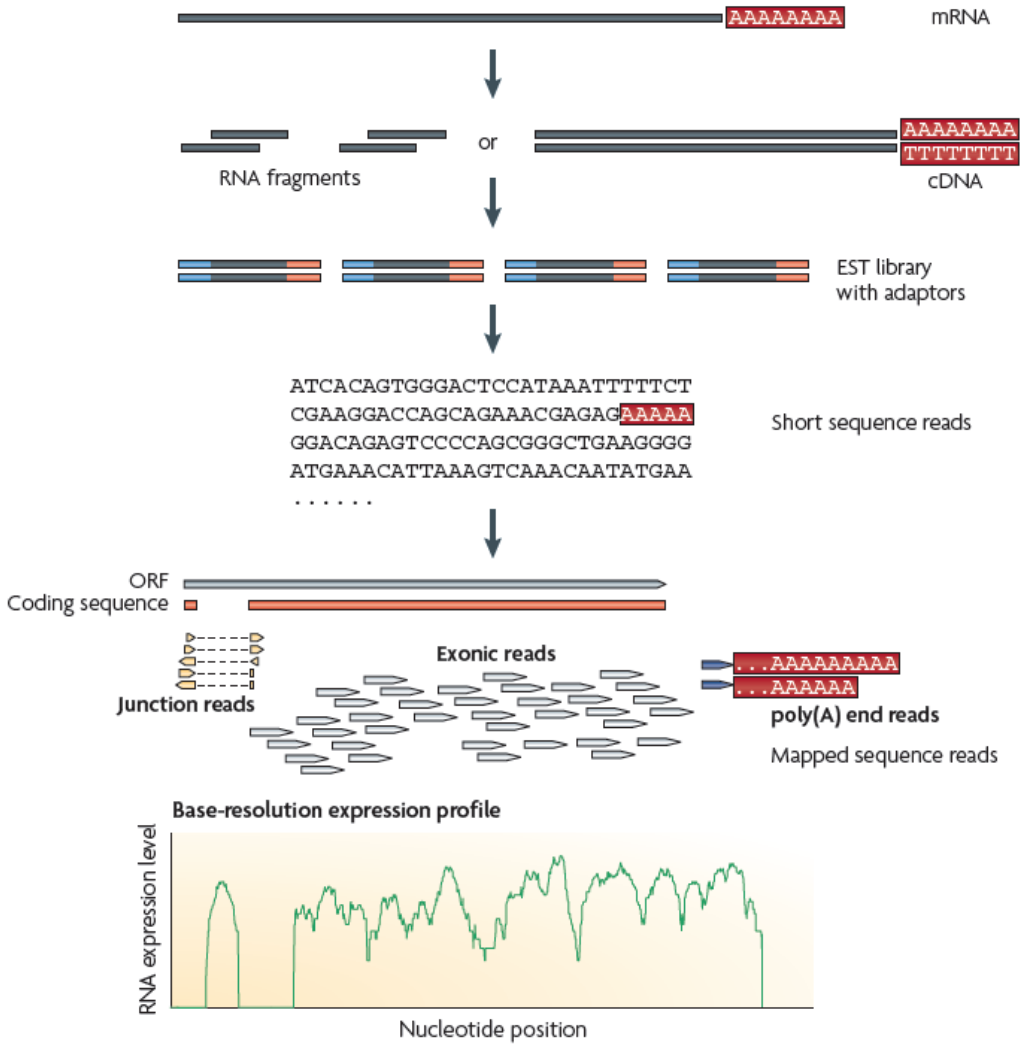
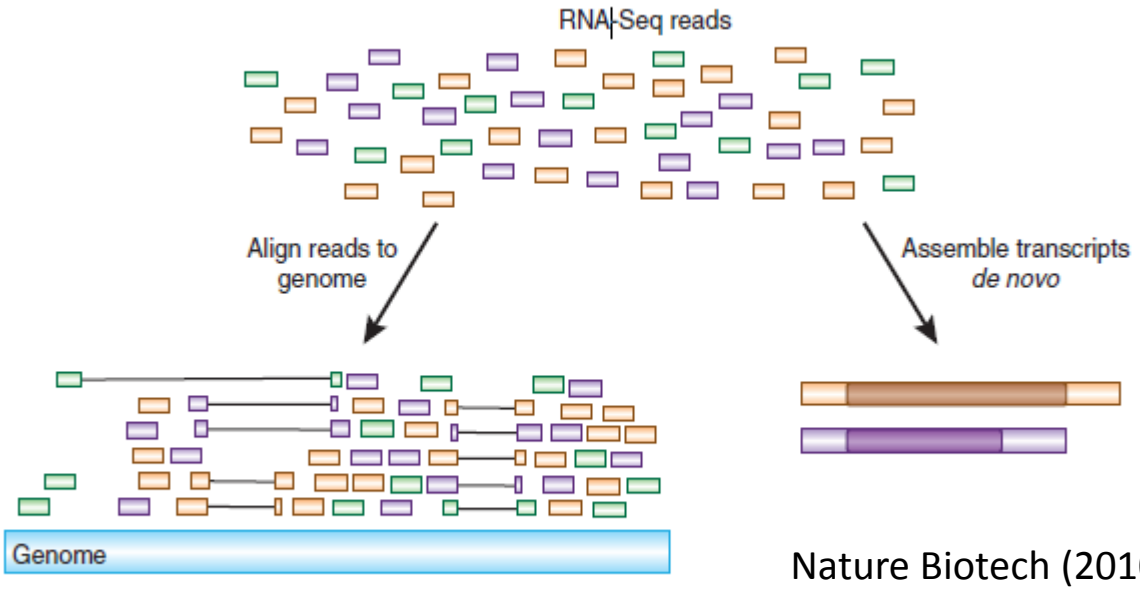
Use Case of a RNA-Seq assembly

