads in the pair have aligned
  - "the pair is properly aligned" meaning that they mapped within a proper distance from each other
- % of **"All Bad"** = neither the read nor its mate mapped
- % of **Only one read maps =** only one read in a pair is mapped

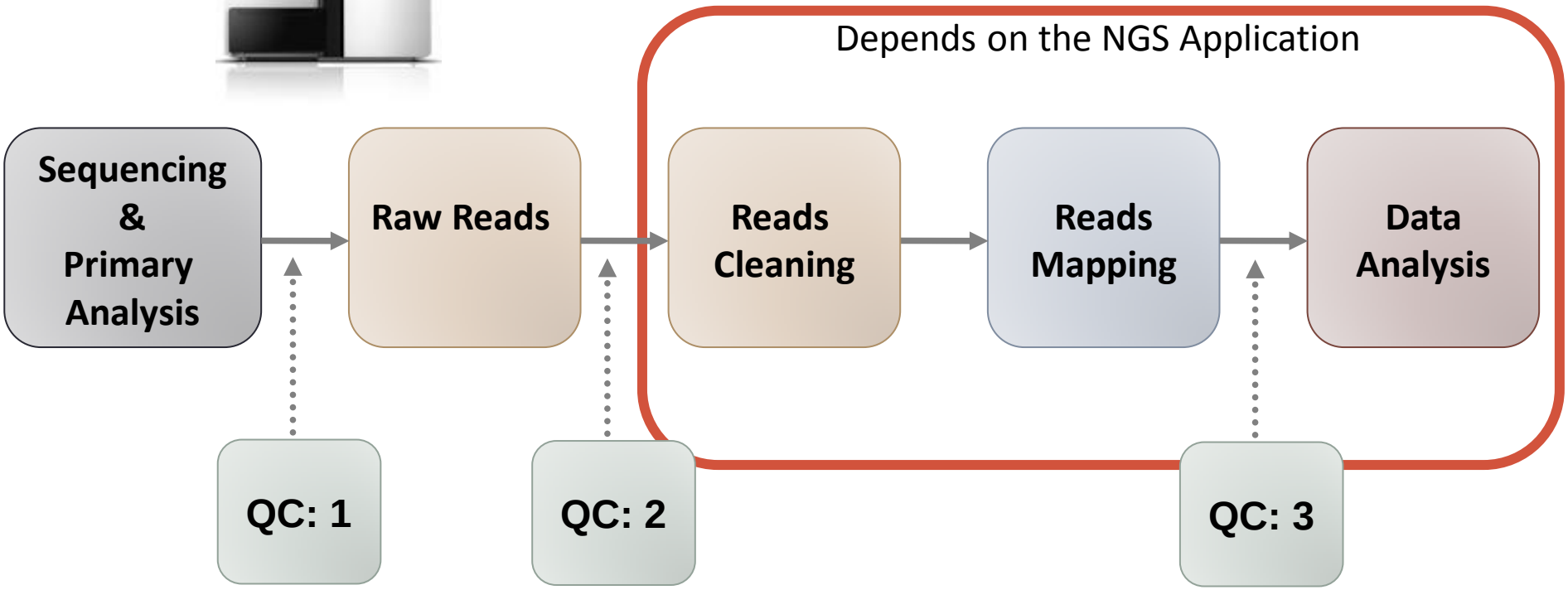
# From sequence to protein to disease



# NGS Applications



# Standard Workflow for NGS Analysis



# About today …

**Overview of the major sequencing applications and their bioinformatics solutions**

- DNA-seq
- RNA-seq
- ChIP-seq
- Chromosome conformation capture
- Clinical applications

institutCurie

# DNA SEQUENCING

Whole Genome-seq

Exome-seq

Target-seq

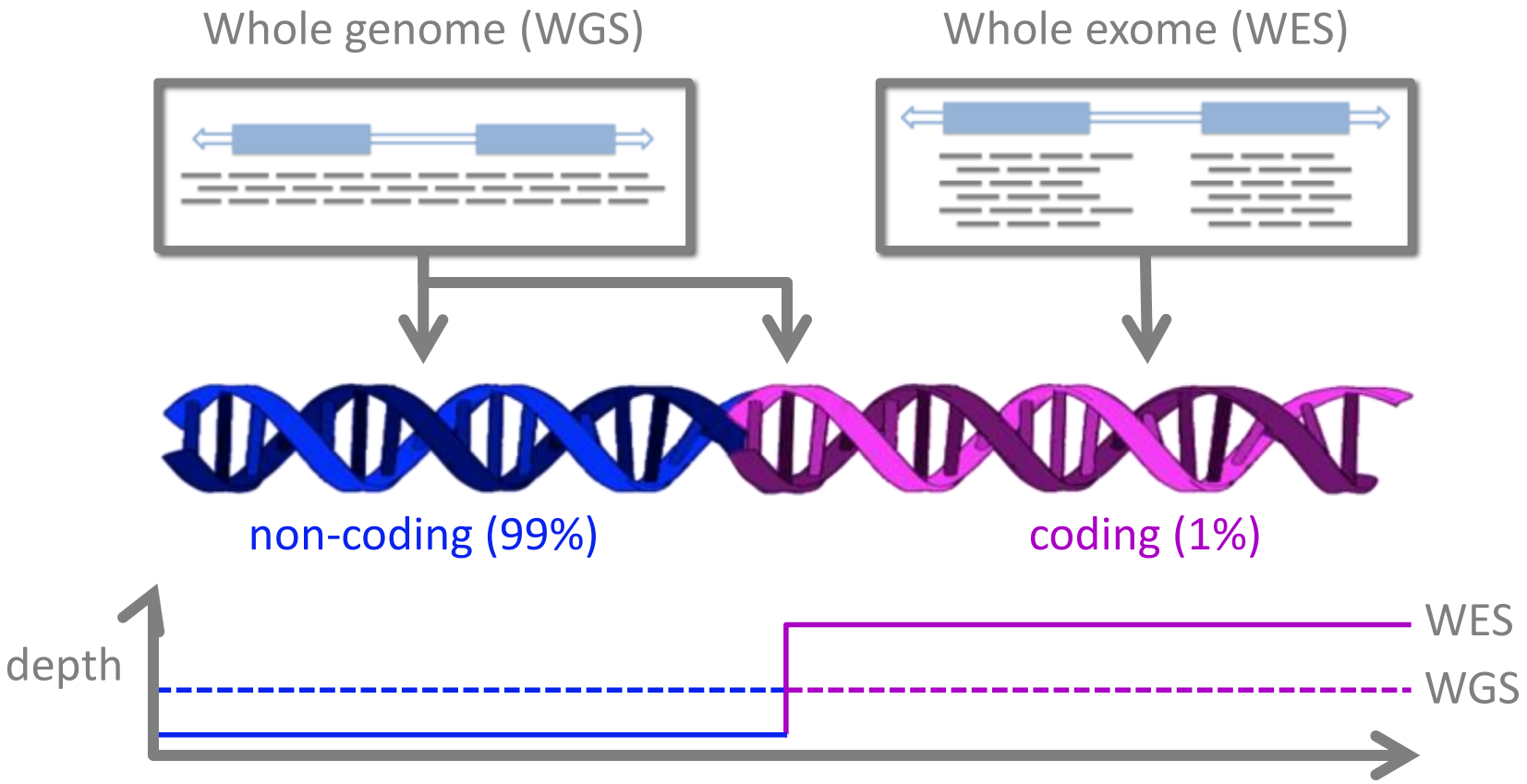
# Variants Calling

# Whole Genome or Target sequencing

- **Whole Genome sequencing** is used to determine the complete DNA sequence of an organism at a single time (associated with a depth of coverage)
  - Applications: Genome assembly - Detection of SNVs, SVs, CNVs

- **Target sequencing** is used to determine the DNA sequence of a targeted region (list of genes, exomes, hotspots, etc.). The targeted regions are captured before sequencing using different **enrichment** methods (PCR, arrays, etc.)
  - Applications: Detection of SNVs, (CNVs)

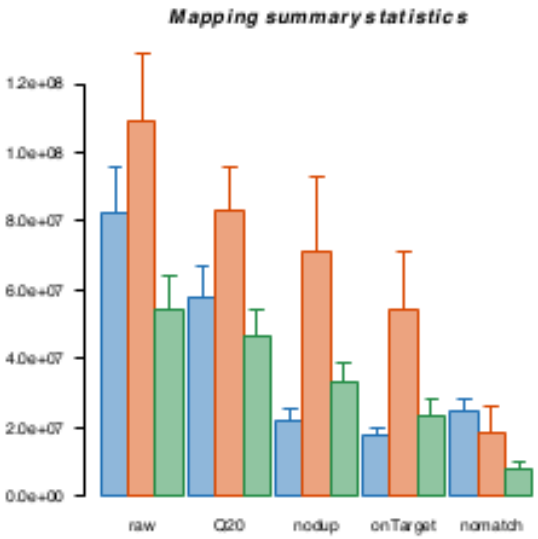
# A quick diversion about whole genome and exome sequencing

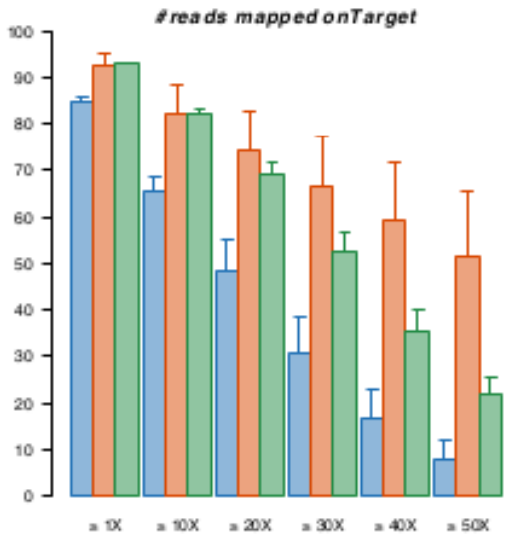


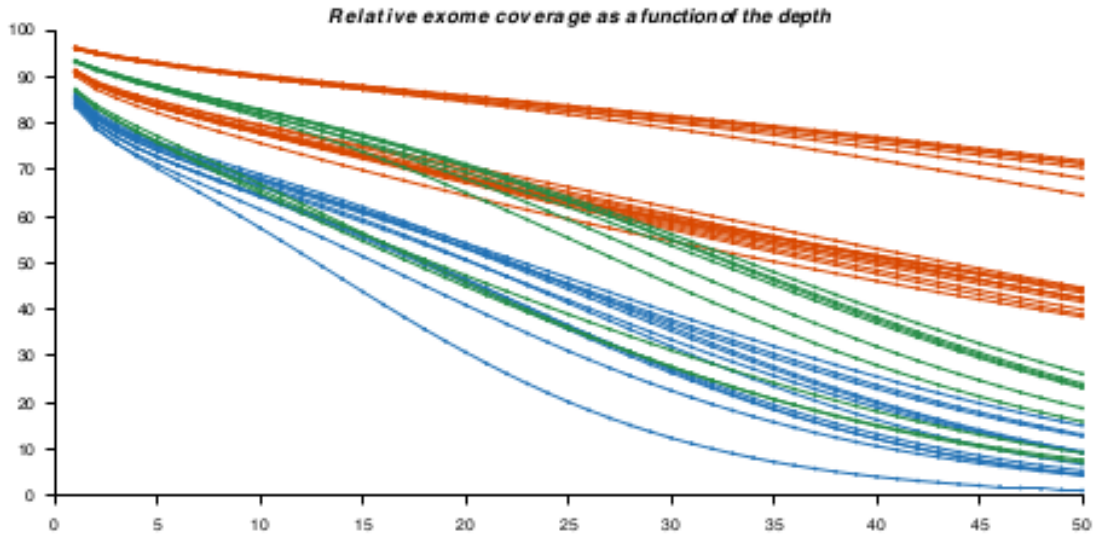

- WES generally provides higher depth and is thus more sensitive
- WGS allows variant detection outside the coding sequences (CDS)

From R. Daveau

# Checkpoints for Exome sequencing



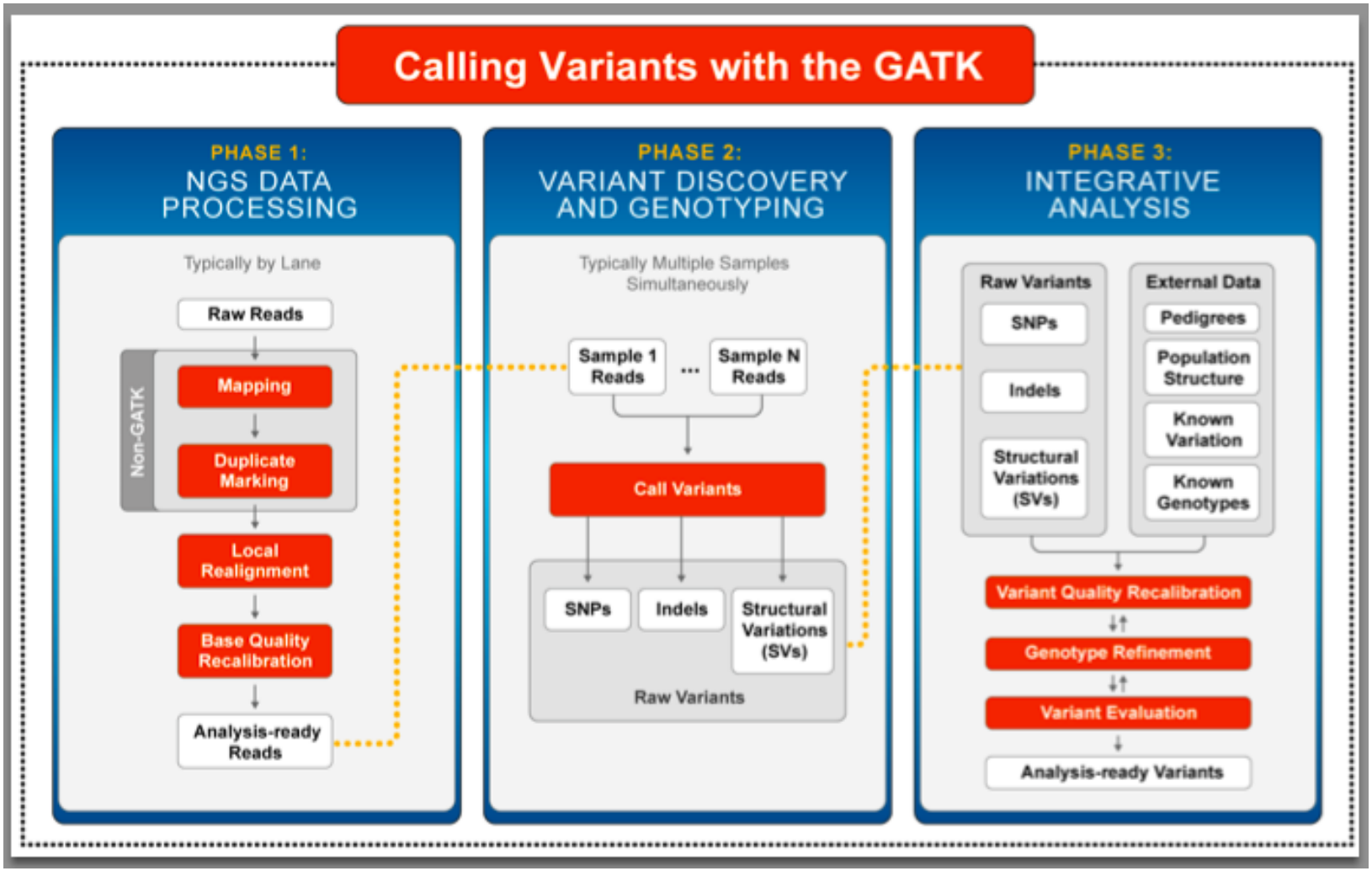

From R. Daveau

# Main DNA variants detectable through NGS

- SNVs and short indels are the most frequent events:
  - intergenic
  - intronic
  - *cis*-regulatory
  - splice sites
  - frameshift or not
  - synonymous or not
  - begnin or damaging etc…

- Example of SNV one want to pinpoint:
  - non-synonymous + highly deleterious + somatically acquired

# Overview of the GATK workflow for variant discovery

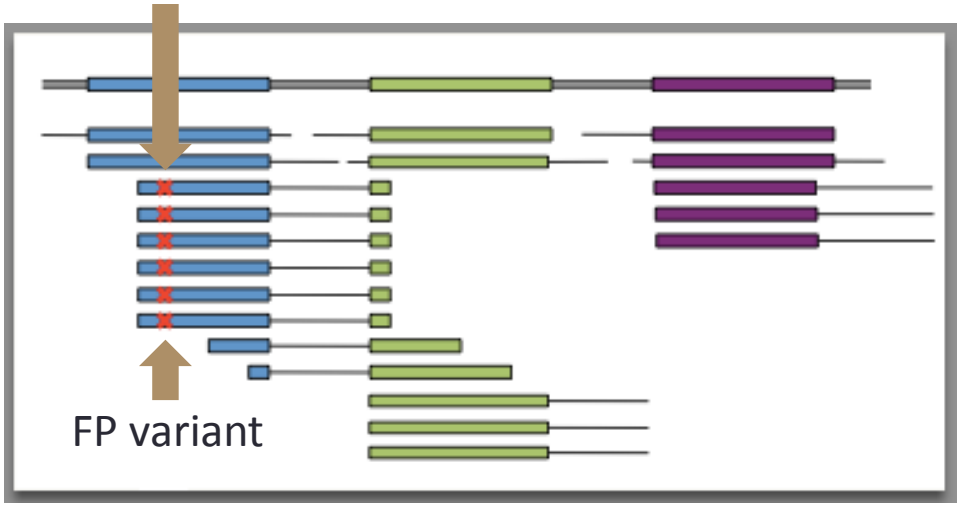



→ Y. Duffourd, JB Rivière – Module 3 – 12-13/12/13

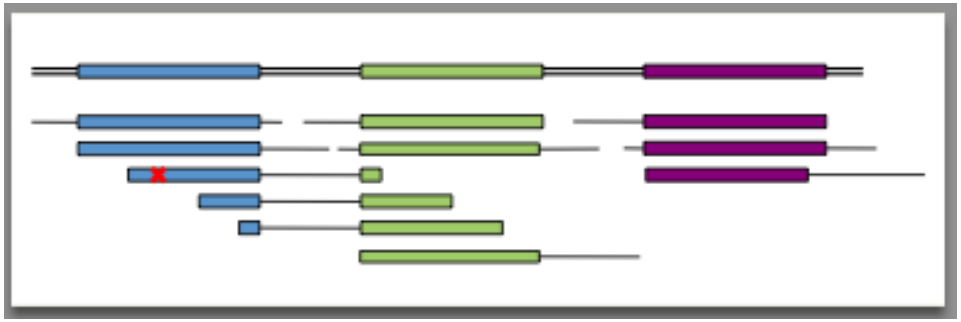
From R. Daveau

# The reason why PCR duplicates are bad

Sequencing error propagated in duplicates




After duplicates marking, GATK will see:



… and thus be more likely to make the right call

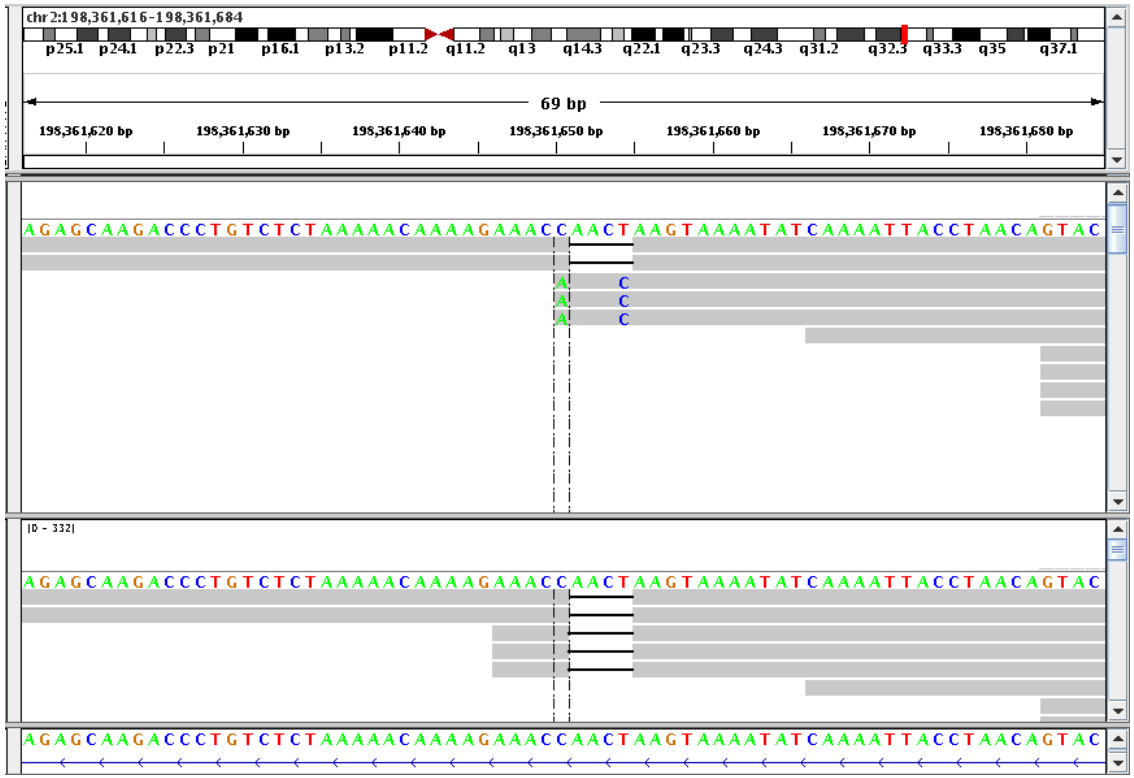
# Duplicates have the same starting position and CIGAR


**Picard Tool can be used for that!**

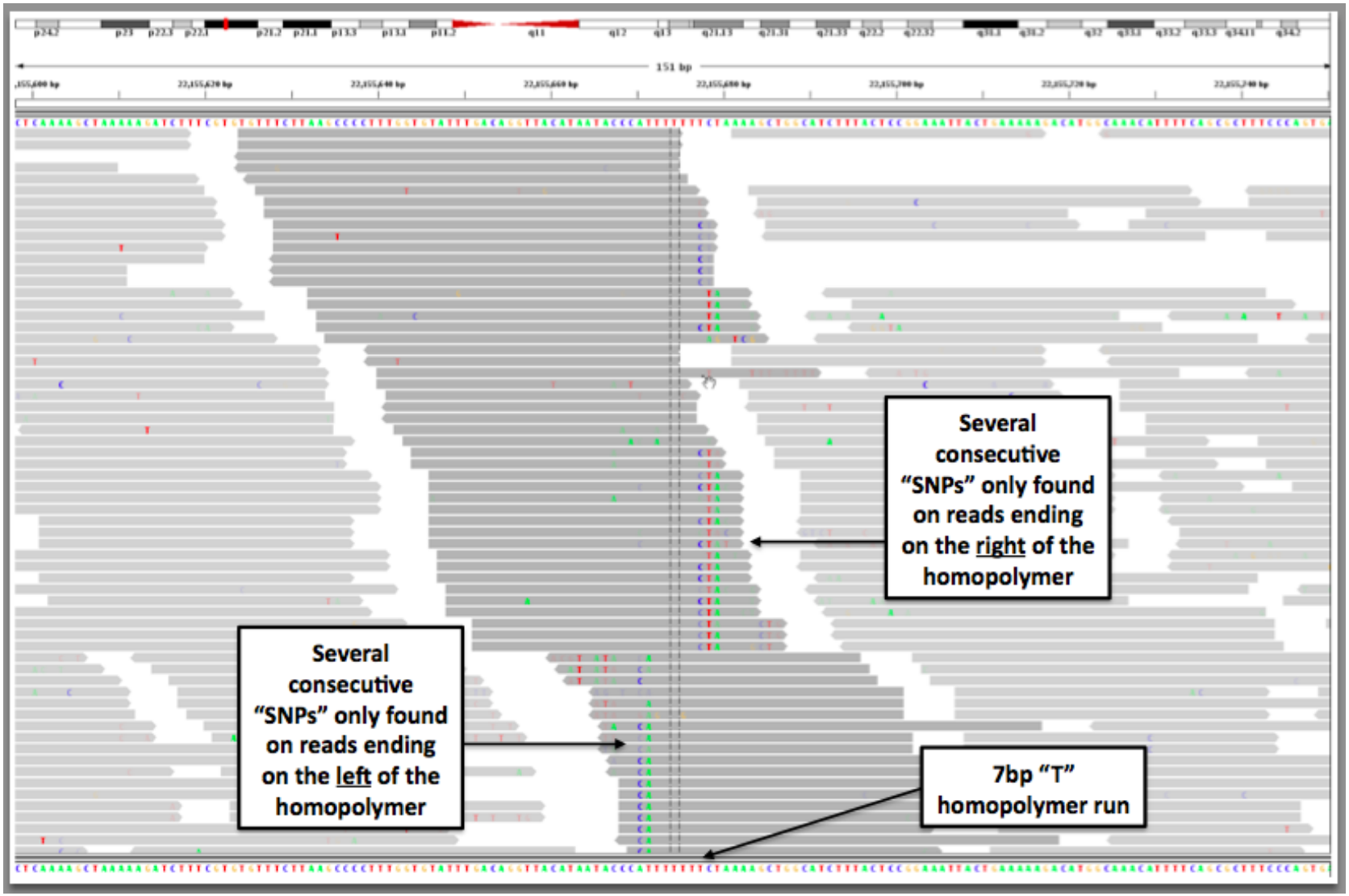
From R. Daveau

# Indels and Local Realignment

- Simple models for calling indels based on the initial alignments show a high number of false positives and negatives
- More sophisticated algorithms have been developed
  - Dindel, GATK, etc.



# Indels and Local Realignment




From R. Daveau

# Indels and Local Realignment


From R. Daveau

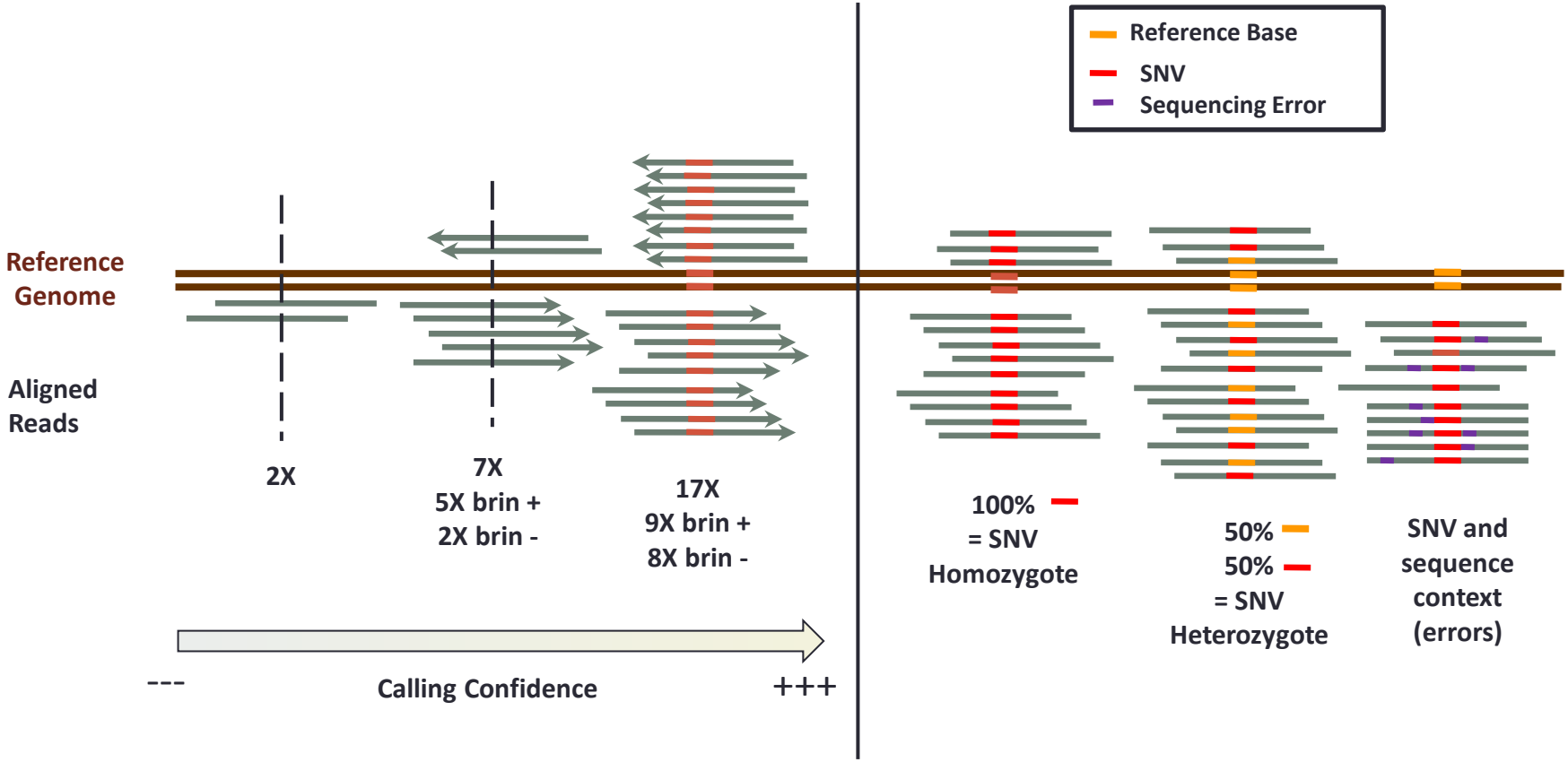
# SNVs Calling

- SNVs – Single Nucleotide Variations
  - Examine the bases aligned to position and look for differences
  - Look at the sequence context of the SNVs (errors, homopolymer, etc.)
- Homozygous vs heterozygous SNVs
- Factors to consider when calling a SNVs:
  - Base call qualities of each supporting base
  - Proximity to small indels, or homopolymer run
  - Mapping qualities of the reads supporting the SNP
    → Low mapping qualities indicates repetitive sequence
  - Read length
    → Longer reads = higher mapping confidence
  - Sequencing depth
    → More than 30X for a SNVs
  - SNVs position within the reads
    → Higher error rate at the reads ends
  - Look at strand bias

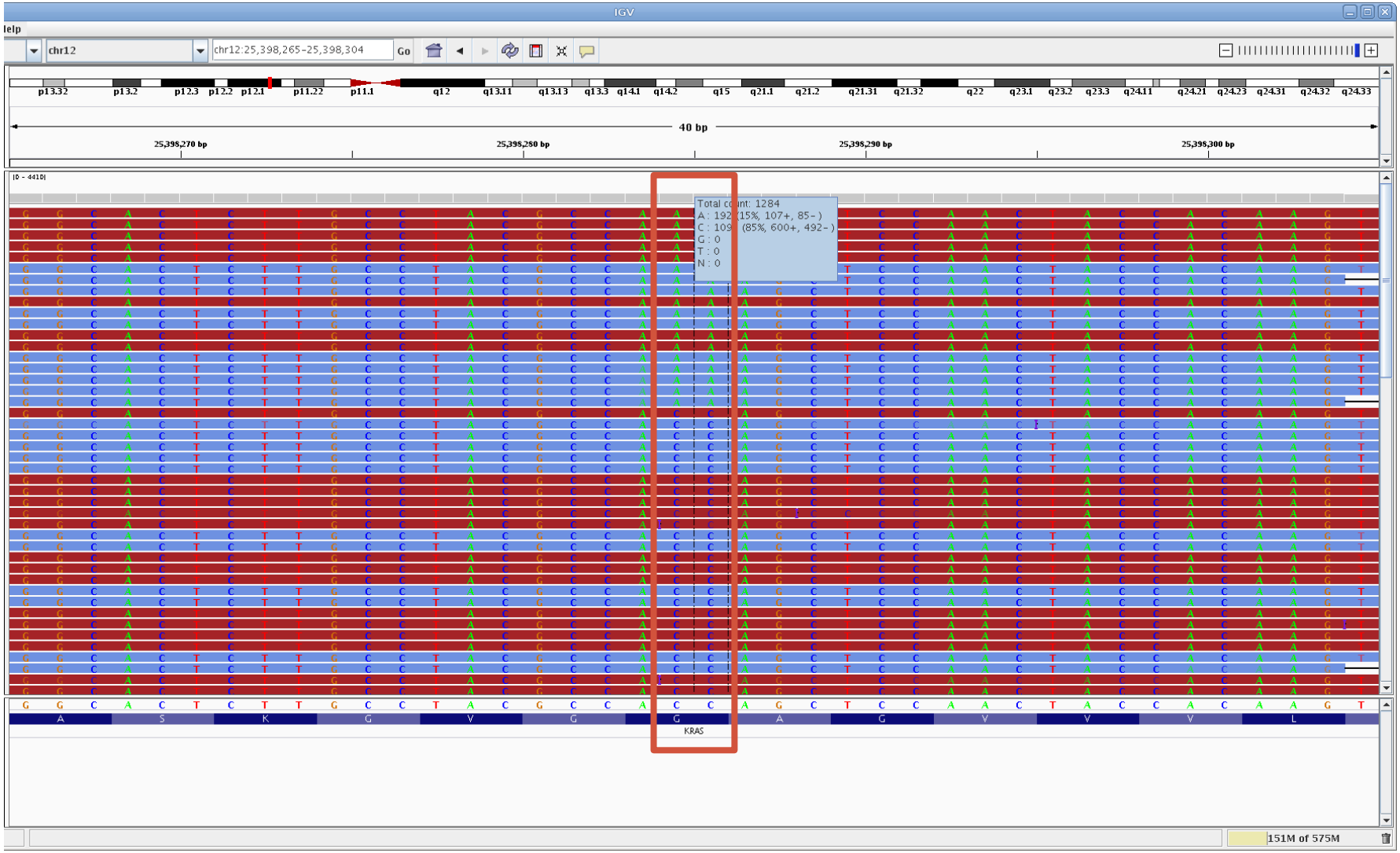
# SNVs Calling and Depth of Coverage

Depth of Coverage = number of reads supporting one positions

ex: 1X, 5X, 100X… >1000X



# SNVs Calling - example



# Short Indels Calling

- Small insertions and deletions observed in the alignment of the read relative to the reference genome

- Factors to consider when calling indels
  - Misalignement of the read
    - Alignment scoring – cheaper to introduce multiple SNPs than an indel
    - Sufficient flanking sequence quality at each side of the read
  - Homopolymer runs
  - Length of the reads
  - Homozygous or heterozygous

# Variant Calling Format (VCF)

- 2 formats are widely used:
  - Pileup format
  - VCF format
- Pileup format is Samtools Mpileup tool specific
- VCF format is the default output from all others SNP caller

# VCF format

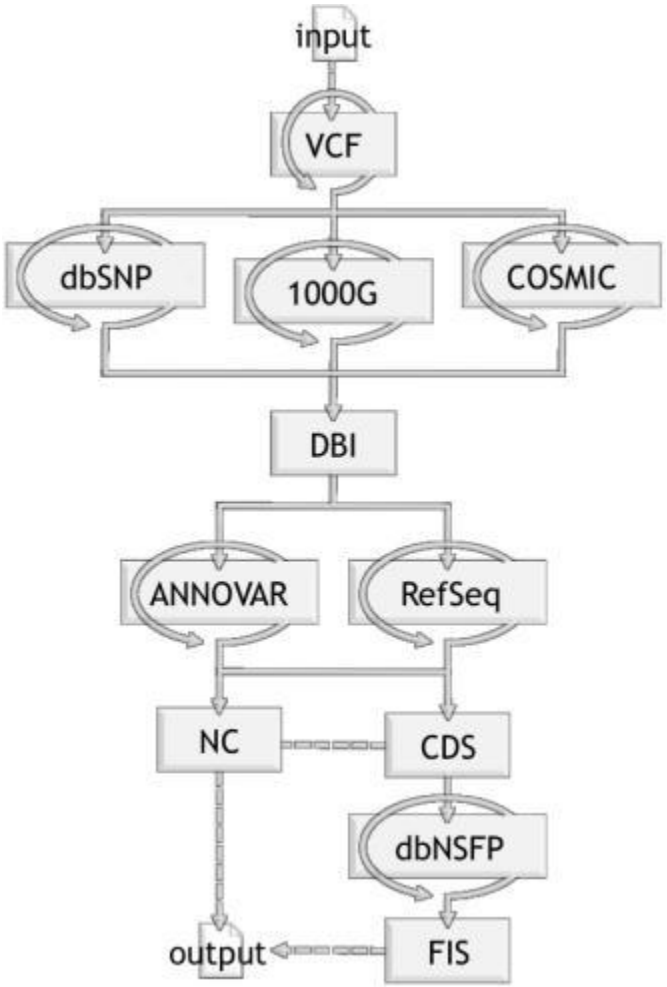
- VCF for Variant Calling Format
- Describe variations such as:
  - Unique base mutation (SNV)
  - Insertion/deletion (indels)
- Tabulated text file:
  - Chromosome/Position
  - Identifier
  - Nucleotide in référence
  - Nucleotide in sample
  - Variation quality
  - Filter
  - Supplementary informations (allelic ratio, annotation, …)

```
20     14370   rs6054257 G    A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2
20     17330   .         T    A       3    q10    NS=3;DP=11;AF=0.017
20     1110696 rs6040355 A    G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20     1230237 .         T    .       47   PASS   NS=3;DP=13;AA=T
20     1234567 microsat1 GTCT G,GTACT 50   PASS   NS=3;DP=9;AA=G
```

# Resources dedicated to human genetic variations

- dbSNP and 1000-genomes
  - Population-scale DNA polymorphisms
- COSMIC
  - Catalogue Of Somatic Mutations In Cancer
- dbNSFP
  - Database of non-synonymous SNV functional predictions
- ANNOVAR
  - Tools to annotate genetic variations
- Automatic annotation-based DNA variant prioritization can be done by integrating these resources in a single workflow
  - The GFAP we have developed is such a tool

# Filtering and Annotation of Variants




**http://gfap.curie.fr**

GFAP features

| Features | NC | CDS | FIS |
|---|---|---|---|
| chr:start-end | ☑ | ☑ | ☑ |
| ATG-based positions | ☐ | ☑ | ☑ |
| ref./alt. alleles | ☑ | ☑ | ☑ |
| #F-ref./alt. reads | ☑ | ☑ | ☑ |
| #R-ref./alt. reads | ☑ | ☑ | ☑ |
| total #reads | ☑ | ☑ | ☑ |
| #alt. alleles | ☑ | ☑ | ☑ |
| alt. allele ratio | ☑ | ☑ | ☑ |
| alt. allele likelihood | ☑ | ☑ | ☑ |
| strand-bias p-values | ☑ | ☑ | ☑ |
| GFAP/VCF-filter fields | ☑ | ☑ | ☑ |
| alt. allele frequencies in 1000G | ☑ | ☑ | ☑ |
| COSMIC/dbSNP links | ☑ | ☑ | ☑ |
| genomic annotation | ☑ | ☑ | ☑ |
| gene symbol(s) | ☑ | ☑ | ☑ |
| RefSeq identifiers | ☐ | ☑ | ☑ |
| amino-acid substitution(s) | ☐ | ☐ | ☑ |
| functional impact score(s) | ☐ | ☐ | ☑ |
| functional impact prediction | ☐ | ☐ | ☑ |

Fully automated pipeline for filtering and annotation

From R. Daveau

# More informations

## SNVs/Indels Calling

- Samtools
  - http://bioinformatics.oxfordjournals.org/content/25/16/2078.long
  - http://samtools.sourceforge.net/
- GATK
  - http://www.broadinstitute.org/gatk/
- ANNOVAR
  - http://www.openbioinformatics.org/annovar/
- More : varscan2, SOAP2, etc.

