Bulk RNA-seq

Introduction to RNAseq Methods

Xabier Bujanda Cundin
12/07/2023

HTS Applications - Overview

DNA Sequencing

- Genome Assembly
- SNPs/CNVs identification
- DNA methylation
- DNA-protein interactions (ChIPseq)
- Chromatin Modification (ATAC-seq/ChIPseq)

RNA Sequencing

- Transcript Assembly
- Differential Gene Expression
- Fusion Genes
- Splice Variants
- Protein-RNA interactions (iCLIP)

Single Cell

- RNA/DNA
- Low RNA/DNA detection level
- Cell-type identification
- Dissection of heterogeneous cell populations

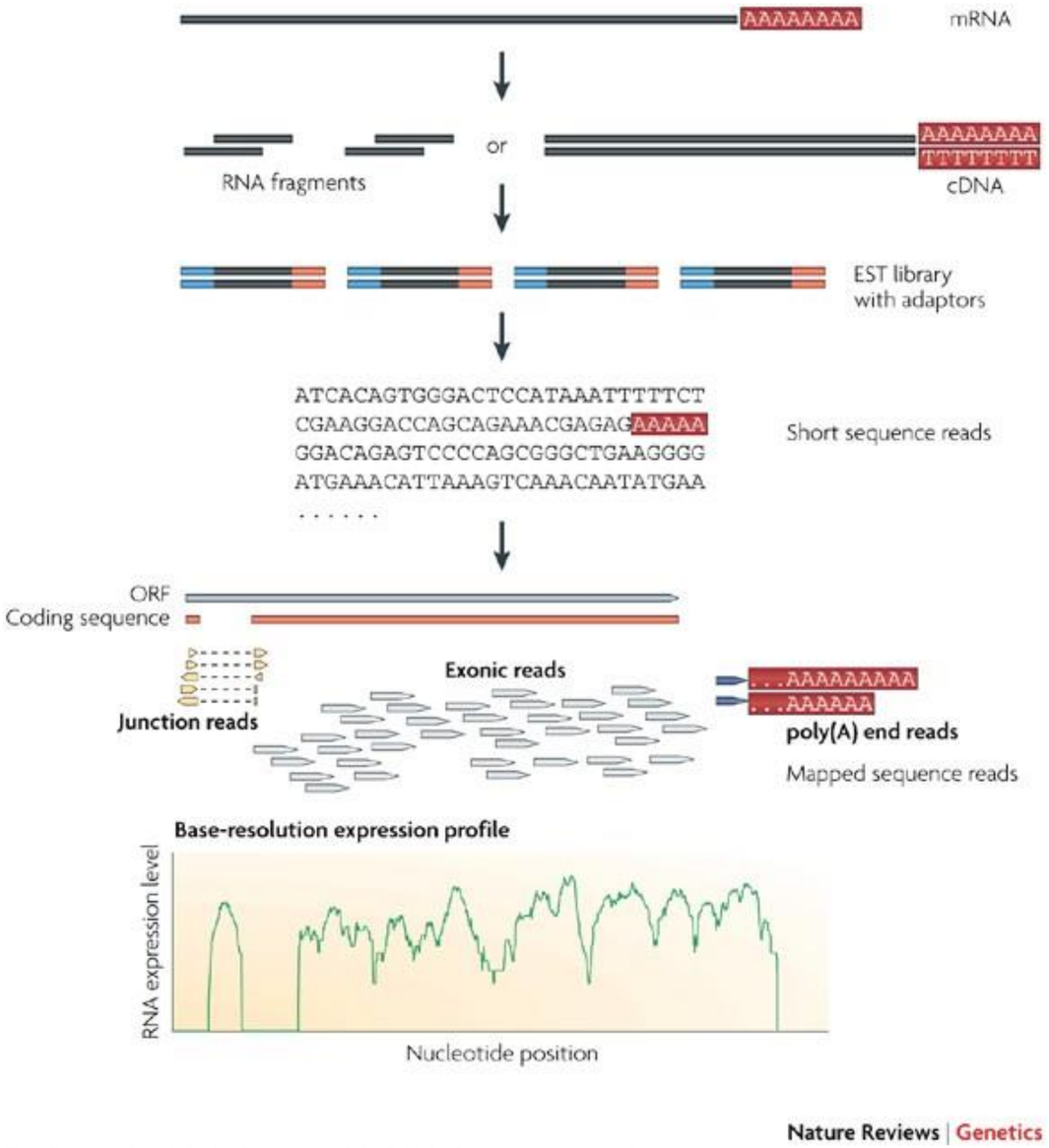
RNAseq Workflow

Experimental Design

Library preparation

Sequencing

Bioinformatics Analysis



Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009

Designing the right experiment

A good experiment should:

- Have clear objectives
- Have sufficient statistical power
- Be amenable to statistical analysis
- Be reproducible

Designing the right experiment

A good experiment should:

- Have clear objectives
- Have sufficient statistical power
- Be amenable to statistical analysis
- Be reproducible

Practical considerations:

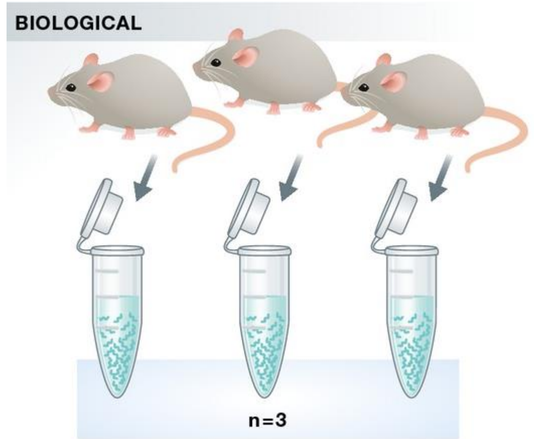
- Coverage → How many reads?
- Read length and structure
- Batch effect consideration
- Library preparation method selection

Designing the right experiment – Replication

Biological Replication

- Accounts for biological variations between individuals
- Sampling bias

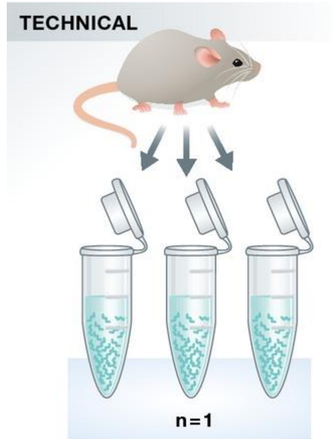
Each replicate comes from an independent individual



Technical Replication

- Accounts the variation due to imprecision in the technique
- Technical noise

Replicates are from the same individual but processed separately



Please, process as many samples as possible

https://doi.org/10.15252/embj.201592958

Designing the right experiment – How many reads?

Coverage is defined as:

