Module 5/5: Analyses RNA-seq

RNA-seq bioinfo analysis

13 & 14 Oct. 2022

Camille Marchet - Pierre Pericard

Jour 2 & 3 : RNA-seq bioinfo analysis

Jour 2

- Cours: with reference RNA-seq
  - RNA-seq QC + Cleaning
  - Mapping on reference
  - Assembly with reference
  - Quantifying gene expression
- TP: with reference and de-novo RNAseq (Jour 2 & 3)

Jour 3

- Cours: *de-novo* RNA-seq
  - de-novo assembly
  - local assembly for variant calling
- TP: with reference and de-novo RNAseq (Jour 2 & 3)
- Cours: Introduction to long-reads RNA-seq
- Cours (15h-17h): Introduction to RNA-seq biostatistics analysis

General Introduction

Goals

This course main goals:

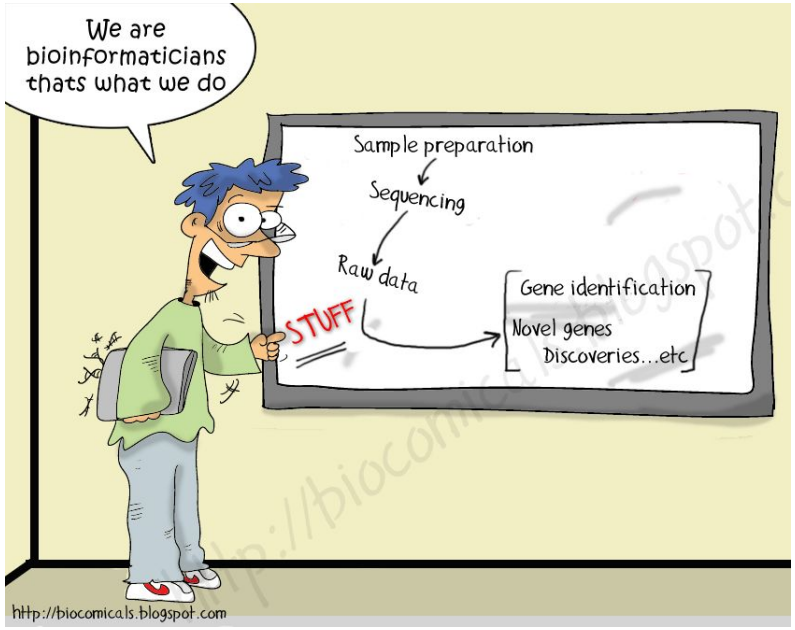
- An overview of RNA-seq data analysis

- Identify the (key issues/points) (critical steps/parameters)

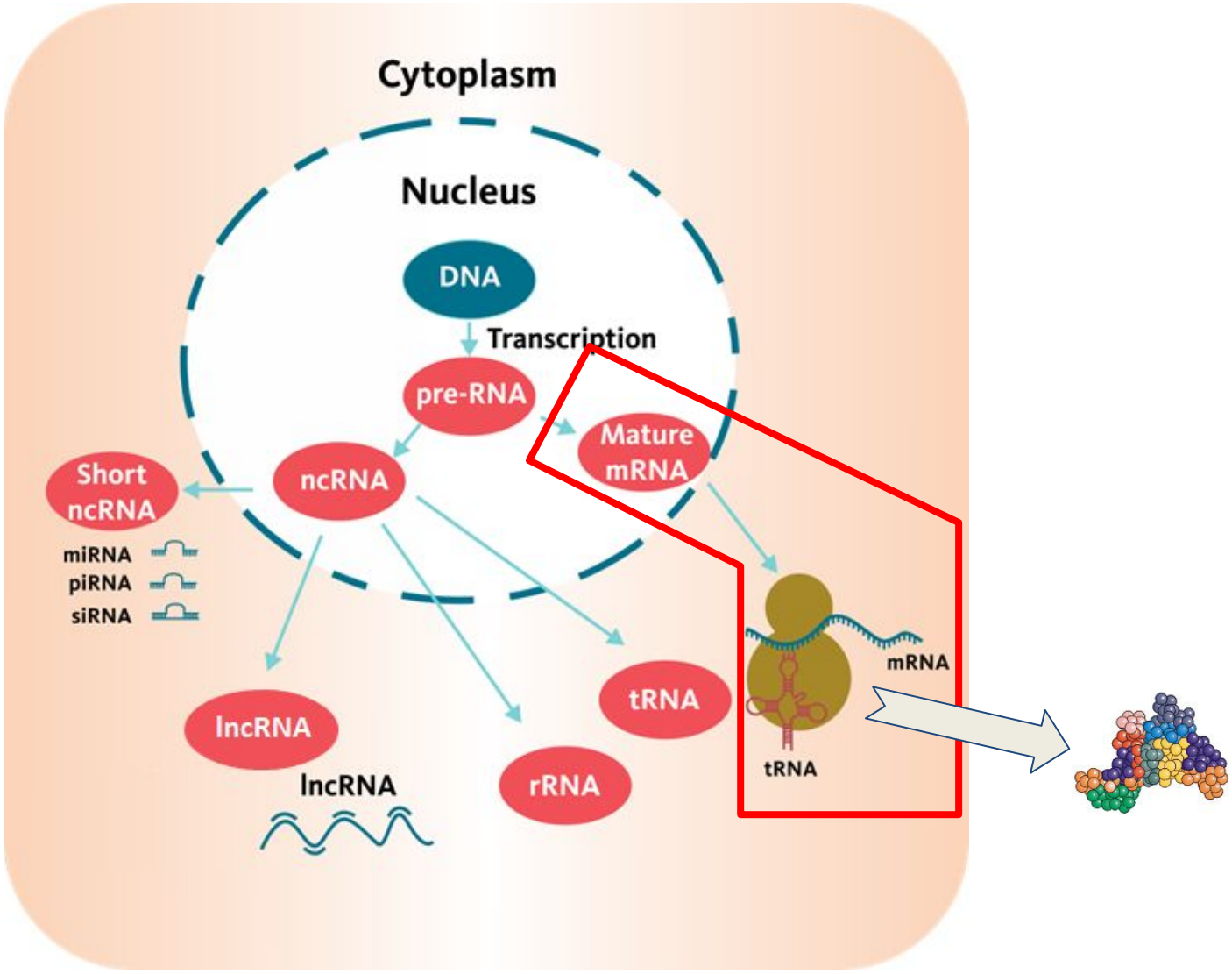
Warning !

This is NOT a course to train you as a bioinformatician, and this course will NOT allow you to design an analysis pipeline set-up for your specific needs

This course WILL give you the basis information to understand and run a generic RNA-seq analysis, its key steps and problematics, and how to interact with bioinformaticians/bioanalysts that can analyze your RNA-seq datasets

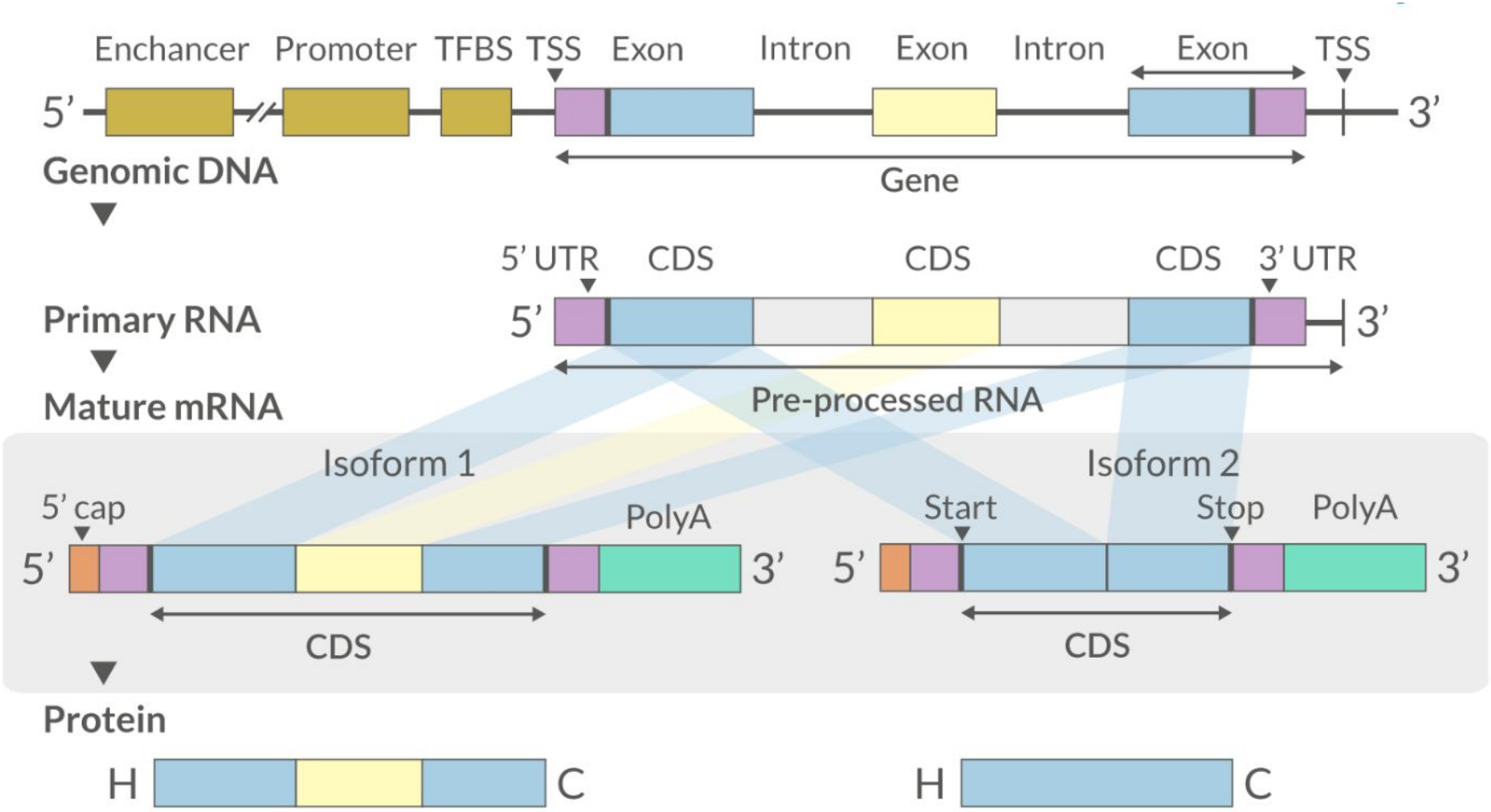


Preliminary



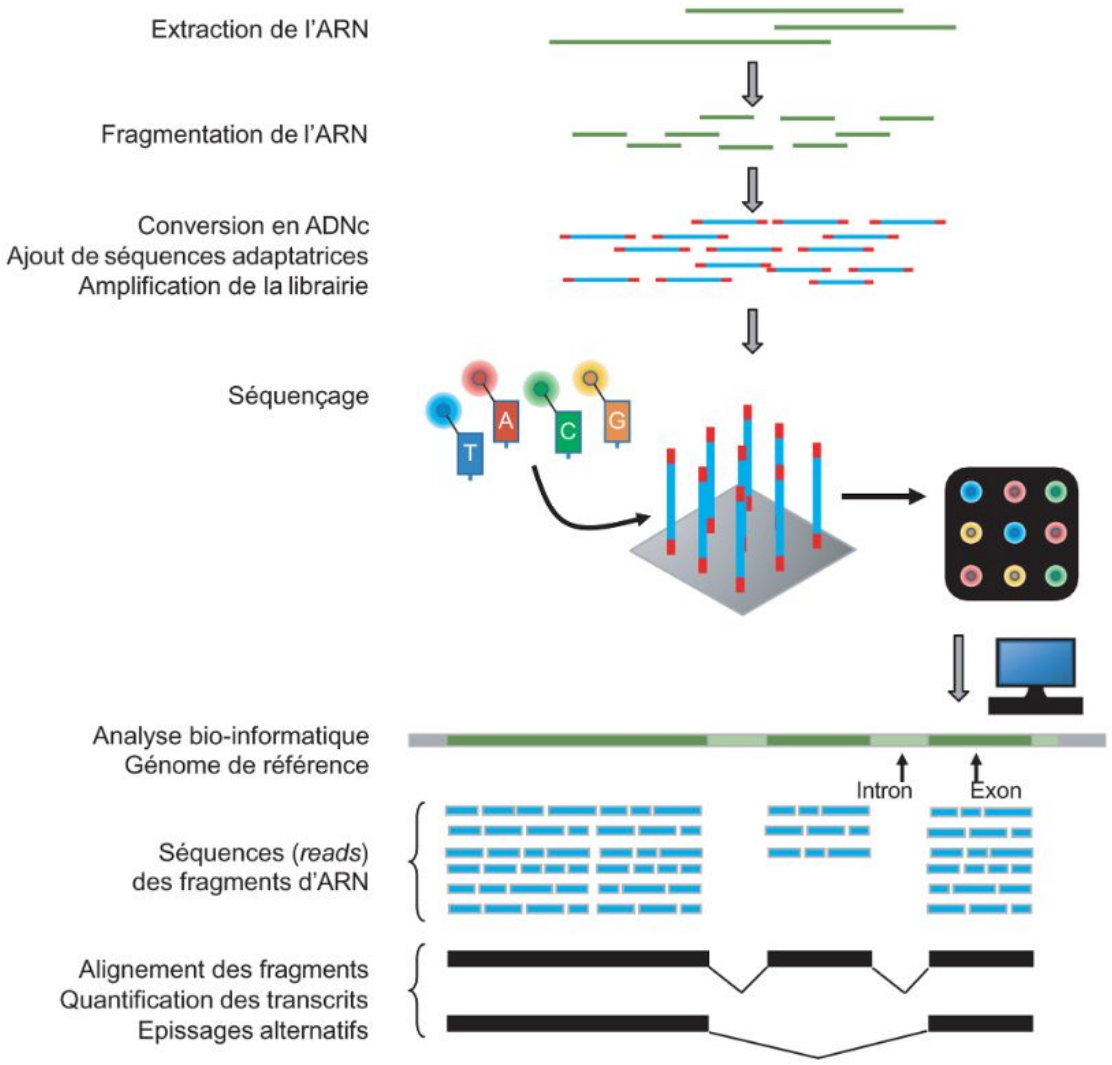
source: prez journée RNA-seq Go@L / Bilille - David Hot

Maturation and variability of RNA



https://scilifelab.github.io/courses/ngsintro/1905/slides/rnaseq/presentation.html#3

Sequencing overview



source: prez journée RNA-seq Go@L / Bilille - David Hot

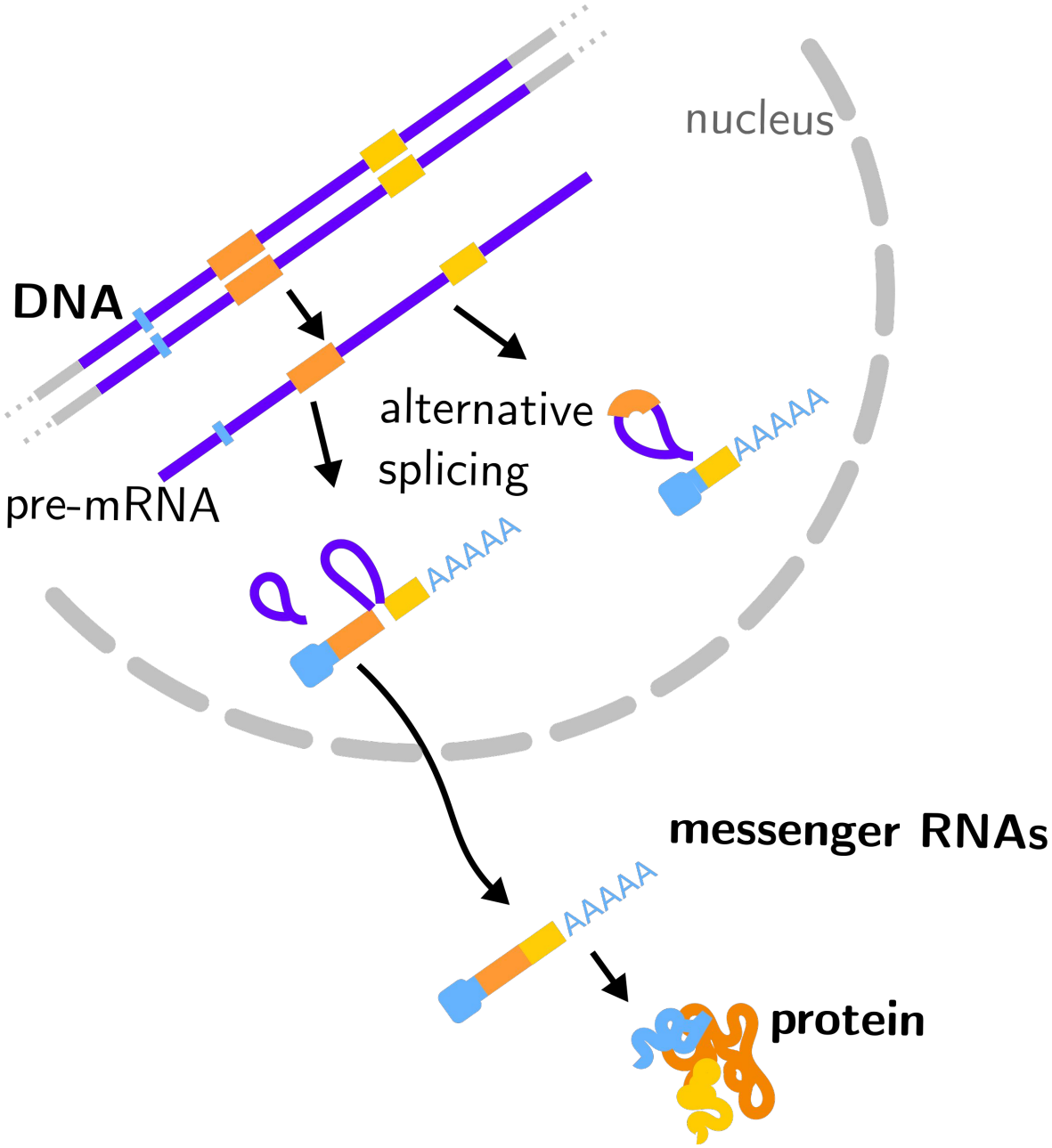
Preliminary

Transcriptome/transcript

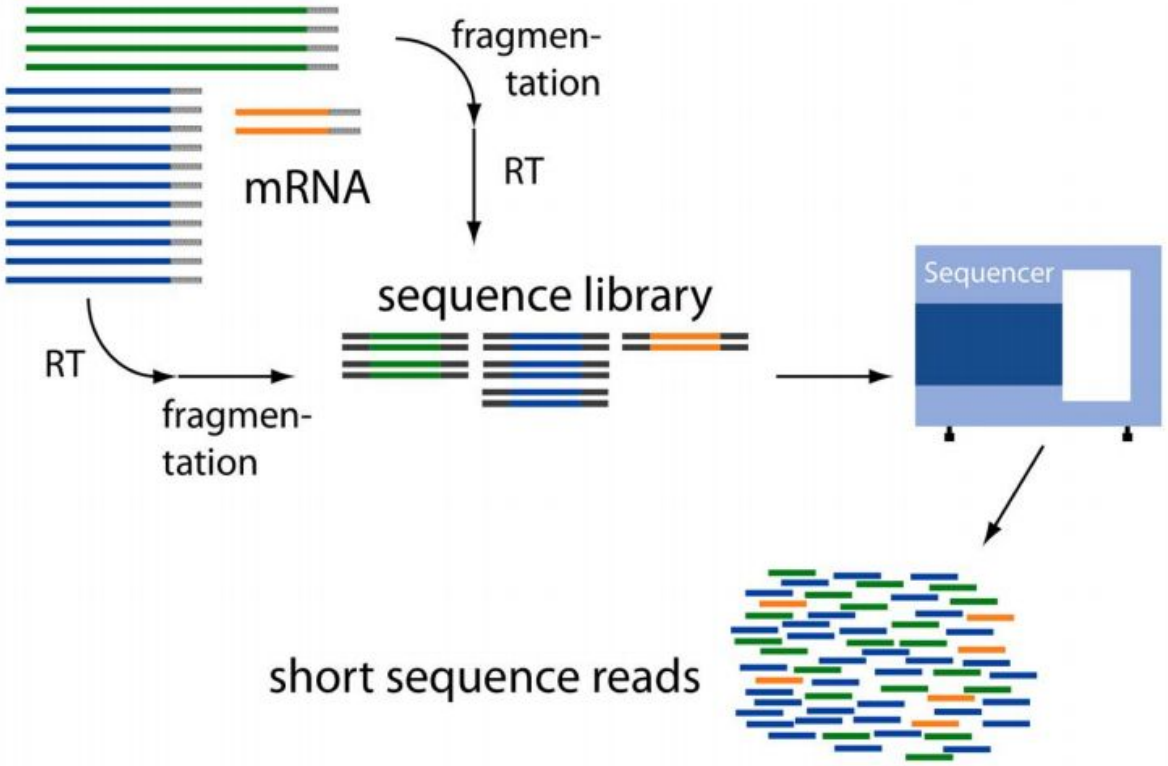
Transcriptomics

(Alternative) isoform

Splicing



Sequencing: overview



From: http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html

How to make cDNA libraries

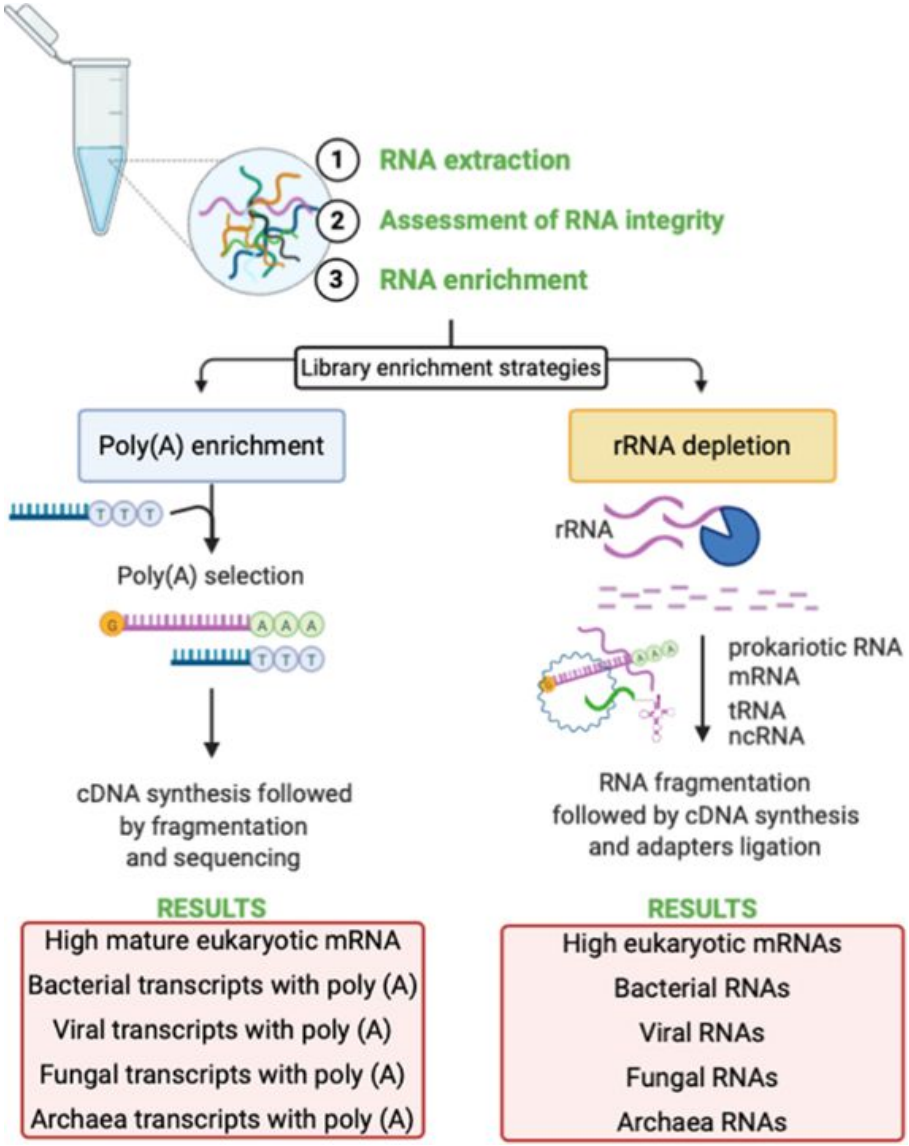
- Extract RNA, convert to cDNA
- pass to next gen sequencer
- millions to billions of reads

make cDNA?