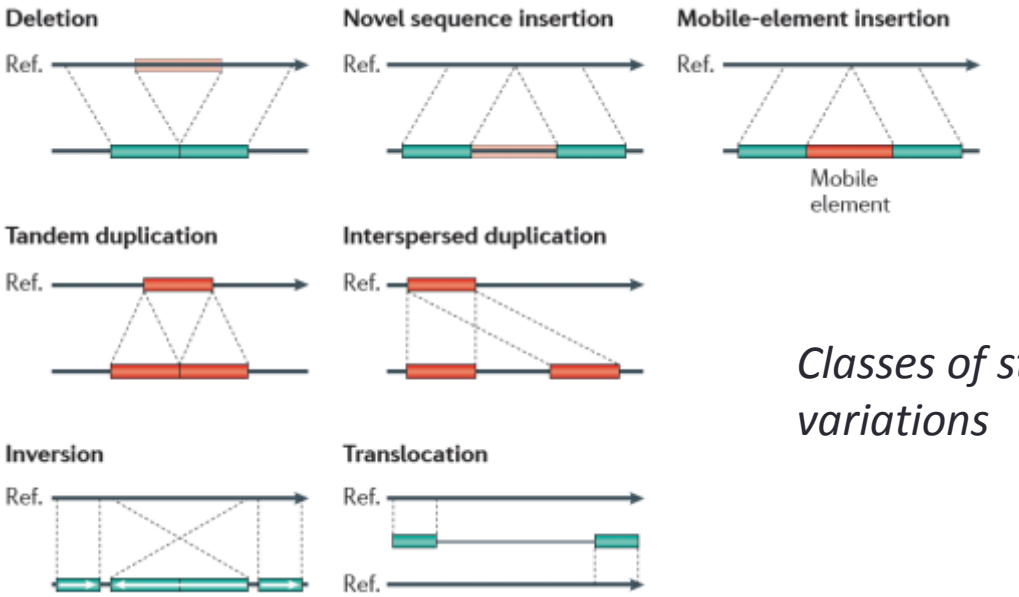
## VCF File

- http://vcftools.sourceforge.net/

# Structural Variation (SV)

# Structural variations and Cancer

A number of human tumors have been shown to be associated **with chromosomal mutations**. This could append at a chromosome level **(chromosome loss or duplication)** or through a change in chromosome structure **(deletions, duplications, inversions and translocations).**

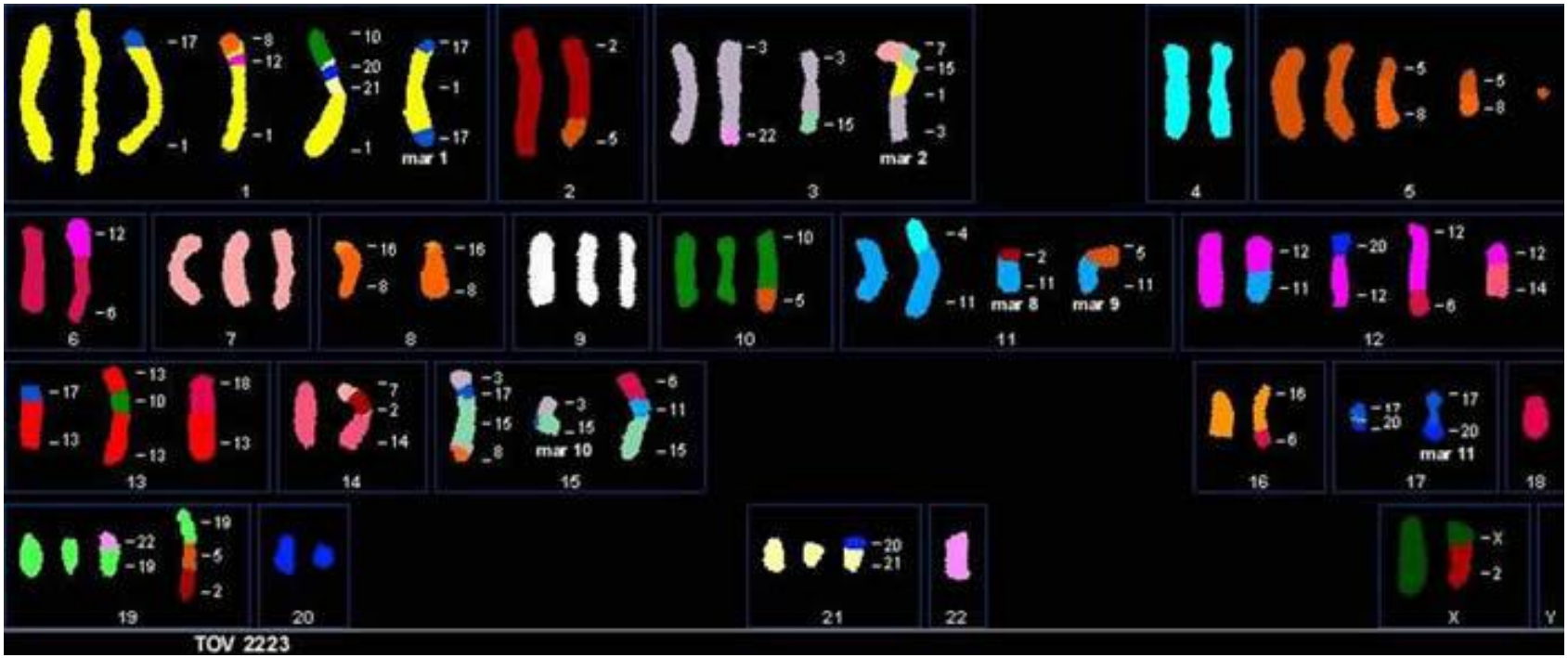



*Classes of structural variations*

*Alkan et al 2012*

From B. Zeitouni

# Genomic Rearrangements and Cancer

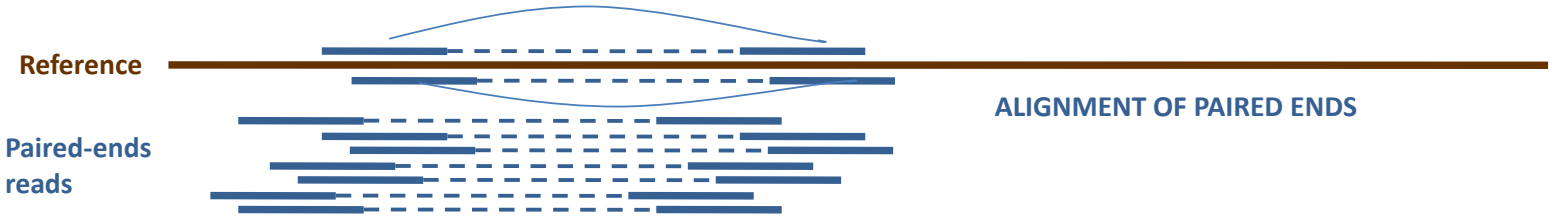


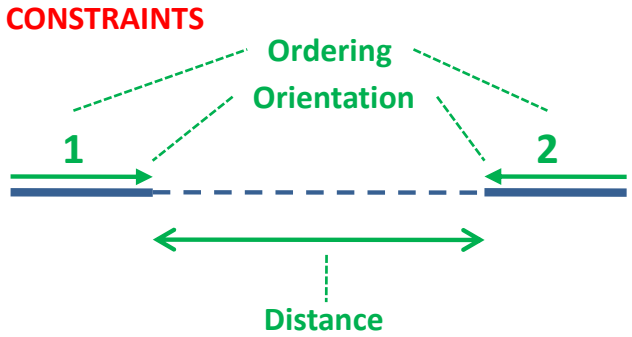

From B. Zeitouni

# Detection of Structural Variation (SVs)

## Paired-end based method (Breakdancer, PEMer, SVDetect, etc.)

Read-pair methods assess the span and orientation of paired-end reads and cluster 'discordant' pairs in which the mapping span and/or orientation of the read pairs are inconsistent with the reference genome

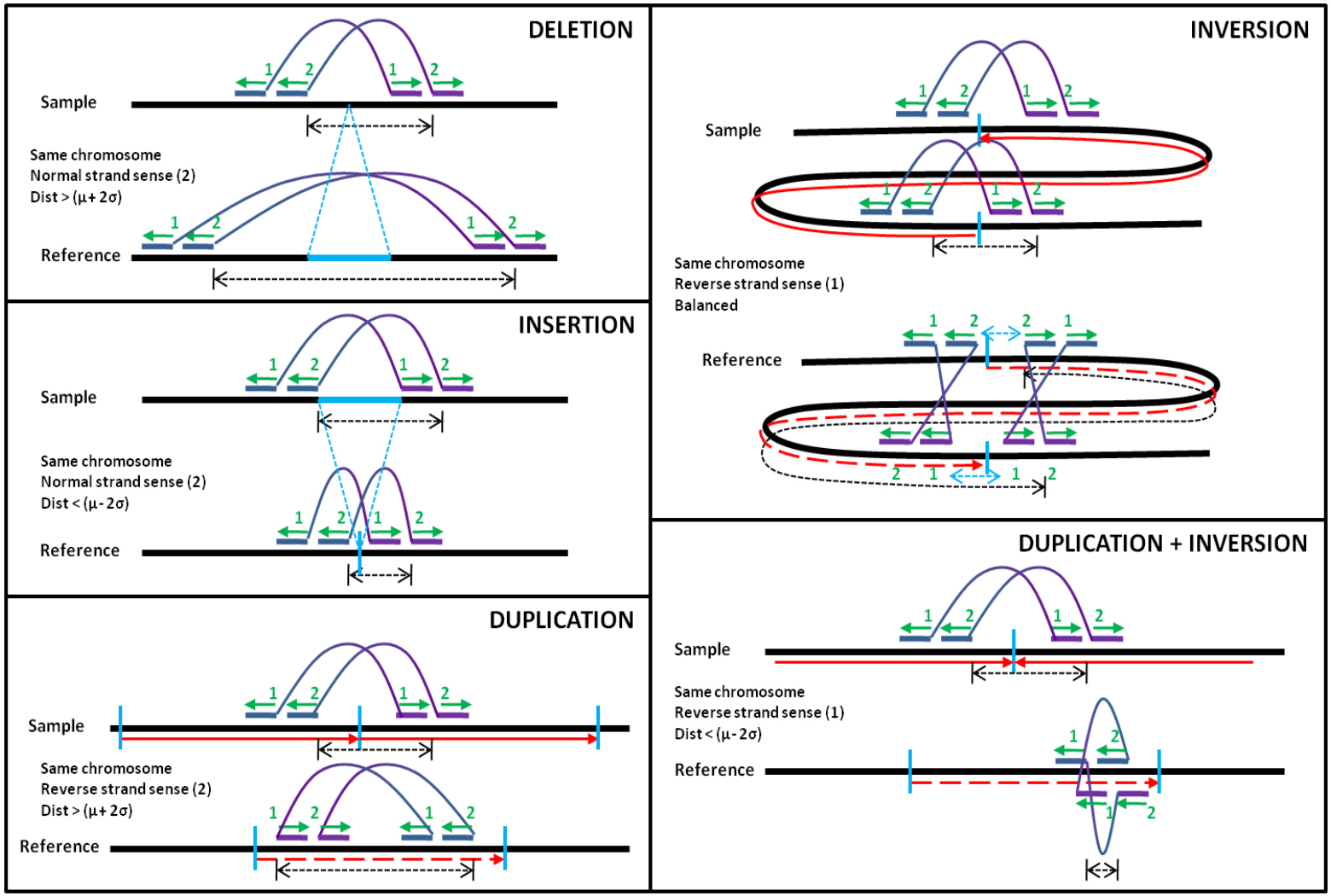



From B. Zeitouni

# Detection of Structural Variation (SVs)

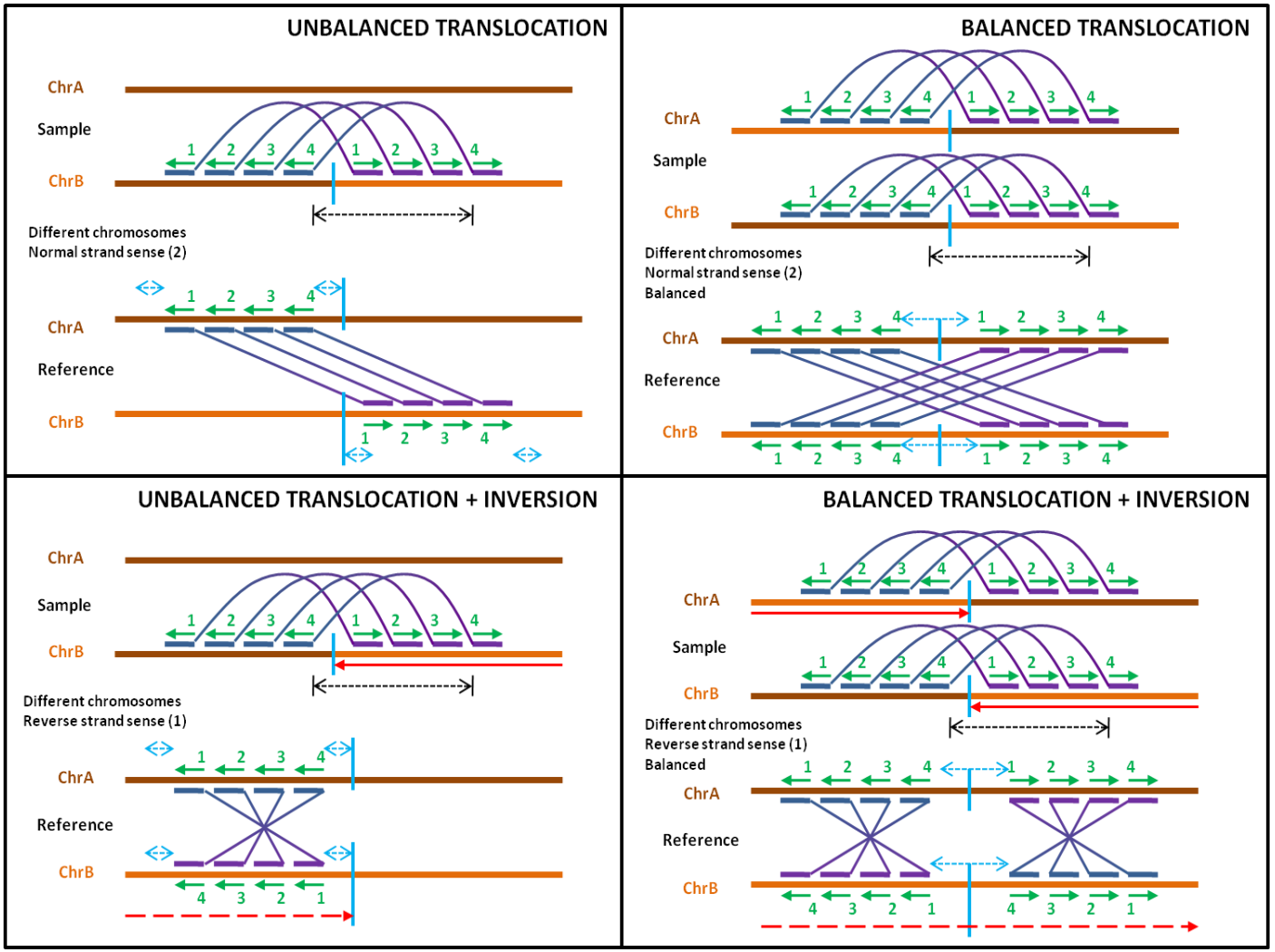
## Intra-chromosomal SVs




From B. Zeitouni

# Detection of Structural Variation (SVs)

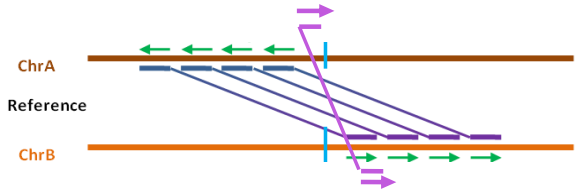
## Inter-chromosomal SVs

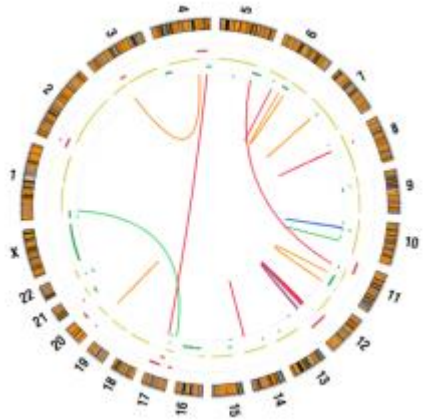



From B. Zeitouni

# SVDetect (Zeitouni et al.)

- Prediction of **>10 types of structural variants** (SVs), Inter/intra-chromosomal rearrangements
- **Many filters** to remove pairs inconsistent with the PE signature of the predicted SV



- **Compare SVs** predicted for different samples (Tumor vs Normal)
- **Graphical** view: BED/Circos
- **Total compatibility** (PE, technology), easy set-up and usage, parallel computing
- **Performance**: 3 millions of abnormal pairs and 8 CPU : 10 min 16 sec, 15 GB of memory usage

Zeitouni B., Boeva V. et al. SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. 2010. Bioinformatics 26: 1895-1896

From B. Zeitouni

# Copy Number Variation (CNV)

# Detection of Copy Number Variation (CNV)

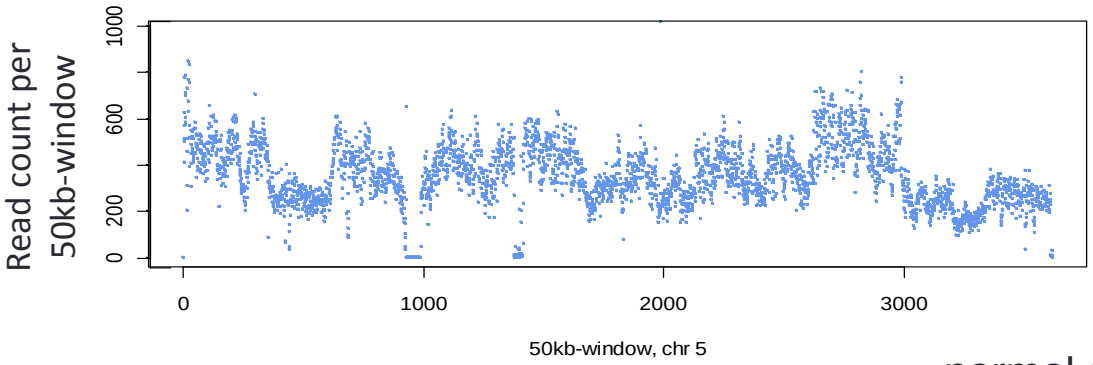
**The copy number variations can be assessed from Whole Genome or Exome data**

Broadly, there are four methods for CNV detection using NGS data
- Depth of coverage (DOC) methods
- Paired-end mapping (PEM) methods
- Split-read (SR) methods
- Assembly-based (AS) methods

But also some bias to normalize :
- GC content
- Mappability
- Contamination with normal cells
- Capture bias

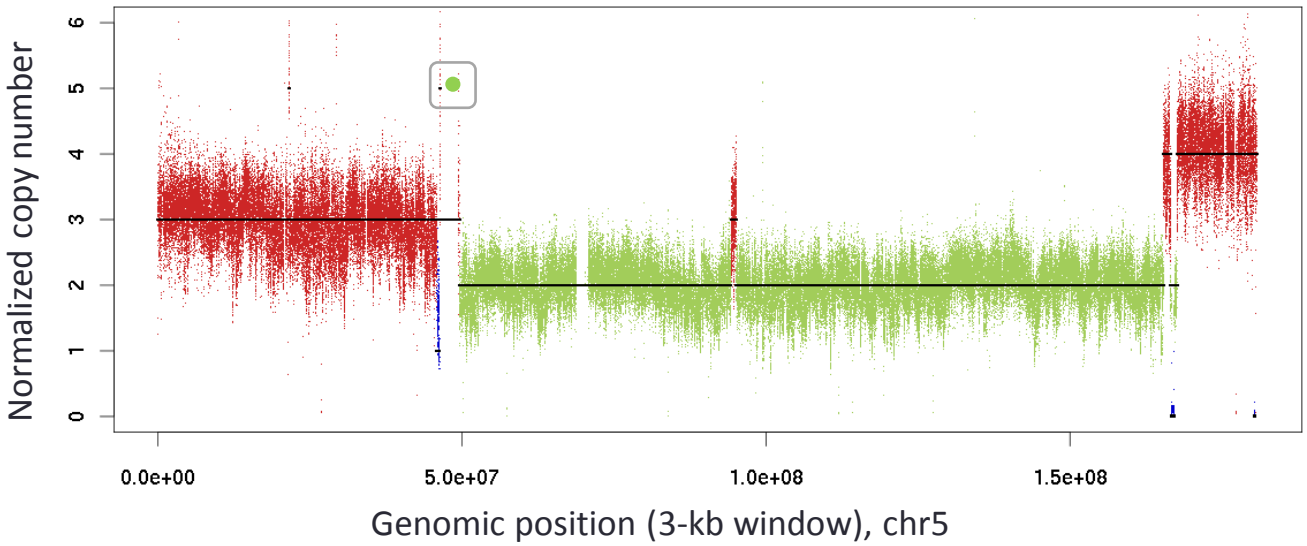
# Control-freec (Boeva et al.)

**Control-Freec** is an application for annotation of copy-numbers in cancer sequencing data



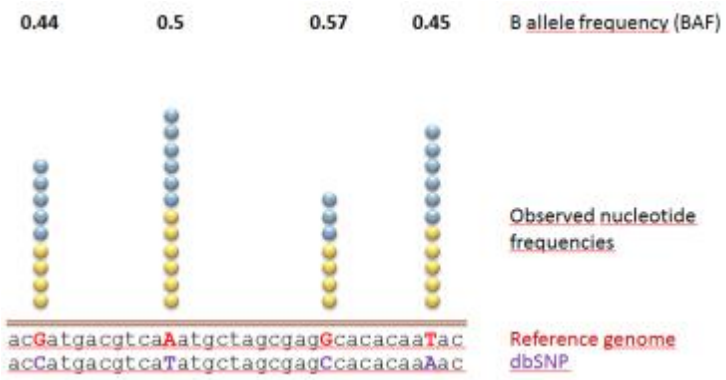

Normalization:
- GC-content,
- mappability,
- Contamination by normal cells

– normal copy number – loss – gain

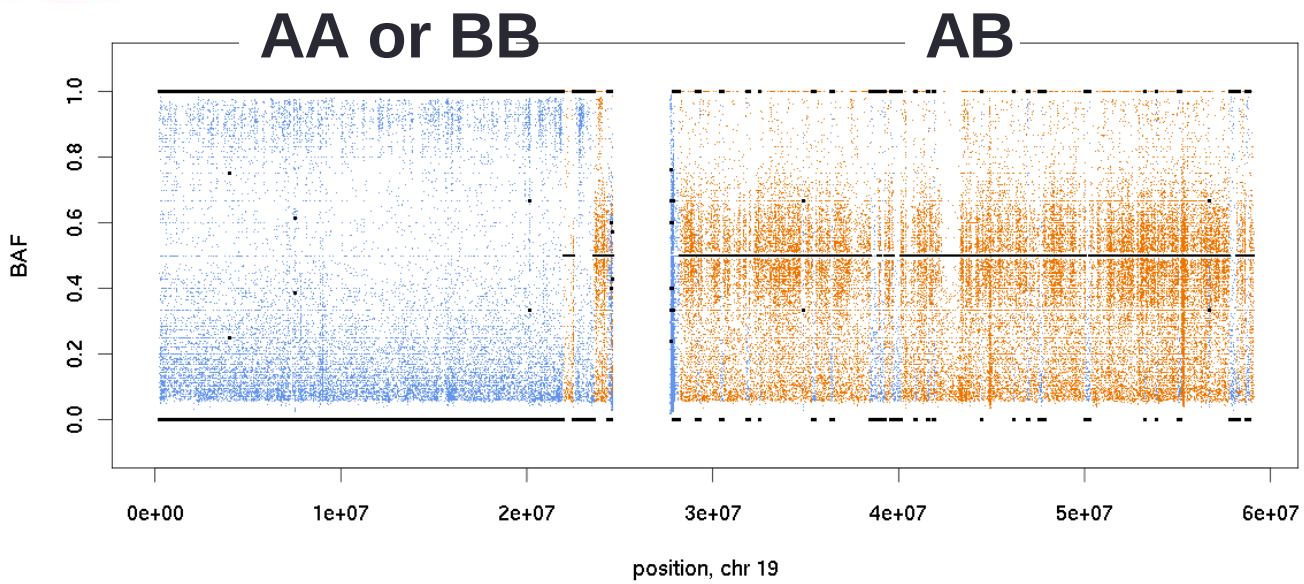



From V. Boeva

# Control-freec (Boeva et al.)




Normalization:
- mappability,
- Contamination by normal cells




Boeva V., Zinovyev A. et al. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. 2011. Bioinformatics 27(2):268-9.
Boeva V., Popova T. et al. Control-FREEC : a tool for assessing copy number and allelic content using next generation sequencing da ta. 2011. Bioinformatics 28(3):423-5 .

# DNA-seq - FAQ

- How many reads do I need ?
- What is the difference between Whole Genome sequencing and DNA microarrays analysis ?
- Whole Genome-seq or exome-seq ?
- Can I find CNV with Exome-seq ?
- Which control(s) do I need ?

institutCurie

# RNA SEQUENCING

RNA-seq

Small RNA-seq

# The RNAs World

# Eukaryotic Genome Complexity

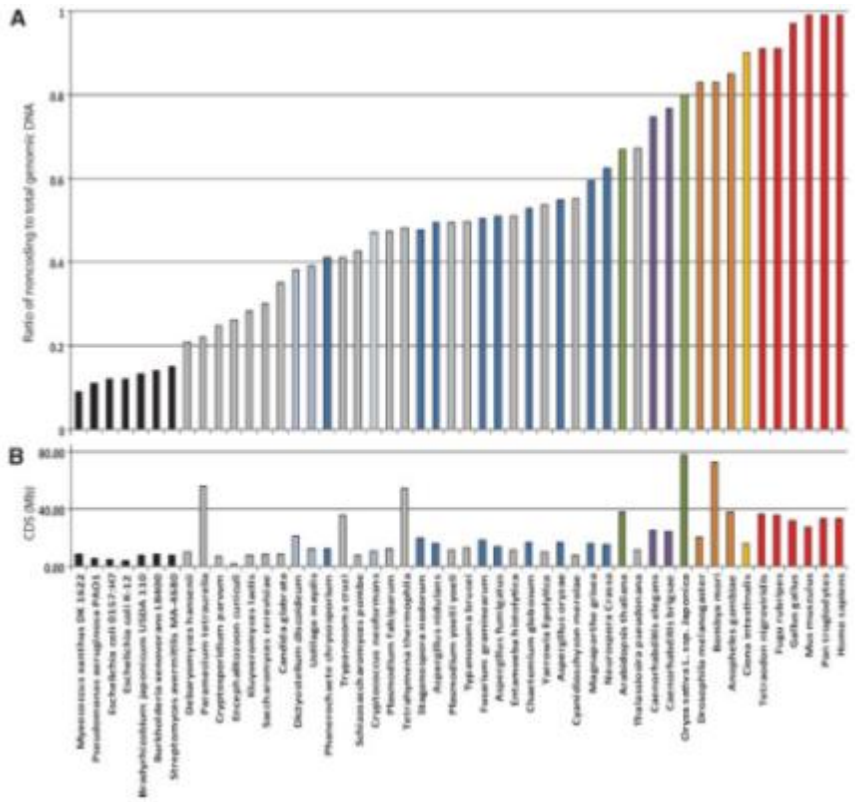
- Protein coding genes compose a small fraction of eukaryotic genomes, e.g 2% in human (Lander et al. 2001)

- Messenger RNA only accounts for less then 5% of total RNA in a single cell (Kampers et al. 1996)

- 92%-95% of human genes undergo alternative processing (Pan et al. 2008; Wang et al. 2008)
  - Alternative promoter
  - Alternative splicing
  - Alternative polyadenylation

# Eukaryotic Genome Complexity

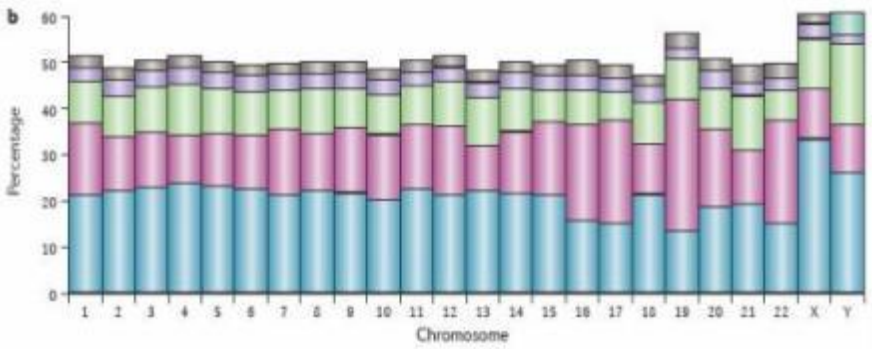
Correlation of non-coding regions and organism complexity




Taft et al. 2007

Approximately **42%** of the human genome is comprised of repeats

| Repeat class | Repeat type | Number (hg19) | Cvg | Length (bp) |
|---|---|---|---|---|
| Minisatellite, microsatellite or satellite | Tandem | 426,918 | 3% | 2–100 |
| SINE | Interspersed | 1,797,575 | 15% | 100–300 |
| DNA transposon | Interspersed | 463,776 | 3% | 200–2,000 |
| LTR retrotransposon | Interspersed | 718,125 | 9% | 200–5,000 |
| LINE | Interspersed | 1,506,845 | 21% | 500–8,000 |
| rDNA (16S, 18S, 5.8S and 28S) | Tandem | 698 | 0.01% | 2,000–43,000 |
| Segmental duplications and other classes | Tandem or interspersed | 2,270 | 0.20% | 1,000–100,000 |



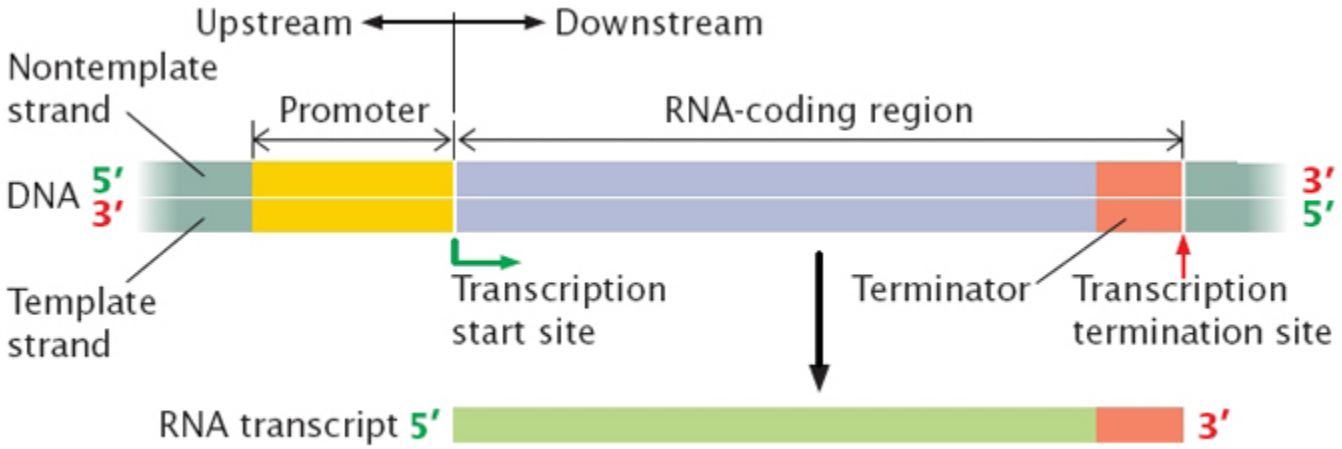

Treangen T.J. and Salzberg S.L, 2012

# Messenger RNAs (mRNAs)

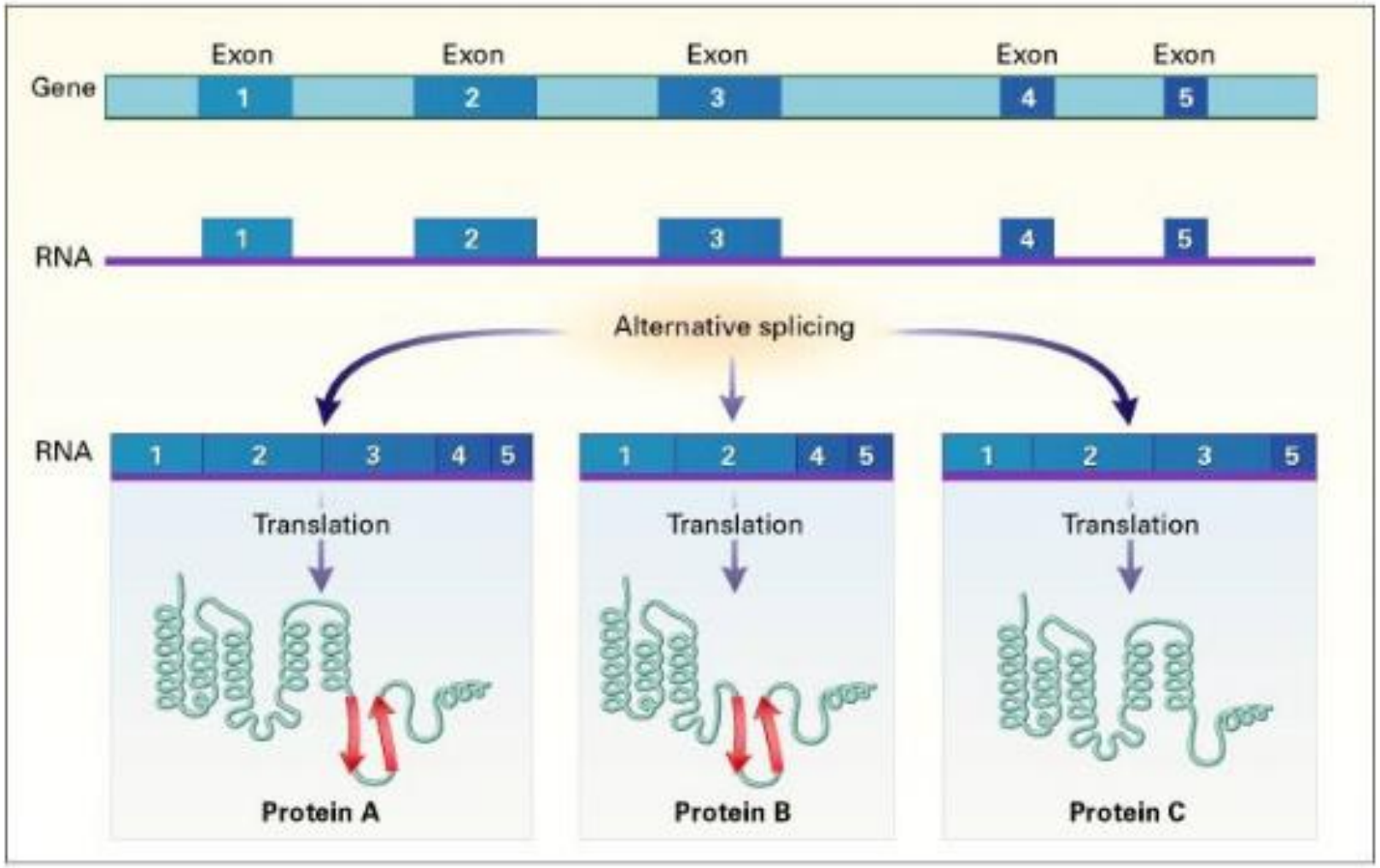
**Messenger RNAs (mRNAs)** are functional transcripts that are translated into proteins.