$$\frac{Read\ Length\ x\ Number\ of\ reads}{Length\ of\ Target\ Sequence}$$

Considerations

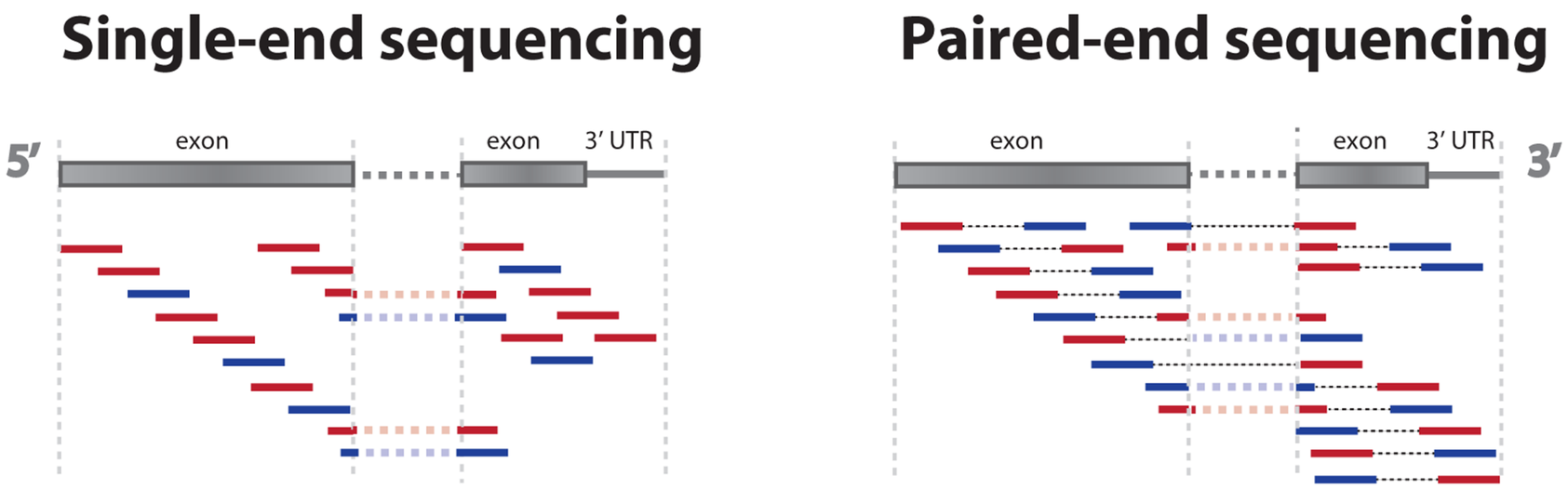
- For general differential expression: 5-25 million reads per sample
- For alternative splicing and low expressed genes: 30-60 million reads per sample
- In-depth view of the transcriptome/assemble new transcripts: 100–200 million reads
- Targeted RNA expression requires fewer reads.
- miRNA-Seq or Small RNA Analysis require even fewer reads.

If working with tight budget: Samples >>> Coverage

Designing the right experiment – Read length

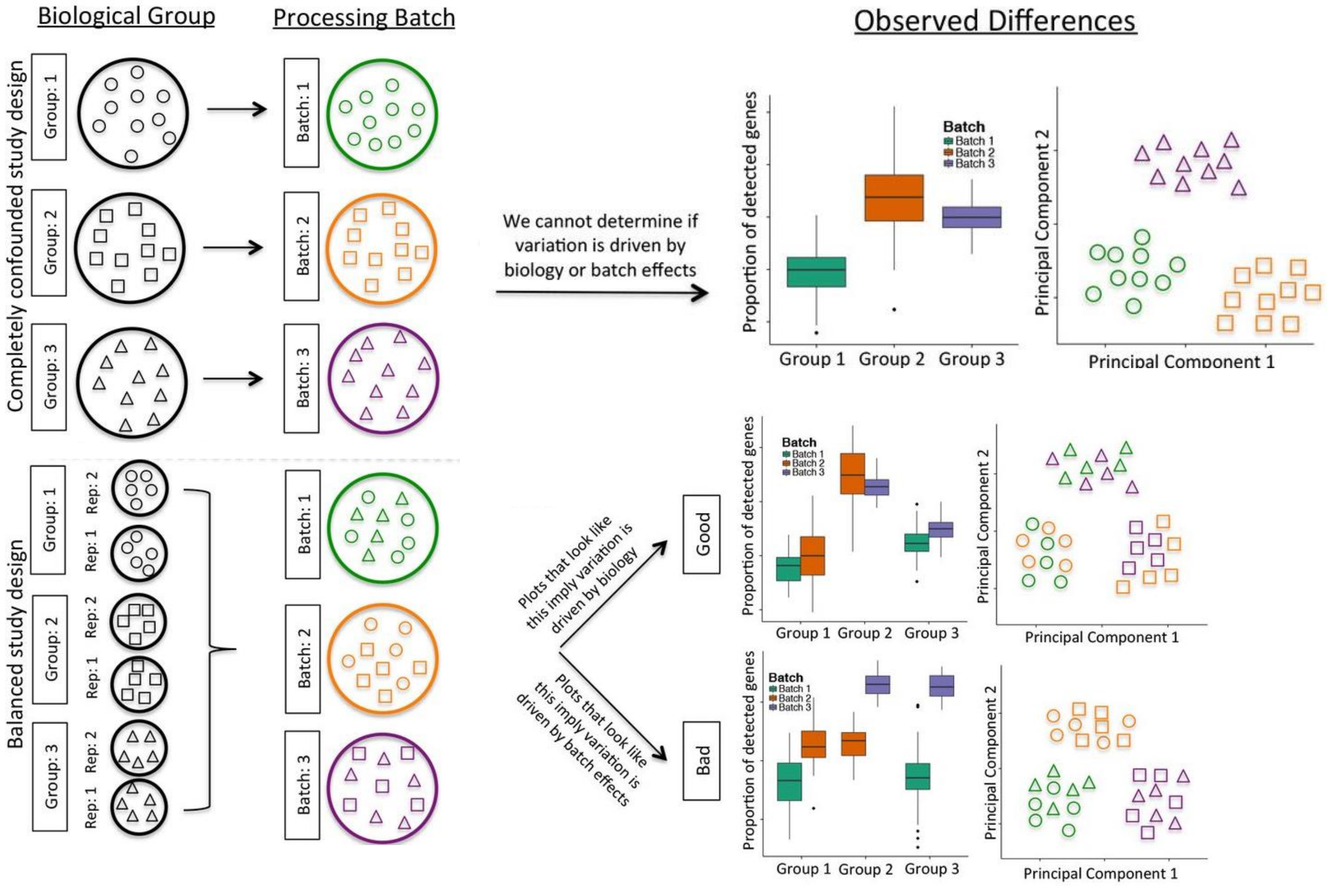
Short or long read sequencing? Paired or Single end reads?

- Gene expression → 75 bp; Short read
- Transcriptome Analysis → longer paired-end reads (2 reads x 75 bp each)
- Small RNA Analysis → short single-end reads
- Novel isoforms and splicing regulation → Long read sequencing (10.000 bp)



Designing the right experiment – Batch effects

- Batch effects are technical sources of variation that have been added to the samples during handling
- Batch effects are problematic if they are confounded with the experimental variable.

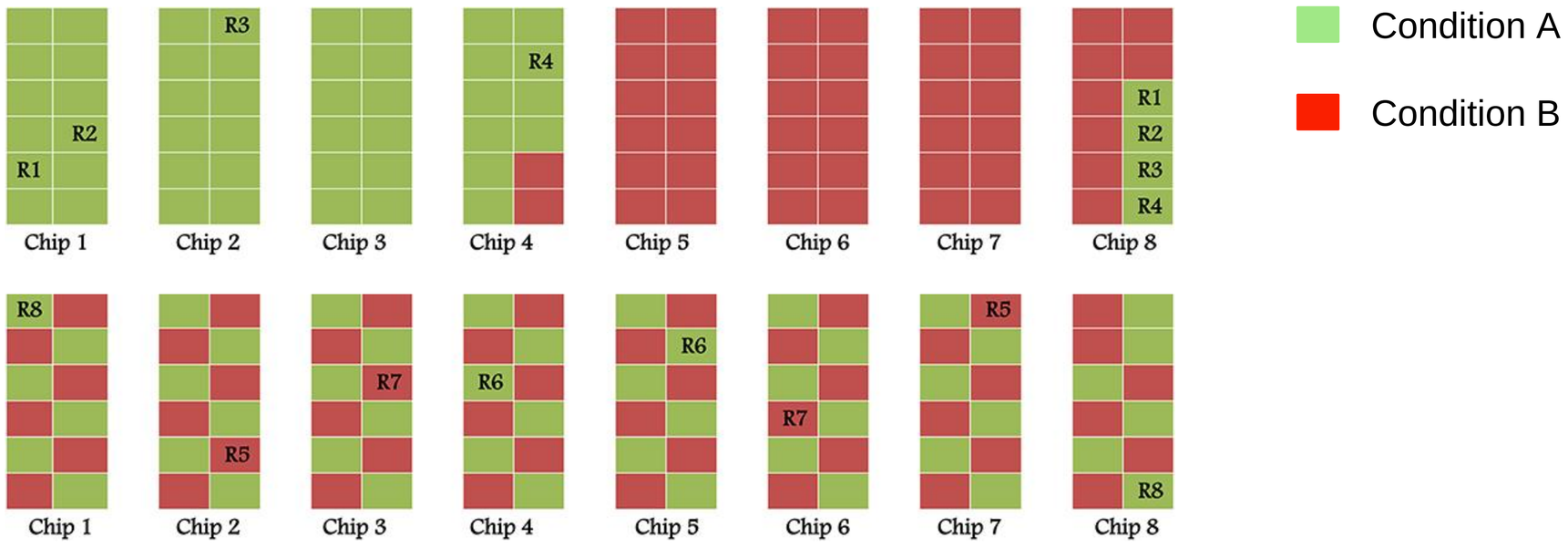


Designing the right experiment – Batch effects

- Batch effects are technical sources of variation that have been added to the samples during handling
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects randomly distributed across variables can be controlled
- Randomize all steps in order to avoid batch effects