- Prime mRNA with random hexamers R6
- reverse transcriptase => cDNA first strand synthesis
- then second strand

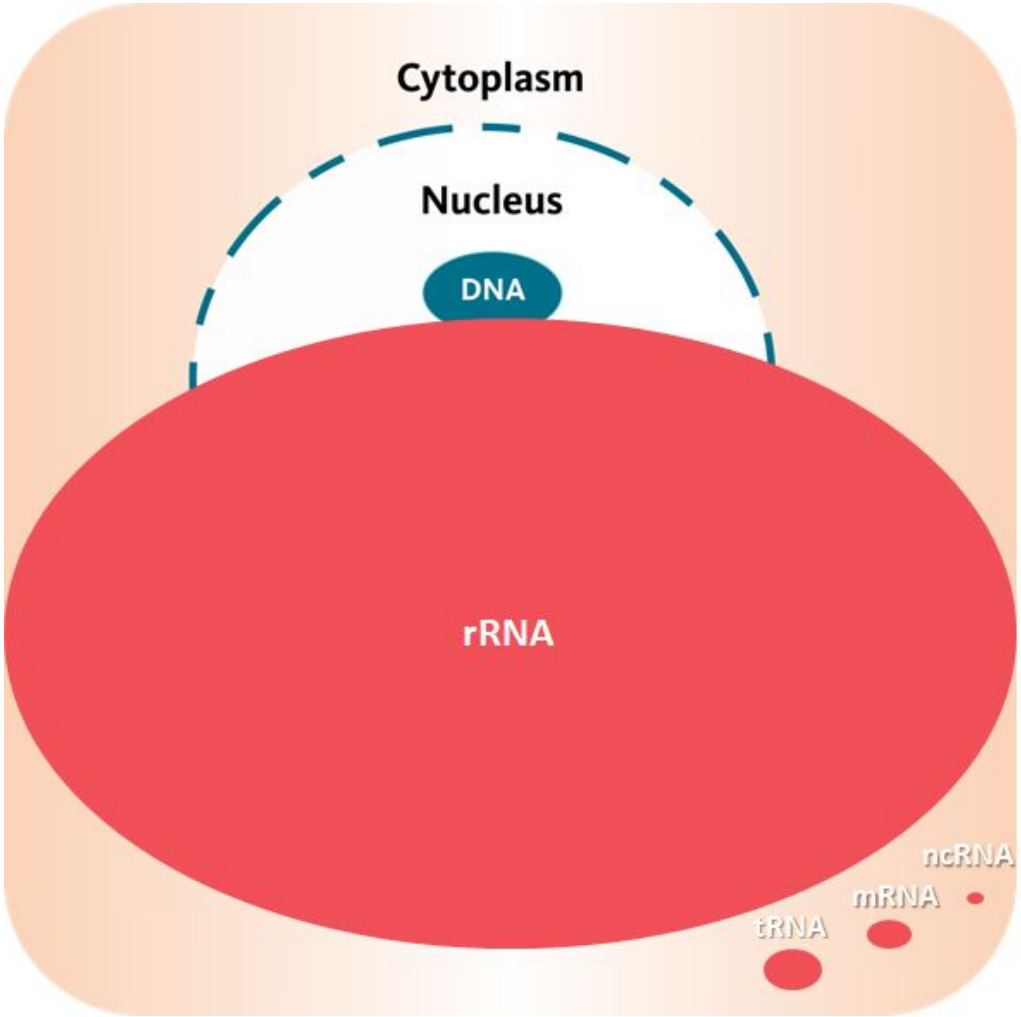
=> illumina cDNA library

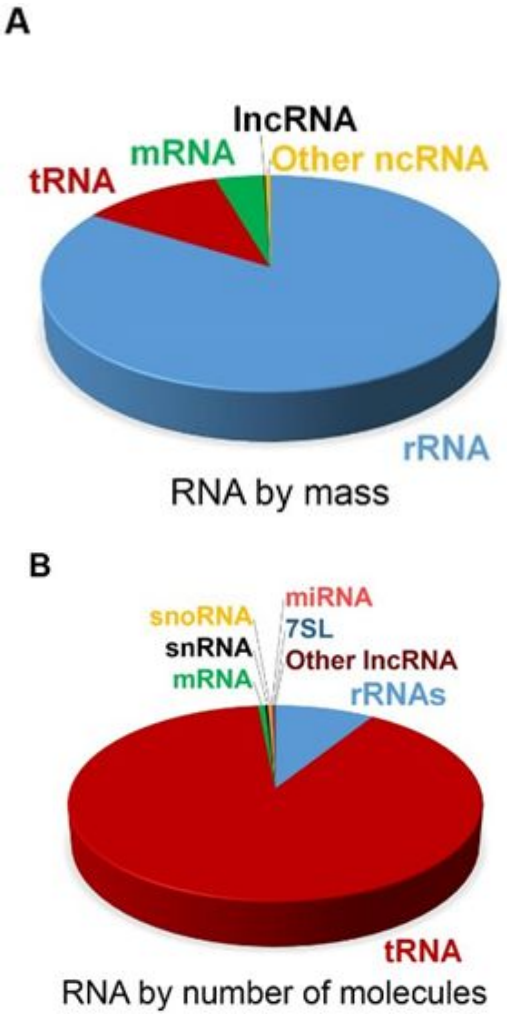
How to make cDNA libraries



source: journée RNA-seq Go@L / Bilille 2022 - David Hot

Do you want *all* RNAs?





source: prez journée RNA-seq Go@L / Bilille - David Hot

How to sequence (1)

RNA selection

- polyA+
- Ribo-Zero (human, mouse, plants, bacteria, ...)

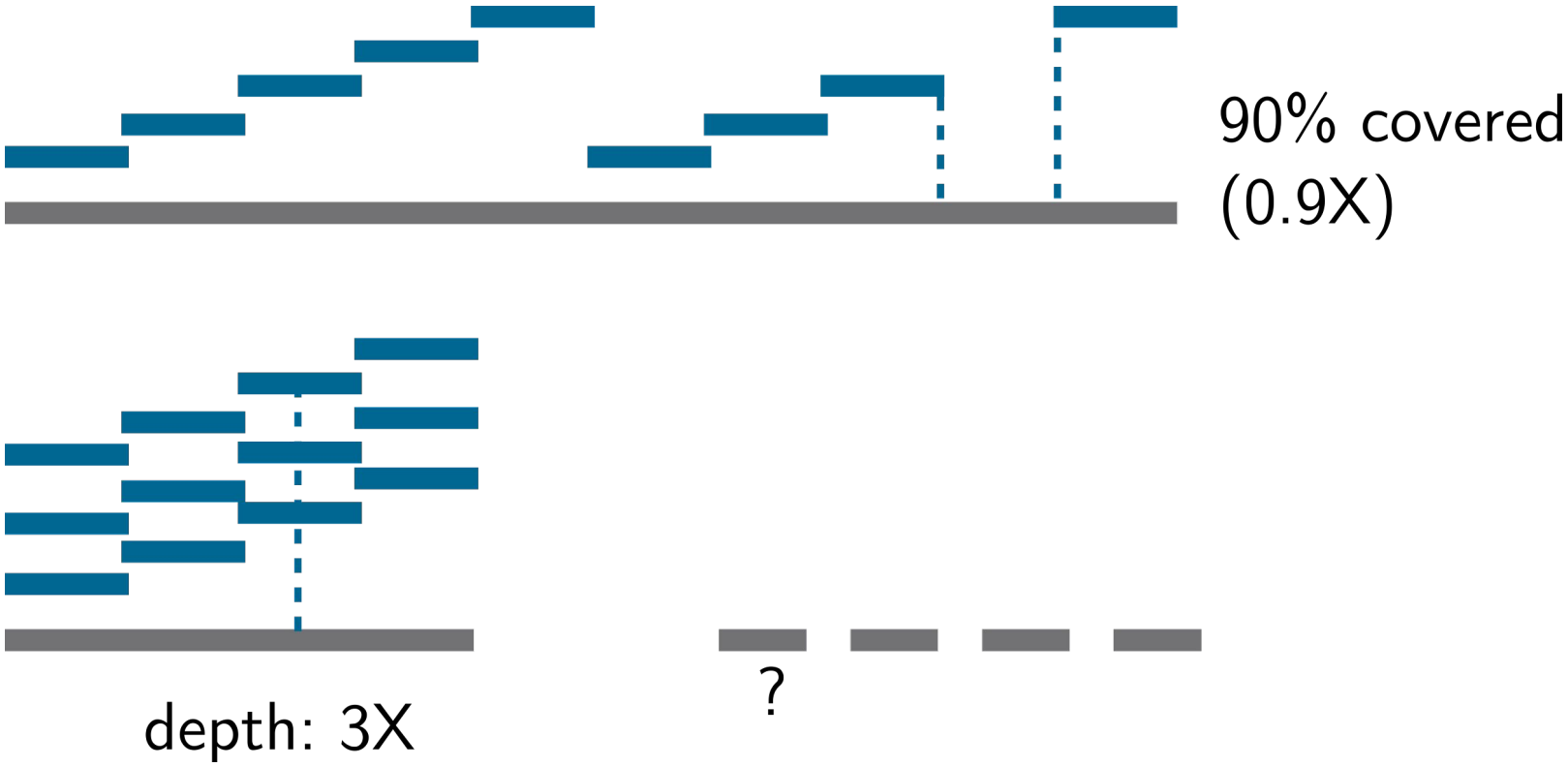
  (ARN = 90% of ARNr, 1-2% of ARNm)

- in prokaryotes: no polyA (= no capture), no splicing (= less complex)

Types of reads, experimental design

- paired-end
- replicates

How to sequence (2)



RNA-seq

- reads around **150-200** bp
- the number of **detected transcripts increases with the sequencing depth**
- the **expression** measure is **more precise with more depth**
- 5 million reads can be enough to detect genes mildly-highly expressed in human
- 100 million must be preferred to detect lowly expressed genes (see for instance **saturation curves** in “Differential expression in RNA-seq: a matter of depth.” *Genome Res.* 2011)
- these numbers depends on the species/tissues (complex splicing…)
- keep **replicates** in mind

There are plenty of protocols...

Méthode	Description	Référence
mRNA-seq	Identification les ARN messagers.	[Mortazavi et al., 2008]
miRNA-seq	Identification les micro ARN.	[Ruby et al., 2006]
GRO-Seq (Global Run-On Sequencing), PRO-Seq (Precision Run-On Sequencing) et NET-Seq (Native elongation transcript sequencing)	Sélection et séquençage uniquement les ARNs en cours de transcription par l'ARN polymérase II.	[Core et al., 2008] [Kwak et al., 2013] [Churchman and Weissman, 2011]
Ribo-Seq (Ribosome profile sequencing) et TRAP-Seq (Targeted purification of polysomal mRNA sequencing)	Identification les ARNs messagers en cours de traduction.	[Ingolia et al., 2009] [Reynoso et al., 2015]
RIP-Seq (RNA immunoprecipition sequencing), CLIP-Seq (Cross-linking and immunoprecipitation sequencing), PAR-CLIP (Photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation) et iCLIP (individual-nucleotide resolution CLIP)	Détermination des régions d'ARN liées à une protéine d'intérêt.	[Cloonan et al., 2008] [Chi et al., 2009] [Hafner et al., 2010] [Huppertz et al., 2014]
ChIRP-Seq (Chromatine isolation by RNA purification)	Identification des régions du génome qui interagissent avec l'ARN.	[Chu et al., 2011]
PARE-Seq (Parallel analysis RNA ends sequencing)	Etude des sites de clivage des micro-ARNs ainsi que de la dégradation des ARNs.	[German et al., 2009]

from Clara Benoit Pilven's PhD thesis

Resources: genomes, transcriptomes, annotations

Common databases

Ensembl

UCSC

National Center for Biotechnology Information
NCBI

Specific databases

FlyBase

EcoCyc

MicroScope

GENCODE

SGD Saccharomyces GENOME DATABASE

AmoebaDB
Amoeba Genomics Resource

VectorBase
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

From Rachel Legendre (Institut Pasteur)

Sequencing reads file formats

FastQ

1. Identifier
2. Sequence
4. Quality scores *(as ASCII chars)*

READ

```
@SRR062641.6751359
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT
+
CBLNPGJQQQJPPQPPQPQRGPPPPRRQQRPSPGRQQQRLRRRMEPQQPMJHQQEHKMMFIIRH?SIIHKNJIKRLJJIKHEABHIFGCGGEFCGDGDCE
```

```
@SRR062634.16249693
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC
+
ALKMOOOOPPQJQOPPPPPQPPPPPPPRJQRQQQQQRPQPRQQPFQSQQPRLIMHKSNRJQORMFELRPQNQRQJQRRPQQLIRKDMKQJPN8CFDGCCCB

@SRR062634.20060465
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAGTGTTGGCCTGATACCTACCAGTGGAGCGAGGGGAACCCGAGGACTGCCAAGGGCA
+
D?KMPQEPGCPQQNPQIQIGR@DPERQHEKBED=HCHG8EHFDCD6<329@<:69A<6,;<967>;=C:>AA8BBED######################
```

FastA

```
>SRR062641.6751359
CGCCCGGCCAATCATTGTGGTTTTAAGTCACTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCT

>SRR062634.16249693
CTAAGTTTGAGGCTATTTTGTTTTACAGCAAAAGCTAACTGATGCAGACAGGGACAAGTCAGTCTCATCTCTGTGCACCCAGCATTGCCCAGAACAGGGC

>SRR062634.20060465
CTCCCAGCTTCCAACAGACCCTGTCCCAGCTCCCTCCAAGCTGAGTGTTGGCCTGATACCTACCAGTGGAGCGAGGGGAACCCGAGGACTGCCAAGGGCA
```

Mais aussi: FAST5, BAM, ...