## Processing of mRNA :

- 5' capping
- Splicing
- Editing
- Polyadenylation



# mRNA splicing



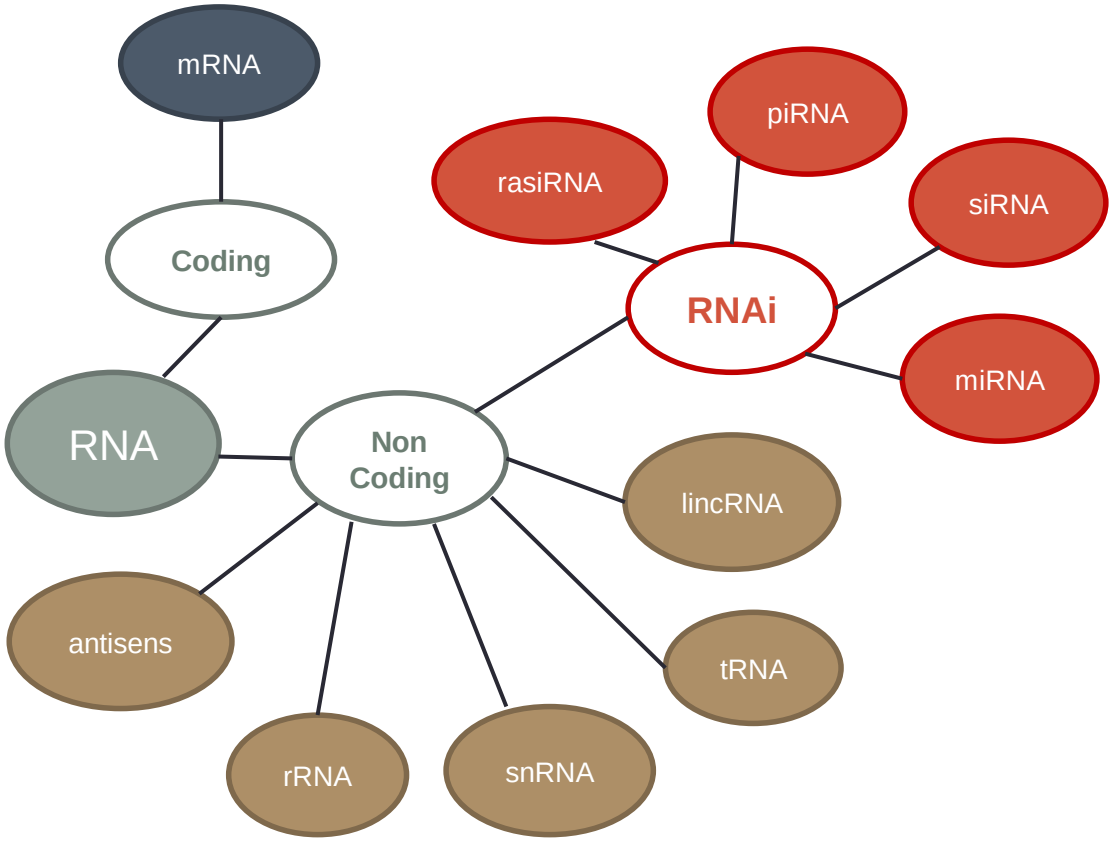

Guttmacher and Collins, 2002

# Non Coding RNAs (ncRNAs)

**Noncoding RNAs (ncRNAs)** are functional transcripts that are **not** translated into proteins, i.e., that always exist in the form of RNA during their lifespan.

## Functions of ncRNAs :

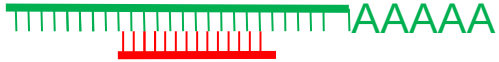
- Protein synthesis: rRNA, tRNA, …
- Post-transcriptional modification: snRNA, snoRNA, …
- Regulatory: miRNA, siRNA, piRNA, …
- …



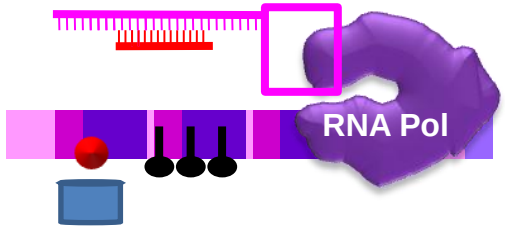
# What are small RNAs ?

• Small RNAs are a pool of 21 to 24 nt RNAs that generally function in **gene silencing**

• Small RNAs contribute to **post-transcriptional gene silencing** by affecting mRNA translation or stability



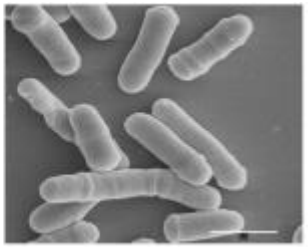

• Small RNAs contribute to **transcriptional gene silencing** through epigenetic modifications to chromatin




Histone modification, DNA methylation

# RNAi proteins in different organisms

| | Human | Mice | Drosophila | C. Elegans | Pombe | A. Thaliana |
|---|---|---|---|---|---|---|
| Dicer | 1 | 1 | 2 | 1 | 1 | 4 |
| Argonaute | 4+4 | 4+3 | 3 | 27 | 1 | 10 |



# Small RNAs as ubiquitous regulators of gene expression in Mice

**MicroRNA (miRNA):**

- endogenously expressed as stem loop structured precursors
- could be processed independently of Drosha by splicing event (miRtron)
- processed to **22-23 nt** mature miRNA
- post-transcriptional regulation of transcripts from a wide range of genes

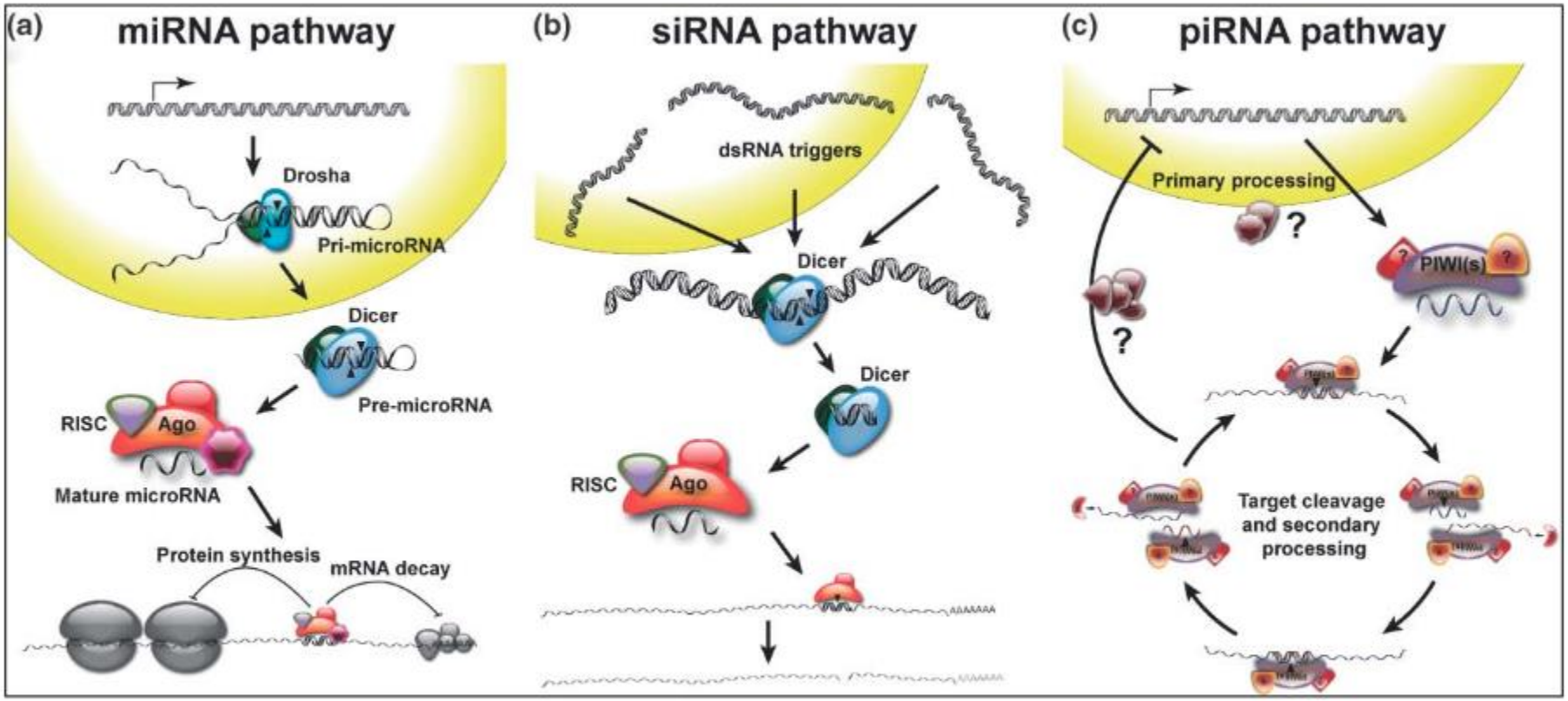
**Piwi-associated RNA (piRNA, rasiRNA)**

- biogenesis is Argonaute dependent but Dicer independent
- processed to **26-30 nt** RNA
- suppression of transposons and retro-elements in the germ lines of flies and mammals
- regulation of gene expression in mice oocytes

**Endogenous siRNA (endo-siRNA):**

- processed to **21-23 nt**
- derived from piRNA clusters in oocytes of mice
- regulation of gene expression

# Small RNAs pathways




(Roca & Karginov, 2012)

# High-Throughput Sequencing and RNAs

# Interest of RNA sequencing

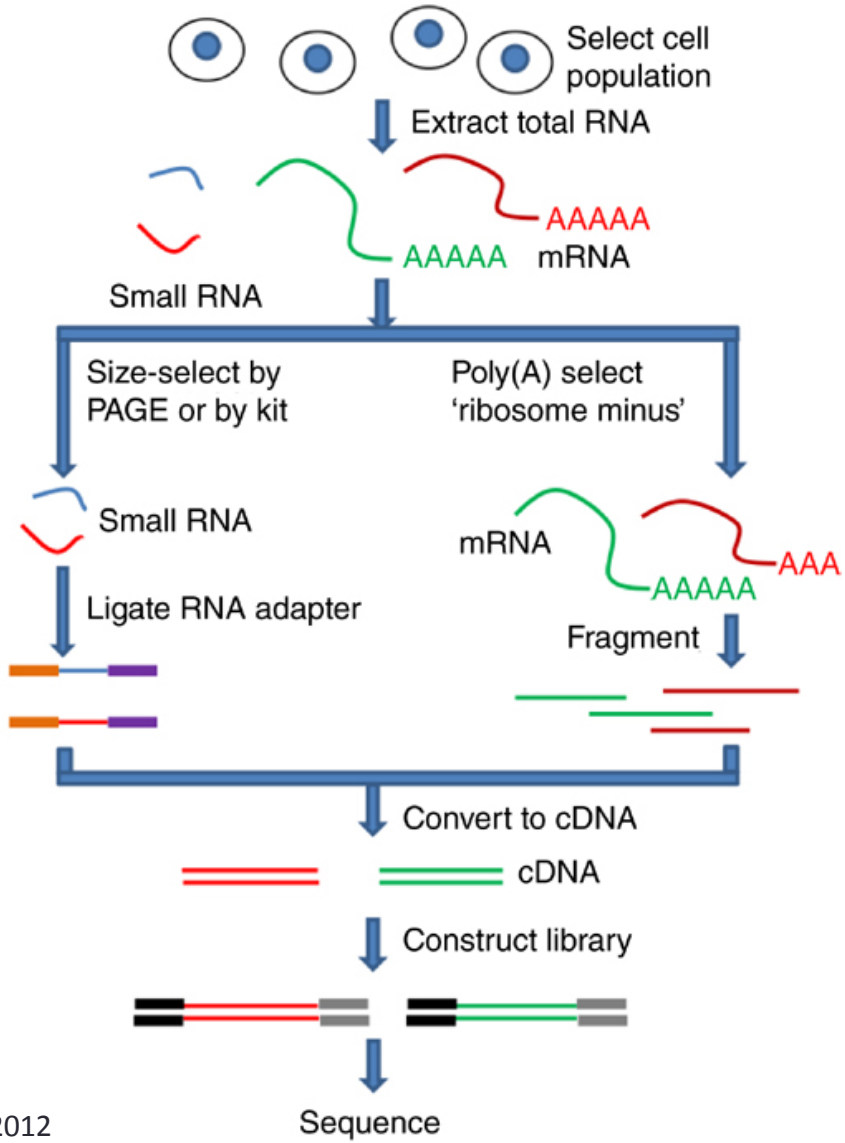
## RNA-seq

- Abundance of mRNA and differential Expression at the gene/transcripts level
- Detection of transcripts
- Detection of fusion genes
- Mutation calling (editing, monoallelic expression)
- …

## sRNA-seq

- Abundance of ncRNAs and differential expression
- Enrichment of ncRNAs
- Profiling of ncRNAs families
- De novo prediction of miRNAs (miRDeep)
- Detection of piRNAs
- …

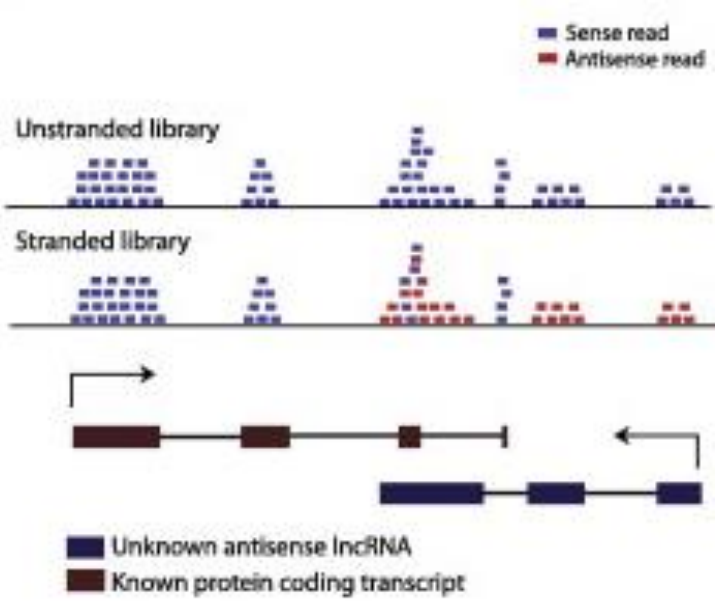
# RNA vs sRNA sequencing



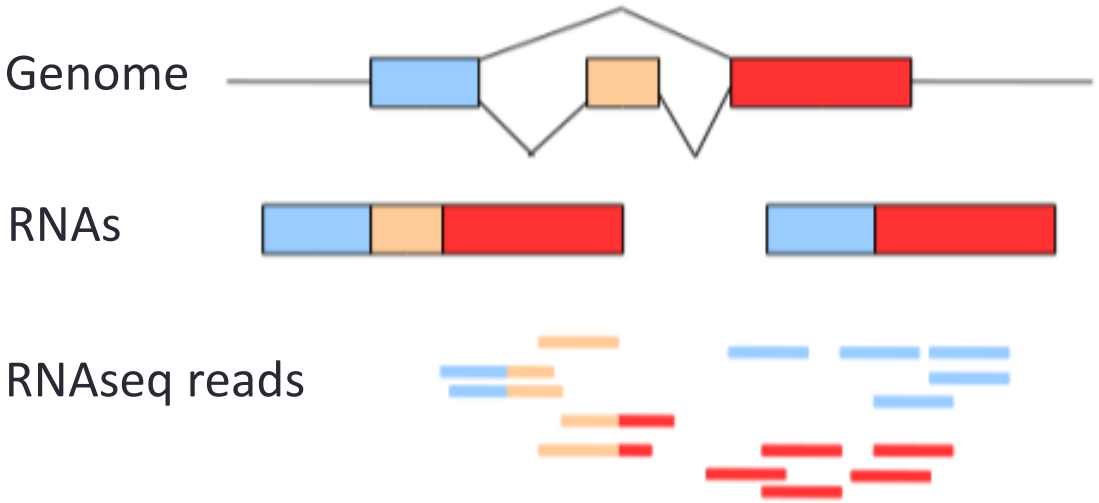

Adapted from Weihua Zeng & Ali Mortazavi. 2012

# RNA stranded sequencing

Most of the RNA-seq protocol are now directional.
This information is important for the downstream analysis.

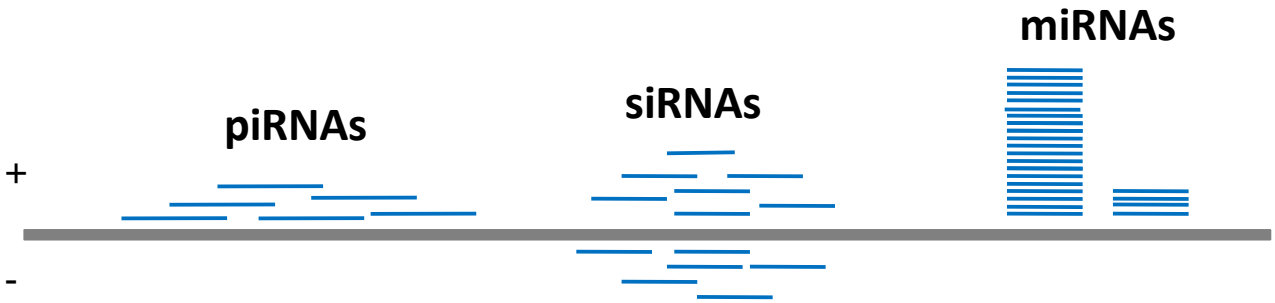


# RNA-seq : What do we expect ?




Two types of reads :
- Exonic reads
- Splicing reads

# sRNA-seq : What do we expect ?




| | miRNAs | siRNAs | piRNAs |
|---|---|---|---|
| **Size** | 22/23 | 21 | 25/30 |
| **Strand** | One strand | Both strand | One strand |
| **Distribution** | Unique location | cluster | cluster |
| **Sequence** | *U1(*)* | *(*)* | *U1(*)A10(*)* |

# Sequencing protocol and sRNA-seq

Comparison of small-RNA sequencing libraries generated **from the same embryonic stem cell lines, using different sequencing platforms, or library preparation protocols**.

OPEN ACCESS Freely available online PLoS one

## Deep-Sequencing Protocols Influence the Results Obtained in Small-RNA Sequencing

**Joern Toedling[1,2,3,4,5☯], Nicolas Servant[1,2,3*☯], Constance Ciaudo[1,4,5,6*☯], Laurent Farinelli[7], Olivier Voinnet[6,8], Edith Heard[1,4,5¶], Emmanuel Barillot[1,2,3¶]**

**1** Institut Curie, Paris, France, **2** INSERM U900, Paris, France, **3** Mines ParisTech, Fontainebleau, France, **4** CNRS UMR3215, Paris, France, **5** INSERM U934, Paris, France, **6** Department of Biology, Swiss Federal Institute of Technology Zürich, Zürich, Switzerland, **7** Fasteris, Plan-les-Ouates, Switzerland, **8** Institut de Biologie Moléculaire des Plantes, CNRS UPR2357 – Université Louis Pasteur, Strasbourg, France

# Sequencing protocol and sRNA-seq

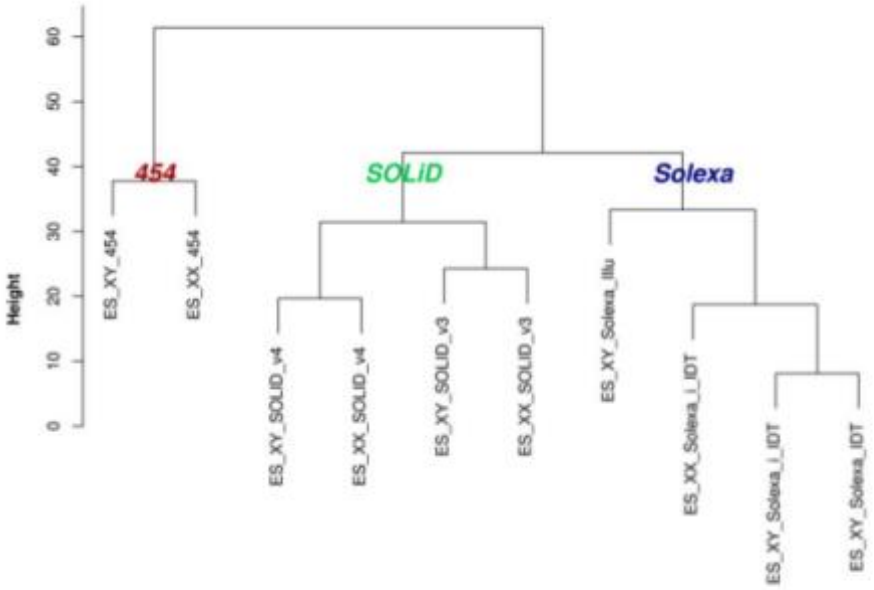



**Figure 5. Clustering of sRNA-seq librairies.** Hierarchical clustering dendrogram visualising the pair-wise distances between the 10 libraries after normalisation of miRNA read counts. The library identifiers correspond to the identifiers used in Table 1.
doi:10.1371/journal.pone.0032724.g005

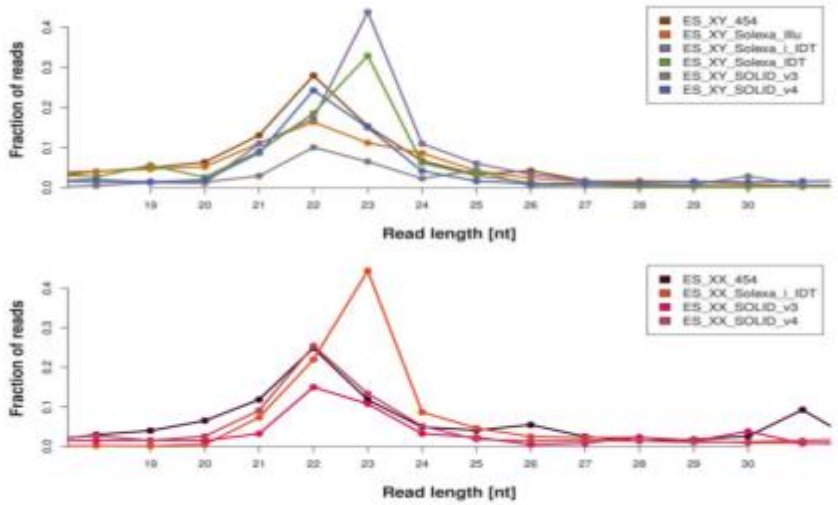



**Figure 1. Read length distributions after adapter removal.** The upper panel shows the E14 XY libraries while the lower panel displays the PGK XX libraries. See Table 1 for details about the libraries.
doi:10.1371/journal.pone.0032724.g001

Two libraries made with different protocols show more differences than two different biological samples, male versus female ES cell lines !

# Sequencing protocol and sRNA-seq

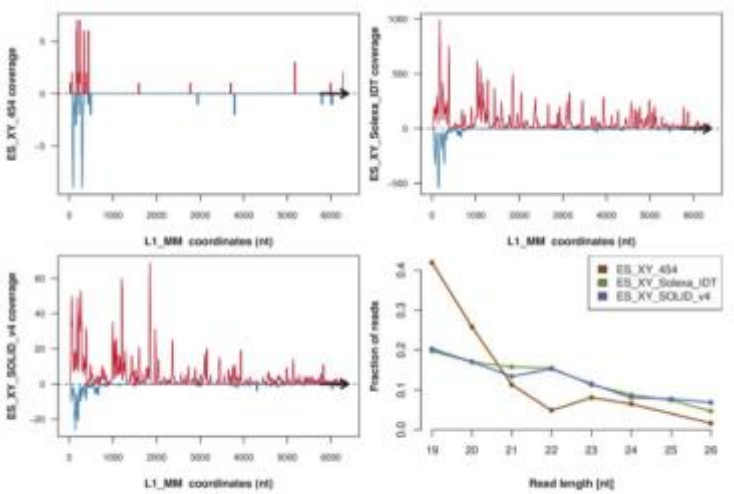
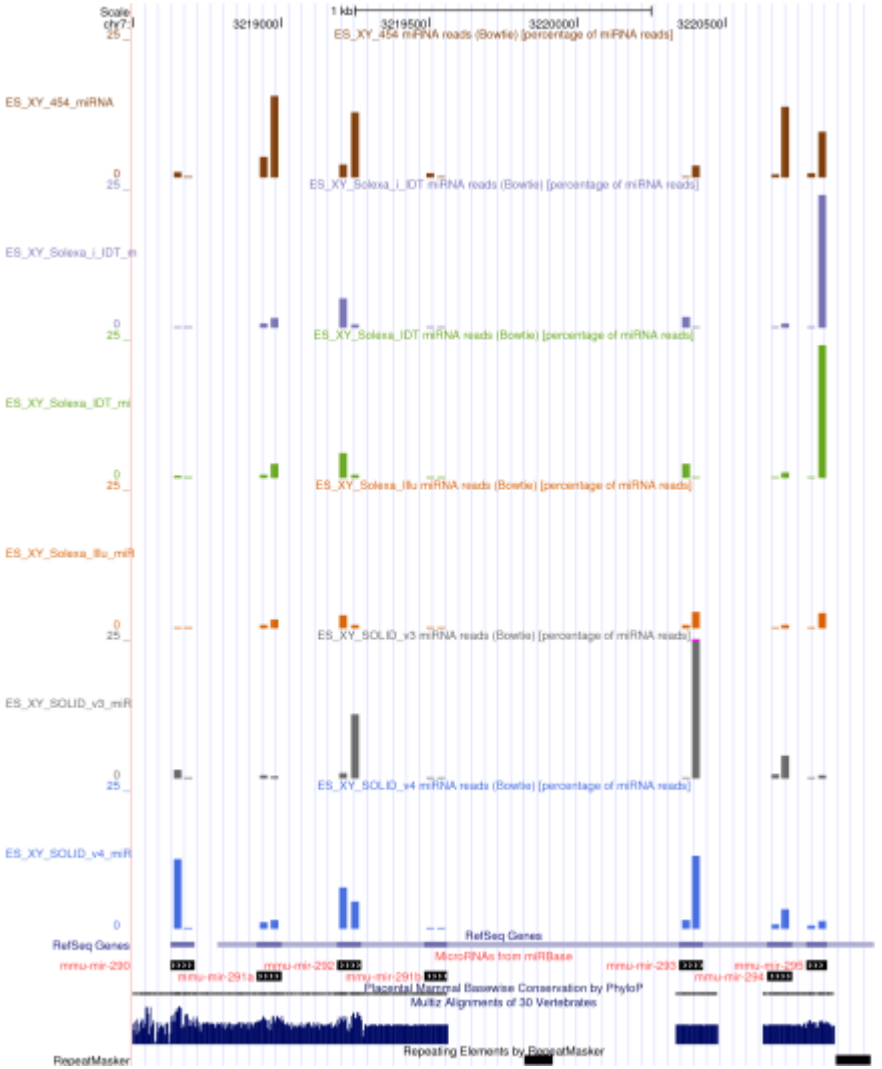



**Figure 7. Reads coverage of L1 consensus sequence.** The reads with a size of 19-26 nt were aligned on the L1_MM consensus sequence extracted from Repbase. The coverage from the ES_XY_Solexa_IDT, ES_XY_454 and ES_XY_SOLID_v4 libraries, on the sense orientation is represented in red, whereas the coverage in antisense orientation is represented in blue. The size distribution of the reads aligned on the L1_MM consensus is shown for the three libraries.
doi:10.1371/journal.pone.0032724.g007

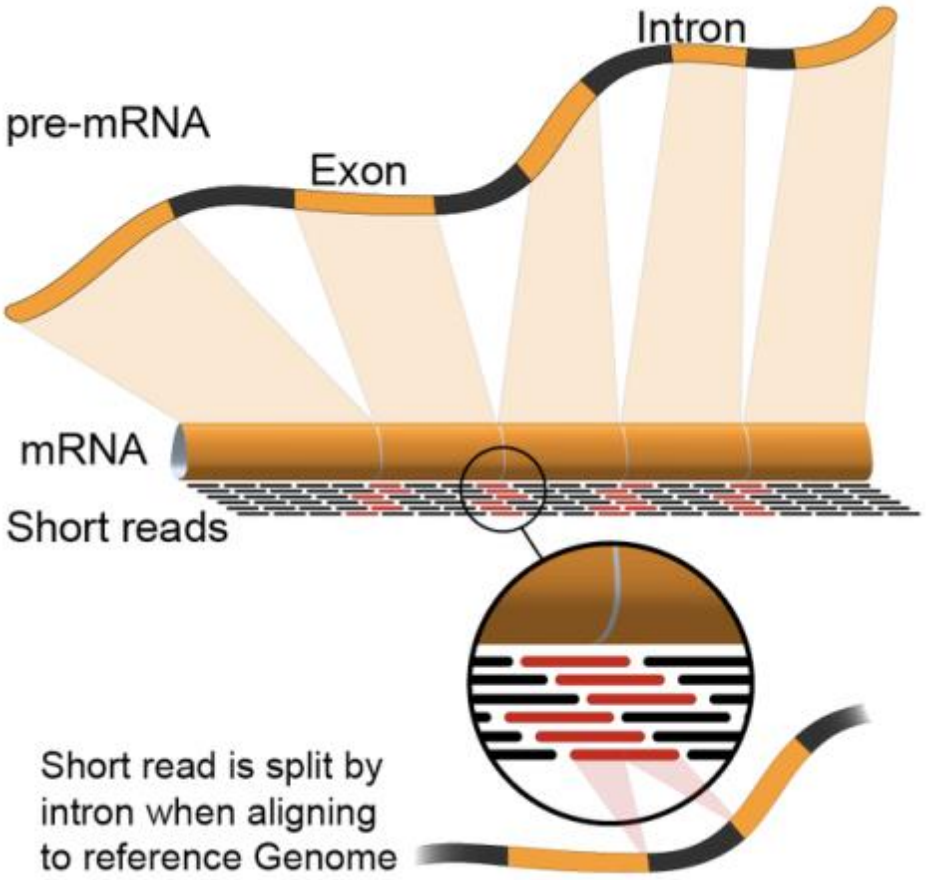



These differences are mainly true for gene-based analysis.

**Only two samples done with the same protocols at the same time can be compared !**

# RNA-seq Analysis

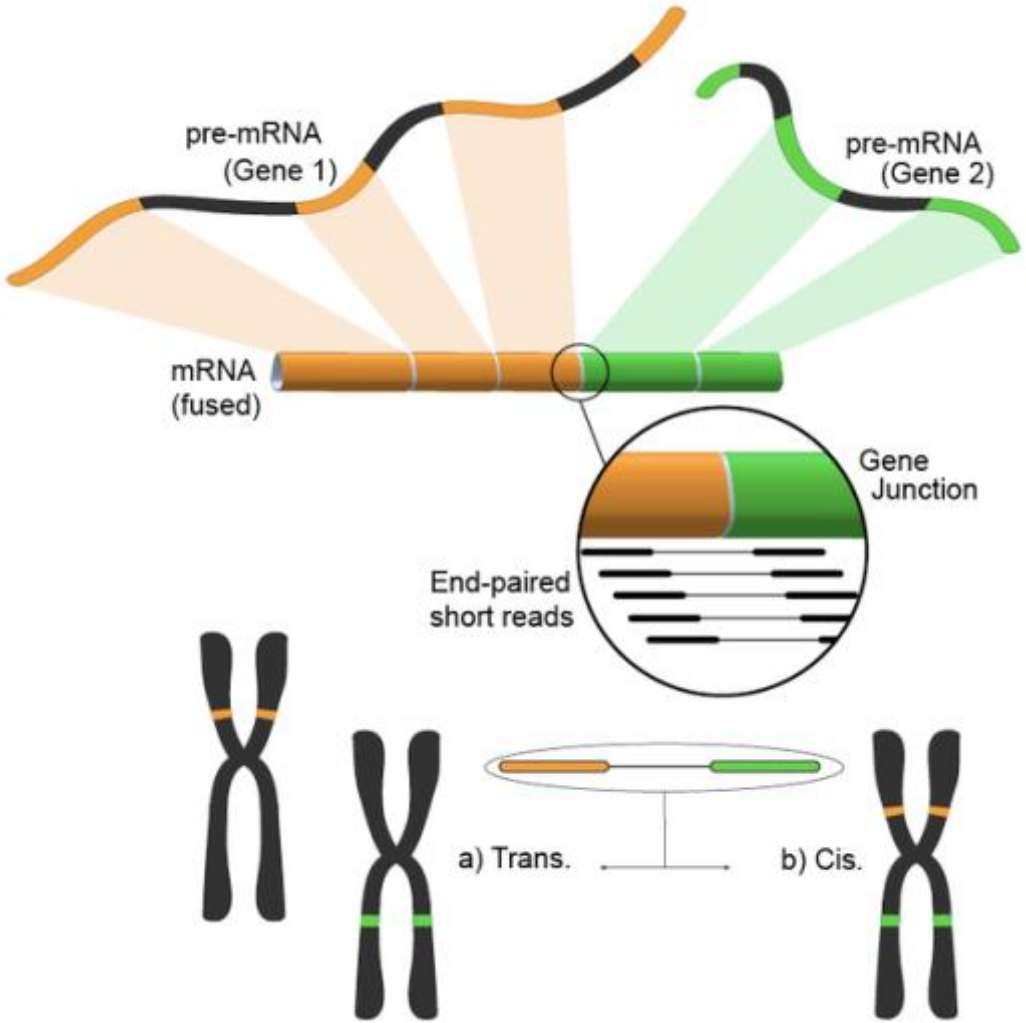
# Biological Applications



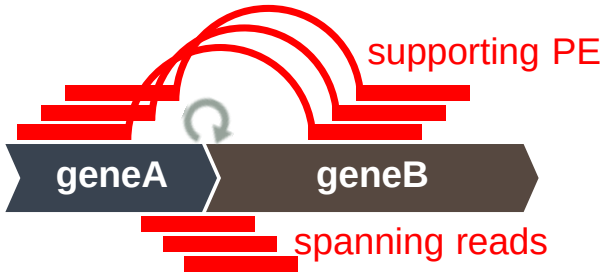

- **Gene/Transcripts expression**
- **Fusion genes**
- **Alternative splicing**
- **Novel transcripts**
- **Variant calling/Editing**
- **ncRNAs**

From wikipedia.org

# RNA-seq : Fusion genes

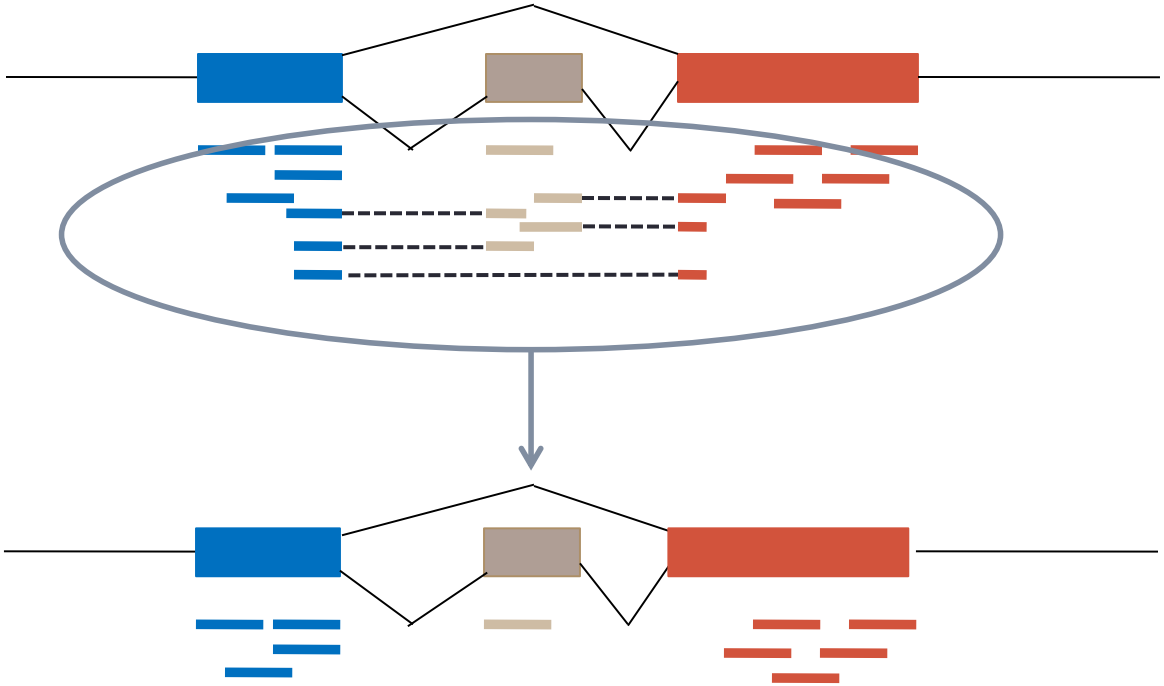



TopHat-fusion
FusionMap
DeFuse
FusionFinder
...