Experimental design 1

Experimental design 2



Designing the right experiment – Batch effects

- Batch effects are technical sources of variation that have been added to the samples during handling
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects randomly distributed across variables can be controlled
- Randomize all steps in order to avoid batch effects

Record every single potential batch effect condition: Technician, days of sample extraction, cell passage…

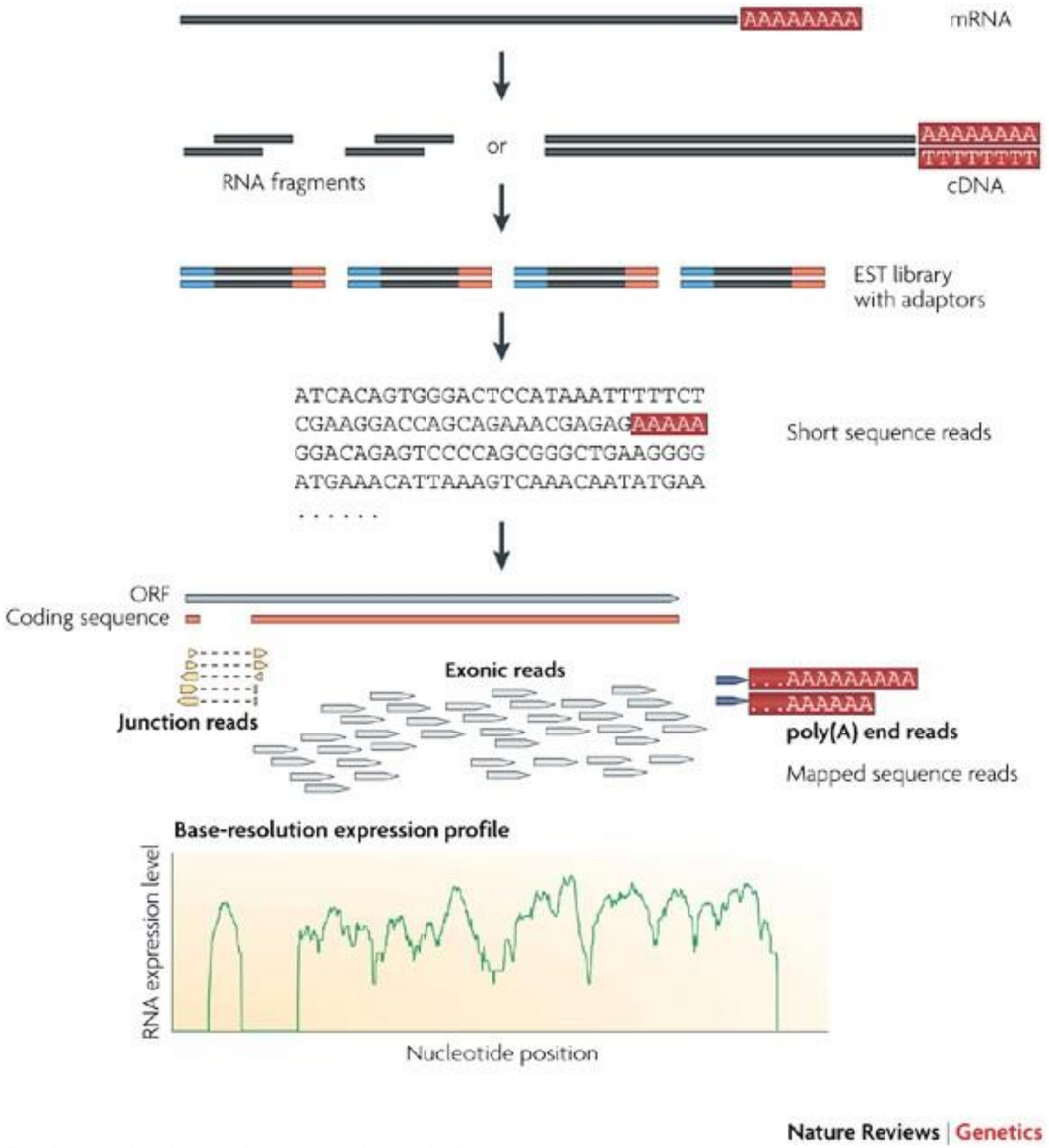
RNAseq Workflow

Experimental Design

Library preparation

Sequencing

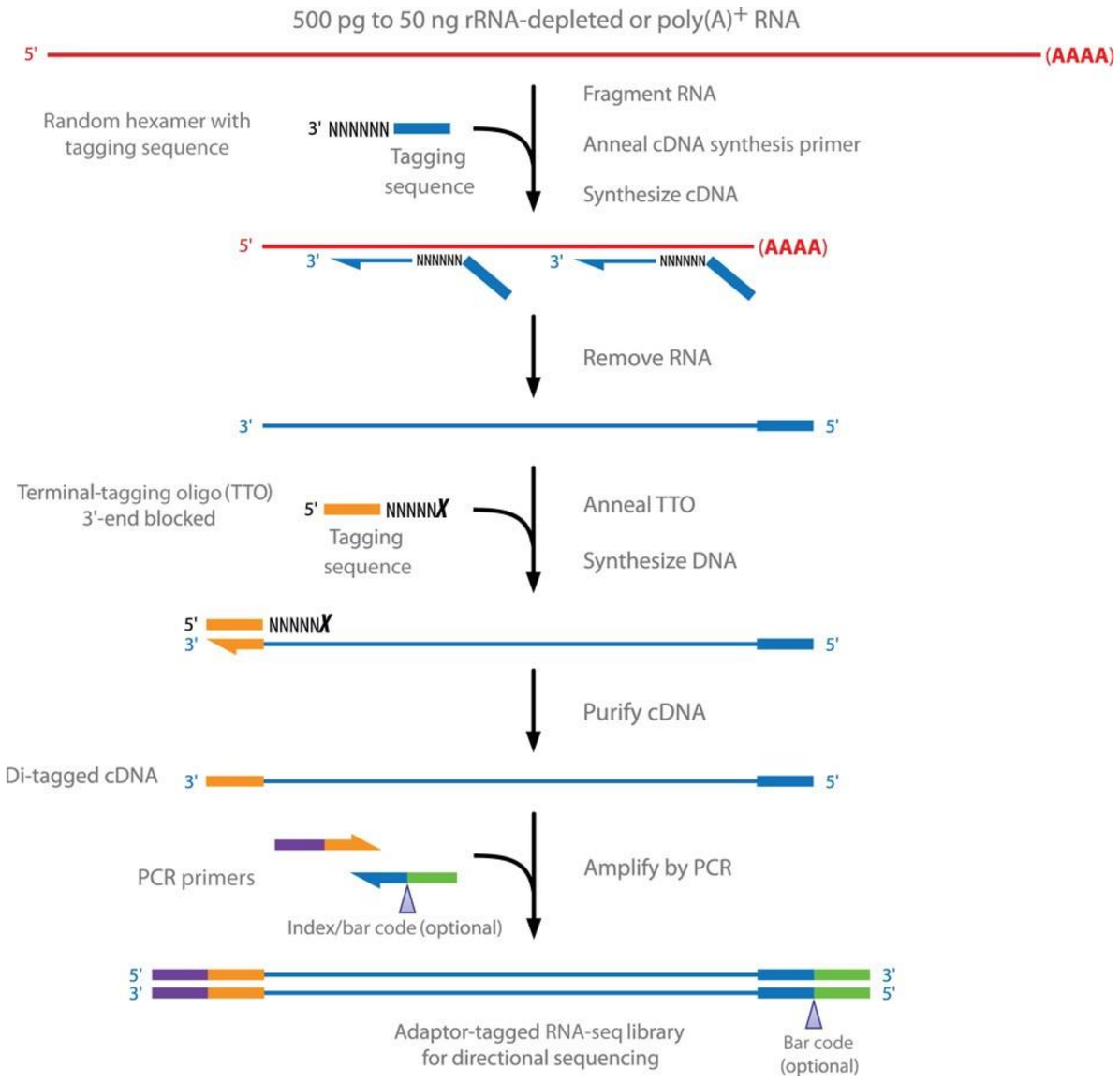
Bioinformatics Analysis



Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009

Library Preparation

1. PolyA + RNA capture
2. RNA fragmented and primed
3. First strand cDNA synthesized
4. 3’ and 5’ ends repaired
5. Adapters ligation
6. PCR amplification



