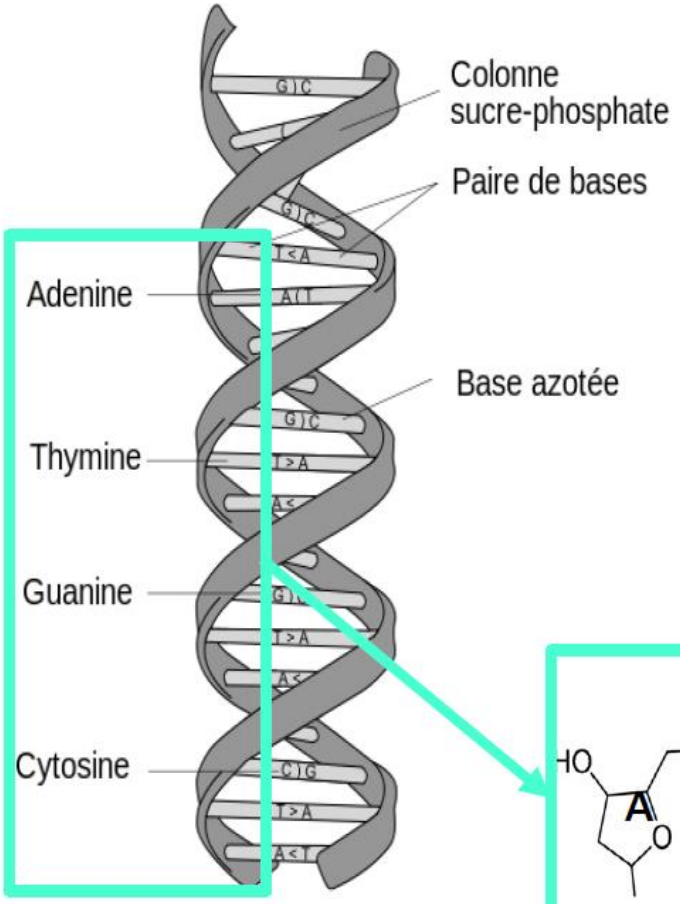
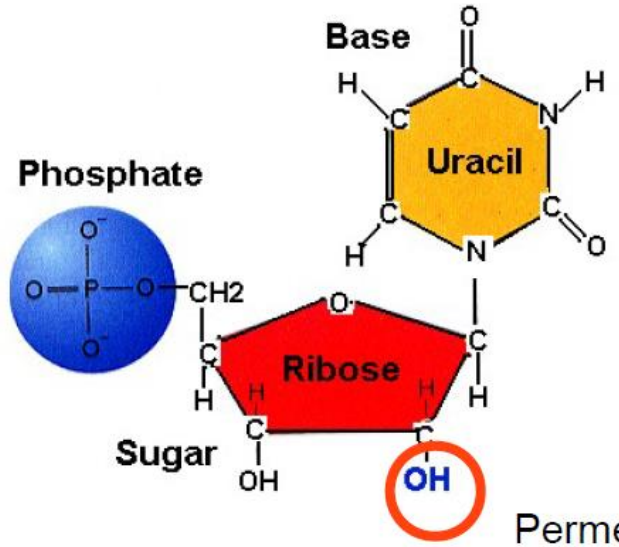
# ChIP-Seq / RNA-Seq

Kevin Cheeseman
Damien Ulveling
Olivier Kirsh

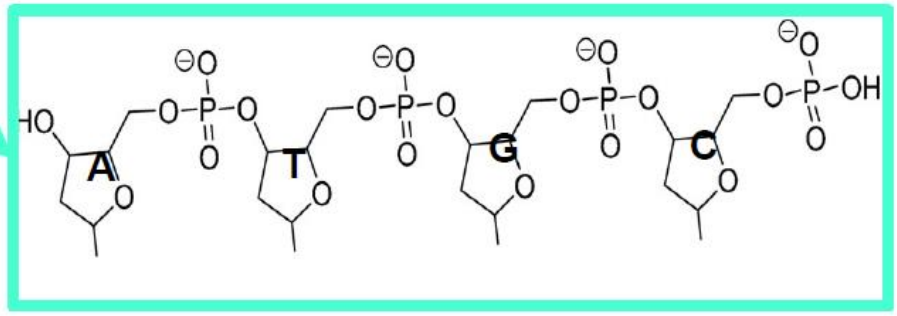
# SEQUENCING INTRODUCTION

# DNA STRUCTURE

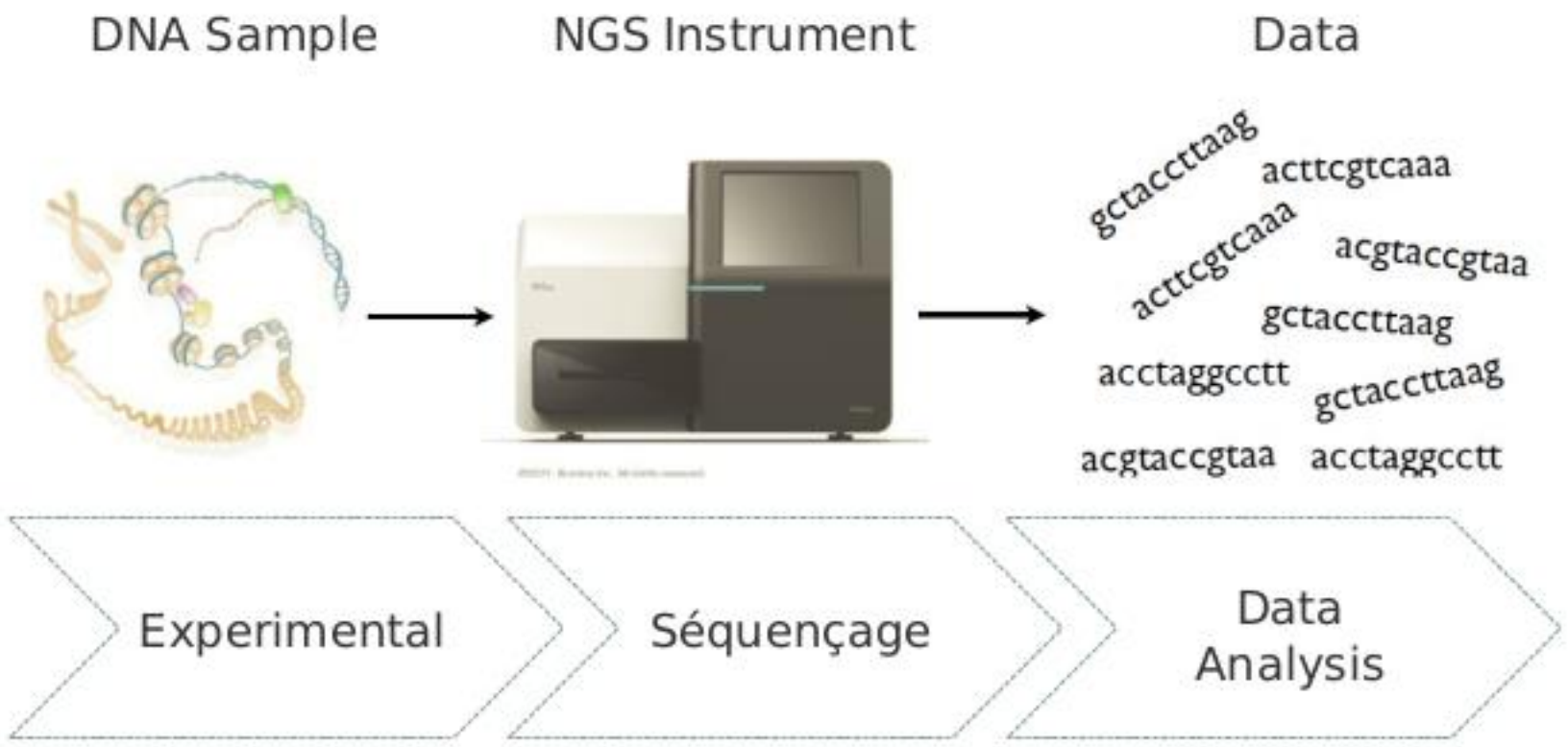



Permet la liaison
nucléotide à nucléotide



# SEQUENCING OVERVIEW

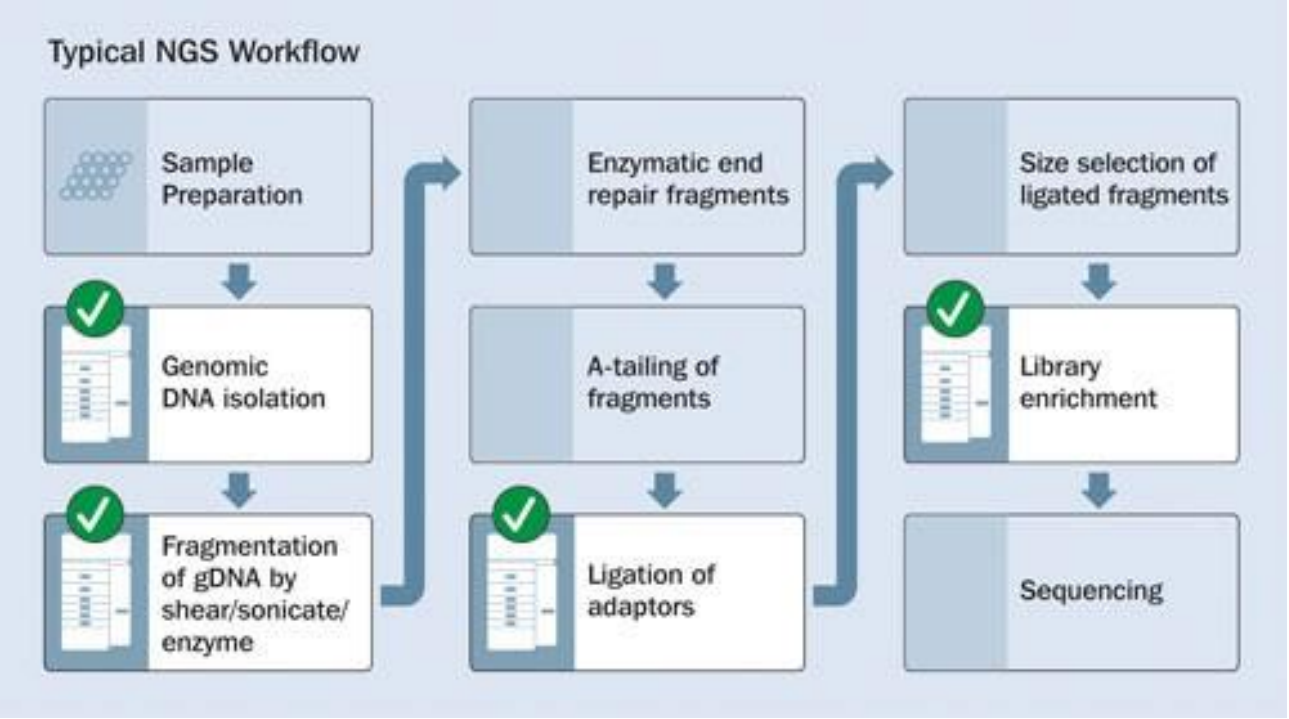


# NGS: COMPARISON

| 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 | 200 | >12 | 2-4h | Base |

# LIBRARY

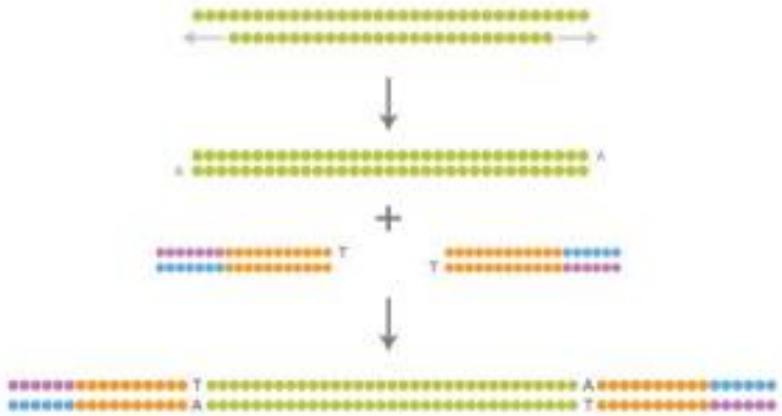
A library is a collection of DNA fragments representative of cellular state, genome,development stage etc…




http://www.genengnews.com/gen-articles/dna-library-qualification-quantification/3960/

# LIBRARY PREPARATION

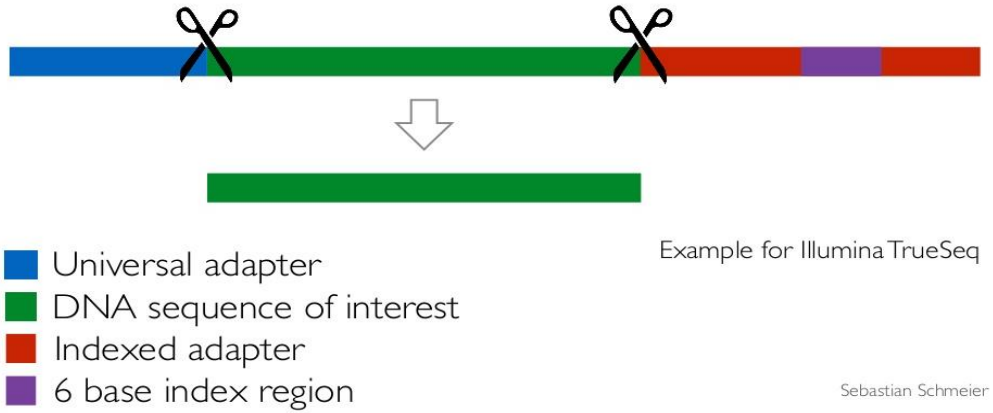
- Exemple par Illumina



1. Selection des fragments (experience)
2. Ajout d'un A sur les bouts francs en 3'
3. Ligation des adaptateurs grâce au T

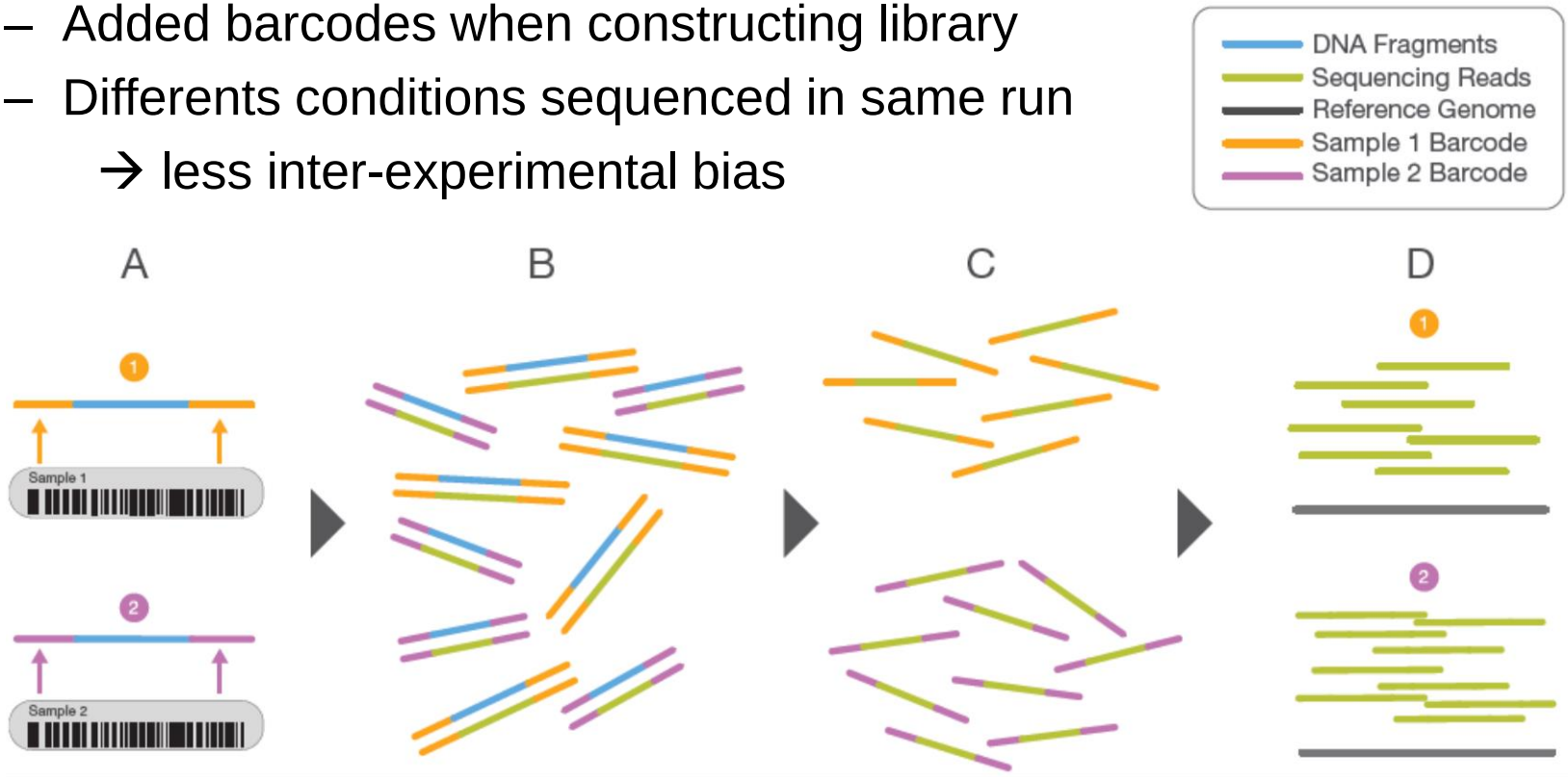
# ADAPTERS

- Known sequences at ends of reads
- Role
  - Link reads to the flowcell
  - Reads specific PCR
  - Barcodes used with multiplexing
- Usely removed by sequencer
- Tools
  - Cutadapt
  - Trimmomatic
  - RmAdapter etc…



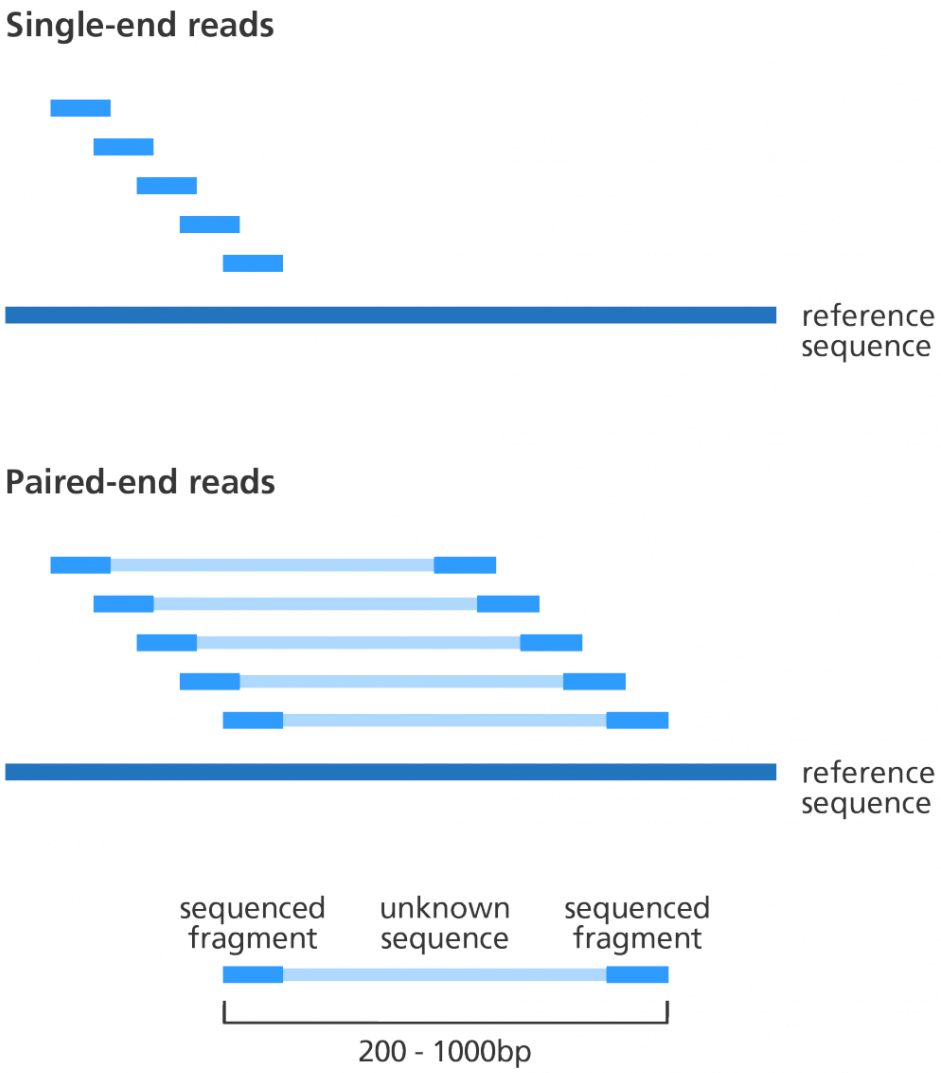
# MULTIPLEXING

- Parallel sequencing:
  - Added barcodes when constructing library
  - Differents conditions sequenced in same run
    → less inter-experimental bias