FASTQ format: Quality

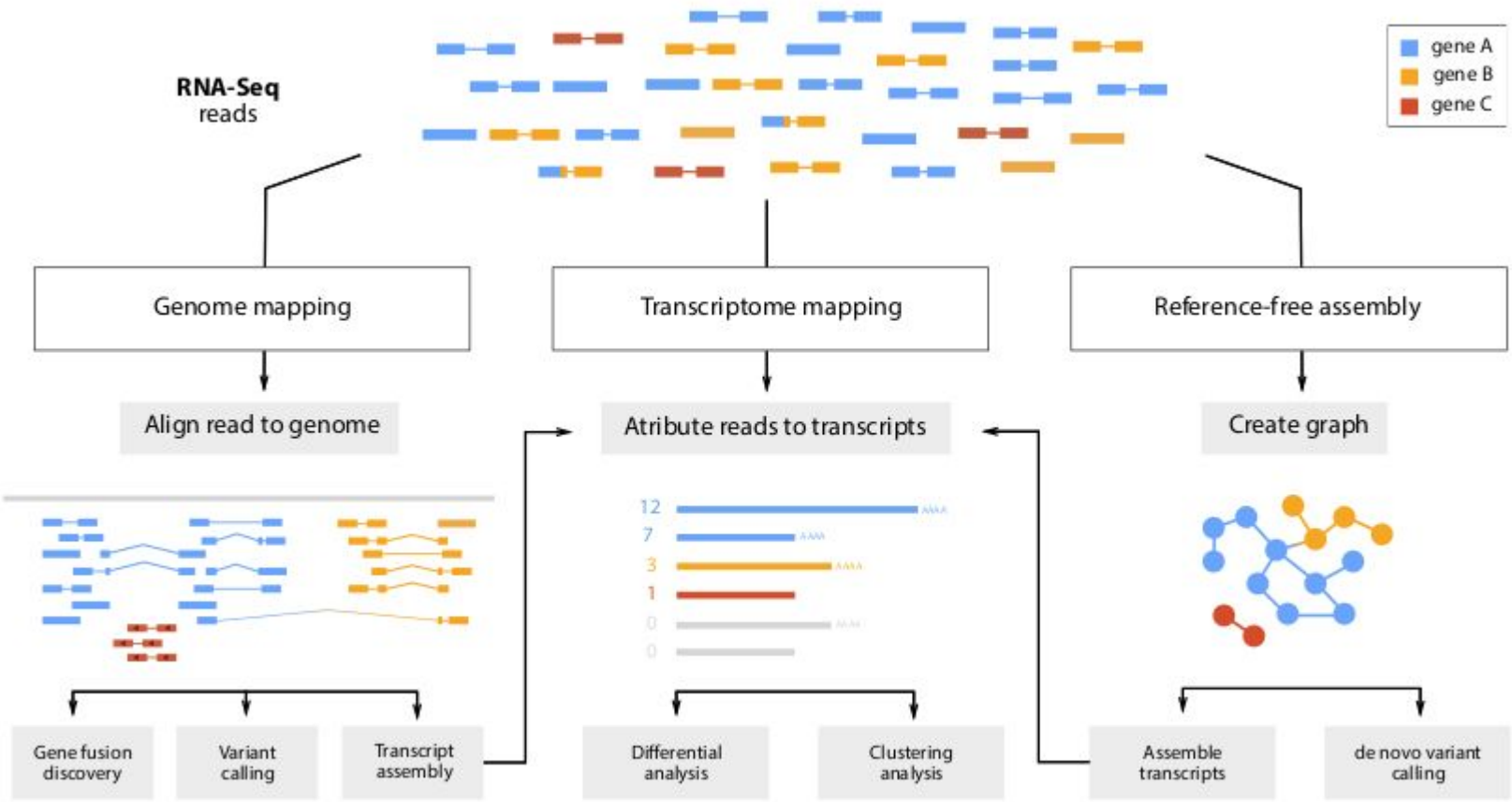
```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 .........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 .............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 .............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                          |    |       |                                  |                     |
33                         59   64      73                                 104                   126
  0........................26...31.......40
                           -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................41
  0.2......................26...31........41

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, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

$$Q_{\text{sanger}} = -10 \log_{10} p$$

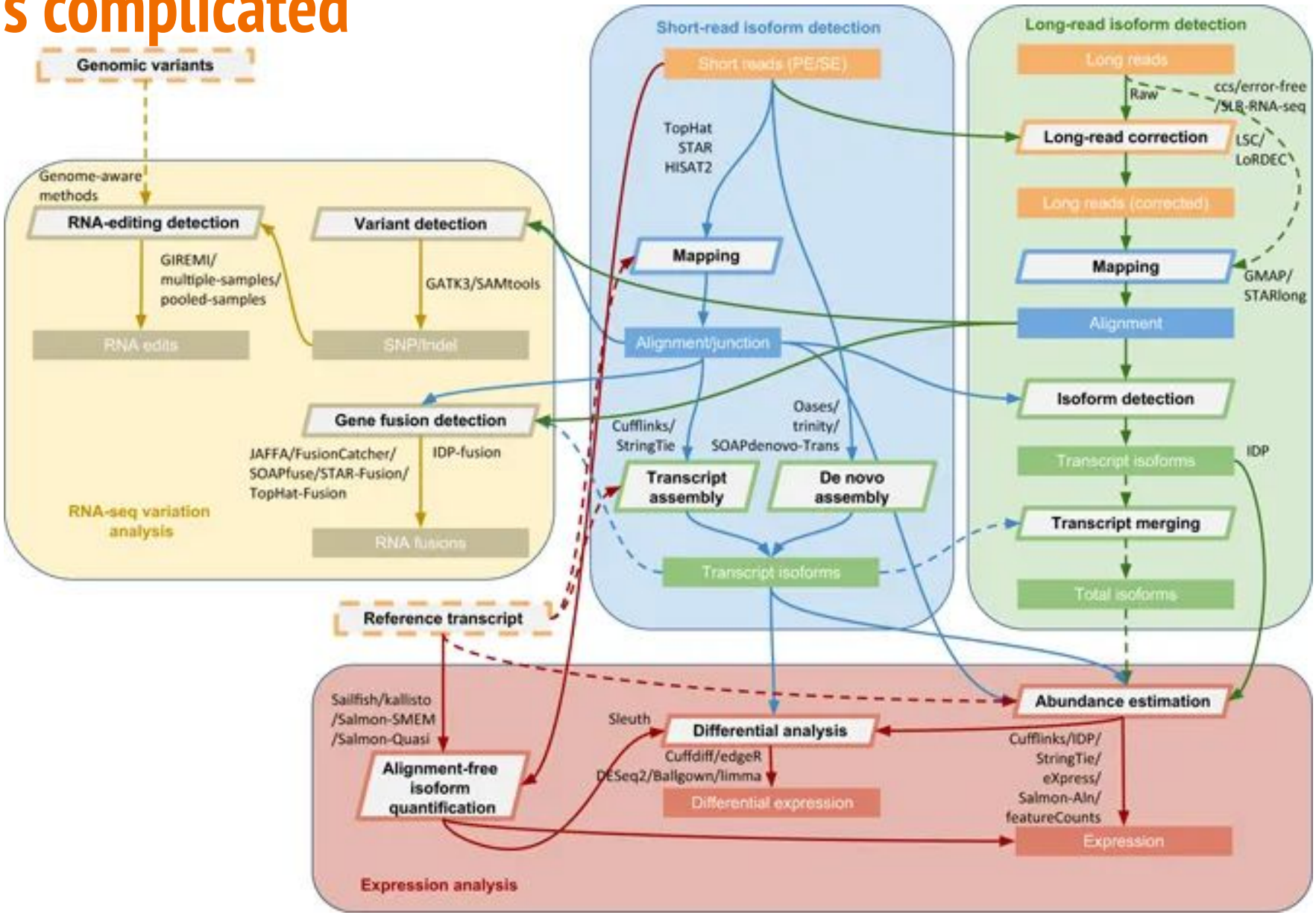
Quality	Error rate
10	10%
20	1%
30	0.1%
40	0.01%

What people do with their RNA-seq



From J. Audoux's PhD thesis

It's complicated



Nature Communications **8**, Article number: 59 (2017)

Outcomes of RNA-seq studies

- gene annotation
- protein/function prediction
- gene/splicing quantification
- isoform discovery/fusion transcripts/lncRNA...
- variant calling
- methylations
- RNA structures
-

QC - Cleaning - Preprocessing

Known biases in RNA-seq

?

Known biases in RNA-seq

Biological sample:

- presence of pre-mRNA
- 3’ bias over-represented (RNA degradation)
- contaminations

Library preparation:

- DNAse fail
- rRNA depletion not effective
- pcr bias
- variable insert size (smaller than sequencing length)
- reads with no inserts

Sequencing:

- quality drops at the end of reads

Quality Control (QC)

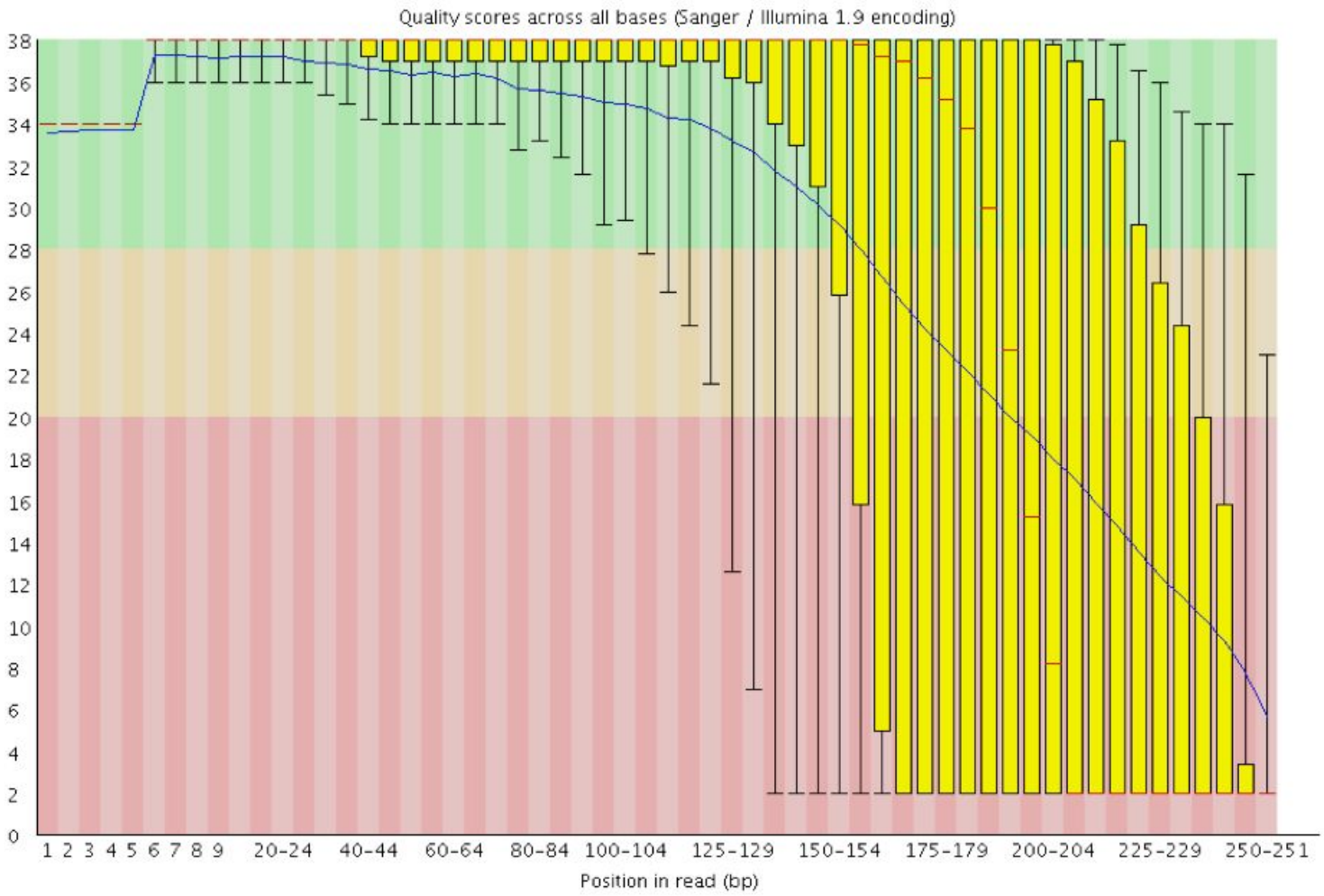
Quality Control (QC) is important to:

- Check if your sample sequencing went well
- Know when you need to sequence again (sequencing platform QC fail)
- Identify potential problems that can be fixed, or not
- Follow the impact of preprocessing steps

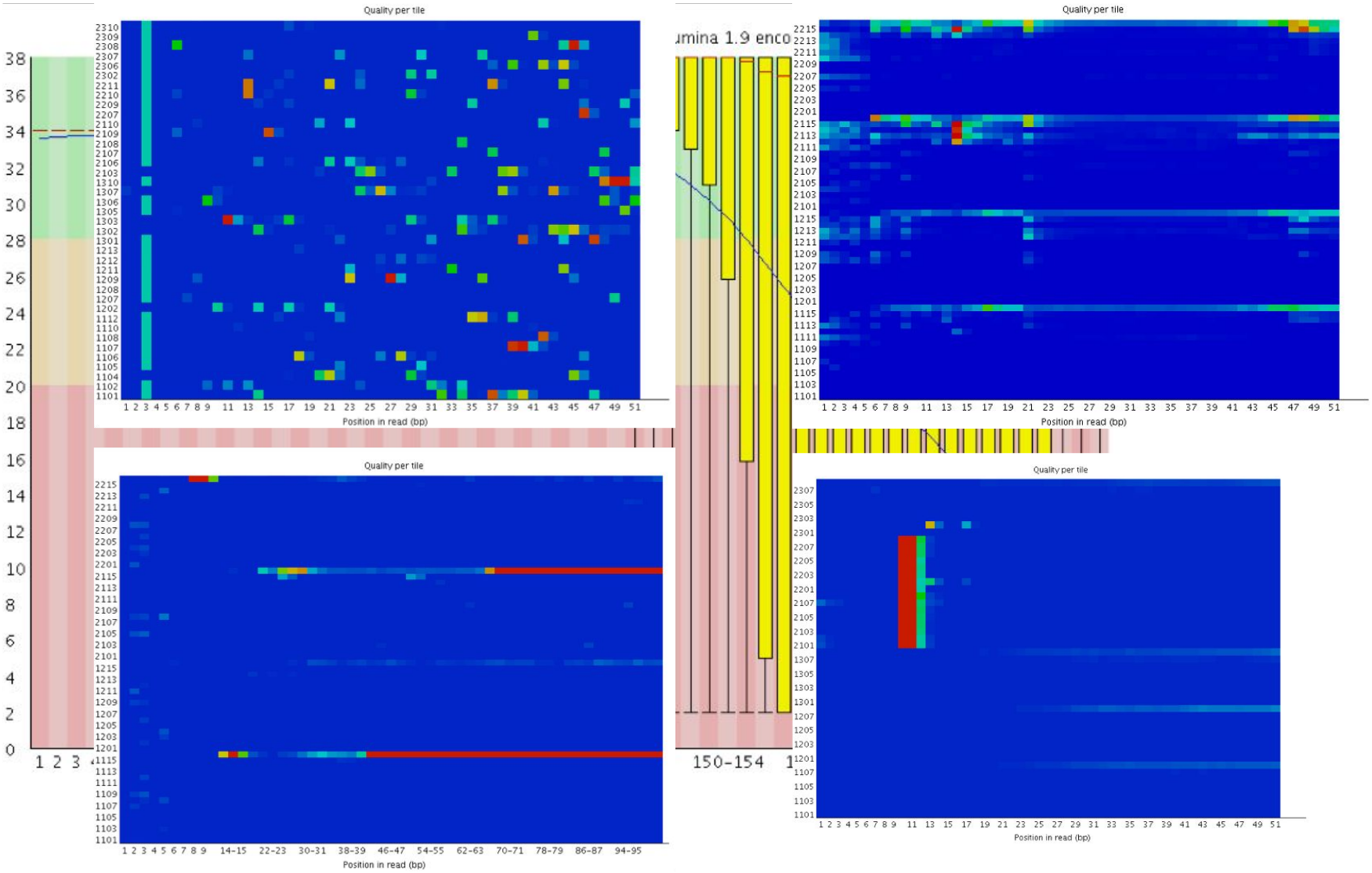
⇒ FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

+ MultiQC (https://multiqc.info/) when comparing multiple datasets

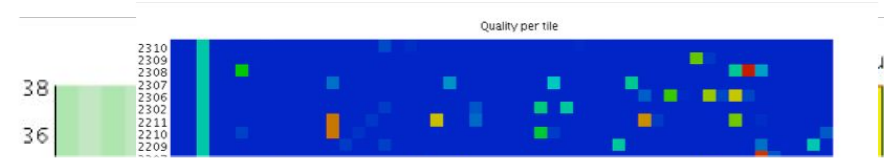
Quality Control

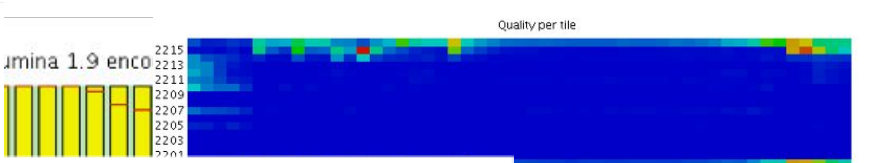


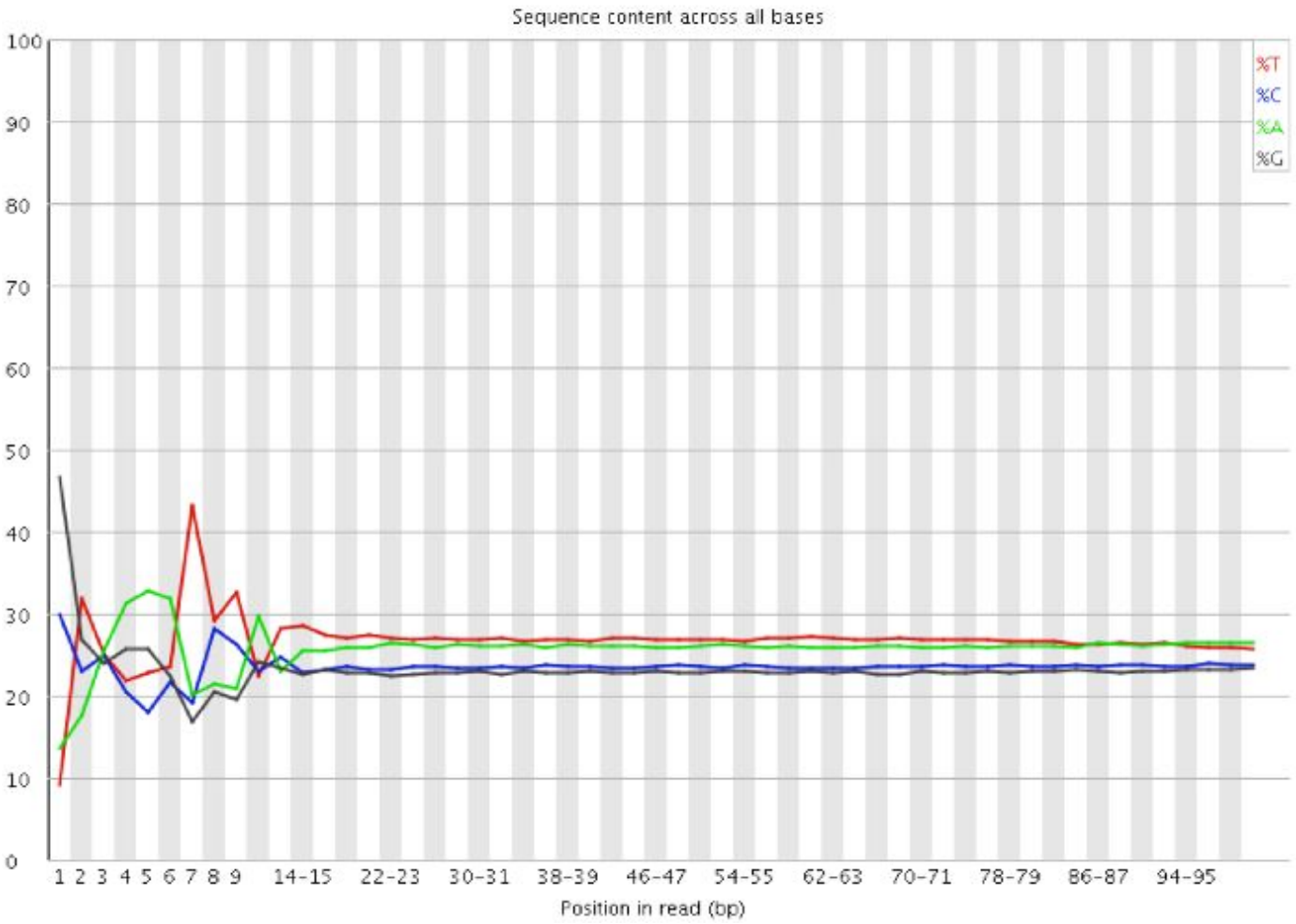
Quality Control

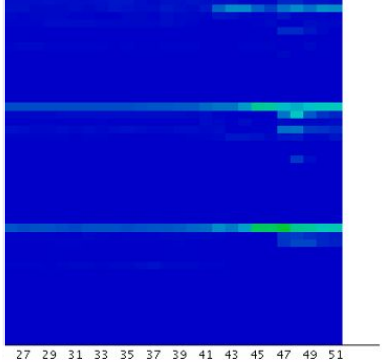


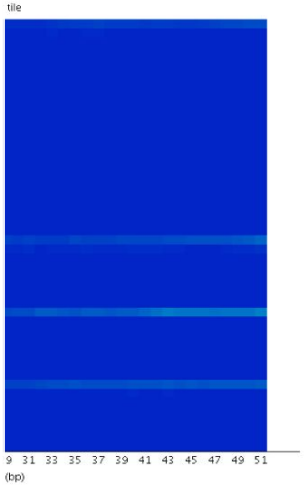
Quality Control



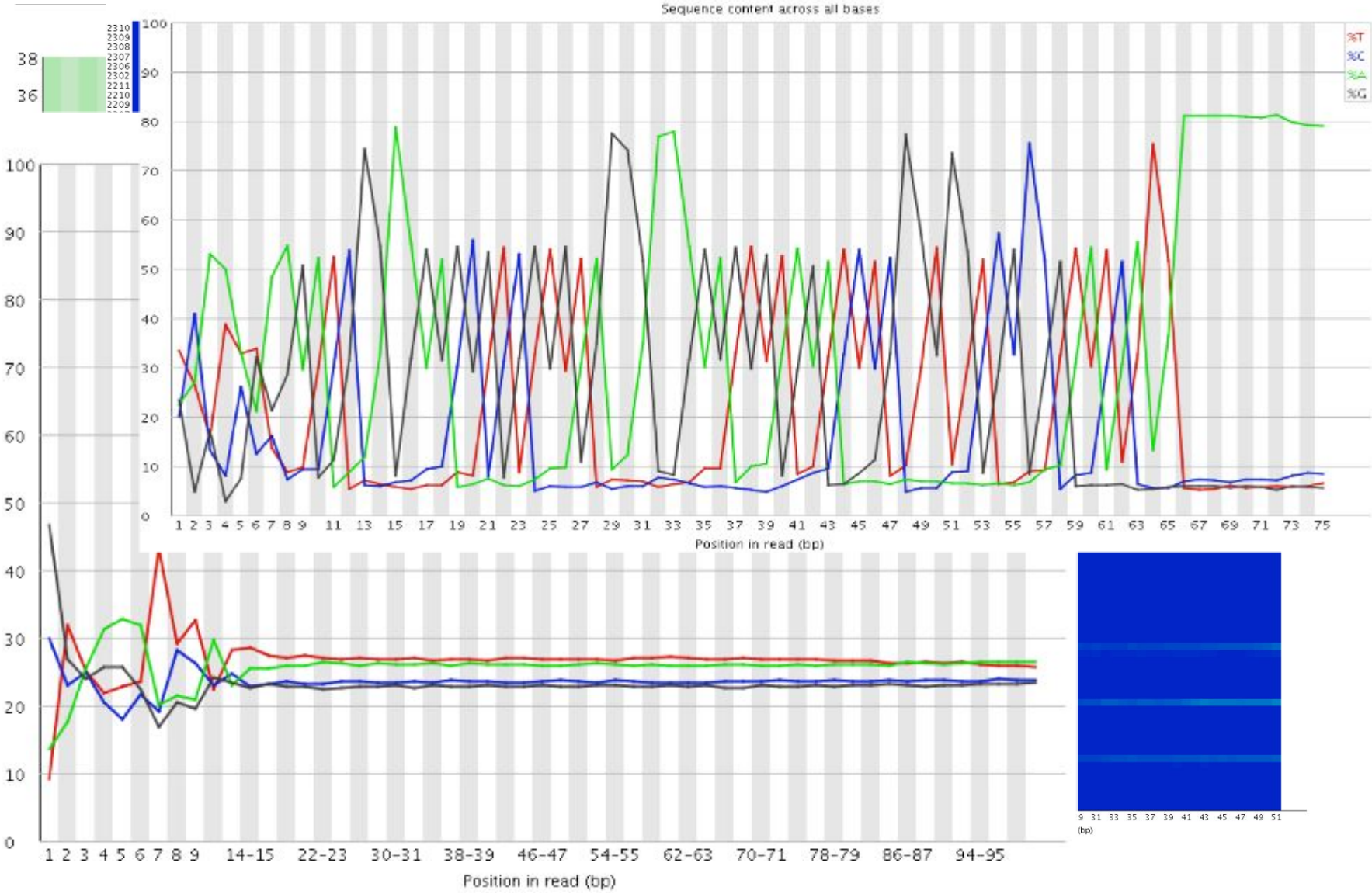




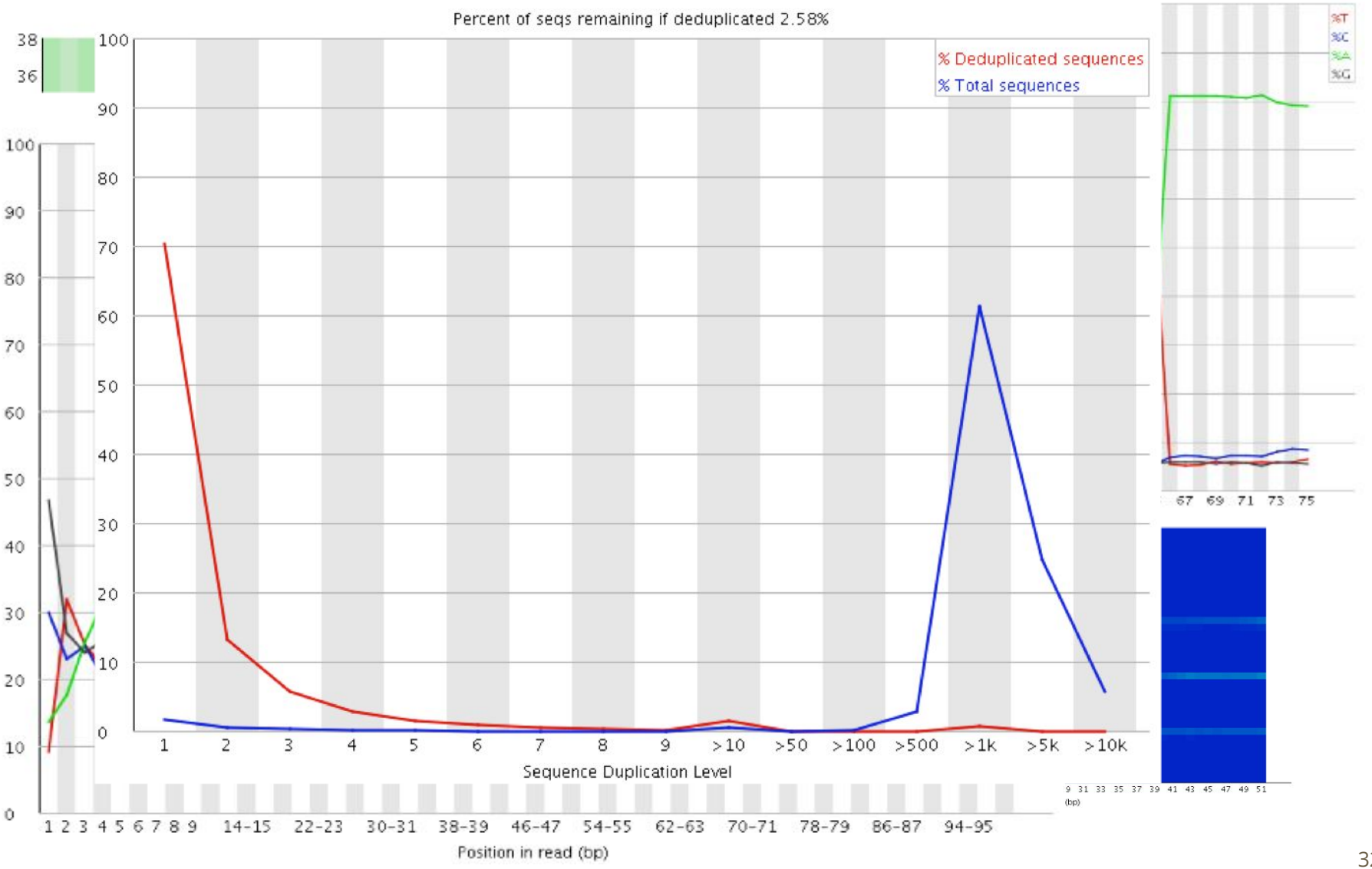




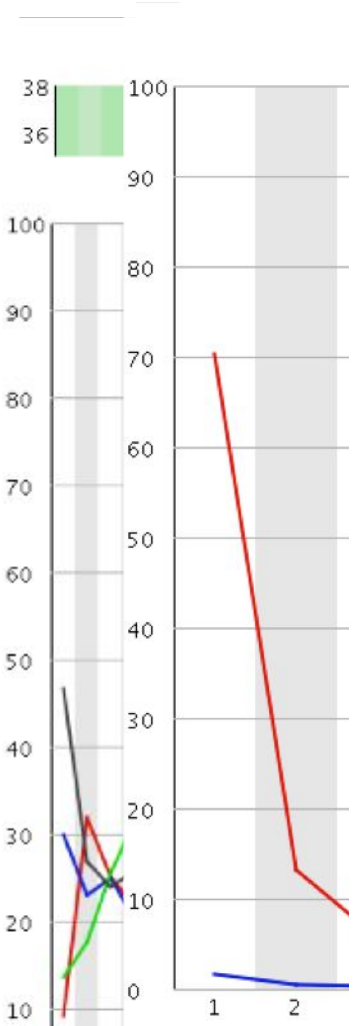
Quality Control

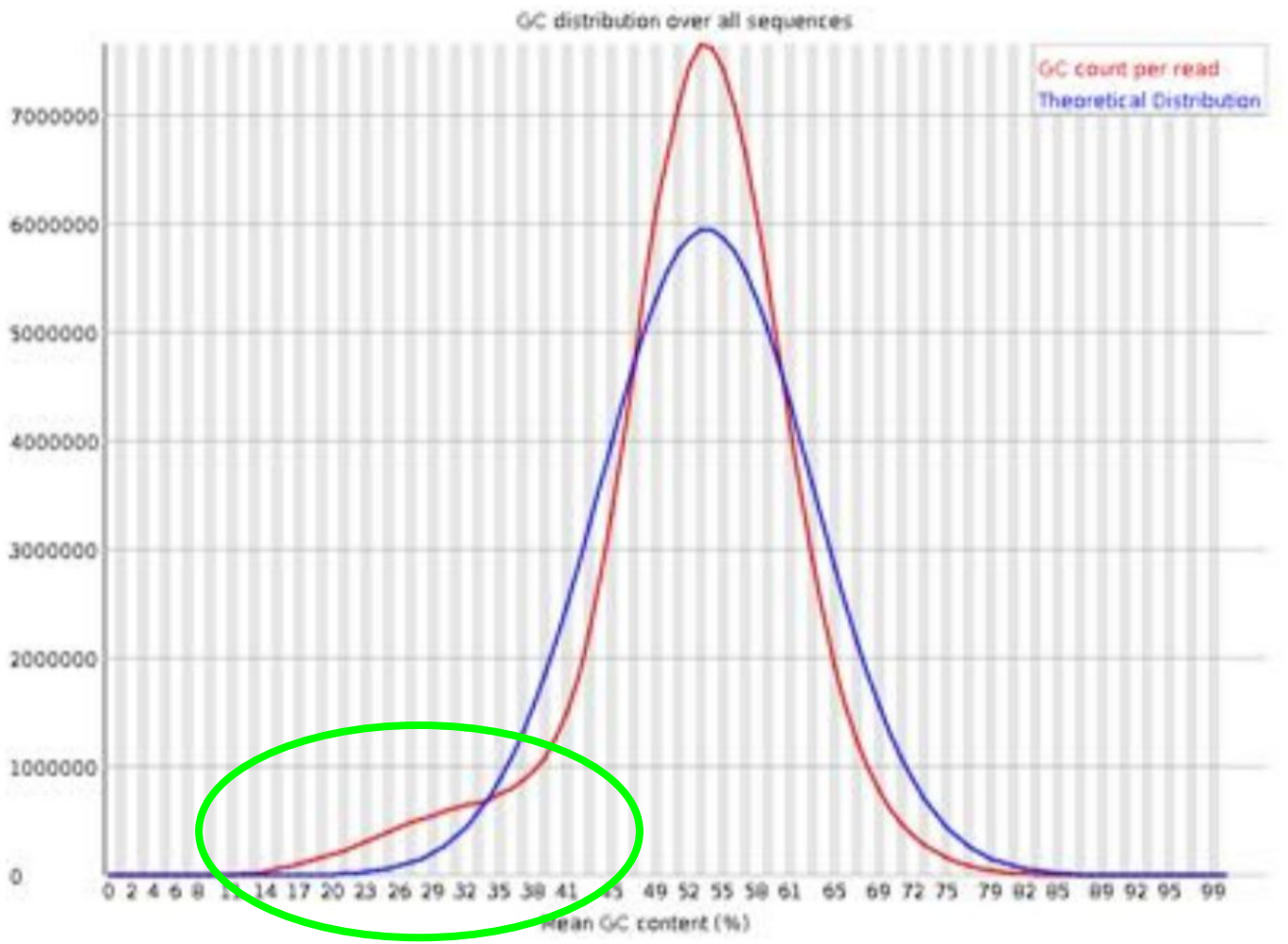


Quality Control

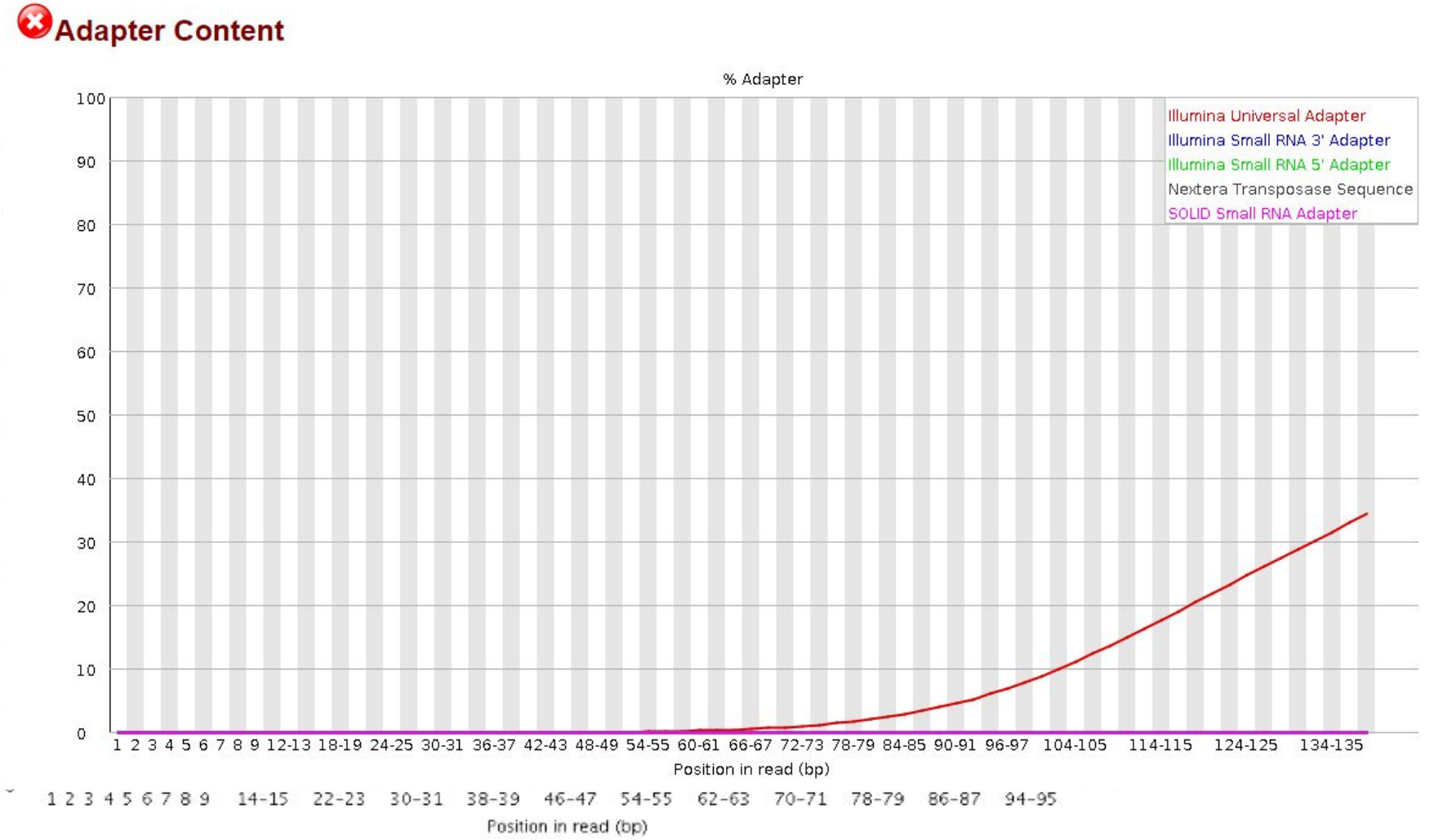


Quality Control





Quality Control



Cleaning - Preprocessing