IPS2 Plant Institute of Paris Saclay
Genomic Networks
V. Brunaud

RNASeq technology



RNASeq technology

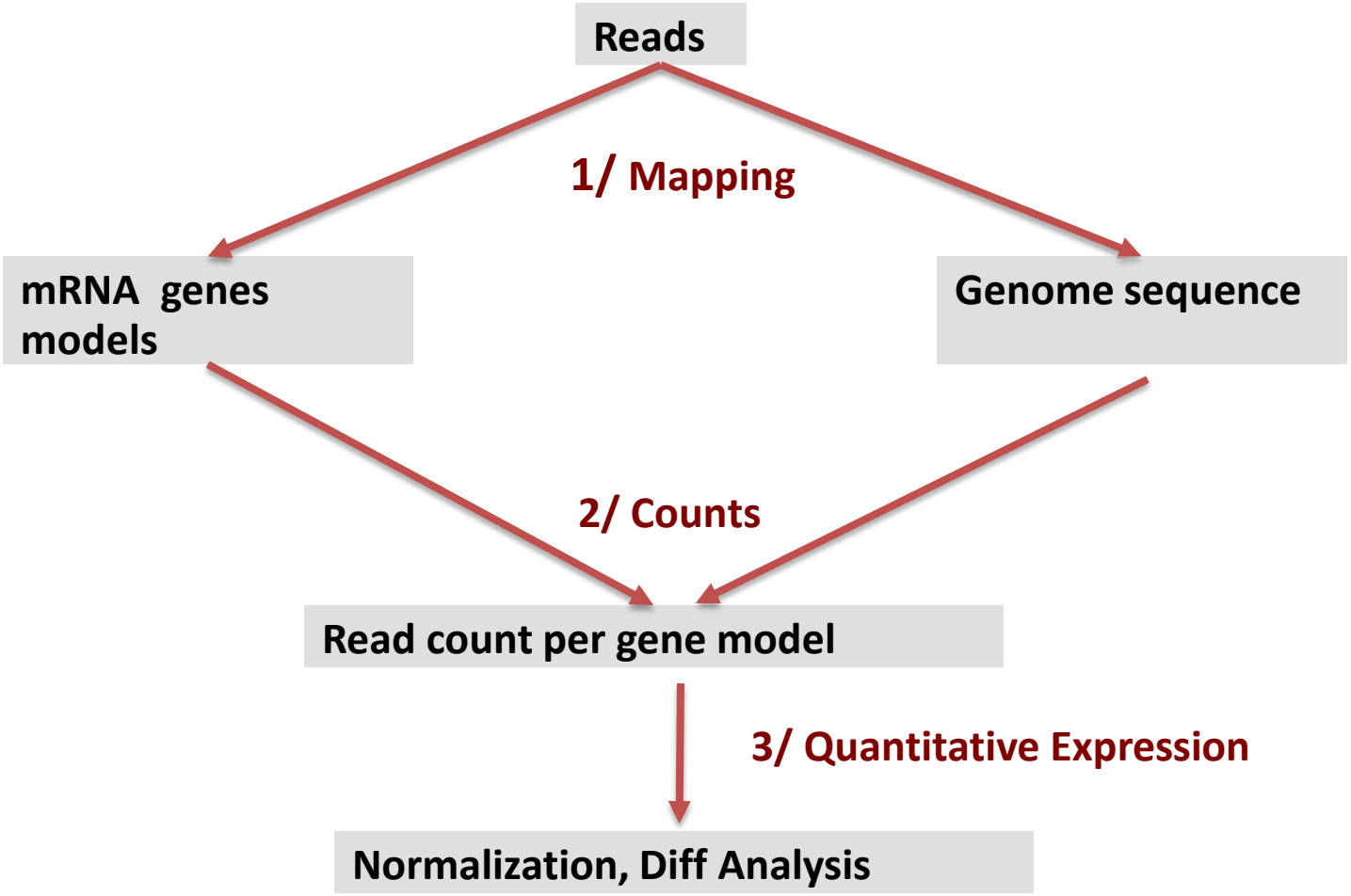


Nature Biotech (2010)

Many Applications :

- quantification,
- detection de novo “new gene, new transcript”
- Meta-transcriptomic (TARA Oceans project Genoscope)