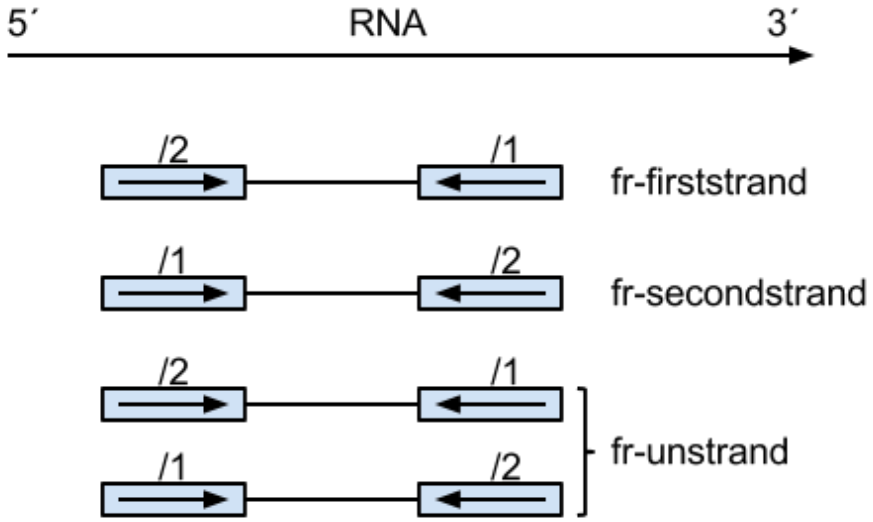

www.illumina.com/Documents/products/Illumina_Sequencing_Introduction.pdf

# DIFFERENTS LIBRARIES

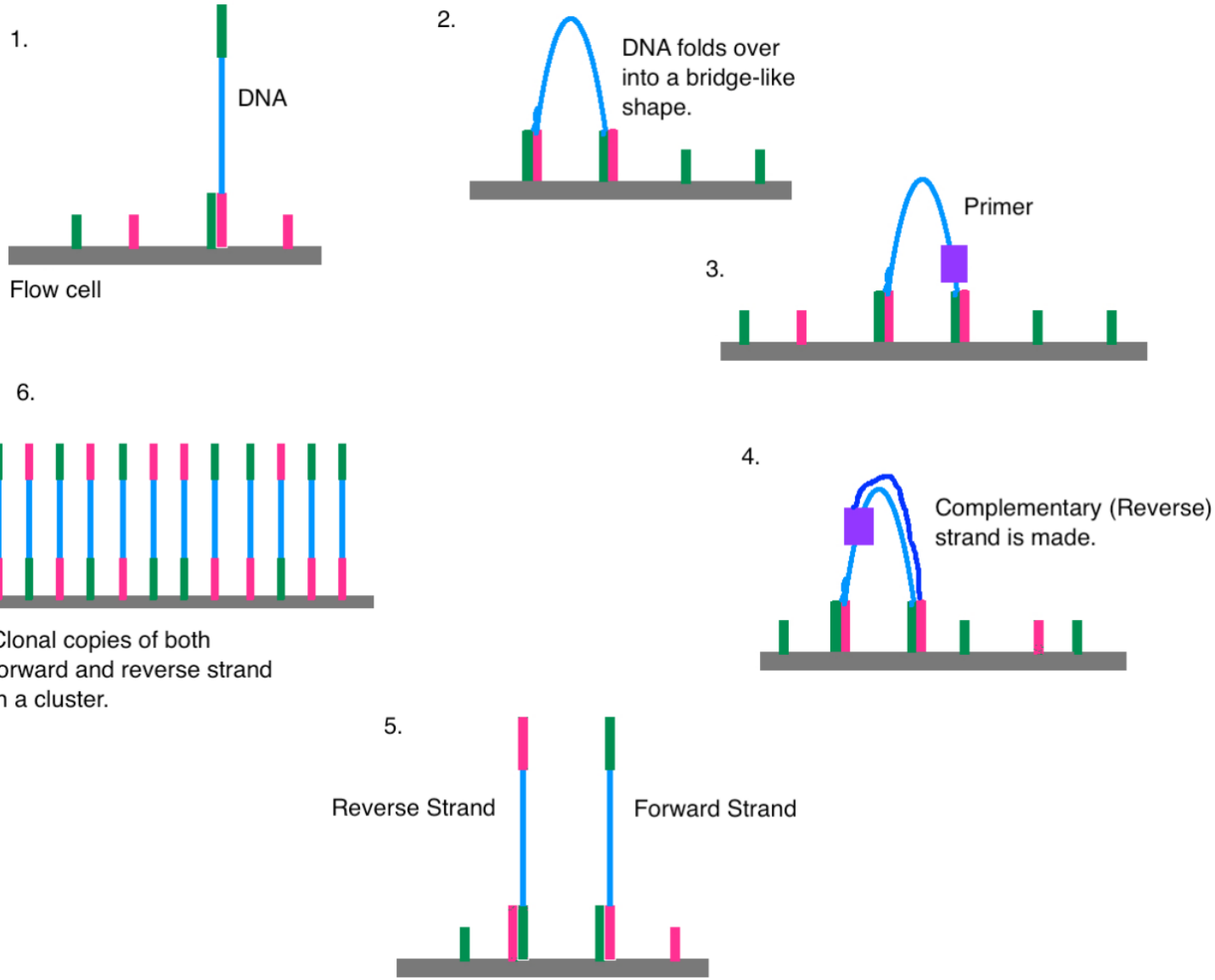


# ORIENTATION PROTOCOL

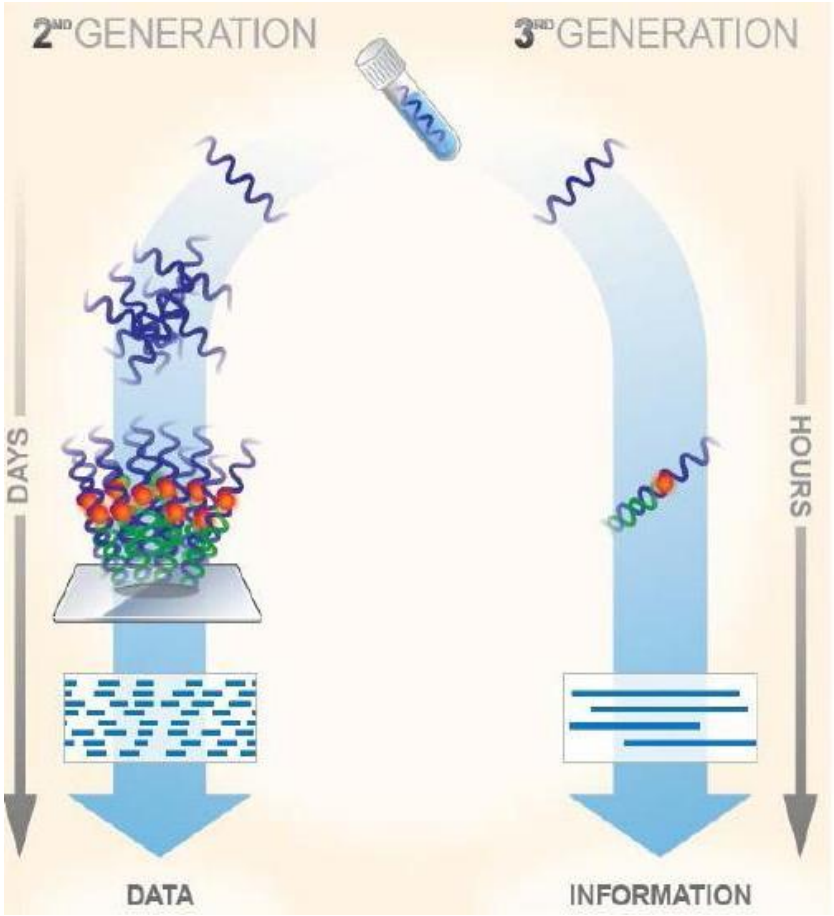
- Most of RNA protocols are directionals
- This information is very important for further analysis



# NGS: ILLUMINA



# NNGS: Single Molecule Sequencing (SMS)



# NNGS: Oxford Nanopore

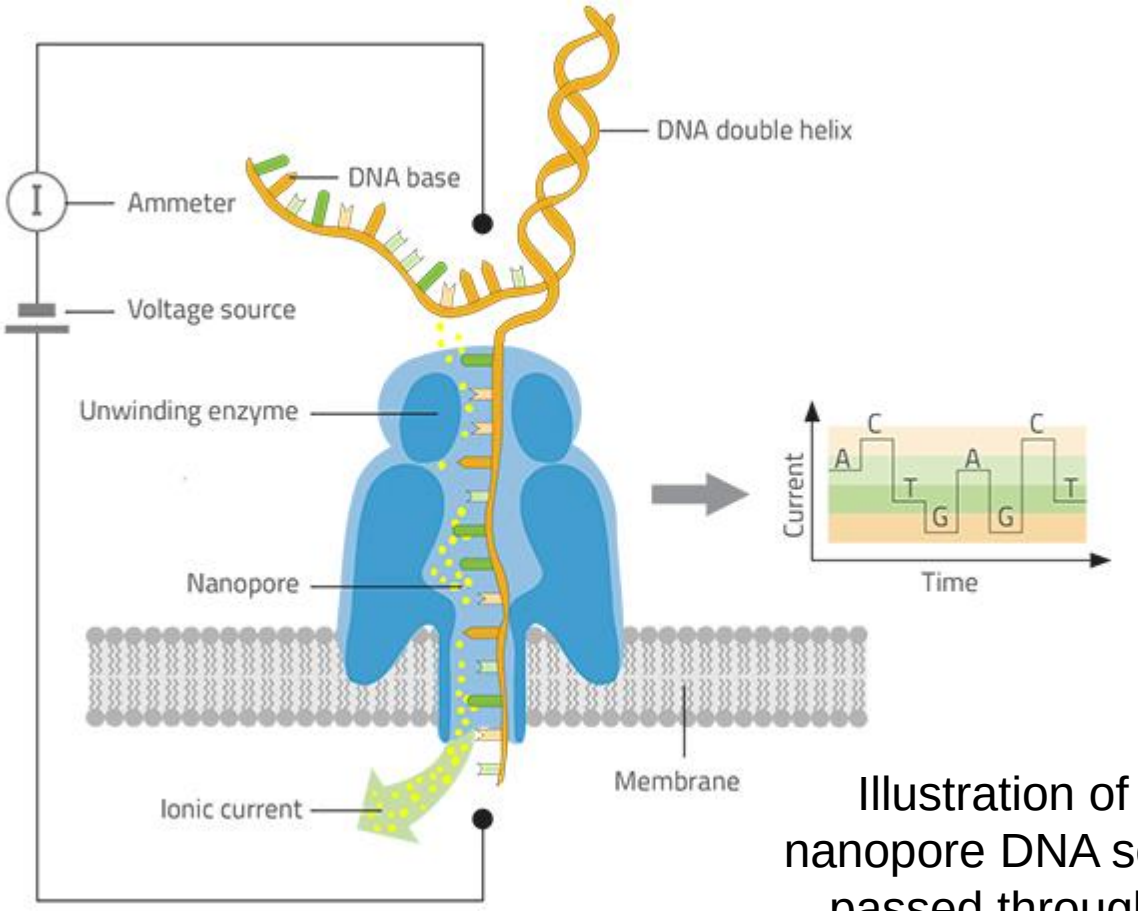
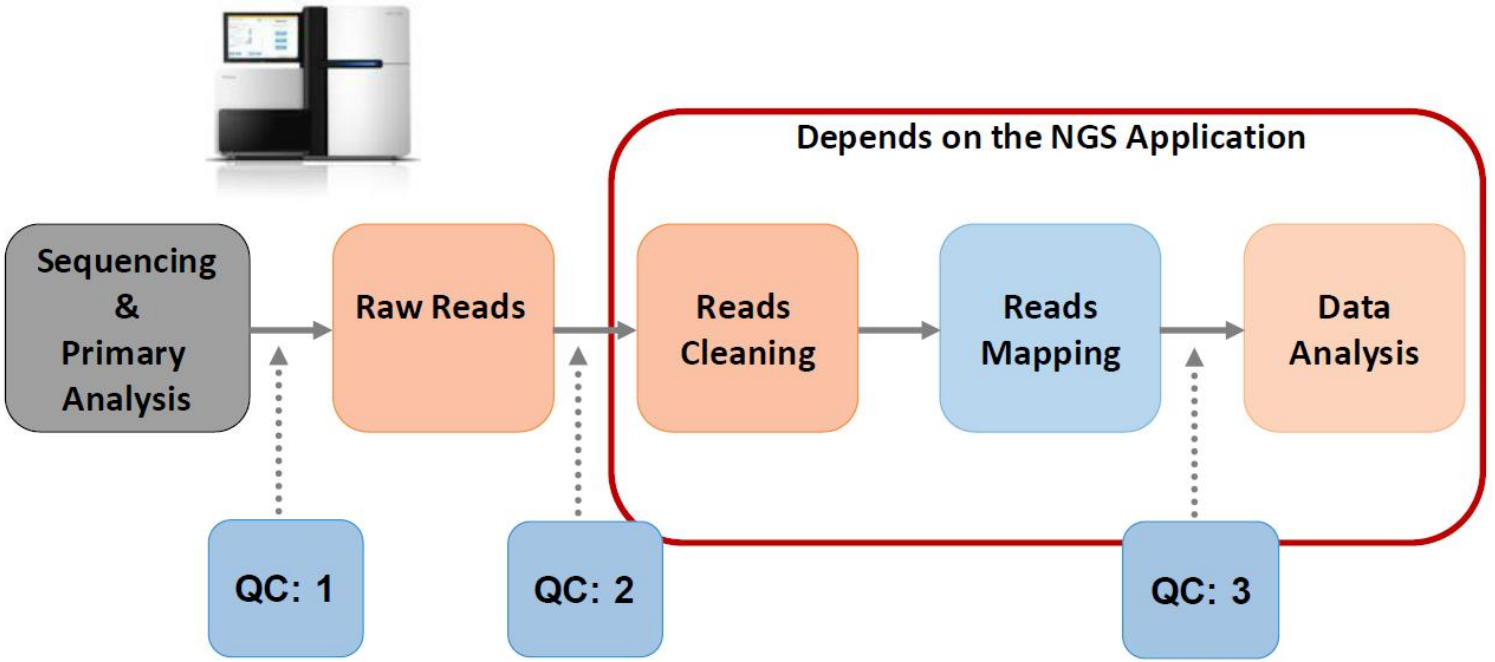



Illustration of the working principle of a nanopore DNA sequencer. A strand of DNA is passed through a nanopore, and an ionic current is measured and translated into the DNA sequence.

Kerstin Göpfrich

# RAW DATA PROCESSING

# STANDARD PIPELINE

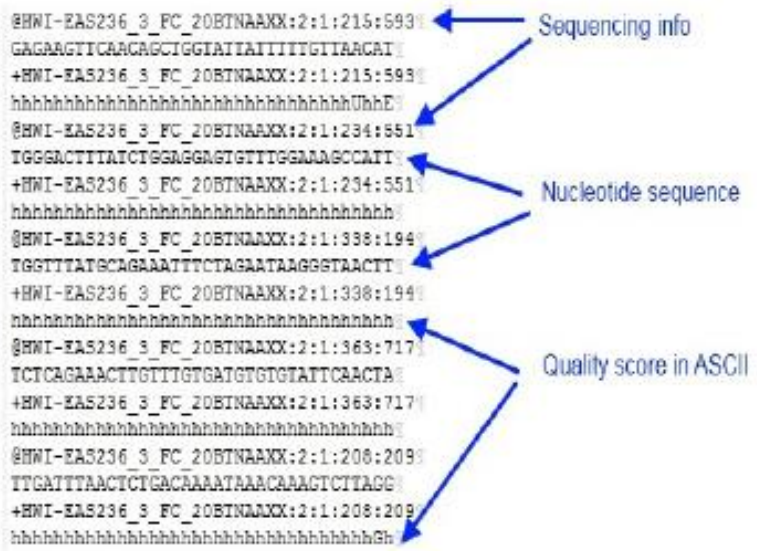


# RAW DATA

- Raw files (.fastq, .fa etc…) with millions of tags (reads) with same size (SOLiD, HiSeq)

# FASTQ FORMAT

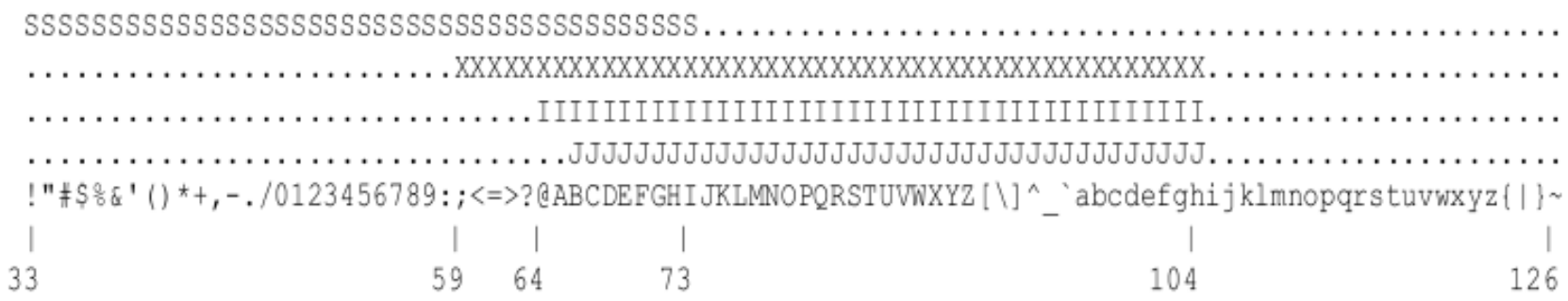
- Block of 4 lines per read:
  - Read ID
  - Read sequence
  - Comment
  - Quality PHRED score




$Q_{Phred} = -10 \log_{10} P(\text{error})$.

| PHRED score | Probability of error |
|---|---|
| 10 | 0.1 |
| 20 | 0.01 |
| 30 | $10^{-3}$ |
| ... | |
| 60 | $10^{-6}$ |

# ENCODING QUALITY




S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)

# QUALITY CONTROL (QC)

- To check
  - Statistics: amount, size of reads …
  - Per base quality: N presence?
  - Overrepresentation of reads: contamination?
  - GC%, ATCG proportion
  - Linkers, Adapters, barcodes, UMI etc…

# QC: TOOLS

- FASTQC = html report of fastq files
- PRINSEQ
- Multi-QC = aggragate bioinfo results in html report

PRIN
SEQ

MultiQC

# QC: METRICS

- Distribution of reads lengths
- ATCG proportion, GC%
- Quality score per base along reads
- K-mers content
- Overrepresented sequences
- Read duplication
  - Technical or biological effect?
  - To mark or remove up to experiment design

# FASTQC

## FastQC Report

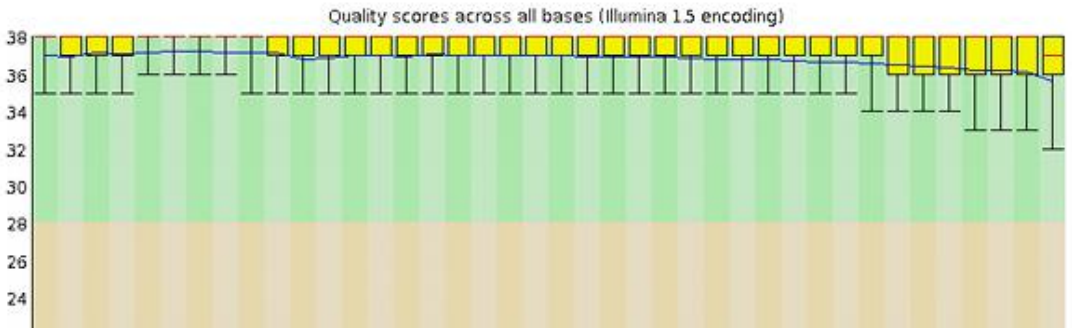
### Summary

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

### Basic Statistics

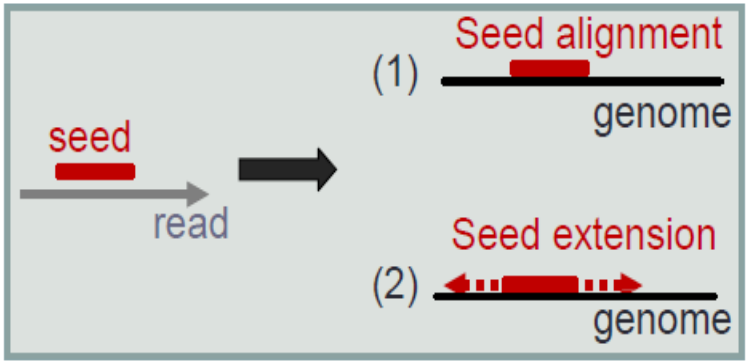
| Measure | Value |
|---|---|
| Filename | good_sequence_short.fastq |
| File type | Conventional base calls |
| Encoding | Illumina 1.5 |
| Total Sequences | 250000 |
| Filtered Sequences | 0 |
| Sequence length | 40 |
| %GC | 45 |

### Per base sequence quality



# ALIGNMENT VOCABULARY

- **Mapping method**: seed & extend
  1/ Align the seed (small part of read)
  2/ Extend the seed to align the whole read

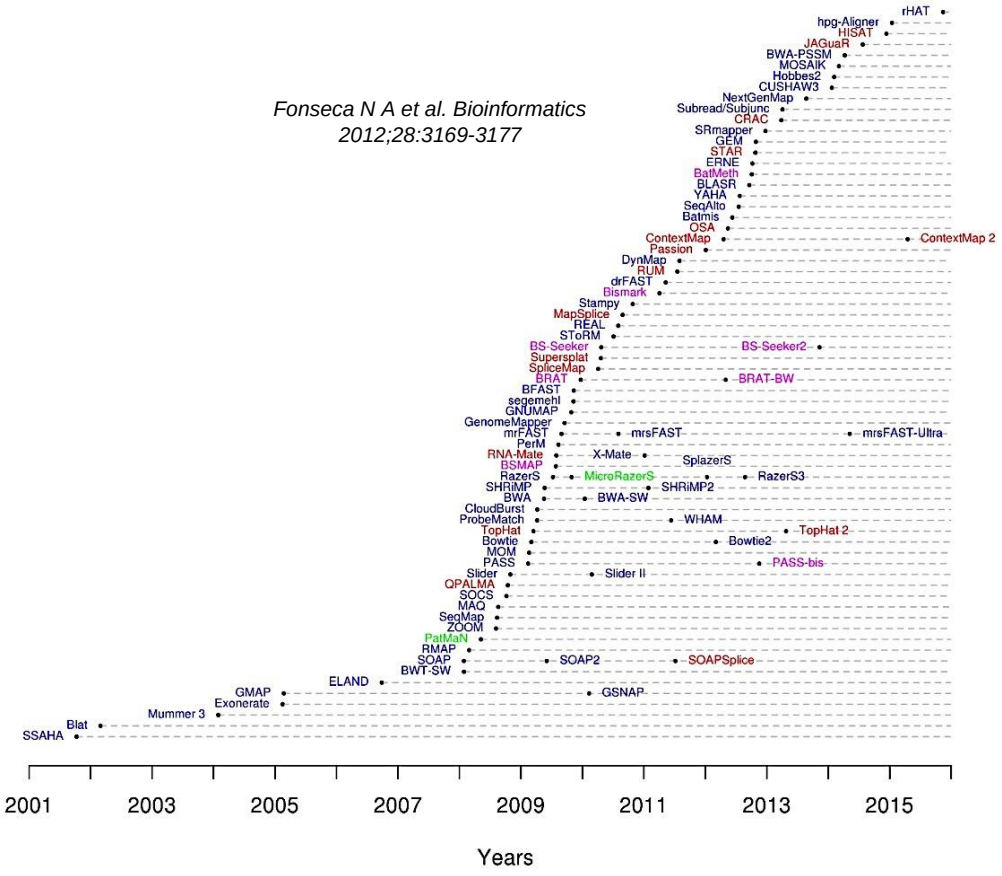



- **Mismatch**: Incoherence between ref/read
- **Gap**: Bridge within the read alignment
- **Mappability**: Uniqueness of a region
  - –Repeated region = low mappability
  - –Unique region = good mappability

# ALIGNMENT: ISSUES

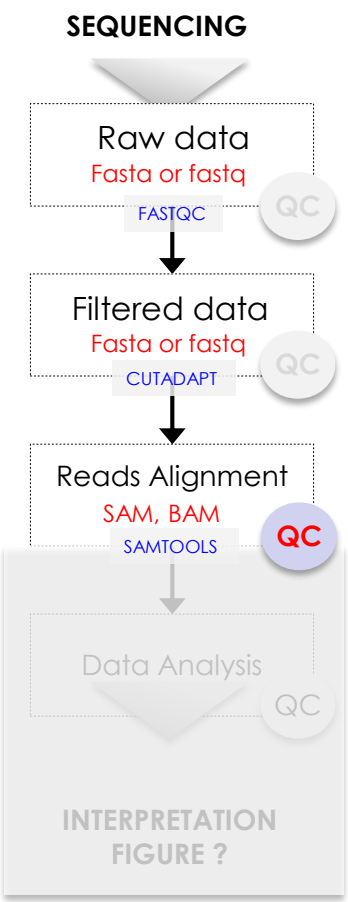
- Million reads / run
- Small length of the read compared to genome
  - human $10^9$
- Type of reads sequenced (SE, PE etc…)
- Sequencing error
- Repeated regions etc…

# ALIGNMENT: TOOLS

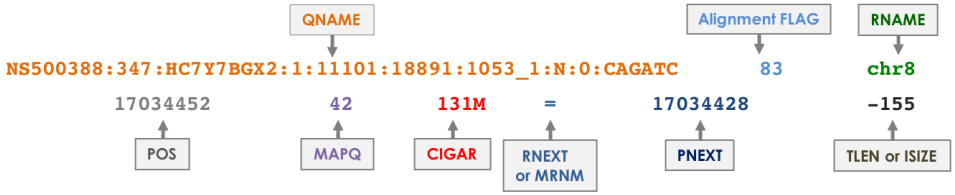


# SAM/BAM FORMAT

## SAM File : Body



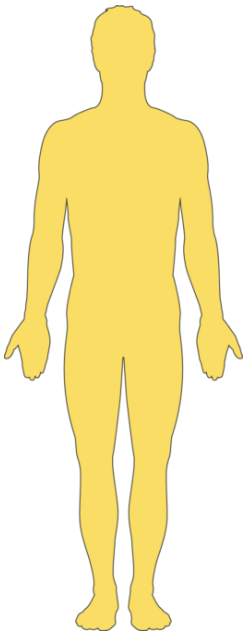
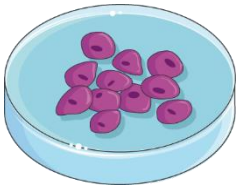

**SAM** = **S**equence **A**lignment **M**ap format (readable)




| | | |
|---|---|---|
| 1 | QNAME | ID of the reads |
| 2 | FLAG | Alignment flag (**specific here** : take the strand a bisulfite read originated from into account) |
| 3 | RNAME | Reference name (Chromosome name) |
| 4 | POS | Position in reference (5' position) |
| 5 | MAPQ | Mapping quality |
| 6 | CIGAR | Alignment description in CIGAR format (M:match ; I:insertion ; D:deletion ; S:substitution) |
| 7 | RNEXT or MRNM | Name of reference sequence where mate's alignment occurs. Set to = if the mate's reference is the same as this alignment, or * if there is no mate |
| 8 | PNEXT | Left most position of where the next alignment in this gap maps to the reference |
| 9 | TLEN or ISIZE | Total length or Insert size (for PE) |

https://samtools.github.io/hts-specs/SAMv1.pdf

# NGS APPLICATIONS

| Sample | Technology | Analysis | Integration / Interpretation |
|---|---|---|---|
|  | Genomic WGS/WES | SNP<br>Indels<br>CNV<br>SV | Effet fonctionnel |
| | Transcriptomic RNA-seq | DEG<br>Gene fusion<br>Alternative Splicing | Analyses : réseaux et pathways ... |
|  | Epigenomic ChIP-Seq Bisulfite-Seq | TFBS<br>Histone marks<br>Methylation | Analyse intégrative<br>etc... |

# NGS: STANDARDS

| **Category** | **Application** | **Reads Coverage** (**M**illion or **X**) |
|---|---|---|
| WGS | SNP/indels | 30 X / 60 X |
| WES | SNP | 100 X |
| Transcriptomic | DEG | 40 M |
| | Alternative splicing | 100 M |
| | De novo assembly | > 100 M |
| DNA target | ChIP-Seq | 15 M (narrow peak)<br>40 M (broad marks) |
| DNA methylation | Bisulfite-Seq | 30 X |