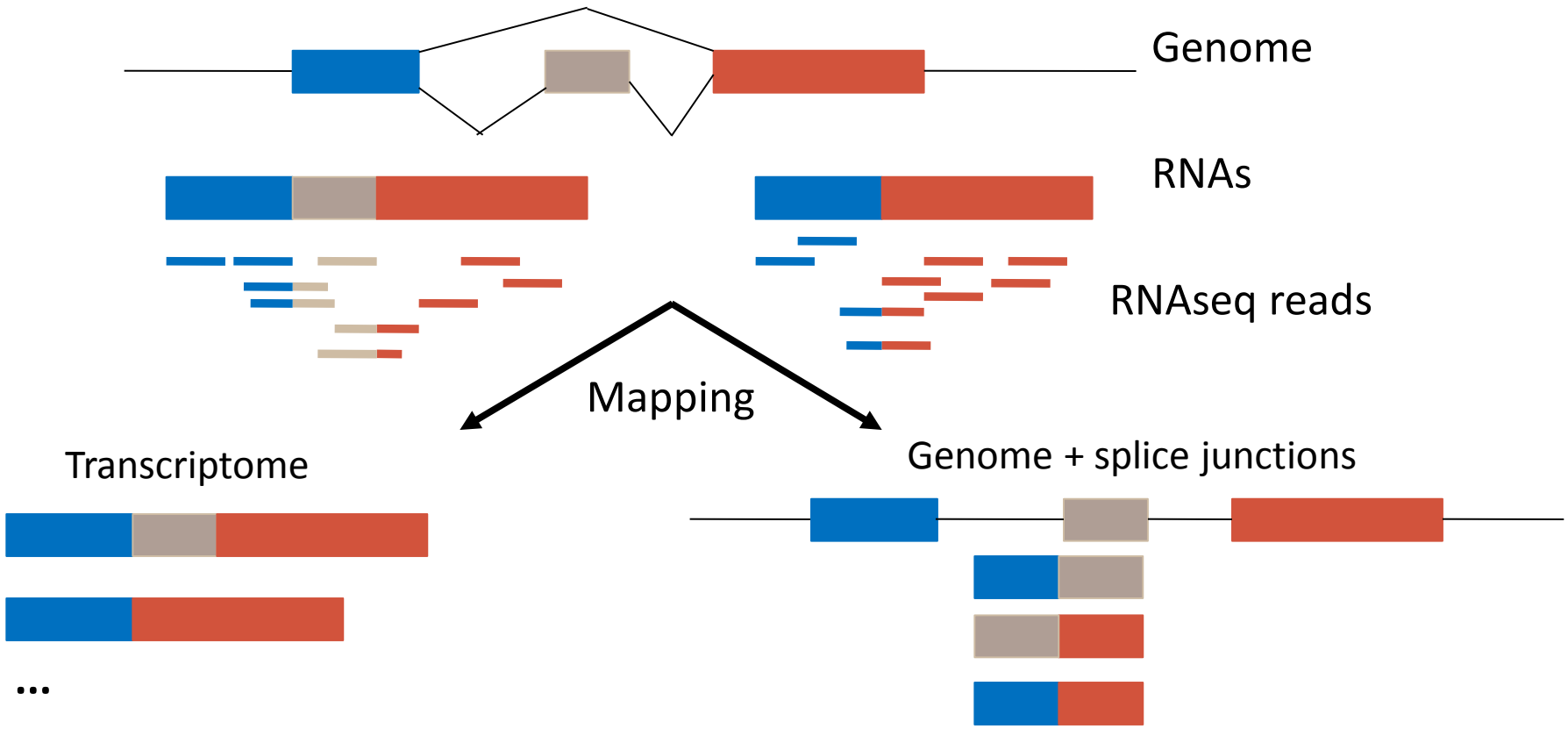

From wikipedia.org

# Challenges with RNA-seq reads

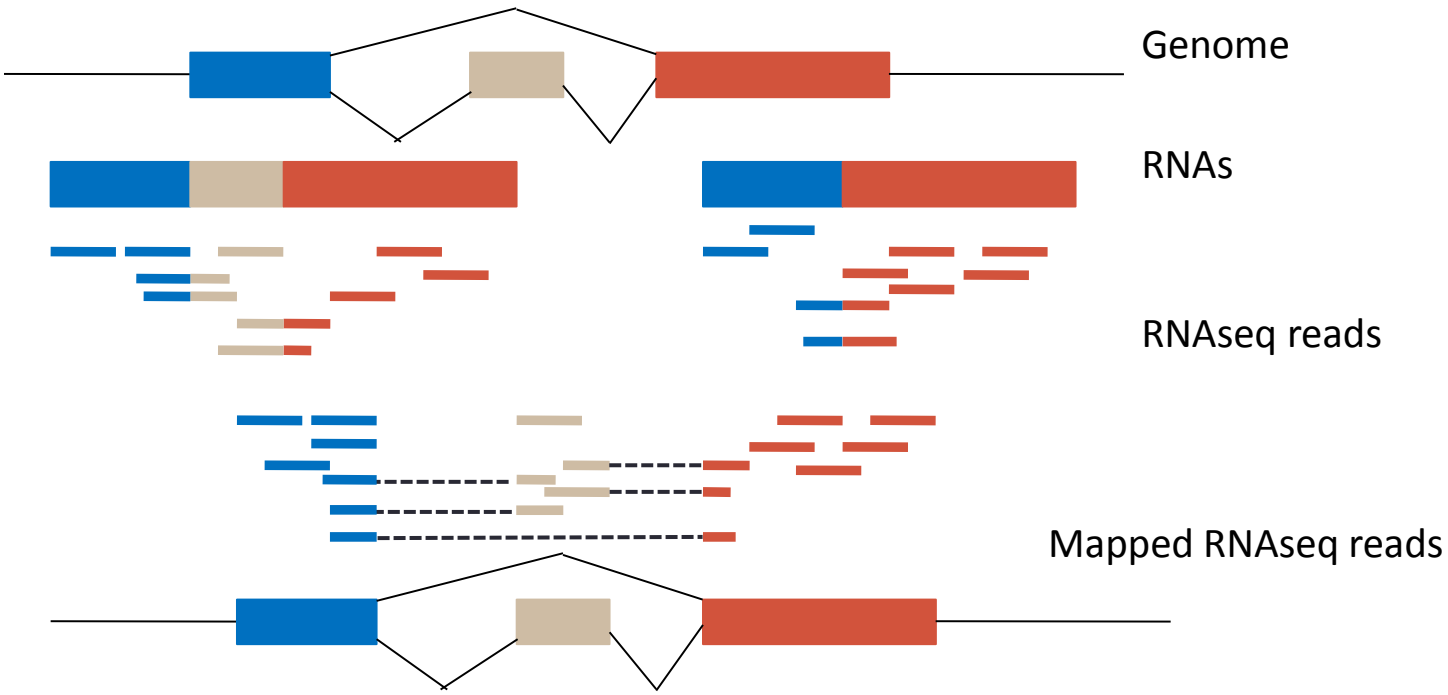


→ Needs for dedicated mapping strategy

# Unspliced alignment of RNA-seq reads



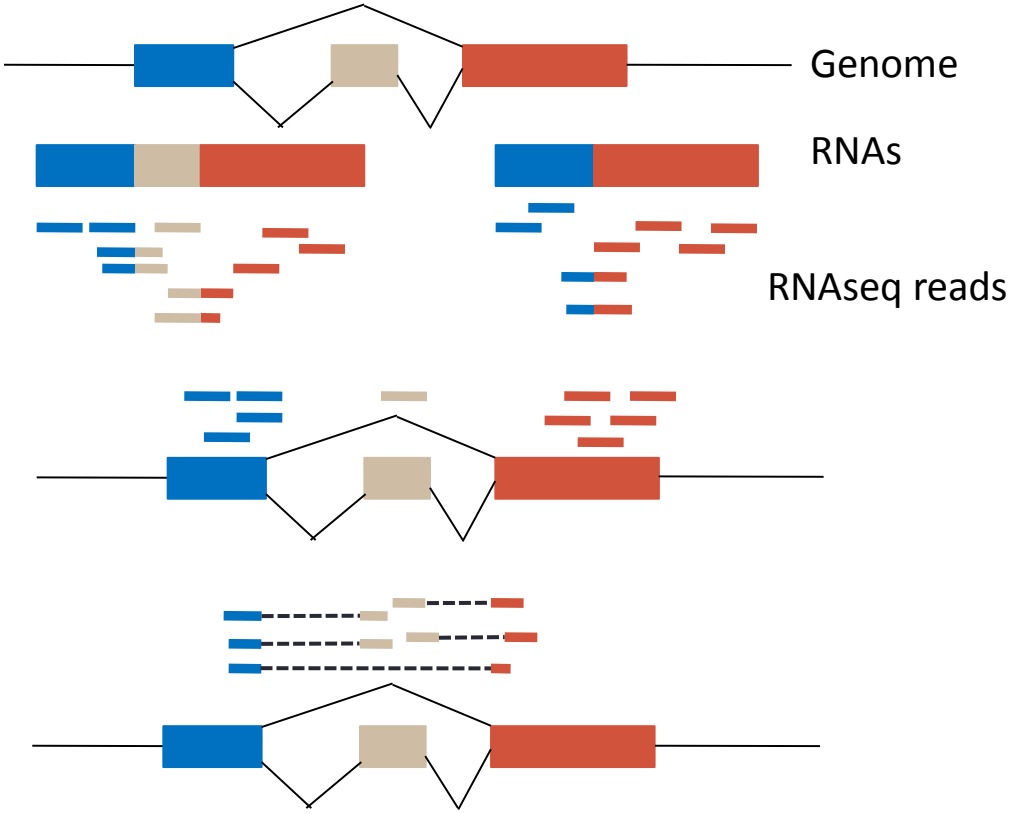
# Spliced alignment of RNA-seq reads




Two strategies:

- Exon-first approach
- Seed and extend approach

# 1- Exon-first approach

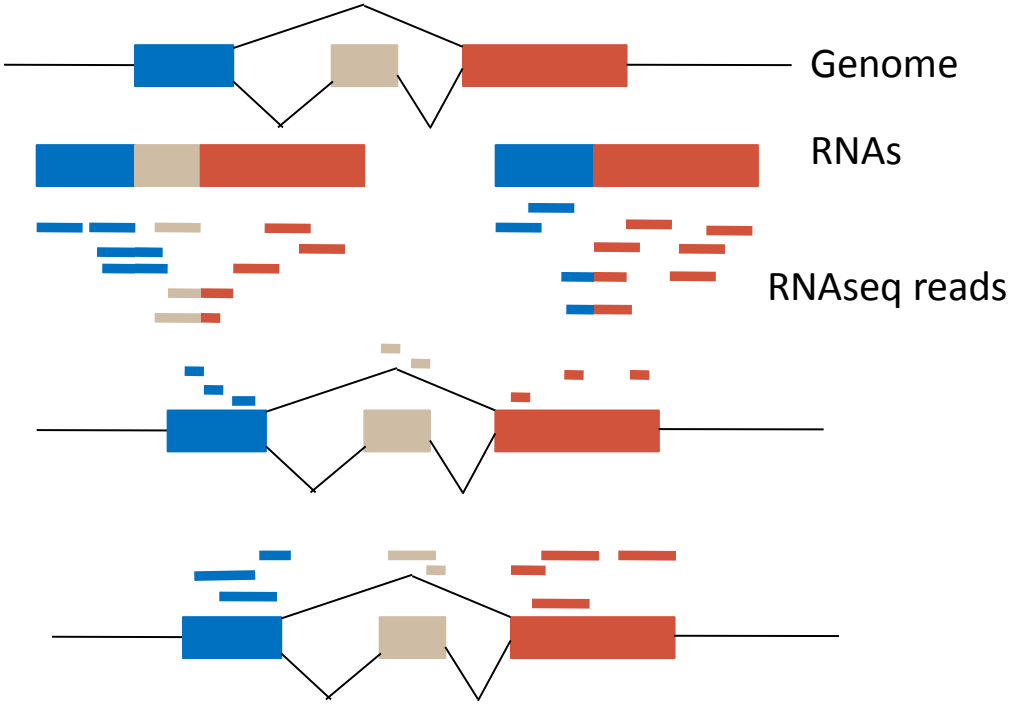



1- Exon read mapping
Map reads continuously onto the genome using an unspliced mapper

2 - Spliced read mapping
Unmapped reads are divided into shorter segments and mapped

→ **Tophat** (*Trapnell et al. Bioinformatics 2009*)

# 2- Seed-extend approach




1- Seed matching
Reads are divided into k-mers that are mapped onto the genome

2 – Seed extend
Candidate regions are examined with a more sensitive method to determine the exact alignment for the read

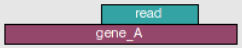
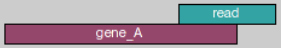
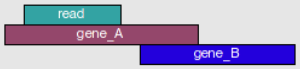
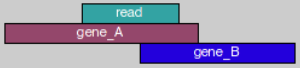
→ **GSNAP** (*Wu et al. Bioinformatics 2010*)
**STAR** (*Dobin et al. Bioinformatics 2012*)

# Quantification

**Goal:** quantify gene expression

**Htseq-count:**

- Count how many reads map to each feature (in RNA-seq, the features are typically genes)
- Input file : file with aligned sequencing reads (bam or sam file) + list of genomic feature ; gtf file

| | union | intersection_strict | intersection_nonempty |
|---|---|---|---|
|   | gene_A | gene_A | gene_A |
|   | gene_A | no_feature | gene_A |
|   | gene_A | no_feature | gene_A |
|   | gene_A | gene_A | gene_A |
|   | gene_A | gene_A | gene_A |
|   | ambiguous | gene_A | gene_A |
|   | ambiguous | ambiguous | ambiguous |

**Options :**

- mode : how you want to count reads
- stranded : protocol are strand specific
- feature type : gene or transcript

# The Cufflinks toolset

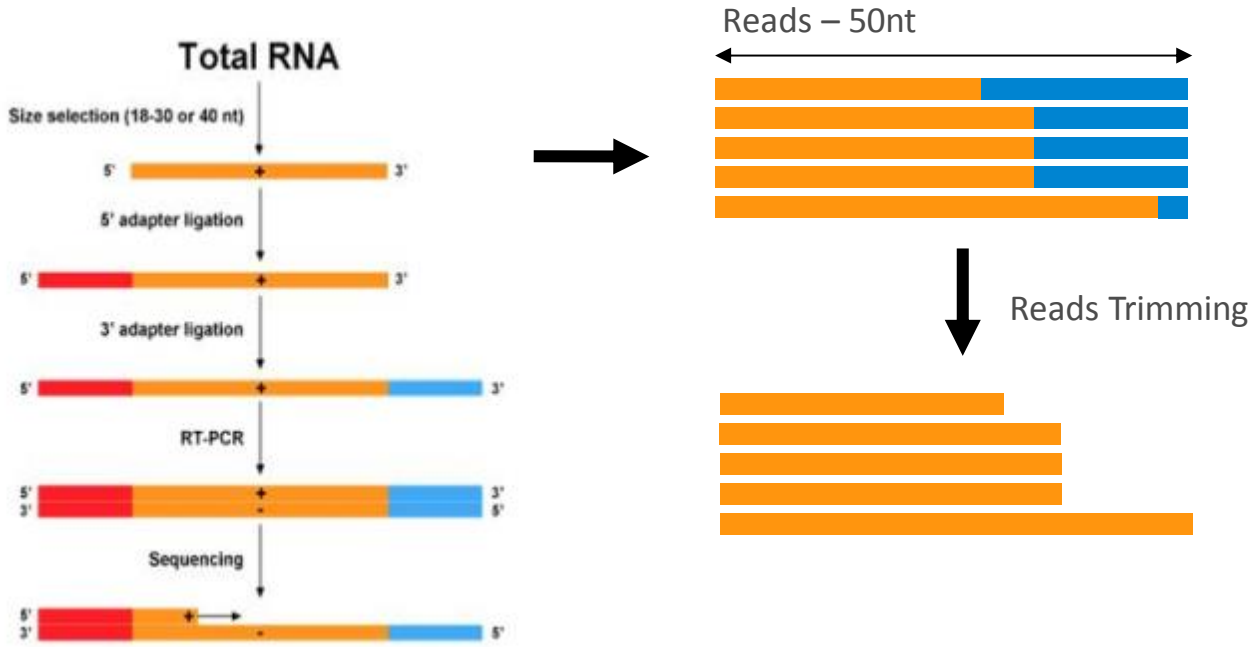
Cufflinks is a package including :

- **Cufflinks :** gene and transcript discovery and quantification
- **Cuffmerge :** merges together cufflinks assemblies and filters artefacts
- **Cuffcompare :** compares the assembled transcript with a reference, tracks cufflinks transcripts across sample
- **Cuffdiff :** differential expression analysis

# Small RNA-seq Analysis

# Adapter removal

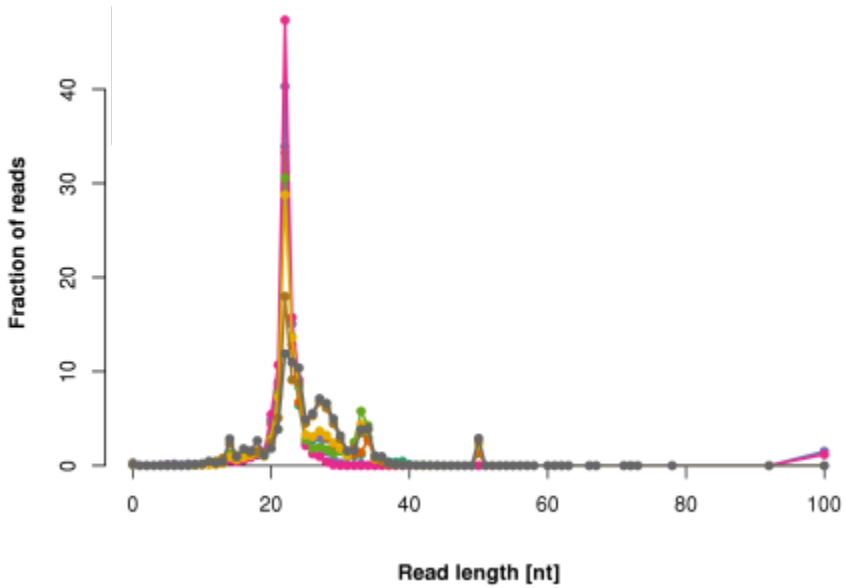
Small RNA sequencing allows the discovery and profiling of small non-coding RNAs

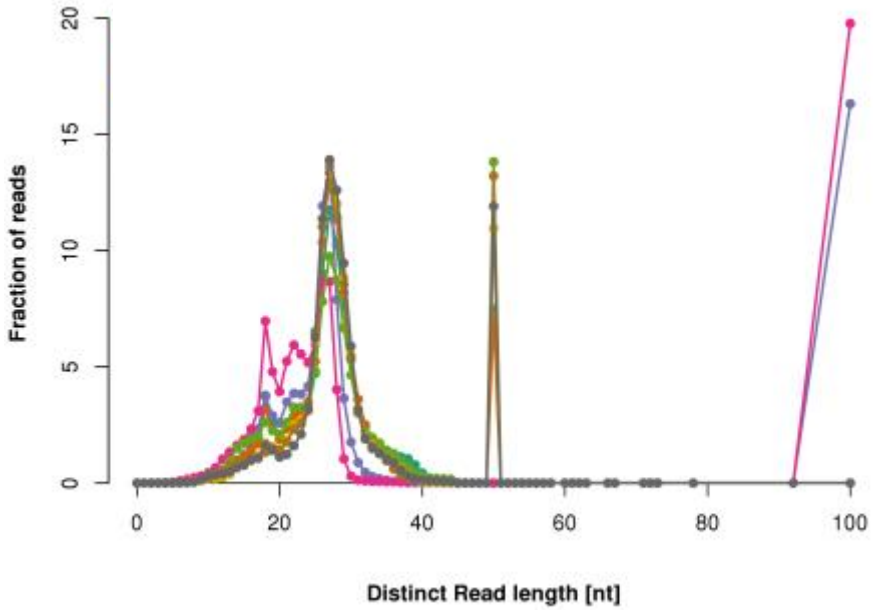


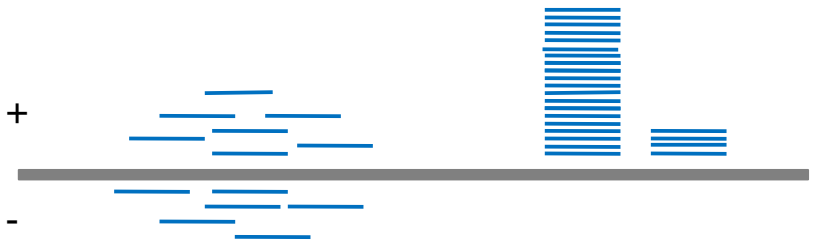

→ Check the number of reads without adapter sequences !

# QC : Insert size distribution

Looking at the insert size distribution : **abundant vs distincts** reads size distribution



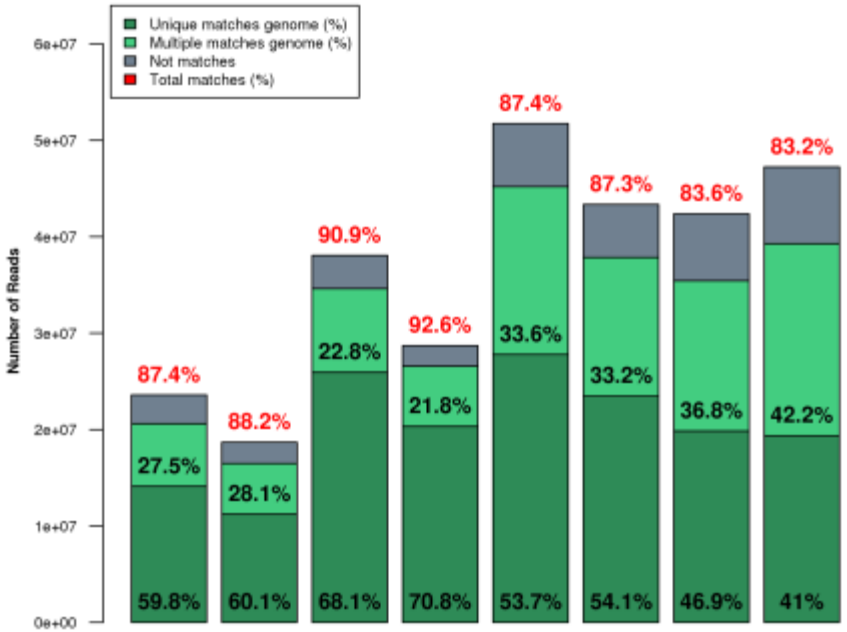



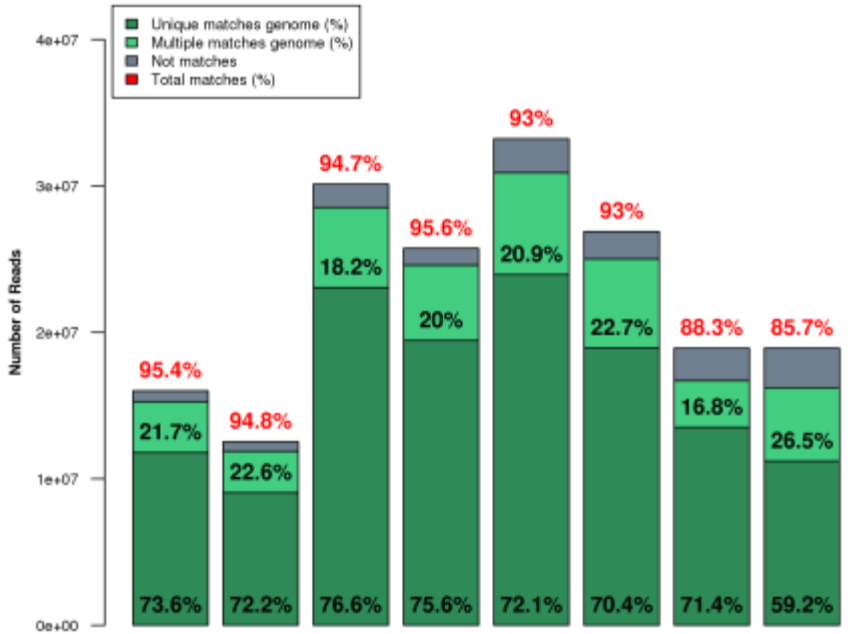

# QC: Reads Alignement

Align all reads or a subset of reads (19-24nt) on the genome

**A large proportion of unique mapped reads is expected**




All reads




19-24nt reads

# Which mapping parameters ?

Align reads on the genome using **Bowtie** (Langmead et al. 2009)

- Fast
- Global alignment
- No gap

**Number of mismatches ?**

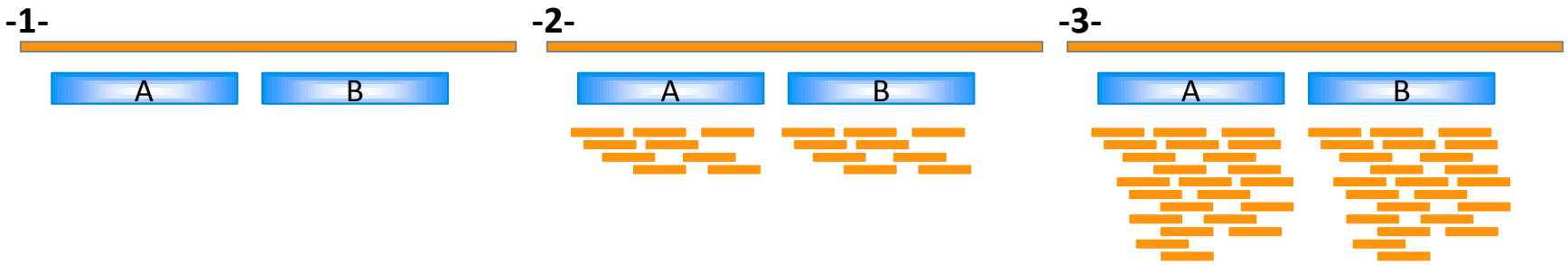
Most of studies use reads mapped with no MM. This does not allow to take into account the sequencing errors, or other expected biological polymorphism as the editing. However, many miRNAs from the same family share a high level of homogeneity.

**Number of mapping sites for a given read ?**

Most of studies used unique mapped reads. However, some miRNAs are repeted on the genome (ex : mmu-mir-16-1/mmu-mir-16-2 or mmu-mir-669a).

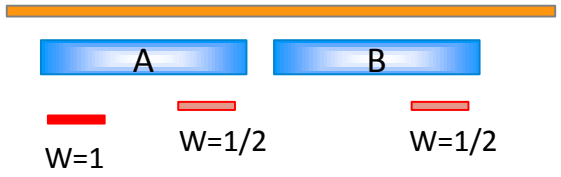
# Reads with multiple mapping sites

**-1-** Report only unique alignment
**-2-** Report best alignments and randomly assign reads across equaly good loci
**-3-** Report all (best) alignments



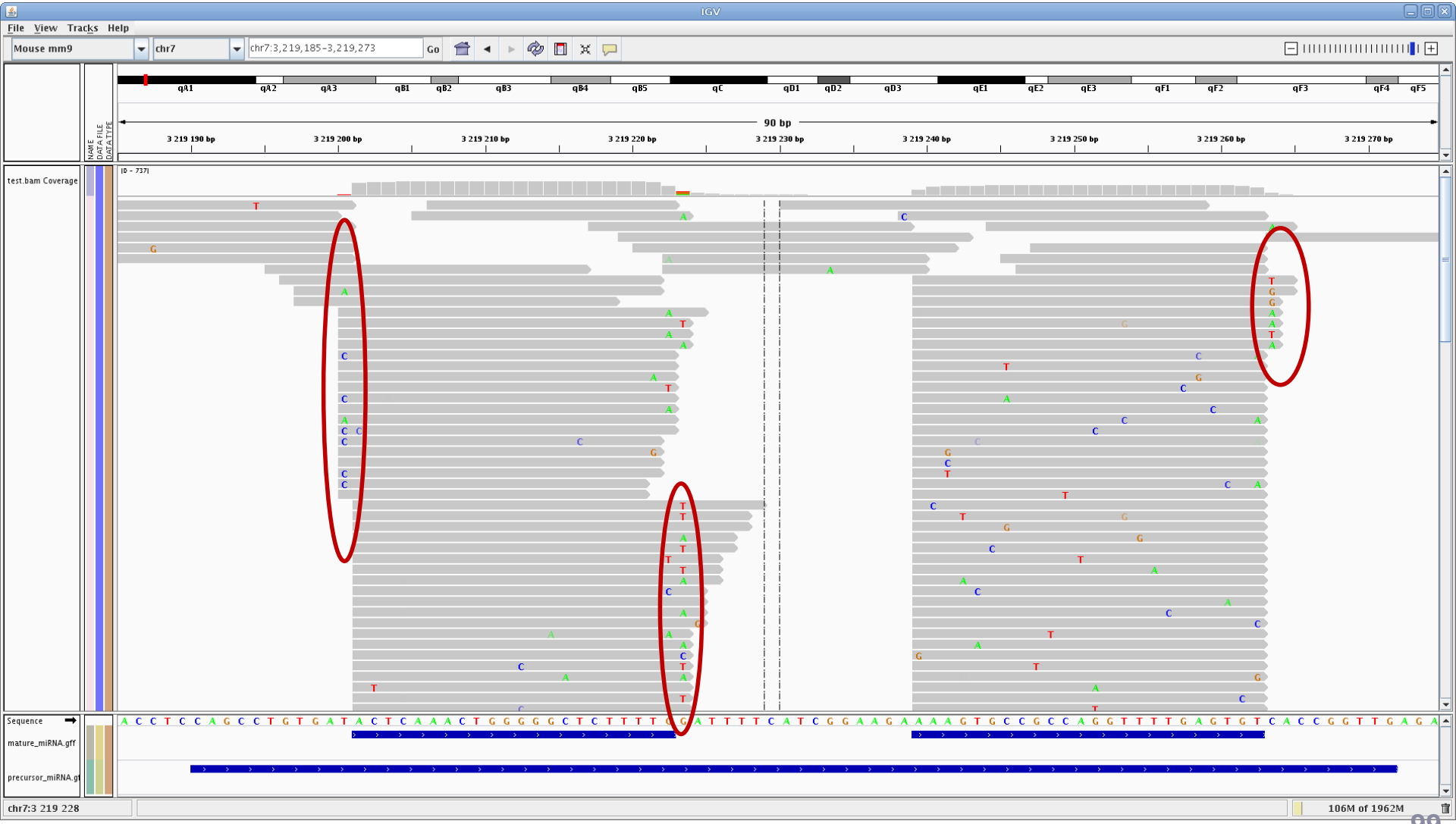

Treangen T.J. and Salzberg S.L. 2012.

**Alternative solution :** keep all best alignments and weight them using their number of mapping sites.

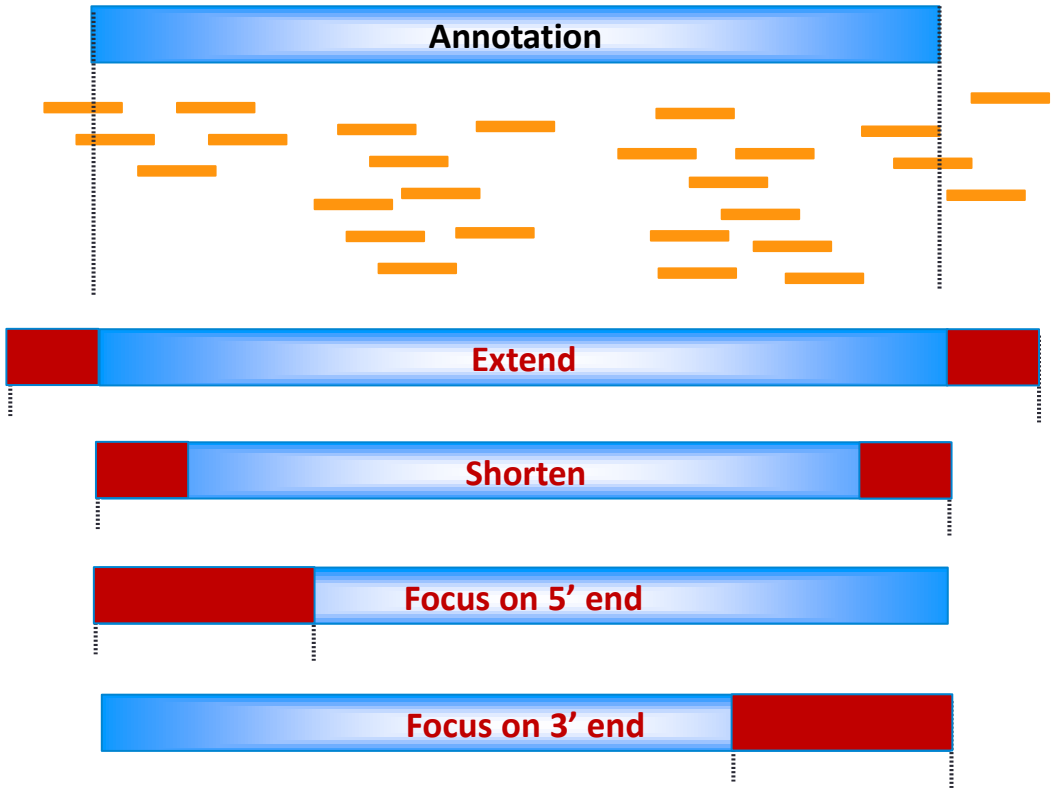


# How to annotate the reads ?

# How to annotate the reads ?



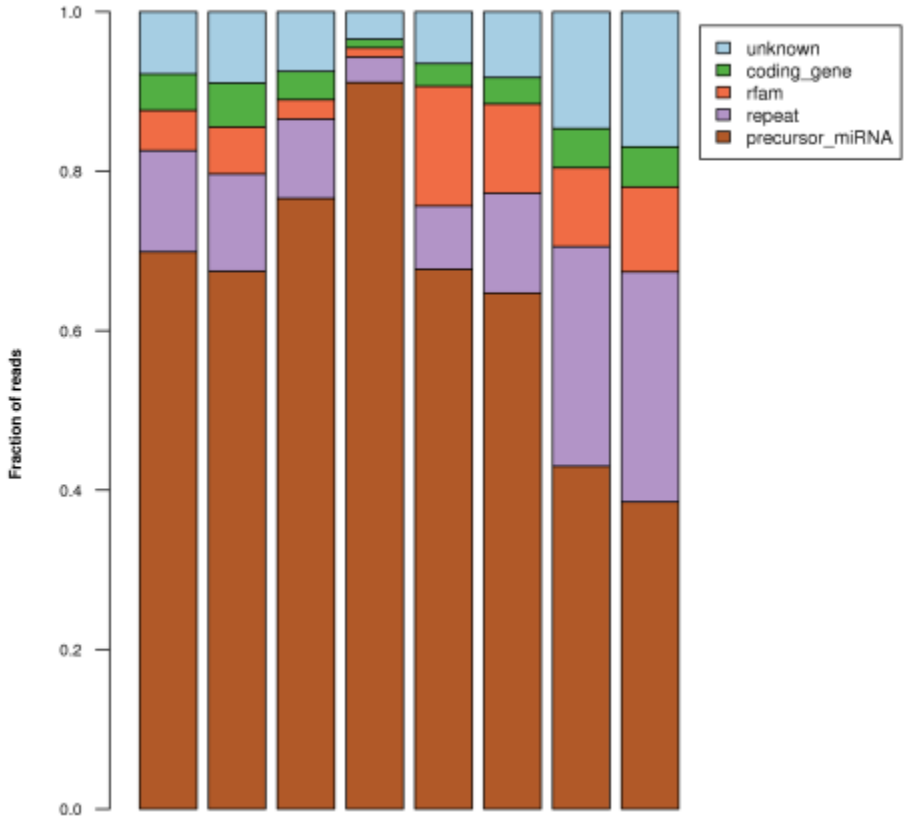
# How to annotate the reads ?