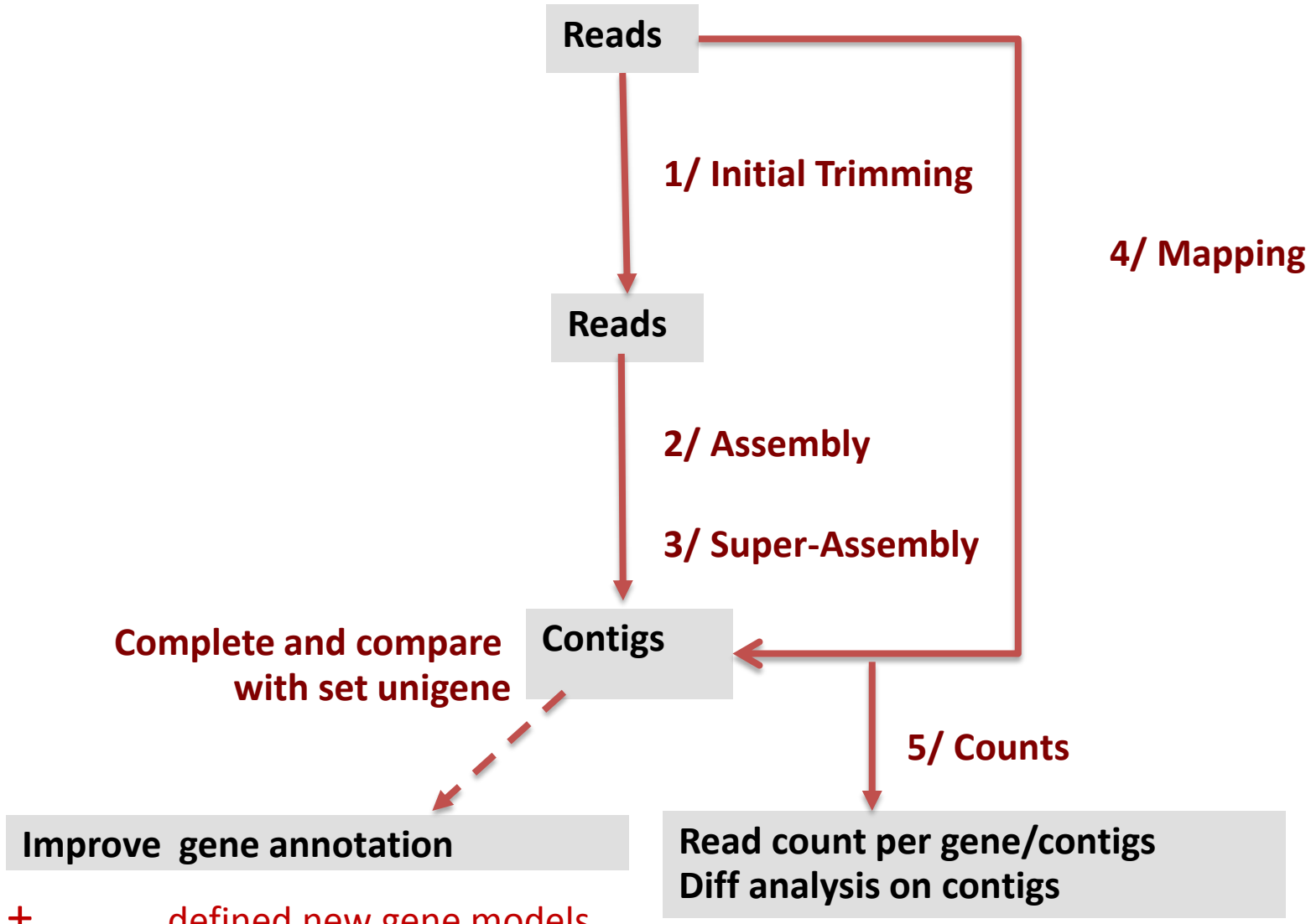
1st strategy : mapping RNA-Seq against a genome (transcripts or genome)