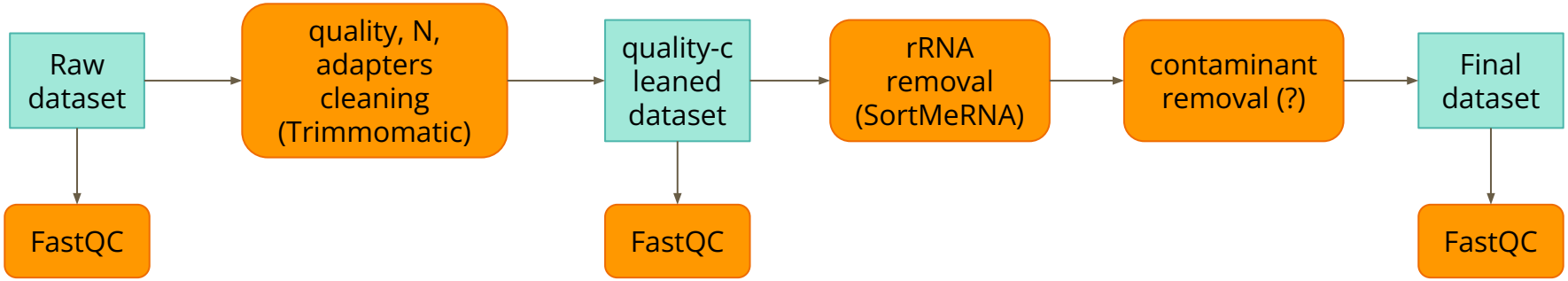
Cleaning should be done in the reverse order that errors were generated.

1. Sequencing errors: quality trimming and filtering, Ns removal
2. Library preparation: adapters removal
3. Sample contamination: rRNA, mito, other contaminants

Note 1: step 1 (quality trimming) is not considered critical anymore and could even hinder downstream tools/algorithms.

Note 2: If the reads are going to be aligned against a reference genome, this whole process can be skipped or applied very lightly

Cleaning - Preprocessing



Practical: Quality Control (QC) & Cleaning

Open Galaxy

Practical: NGS2022_rnaseq_04_tp_bioinfo.pdf

Shared Data → Histories → TP RNAseq bilille Initial datasets

TIAAS: https://usegalaxy.fr/join-training/bilille-2022-rnaseq/

To map or not to map ?

With reference RNA-seq

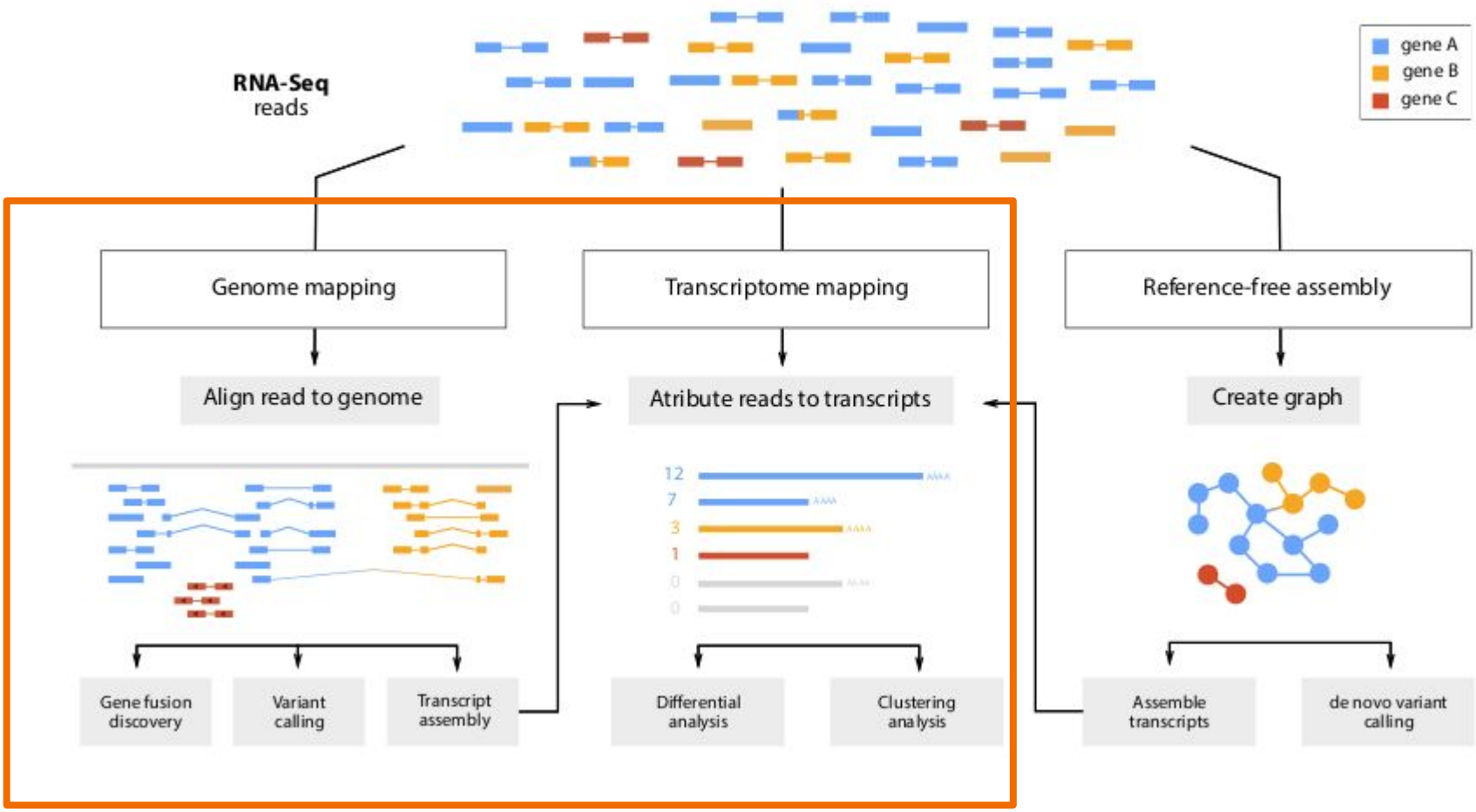
W/ reference RNA-seq. For what purpose ?

Mainly:

- Differential expression
  - between genes
  - between transcripts/isoformes

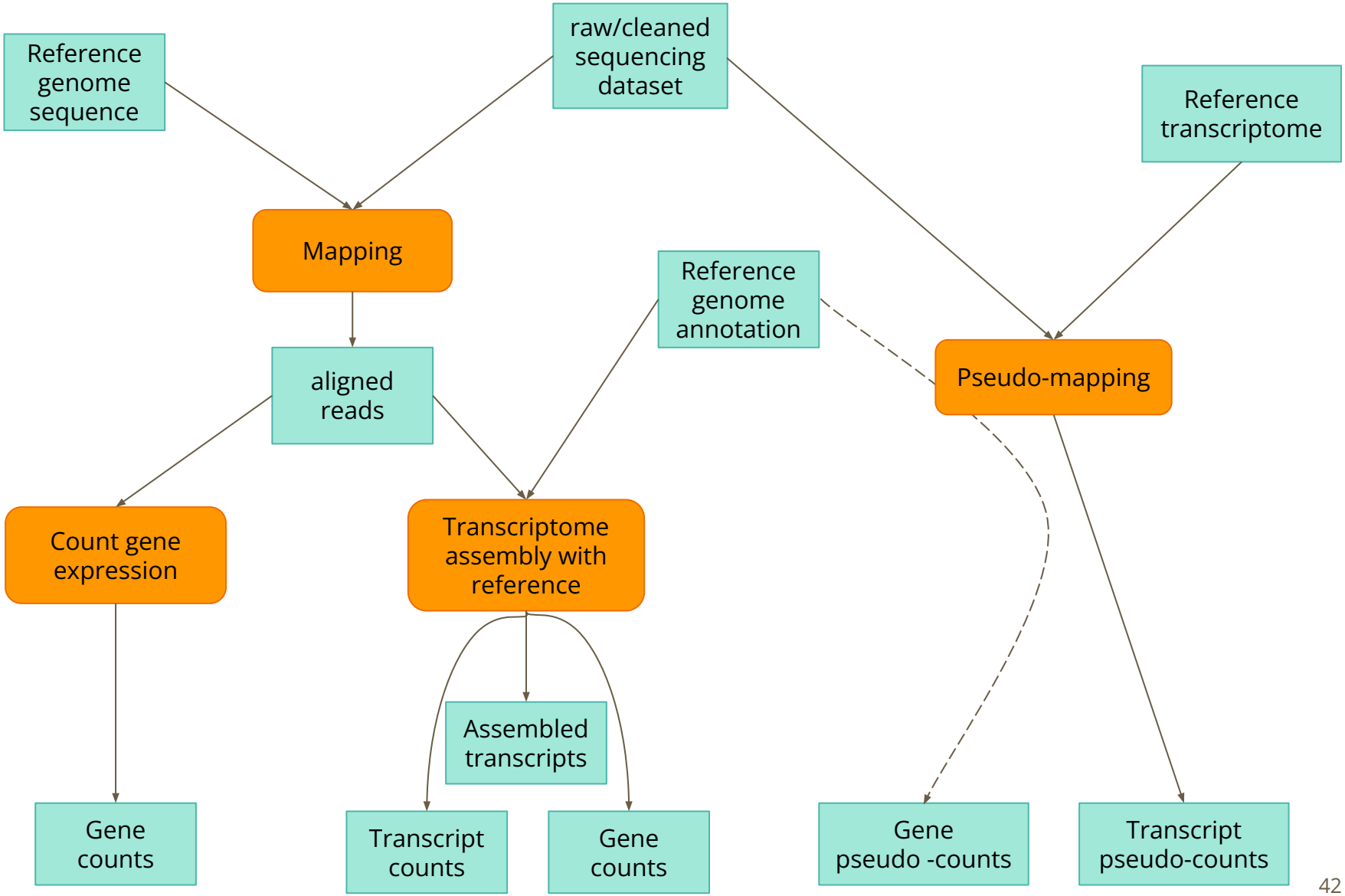
- Transcriptome assembly
  - variant calling
  - isoforme discovery

What people do with their RNA-seq

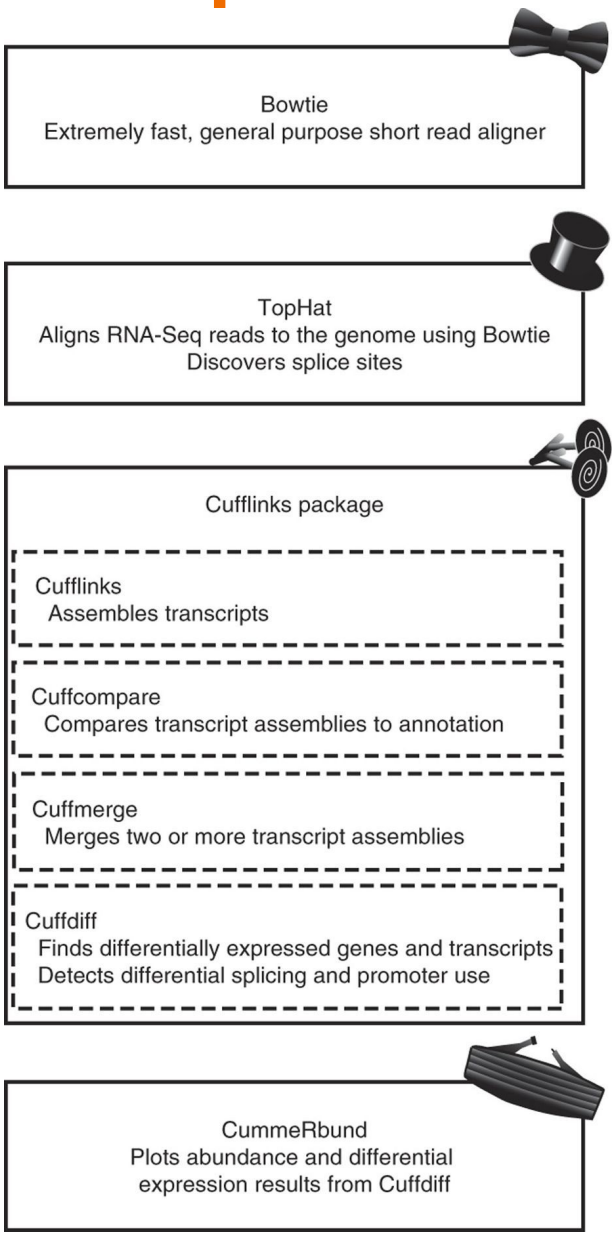


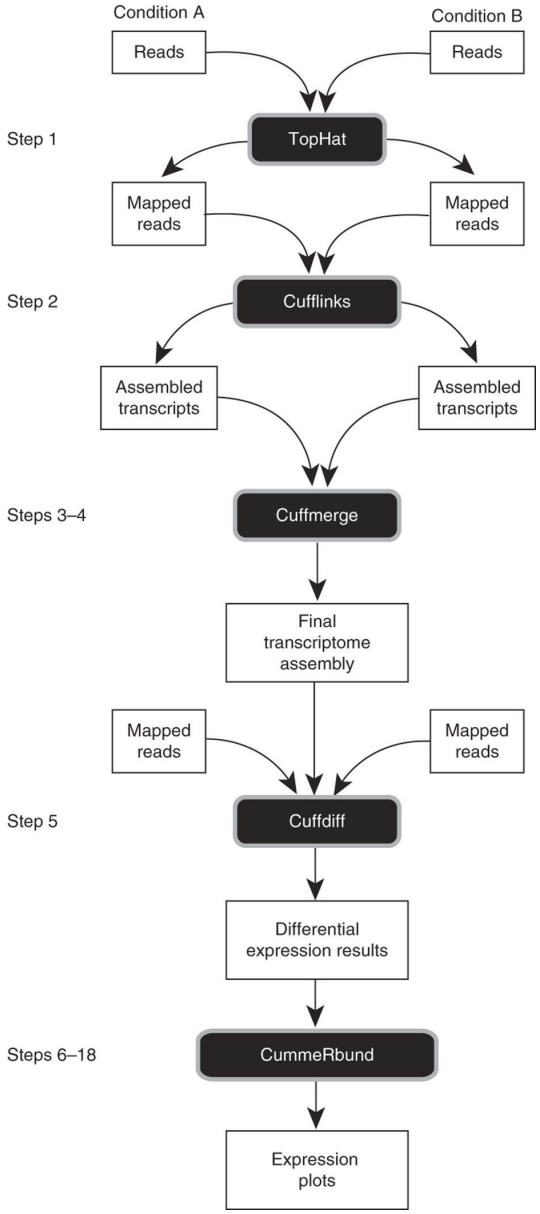
From J. Audoux's PhD thesis

RNA-seq w/ ref


Reference genome sequence
raw/cleaned sequencing dataset
Reference transcriptome
Mapping
Reference genome annotation
aligned reads
Pseudo-mapping
Count gene expression
Transcriptome assembly with reference
Assembled transcripts
Gene counts
Transcript counts
Gene counts
Gene pseudo -counts
Transcript pseudo-counts

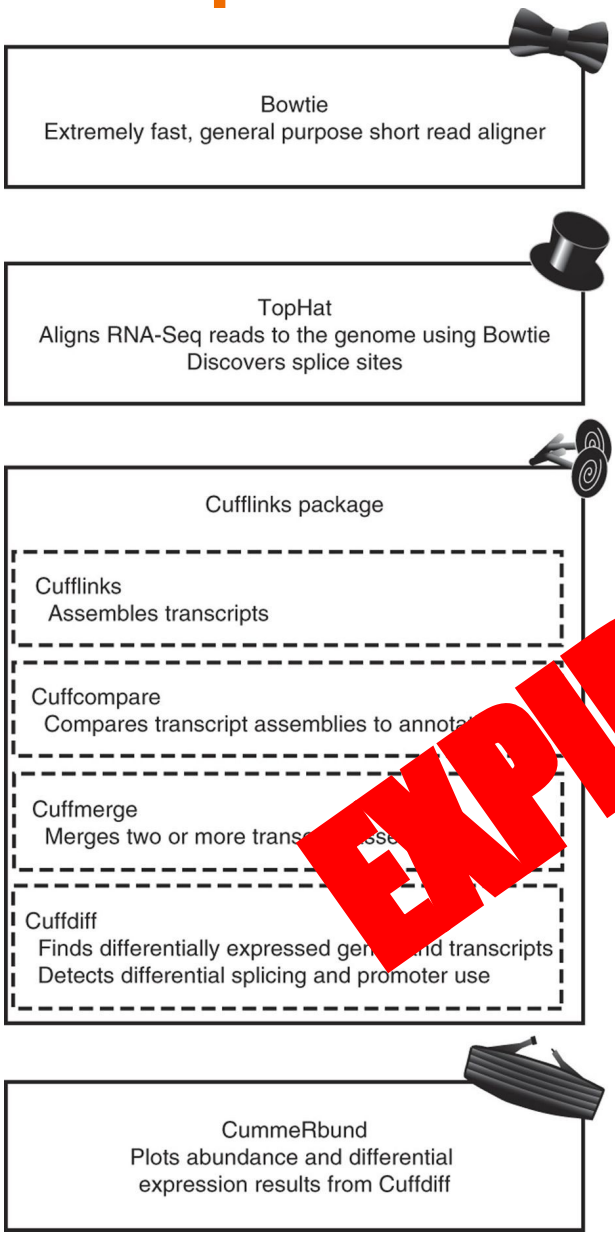
The champion: Tuxedo Suite, "Classic" version



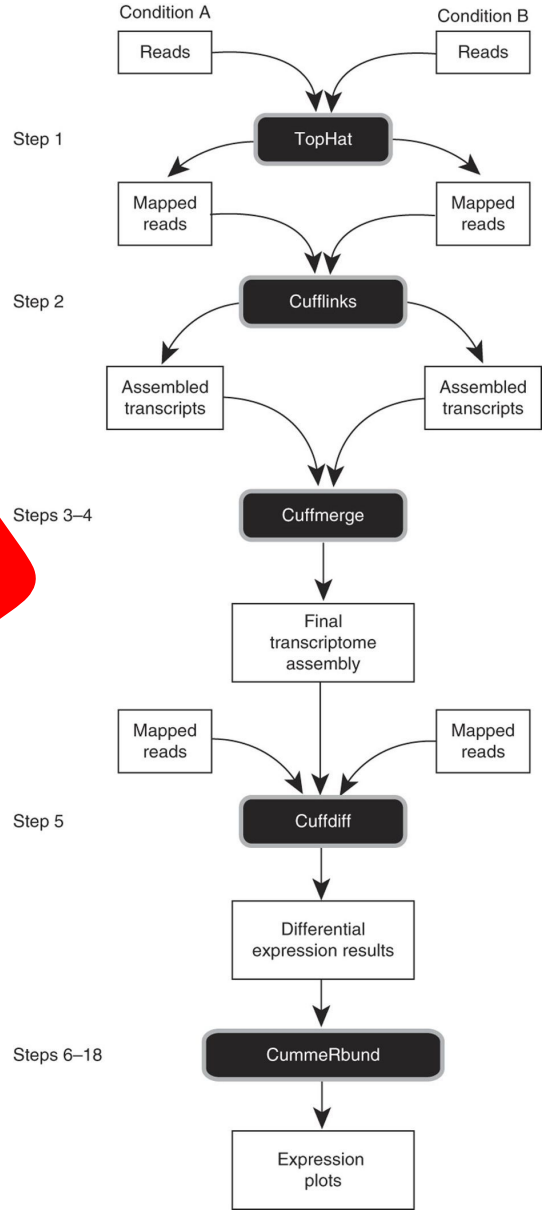


Nat Protoc. 2012;7(3):562–578. doi:10.1038/nprot.2012.016

The champion: Tuxedo Suite, "Classic" version



EXPIRED



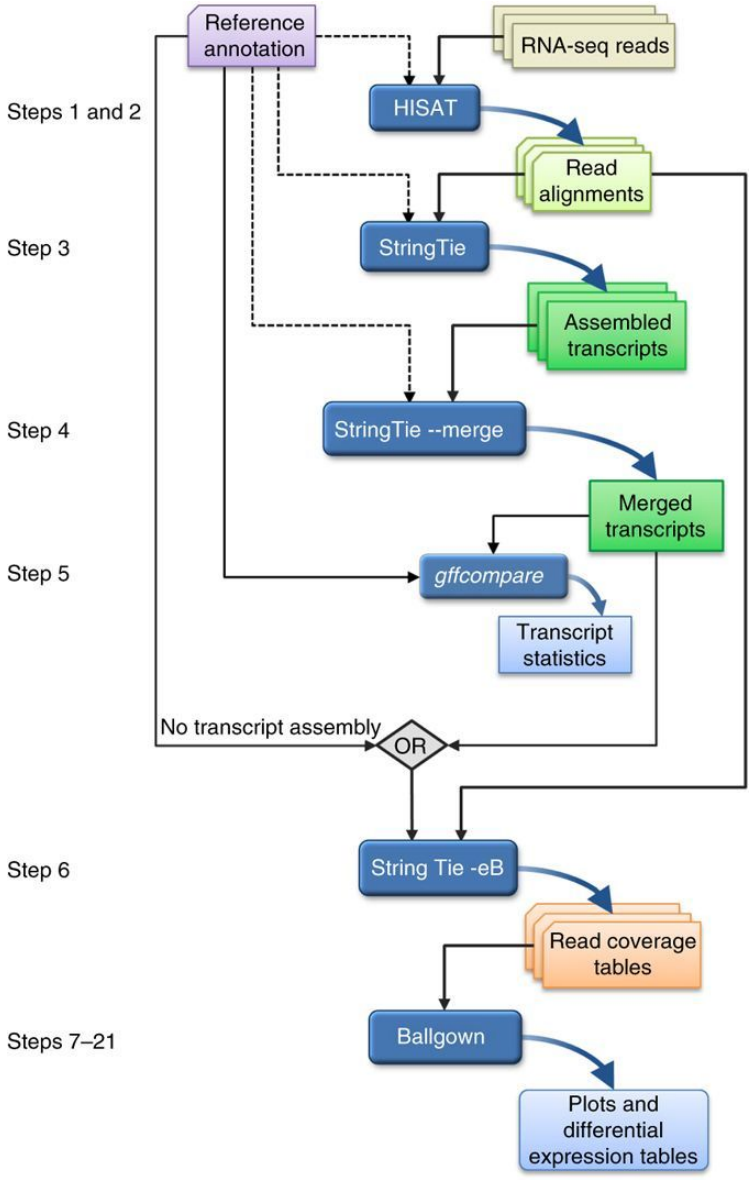
Nat Protoc. 2012;7(3):562–578. doi:10.1038/nprot.2012.016

The champion: Tuxedo Suite, New version

HISAT/HISAT2: splice aware aligner

StringTie: Transcriptome assembler

Ballgown: Differential expression analysis



Nat Protoc. 2016;11(9):1650–1667. doi:10.1038/nprot.2016.095

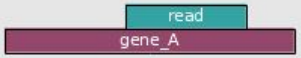
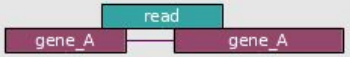
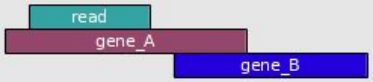
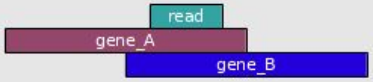
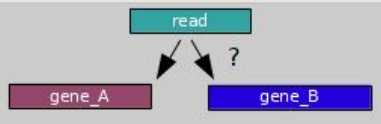
Counting gene expression from alignments

A

B

Approach to handle multireads	Read distribution representation	Counts
Ignore		G1: 10 reads G2: 6 reads
Count once per alignment		G1: 18 reads G2: 14 reads
Split them equally		G1: 14 reads G2: 10 reads
Rescue based on uniquely mapped reads		G1: 15 reads G2: 9 reads
Expectation-maximization	(n)	G1: 15 reads G2: 9 reads