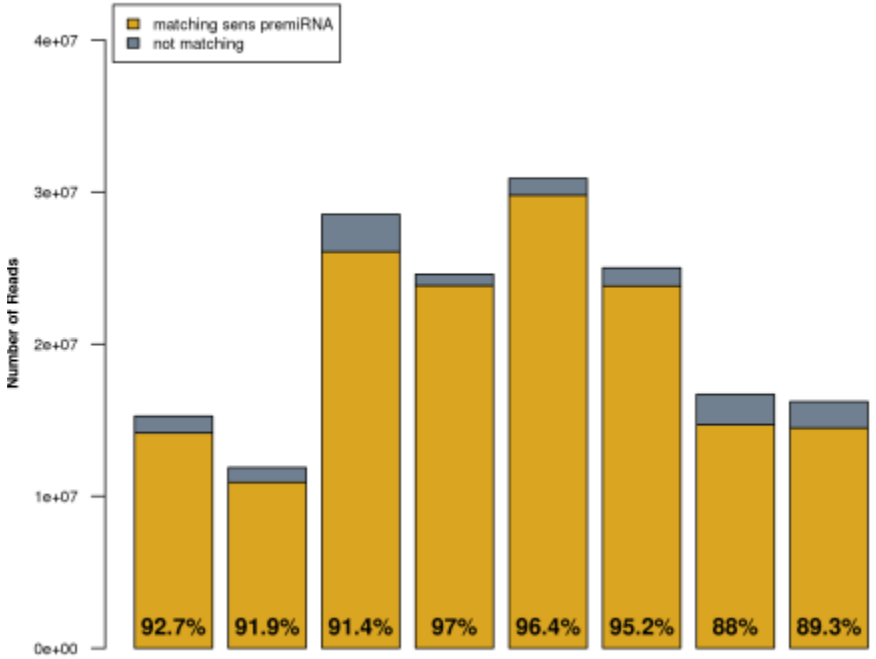

The annotation extension (+-2bp), and 100% overlap allow to be more flexible with the known annotations

# Using reads annotation as a control

Reads annotation based on standard databases (miRBase, RFAM, refseq, repeatMasker, etc.)
**A large proportion of miRNAs is expected**



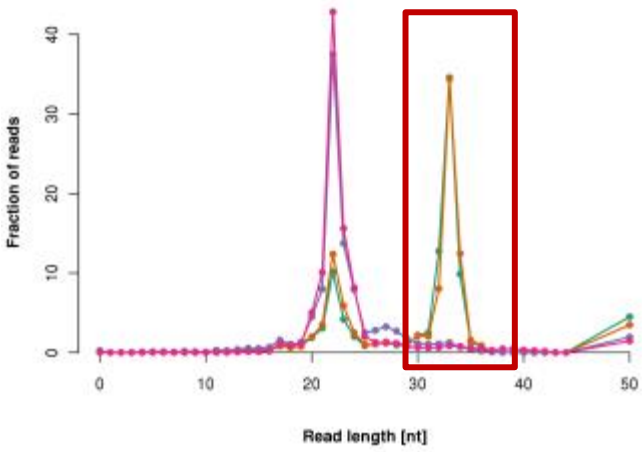

Annotation overview of all reads



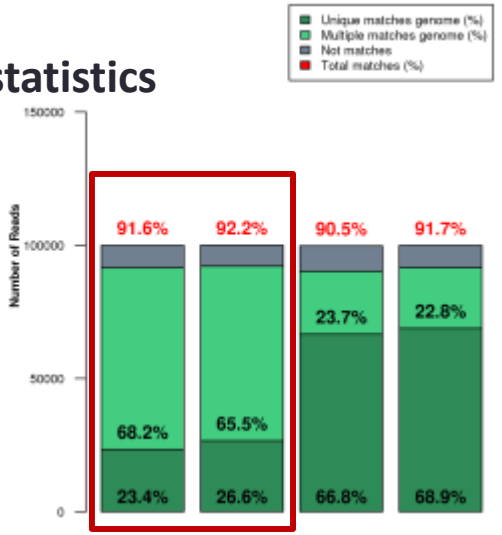

Pre-miRs annotation of 19-24nt reads

# Detecting poor quality samples

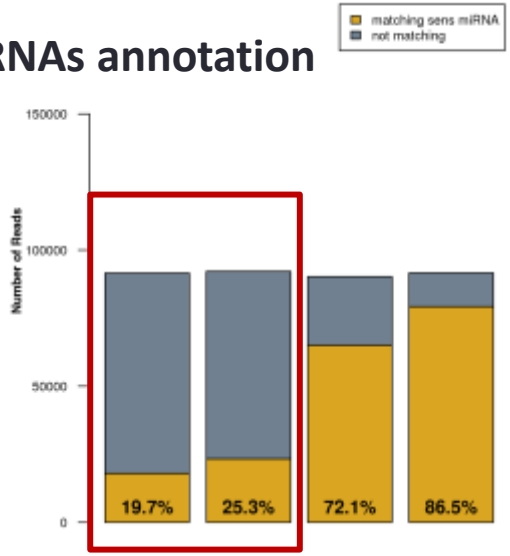
## 1. Inserts size distribution



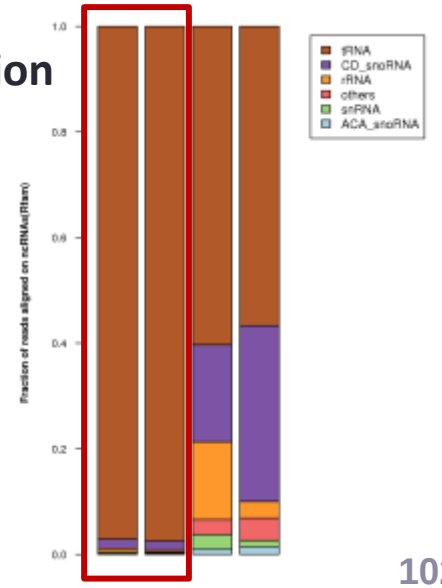

## 2. Mapping statistics




## 3. miRNAs annotation




## 4. ncRNA annotation



# ncPRO-seq (Chen, Servant et al.)

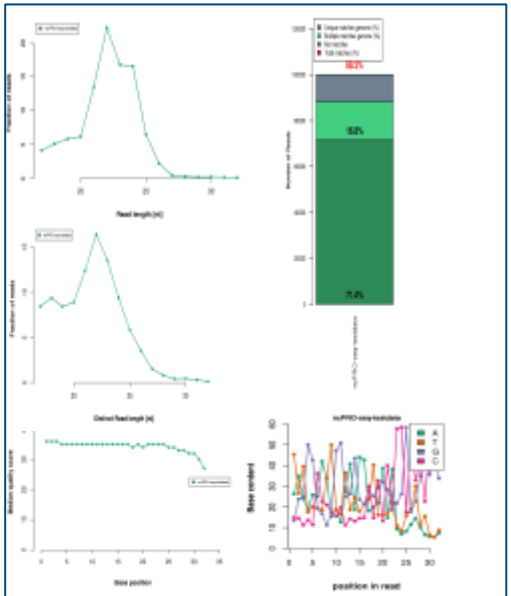
A **flexible** pipeline for **smallRNA-seq analysis**
Analysis of **SOLiD, Solexa, 454** or mapped dataset (**bam** files)
More than **fifteen** annotations files available from **plants** to **human**

→ Reads mapping (Bowtie)
→ Quality Control
→ Perform both **gene and family based analysis of ncRNAs**
→ **Repeats profiling** analysis
→ Search for **new enriched regions** in the genome
→ Data visualisation through **UCSC** genome browser
→ Dynamic view of table files
→ HTML analysis report
→ A **command line** version
→ A **local web** interface
→ An **online** ncPRO-seq version
→ A complete manual

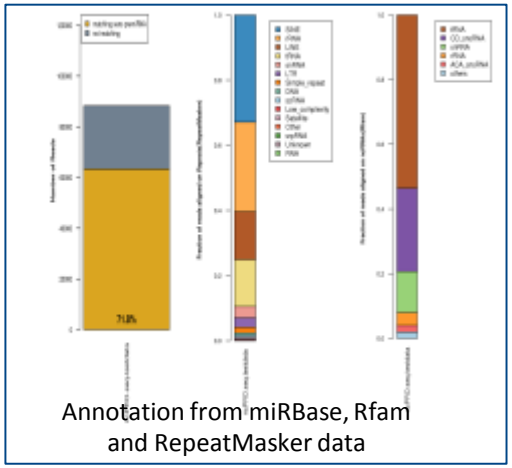
**http://ncpro.curie.fr/**

Chen CJ., Servant N. et al. ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data Bioinformatics (2012) 28(23)

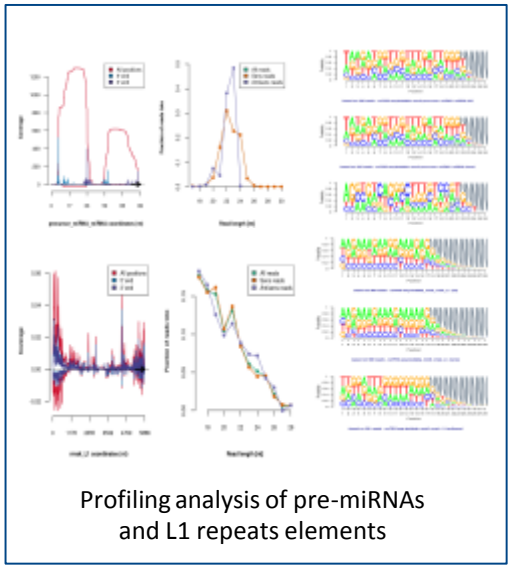
# ncPRO-seq (Chen, Servant et al.)



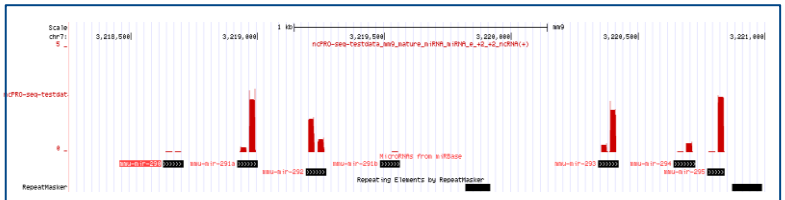
Quality Control (insert size, base composition and quality, mapping, etc.)



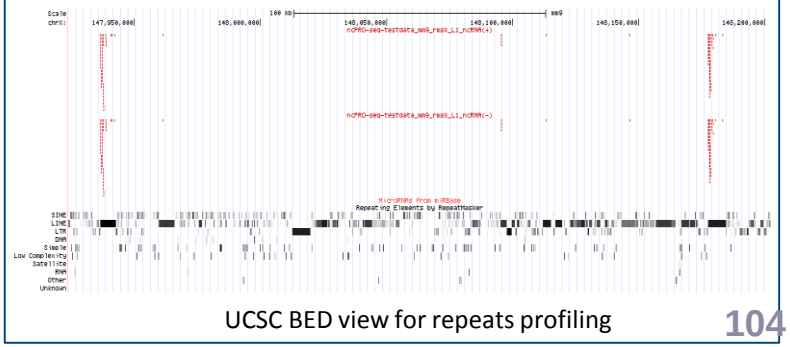
Annotation from miRBase, Rfam and RepeatMasker data



Profiling analysis of pre-miRNAs and L1 repeats elements



UCSC BEDgraph view for miRNAs



UCSC BED view for repeats profiling

Table view of annotation counts

Interactive Analysis Report

# RNA-seq - FAQ

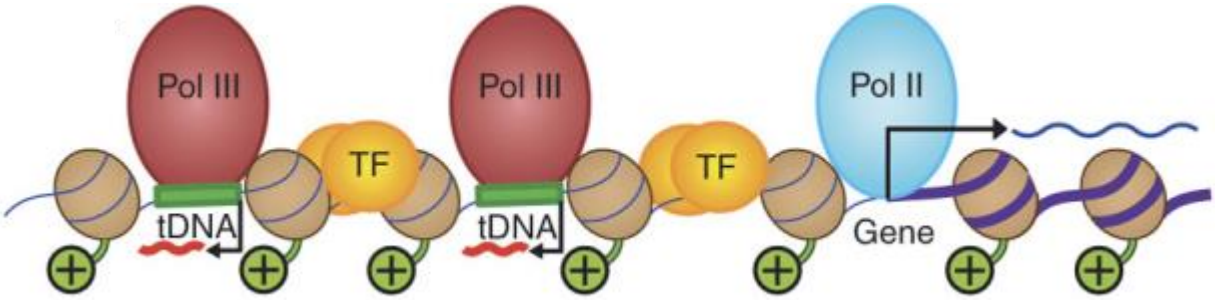
- How many reads do I need ?
- Do I need biological replicates ?
- Can I find variants in RNA-seq data ?
- What are the most expressed genes ?

institutCurie

# CHROMATINE IMMUNO-PRECIPITATION **SEQ**UENCING

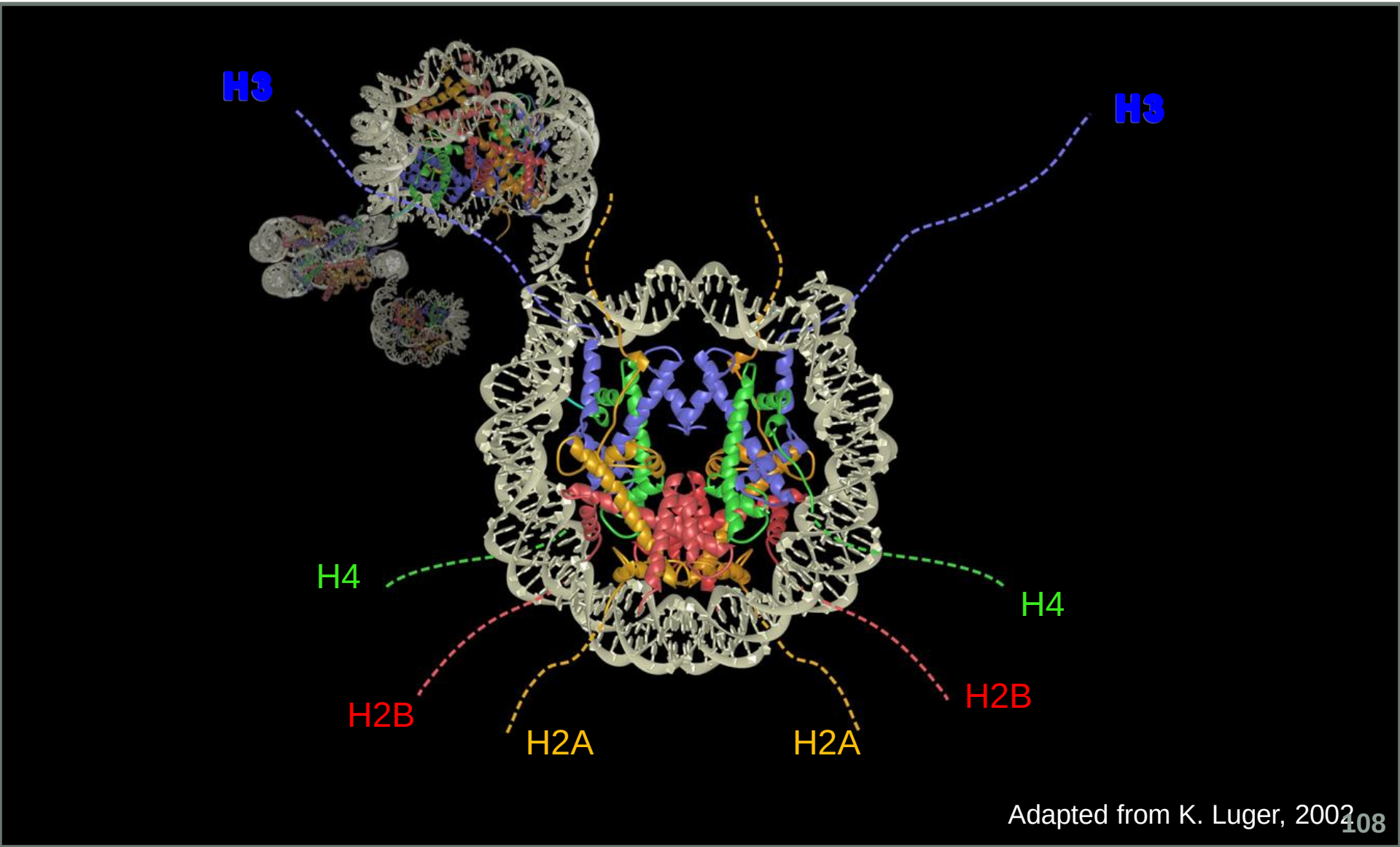
# Biological Applications

- Binding sites of transcription factors involved in regulation of cell growth, DNA repair and cell death pathways
- Histone  modifications
- Binding sites of RNA polymerases




From Oler et al., Nat Struct & Mol Biol 17, 620–8 (2010)

# The nucleosome : chromatin basic unit




Adapted from K. Luger, 2002

# Chromatin Regulation

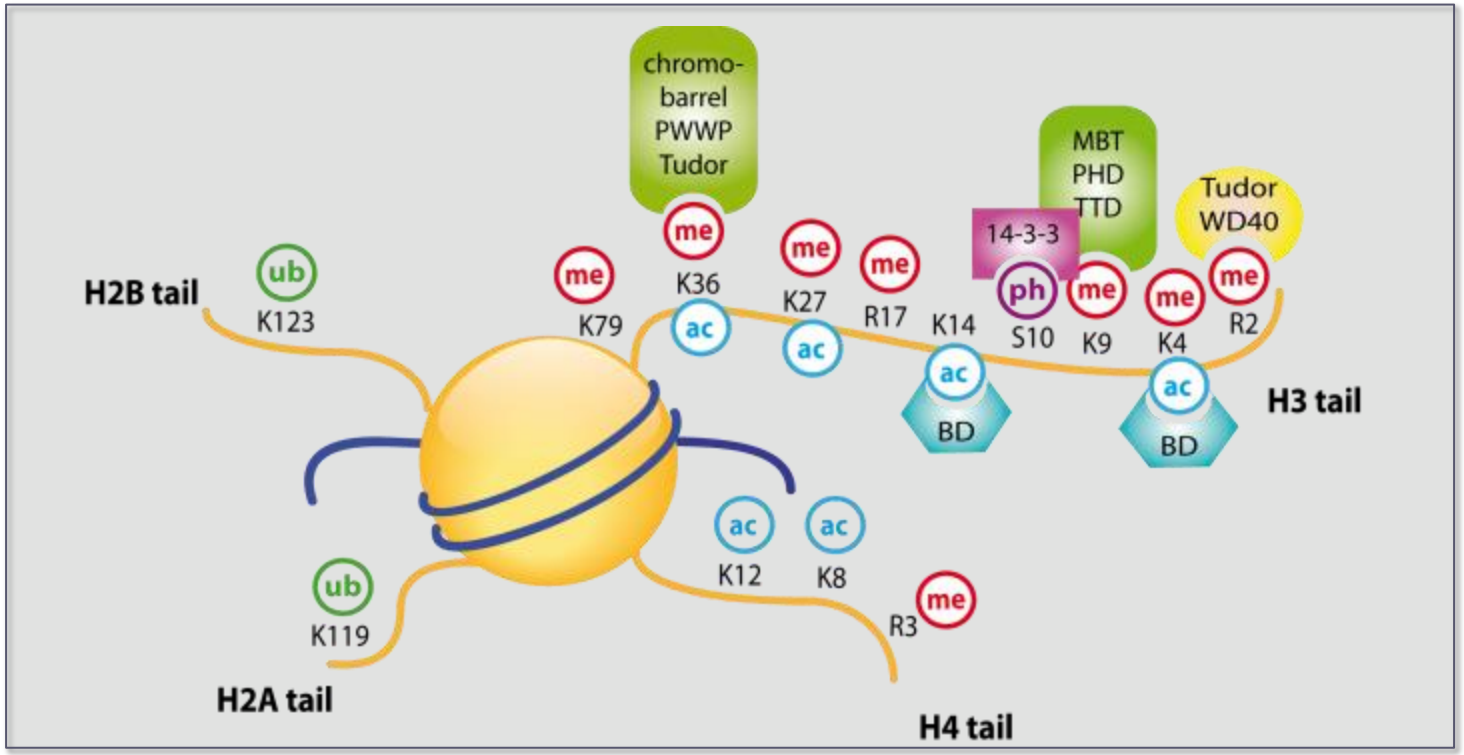
| Type of modification | Histone | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | H3K4 | H3K9 | H3K14 | H3K27 | H3K79 | H3K36 | H4K20 | H2BK5 |
| mono-methylation | activation[22] | activation[23] | | activation[23] | activation[23][24] | | activation[23] | activation[23] |
| di-methylation | | repression[25] | | repression[25] | activation[24] | | | |
| tri-methylation | activation[26] | repression[23] | | repression[23] | activation,[24] repression[23] | activation | | repression[25] |
| acetylation | | activation[26] | activation[26] | activation[27] | | | | |

From Wikipedia

# The nucleosome and histone codes

ub Ubiquitylation
ac Acetylation
me Methylation
ph Phosphorylation

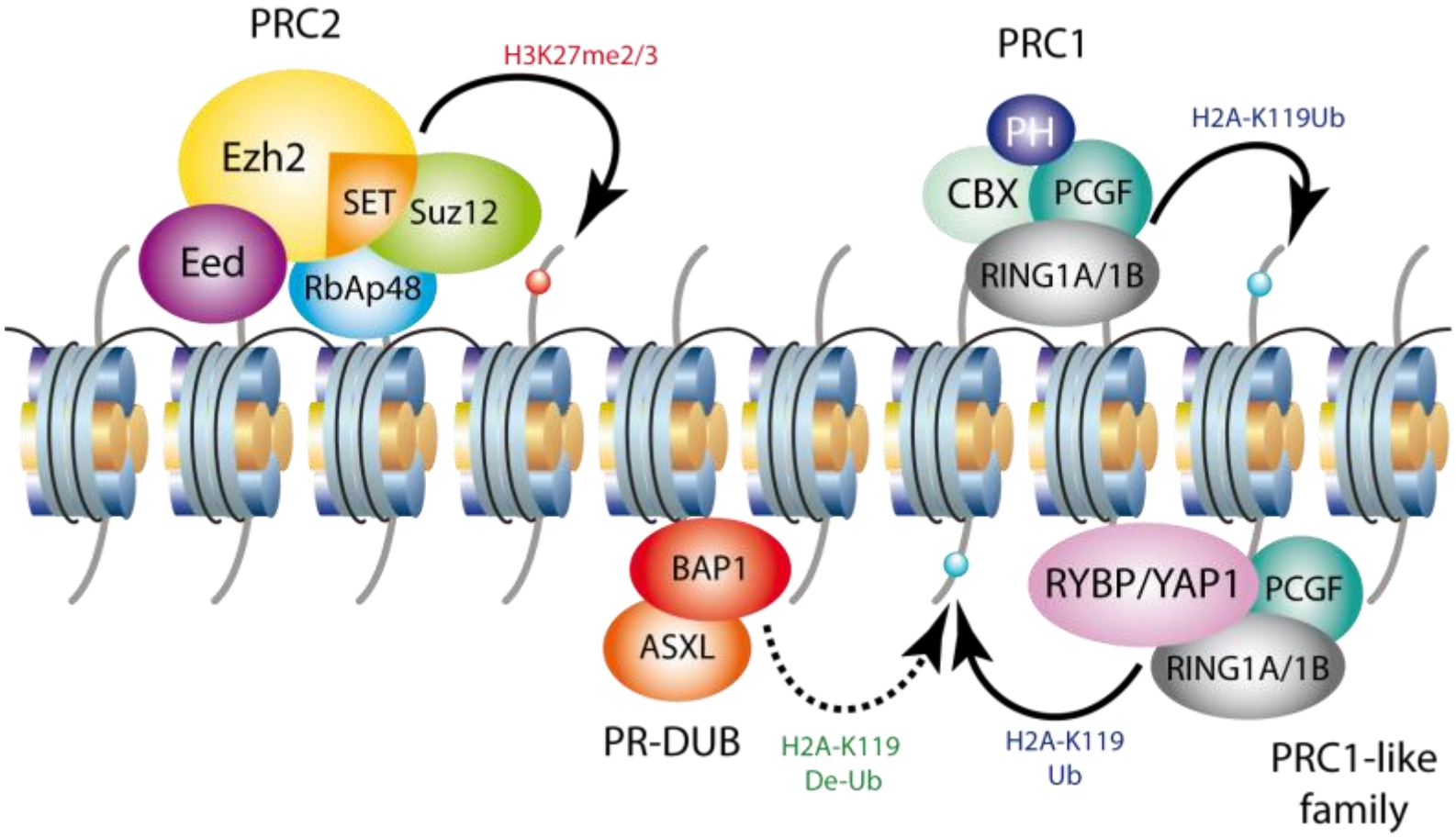
- Chromatin compaction
- Recruitment of protein readers



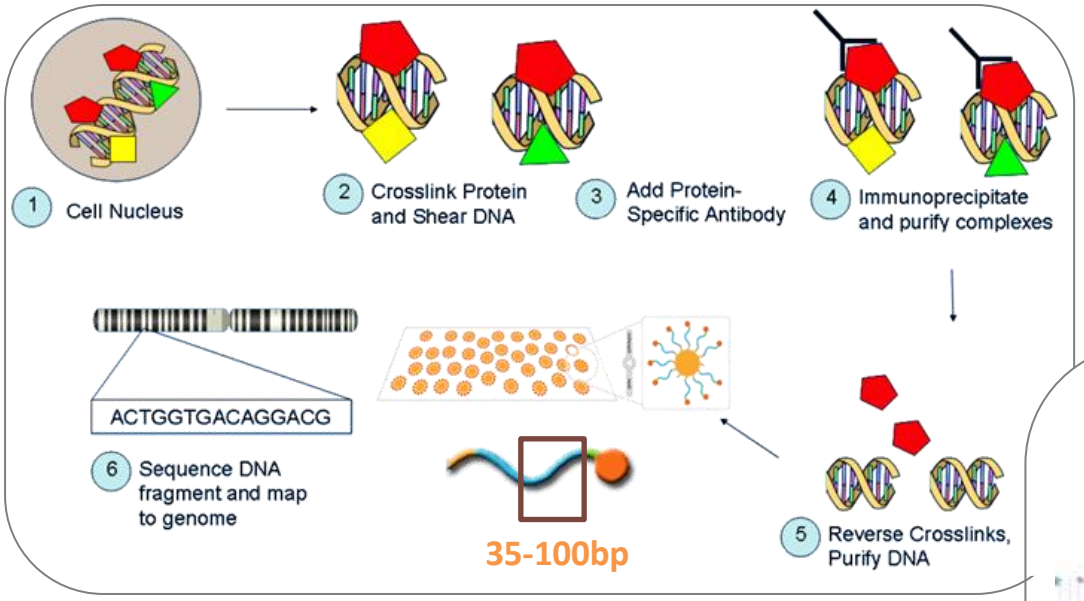

| Transcription Activation | Transcription Silencing |
|---|---|
| **Euchromatin** | **Heterochromatin** |

# Polycomb group of proteins in mammals

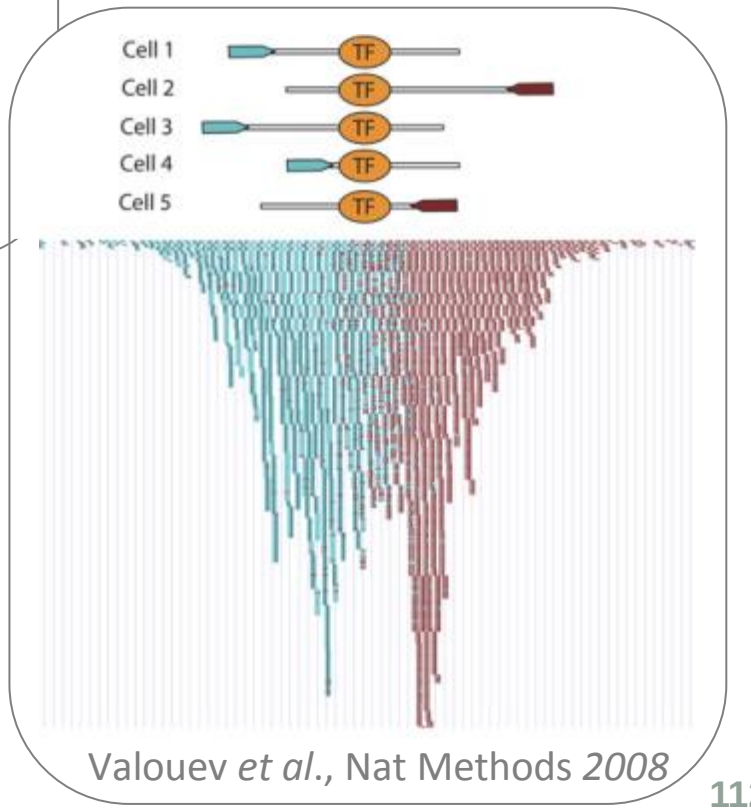
**Maintenance of gene repression**



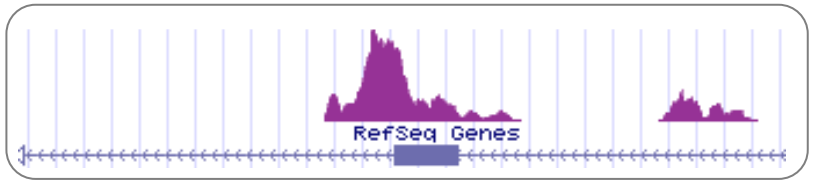
# A typical ChIP-seq experiment




+ Control (e.g., input DNA)

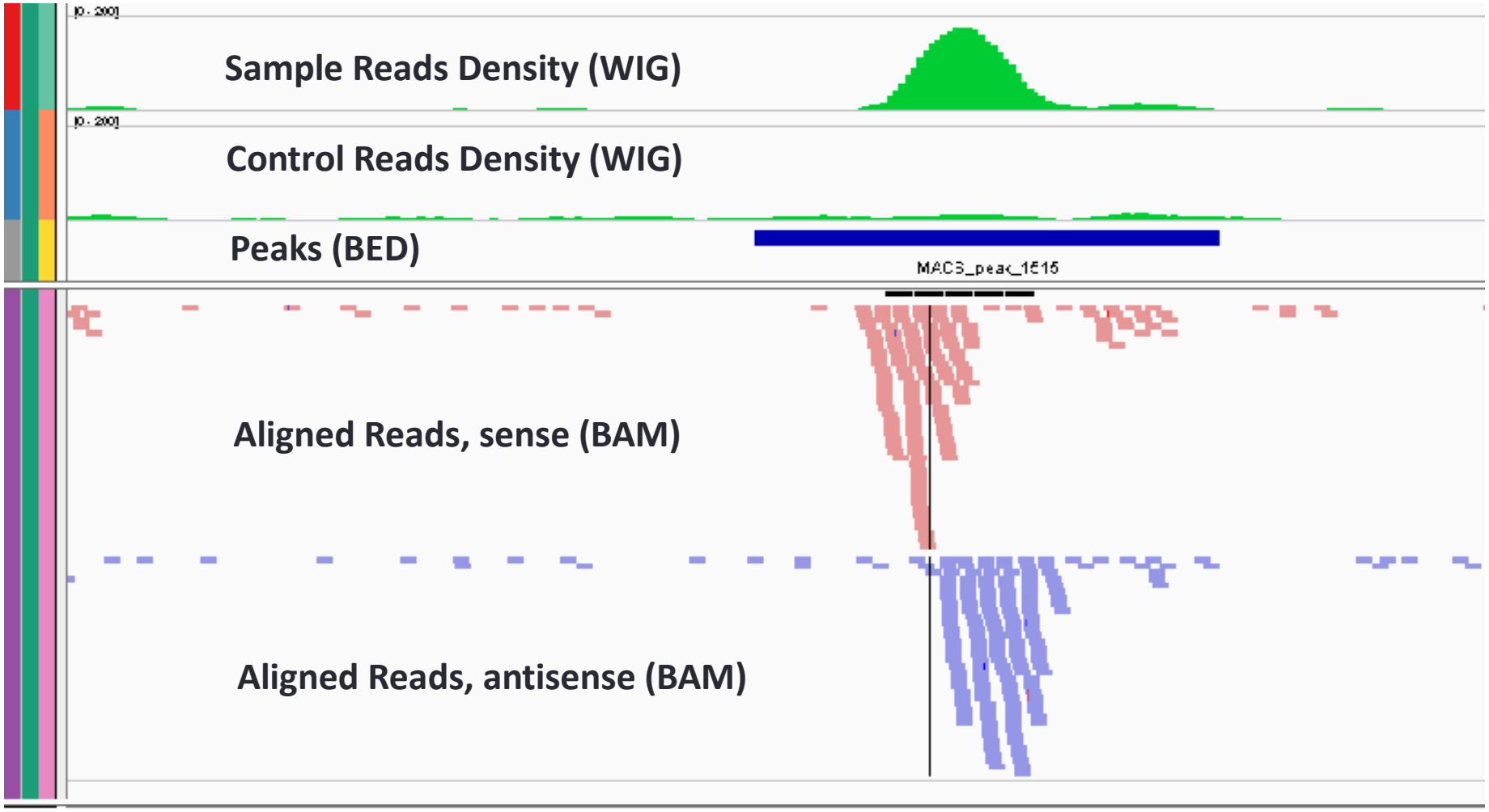



Valouev *et al.*, Nat Methods *2008*




From V. Boeva

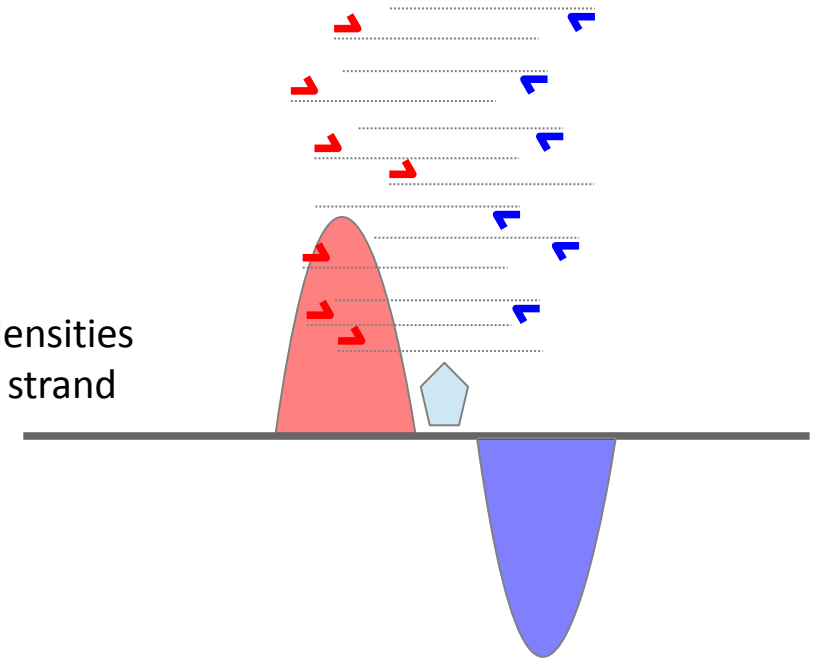
# ChIP-seq signal for Transcription Factor (TF)



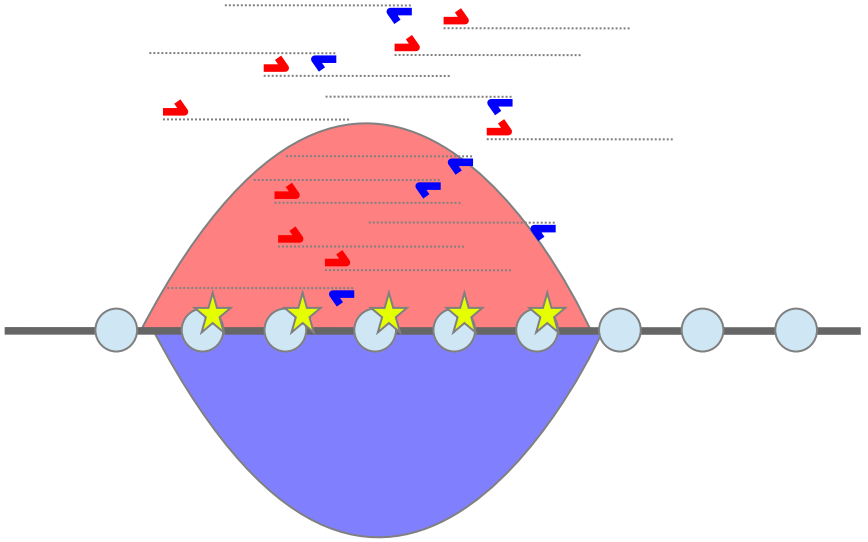

From C. Herrmann

# TF vs Histone modifications

read densities
on +/- strand



**ChIP-seq on TF binding sites**



**ChIP-seq on histone modifications**

The strand asymmetry is completely lost when considering ChIP datasets for diffuse histone modifications

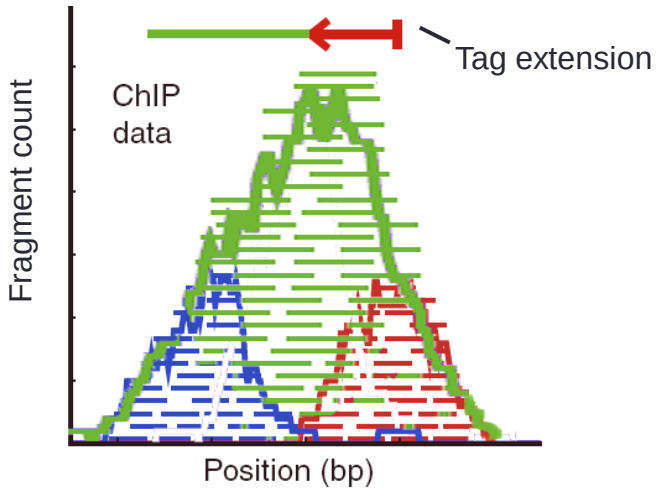
From C. Herrmann

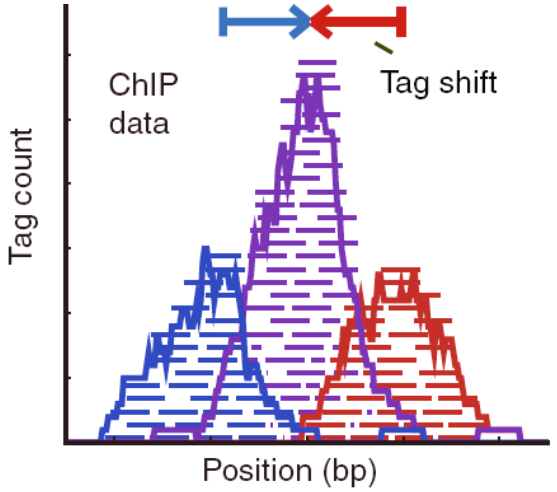
# Peaks Detection

The reads strand distribution do not represent the true location of the binding site.