Adapted from https://genohub.com/recommended-sequencing-coverage-by-application/

# ChIP-Seq

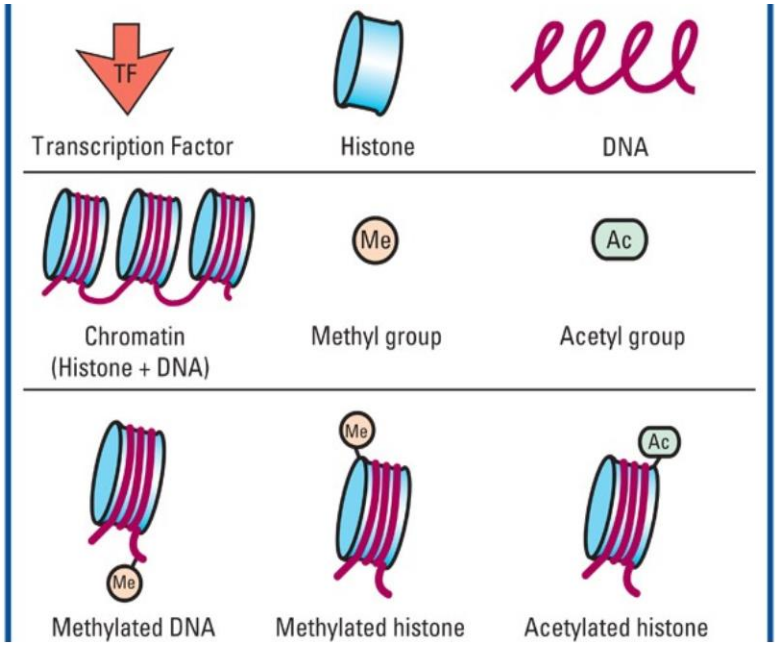
**Ch**romatine

**I**mmuno**P**recipitation

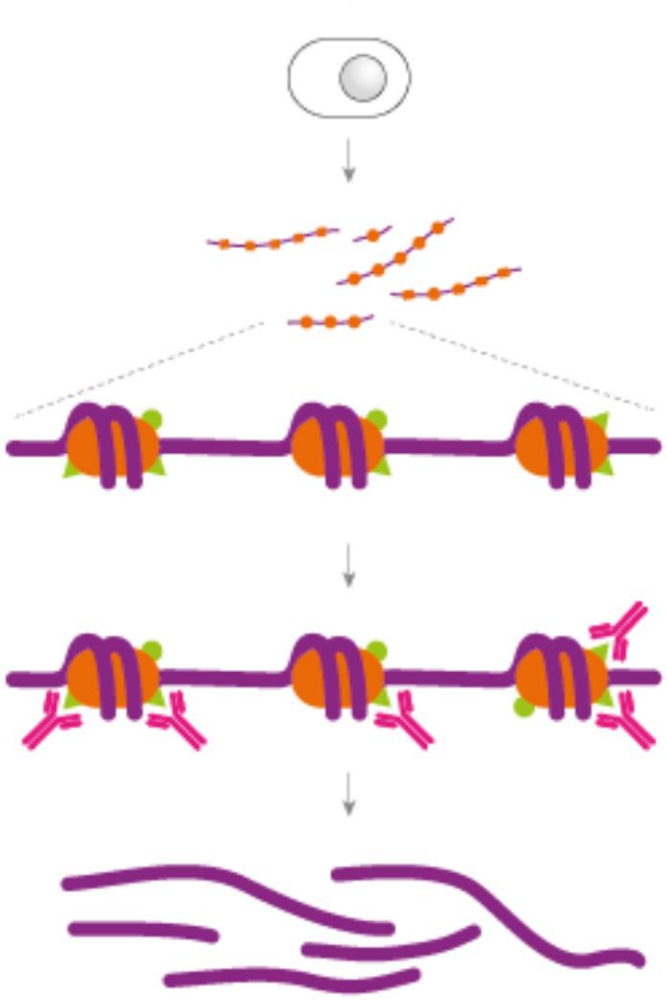
**Seq**uencing

# Chip-Seq: AIMS

- Approach to detect DNA sequences/regions associated to a targeted protein
- Interaction analysis DNA / protein (detection of signal/modification associated to DNA / chromatine)
- Identification of genomic targets for transcriptional factors / DNA binding protein



# Chip-Seq: PRINCIPLE



1- crosslink gDNA/protein

2- gDNA isolation

3- gDNA sonication

4- Specific antibodie IP

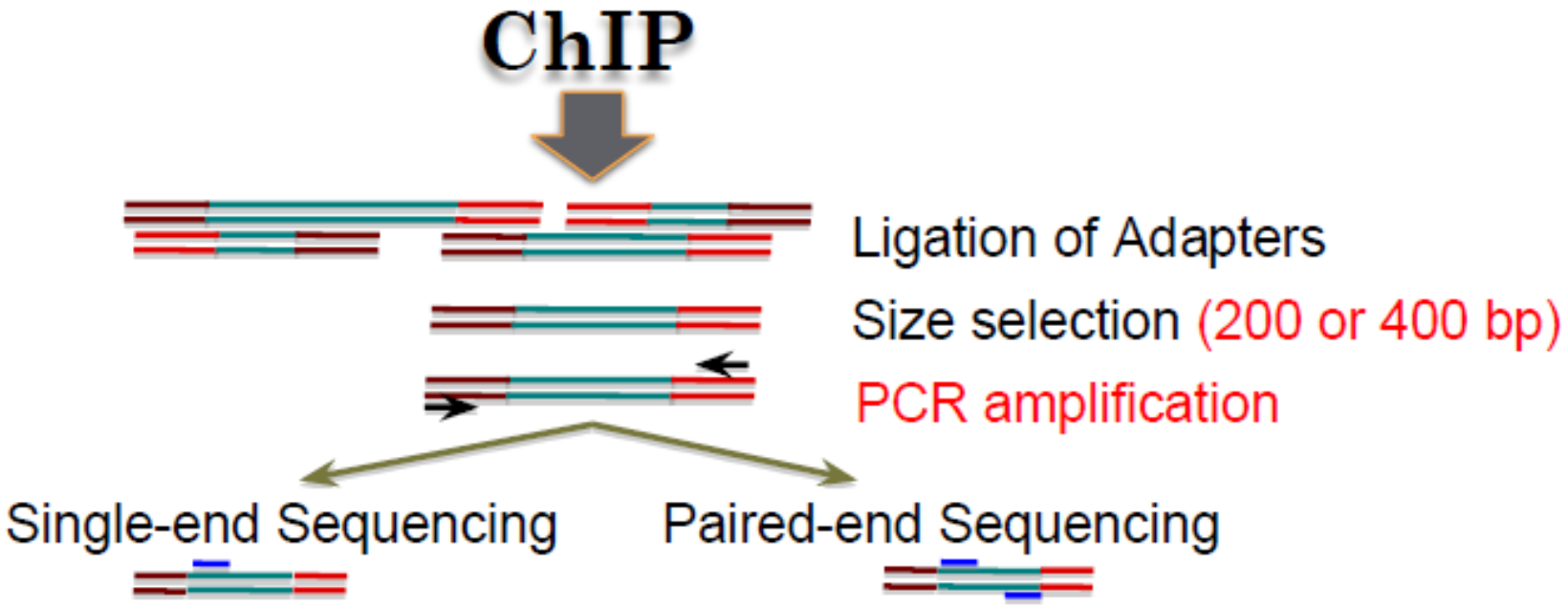
5- Reverse crosslink

6- DNA purification

7- Sequencing library preparation

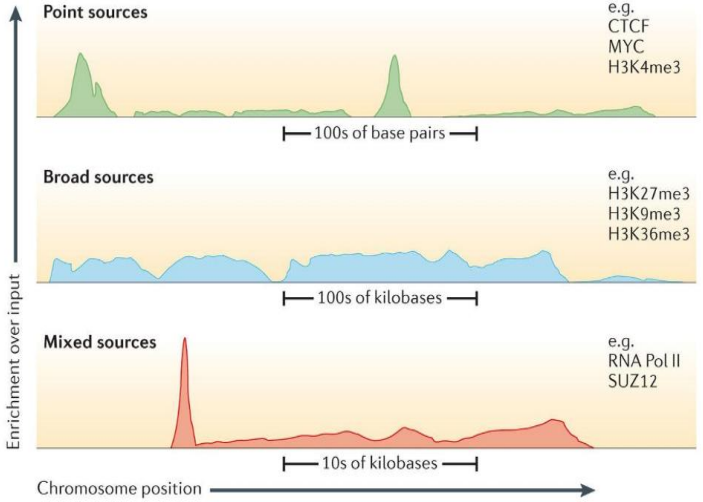
# ChIP-Seq: LIBRARY PREP

- Step between ChIP and sequencing
- Prepare DNA for sequencing
- Starting material: ChIP sample (1-10ng of sheared DNA)



# Chip-Seq: DEPTH

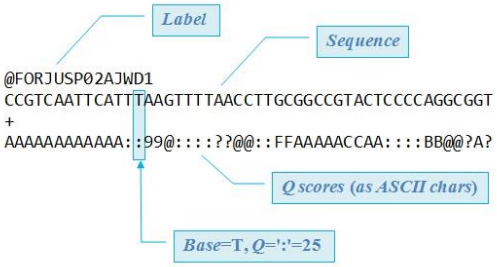
- Depth depend on:
  - Targeted protein
  - Expected profil type
  - Expected number of site
  - Genome of interest size

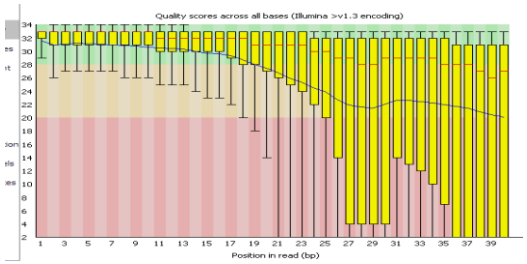


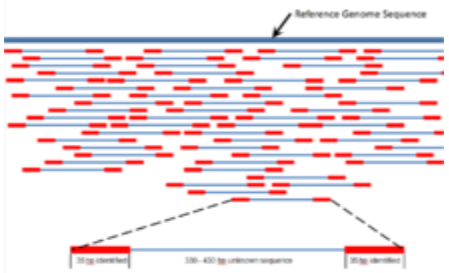

| Human | 15/20M<br>Narrow peak | 20/40M<br>Broad peak |
|---|---|---|
| Fly | 8M | |
| Worm | 10M | |
| Bacterial | 2/3M | |

# Chip-Seq: AFTER SEQUENCING

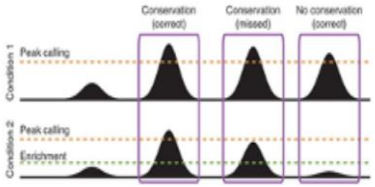
FASTQ → QC → Mapping



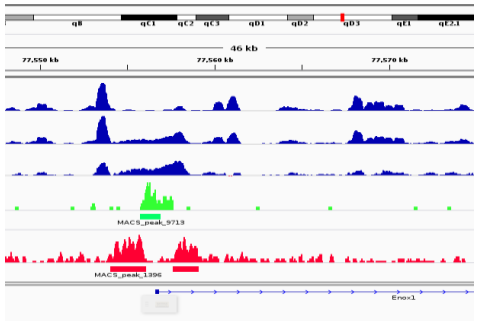




Mapping → Coverage → Peak calling




Peak calling → DOWNSTREAM ANALYSIS

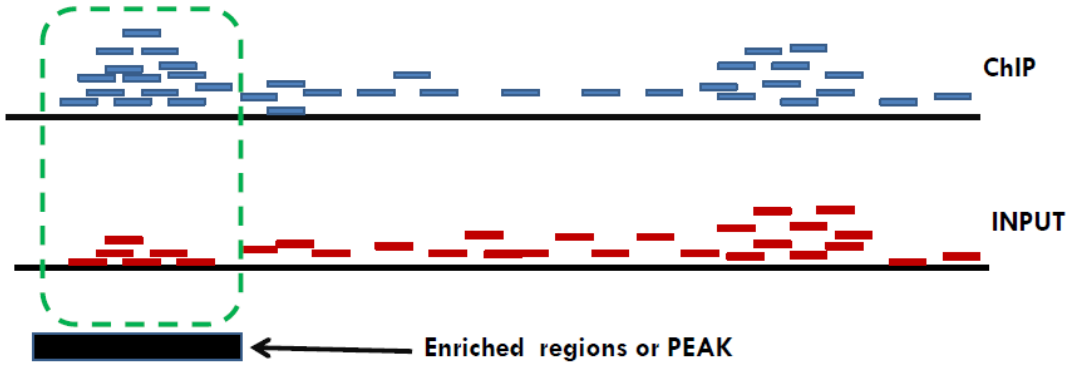


# ChIP-Seq: BEST PRACTICE

- ENCODE project https://www.encodeproject.org/about/experiment-guidelines/
  - The **ENC**yclopedia **O**f **D**na **E**lement : consortium which establish and developp NGS standard
- Replicates
  - 2 replicates minimum / experiment
  - Biological replicate, not technical!!
    - Cell culture, pool, tissue sample: INDEPENDANT

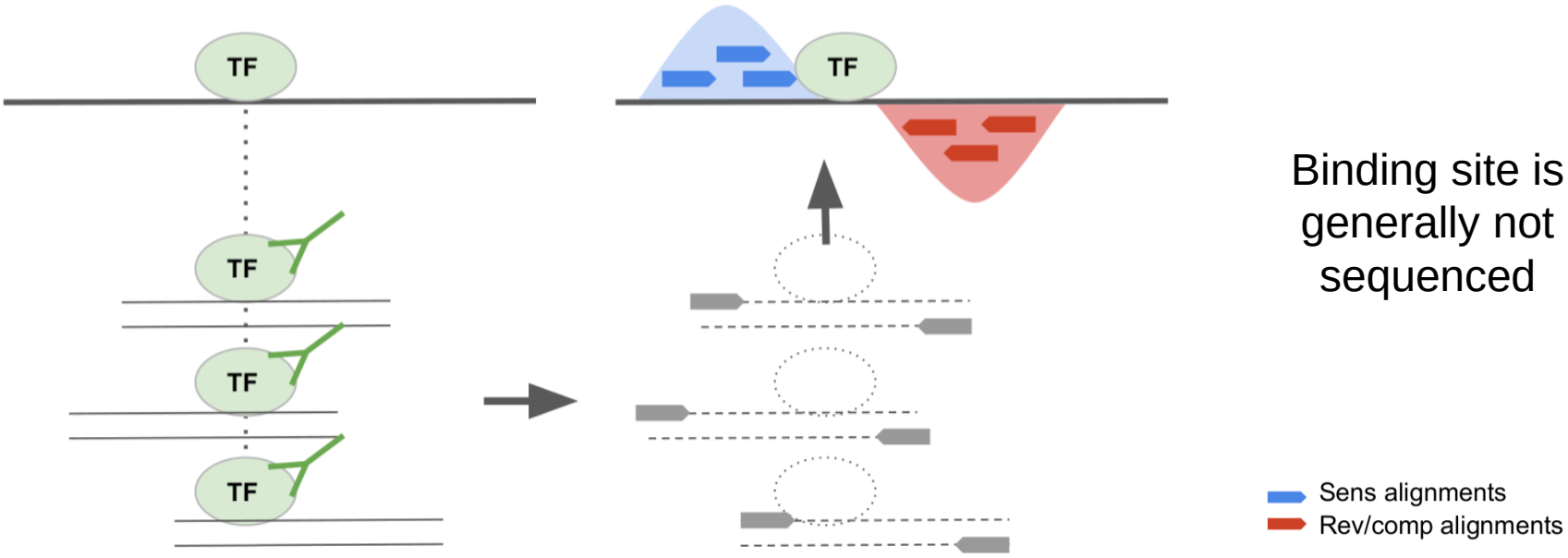
# ChIP-Seq: CONTROLS

- Mandatory to filter out false positive
  - A false positive will be enriched into the ChIP and control
- 3 control types
  - Input DNA: sample removed prior to IP
  - DNA from non spe IP: IP with antibody not known to be involved in DNA binding or chrom modif (IgG)
  - Mock IP DNA: IP without antibodies



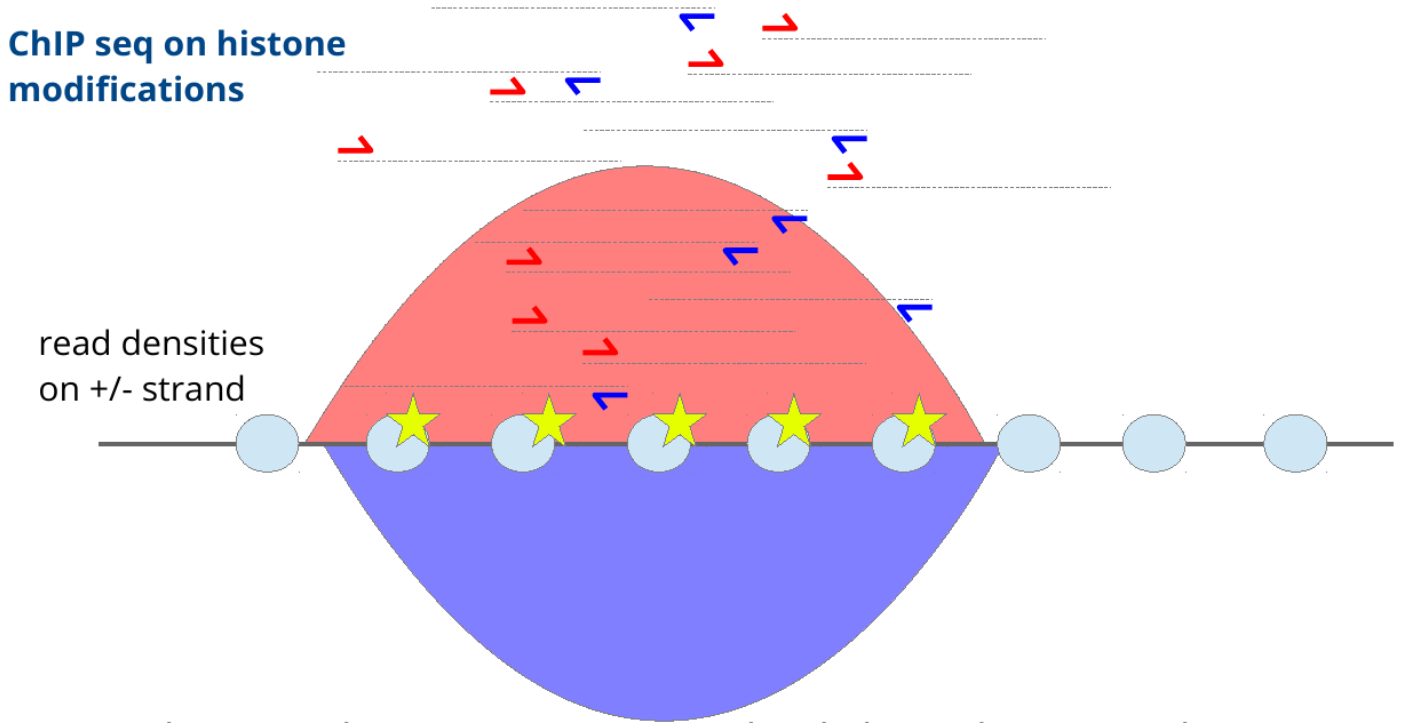
# ChIP-Seq: EXPECTED SIGNAL

- For a transcription factor signal is expected to be **sharp**

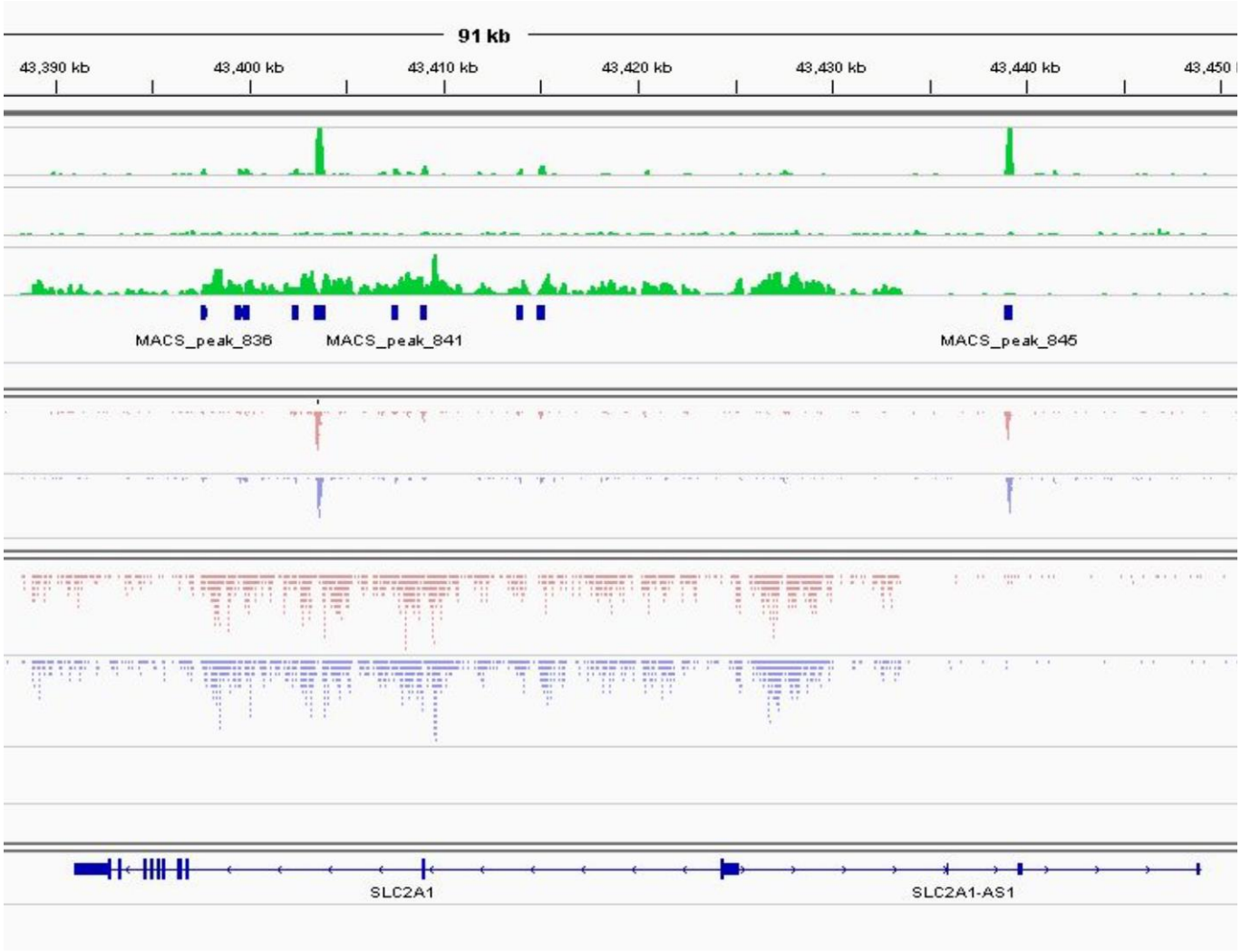


# ChIP-Seq: EXPECTED SIGNAL

- Most histone marks : **broad** signal expected
- Asymetry less/not pronounced
- Peak calling tools need to adapt to these various signals



# ChIP-Seq: OBSERVED SIGNAL




FT

Histone marks

# ChIP-Seq: ALIGNMENT

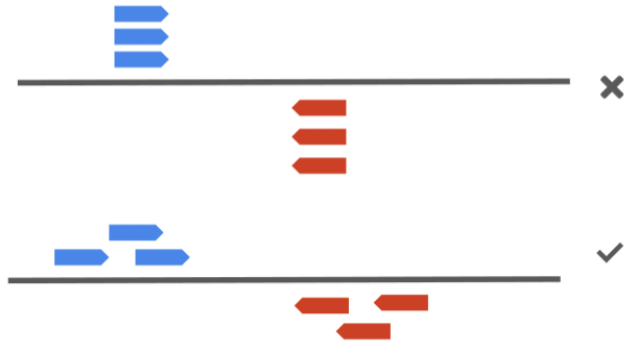
- Localization of reads in the reference genome

# ChIP-Seq: ALIGNMENT

- bowtie 1 & 2
  - bowtie 1 reads length < 50pb
  - bowtie 2, most used soft in ChIP-Seq
- STAR
  - As good as bowtie 2, but more faster
  - Need > 40GB RAM for human genome
- Alignment must be unique

# ChIP-Seq: PCR DUPLICATES

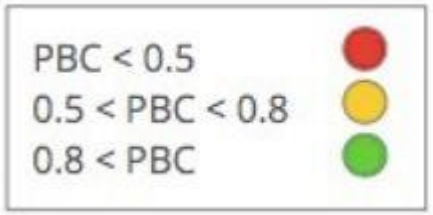
- Low complexity library
- Same set of amplified fragments
  - IP failed?