\+ trimming not necessary, time saving

\- confidence of gene models or assembly genome, no new genes detected

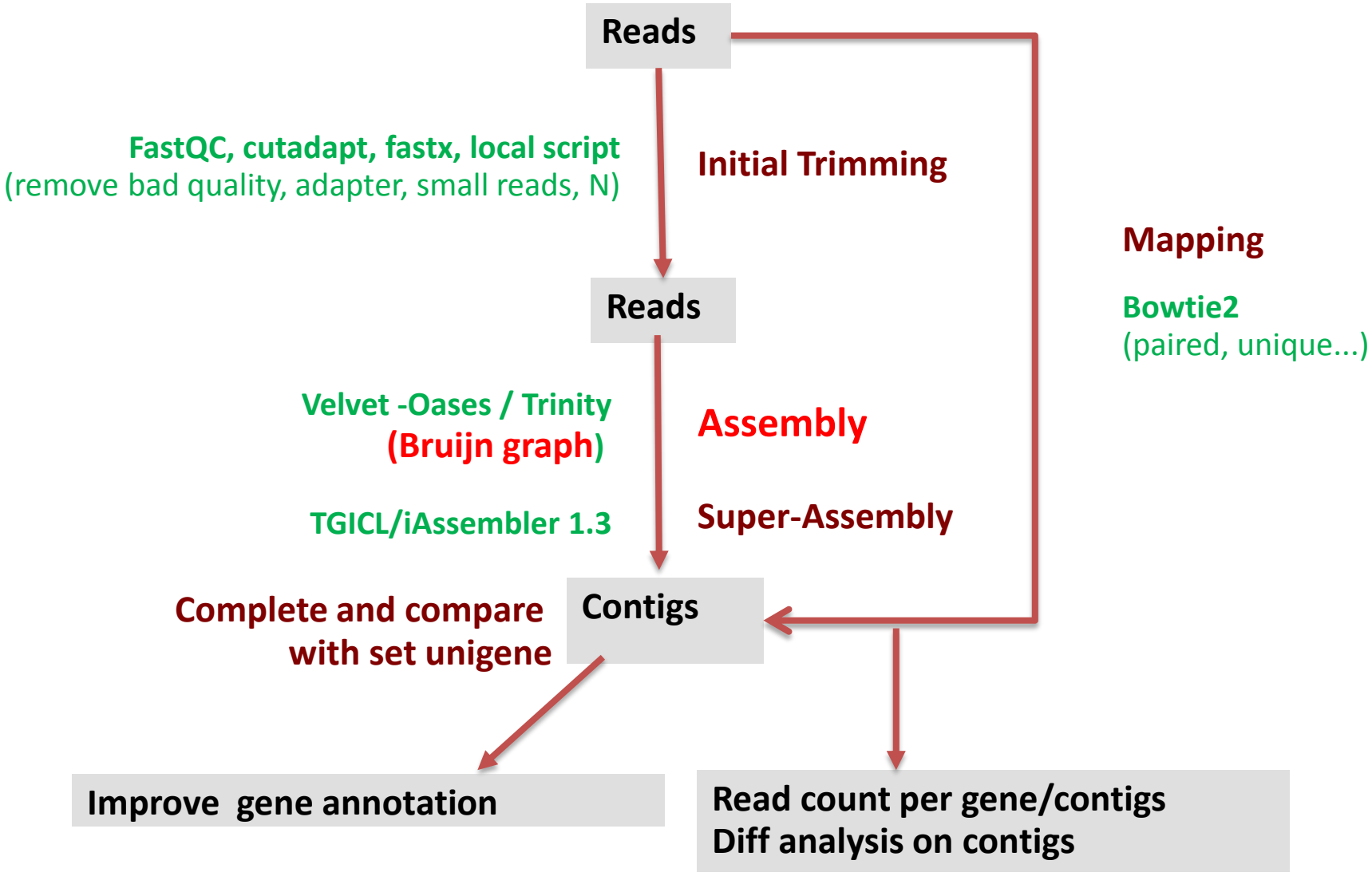
2nd strategy : de novo Assembly of RNAseq (without reference genome)



\+ defined new gene models

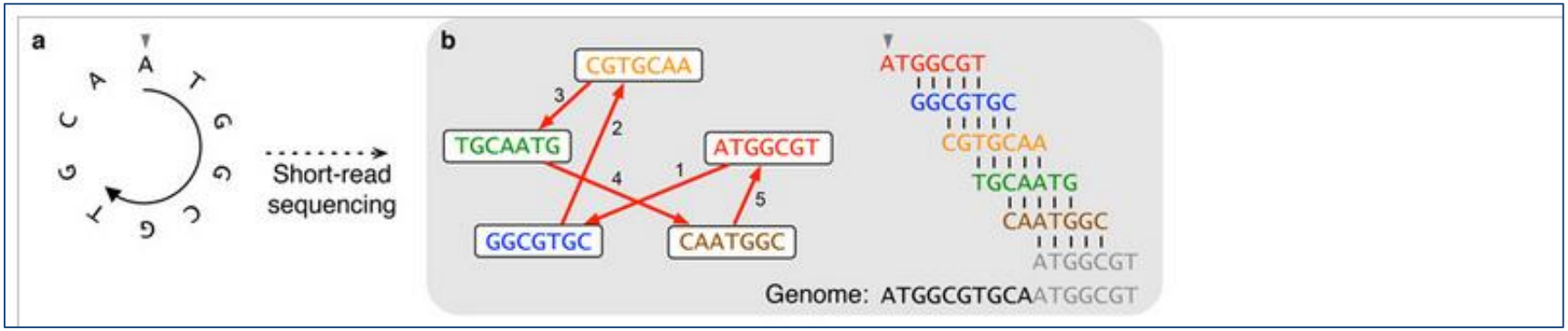
\- Assembly: not perfect, defined kmer, time and memory consuming