If *d* is the fragment length :

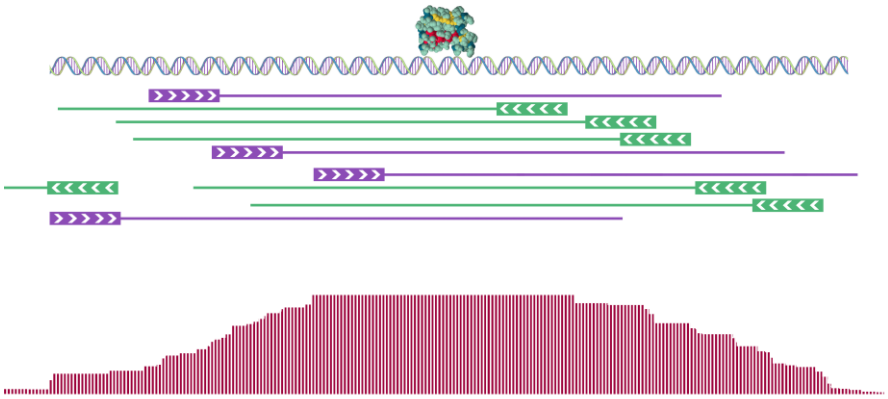
- Reads can be **shifted** by *d/2* (MACS, etc.)
- Reads can be **elongated** to a size of *d* (FindPeaks, etc.)



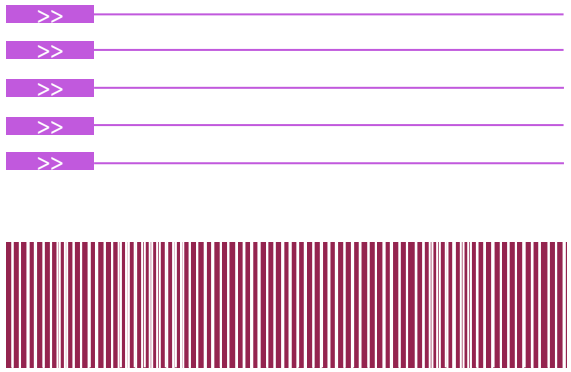




Adopted from S. Pepke et al., 2009 Nat Methods

# It is important to filter out PCR duplicates

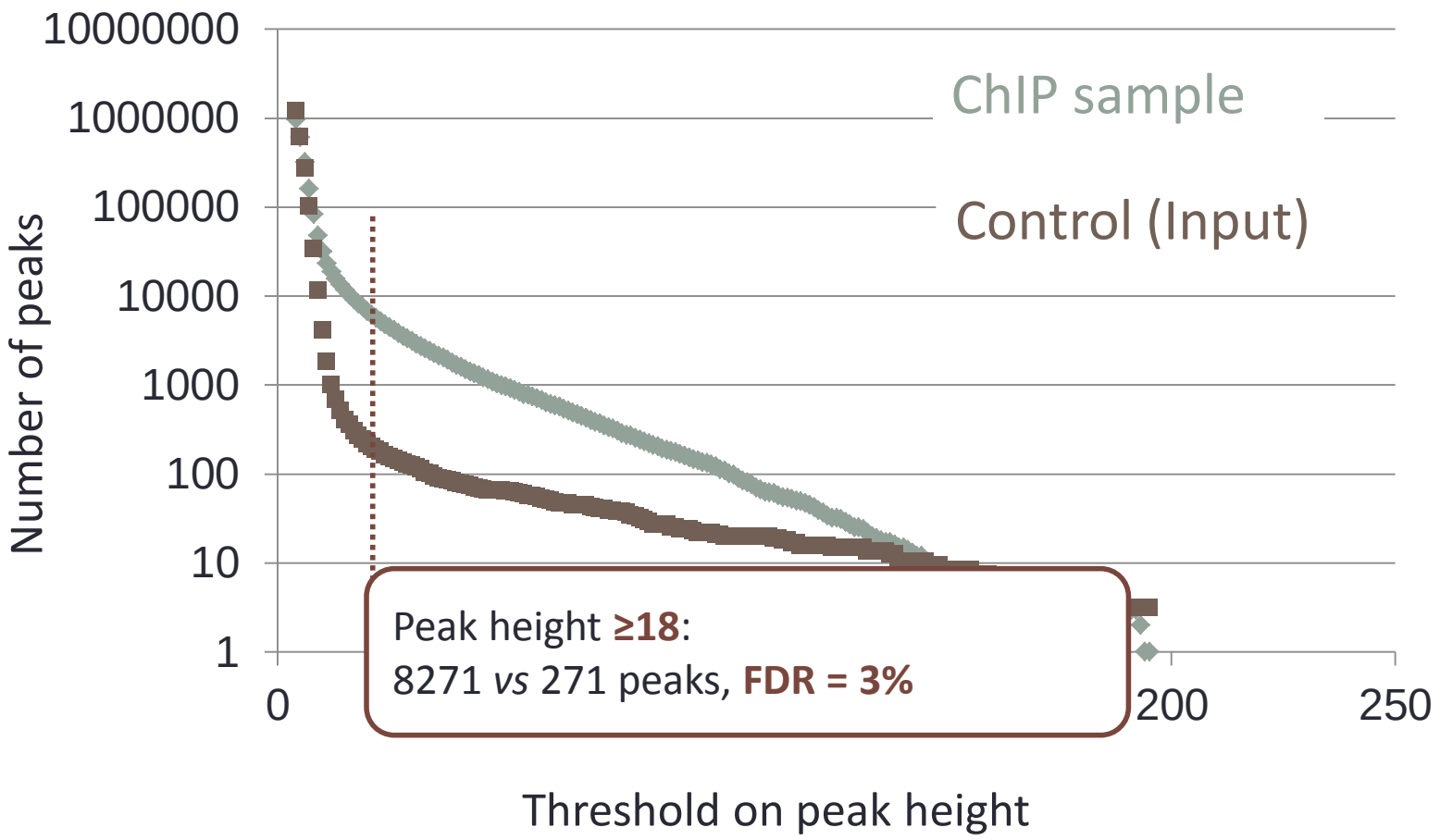


Real Binding Site



A peak due to PCR duplicates

# Peak threshold selection based on peak height



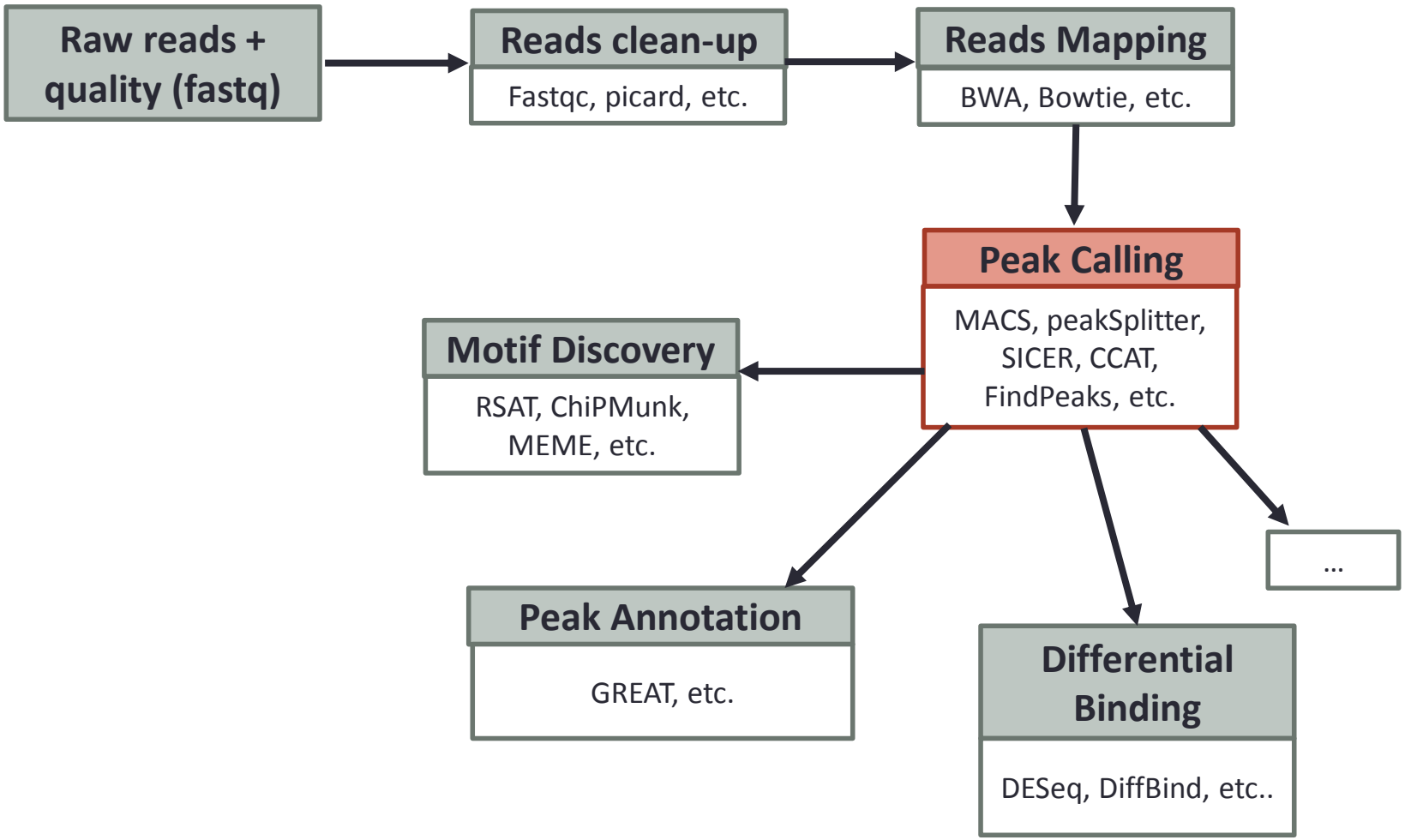

**The input sample is mandatory for ChIP-seq experiments**

# What can I do with my peaks ?

**Functional interpretation of ChIP-seq data**

- Detection of motif enriched in the binding sites (peak motif, MICSA, ChipMunk, etc.)
- Where do peaks localize ? Do they correlate with some functional marks ? promoter ? Intergenic regions ? Intronic regions ? Histone marks ?
- Finding Gene Ontolgy enriched term on genes associated with my peaks (GREAT)

# Workflow for ChIP-seq Analysis

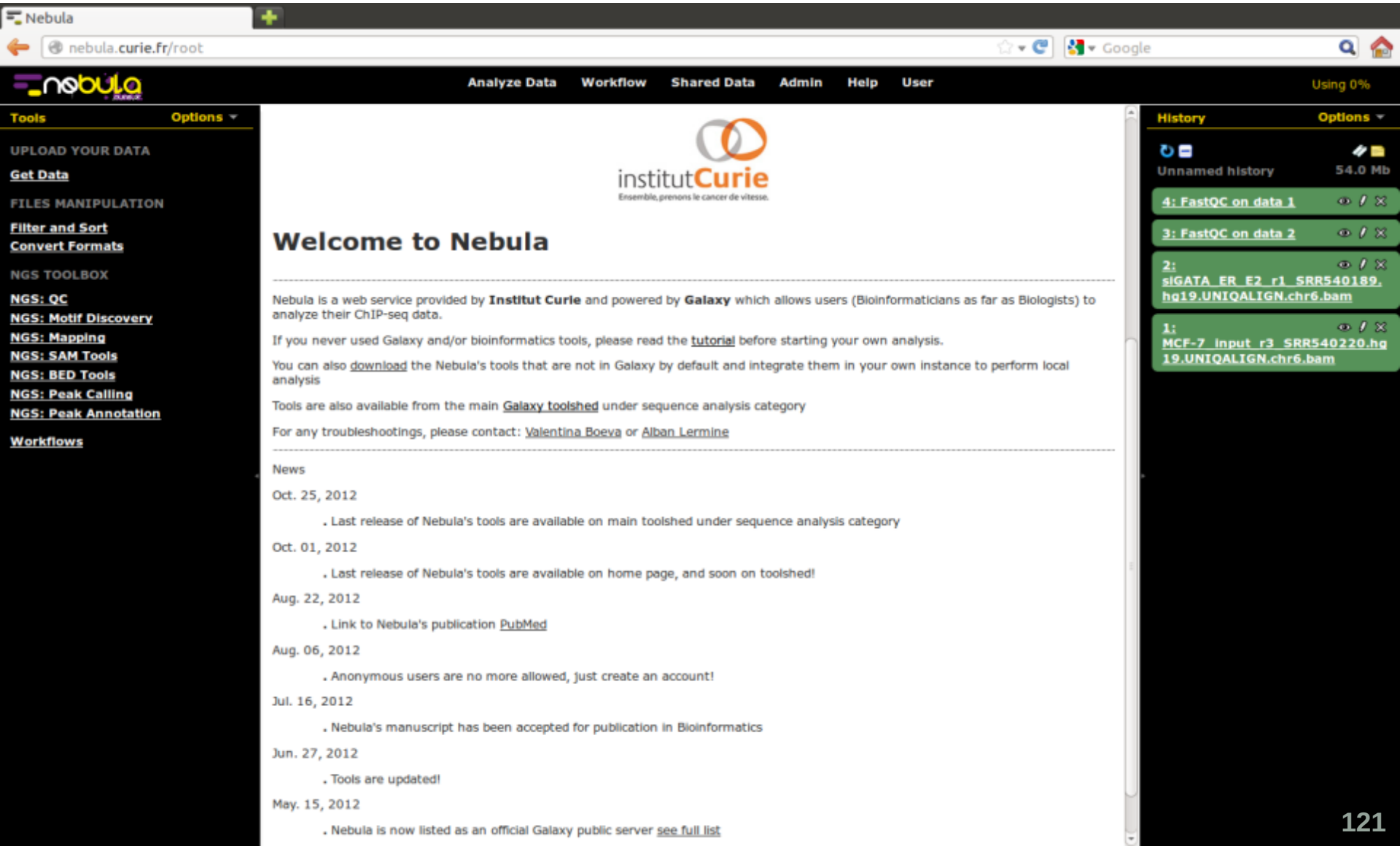


# Nebula (Boeva et al.)

- Nebula allows bioinformaticians as far as biologists to perform by them selves complete analysis of their ChIP-seq data
- Nebula's main goal is to identify transcription factors and associated roles starting from sequencing raw reads

- Nebula is composed of 23 tools:
  - Published tools: Bowtie, Samtools, MACS, FindPeaks, ChIPmunk, Fastqc, BEDtools
  - Homemade tools: Get peak distribution around TSS, Extract central regions of peaks, Genomic annotation of peaks, Gene annotation using peaks, Get peak height distribution
- Galaxy is used for the whole encapsulation and the workflow editor

# Nebula (Boeva et al.)



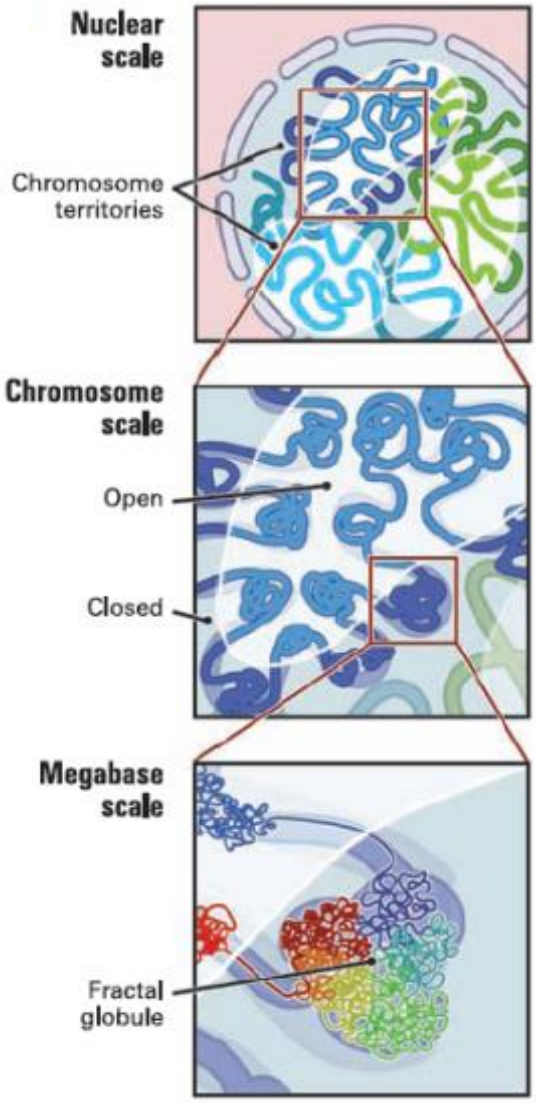
# ChIP-seq - FAQ

- What about ChIP-seq in tumour cells ?
- Can I used the same tools for TF and histone marks ?
- Do I need any control ?
- Do I need to  normalize my data ?
- How can I compare two ChIP-seq samples ?

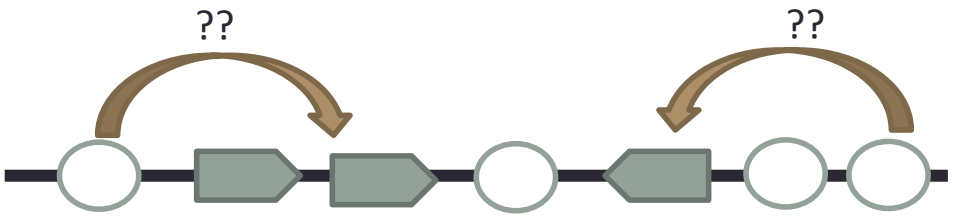
institutCurie

# Chromatine Conformation Capture

# Measuring physical interactions

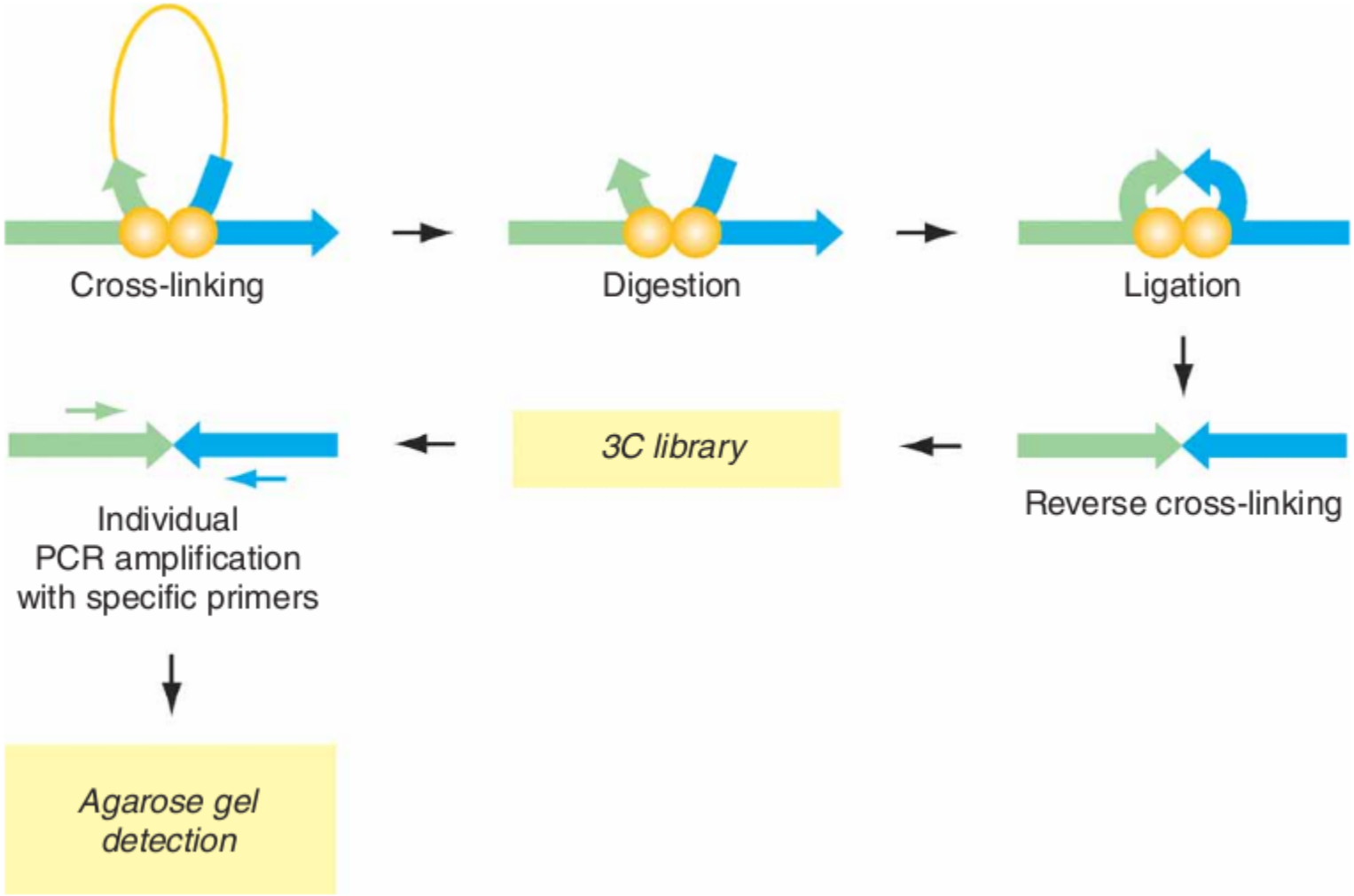



**How is the genome organized ?**
**Which element regulates which genes ?**

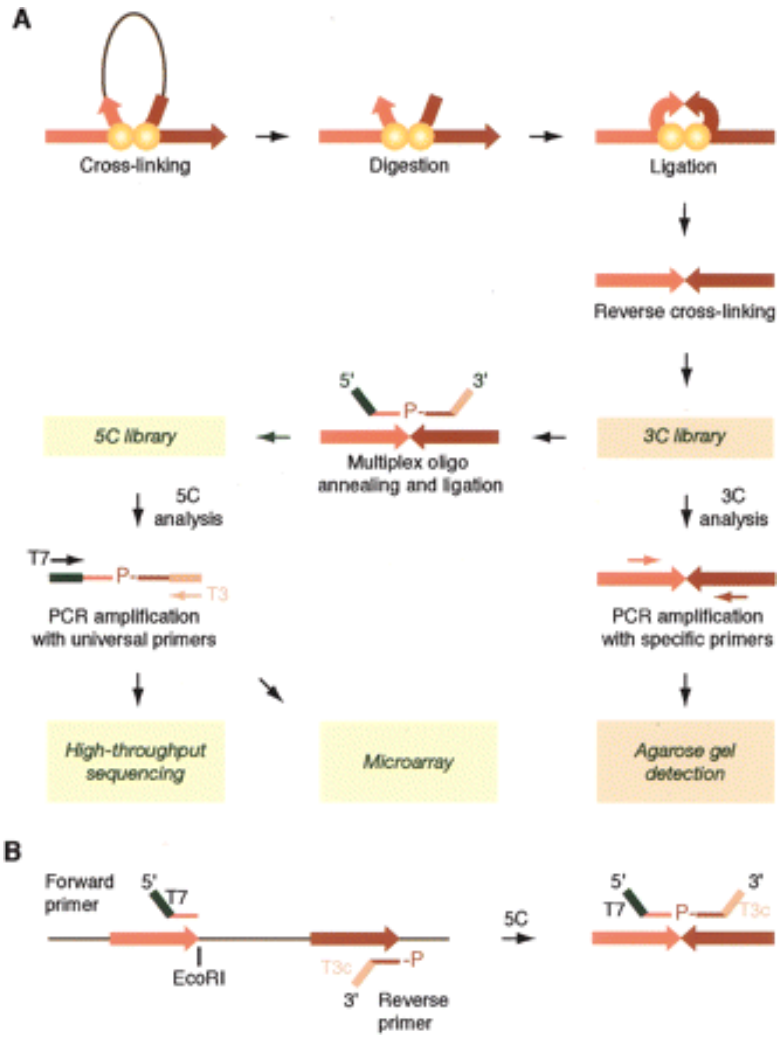



Dekker et al

# Chromosome Conformation Capture (3C)



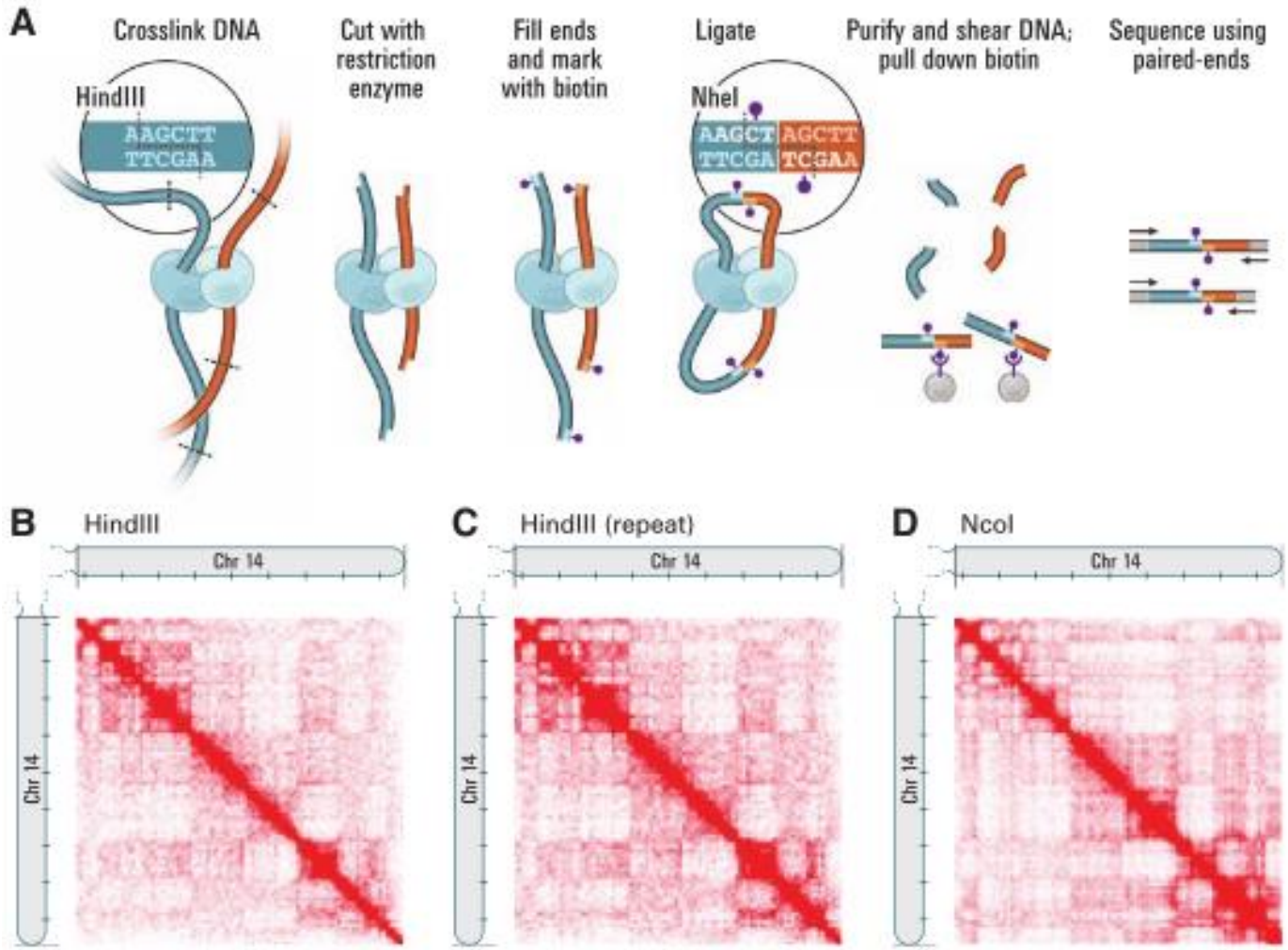
# High-throughput 'C' experiment (5C)




**The 3C Carbone Copy requires the design of reverse and forward primers.**
Forward 5C primers anneal to the sense strand of the 3'-end of restriction fragments and include half of the selected restriction site.
Reverse 5C primers anneal to the antisens strand of the 3'-end of restriction fragments, including half of the restriction site and are phosphorylated at the 5'-end..
5C forward and reverse primers anneal to the same strand of head-to-head ligation products present in the 3C library.