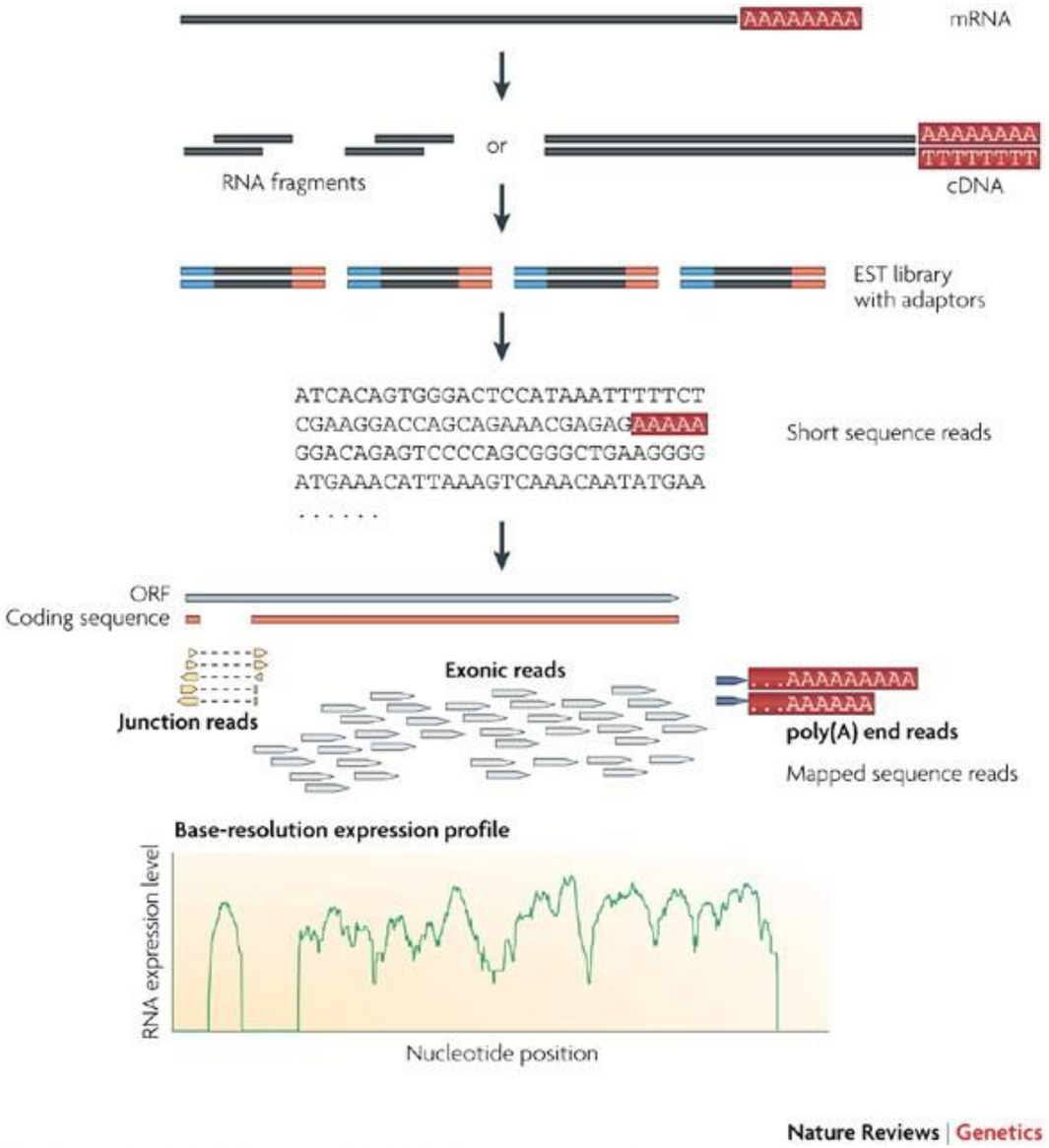
RNAseq Workflow

Experimental Design

Library preparation

Sequencing

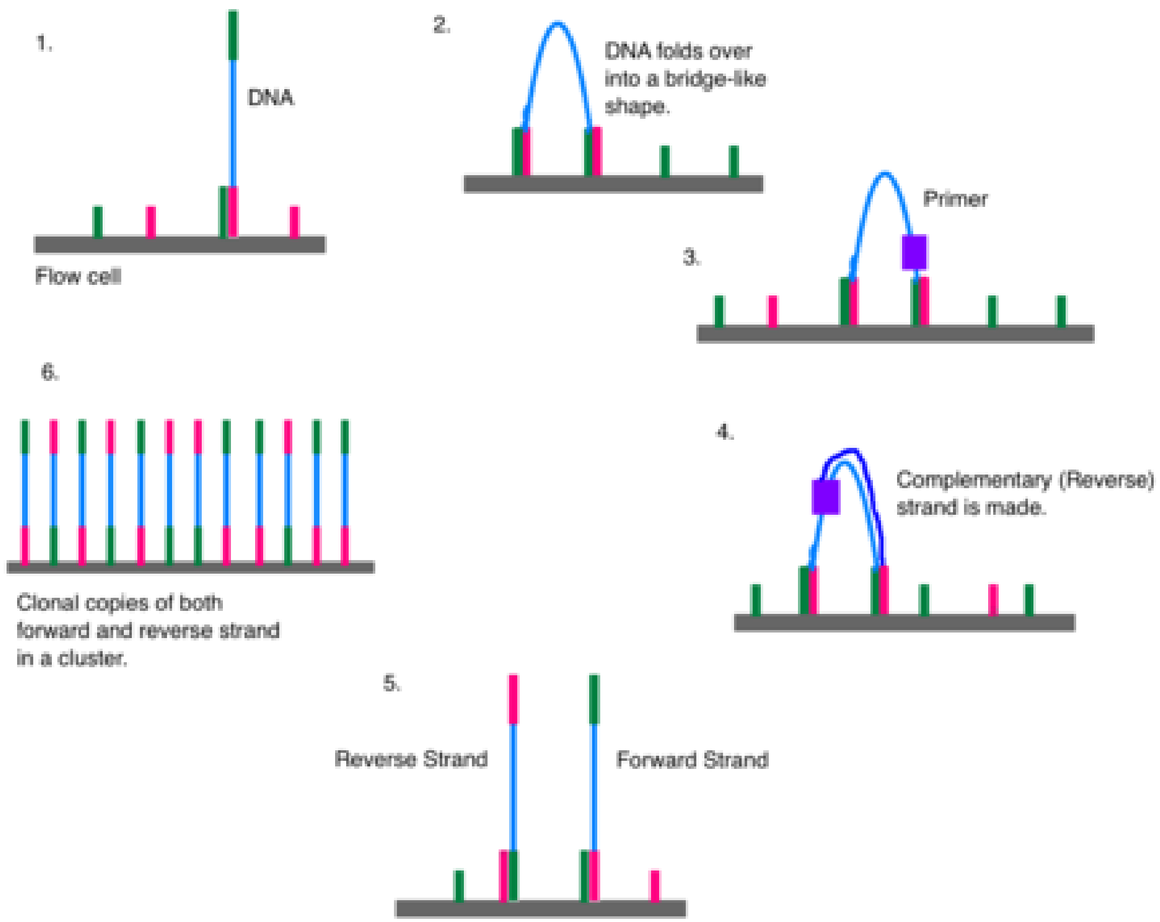
Bioinformatics Analysis



Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009

Sequencing

https://emea.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html



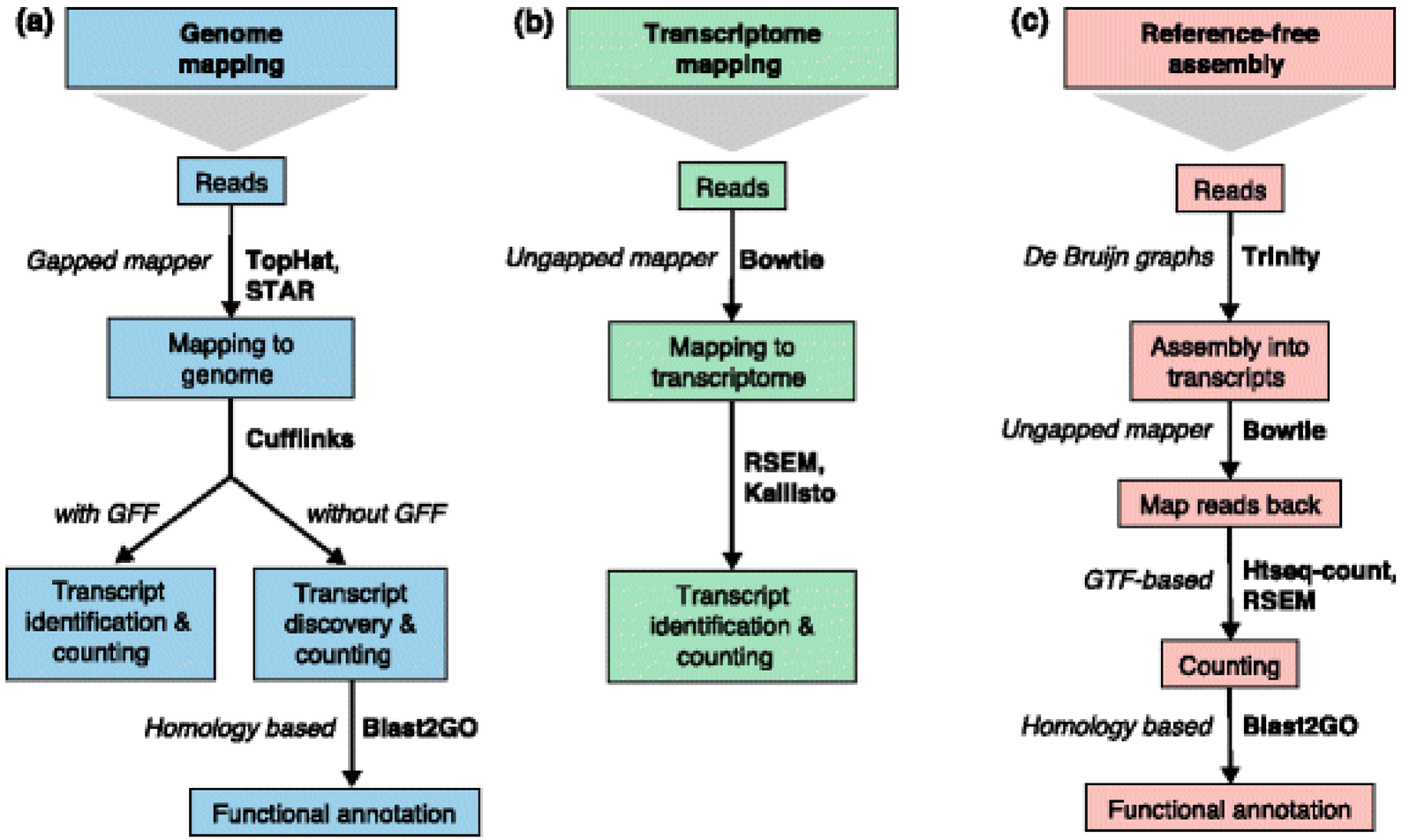
RNAseq Workflow

Experimental Design

Library preparation

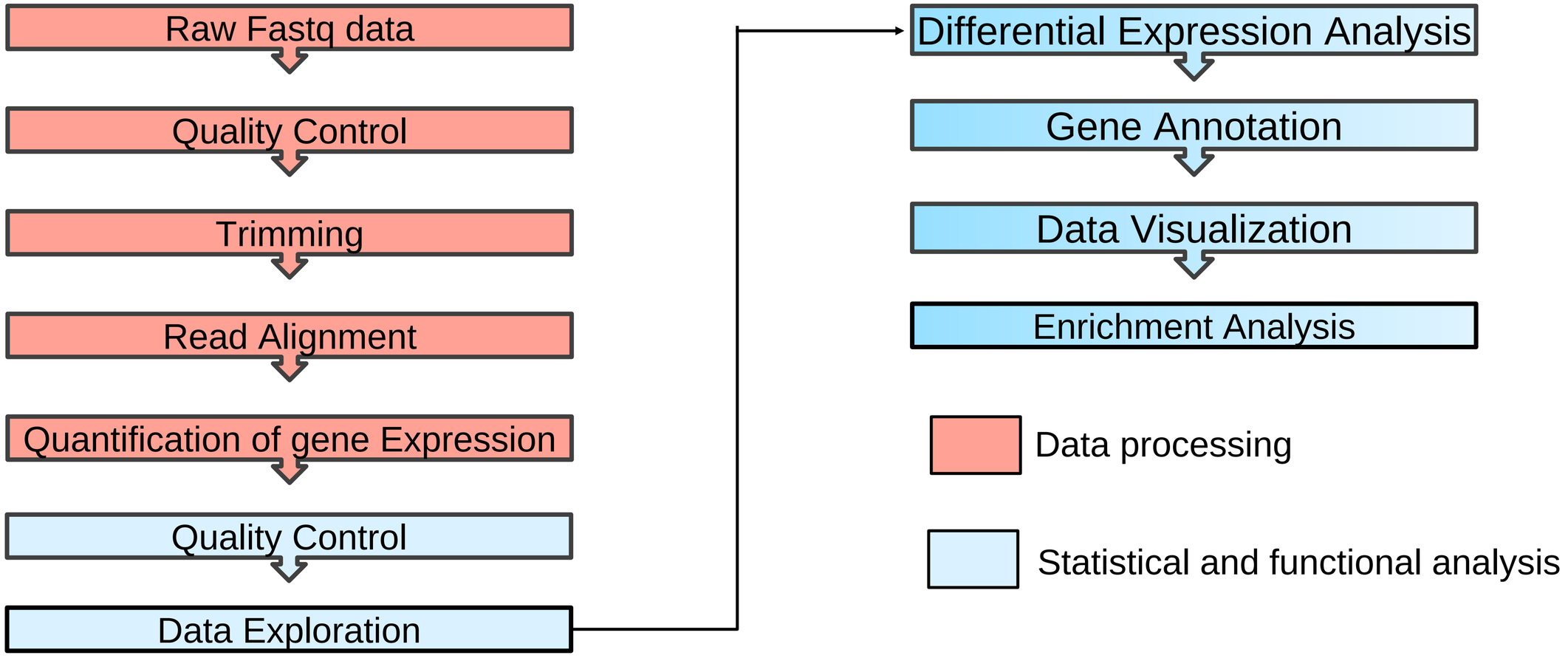
Sequencing

Bioinformatics Analysis



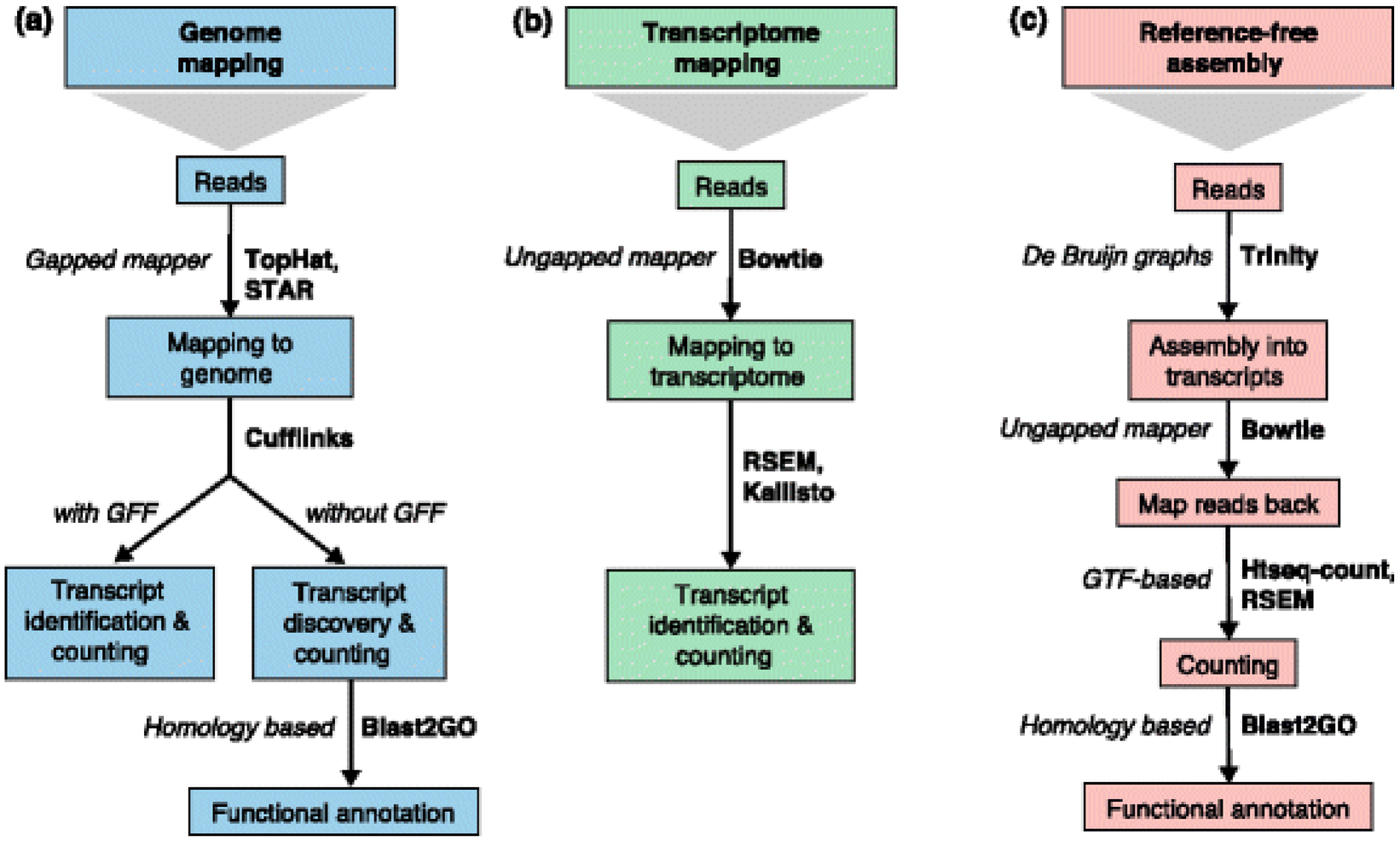
Conesa, A., Madrigal, P., Tarazona, S. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13 (2016)

Bioinformatics Analysis



Bioinformatics Analysis

Different data processing methods answer different biological questions



Bioinformatics Analysis

Starting file: .fastq file

```
@A00560:228:HM7HMDRXY:2:2101:3061:1000 1:N:0:AGTTGA+NTTACA
NAGTGGTTTCGGGGGGAGTATGGTTGCAAAGCTGAAAAAAAAAAAAAAAAGATGGGAAGAGCACACGTCTGAACTCCAGTCACAGTTGAATCTCGGATTGG
+
#FFFFFFFF::FFFFFFFFFFFF:F:FFFFFFFFFFFFFFFFFFFFFFFF,F:,:FFF,FF:F::F,,,F,FF:FFFF,FF:::FFFFFF,,,FF,,:,,,
@A00560:228:HM7HMDRXY:2:2101:3513:1000 1:N:0:AGTTGA+NTTACA