de novo Assembly of RNAseq

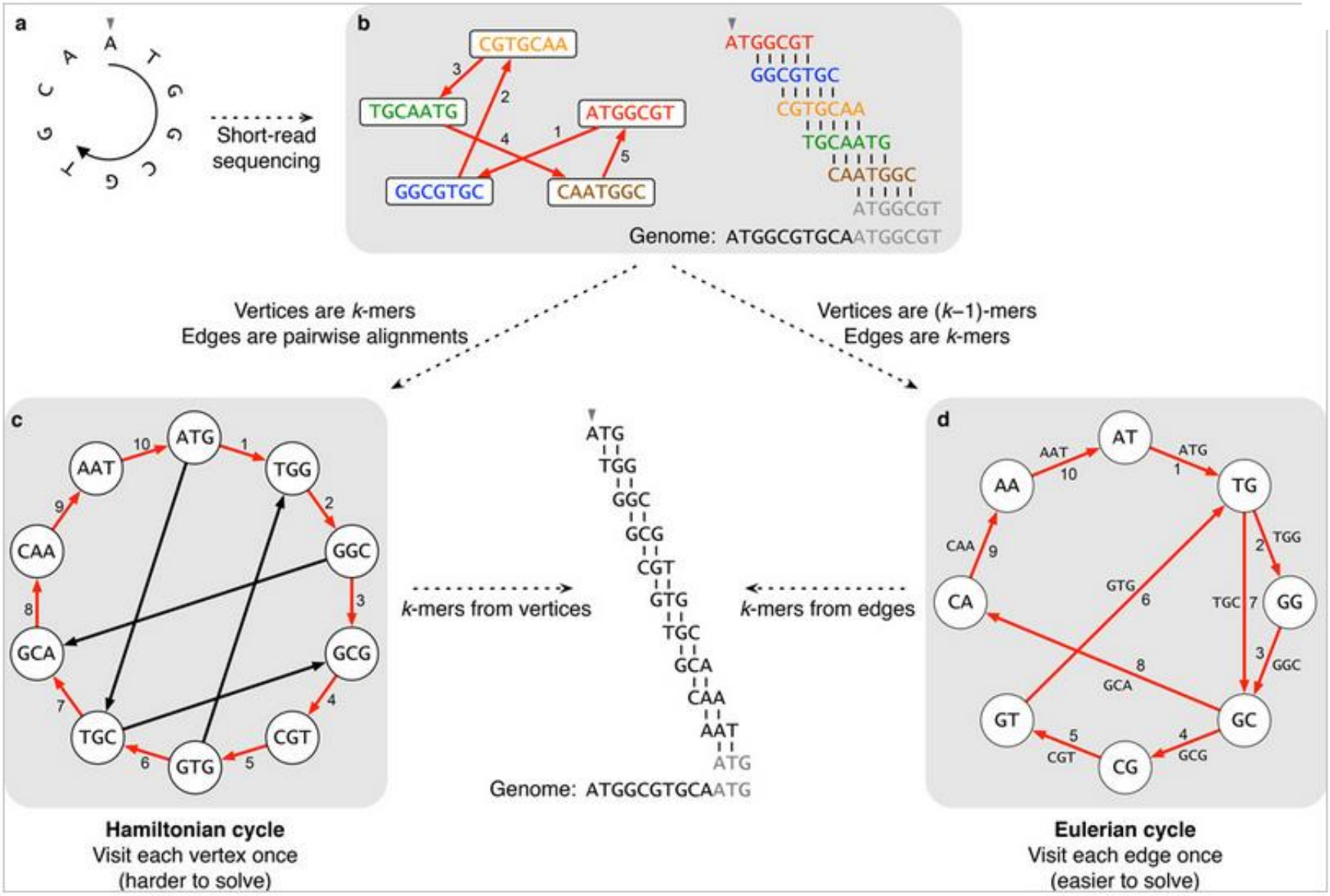


From reads assembly to contigs → Graph construction methods

Why are de Bruijn graphs useful for genome assembly? Nat Biotechnol. Compeau et al. 2017



Why are de Bruijn graphs useful for genome assembly? Nat Biotechnol. Compeau et al. 2017



Two strategies for genome assembly: from Hamiltonian cycles to Eulerian cycles

De novo Assembly : 2 methods for Graph path

Hamiltonian: Node= kmer or read; Edge= pairwise between node

- Aligning pairwise = this step is very time/memory consuming
- Path = Cross by all nodes once a time

→ **Overlap Layout Consensus (OLC method)**

→ **Tools : newbler, cap3 adapted for 454 sequencing…not adapted to 2nd generation of sequencer**

Eulerian: Node = (k-1) mer; Edge = kmer

- Easy to contsruct
- Path = Cross by all edges once a time

→ **Graph de Bruijn method**

→ **Adapted to last NGS with 10 to 40 million of reads (50 to 150 bases)**

→ **using by all popular tools : velvet, Trinity, SOAPdenovo**

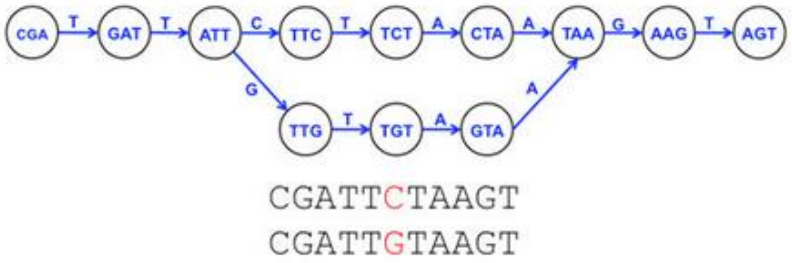
Bruijn Graph :

"Instead of assigning each k-mer to a node, we will now assign each k-mer located within a read to an edge."