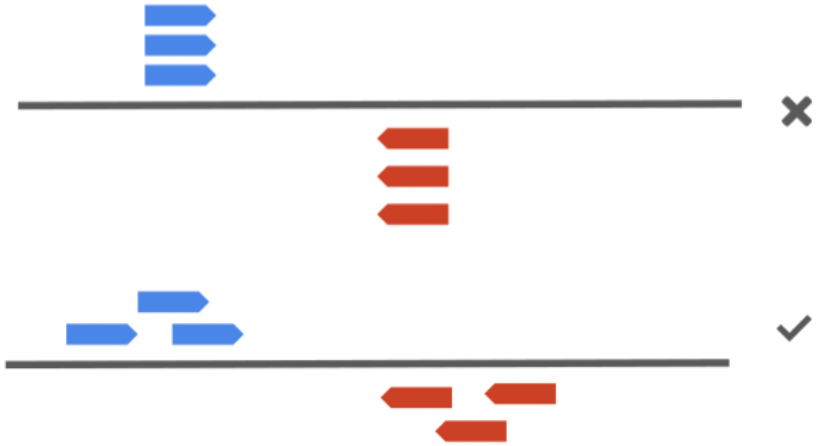


- QC Tools
  - FASTQC
  - PRINSEQ …

# ChIP-Seq: METRICS

- PBC: PCR Bottleneck Coefficient
- Approximation of library complexity
  - PBC = $N_1/N_d$
    - N1= genomic position with only 1 read aligned
    - Nd = genomic position with ≥ 1 read aligned

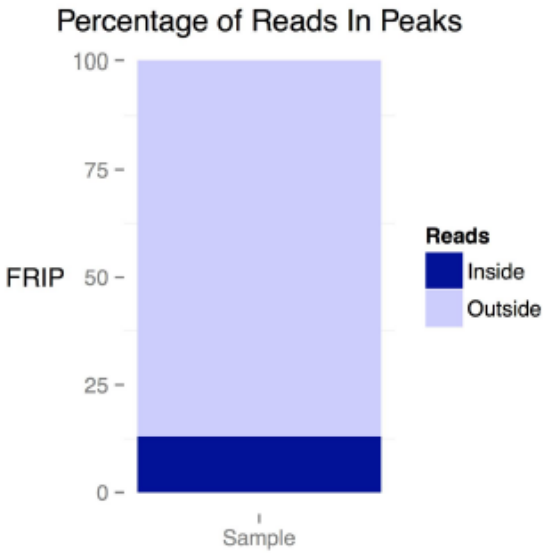


# ChIP-Seq: METRICS

- FRiP: Fraction of Reads in Peaks
- Measuring global ChIP enrichment

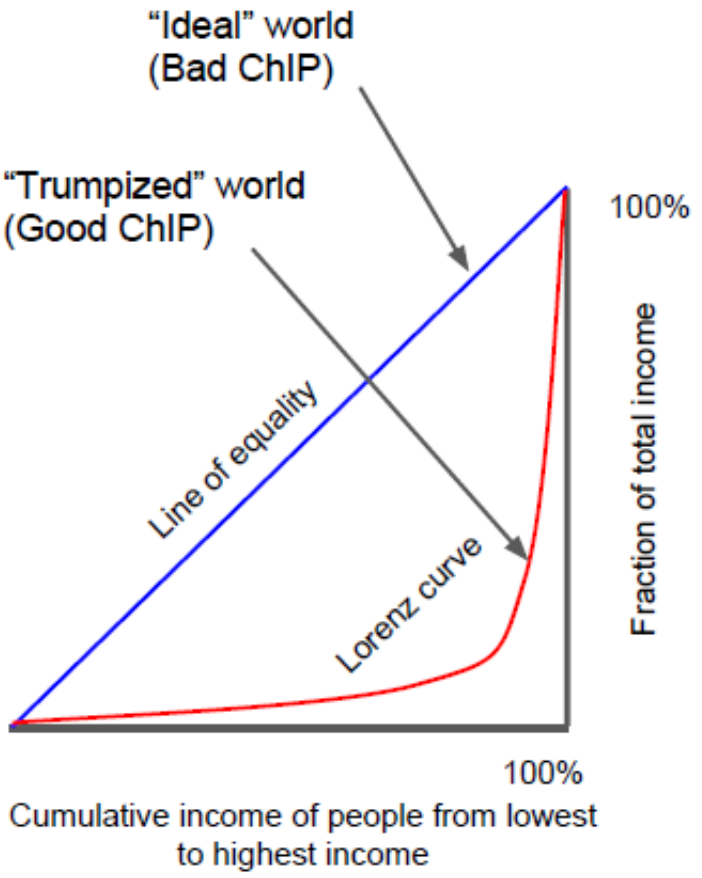
$$FRiP = \frac{reads \in peaks}{total\ reads}$$




  - Depends on ChIP experiment (TF/histone marks)

# ChIP-Seq: METRICS

- Lorenz Curve:
  - Analyze income among workers by cumulative sum
    - If income distribution is uniform: straight line
    - If they were deceived: Lorenz curve




- In ChIP context:
  - Genome == workers
  - Incomes == reads

# ChIP-Seq: Peak calling

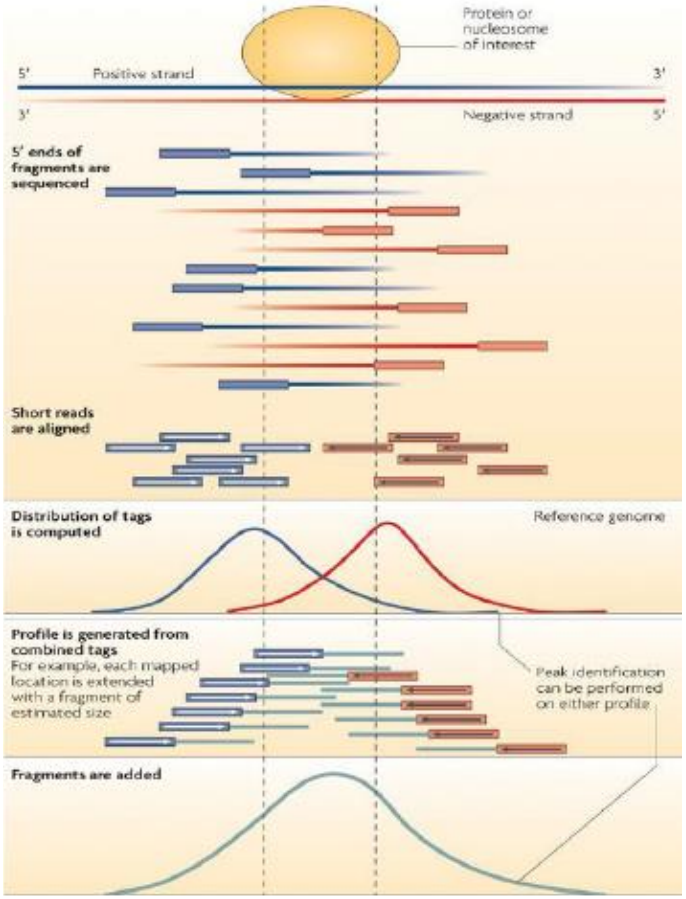
| Tools | Algorithm | Approach | Publication | Language | OS | Last release |
|---|---|---|---|---|---|---|
| BayesPeak | Bayesian | HMM | 2011 | R & C | linux/mac/win | 2018 1.29.0 |
| F-Seq | F-Seq density estimation | Kernel desity estimation | 2008 | Java | linux | 2011 1.84 |
| MACS2 | Shift & sliding window | Model-based analysis | 2008 | Python | linux | 2018 2.1.2 |
| SICER | Scoring scheme | Spactial clustering (histone marks) | 2009 | Python | linux | 2016 |

Adapted from Tran and Huang Biology Direct 2014, 9:4

# ChIP-Seq: FROM READS TO PEAKS

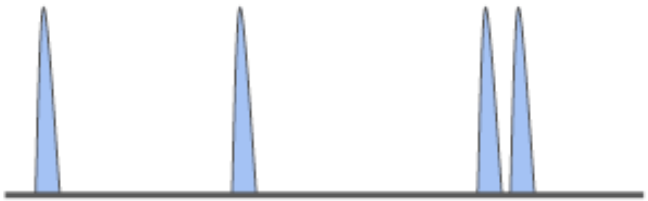
- Peaks are a mixture of two signals:
  - positive strand reads
  - negative strand reads

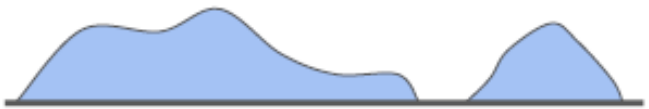
- The sequence tag density accumulates on forward and reverse strands centered around the binding site



# ChIP-Seq: PEAK CALLERS

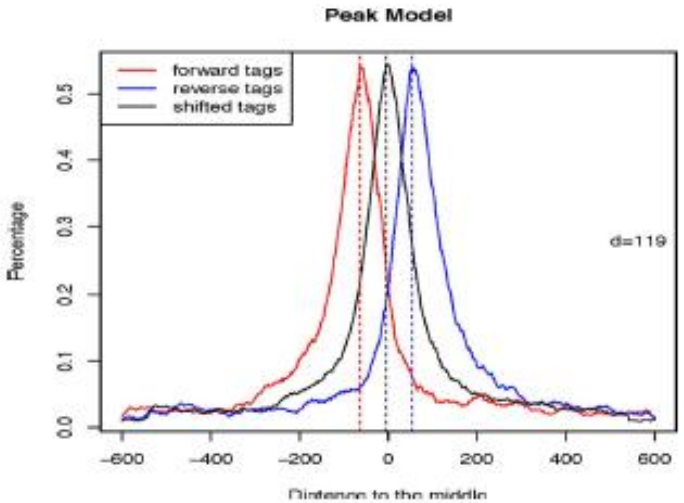
- Peak caller chosen based on:
  - Experimental design
    - SE (MACS1.4) or PE (MACS2)
  - Expected signal
    - Sharp peaks (transcription factors)
      - MACS
    - Broad peaks (epigenetic marks)
      - MACS, SICER …



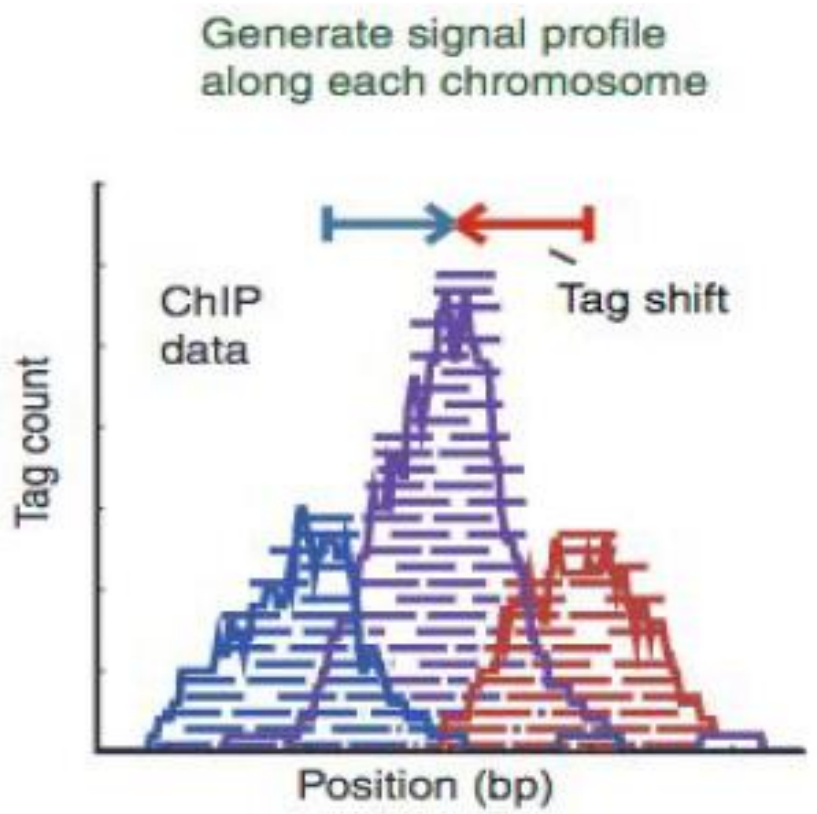
# ChIP-Seq: MACS2

- Step 1: Modeling the shift of ChIP-Seq tags
  - Sliding window 2 x ***bw*** (bandwidth) along genome to find regions with tags > ***MFOLD*** enriched relative to a random tag genome distribution
  - Randomly samples 1000  of highly enriched regions
  - Separation of positive / negative tags, and aligns them by the midpoint
  - Define ***d*** as the distance in bp between the summit of the two distribution



# ChIP-Seq: MACS2

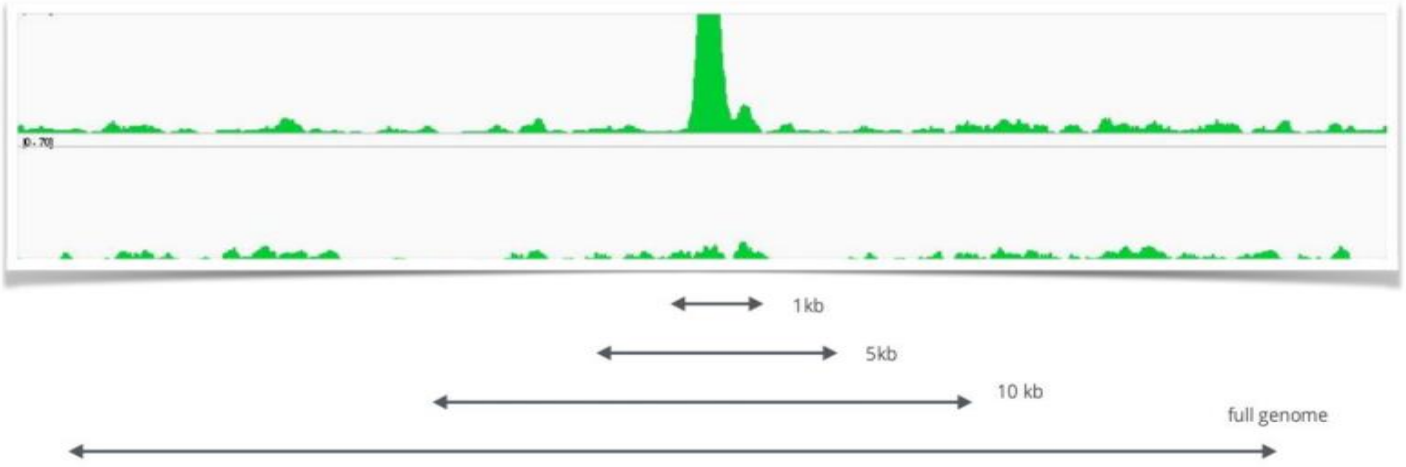
- Step 2: peak detection
  - Scaling the input reads count with ChIP count
  - Duplicate read removal
  - Tags are shifted by ***d***/2




Pepke et al. 2009

# ChIP-Seq: MACS2

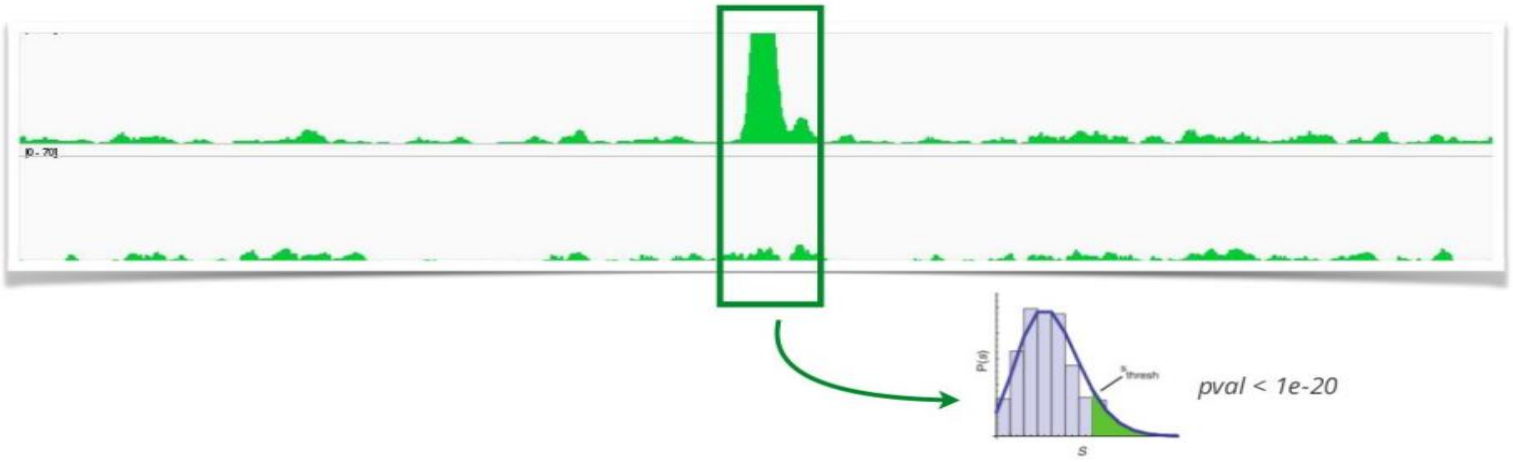
- Step 3: identify local noise
  - Sliding window **2 x *d*** between ChIP and Input
  - Estimate parameter $\boldsymbol{\lambda_{local}}$ of Poisson distribution




estimate parameter $\lambda_{local}$ over different ranges, take max.

# ChIP-Seq: MACS2

- Step 4: identify enriched region
  - Determine regions with pval < threshold
  - Determine summit location in enriched region as the maximum density
  - Compute FDR



# ChIP-Seq: MACS2 options

- Various options to indicate/control input, output, peak modelling and peak calling
- macs2 callpeak

usage: macs2 callpeak [-h] -t TFILE [TFILE ...] [-c [CFILE ]]

[-f {AUTO,BAM,SAM,BED,ELAND,ELANDMULTI,ELANDEXPORT,BOWTIE, BAMPE}] [-g GSIZE] [--keep-dup KEEPDUPLICATES] [--buffer-size BUFFER_SIZE]

[--outdir OUTDIR] [-n NAME] [-B] [--verbose VERBOSE] [--trackline] [--SPMR]

[-s TSIZE] [--bw BW] [-m MFOLD MFOLD] [--fix-bimodal] [--nomodel] [--shift SHIFT] [--extsize EXTSIZE]

[-q QVALUE] [-p PVALUE] [--to-large] [--ratio RATIO] [--down-sample] [--seed SEED] [--nolambda] [--slocal SMALLLOCAL] [--llocal LARGELOCAL] [--broad] [--broad-cutoff BROADCUTOFF] [--call-summits]

# ChIP-Seq: MACS2 output

MACS peaks in bed format

```
chr1 3001827   3002328   MACS_peak_1   55.28
chr1 3067471   3067948   MACS_peak_2   50.67
chr1 3660316   3662844   MACS_peak_3   352.43
chr1 3842462   3842994   MACS_peak_4   59.21
chr1 3877254   3877710   MACS_peak_5   52.72
chr1 3939314   3939679   MACS_peak_6   82.99
```

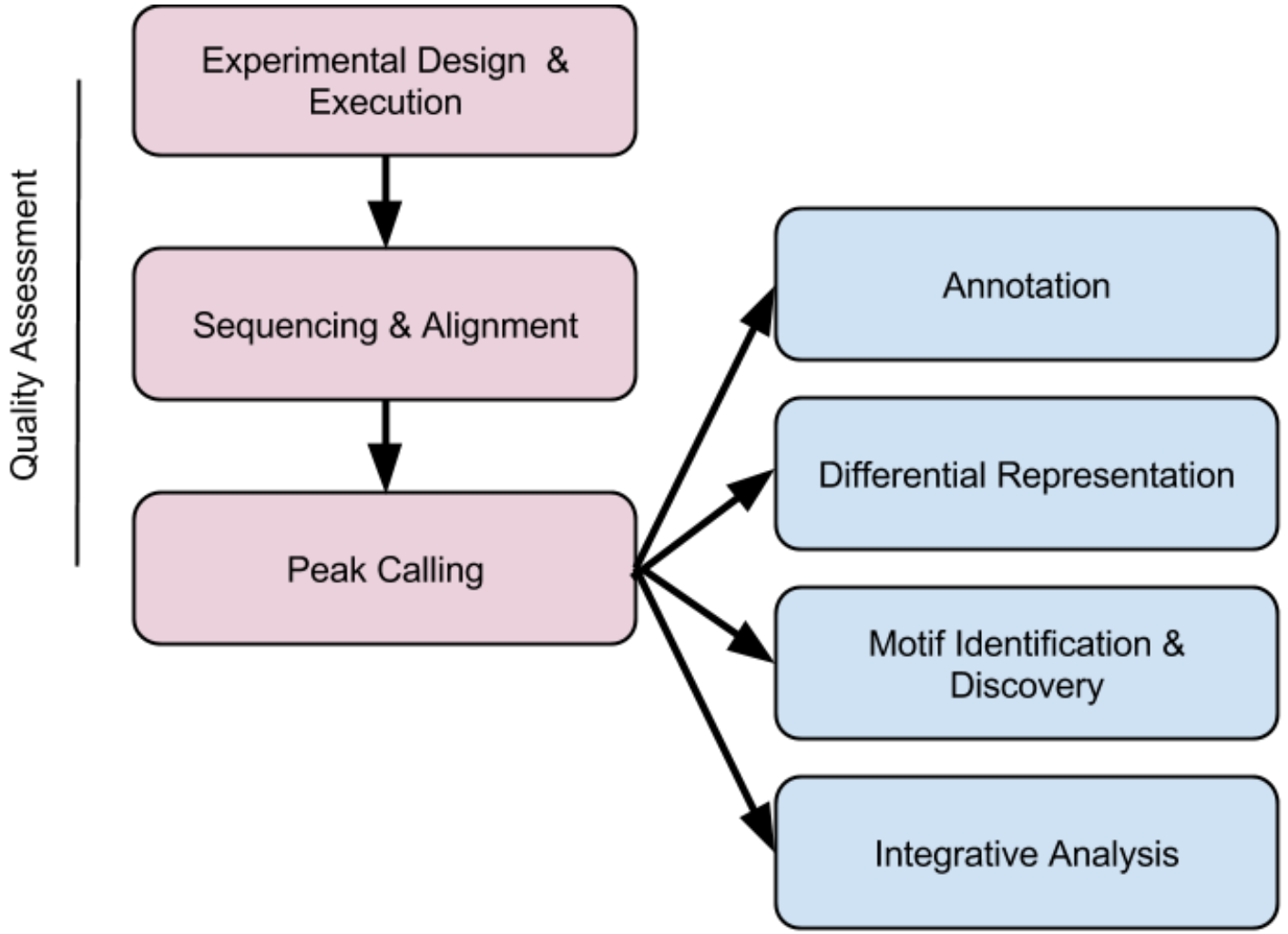
Statistical significance
-10 log(P-value)

MACS peaks extented format

| Chr | Start | End | W | Summit | Tags | Sig | Fold | FDR |
|---|---|---|---|---|---|---|---|---|
| chr16 | 35981451 | 35981951 | 321 | 35981701 | 24 | 1107.07 | 30.55 | 0.0 |
| chr18 | 30784846 | 30785346 | 628 | 30785096 | 40 | 964.91 | 43.62 | 0.0 |
| chr14 | 79381873 | 79382373 | 441 | 79382123 | 29 | 939.17 | 37.2 | 0.0 |
| chr12 | 34467249 | 34467749 | 1160 | 34467499 | 53 | 928.38 | 19.93 | 0.0 |
| chr8 | 90304944 | 90305444 | 1804 | 90305194 | 80 | 883.76 | 10.21 | 0.0 |
| chr15 | 65294343 | 65294843 | 992 | 65294593 | 62 | 824.32 | 13.4 | 0.0 |
| chr17 | 48499365 | 48499865 | 370 | 48499615 | 24 | 798.58 | 20.62 | 0.0 |
| chr18 | 72429446 | 72429946 | 531 | 72429696 | 31 | 790.48 | 39.77 | 10.0 |
| chr15 | 54579253 | 54579753 | 487 | 54579503 | 29 | 781.63 | 32.15 | 9.09 |
| chr13 | 56988583 | 56989083 | 916 | 56988833 | 60 | 777.7 | 9.44 | 8.33 |

https://genome.ucsc.edu/FAQ/FAQformat.html

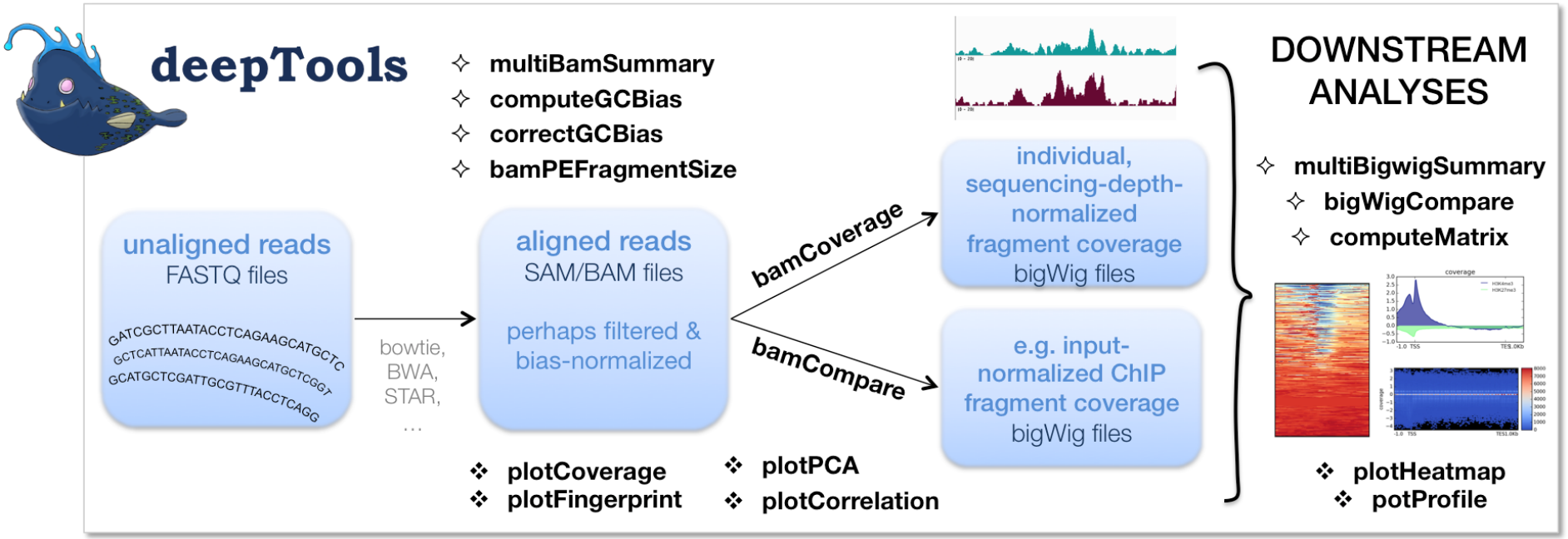
# ChIP-Seq: DOWNSTREAM ANALYSIS



# ChIP-Seq: DOWNSTREAM ANALYSIS

- Normalization and Visualisation
- Annotation of genomic features to peaks
- Feature distribution of binding sites
- Feature overlap analysis
- Functional enrichment analysis
- Motif identification or motif enrichment analysis
- Differential analysis