NAAGCCCCTTGAGCATTATCCTTCCACACTGCATCTCAAAAAAGGAAAAAAAAAAAAAAAAAGATCGGAAGAGCACACGTCTGAACTCCAGTCACAGTTGAA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFF:F,FFFFFFFFFFFFF,FFFF,FFFFFFFFFFF,,F::FF::,FF:F,::FF:FFFF,::FFFF:FF,FFFFF,
@A00560:228:HM7HMDRXY:2:2101:6840:1000 1:N:0:AGTTGA+NTTACA
NTTGTCATGGCTGTTCAGAAAGGAATTAAGGGCTTAGGCCAAAACTTACTTTGAGTAAGTTAAAAAAAAAAAAAAAAAAAGAACGCAAGAGCACACGACTGA
+
#FFFFFF:FFFFFFF,:FFFFFF,,FFFFF:,:FFFFFF,FFF:,FFF,,,F:F,,FFFF,:F:FFFFFFFFFFF:FFF,F,::,FF:FF:F:FF:F:,:F
@A00560:228:HM7HMDRXY:2:2101:29188:1000 1:N:0:AGTTGA+NTTACA
NACTGAGCTATGACTGCTCCCCTGCACTCCAGGCTGGGTGACAGAGTGAGACCCAGTCTCTAAAATAAAAAAAAAAAAAAAAAATAACGGAAGAGCACACG
+
#FF:FFFFFFFFF:FF:FFFFFFFFFFFFFFFFFFFFFFF:FF:FFF,FFFFFFF::FFFF:FFFFFF:FF,FFFFFFFFFF,,,,FF:F::FFFF:,F,
@A00560:228:HM7HMDRXY:2:2101:30581:1000 1:N:0:AGTTGA+NTTACA
NTCCTTGCCTTCTTTACTCGGCGTGCTCCTTTTCTCTTTGGGTTTCTTGTTTACCAAAGAAGAGTTTACAGACAATAAAATGGAAAGGTCCTGCTGTGGAA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFF:FF:FFFFF
@A00560:228:HM7HMDRXY:2:2101:2275:1016 1:N:0:AGTTGA+NTTACA
CGAGGCAGCCGGCTCATAAAGGTTCATTTGGACAAAGCACAGCAGAACAATGTGGAACACAAGGTTGAAACTTTTTCTGGTGTCTATAAGAAGCTCACAGG
+
FFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Bioinformatics Analysis

Starting file: .fastq file