Limits of using Assembly

Time consuming:
for 40 millions of reads with the most popular tool Trinity
=index kmer 27bases + graph +contigs.
→ 23h to obtain contigs , Graph construction 6h

Bubbles=complexities in graph

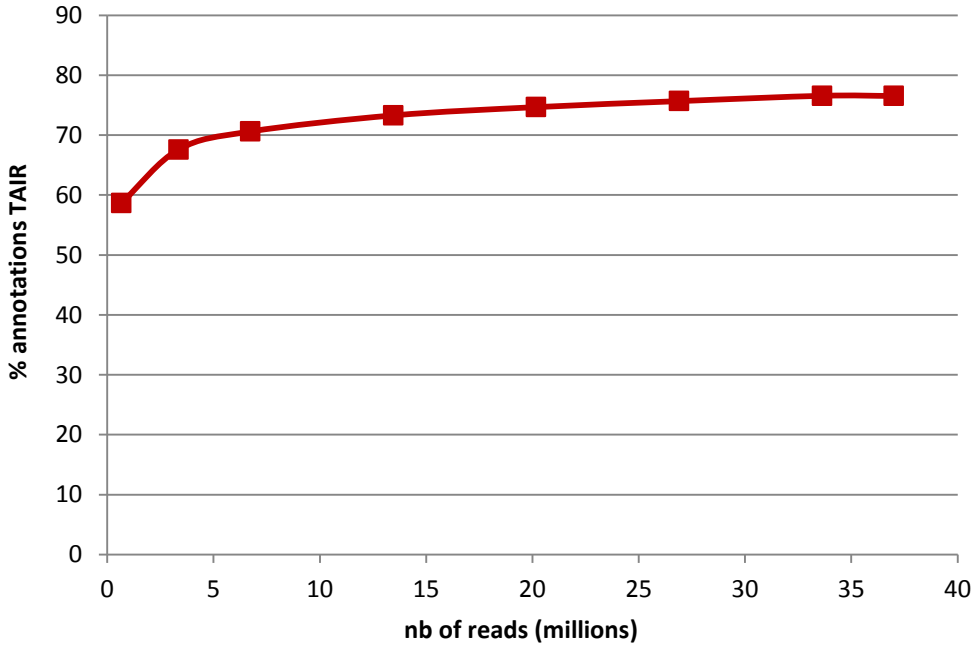


→ **Biological reasons**

- Errors in read sequences : tools exist to remove some errors
- Natural genotype difference : allelic, polyploidy
- Repetitions : many repeat element in eukaryote
- RNA-Seq :
  - No same coverage by gene
  - Alternative transcripts (splicing alternative)
  - Many graph by gene