# ChIP-Seq: NORMALIZE & VISUALIZE




https://deeptools.readthedocs.io

# ChIP-Seq: ANNOTATION

R packages: Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

- ChIPpeakAnno or ChIPseeker
  - Peaks mapping to nearest feature
    - (TSS / gene / exon / custom feature …)
  - Extract peak sequence
  - Find peaks with bidirectional promoters
  - Enriched gene ontology …
- biomaRt pck to get annotation from Ensembl
- IRanges, GenomicFeatures, Go.db, BSgenome etc…

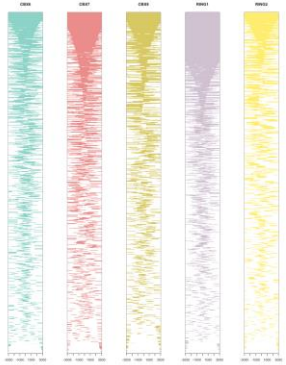
- Also possible to do it in unix with bedtool and gff / gtf
- Or HOMER suite (annotatePeaks.pl)

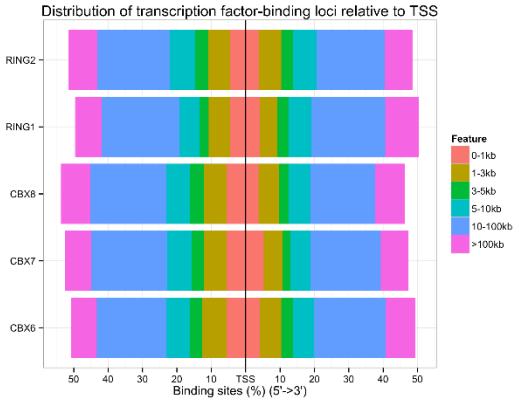

# ChIP-Seq: PEAKS FEATURES SUMMARY

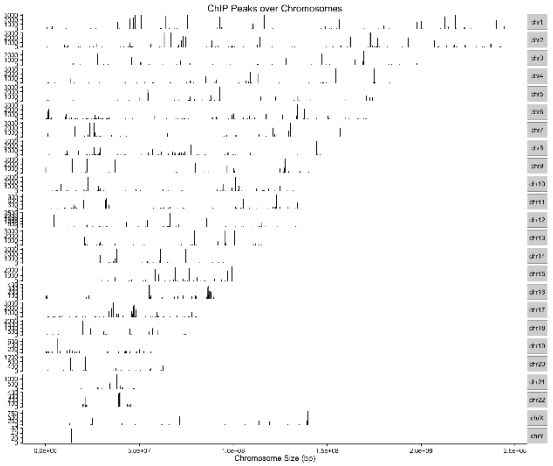
**ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization** FREE

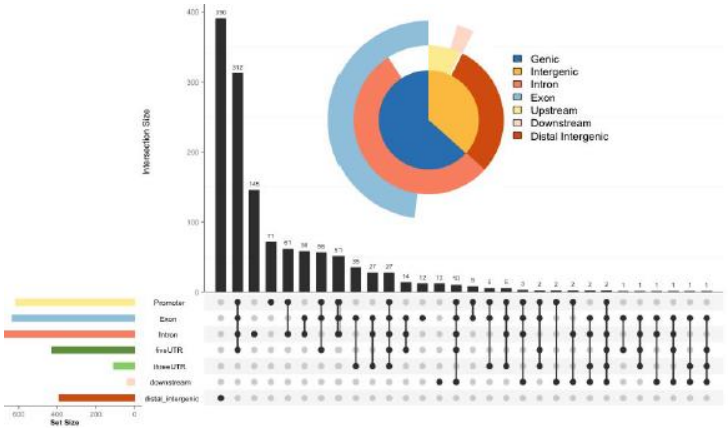
Guangchuang Yu ✉, Li-Gen Wang, Qing-Yu He ✉ Author Notes

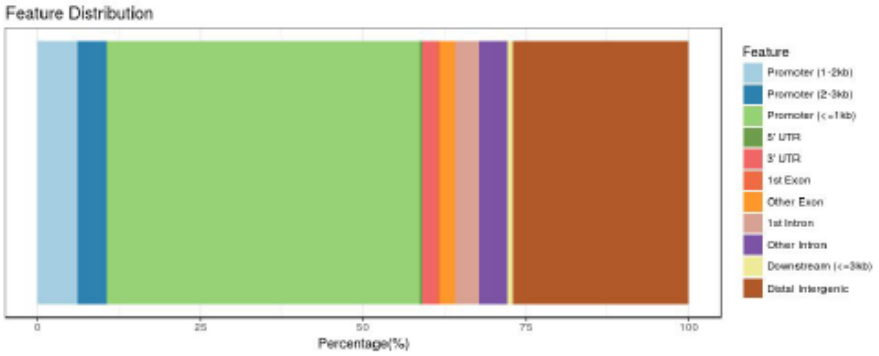
*Bioinformatics*, Volume 31, Issue 14, 15 July 2015, Pages 2382–2383, https://doi.org/10.1093/bioinformatics/btv145



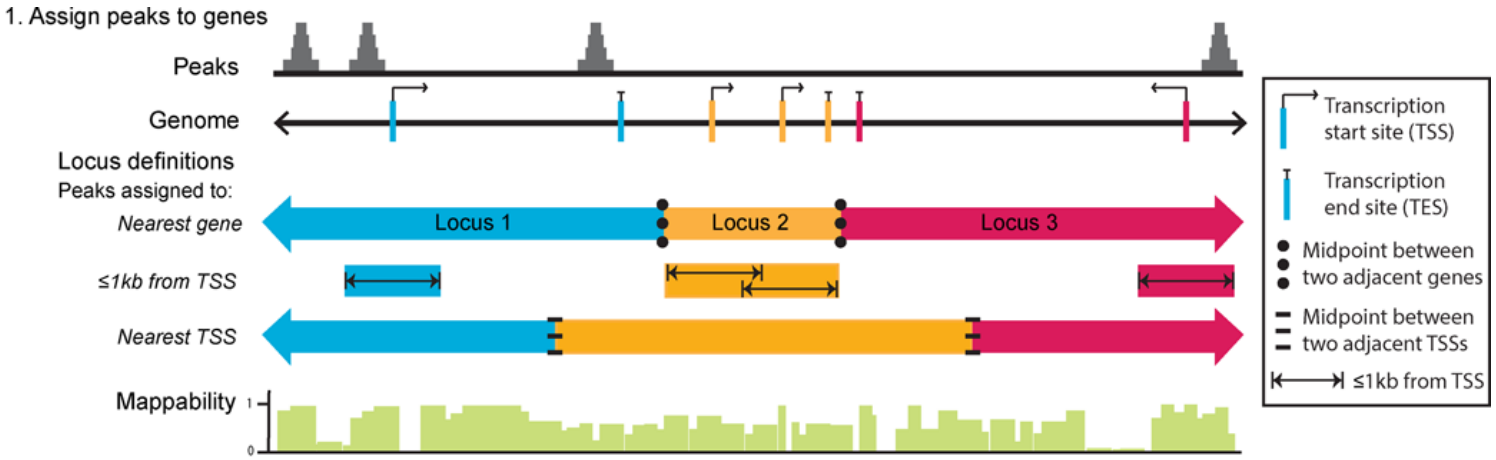


# ChIP-Seq: FUNCTIONAL ENRICHEMENT ANALYSIS




2. Determine presence of peaks in genes

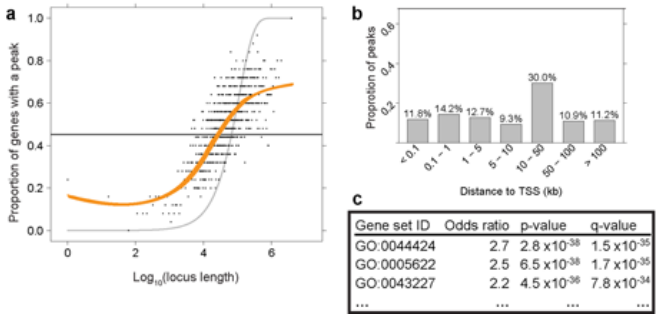
| Gene | Locus length | Presence of peak |
|---|---|---|
| ACP1 | 11,541 | 0 |
| CHL1 | 447,985 | 1 |
| HES4 | 23,485 | 0 |
| ITPR1 | 24,602 | 1 |
| MYT1L | 500,221 | 1 |
| SAMD11 | 266,255 | 0 |
| … | … | … |

3. Test for gene set enrichment

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 g + f\left(log_{10}(mL+1)\right)$$

Logistic regression model
- Adjust for (mappable, $m$) locus length ($L$)
- Estimate gene set ($g$) effect size ($\beta_1$)

4. Summarize data and enrichment results




c

| Gene set ID | Odds ratio | p-value | q-value |
|---|---|---|---|
| GO:0044424 | 2.7 | $2.8 \times 10^{-38}$ | $1.5 \times 10^{-35}$ |
| GO:0005622 | 2.5 | $6.5 \times 10^{-38}$ | $1.7 \times 10^{-35}$ |
| GO:0043227 | 2.2 | $4.5 \times 10^{-36}$ | $7.8 \times 10^{-34}$ |
| … | … | … | … |

Ryan P. Welch et *al*, *Nucleic Acids Research*, 2014

ChIP enrich

# ChIP-Seq: MOTIF SEARCH

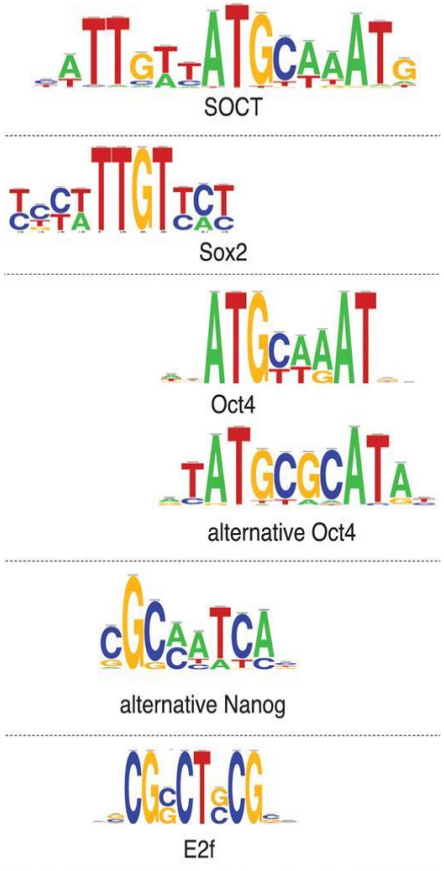
## A lot of tools!! Non exhaustive list:

**Motif Finding Tools**

| Tool | Title of Reference |
|---|---|
| CONSENSUS | Identification of consensus patterns in unaligned DNA sequences known to be functionally related |
| MotifSearch | qPMS7: A Fast Algorithm for Finding (ℓ, d)-Motifs in DNA and Protein Sequences |
| MEME | MEME SUITE: tools for motif discovery and searching |
| PSCAN | Pscan: Finding Over-represented Transcription Factor Binding Site Motifs in Sequences from Co-Regulated or Co-Expressed Genes |
| ALLEGRO | Allegro: Analyzing expression and sequence in concert to discover regulatory programs |
| DIRE | DiRE: identifying distant regulatory elements of co-expressed genes |
| PASTAA | Predicting transcription factor affinities to DNA from a biophysical model |
| oPOSSUM | oPOSSUM-3: advanced analysis of regulatory motif over-representation across genes or ChIP-Seq datasets |
| Cscan | Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets |
| ENCODE ChIP-seq | Relating Genes to Function: Identifying Enriched Transcription Factors using the ENCODE ChIP-Seq Significance Tool |

**Regulatory Sequence Analysis Tools**

| Tool | Title of Reference |
|---|---|
| i-cisTarget | i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules |
| RSAT | RSAT 2011: regulatory sequence analysis tools |



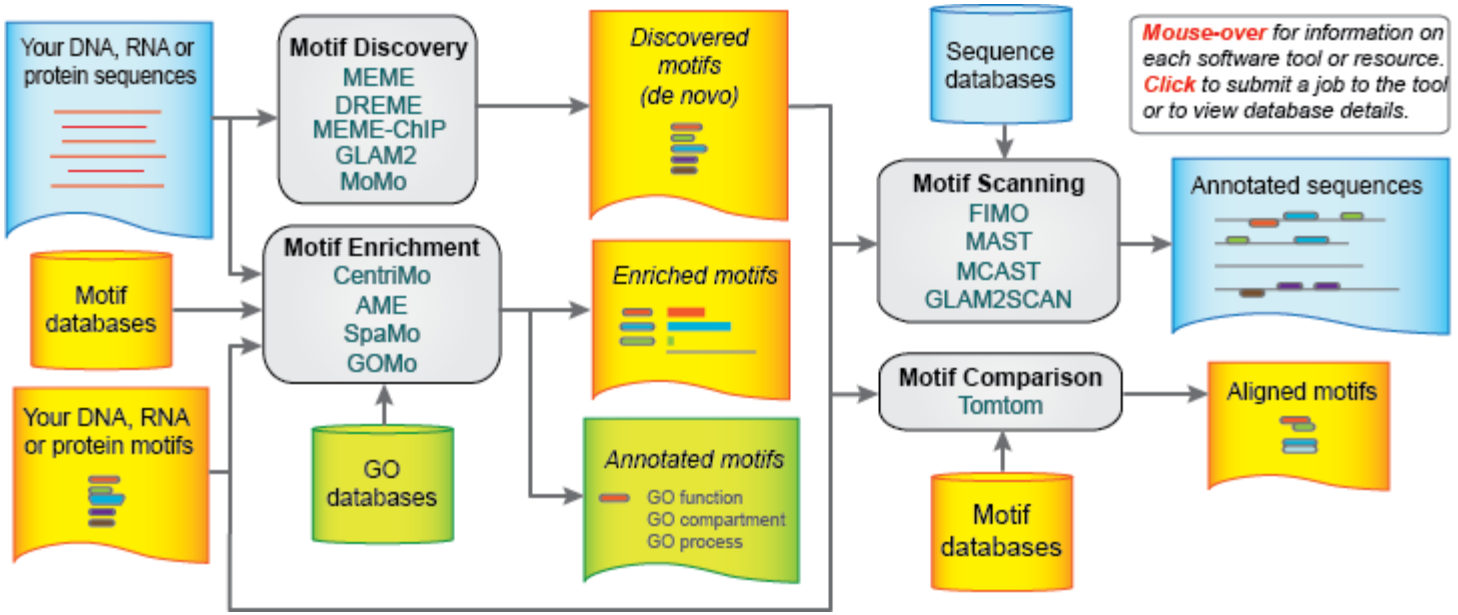
# ChIP-Seq: MOTIF DBs

| Name | Description |
|---|---|
| ChIPBase | ChIPBase a database for Transcription factor-binding sites, motifs (~1290 transcription factors) and decoding the transcriptional regulation of LncRNAs, miRNAs and protein-coding genes from ~10,200 curated peak datasets derived from ChIP-seq methods in 10 species |
| ChEA | transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. |
| CistromeMap | a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. |
| CTCFBSDB | a database for CTCF binding sites and genome organization |
| Factorbook | a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. |
| hmChIP | a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. |
| HOCOMOCO | a comprehensive collection of human transcription factor binding sites models. |
| JASPAR | The JASPAR CORE database contains a curated, non-redundant set of profiles, derived from published collections of experimentally defined transcription factor binding sites for eukaryotes. |

https://en.wikipedia.org/wiki/Transcription_factor_binding_site_databases

# ChIP-Seq: MEME Suite

- Detection and enrichment analysis of motifs




http://meme-suite.org/

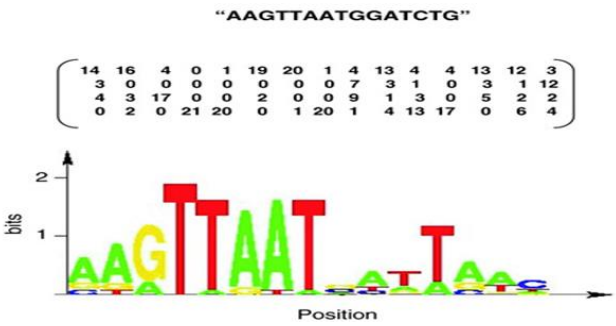
# ChIP-Seq: HOMER

- HOMER v4.10 http://homer.ucsd.edu/homer/
  - Large number of tools (Perl & C++) for ChIP analysis
  - Peak annotation / GO enrichment / extract sequences / Visualization / Motif search etc…
  - Uses a collection of ChIP-seq derived Position Weight Matrix or user can specify PWM

**Homer Known Motif Enrichment Results**

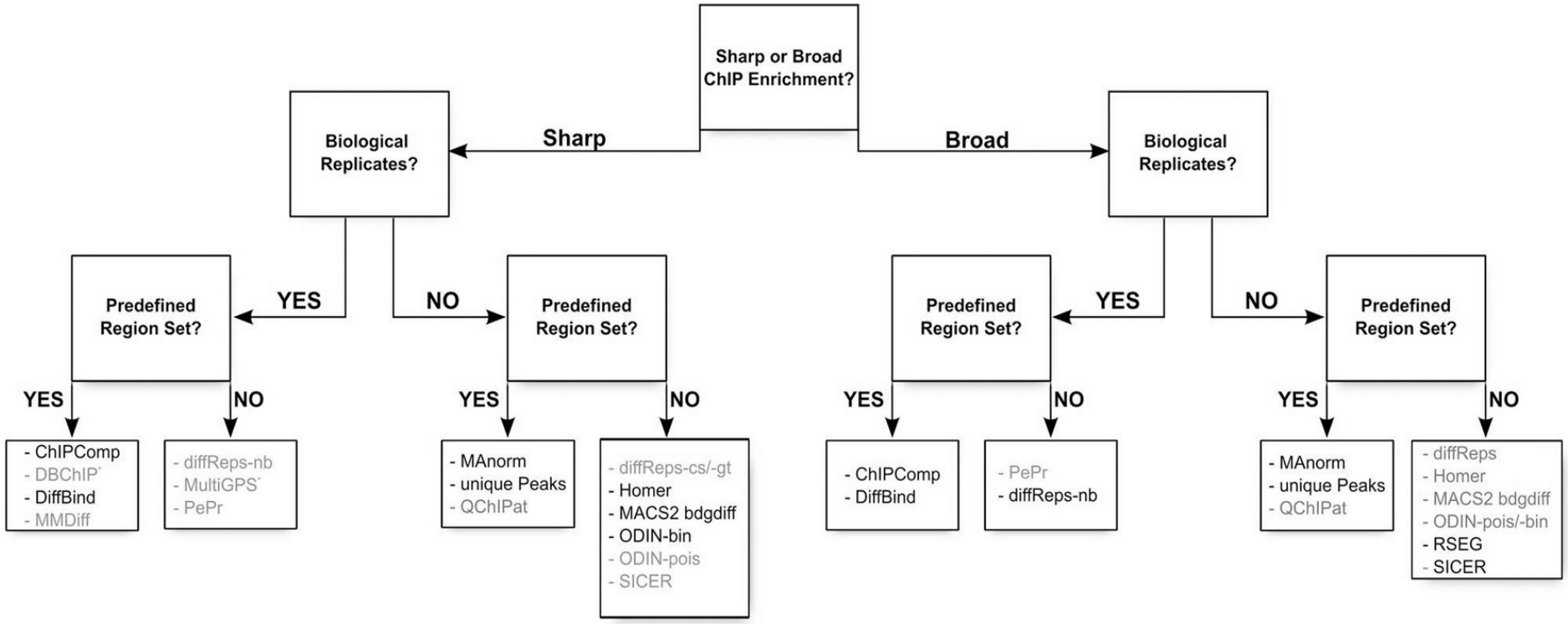
Homer *de novo* Motif Results
Gene Ontology Enrichment Results
Known Motif Enrichment Results (txt file)
Total Target Sequences = 15213, Total Background Sequences = 34081

| Rank | Motif | Name | P-value | log P-pvalue | # Target Sequences with Motif | % of Targets Sequences with Motif | # Background Sequences with Motif | % of Background Sequences with Motif | Motif File | PDF |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 |  | NFkB-p65(RHD)/GM12787-p65-ChIP-Seq/Homer | 1e-1707 | -3.931e+03 | 3855.0 | 25.34% | 1506.4 | 4.42% | motif file (matrix) | pdf |
| 2 |  | CEBP(bZIP)/CEBPb-ChIP-Seq/Homer | 1e-1310 | -3.018e+03 | 3423.0 | 22.50% | 1551.0 | 4.55% | motif file (matrix) | pdf |
| 3 |  | PU.1(ETS)/ThioMac-PU.1-ChIP-Seq/Homer | 1e-1288 | -2.967e+03 | 3413.0 | 22.44% | 1569.7 | 4.61% | motif file (matrix) | pdf |
| 4 |  | AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq/Homer | 1e-947 | -2.183e+03 | 3251.0 | 21.37% | 1900.8 | 5.58% | motif file (matrix) | pdf |
| 5 |  | NFkB-p65-Rel(RHD)/LPS-exp/Homer | 1e-936 | -2.157e+03 | 1163.0 | 7.65% | 166.2 | 0.49% | motif file (matrix) | pdf |
| 6 |  | ETS1(ETS)/Jurkat-ETS1-ChIP-Seq/Homer | 1e-885 | -2.039e+03 | 4447.0 | 29.23% | 3568.0 | 10.47% | motif file (matrix) | pdf |



# ChIP-Seq: DIFFERENTIAL ANALYSIS

- Comparing samples across conditions or time series?
- Overlapping peaks too simple

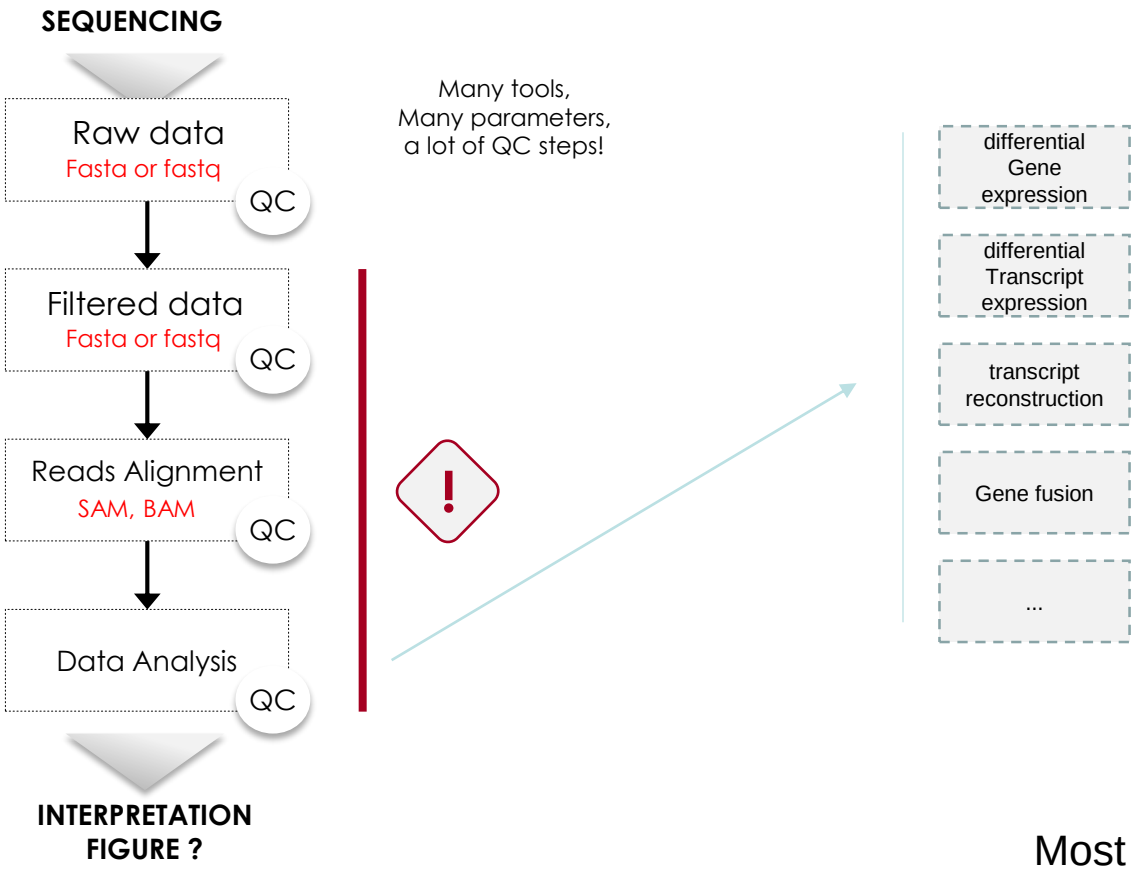



Steinhauser S *et al*, *Brief Bioinform*. 2016;17(6):953-966.