```
@A00560:228:HM7HMDRXY:2:2101:30581:1000 1:N:0:AGTTGA+NTTACA
NTCCTTGCCTTCTTTACTCGGCGTGCTCCTTTTCTCTTTGGGTTTCTTGTTTACCAAAGAAGAGTTTACAGACAATAAAATGGAAAGGTCCTGCTGTGGAA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFF:FF:FFFFF
```

First line: Always begins with an @ followed by an identifier sequence and an optional description

Second line: Raw sequence letters

Third line: Begins with a plus symbol (optionally followed by the sequence identifier)

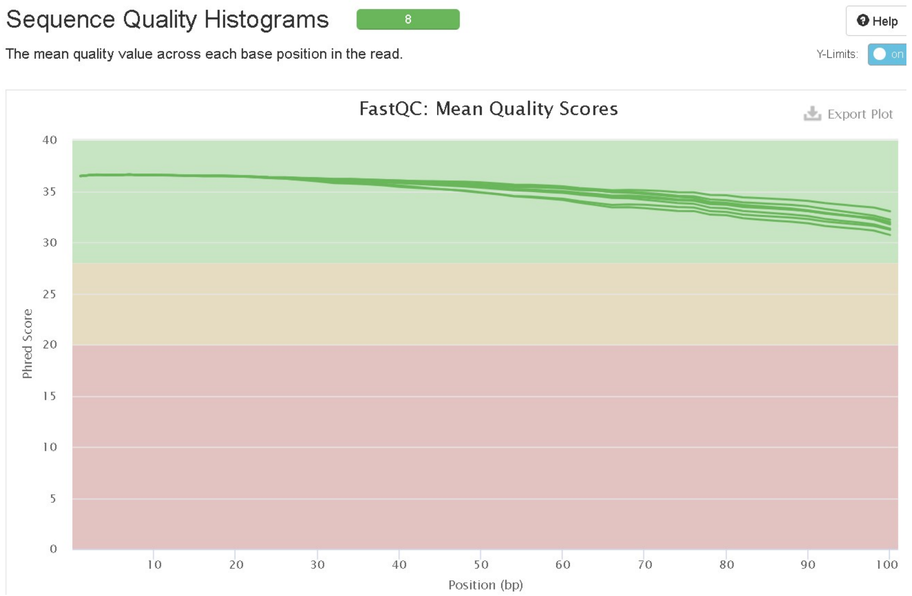
Forth line: Quality scores for each sequence letter. Must have the same length of the second line

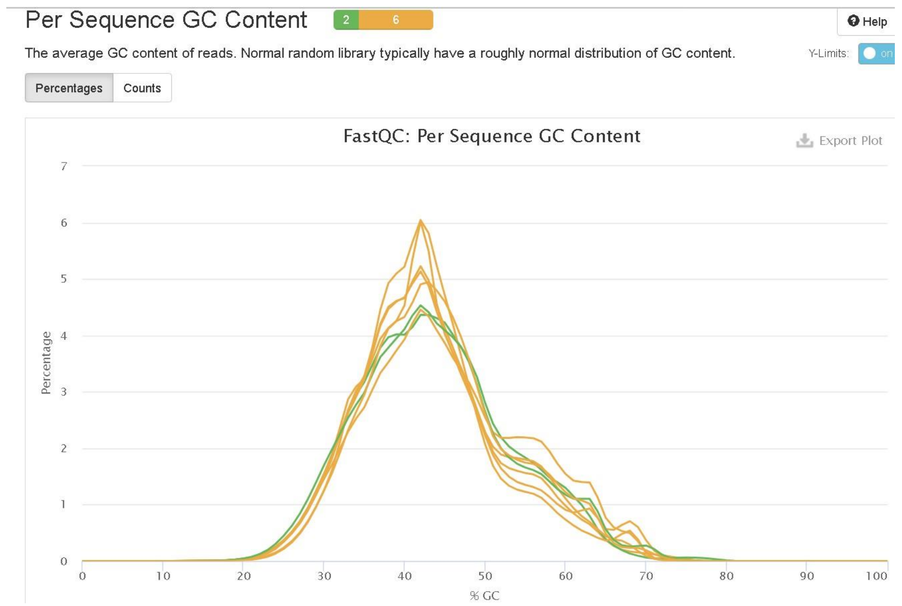
How quality (Q) of a read is computed? → Q = -10*Log (P) ;

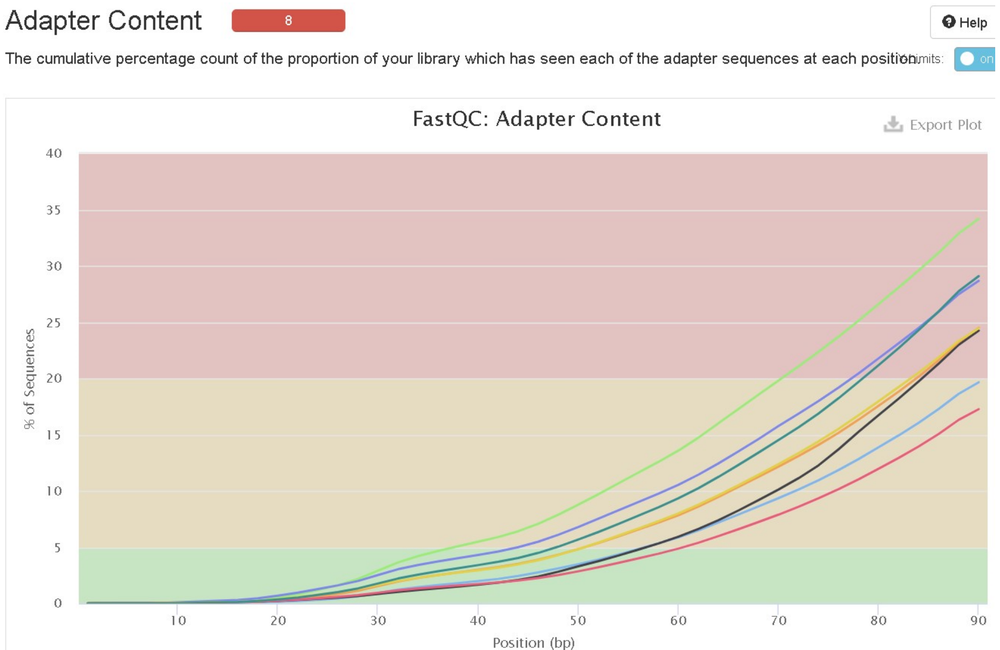