Read coverage based methods		G1: 15 reads G2: 9 reads
Cluster methods		G1:10 reads G2:6 reads Cluster G1/G2: 8 reads

Deschamps-Francoeur, et al. 2020. doi:10.1016/j.csbj.2020.06.014

	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 (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

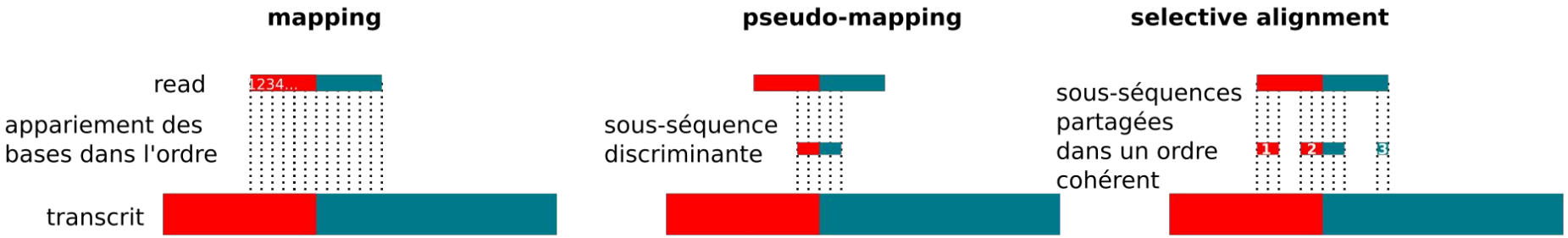
Counting gene expression from alignments

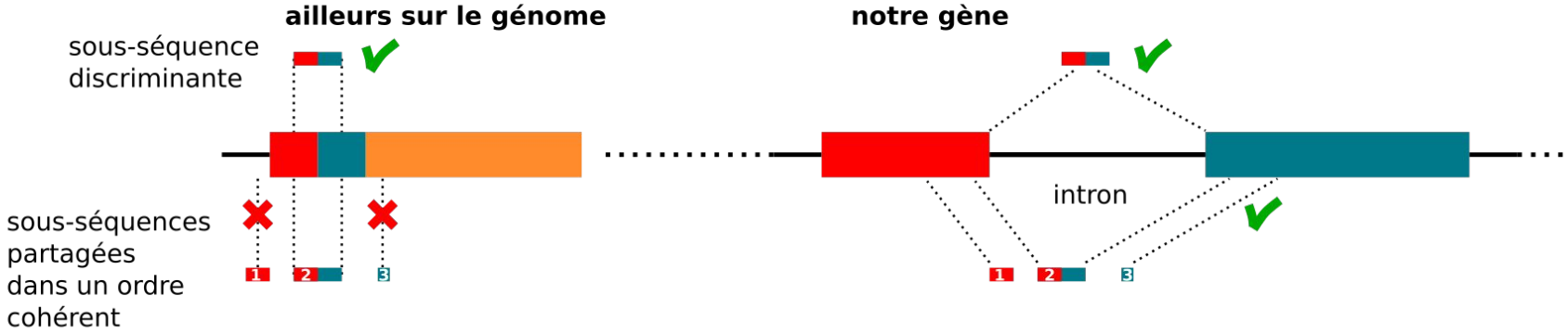
Table 1
Computational strategies and methods that handle multi-mapped reads.

Tool	Quantification level	Input	Strandedness can be specified	Count type	Strategy	Paired end	Confidence level	Focus
HTSeq-count	Gene	BAM	Y	Counts	Ignore	Y	N	Long RNA
STAR geneCounts	Gene	Fastq	Y	Counts	Ignore	Y	N	Long RNA
Cufflinks	Transcript	BAM	Y	RPKM	Split equally, Rescue	Y	N	Long RNA
featureCounts	Gene	BAM	Y	Counts	Ignore, count all, split equally	Y	N	Long RNA
CoCo	Gene	BAM	Y	Counts, CPM, TPM	Rescue	Y	N	Small RNA Long RNA
ERANGE	Transcript	BAM	N	RPKM	Rescue	Y	N	Long RNA
EMASE	Transcript	BAM	N	Counts, TPM	EM	Y	N	Long RNA
IsoEM2	Both	SAM	Y	FPKM, TPM	EM	Y	Confidence intervals	Long RNA
Kallisto	Transcript	Fastq	Y	TPM	EM	Y	Bootstrap values	Long RNA
RSEM	Both	Fastq, BAM	Y	Counts, TPM, FPKM	EM	Y	95% credibility intervals	Long RNA
Salmon	Transcript	Fastq	Y	Counts, TPM	EM	Y	Bootstrap values	Long RNA
MMR	N/A	BAM	Y	N/A	Read coverage	Y	N/A	Long RNA
MuMRescueLite	Genomic loci	Custom format	N	Counts	Read coverage	N	N	Short sequence tags
Rcount	Gene	BAM	Y	Counts	Read coverage	N	N	Long RNA
ShortStack	Gene	Fastq, BAM	N	Counts, RPM	Read coverage	N	N	Small RNA
mmquant	Gene	BAM	Y	Counts	Gene Clustering	Y	N	Small RNA Long RNA
SeqCluster	Gene	BAM	N	Counts	Gene clustering	N	N	Small RNA
Fuzzy method	Gene	Custom format	N	Fuzzy counts	Fuzzy sets	N	Fuzzy counts	Small RNA Long RNA
geneQC	Gene	SAM	Y	NA	ML	Y	Mapping uncertainty level	Small RNA Long RNA

Deschamps-Francoeur, et al. 2020. doi:10.1016/j.csbj.2020.06.014

Salmon selective alignment





Counting gene expression with Salmon

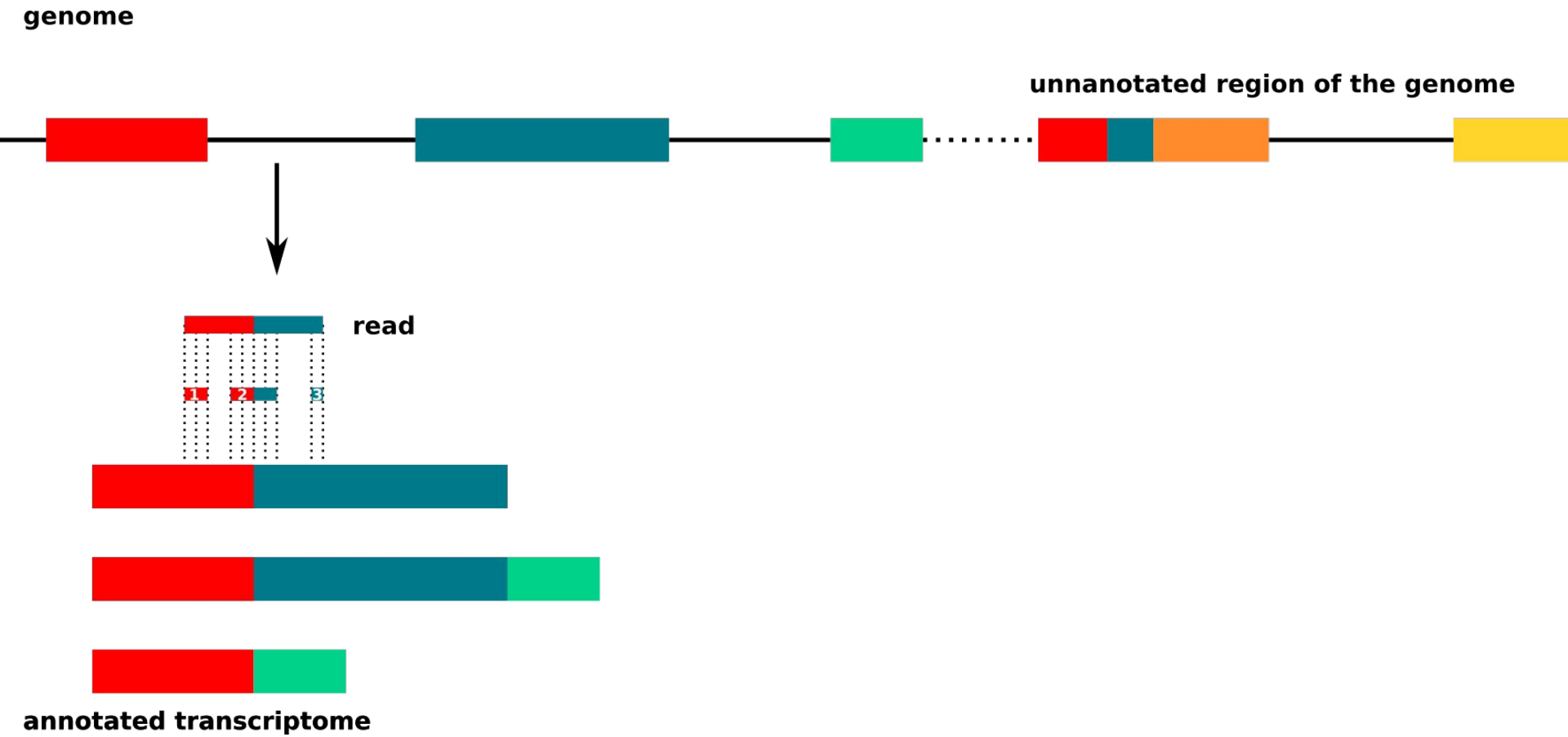
Salmon Selective alignment

- inspiré des techniques d'alignement long-reads
- au lieu de "simplement" assigner les reads à des transcrits en utilisant des sous séquence partagées **sans notion de position**, on vérifie que les les **sous-séquences partagées s'ordonnent correctement** sur le transcrit avant de proposer une quantification

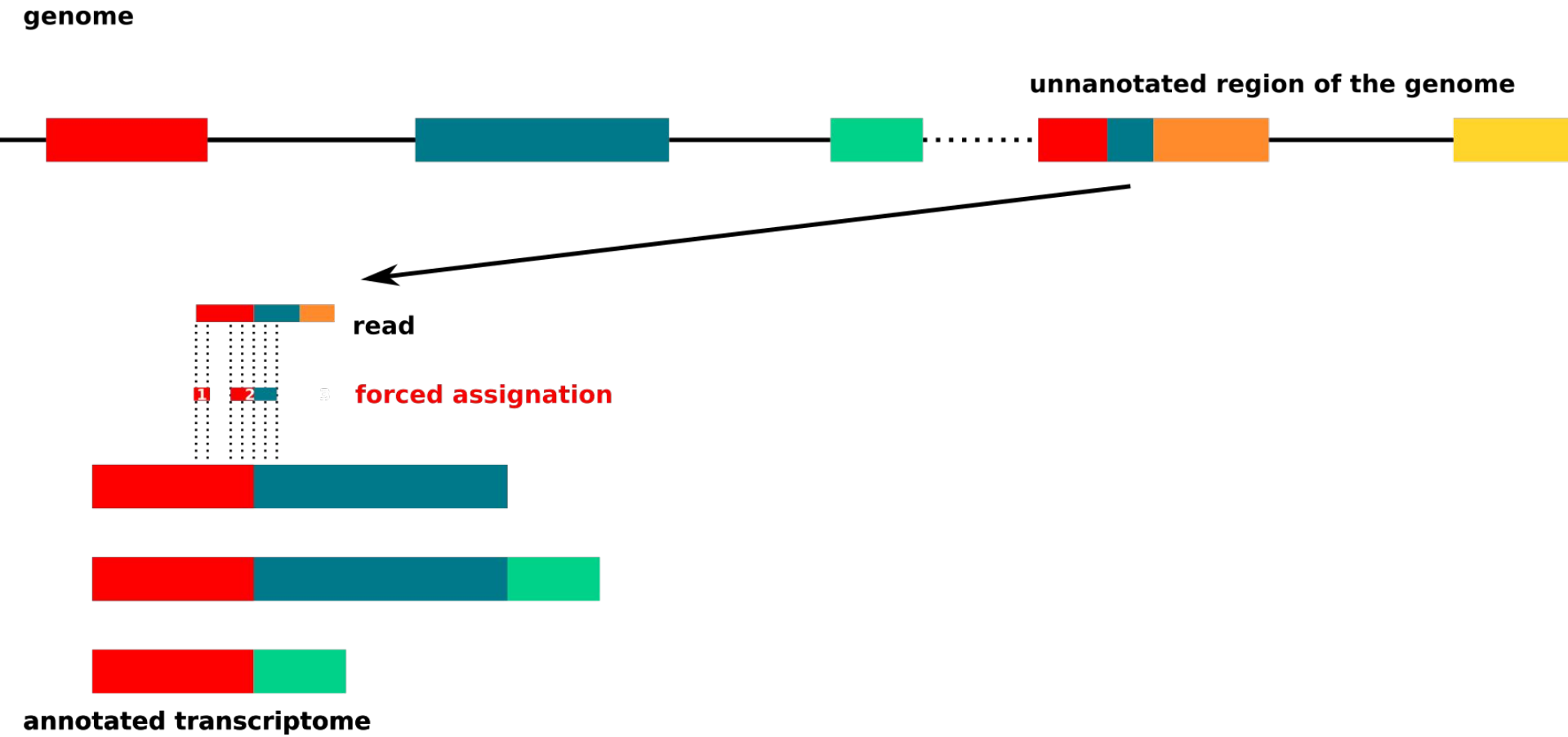
- des résultats récents* suggèrent que cette méthode est une 3ème voix, avec une précision de la quantification qui se rapproche du mapping "traditionnel" tout en étant plus rapide

*Alignment and mapping methodology influence transcript abundance estimation, Srivastava et al. 2020

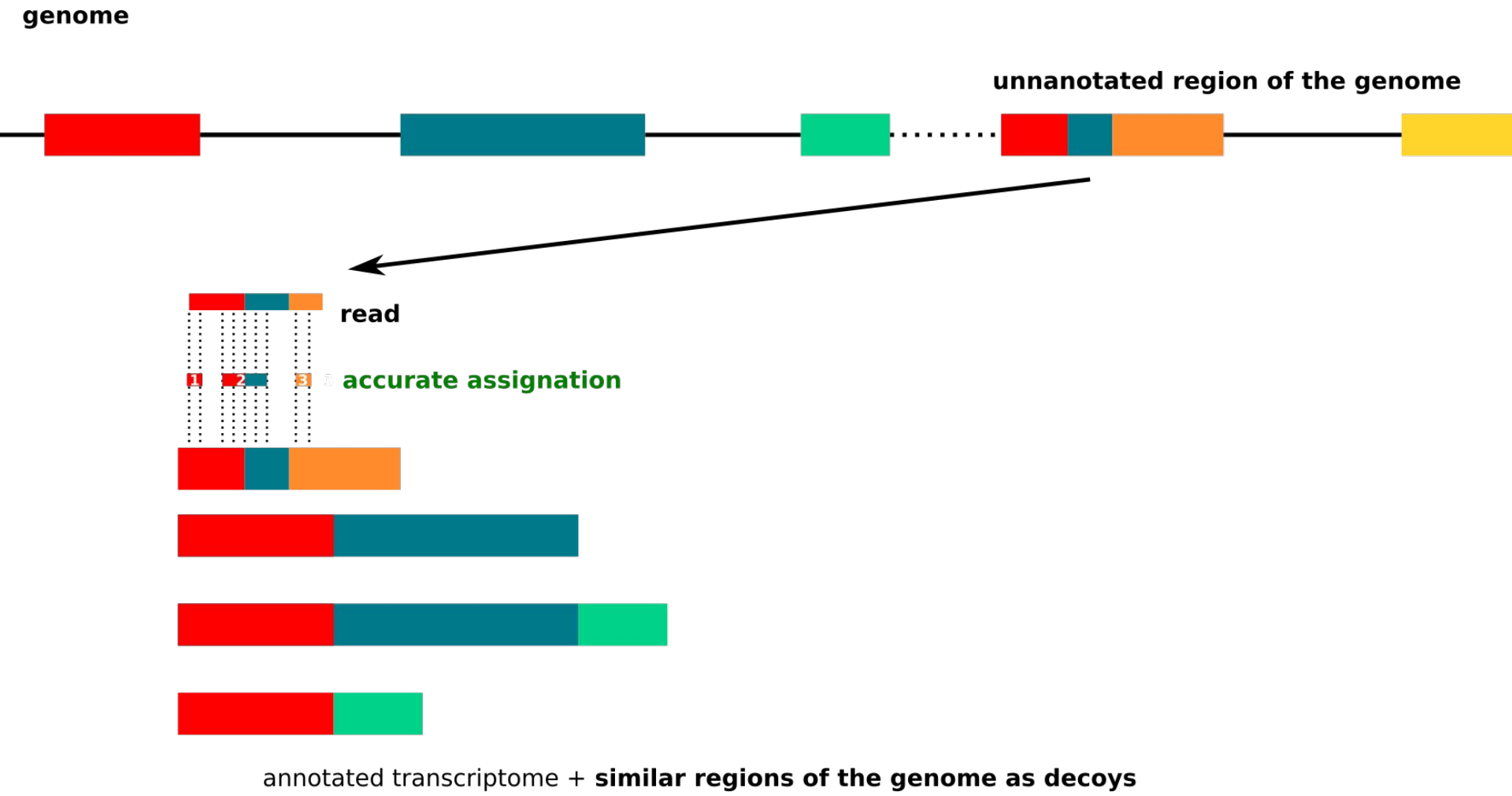