# TP

# RNA-seq

# NGS : general data analysis workflow




**NGS** = **N**ext **G**eneration **S**equencing

**QC** = **Q**uality **C**heck

Most steps are common to most analysis workflow, and most tools/software are common. **Yet** some analyses require specific care in experimental design, tool choice and parameter settings.

# Questions to ask yourself before sequencing?

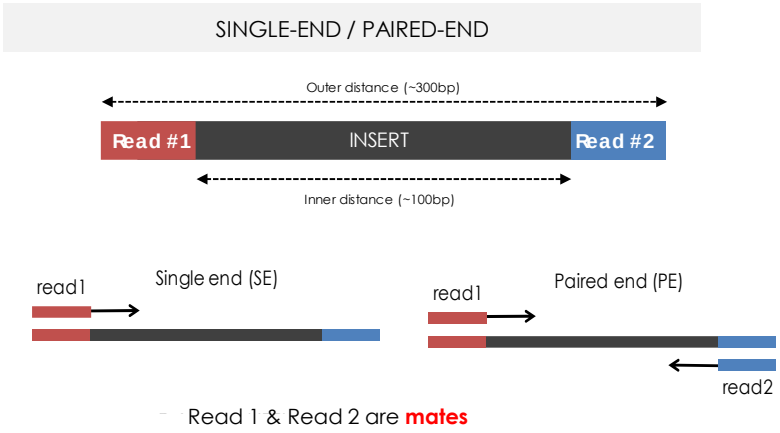
# Questions to ask yourself before sequencing?

- application
- depth
- cost
- library prep
- mate-pair/paired-end
- read length
- How many replicates?

At least 3!

- What is the biological questions I want to address?
- Is the selected approach appropriate to address the question?
  - seek advice and discuss with other biologists/bioinformaticians/sequencing platforms.

# Sequençing considerations




| | rna-seq | depth | ref |
|---|---|---|---|
| Transcriptome Sequencing | Differential expression profiling | 10-25M | Liu Y. et al., 2014; ENCODE 2011 RNA-Seq |
| | Alternative splicing | 50-100M | Liu Y. et al., 2013; ENCODE 2011 RNA-Seq |
| | Allele specific expression | 50-100M | Liu Y. et al., 2013; ENCODE 2011 RNA-Seq |
| | De novo assembly | >100M | Liu Y. et al., 2013; ENCODE 2011 RNA-Seq |
| Small RNA Sequencing | Differential Expression | ~1-2M | Metpally RPR et al., 2013; Campbell et al., 2015 |
| | Discovery | ~5-8M | Metpally RPR et al., 2013; Campbell et al., 2015 |

https://genohub.com/recommended-sequencing-coverage-by-application/

**Key considerations (Sims et al., 2014, Nature reviews)**

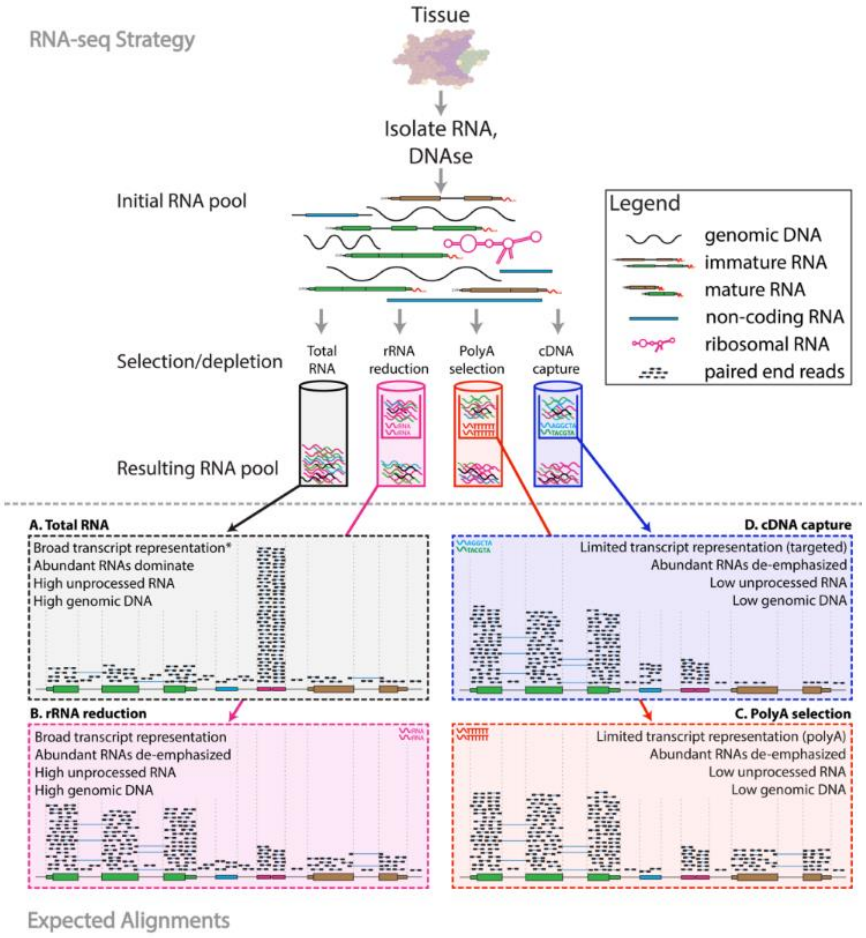
- Biological question : qualitative or quantitative informations ?
- Data quality : *PCR duplicates, reads length, paired/signle-end reads…*
- Model organism genome complexity
- …

# Interesting resources and protocols :

- Ignacio Gonzalez's guide : http://www.nathalievialaneix.eu/doc/pdf/tutorial-rnaseq.pdf
- ucsc (formats), broad, ensembl, ncbi, Sanger Institute
- https://f1000research.com/gateways/bioconductor ==> Three differential expression protocols, mostly done using R. Many other interesting protocols.
- Pertea & al, Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. nature protocols, 2016.
- Bioconductor vignettes & workflow .
- Griffith lab tutorials and protocols (excellent all track resource). https://github.com/griffithlab/rnaseq_tutorial/wiki
- https://bioinformatics.ca/workshops/

# RNA-seq library preparations

looking for similarity and differences across samples



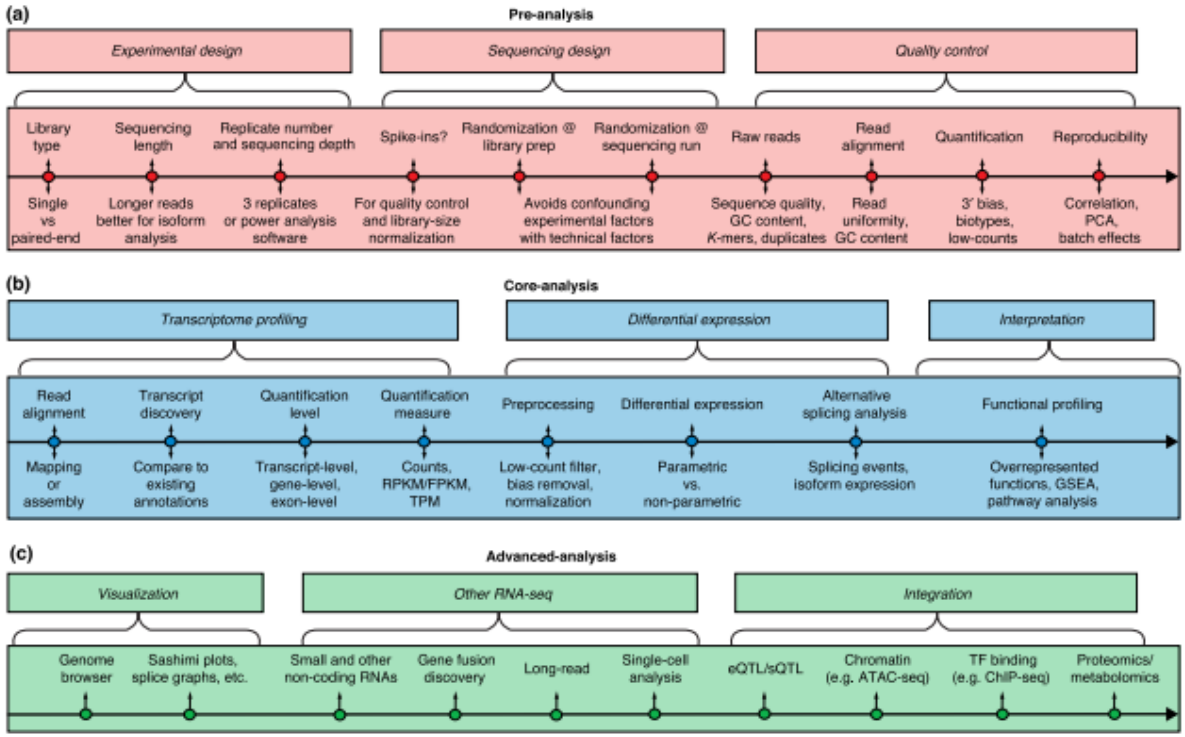

- which RNA population isolated?
- Ribo depletion?
- capture/enrichment?
- size selection? (miARN)
- Reference available?
- old datasets

experimental set-up is critical and will guide the bioinformatics

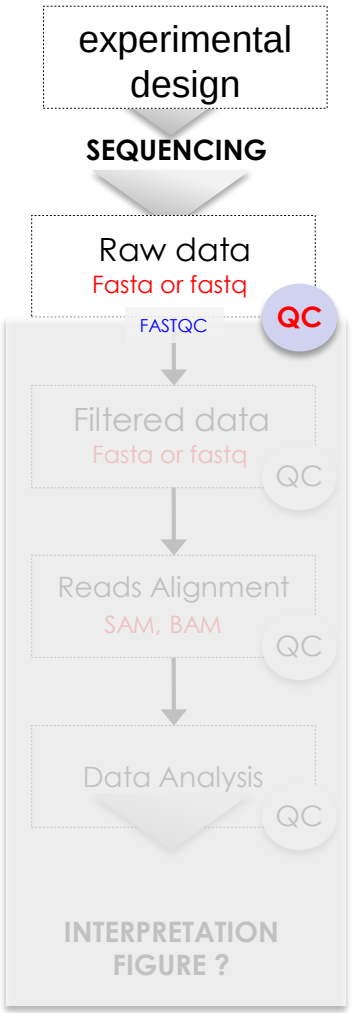
REVIEW Open Access

# A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15] and Ali Mortazavi[16,17*]



# FASTQC report




FastQC Report

**Summary**

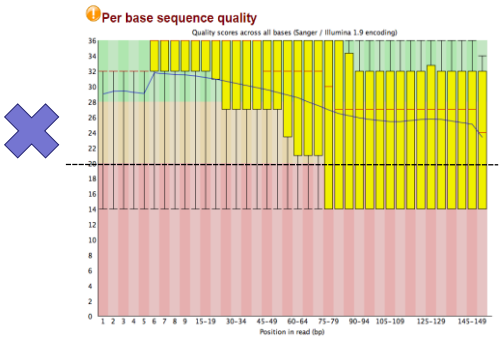
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

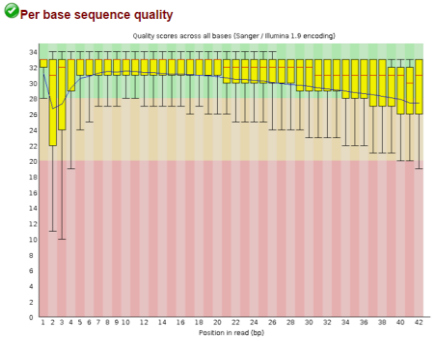
**Basic Statistics**

| Measure | Value |
|---|---|
| Filename | fus_3M_R2.fastq.gz |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 66654738 |
| Sequences flagged as poor quality | 0 |
| Sequence length | 35-151 |
| %GC | 24 |

$$Q = -10 \log\_10\ (p)$$

Where $Q$ is the quality and $p$ is the probability of the base being incorrect.



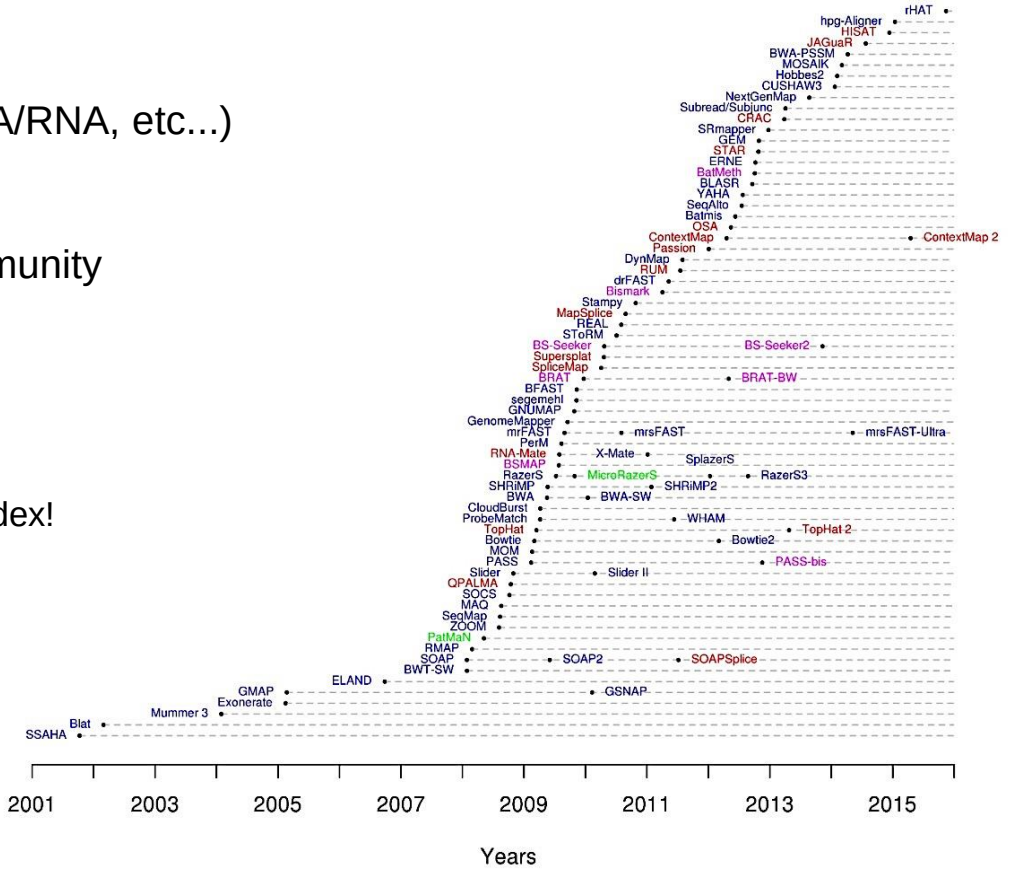




→ read filtering & trimming

# Mapping :

- mapper selection?
- experiment dependant (splicing aware, DNA/RNA, etc...)
- Most recent, fast and accepted by the community probably are (for rna-seq)
  - Star
  - Hisat-2
  - Rsubread

! beware of RAM usage if you need to build the index!

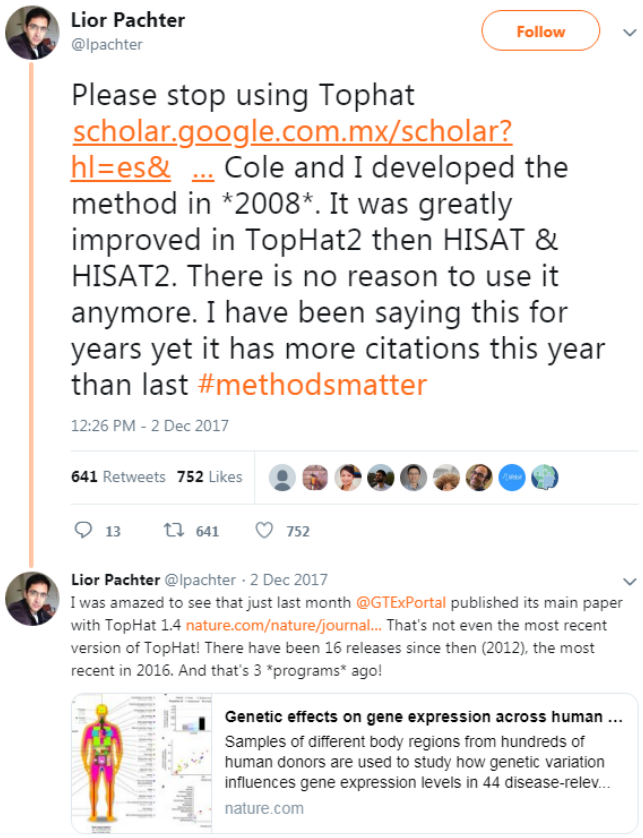



*Fonseca N A et al. Bioinformatics 2012;28:3169-3177*

# A word of caution about standard pipelines...

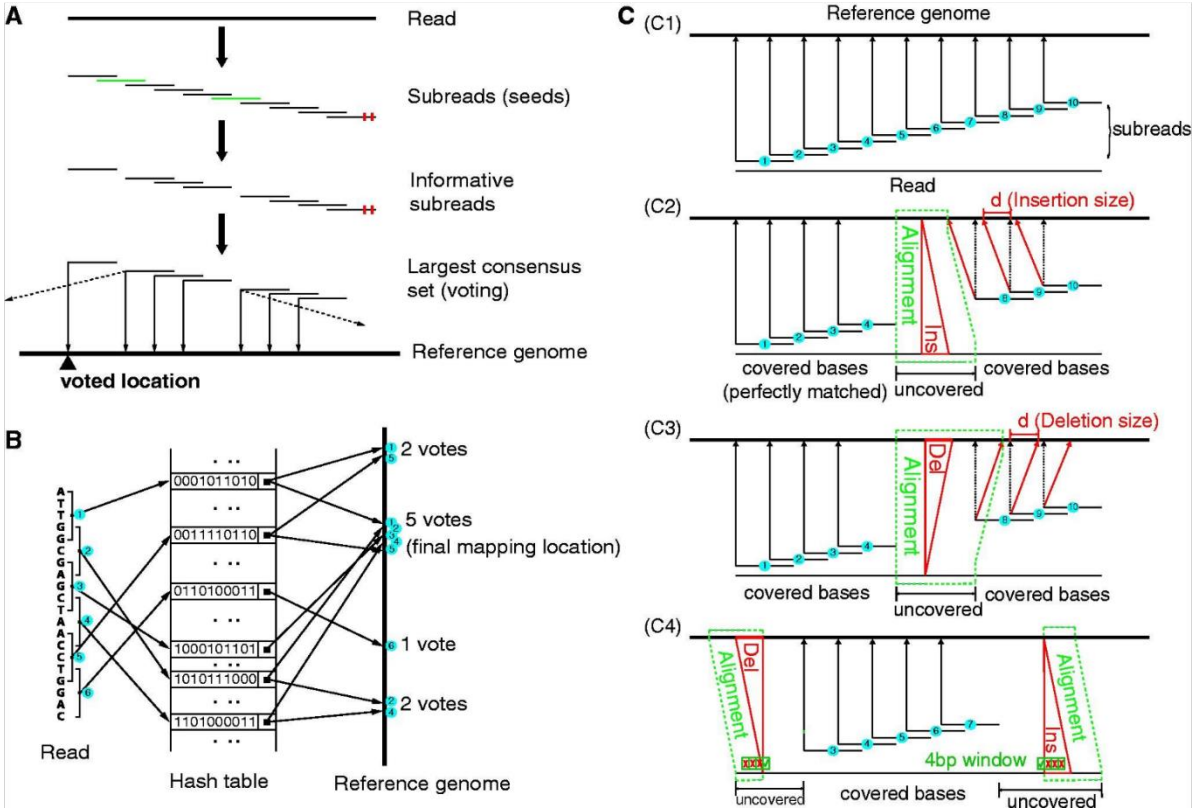
| | map | count | normalise | expression |
|---|---|---|---|---|
| DEG | bowtie(2)<br>Subread<br>STAR | FeatureCounts<br>HTseq-count<br>Subread<br>RSEM | Edger<br>DESeQ(2) | |
| Alternative splicing | HISAT2 | StringTie | ballgown | |
| pseudo alignments | Kallisto | Sleuth | | |

other sequencing technologies will (can) change the way the analysis is performed (PacBio, Oxford Nanopore)




Be wise in your selection of tools and softwares:
a balance between proven « good old » methods, widely used protocols and more efficient & recent methods.

# mappers : how do they work?




- alignment rarely optimal? SNP/mismatches/repeats/reference is another strain or species/etc...

Subread uses the “seed and vote” principle, other mappers rely on similar approaches (“seed & extend”).

genome shred into a binary dictionary (speed up the alignment search).

# STAR

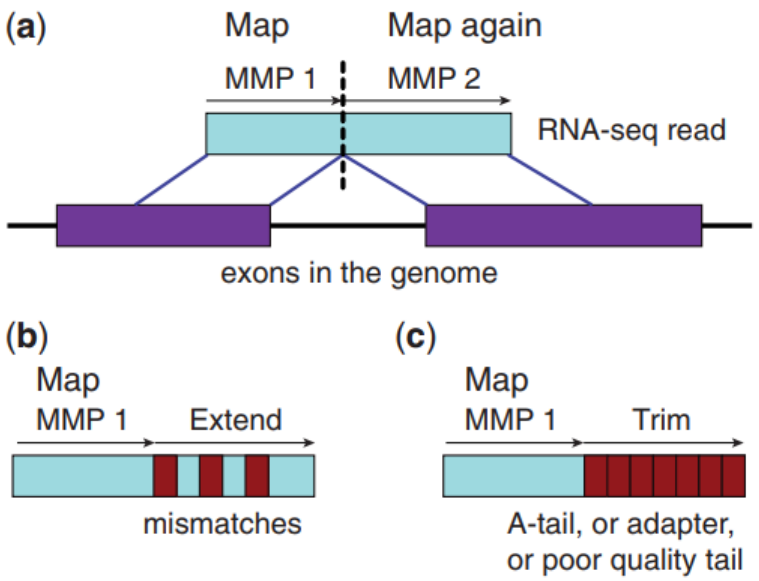



**Fig. 1.** Schematic representation of the Maximum Mappable Prefix search in the STAR algorithm for detecting (**a**) splice junctions, (**b**) mismatches and (**c**) tails

reads are split into Maximum Mapping Units and extended or trimmed appropriately.
➔ easy exon junction mapping.

# A good reference genome ?

https://genome.ucsc.edu/
http://www.ensembl.org/index.html
http://www.biomart.org/

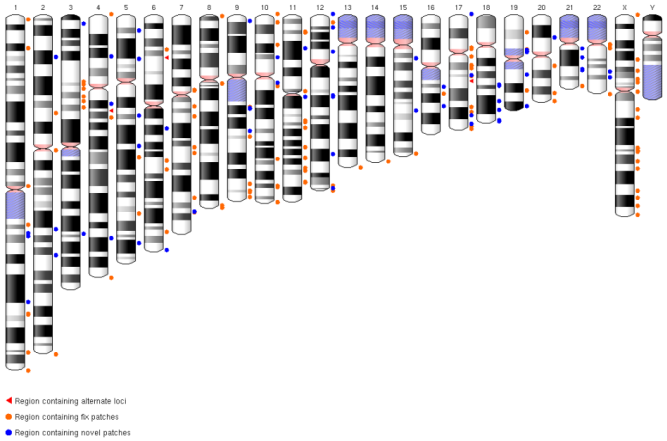
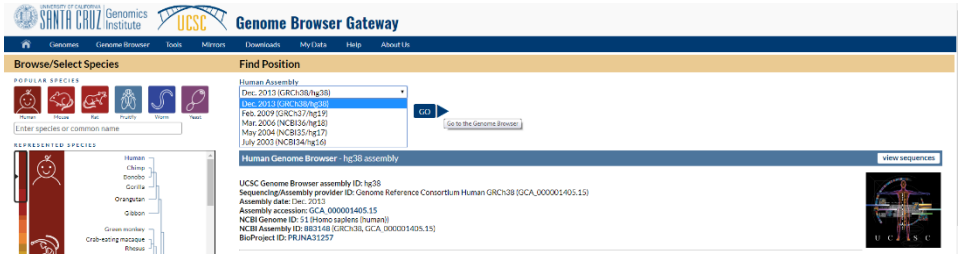



**Figure 8: State of the human Genome Reference sequence as of 2013.** Large regions of the genome are still not part of the assembly (blue and pink coloured regions) including the short arms of accrocentric chromosomes 13, 14, 15, 21 & 22. Centromeres are also not part of the assembly. Additionally a number of gaps, misassemblies and alternate loci are still present. From www.ncbi.nlm.nih.gov/projects/genome/.../human.