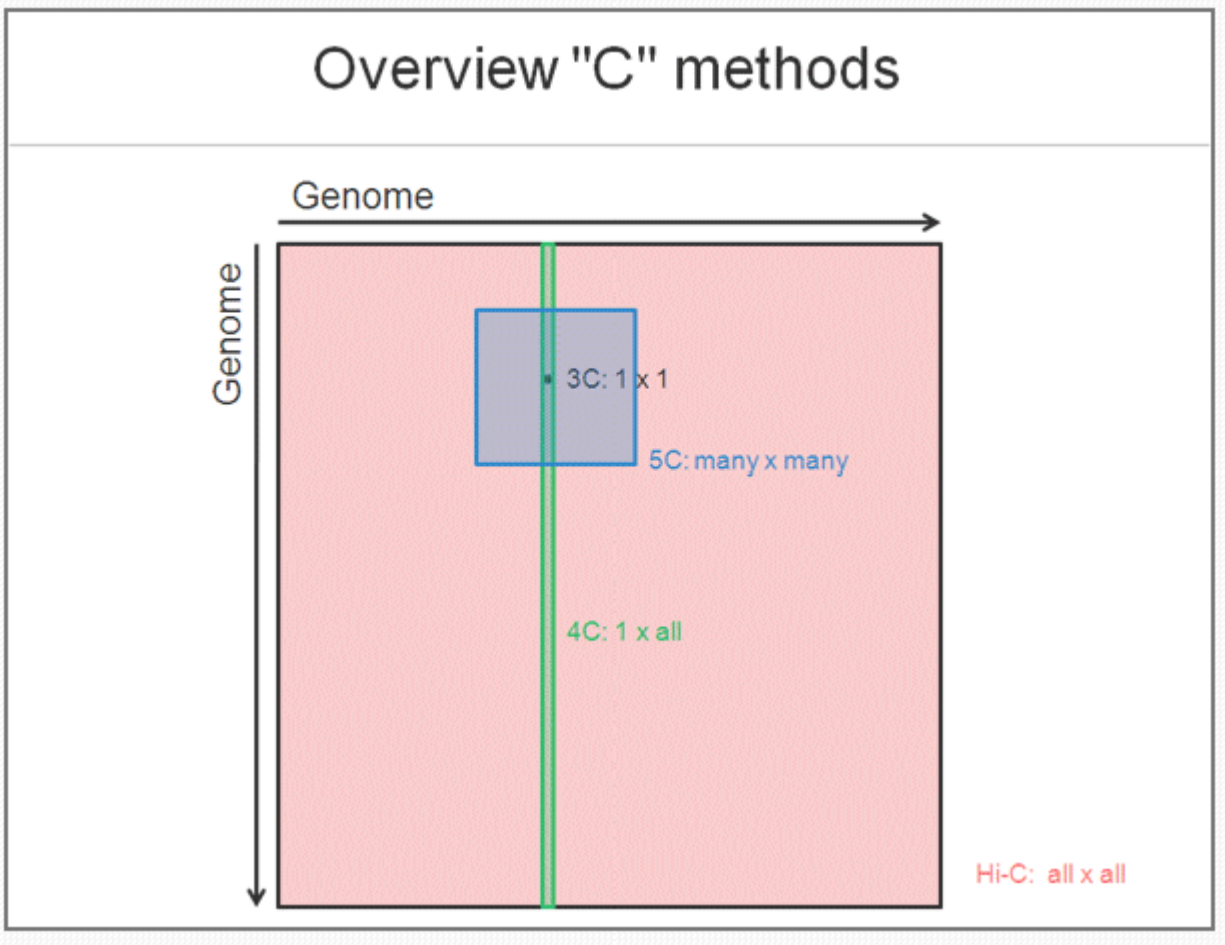
Dostie et al. 2006

# Hi-C Protocol



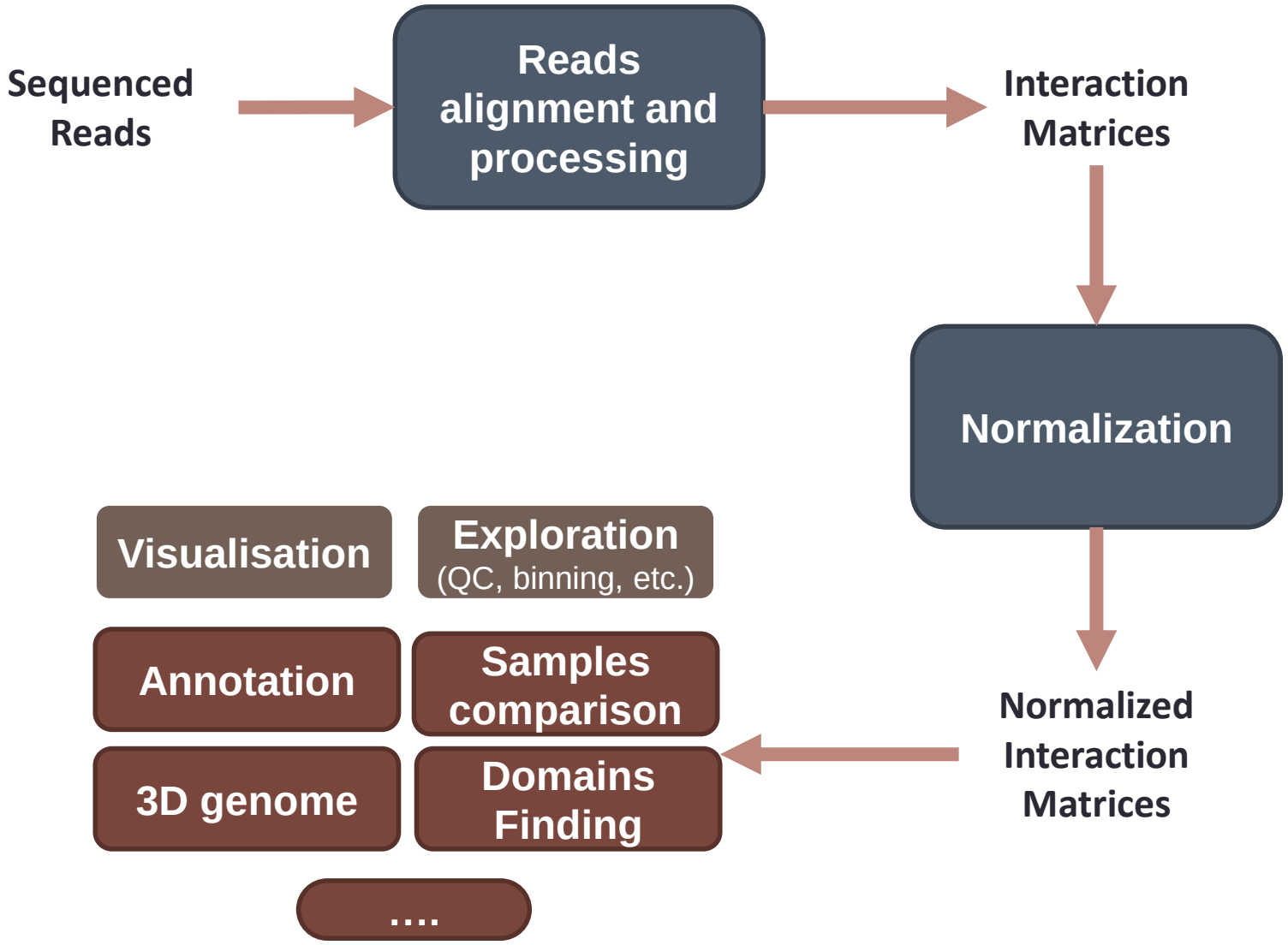

Lieberman-Aiden et al. 2009

# The ‘C’ World



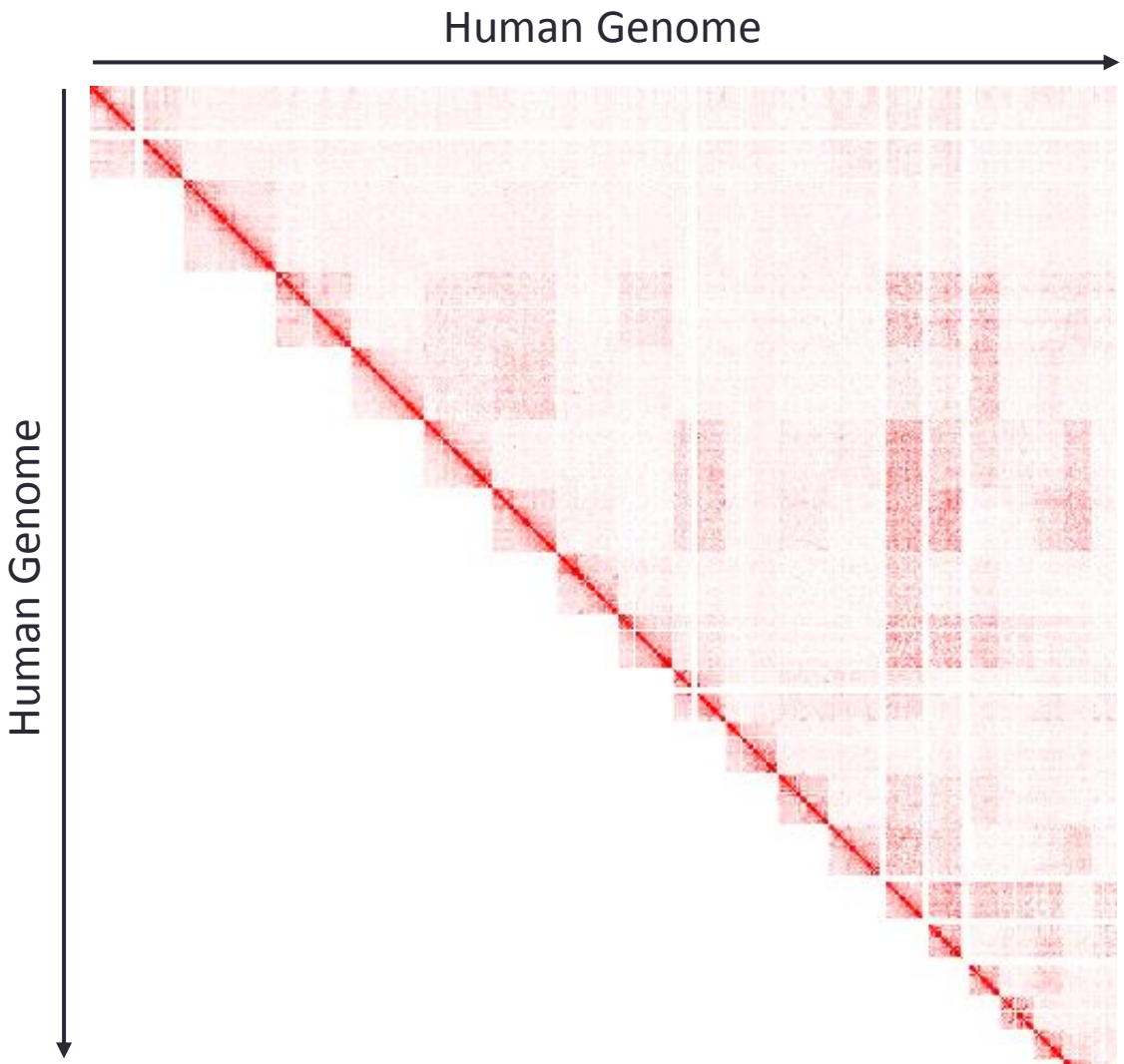

Dekker et al.

# Standard analysis strategy



# What the data look like ?

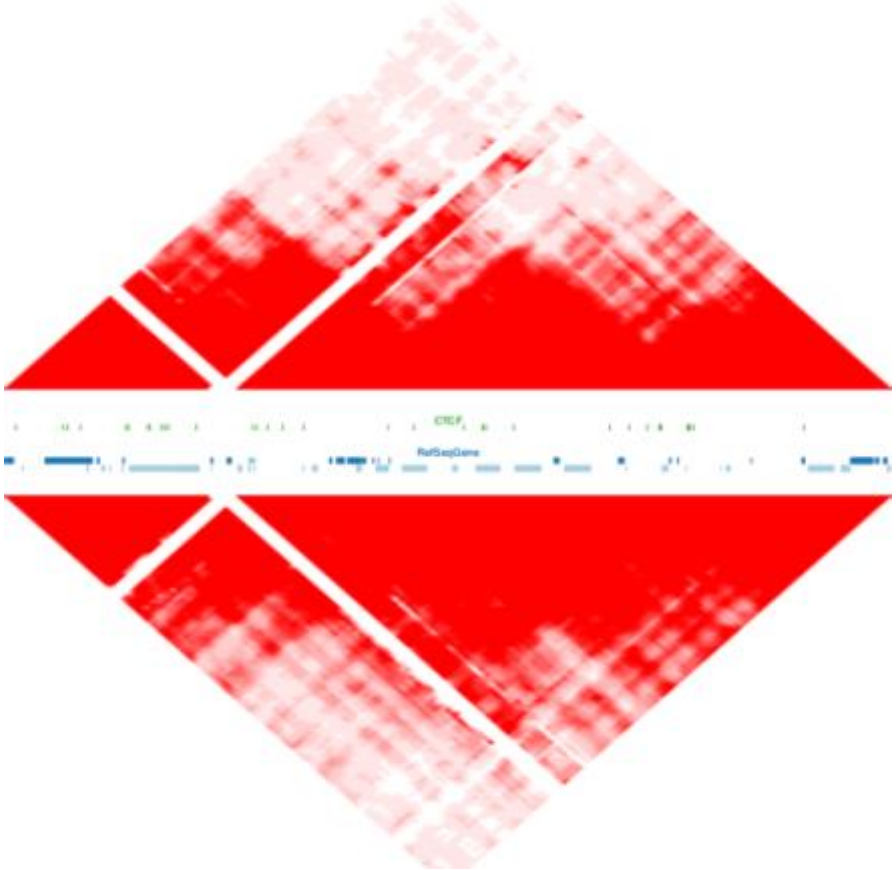
## Hi-C genome view

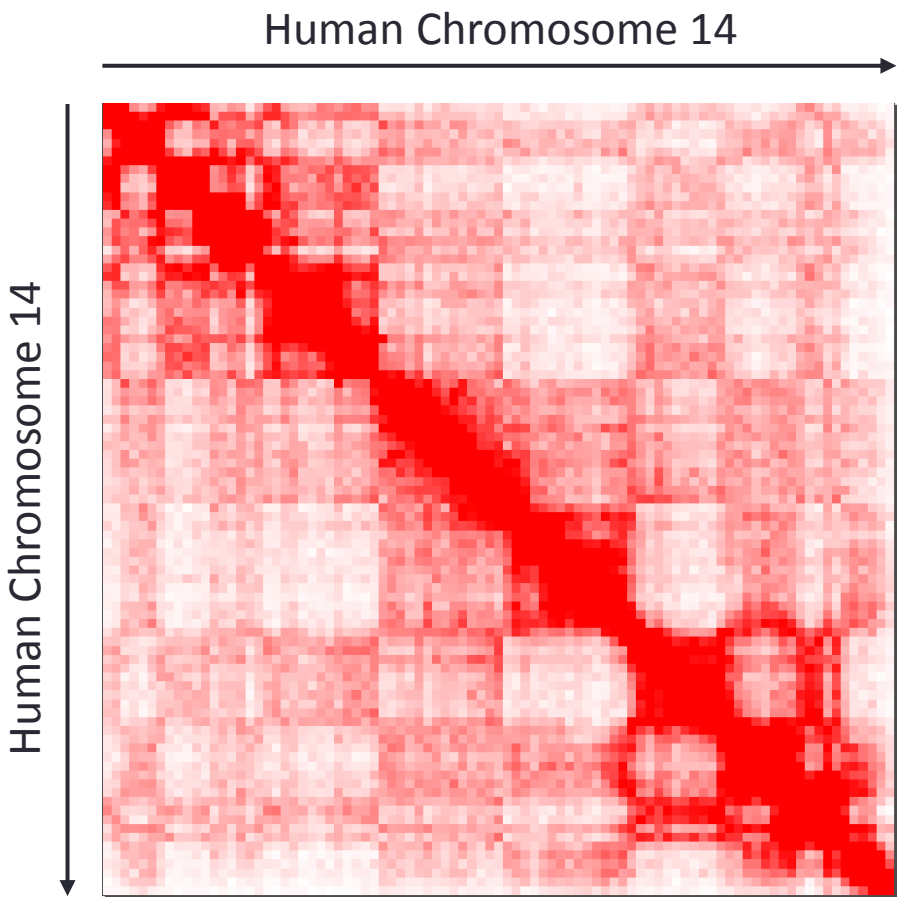



Hi-C data from Lieberman et al

# What the data look like ?

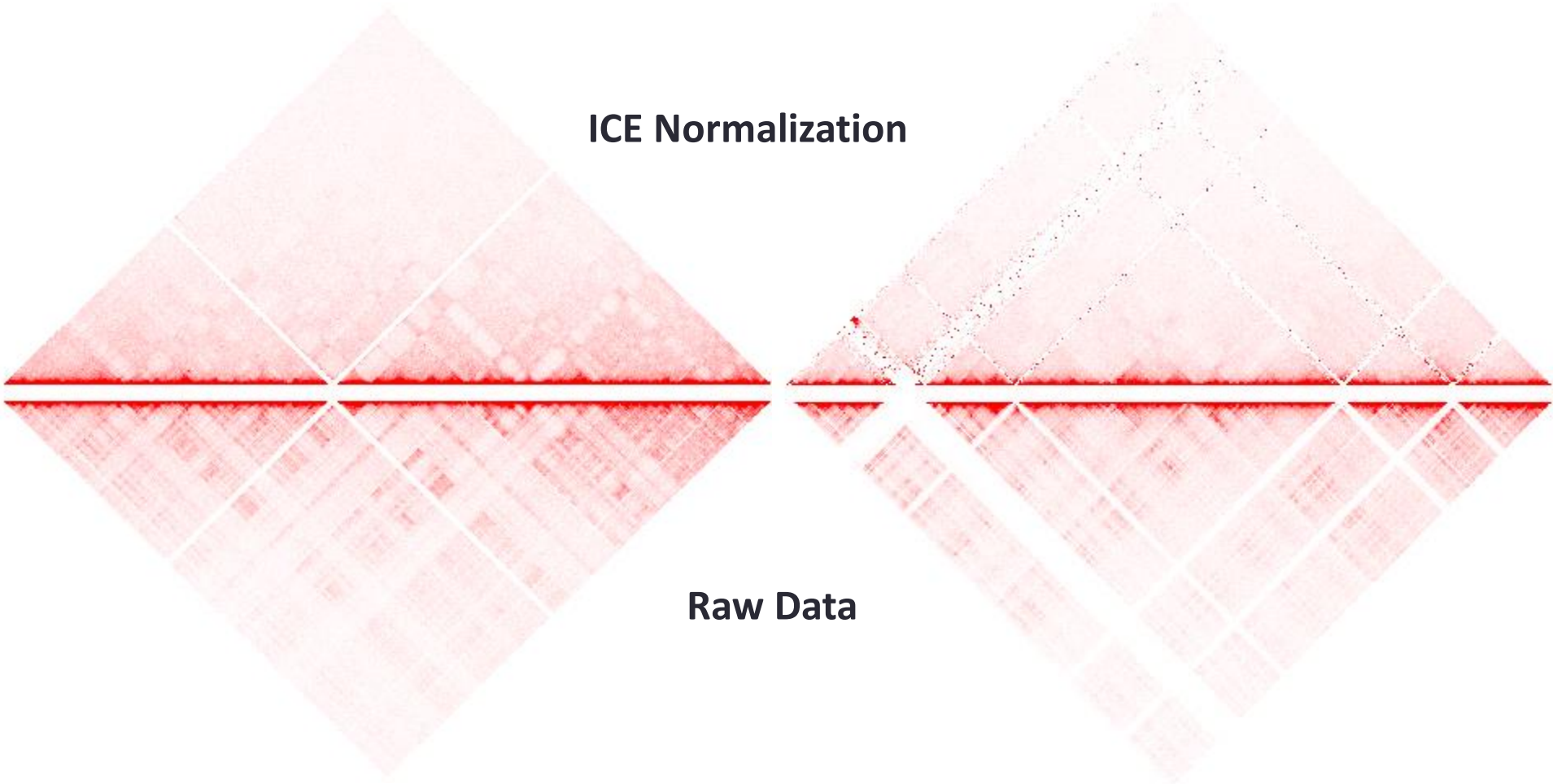
## CIS interaction map



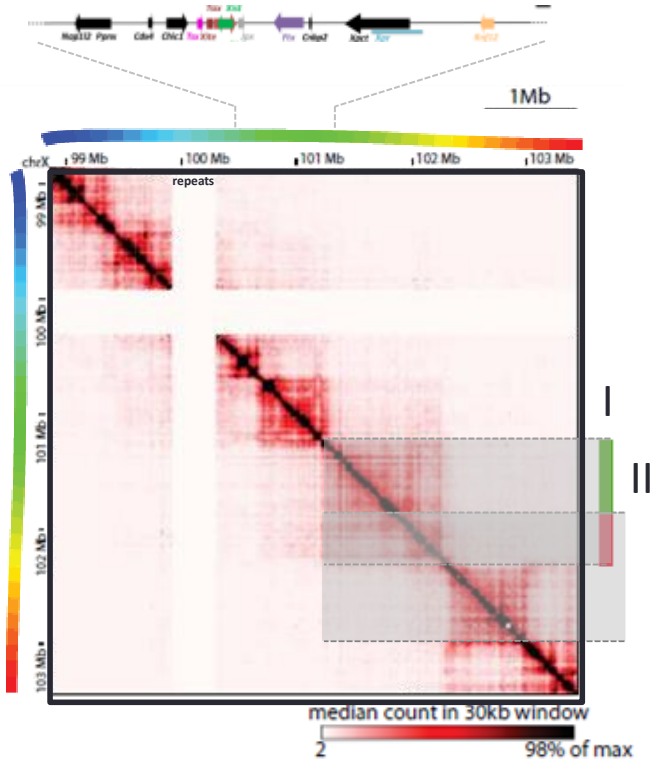



Hi-C data from Lieberman et al

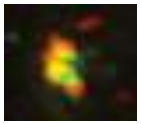
# Hi-C ICE Normalization (Imakaev et al.)



# Identification of Topological Domains (TADs)



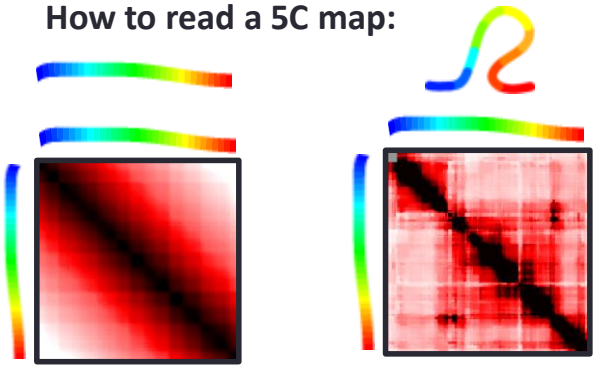

Male mouse ES cells



I – Same domain

II – Neighboring domain

**How to read a 5C map:**



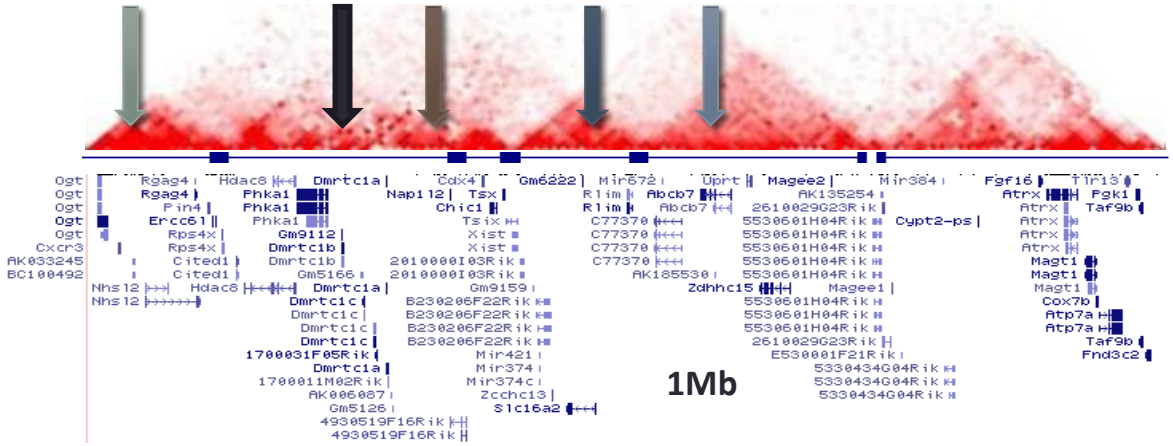
**The Xic locus is organised into sub-megabase Topologically Associated Domains (TADs) of interacting sequences, with multiple specific long range interactions**

**Nora et al** (2012) Spatial partitioning of the regulatory landscape of the X-inactivation center. *Nature* 485:381-5

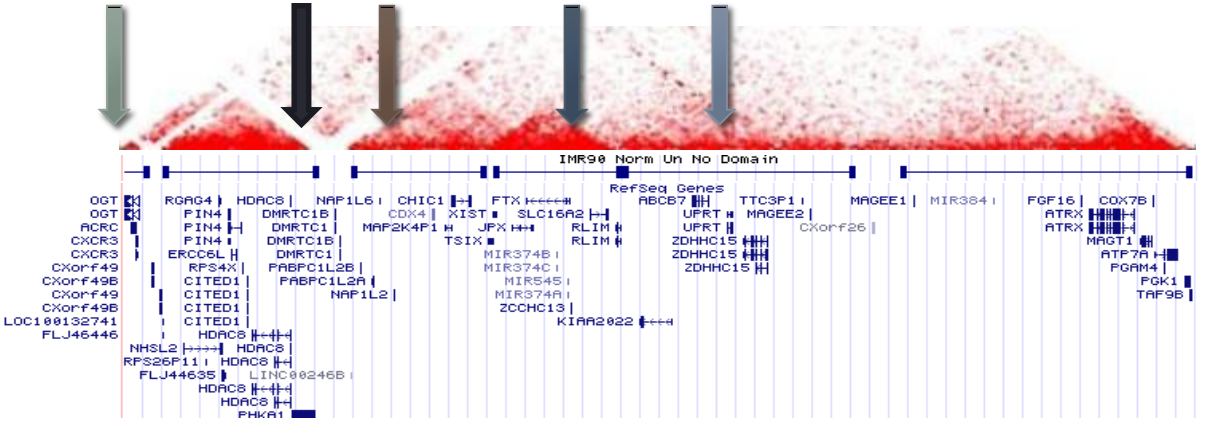
# Identification of Topological Domains (TADs)

Topological domain (TAD) boundaries are highly conserved between mouse and human and remain unchanged during differentiation