Problem of depth sequencing : counts of reads → coverage by gene

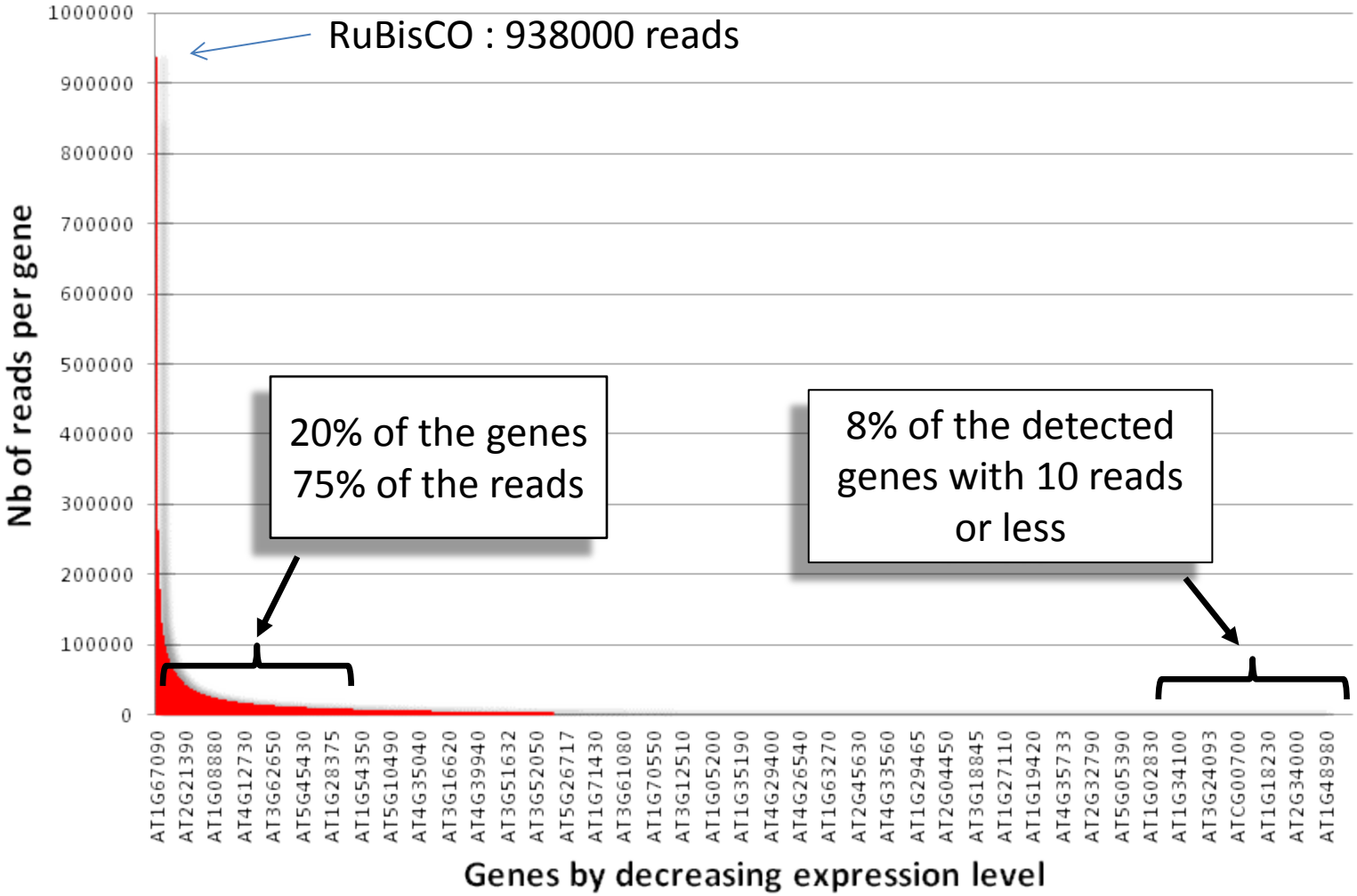
→ Bowtie2 mapping : 98% of reads mapped



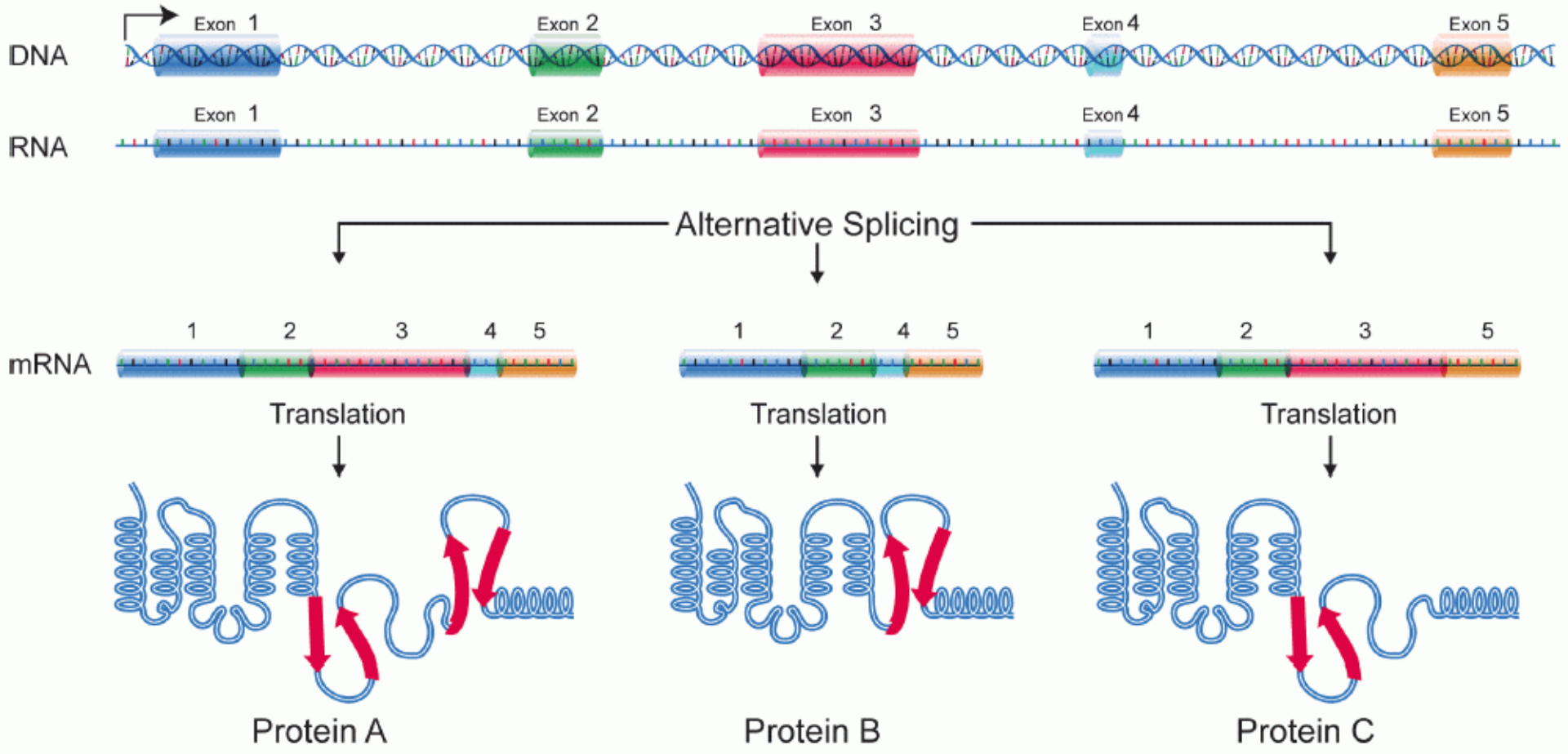
Among 27884 genes on nuclear chromosomes

1m	--> **16507 genes (59%)**
5m	--> 19041
10m	--> 19926
20m	--> 20748
30m	--> 21177
40m	--> 21487
50m	--> 21733
55m	--> **21739 (78%)**