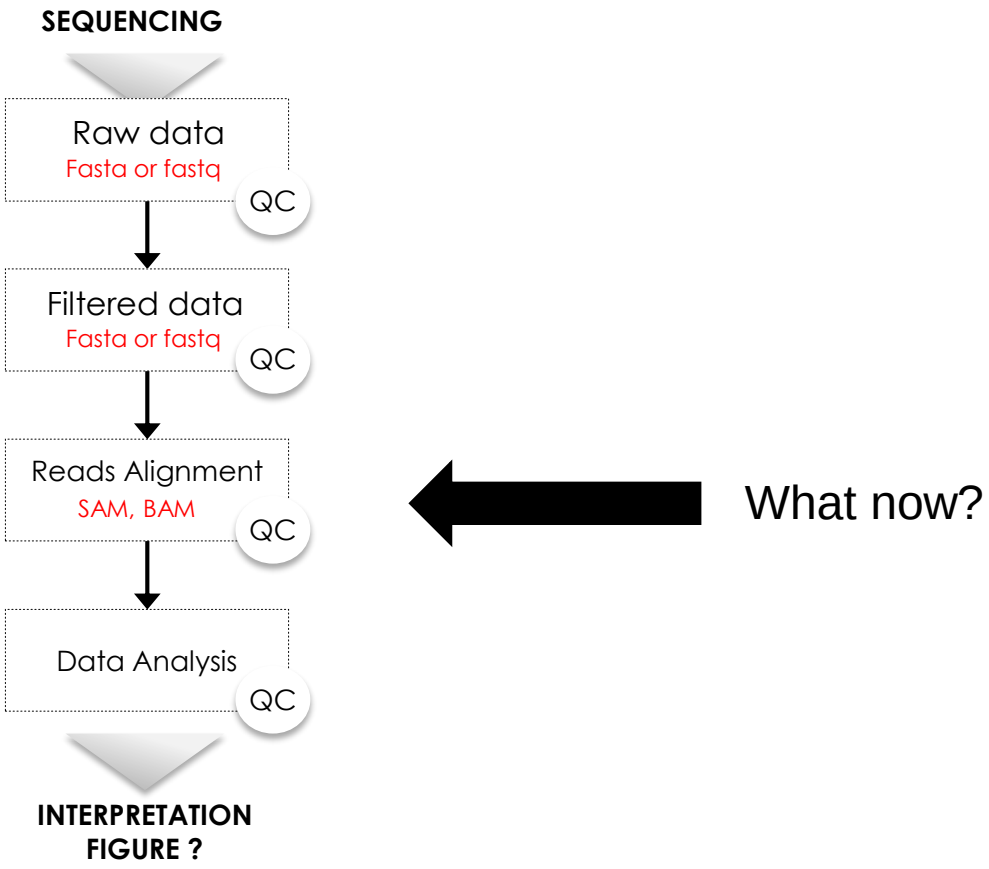



- take into account the complexity of the genome(repeats, polyploidy, etc) and the quality of the annotation.
- Depending on the databases, assembly quality and number of annotated genes can dramatically affect the end results.
- coordinates vary from one version to the other.

biomart, annotationDBi, gtf/GFF from databases, custom gtf...

If no reference available?
=>
genome sequencing, transcript assembly and annotation.

# conventional NGS workflow:




What now?

**NGS** = **N**ext **G**eneration **S**equencing

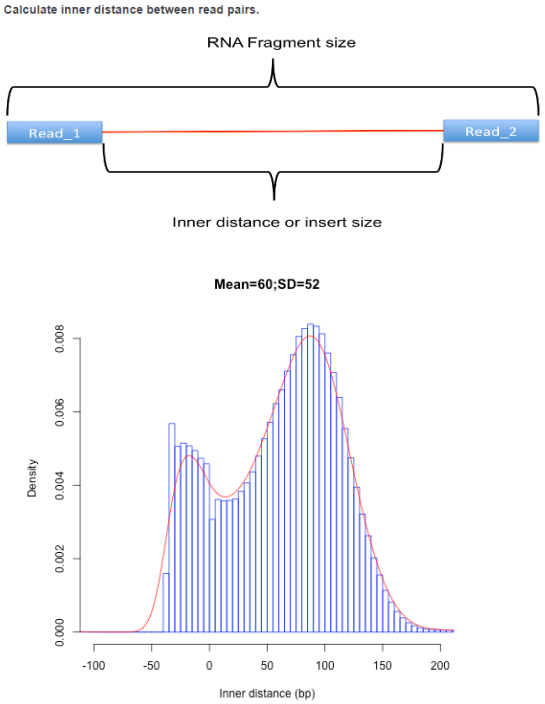
**QC** = **Q**uality **C**heck

# What shall we do after generating a bam file?

& Picard tools

## 1/ mapping QC :

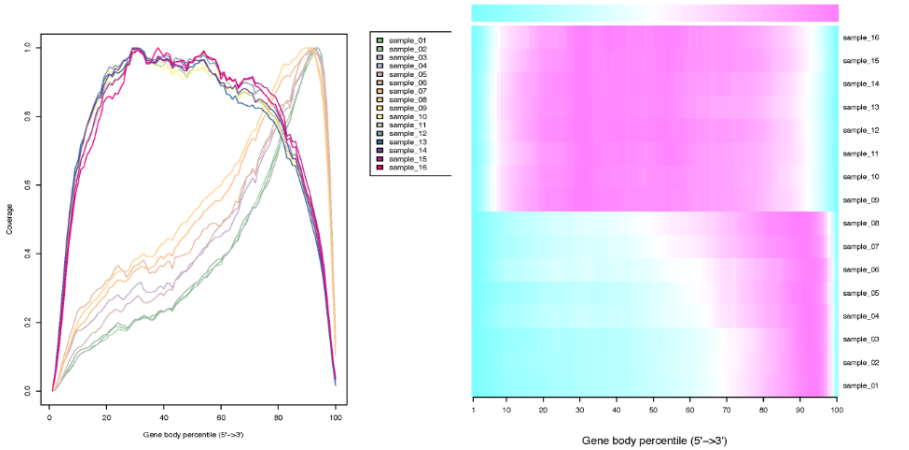
RseQC – 1
a suite of tools to easily estimate the quality of your mapping




polyA/3', library type...

```
$ geneBody_coverage.py -r hg19.housekeeping.bed -i /data/alignment/  -o output
```

Output:

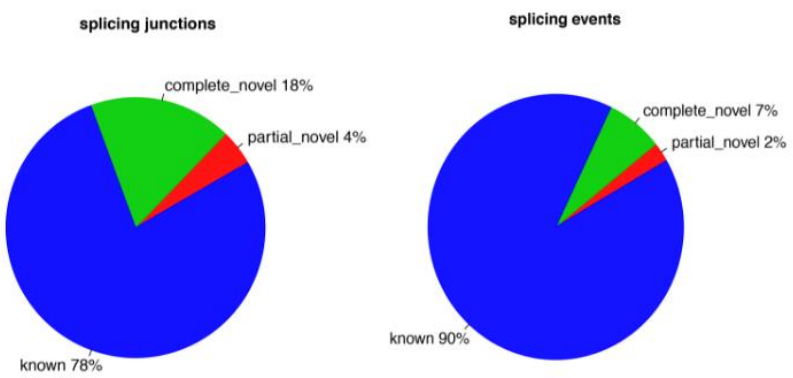



http://rseqc.sourceforge.net/

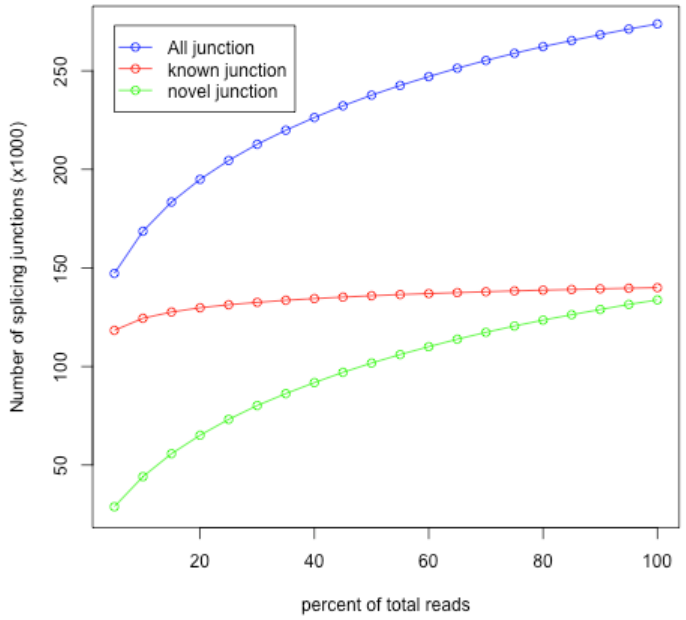
# Mapping QC:

## RseQC - 2

Output:

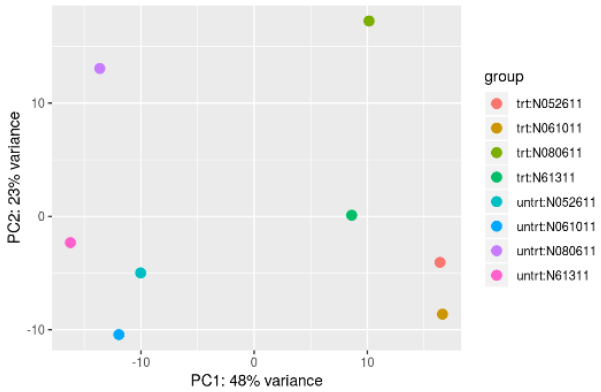



Output:

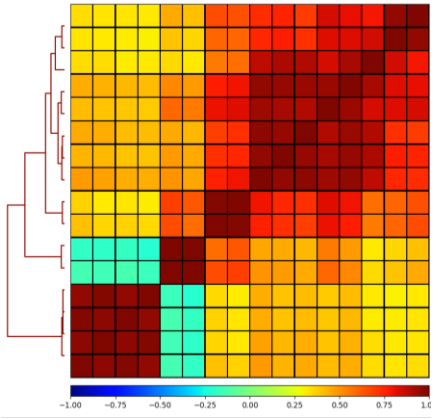


# What shall we do after generating a bam file?

- 2/ Assess reproducibility: Similarity between biological replicates?




Principal Component Analysis (PCA)



Correlation Heatmap (Pearson ou Spearman)

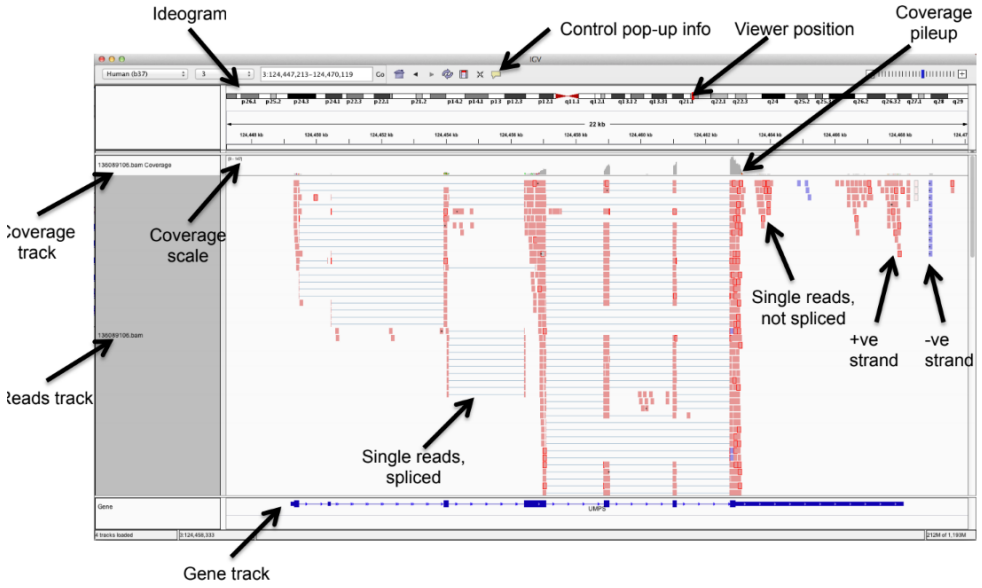
software: DeepTools

# What shall we do after generating a bam file?

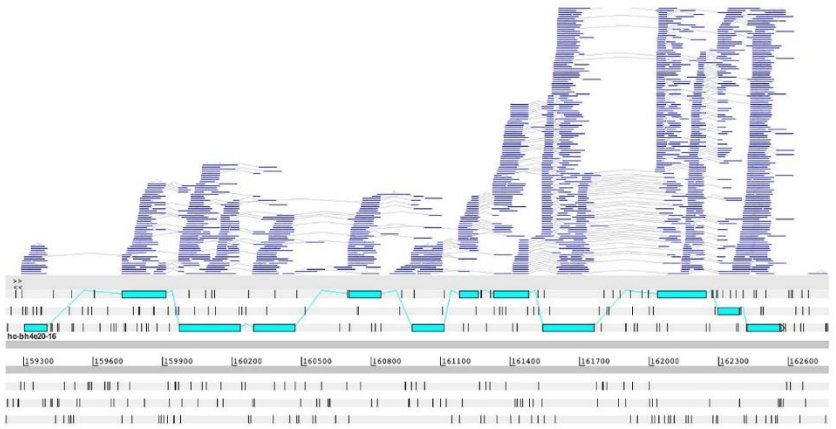
- 3/ Visualise your data!!!

several software:
IGV,
IGB,
Artemis,
Tablet,
UCSC genome browser

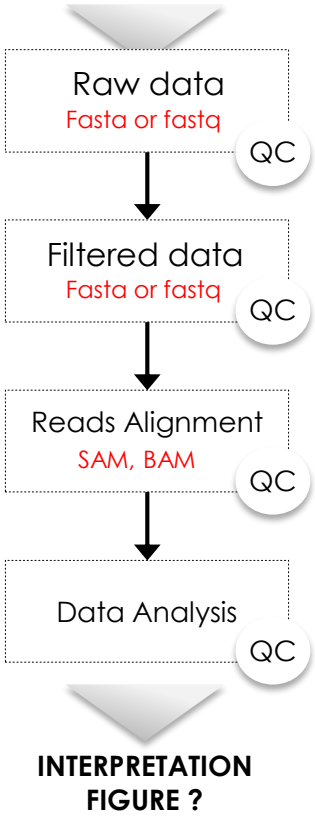
igv




Artemis



# Differential analysis:




Many tools
Many parameters

count reads
normalise
data interpretation

differential Gene expression

DESEQ2
EDGER
Stringtie/Ballgown

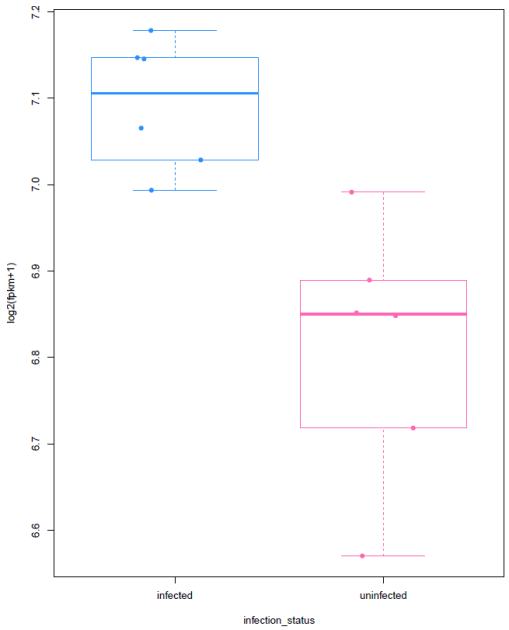
differential transcripts expression

DEXSEQ
Stringtie/Ballgown

- RSEM/Feature Counts/HTSeq: annotations are critical and rely on the GTF you gave the software

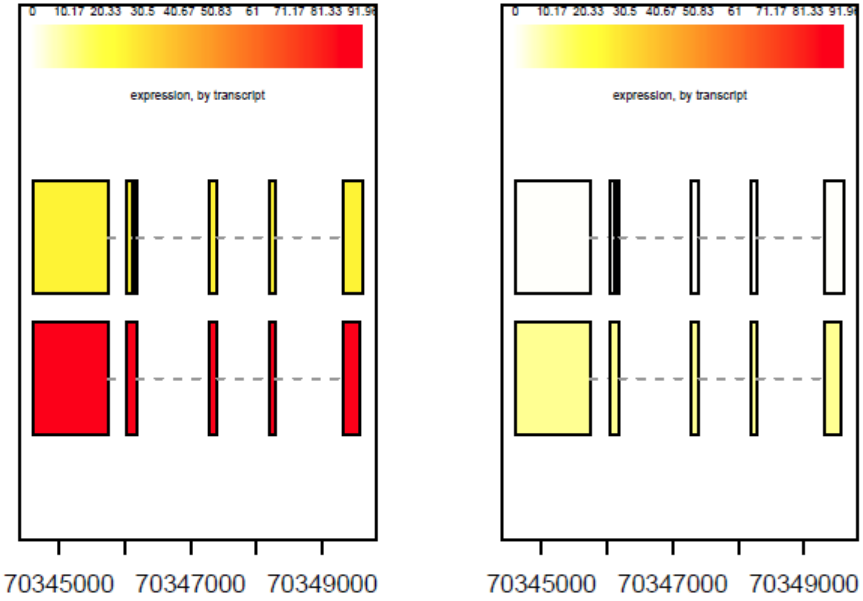
# DEG versus Alternative splicing : an example

infected uninfected




gene-level

infected uninfected




transcript level

# Count reads on features and normalise counts:

software:
htseq-counts
featureCounts
RSEM
summarizeOverlaps
tximport

the easiest way to quantify reads is to transform raw read counts into one of the following metrics per gene(exon):

RPKM : Reads per kilobase of exon model per million reads

FPKM : Fragment per kilobase of exon model per million read (for paired-end).

TPM : Transcripts per million

need for a GFF/GTF file (tabulated file with fields per feature defined by conventions).

```
chr22  TeleGene  enhancer  10000000  10001000  500  +  .  touch1
chr22  TeleGene  promoter  10010000  10010100  900  +  .  touch1
chr22  TeleGene  promoter  10020000  10025000  800  -  .  touch2
```

Name (chrom) | source | feature | Start | end | score | strand | frame | group

gene size? Library length?

➔ to be discussed in detail in another lecture

# Normalise and correct biases, then visualise!

inter individual

**EdgeR ou DESeq2 (ou BaySeq)**
**(explore bioconductor vignettes F1000Research protocols)**

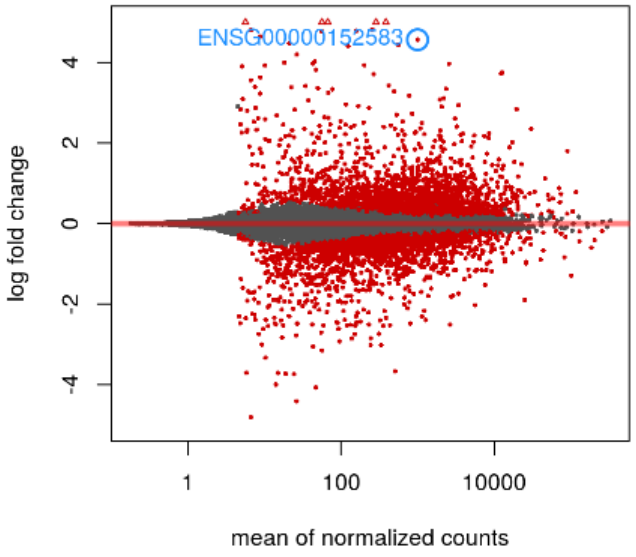
count table: per gene & per sample → log fold-change, P-value, adjusted p-value, etc... in a data frame, SummarizedExperiment or DGEList objects. → Visualisation and functional biology.

DeSeq2 example

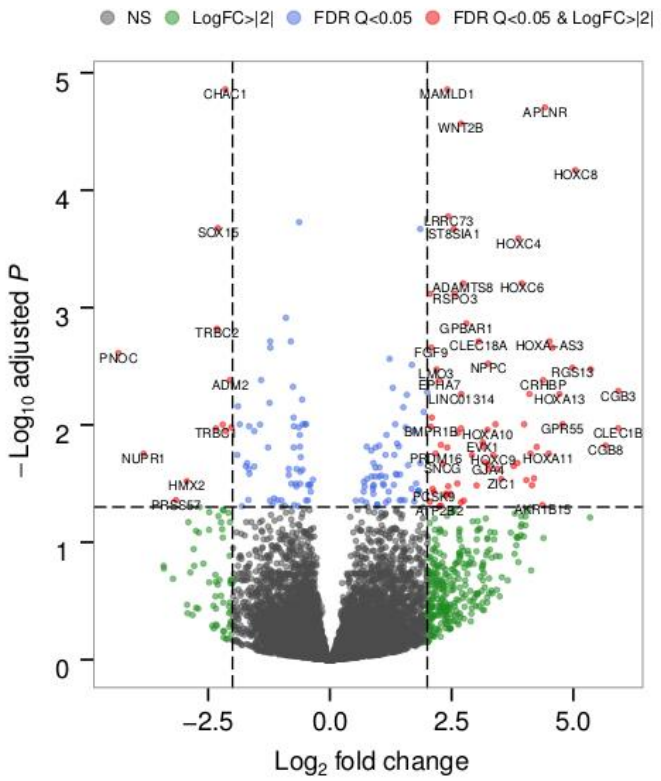
| | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj |
|---|---|---|---|---|---|---|
| Aats-phe | 1439.85510 | -0.3915108 | 0.10641530 | -3.679084 | 2.340731e-04 | 8.682721e-03 |
| achi | 1114.41542 | -0.4206245 | 0.10794425 | -3.896682 | 9.751936e-05 | 4.128319e-03 |
| Act42A | 25233.52971 | -0.4144380 | 0.07727588 | -5.363096 | 8.180730e-08 | 8.283542e-06 |
| Ada | 514.03083 | -0.6321073 | 0.13696097 | -4.615236 | 3.926483e-06 | 2.724179e-04 |
| ade5 | 3620.63094 | 0.8756724 | 0.12531134 | 6.987974 | 2.788849e-12 | 5.804679e-10 |
| Adgf-A | 4432.04719 | -0.3219797 | 0.08413694 | -3.826854 | 1.297917e-04 | 5.173027e-03 |
| alphaTub84B | 22505.94872 | -0.3424581 | 0.09110146 | -3.759085 | 1.705361e-04 | 6.423805e-03 |
| Ama | 198.23454 | -2.0373321 | 0.21461357 | -9.493026 | 2.244235e-21 | 1.401338e-18 |
| Ance | 1513.97966 | -0.3010685 | 0.09949477 | -3.025973 | 2.478346e-03 | 4.876555e-02 |

➔ see corresponding lecture.