Being P the probability of a base being called incorrectly

Bioinformatics Analysis

Quality Control: FastQC

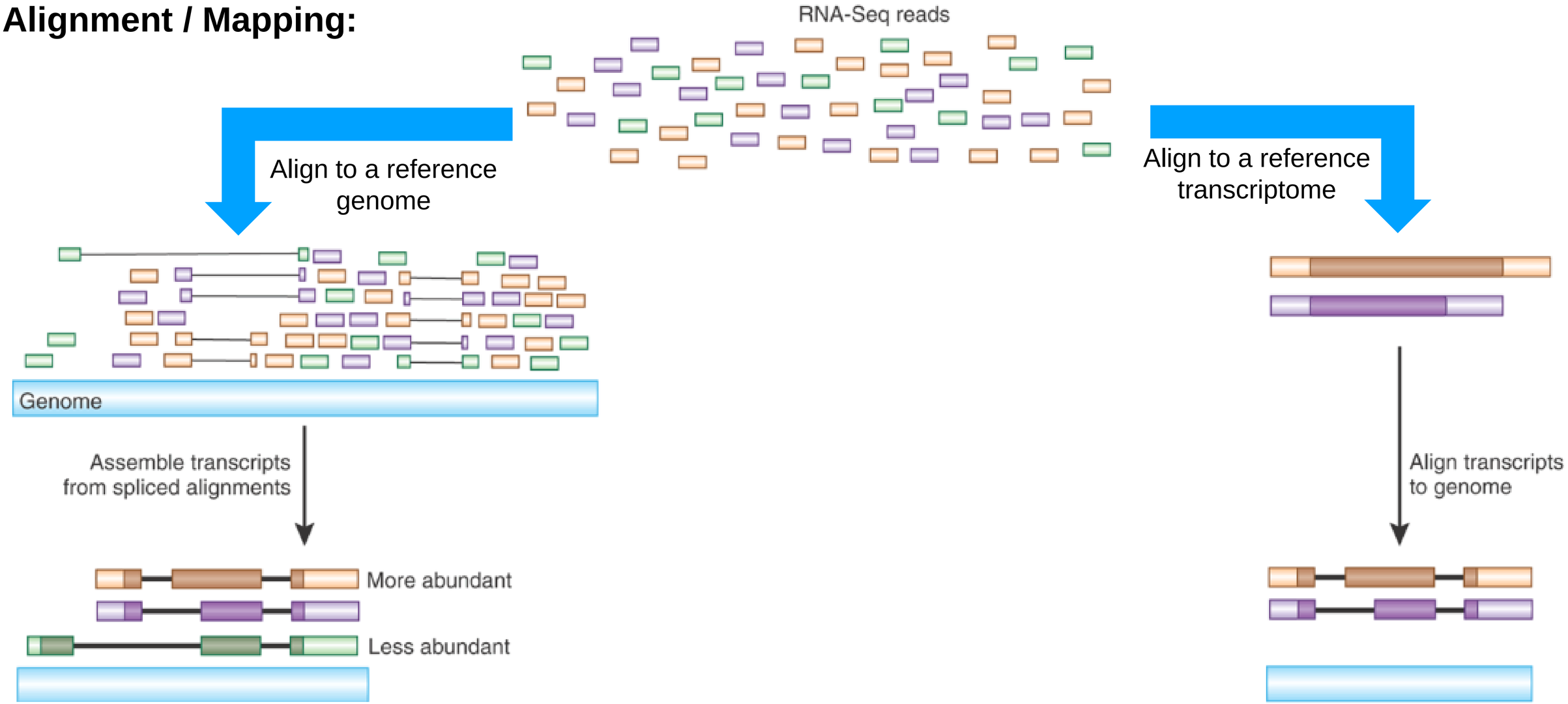






Bioinformatics Analysis

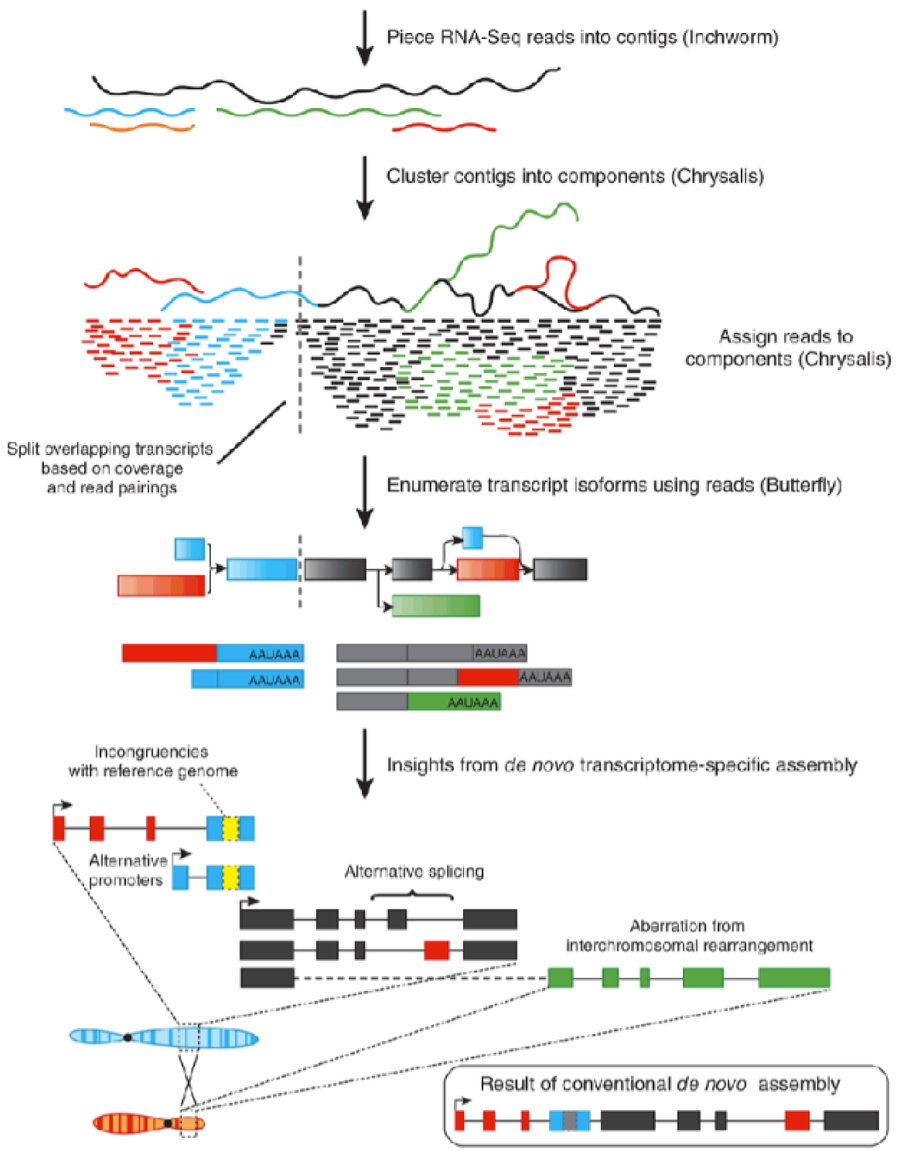
Alignment / Mapping:



Bioinformatics Analysis

Alignment / Mapping: *De novo* Alignment

1. Extract and count K-mers (substrings of length *k* contained in a sample)
2. Assemble initial contigs
3. Cluster overlapping contigs
4. Resolve alternating splicing and paralogous transcripts for each cluster



Bioinformatics Analysis

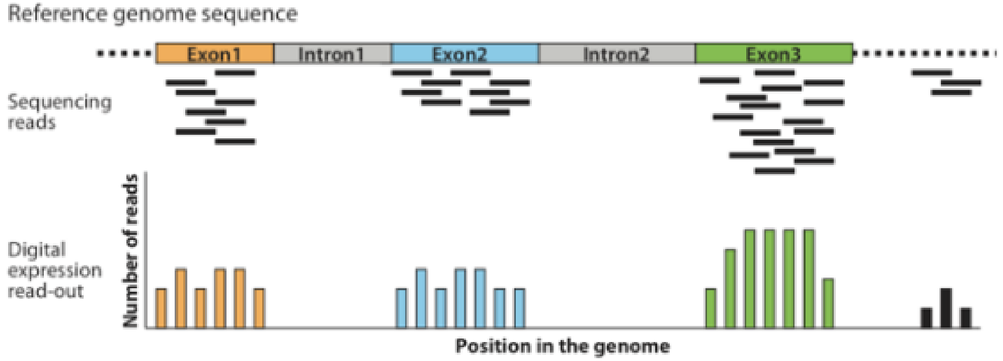
Alignment / Mapping: BAM file

Bioinformatics Analysis

Gene Quantification

- The abundance level of a gene is measured by the number of reads that map to that gene