Salmon: using genomic regions as decoys



Salmon: using genomic regions as decoys

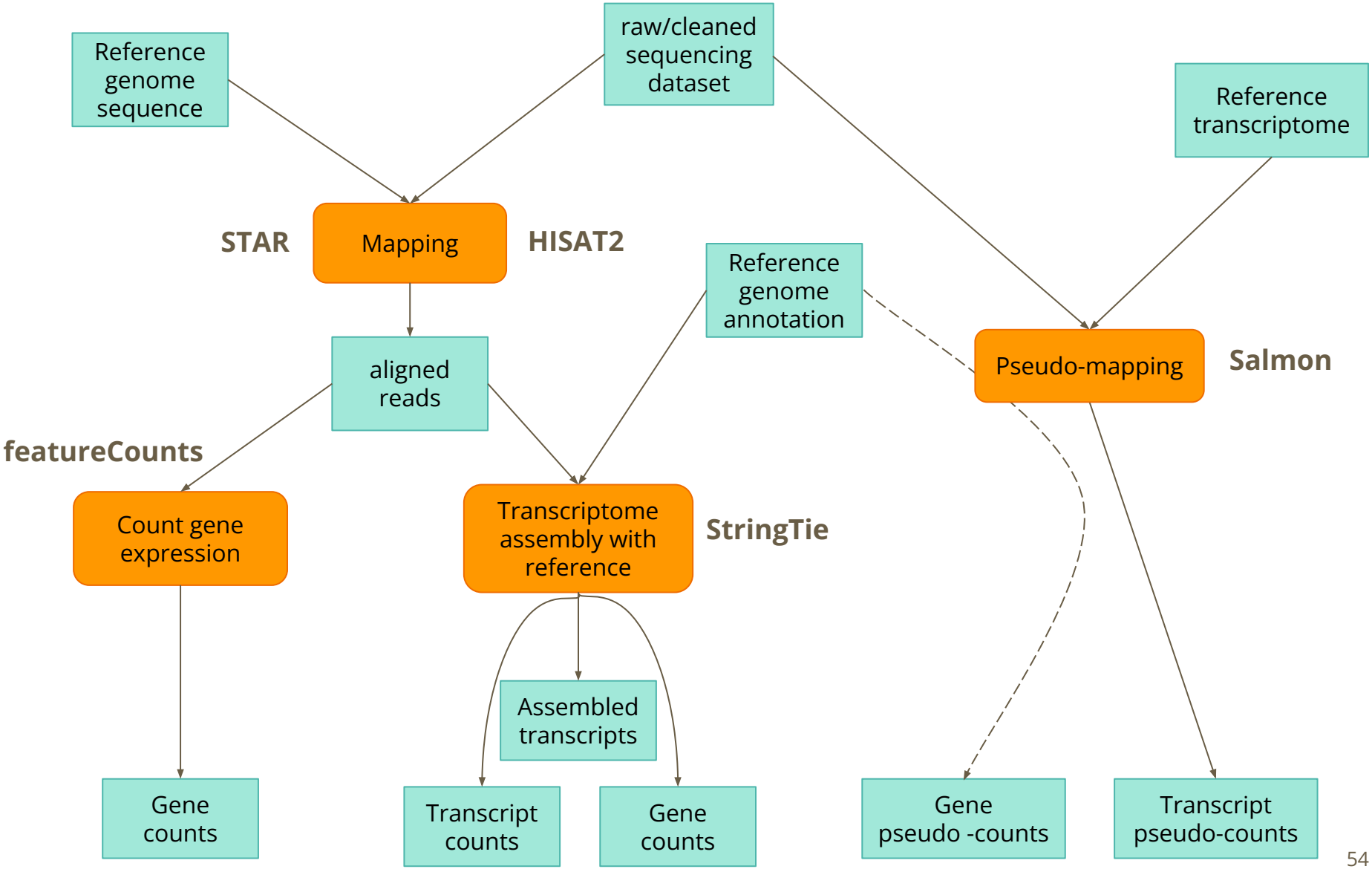


Salmon: using genomic regions as decoys



annotated transcriptome + **similar regions of the genome as decoys**

RNA-seq w/ ref


Reference genome sequence
raw/cleaned sequencing dataset
Reference transcriptome
STAR
Mapping
HISAT2
Reference genome annotation
aligned reads
Pseudo-mapping
Salmon
featureCounts
Count gene expression
Transcriptome assembly with reference
StringTie
Assembled transcripts
Gene counts
Transcript counts
Gene counts
Gene pseudo -counts
Transcript pseudo-counts

Practical: Mapping and Quantification

Recommended pipeline (as of Oct 2022)

- Transcriptome assembly: HISAT2 + StringTie (+ Ballgown ?)
- Transcript/Gene quantification with mapping: STAR + Salmon
- Mapping-less transcript quantification: Kallisto or Salmon

De novo RNA-seq

De novo approaches

- *De novo* methods are approaches that are **free from a reference** for producing results
- Reference-based approaches have limitations as **results depends on the quality of the reference**
- Sometimes we don't even have a reference
- *De novo* and reference-based are **complementary**

Why do we need *de novo* approaches

Aren't references good enough?

- Disease-associated transcripts
- Genetic polymorphism in transcripts
- *de novo* methods are helping creating tomorrow's references

Abstract

Reference transcriptomes: the making of

Enter direct RNA-seq assembly

Shall we ever reach a complete reference transcriptome?

Ignore non-reference transcripts at your own risks

Opinion | Open Access

Bridging the gap between reference and real transcriptomes

Antonin Morillon and Daniel Gautheret

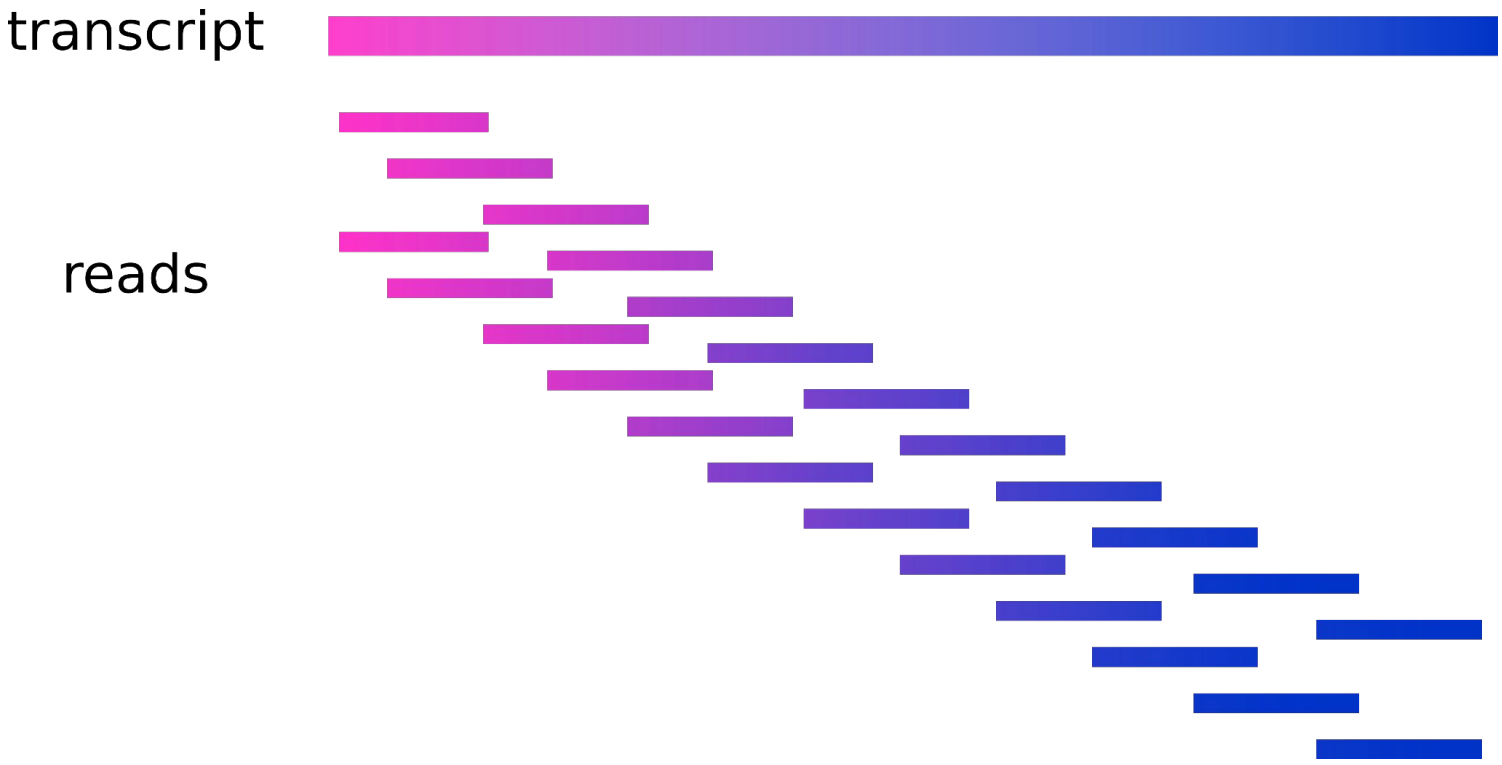
Genome Biology 2019 **20**:112
https://doi.org/10.1186/s13059-019-1710-7
Published: 3 June 2019

The more novel and specific is your need, the more likely you need new bioinformatics (and *de novo*)

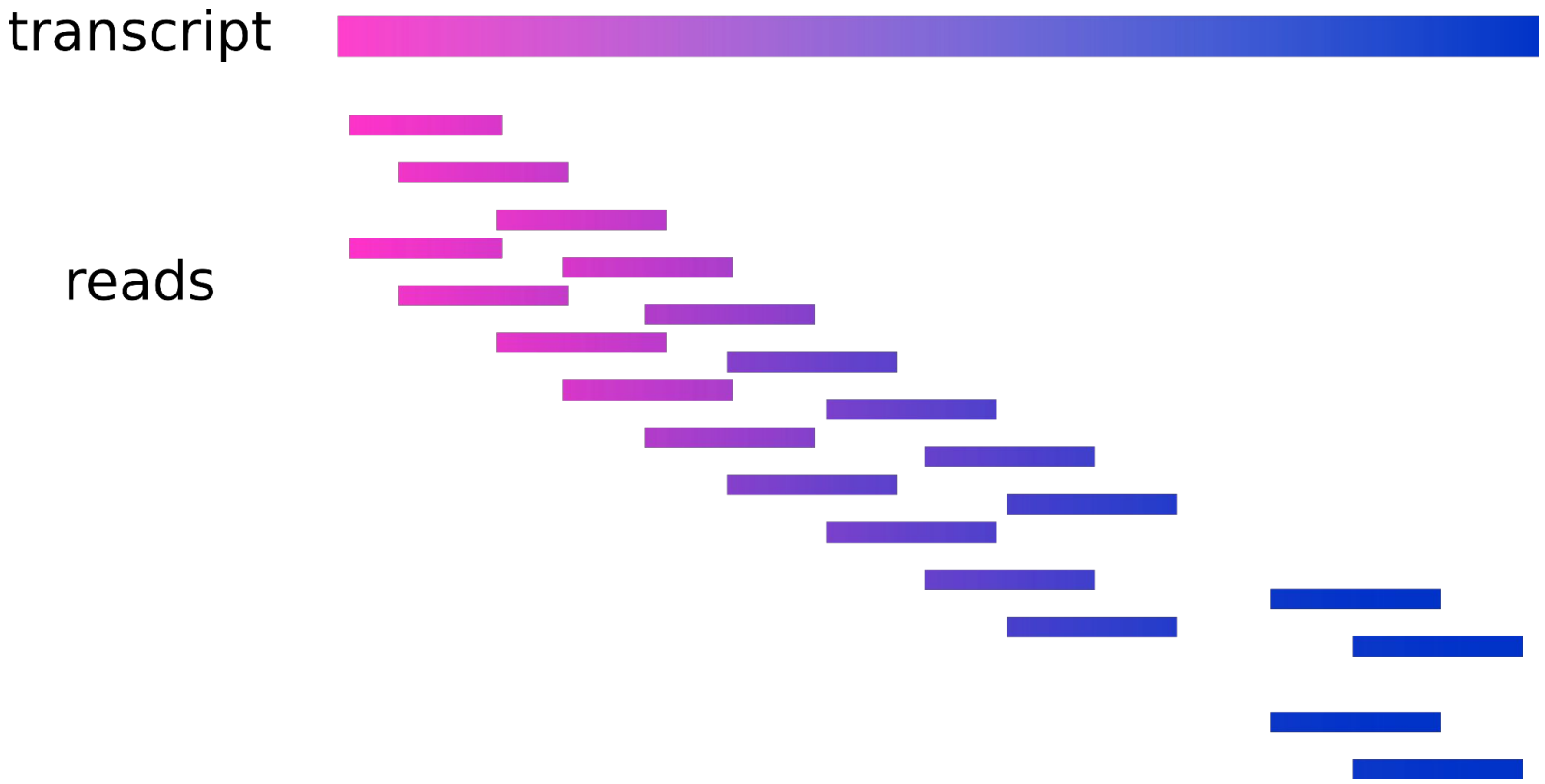
What can be done with *de novo* methods

- transcript assembly + quantification
- genetic polymorphism detection
- alternative transcript detection + quantification

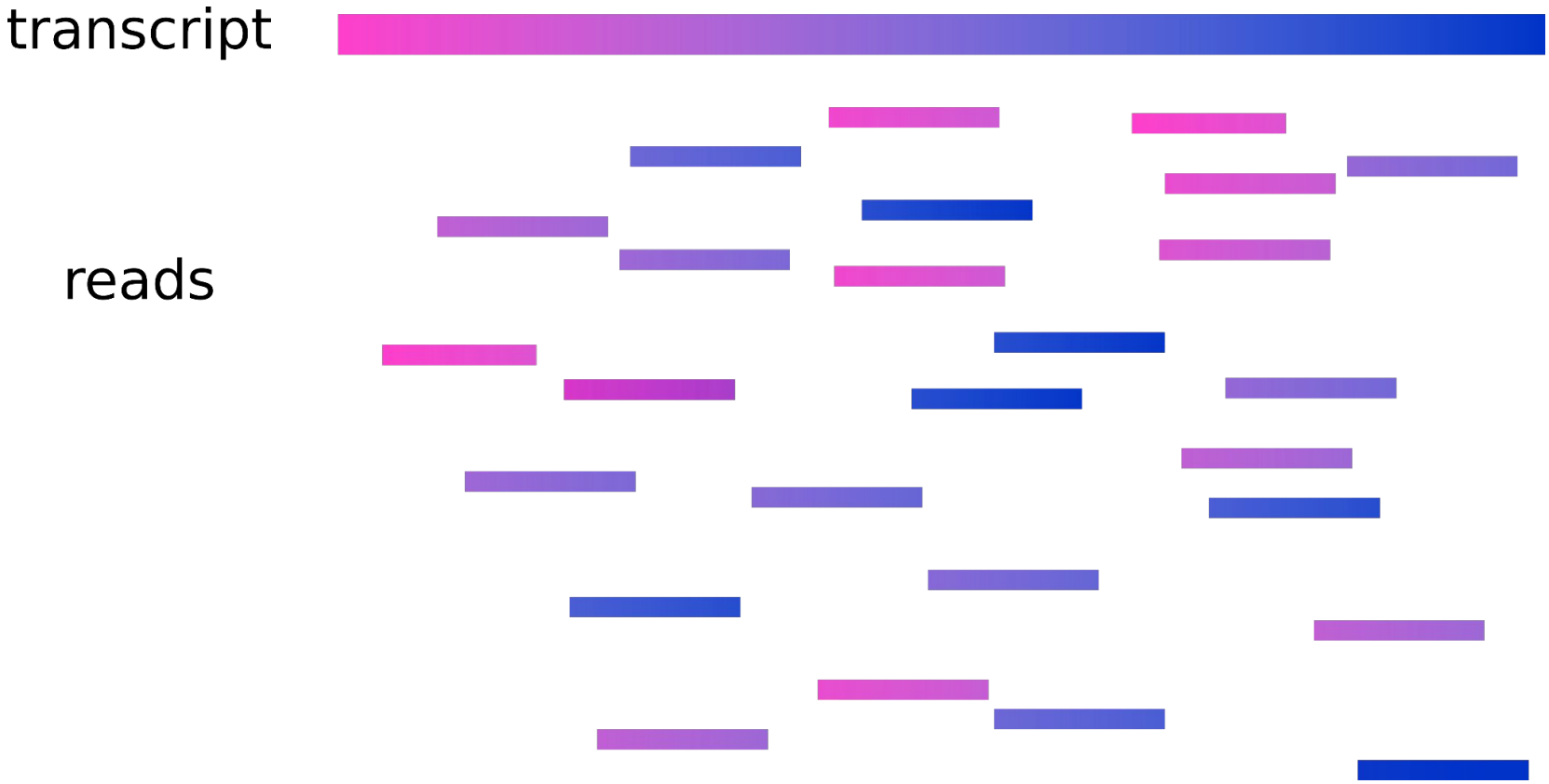
The *de novo* assembly challenge



The *de novo* assembly challenge



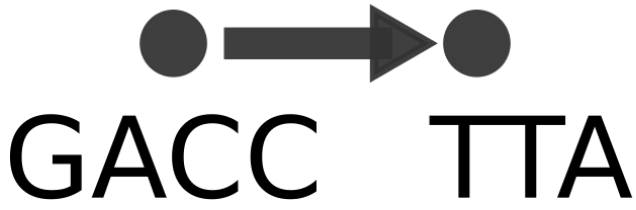
The *de novo* assembly challenge



Assembly recap

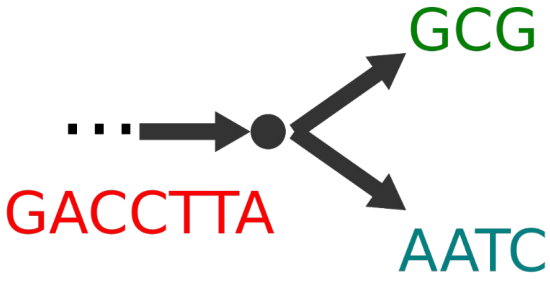
Assembly is like taking a step after another in a maze

One step is a group of nucleotides



Assembly recap

Until you have a choice to make :



why does this happen? check the reads:

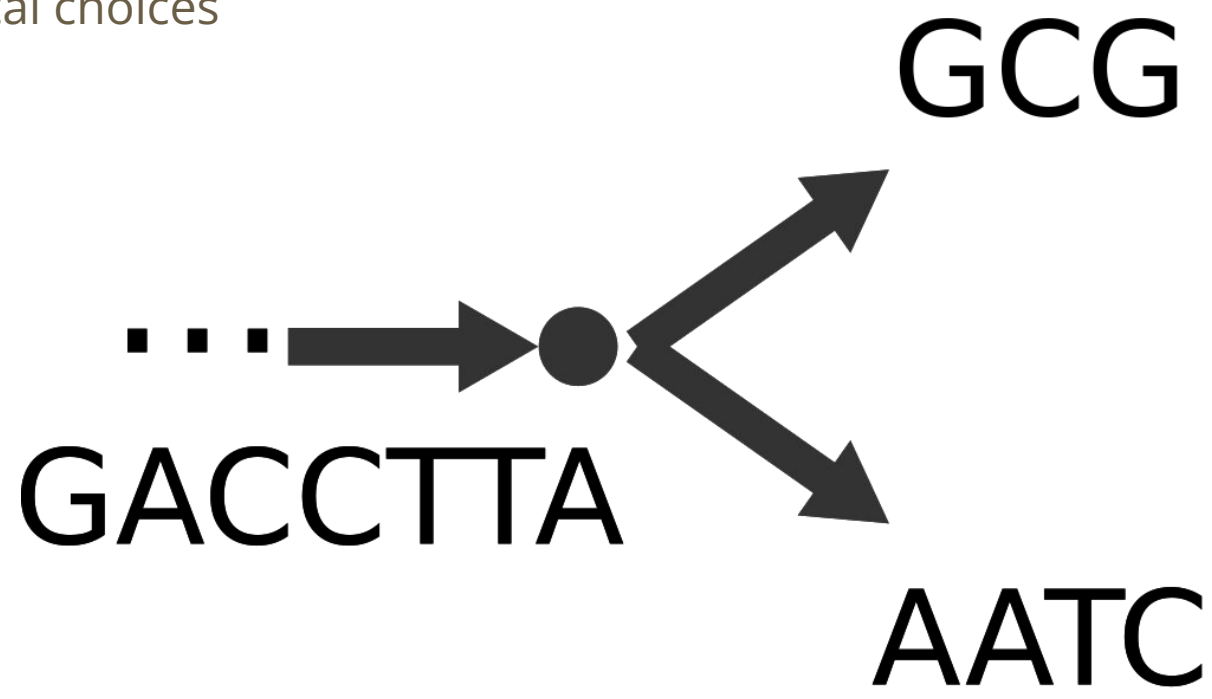
CTTAGCG

TTAAATC

and in the initial molecules, an exon is shared:

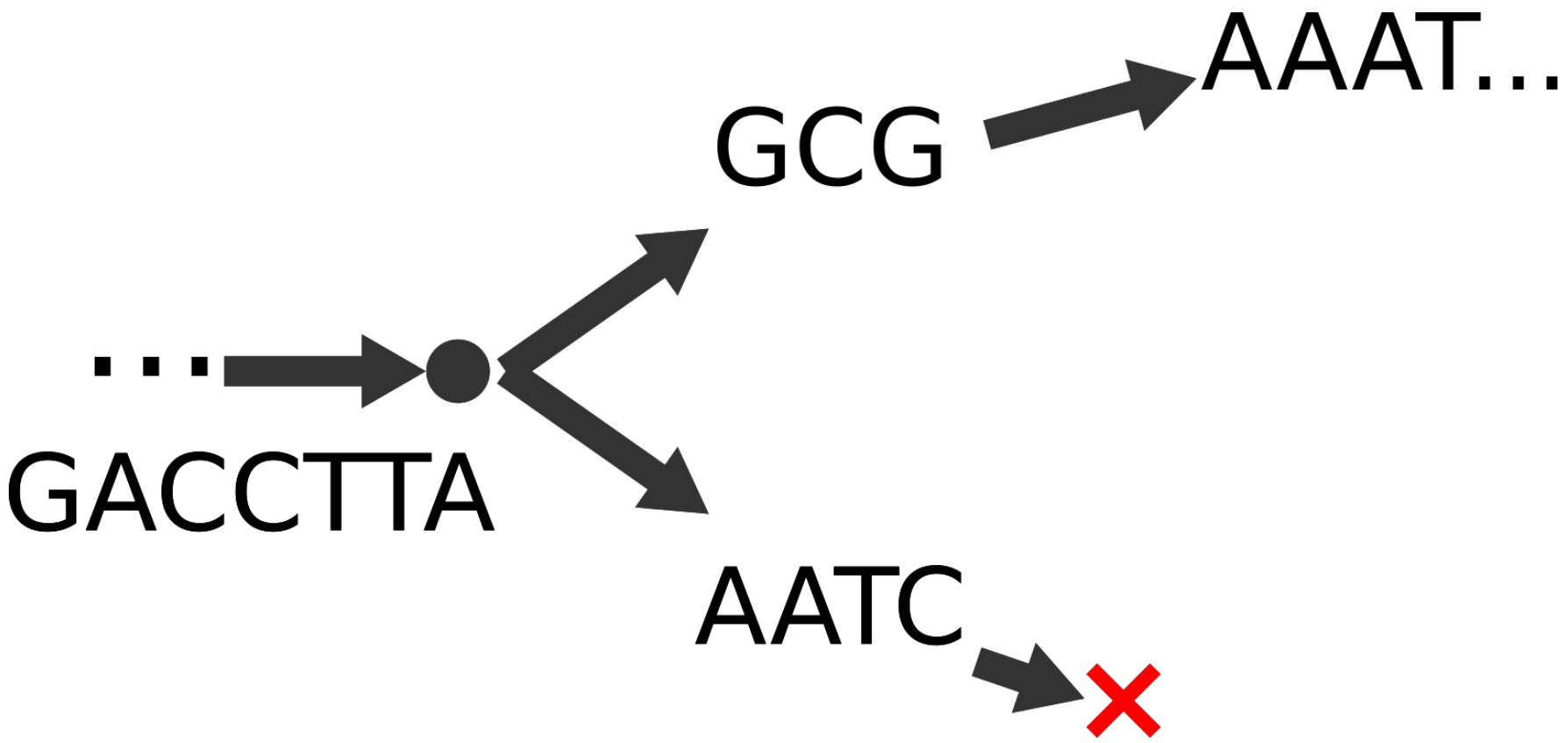