Problem of non uniform gene expression
→Biased distribution of the reads by gene



Problem of Isoforms : Alternative splicing



Wikipedia source

Arabidopsis : 30% of genes with 2 to 3 mRNA
Human : 85% of genes encode isoform proteins

Assembly Results on Arabidopsis RNA-Seq

F1_Mplex	
Nb of PE reads	43 030 388 PE
Nb of contigs from assembly	33 736 (length mean 1360)
Nb of mapped contigs Genome TAIR10	33 072 98%

Data from Illumina HiSeq2000

- **Velvet/oases (kmer 61,71)**
- **iAssembler**

Comparison of annotations
TAIR10 annotation versus Contigs from assembly

F1_Mplex	
Nb of PE reads	43 030 388 PE
Nb of contigs	33 736 (length mean 1360)
Nb of mapped contigs Genome TAIR10	33 072 98%
Comparaison annotations	
Nb tagged Genes Nb contigs	17 783 32 220 (97%)
Nb of genes with confirmed structure Nb Contigs	15 613 (88%) 20 881 (65%) contigs

Data from Illumina HiSeq2000

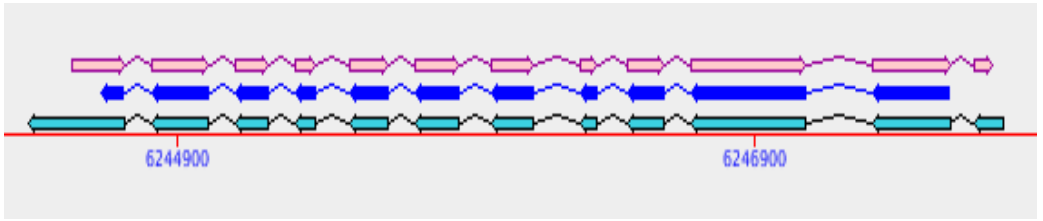
- **Velvet/oases (kmer 61,71)**
- **iAssembler**

→Gene=Locus

→ Model of genes with confirmed exon/intron structure

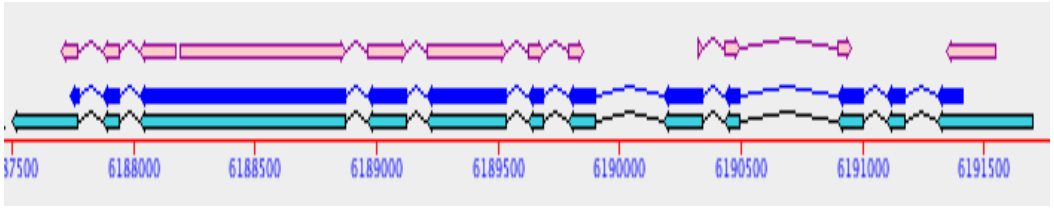
Quality of Assembly : contig versus gene annotation

1 gene – 1 contig
same gene model

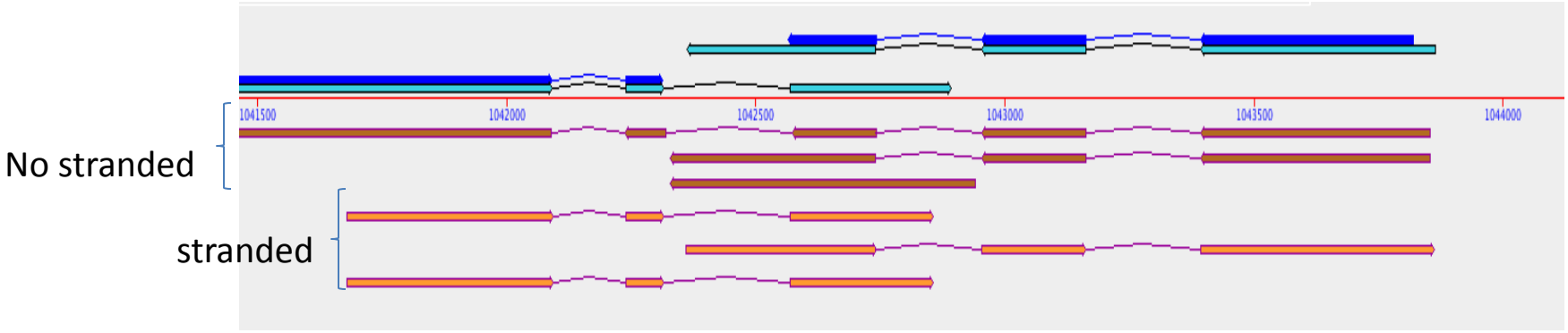


2/3 of contigs correspond to gene structure (partial)

1 gene – 2 or n contigs
same gene model



Library Stranded → remove chimeric genes



Quality of Assembly

35% of contigs with other gene models (isoforms)

1 gene – 1 or n contigs
with other gene models

→3% of contigs without annotation = new genes

Conclusion of assembly on transcriptomic and simple model

→ A good quality of contigs, efficient to detect new gene models

→ Problems: distinct false/good gene models,
chimera that increases with read number

With more complex transcriptome/genome

→ Can not work: too complex graph path

→ Generate too much contigs