# Genome wide visualisations : log fold changes



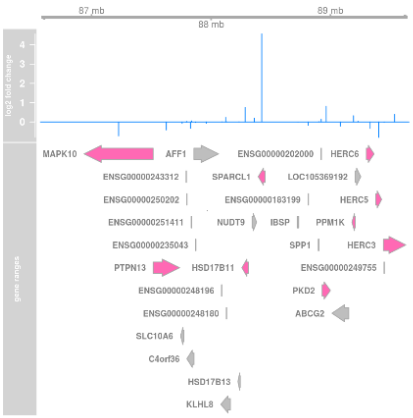

Mean-Average plot



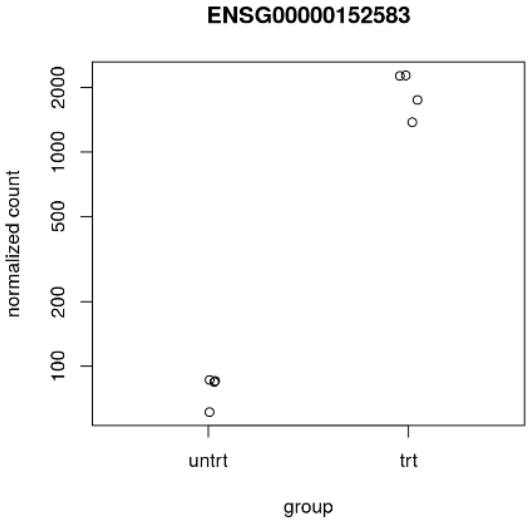

Volcano plot

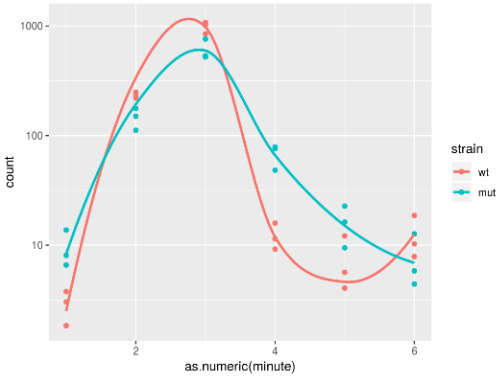
ggplot2
Edger
Deseq

# gene centric visualisations:



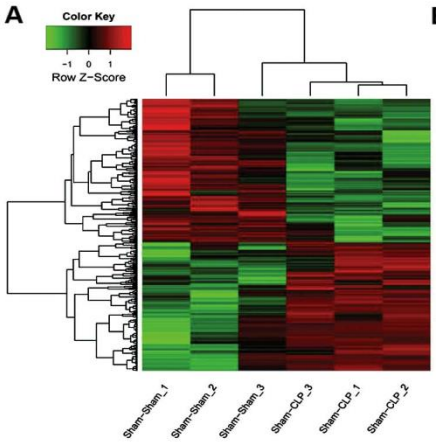

**log2 fold changes in genomic region surrounding the gene with smallest adjusted *p* value.**
Genes highlighted in pink have adjusted *p* value less than 0.1.

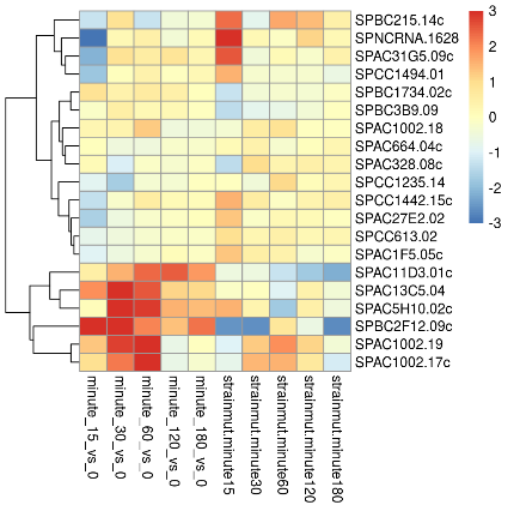



Gviz

# Gene expression heatmaps




Heatmap.2
Pheatmap
ggplot2

# A list of differentially expressed genes (isoforms), or newly discovered transcripts...So what?

➔ **Which pathways & biological processes are involved?**

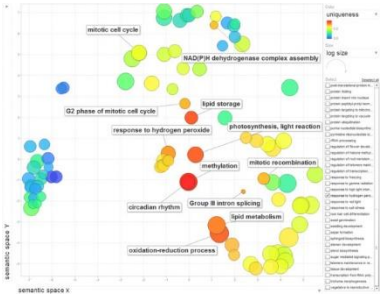
**Gene Ontology**

Gene Ontology Consortium

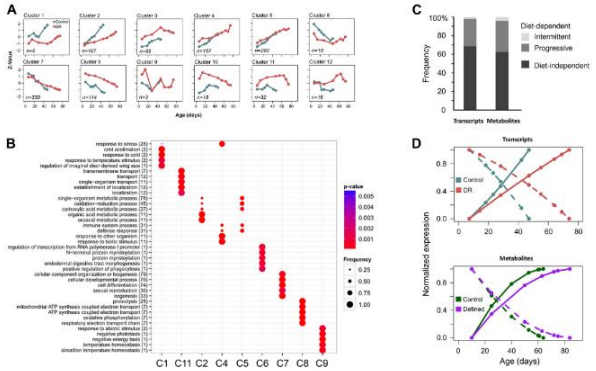
Enrichr

REVIGO

reduce + visualize Gene Ontology



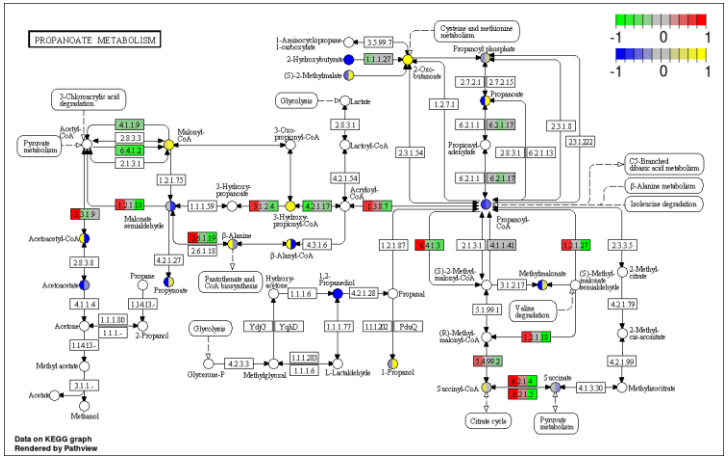

clusterProfiler



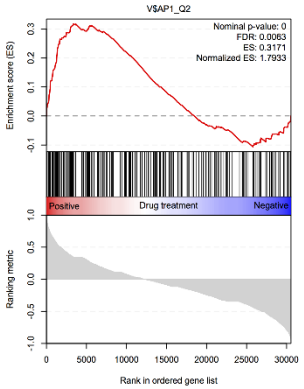

stringDB

R

**Kegg pathway analysis**




**MsigDB/GSEA**



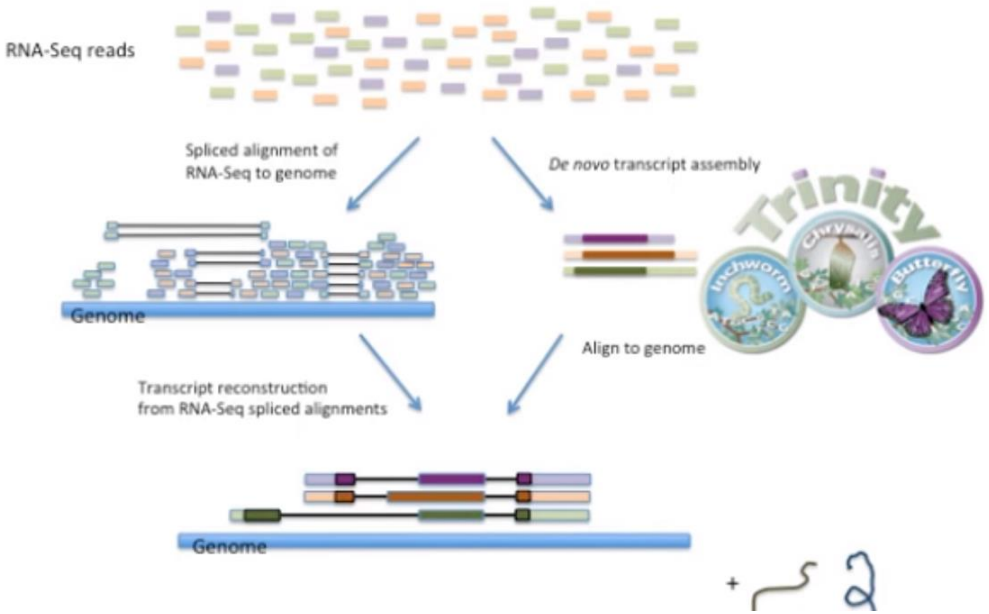

➔ see specific lecture

We had a look at known transcripts...

# How to find new transcripts?

# de novo transcript reconstruction

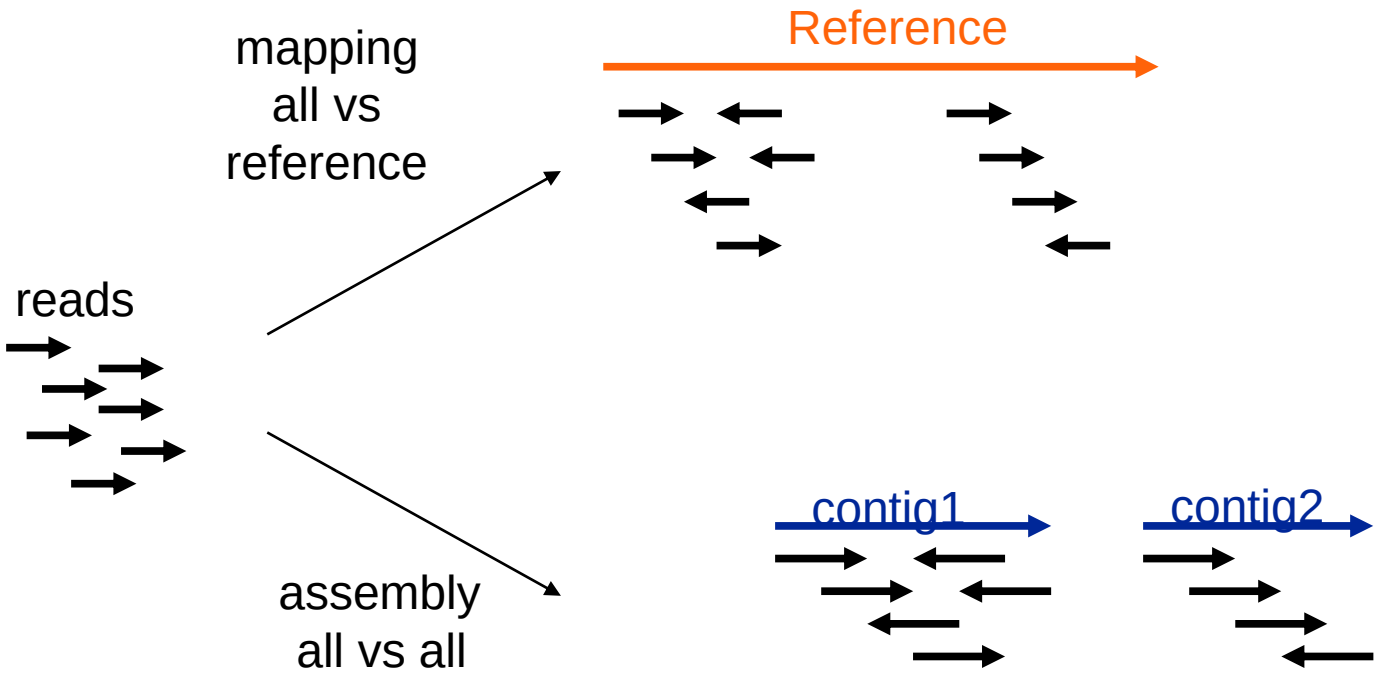
mapping




assembly

https://www.youtube.com/watch?feature=player_embedded&v=9ky5NwV45qY

# assembly vs. mapping



# Assembly: principle

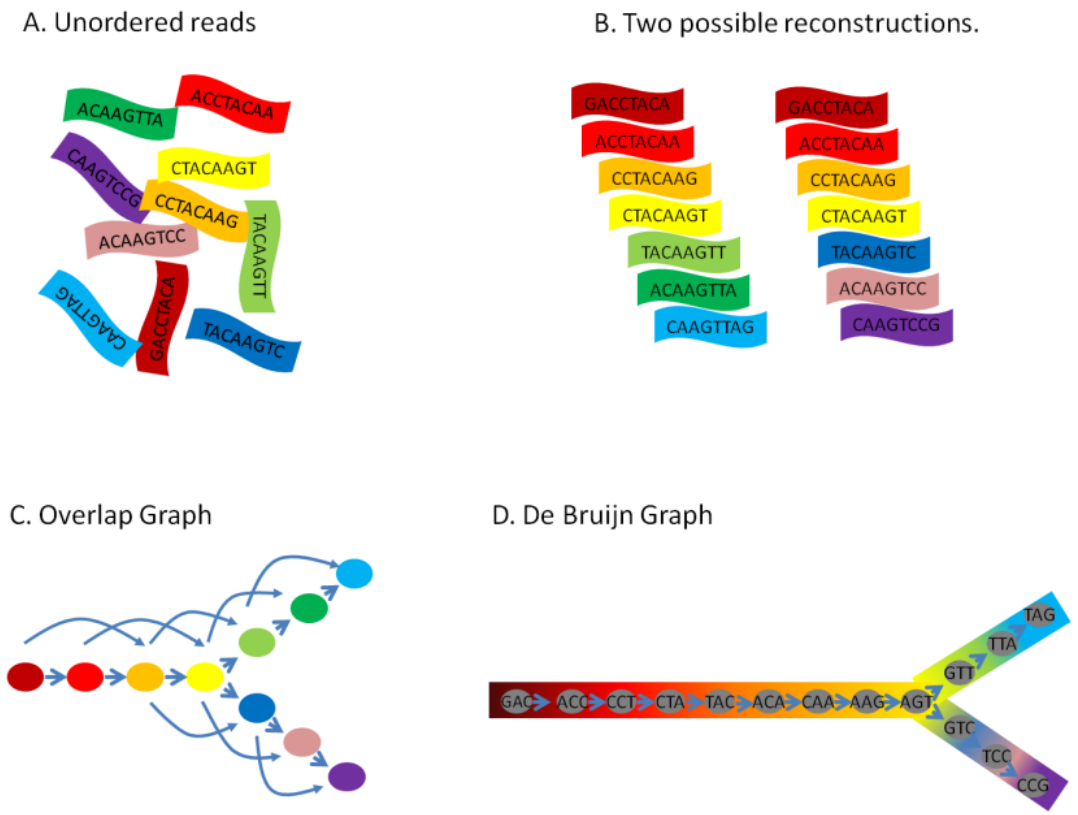



**Figure 4: Different graphical representation of the assembly paradigm.** An unordered set of reads (A) leading to two possible sequence reconstructions (B) can be abstracted as an Overlap graph, keeping into account relationships between reads (as shown by the colour code) or as a De Bruijn graph, where a set of k-mers (here k=3) and reads are implicitly depicted as paths through several nodes (as shown by the graduation in colour along the path). Of note a greedy algorithm would not have been able to resolve this assembly because of the same degree of difference at the fork. The fork illustrates the two sequences. Inspired from (Schatz et al. 2010).

# transcript reconstruction from mapping

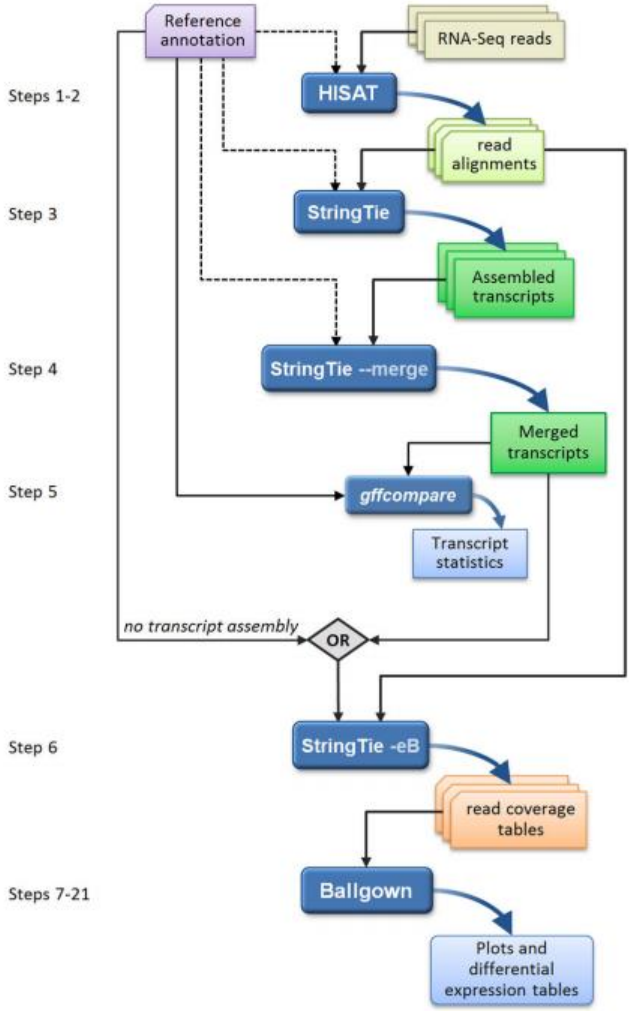
nature protocols

Protocol | Published: 11 August 2016

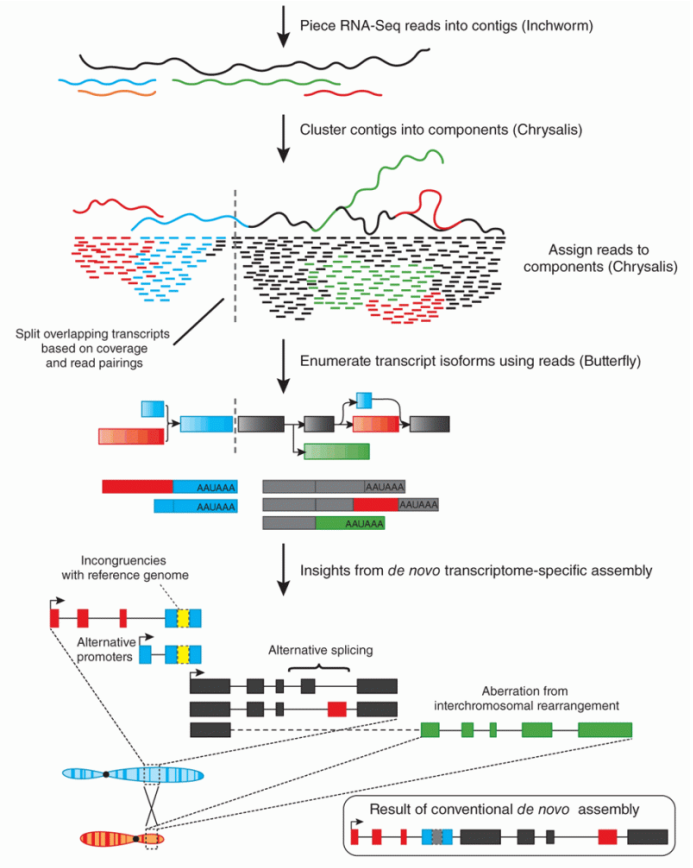
## Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek & Steven L Salzberg

*Nature Protocols* **11**, 1650–1667 (2016) | Download Citation



# transcript reconstruction from assembly



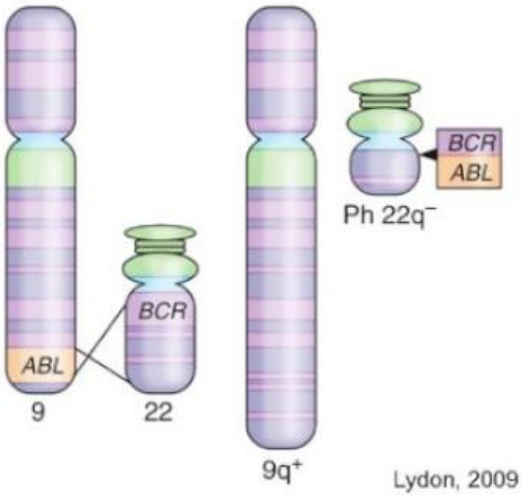

*Inchworm* assembles the RNA-seq data into the dominant isoform and reports unique portions of alternatively spliced transcripts.

*Chrysalis* constructs complete de Bruijn graphs for each cluster. Each cluster represents the full transcriptional complexity for a given gene.

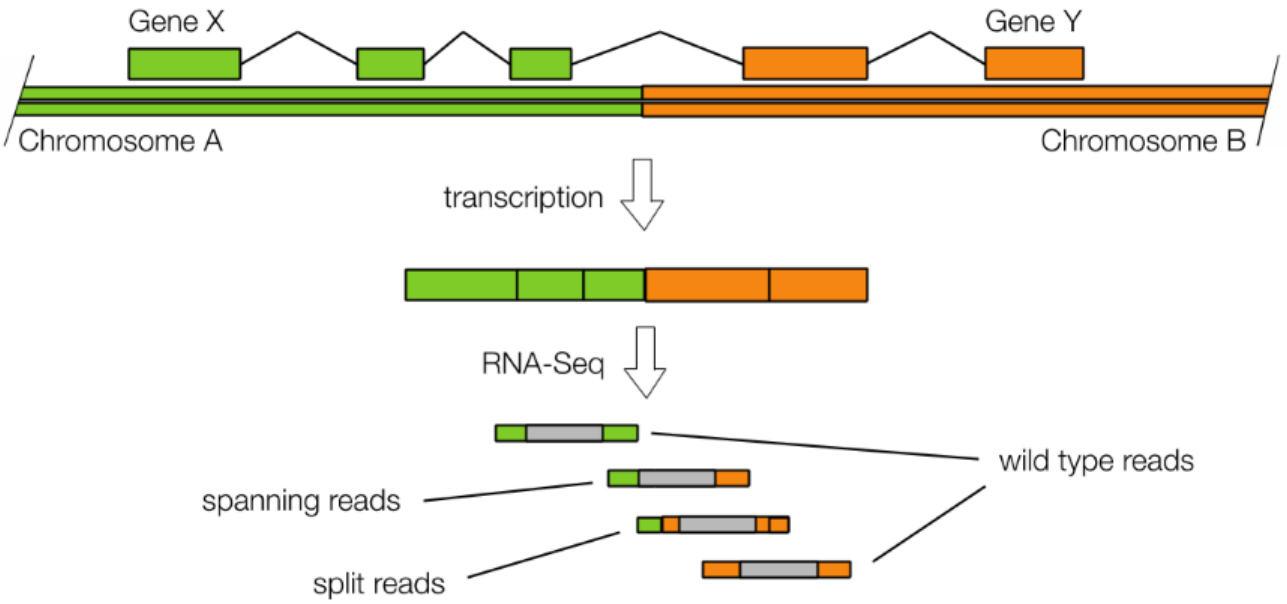
*Butterfly* then processes the individual graphs in parallel, reporting full-length transcripts for alternatively spliced isoforms

# Gene Fusion

- Novel gene formed by fusion of two distinct wild type genes
- In Cancer: produced by somatic genome rearrangements

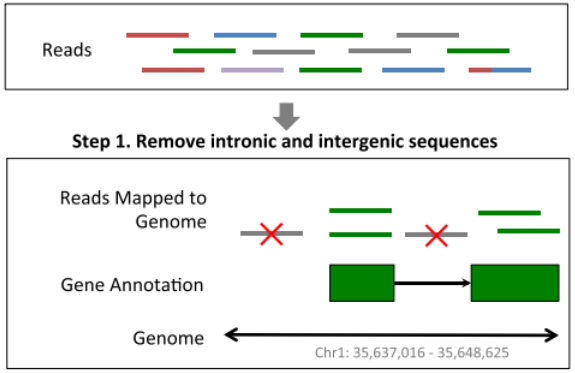


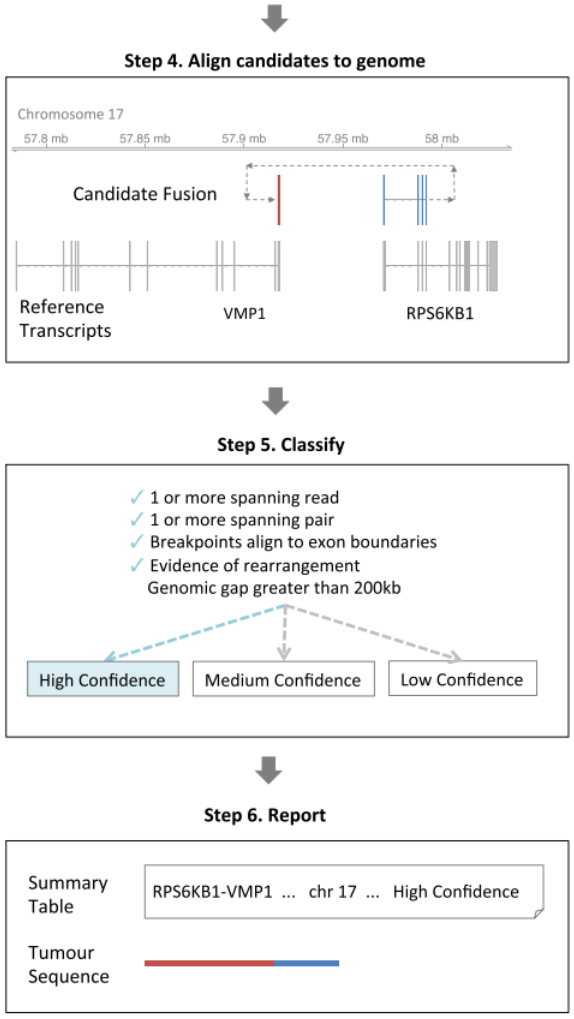
# Chimeric Fusion genes produce chimeric RNA-seq reads




Andrew McPherson, bioinformatics.ca

# An example:




METHOD Open Access

JAFFA: High sensitivity transcriptome-focused fusion gene detection

Nadia M Davidson[1*], Ian J Majewski[2,3] and Alicia Oshlack[1,4*]

Thank you for
your attention!!!

special thanks to:
Olga Rospopoff,
Agathe Duchateau,