Greedy algorithms

local choices



Greedy algorithms

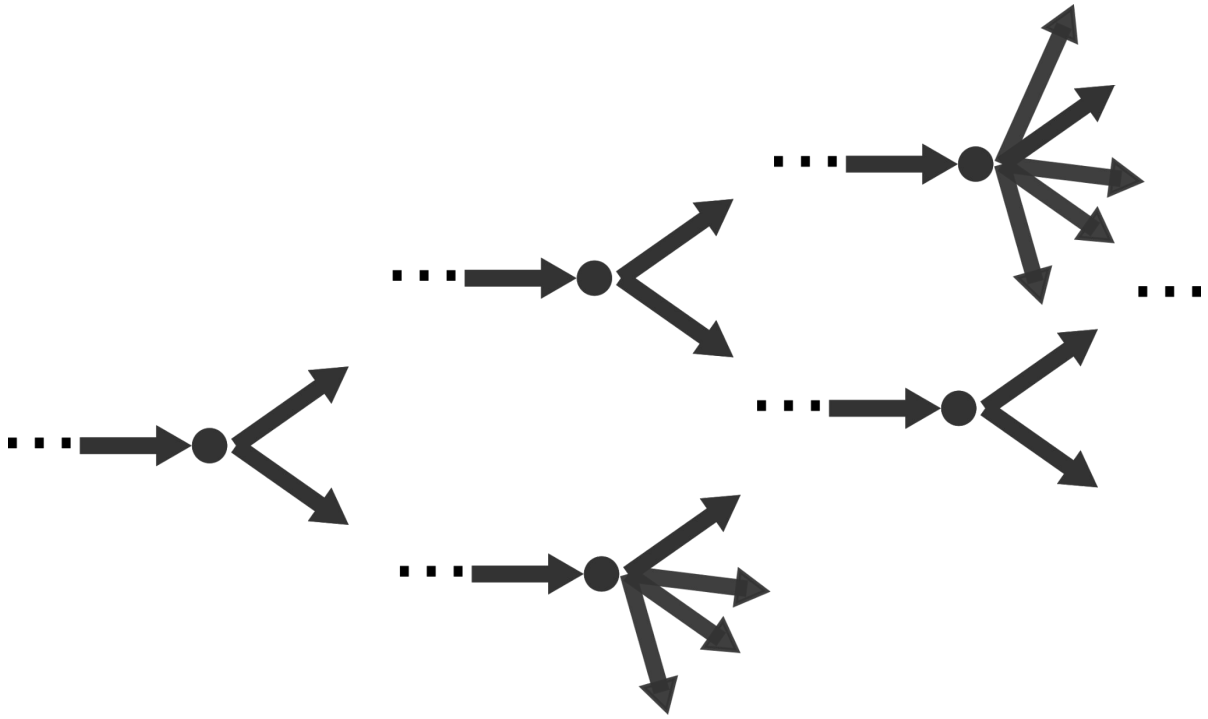
local choices can lead to bad decisions



All vs all overlaps algorithms

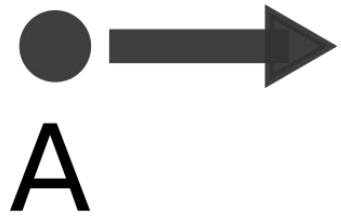
Have a global view of the possibilities in the “maze”

Ideal but... **quadratic**



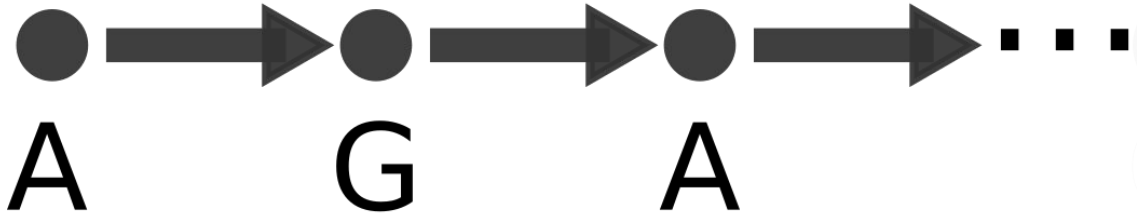
de Bruijn graph assembly

With de Bruijn graphs we walk in the maze nucleotide by nucleotide:



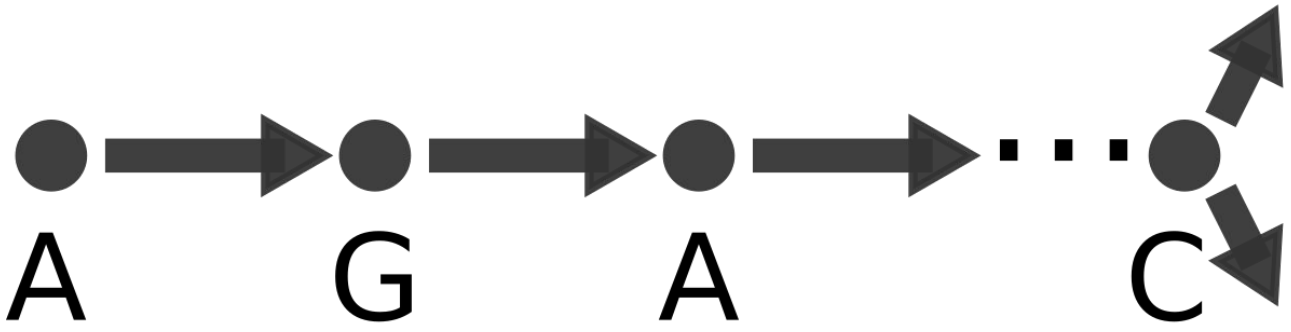
de Bruijn graph assembly

Your next step must correspond to the nucleotide that comes after in the original transcript



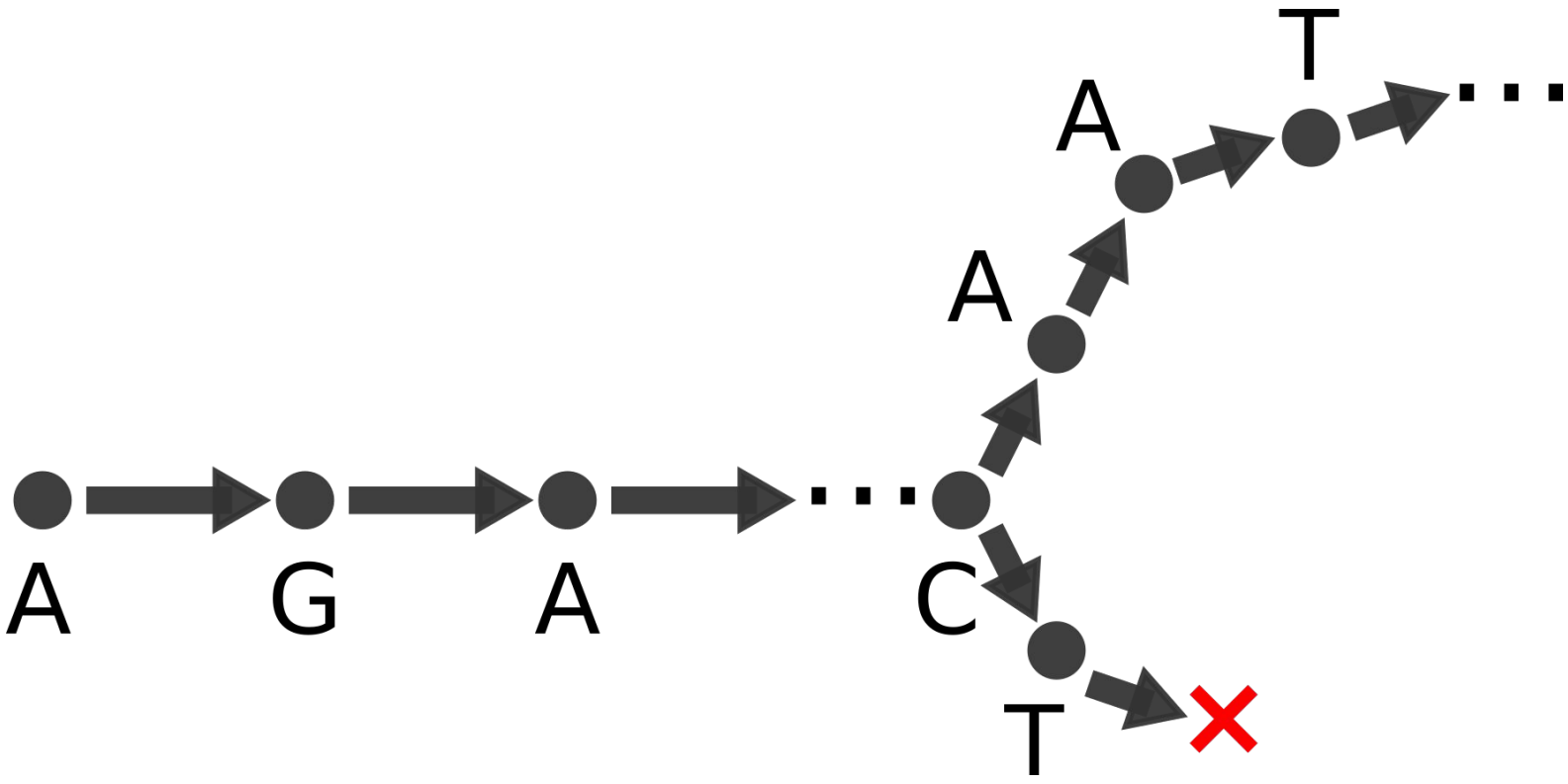
Result: concatenation of the nucleotides (AGA...)

de Bruijn graph assembly



de Bruijn graph assembly

Some dead ends and other bifurcations can be seen



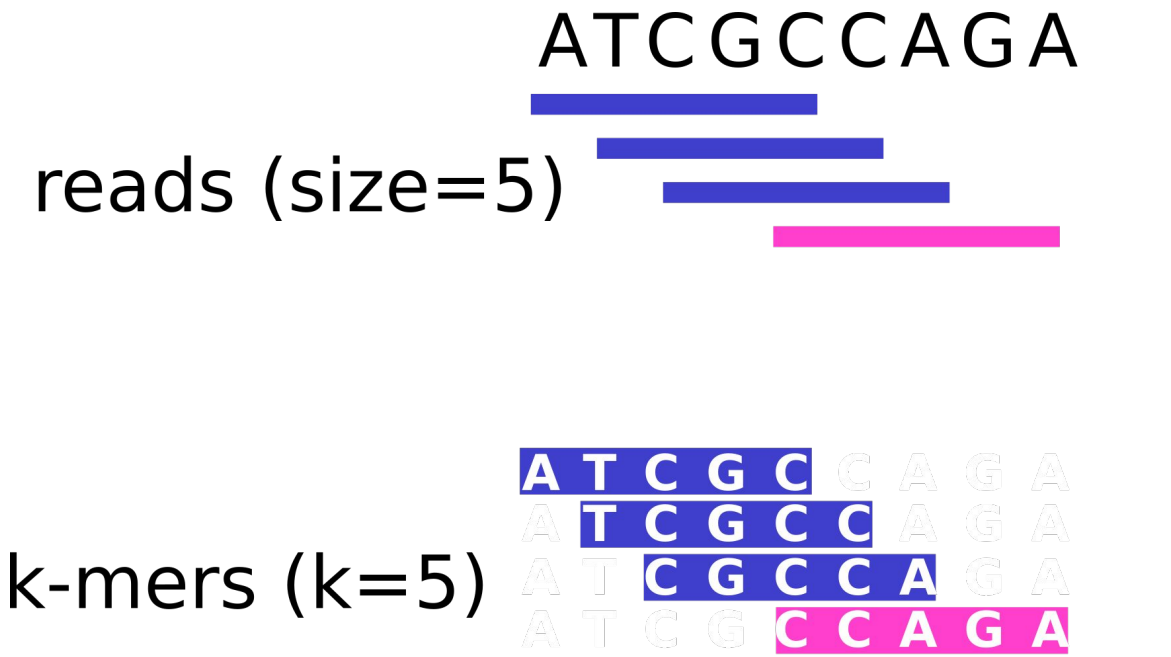
de Bruijn graph assembly

Store the “maze” in a graph structure (de Bruijn graph)

- helps with local choices
- cost efficient (RAM & runtime)

de Bruijn graph in practice: k-mers

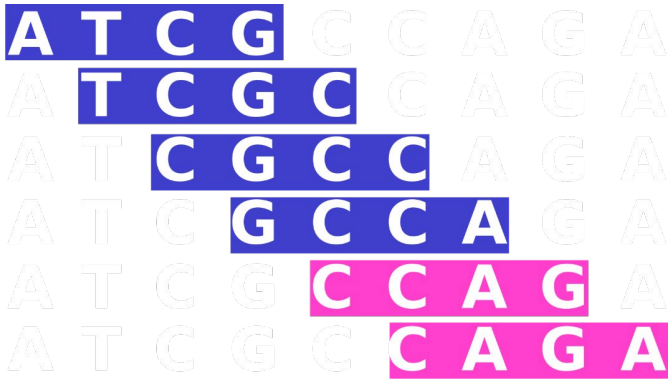
k-mers: why don't we use reads



result: ATCGCCA, CCAGA

de Bruijn graph in practice: k-mers

k-mers (k=4)



result: ATCGCCAGA

de Bruijn graph in practice: k-mers

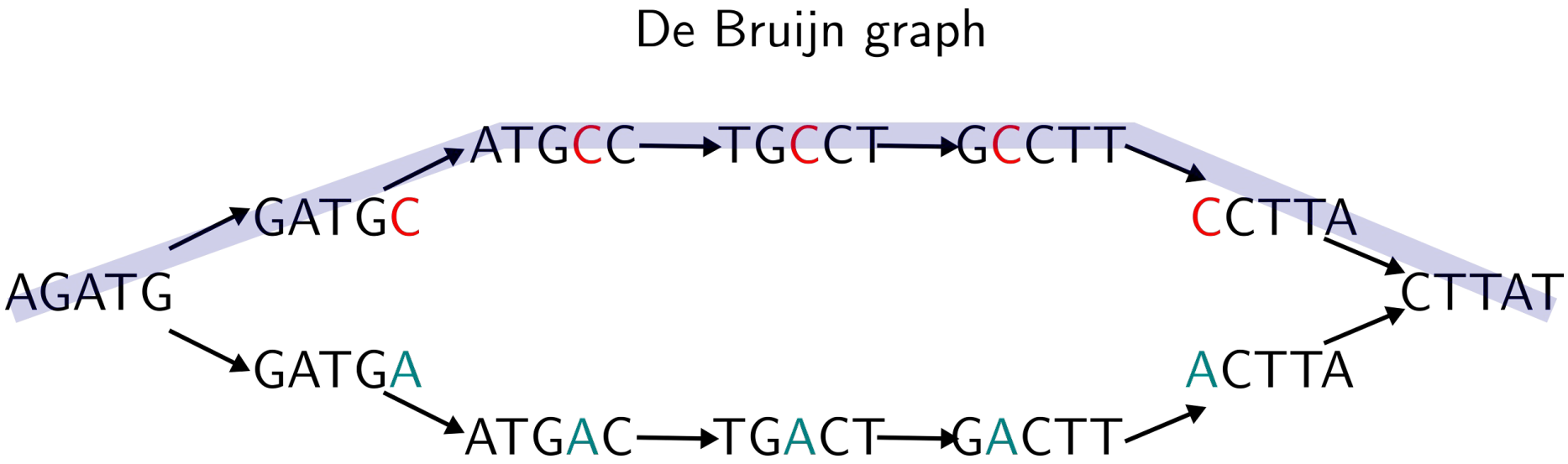
k-mers help bridging the assembly

they are key elements to work with the dBG

in practice implementations allow using several k sizes

tradeoff larger k: more conservative /smaller k: more gaps filled in the graph

Path in the De Bruijn graph

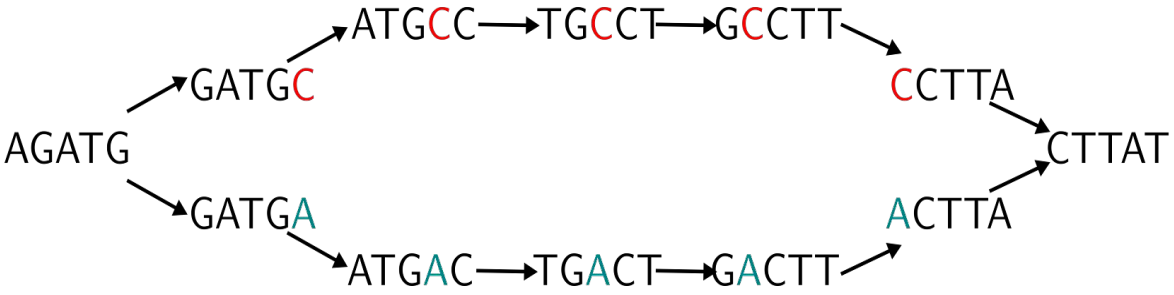


assembly : a set of gap-less sequences extracted from paths covering the graph (after some modifications to the graph...)

Vocabulary: bubbles/bulges

AGATGCCTTAT

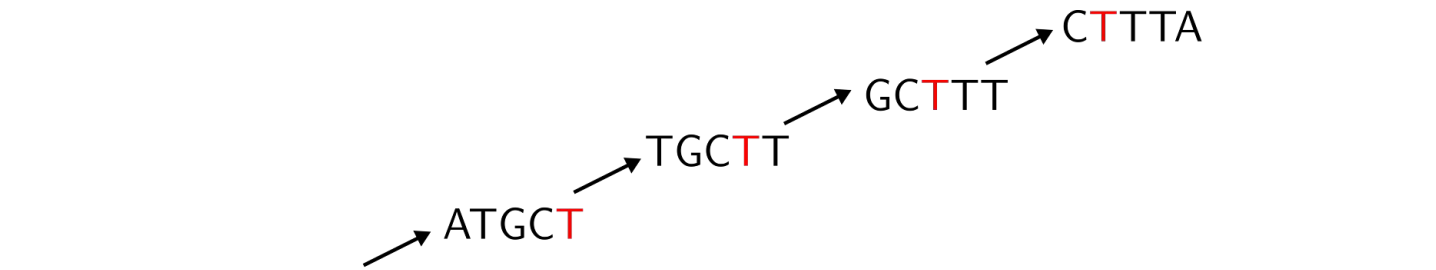
AGATGCCTTAT
AGATGACTTAT

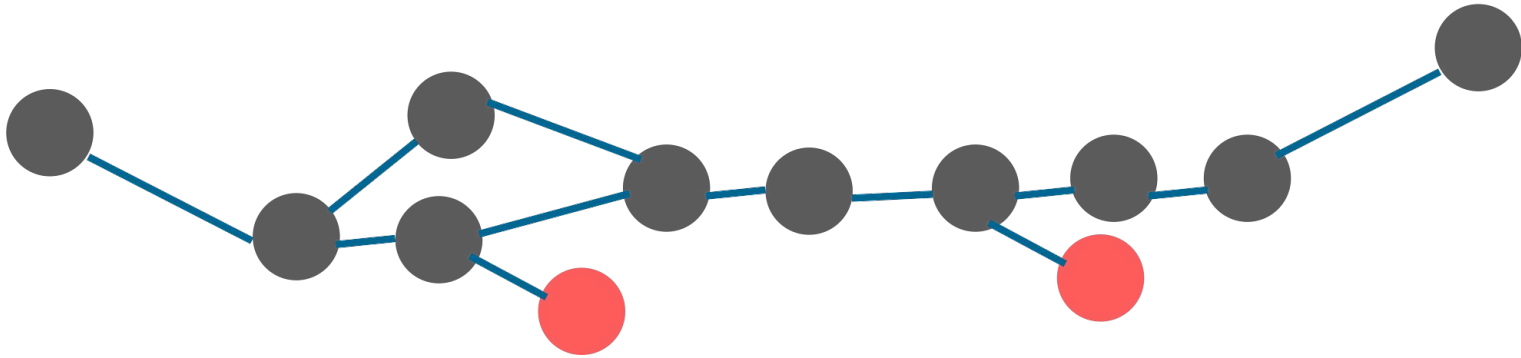


Vocabulary: tips/dead ends

reads

AGATGCCTTA
AGATGCTTTA
AGATGCCTTA
GATGCCTTAT
GATGCCTTAT

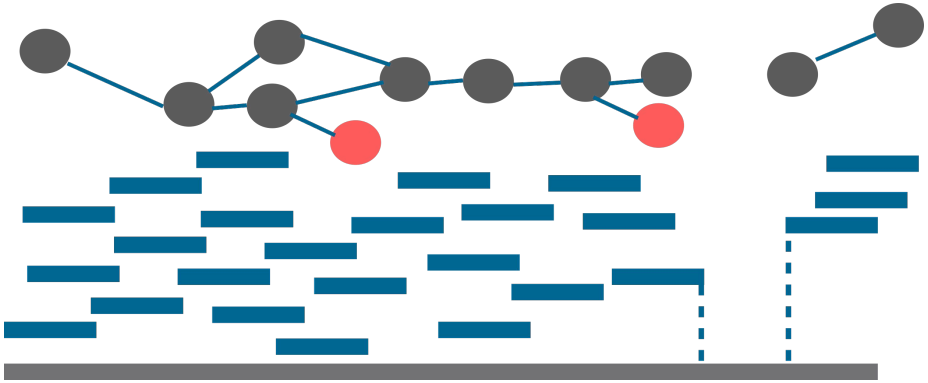




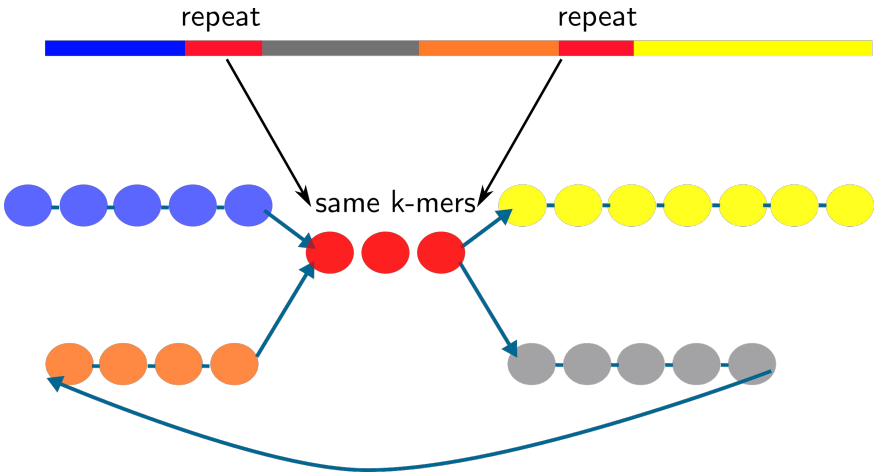
An assembly generally is

- smaller than the reference,
- fragmented

- missing reads create gaps



- repeats fragment assemblies and reduce total size



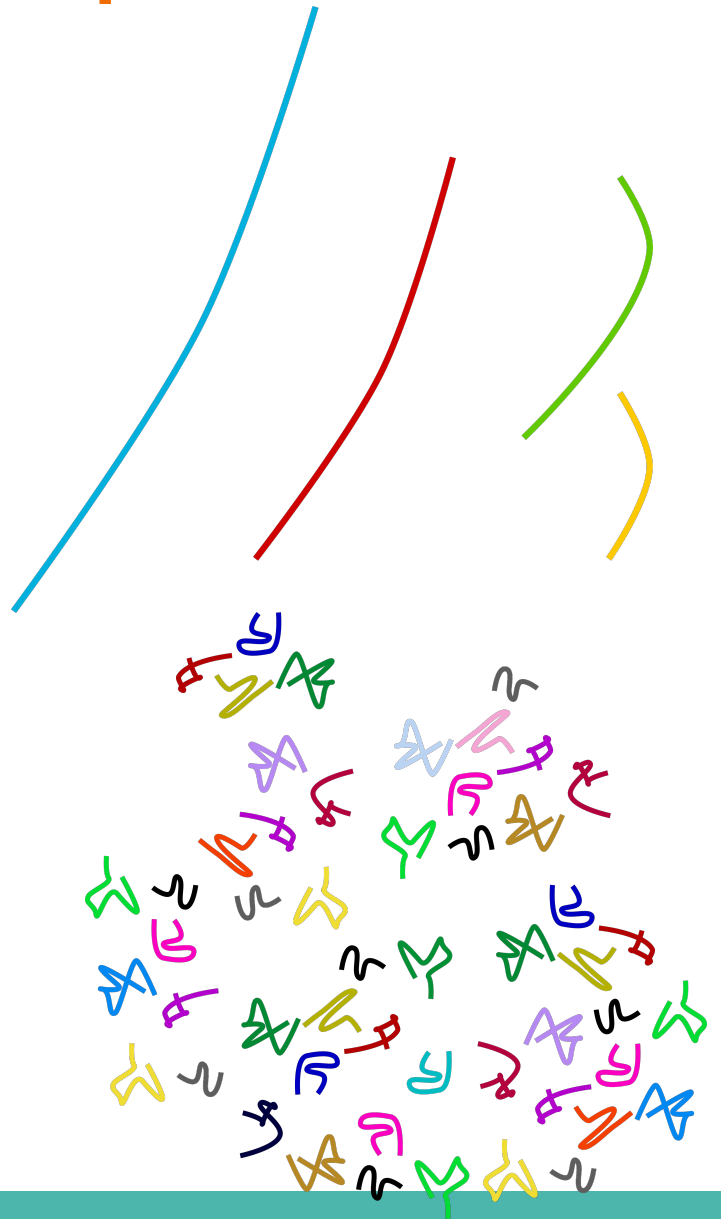
Contrasting genome and transcriptome assemblies

genome

-uniform coverage
-single contig per locus
-double stranded
-theory: one massive graph per chromosome
-practice: repeats aggregate, contigs smaller than chromosomes

transcriptome

-exponentially distributed coverage
-multiple contigs per locus
-strand specific
- theory: thousands of small disjoint graphs, one per gene
-practice: gene families, ALU & TE, low covered

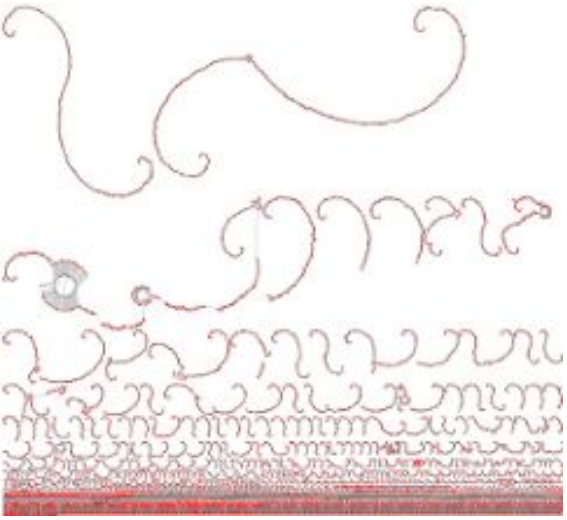


Contrasting genome and transcriptome assemblies

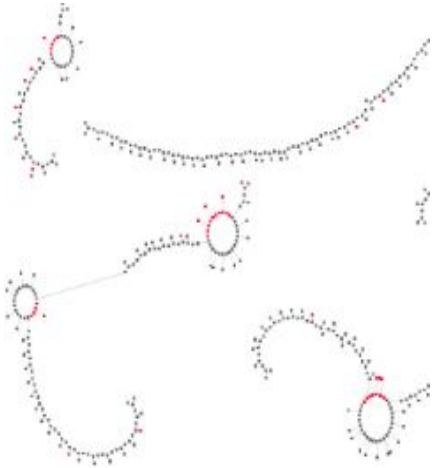