**Mouse**
(ESCs)



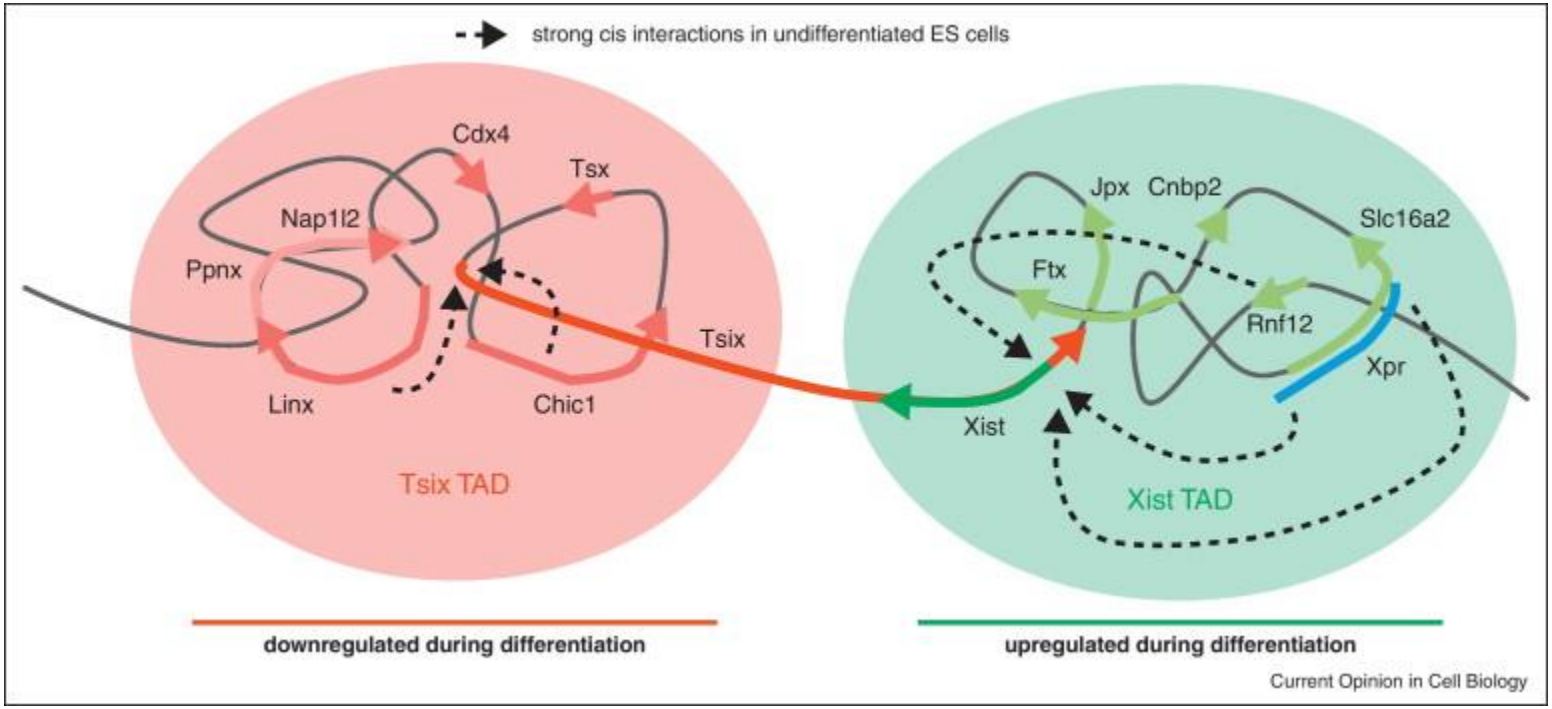

**Human**
(IMR90)




**Hi-C data from Dixon et al., 2012**

# Identification of Topological Domains (TADs)

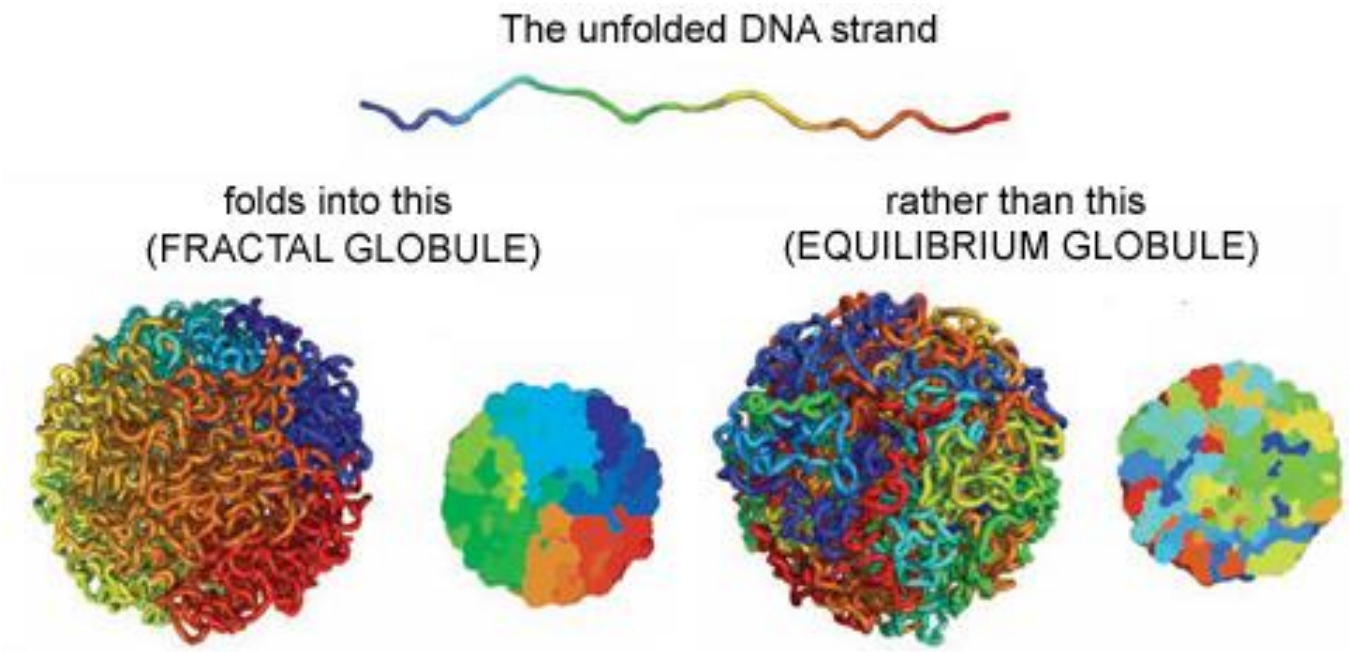
## Inference of 3D structure




Pollex T., Heard E. 2012

# Genome Architecture based on Hi-C data

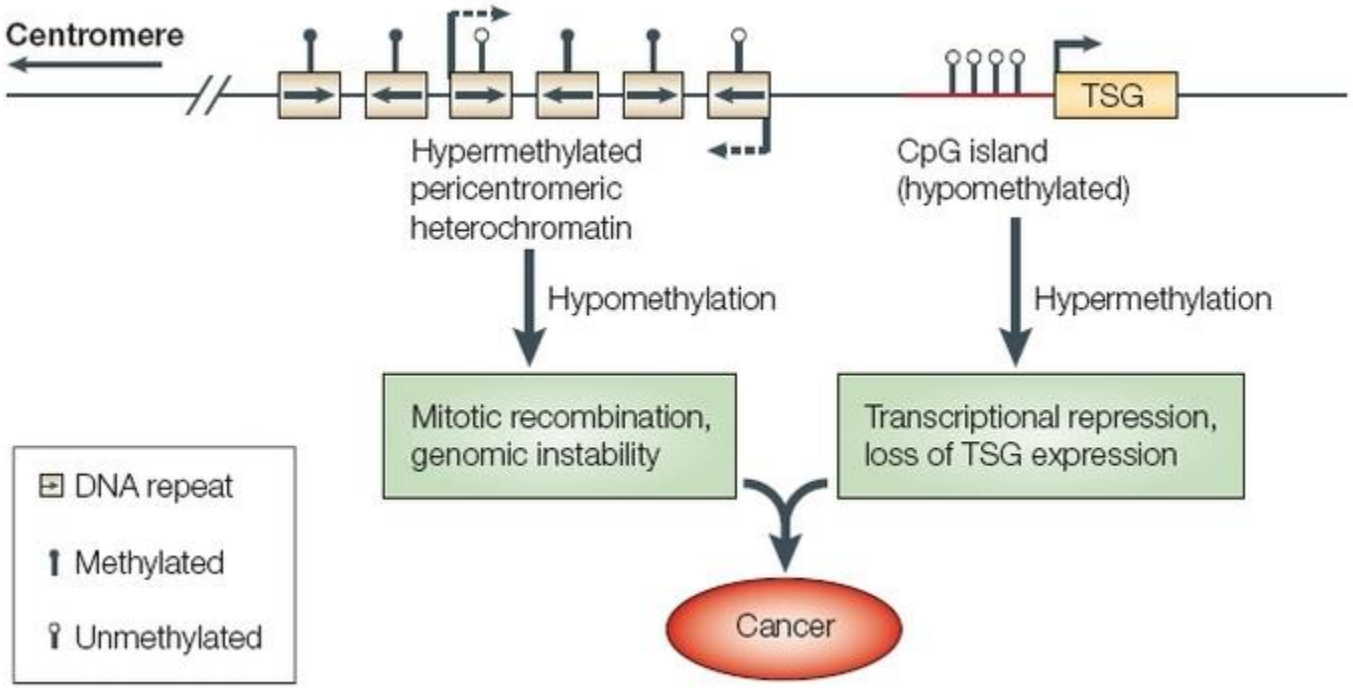
## The fractal globule



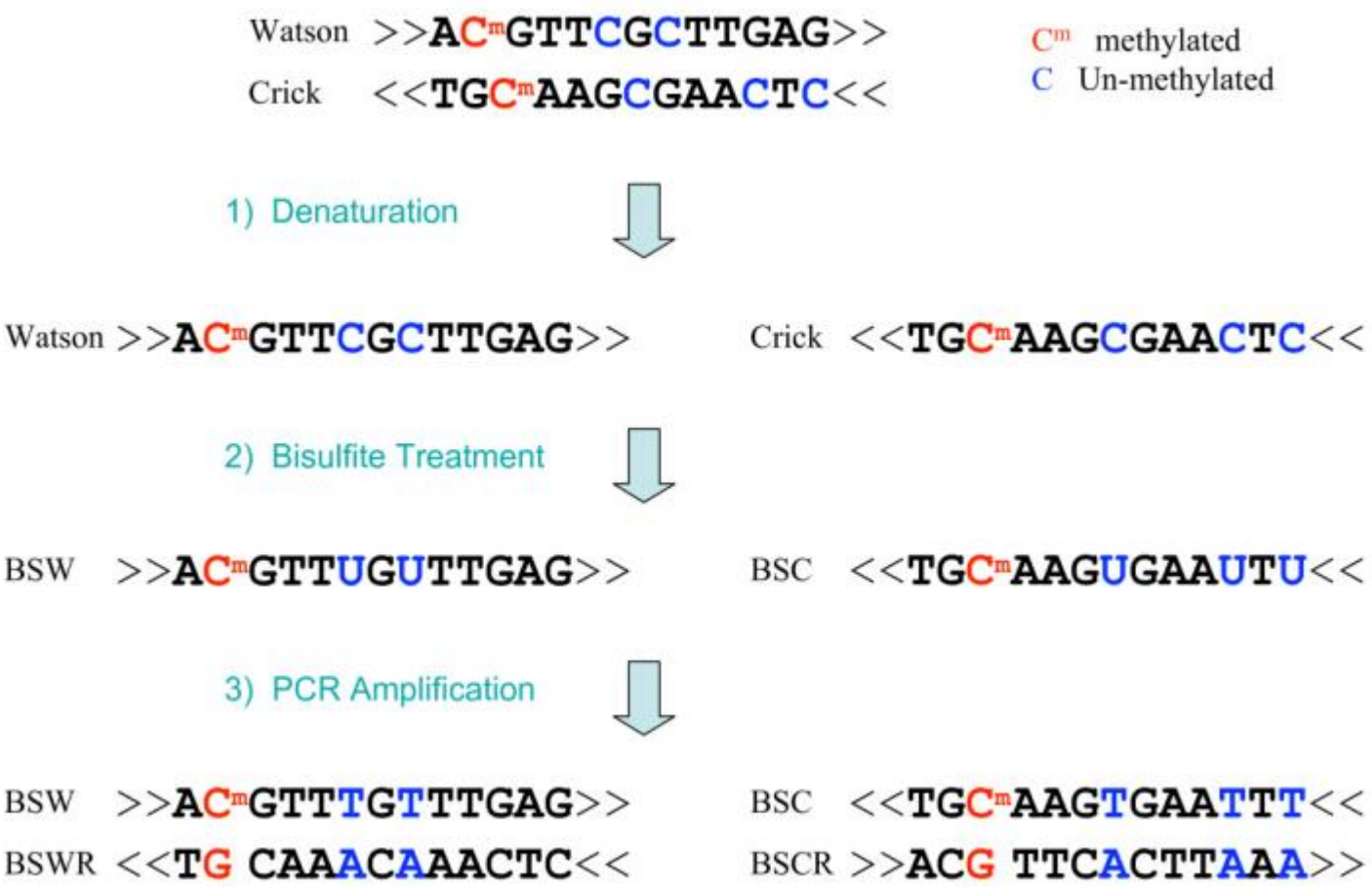

Lieberman-Aiden et al. 2009

# OTHERS NGS APPLICATIONS

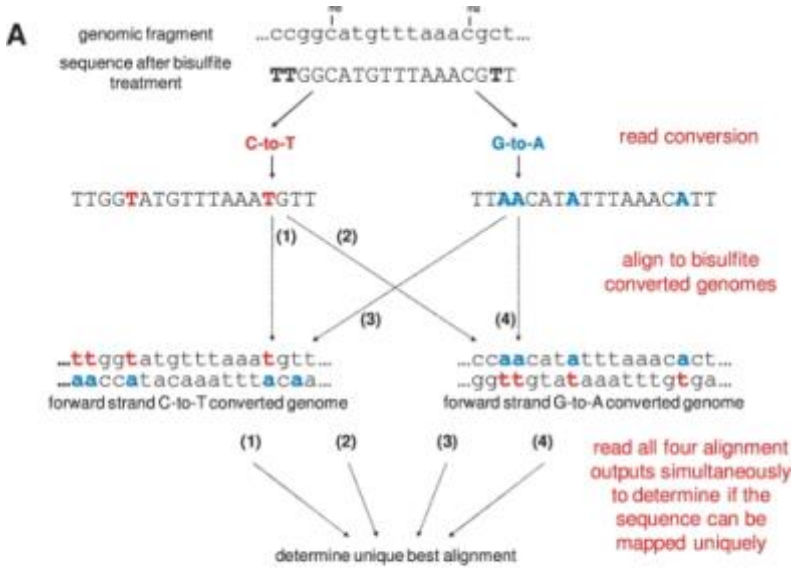
# DNA methylation and Cancer

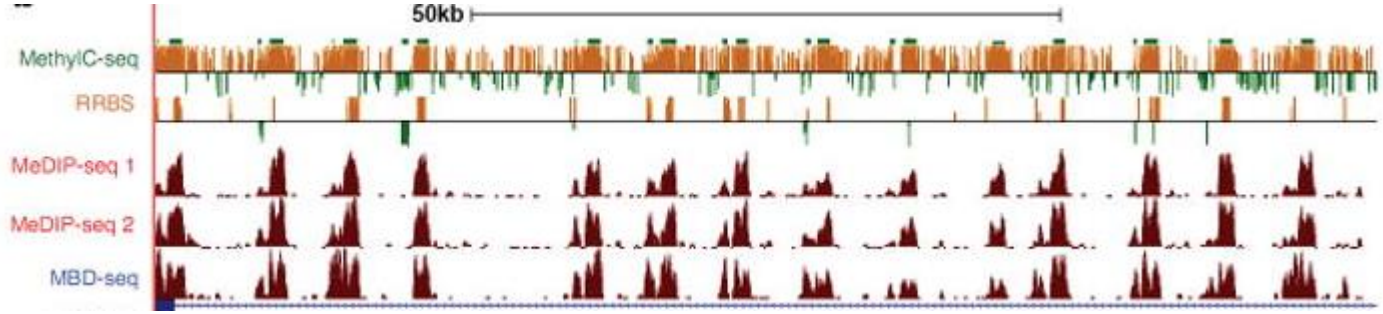


# Bisulfite-seq

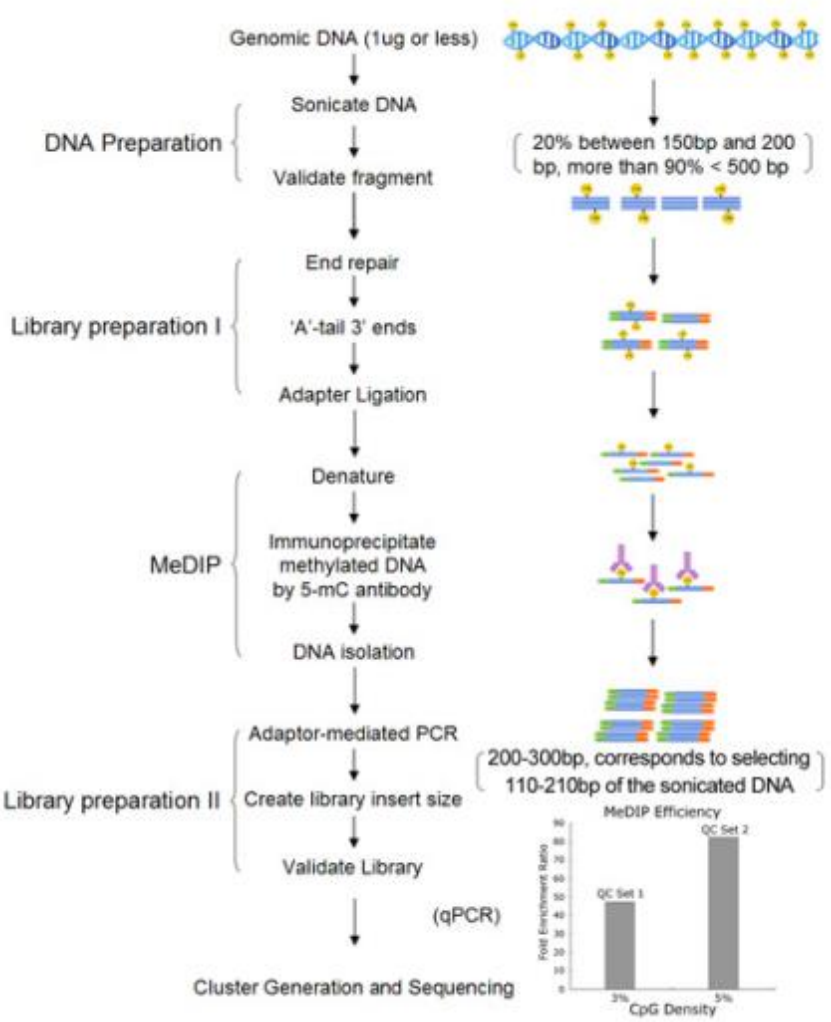


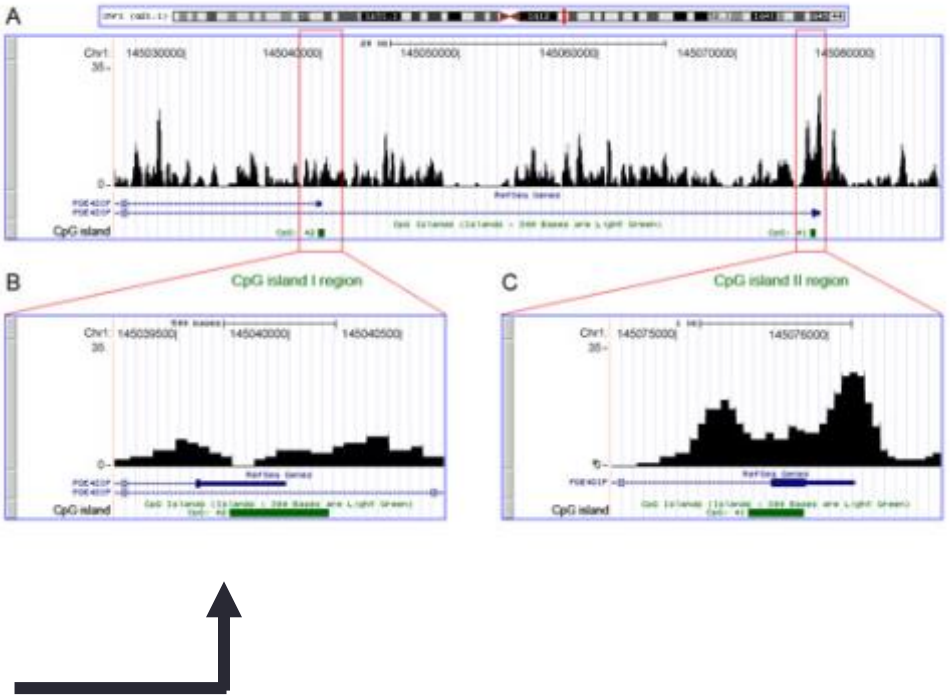
# Bisulfite-seq



# MeDIP-seq (Methylated DNA Immunoprecipitation)



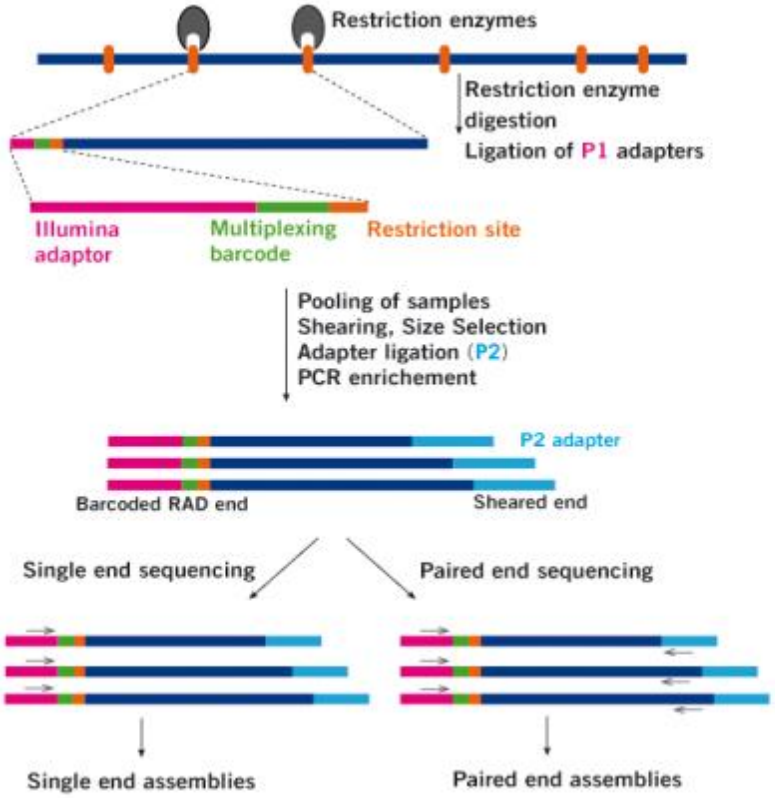



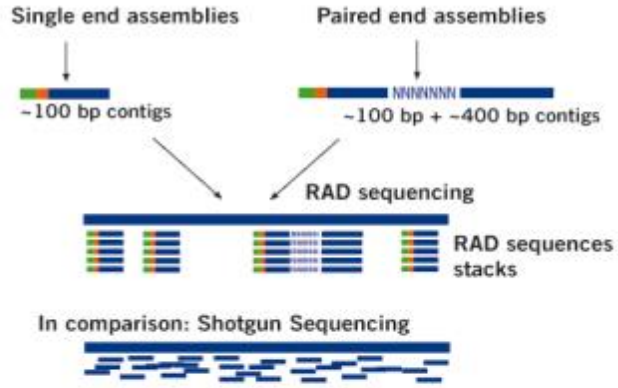
# CLIP-seq (Cross Linking and Immunoprecipitation)

- ChIP-seq for RNA binding sites
- Mainly used to deteect miRNA target interaction maps, or mRNA splicing proteins
- The choice of the control is crutial (RNA-seq)

# RAD-seq (Restriction Associated DNA)

Benefit from all the advantages of the precise restriction site associated DNA marker genotyping (RAD) with HT sequencing. By focusing on regions flanking particular restriction enzymes (genome complexity reduction) instead of the entire genome, we can discover and screen thousands of SNPs and genotype large populations. Since RAD-Seq requires less reads per sample, you get your results faster since we perform high throughput sequencing of numerous samples simultaneously.



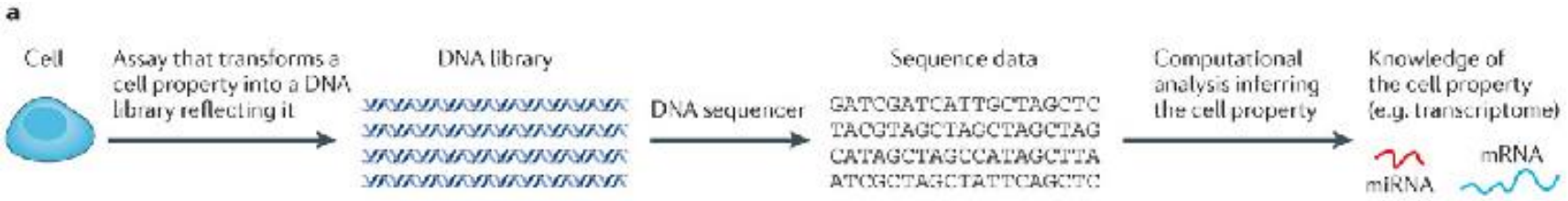




DNA is digested with a specific set of restriction enzymes and two adaptors ligated to the fragments afterwards.

Sequencing the fragments from either one side or both sides results in RAD sequences stacks.

The stacks are the perfect starting point to identify and analyse genetic markers.

# Single Cell Sequencing




Current implementations include single-cell genomics (targeted exome or mutational, copy-number variation and recombination analysis in germ cells), transcriptomics (transcriptome analysis and recombination analysis in the immune system) and epigenomics (Shapiro E. et al).

**But :**

- Bioinformatics and statistical analysis ?
- Noise
- Coverage

institutCurie

# NGS FOR CLINICAL APPLICATIONS

# Personalized Medicine

The phrase “personalized medicine” is commonly used to refer to genomic medicine; defined as “the use of information from genomes (from humans and other organisms) and their derivatives (RNA, proteins and metabolites) to guide medical decision-making”.

Personalized medicine may, however, be defined more broadly to be a model of healthcare that is predictive, personalized, preventive and participatory (“P4 Medicine”), and that also applies technologies to customize and deliver care

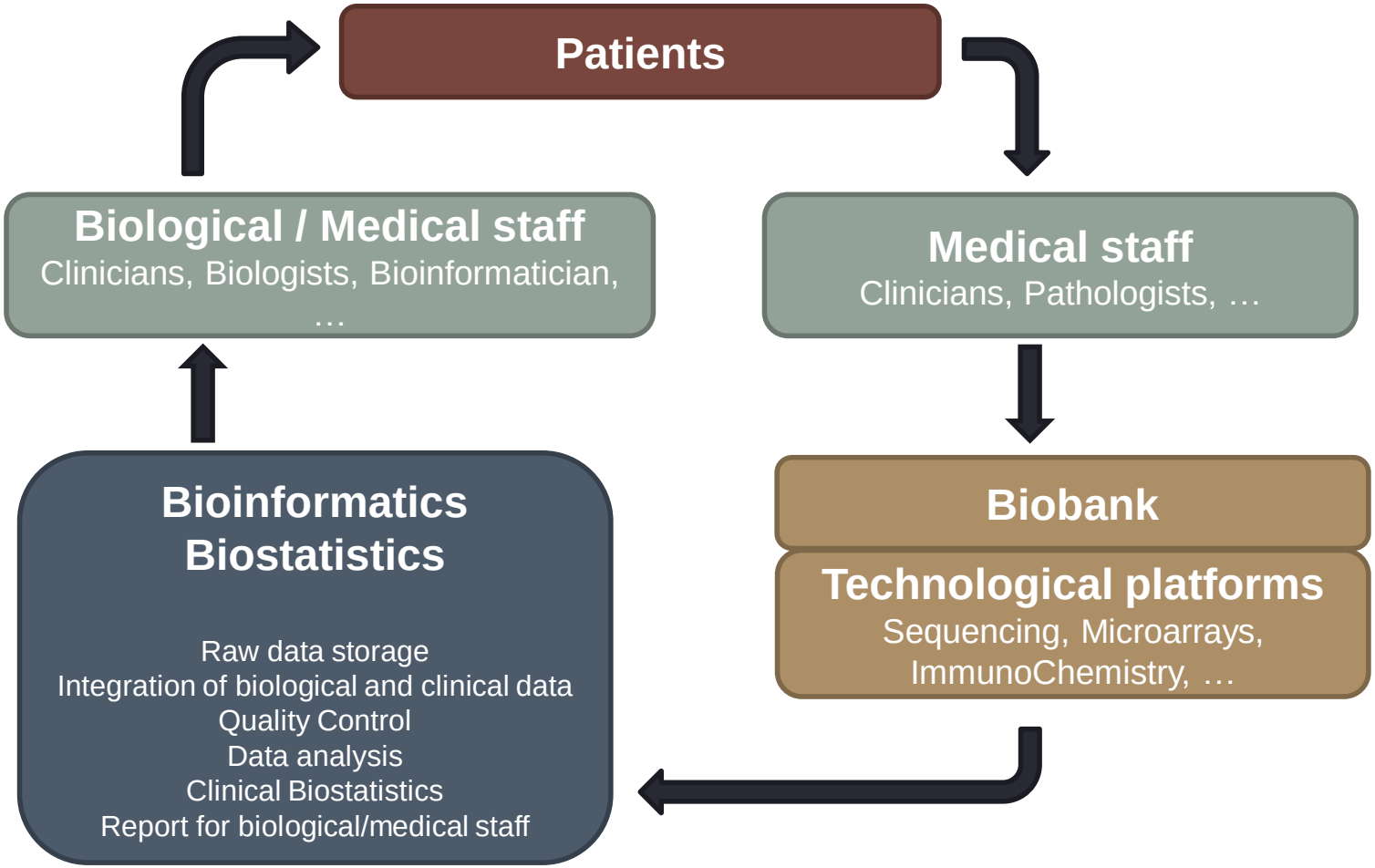
Overby and Tarczy-Hornoch, 2013

# Bioinformatics Challenges for PM

- Data storage and management
- Computational ressources
- Define dedicated quality controls
- Define appropriate analysis workflow
- Define variations of interest for clinical decision
- Interdisciplinary collaboration
- Technical, organisational, legal and scientific levels
- Integration of clean public data

# Interdisciplinary Collaboration

Bioinformatics acts as a hubs between the different fields. Trust between partners is needed, training is needed as well for efficient understanding.



# Data storage and Computing Ressources

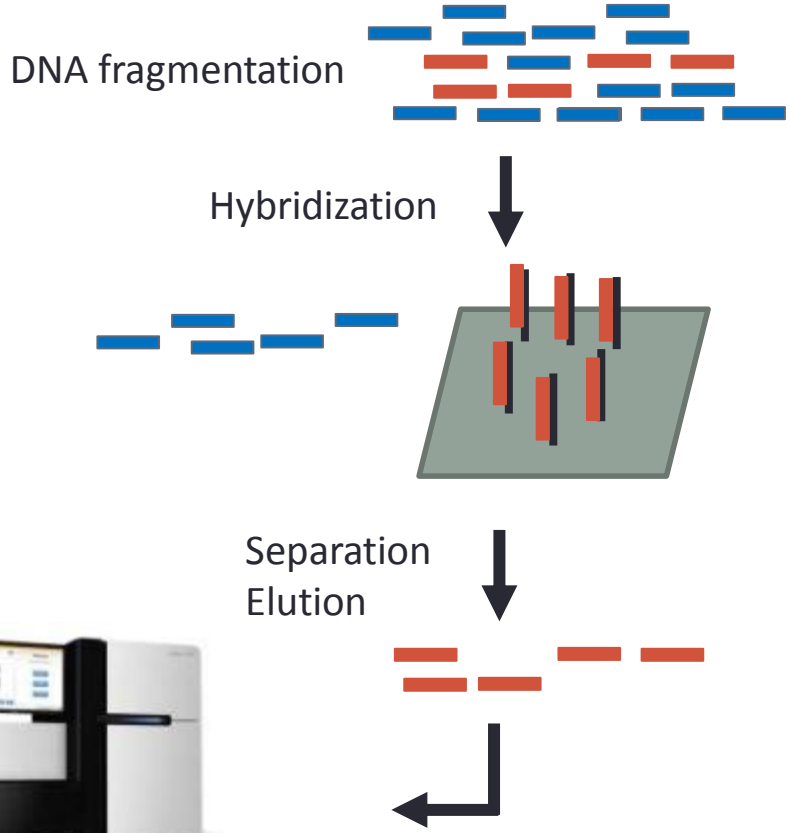
- NGS technolgies generate a huge amount of data
- High computational ressources and large data storage are mandatory
- If clinical applications are today mainly focus on targeted sequencing, next level application (exome, whole genome) should arrive very soon

- Policy of clinical data storage are not yet defined
- Acces to high performance computing ?

# Next Generation Sequencing for clinical assay

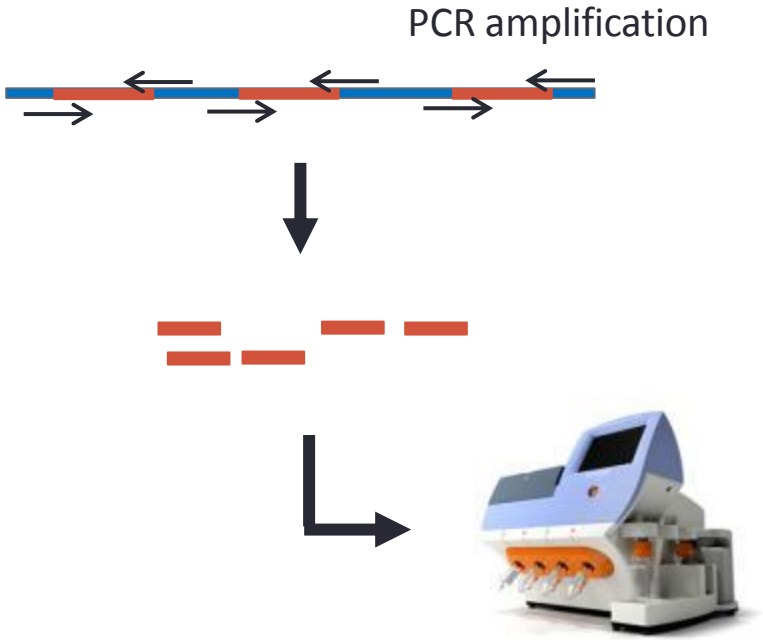
## EXOME SEQUENCING

Sequencing of all captured coding regions

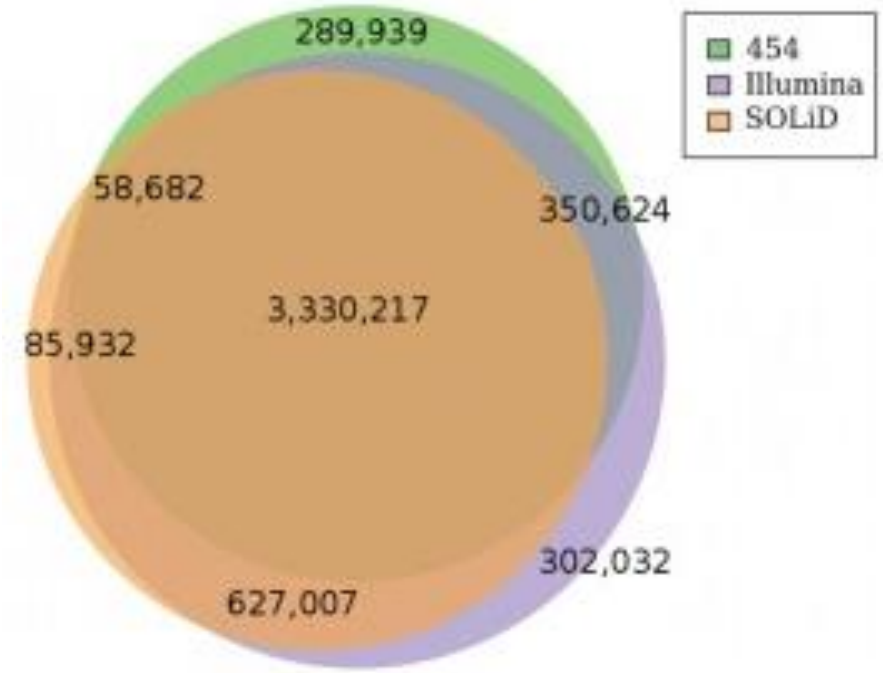



## AMPLICON SEQUENCING

Sequencing of a dedicated panel of genes/hotspots



# Ability to detect SNVs

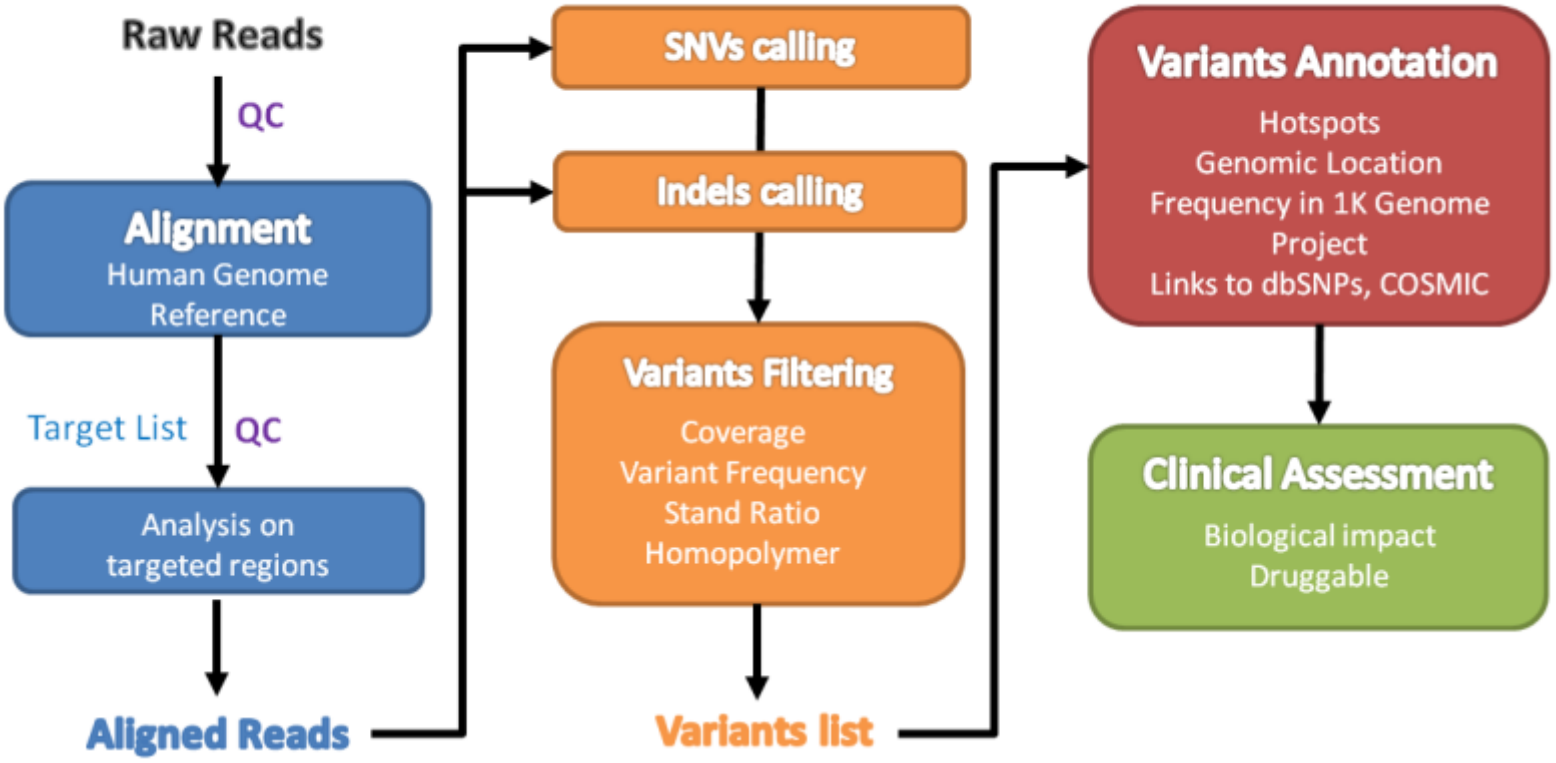



**Between 1.4 and 8.9 % of the variants are technology specific**

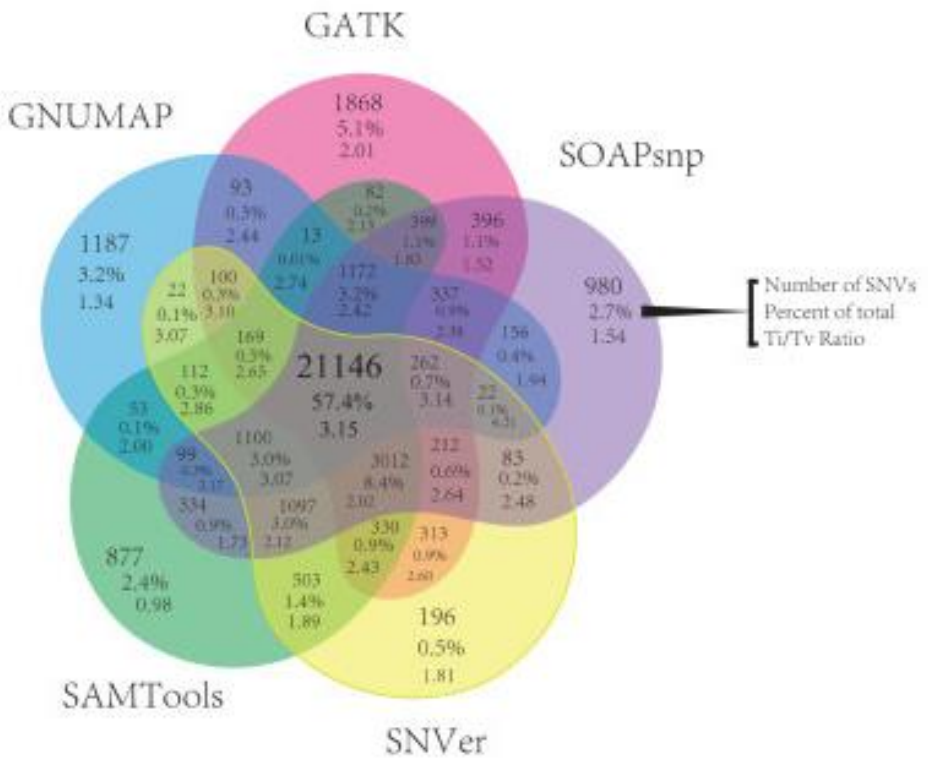
Coverage, alternate allele, variant calling

Ratan et al. PlosOne 2013

# Bioinformatics analysis workflow



# Stability of Bioinformatics Workflows




**Results on 15 exomes using 5 differents bioinformatics pipelines**

O'Rawe et al. 2013

# Challenges for bioinformatics analysis

- A significant number of false positive variants (stretch, low frequency, sequencing error, …)
- A large number of available bioinformatics tools to call variants
- A large number of detected variants
- Variable tumor cellularity and heterogeneity

# Clinical Utility

**NGS data density = frequently call variants of unknown significance**

- Which variants are clinicaly actionable ?
- Risk of over-interpretation not necessary for medical treatment
- Careful selection of patients for genome sequencing and genetic counseling is crucial

**Biology is the key of the personalized medicine assays**

# NGS based Diagnosis at Institut Curie

First projects in 2012 at Institut Curie

Tends to increase in the coming months

Design specific tools for diagnosis (Specificity/sensibility)

**BRCA Diagnosis at Institut Curie** (C. Houdayer, Institut Curie)

→Test of enrichment methods (Agilent, HaloPlex, RainDance, Multiplicom, PCR)

→Test of sequencing machine (SOLiD, Ion Torrent PGM)

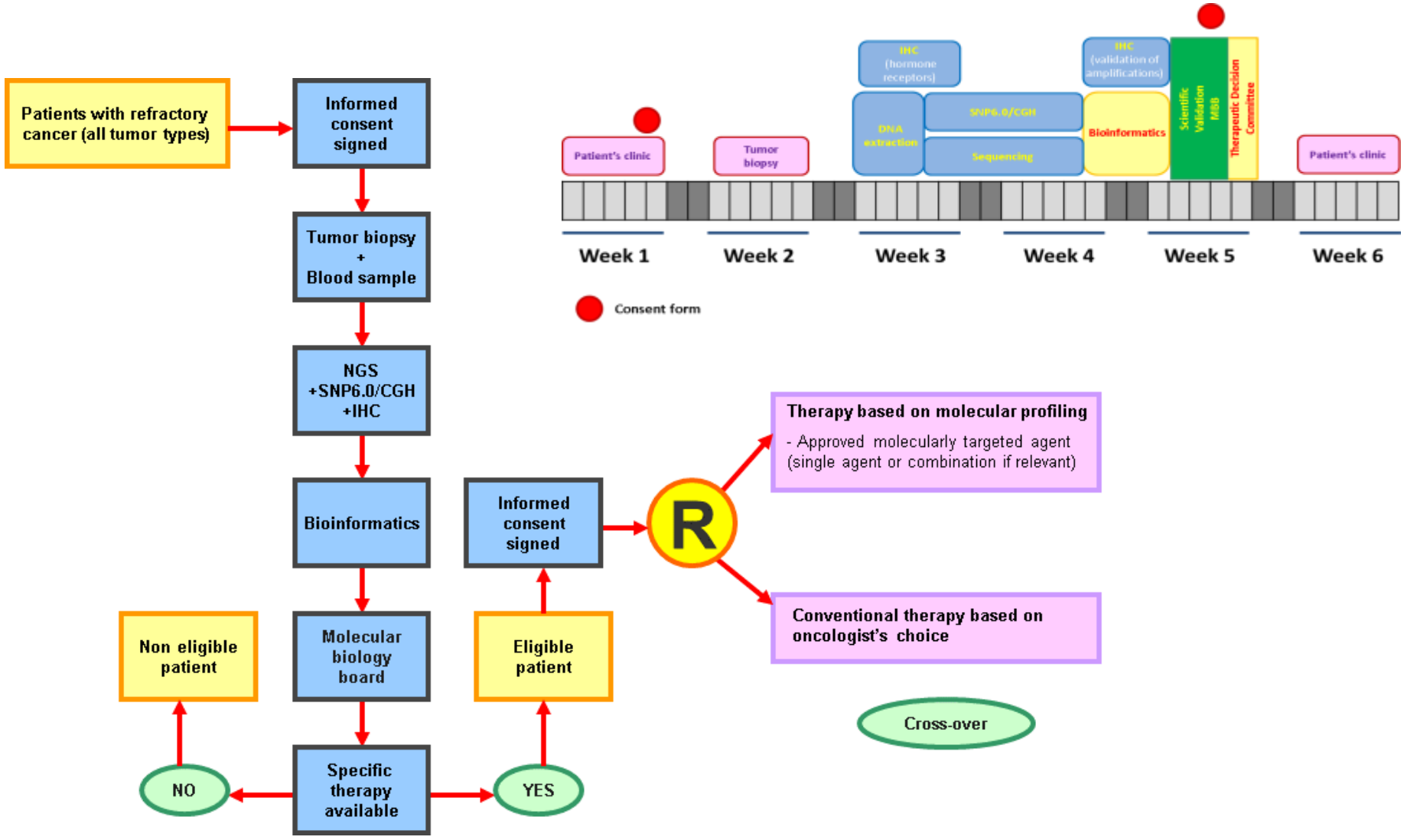
→Test of Bioinformatics solutions (NextGene, Home made pipeline)

# The SHIVA clinical trial

**SHIVA: A phase II clinical trial** (C. Letourneau, Institut Curie)

- A **multicentric** open randomized **phase II trial** involving patients (adults and children) with refractory cancer to standard treatment
- Molecular profiling based on IHC, Affymetrix Cytoscan arrays, and Ion Torrent PGM sequencing
- NGS in daily pratice. Real time analysis
- More than 400 included patients

# Design of the SHIVA clinical trial



# TAKE HOME MESSAGE

# Take home message

- The previous list of applications is not exhaustive. Sequencing is expanding very fast
- Exome/RNA/ChIP are the most widely used applications
- Normal samples/controls are mandatory for any kind of applications
- Replicates are also now strongly recommanded especially for quantitative purpose
- « Personalized medicine » is now a reality
- Clinical NGS requires dedicated bioinformatics worflow and visualization tool

# Many Thanks

**INSERM U900**

Valentina Boeva

Bruno Zeitouni

Aurélie Teissandier

Romain Daveau

Alban Lermine

Séverine Lair

Chongjian Chen

Emmanuel Barillot

**Collaborators**

Edith Heard

Serena Sanulli

Constance Ciaudo

Christophe LeTourneau

Maud Kamal