```
1       ensembl_havana  gene    1471765 1497848 .       +       .       gene_id "ENSG00000160072"; gene_version "20"; gene_name "ATAD
3B"; gene_source "ensembl_havana"; gene_biotype "protein_coding";
```



	sample1	sample2	sample3	sample4	...
gene1	999	701	616	595	
gene2	532	520	41	26	
gene3	14	36	305	322	
...					

Bioinformatics Analysis

Gene Quantification

	SH_WT_ND_1	SH_WT_ND_2	SH_WT_ND_3	SH_WT_ND_4	SH_CDKL5_KO_ND_1	SH_CDKL5_KO_ND_2	SH_CDKL5_KO_ND_3	SH_CDKL5_KO_ND_4	SH_WT_D_1
ENSMUSG00000001023	0	0	1	0	0	0	0	0	0
ENSMUSG00000001025	141	61	88	88	241	149	96	62	48
ENSMUSG00000001027	0	0	0	0	0	0	0	0	0
ENSMUSG00000001029	0	0	0	0	0	0	0	0	0
ENSMUSG00000001034	8	0	3	4	7	2	3	3	1
ENSMUSG00000001036	55	31	38	61	31	38	46	18	32
ENSMUSG00000001039	2	0	0	0	0	1	1	0	0
ENSMUSG00000001052	19	6	9	10	16	18	9	4	0
ENSMUSG00000001053	0	0	0	0	0	0	0	0	0
ENSMUSG00000001054	0	0	0	0	0	0	0	0	0
ENSMUSG00000001056	39	16	10	27	49	62	43	59	37
ENSMUSG00000001062	0	0	0	0	0	1	0	0	0
ENSMUSG00000001065	1	0	0	0	0	0	0	0	0
ENSMUSG00000001076	0	0	0	0	0	0	0	0	0
ENSMUSG00000001082	0	0	0	0	0	0	0	0	0
ENSMUSG00000001089	111	76	95	106	95	120	115	86	66
ENSMUSG00000001095	0	0	0	0	0	0	0	0	0
ENSMUSG00000001098	0	0	0	0	0	0	0	0	0

Bioinformatics Analysis

Gene Quantification

	SH_WT_ND_1	SH_WT_ND_2	SH_WT_ND_3	SH_WT_ND_4
ENSMUSG00000001023	0	0	1	0
ENSMUSG00000001025	141	61	88	88
ENSMUSG00000001027	0	0	0	0
ENSMUSG00000001029	0	0	0	0
ENSMUSG00000001034	8	0	3	4
ENSMUSG00000001036	55	31	38	61
ENSMUSG00000001039	2	0	0	0
ENSMUSG00000001052	19	6	9	10
ENSMUSG00000001053	0	0	0	0
ENSMUSG00000001054	0	0	0	0
ENSMUSG00000001056	39	16	10	27
ENSMUSG00000001062	0	0	0	0
ENSMUSG00000001065	1	0	0	0

How to represent the results?

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**

Grazie

TELETHON INSTITUTE OF GENETICS AND MEDICINE

Tigem