Despite these differences, DNA-seq assembly methods apply:

- Construct a de Bruijn graph (same as DNA)
- Output contigs (same as DNA)
- Allow to re-use the same contig in many different transcripts (new part)

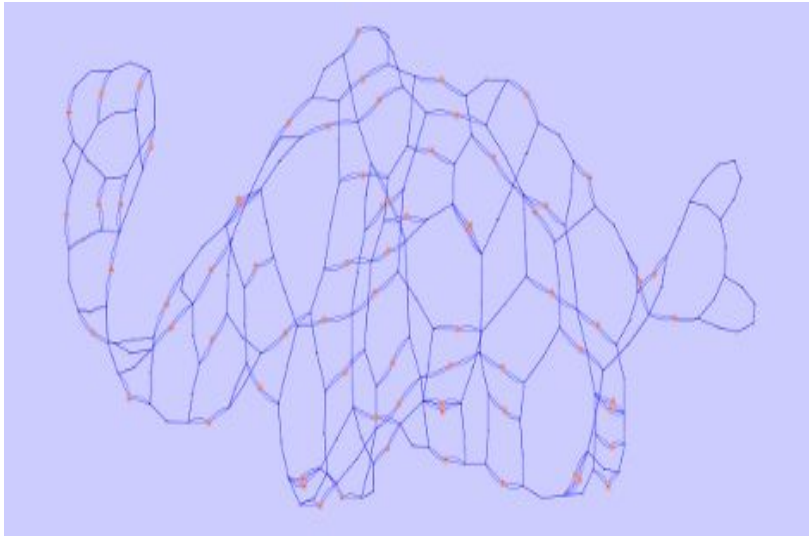
Real instance graphs



graph from shallow covered Drosophila dataset



zoomed-in bubbles (+ tips)



gene family

Credit: ERABLE team (Lyon)

There is no single solution for assembly...

Conclusions of the GAGE benchmark : in terms of assembly quality, there is no single best assembler. Applies to RNA-seq.

Main tools:

-**TransAbyss**, Robertson et al. *Nat. Met* 2010 https://github.com/bcgsc/transabyss

-**Bridger**, Chang et al. *Genome Biol.* 2015 https://github.com/fmaguire/Bridger_Assembler

-**SOAPdenovo-Trans**, Xie et al. *Bioinformatics* 2014
https://github.com/aquaskyline/SOAPdenovo2

-**Trinity**, Grabherr et al. *Nat. Biotechnol.* 2011
https://github.com/trinityrnaseq/trinityrnaseq/wiki

- **rnaSPAdes**, Bushmanov et al. *GigaScience* 2019 http://cab.spbu.ru/software/spades/

The main building blocks in theory

1. (optional) correct the reads (for instance BayesHammer in rnaSPAdes)
2. build a graph from the reads (remove k-mers seen once)
3. remove likely sequencing errors (tips)
4. remove known patterns (bubbles)
5. return simple paths (i.e. contigs), **allow nodes to be used several times**

Warning: what's in the paper is different than what's in the implementation…

2. Assembly in SPAdes: An Outline

Go to:

Below we outline the four stages of SPAdes, which deal with issues that are particularly troublesome in SCS: sequencing errors; non-uniform coverage; insert size variation; and chimeric reads and bireads:

(1) Stage 1 (assembly graph construction) is addressed by every NGS assembler and is often referred to as de Bruijn graph *simplification* (e.g., *bulge/bubble* removal in EULER/Velvet). We propose a new approach to assembly graph construction that uses the *multisized de Bruijn graph*, implements new bulge/tip removal algorithms, detects and removes chimeric reads, aggregates biread information into *distance histograms*, and allows one to backtrack the performed graph operations.

(2) Stage 2 (*k*-bimer adjustment) derives accurate distance estimates between *k*-mers in the genome (edges in the assembly graph) using joint analysis of distance histograms and paths in the assembly graph.

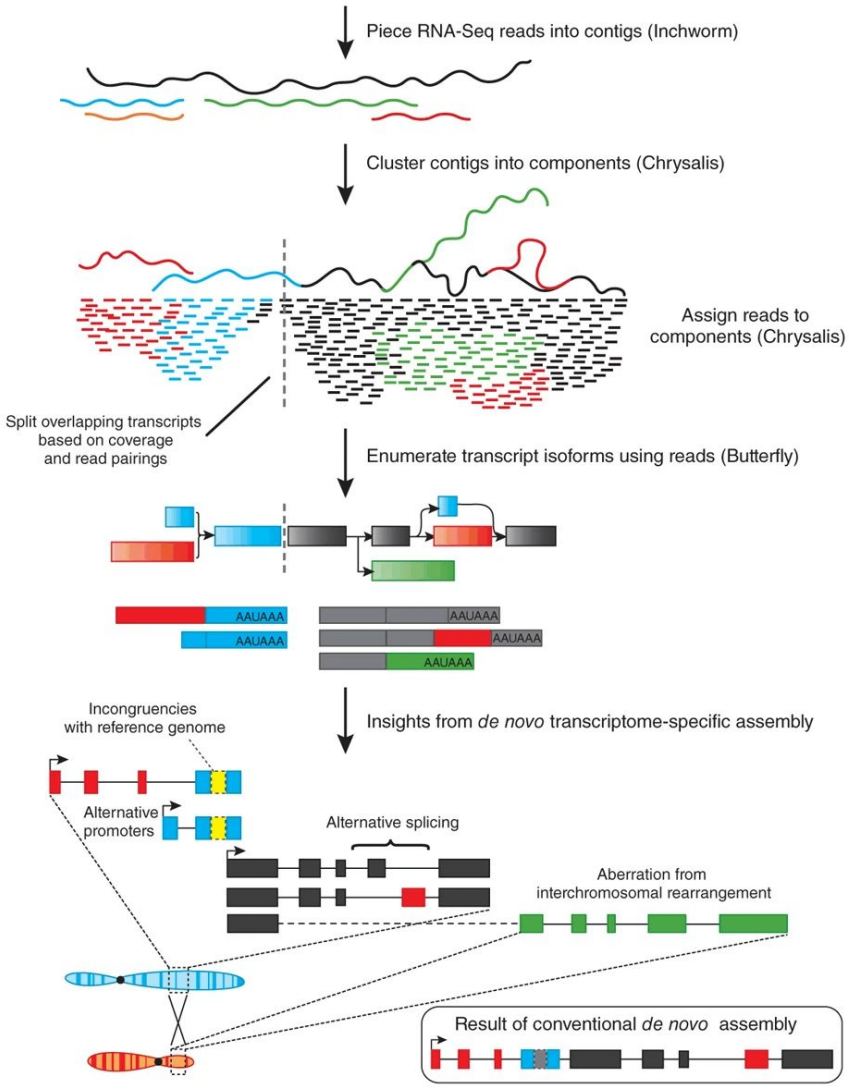
Trinity assembler

- Inchworm de Bruijn graph construction, part 1

- Chrysalis de Bruijn graph construction, part 2

- Butterfly Graph traversal using reads, isoforms enumeration

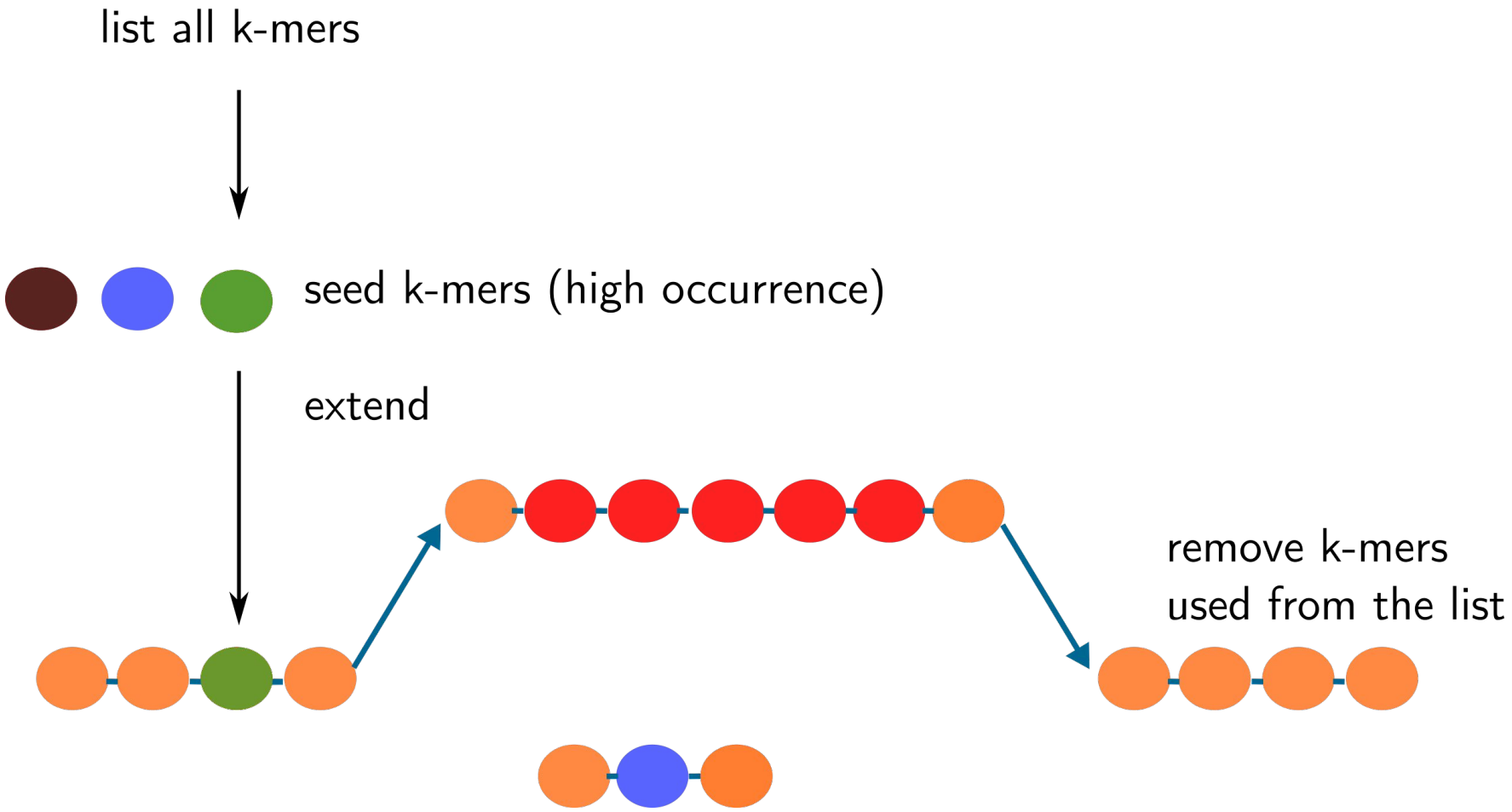
Trinity overall



Iyer et al, Nature Biotech., 2011

Trinity: detail

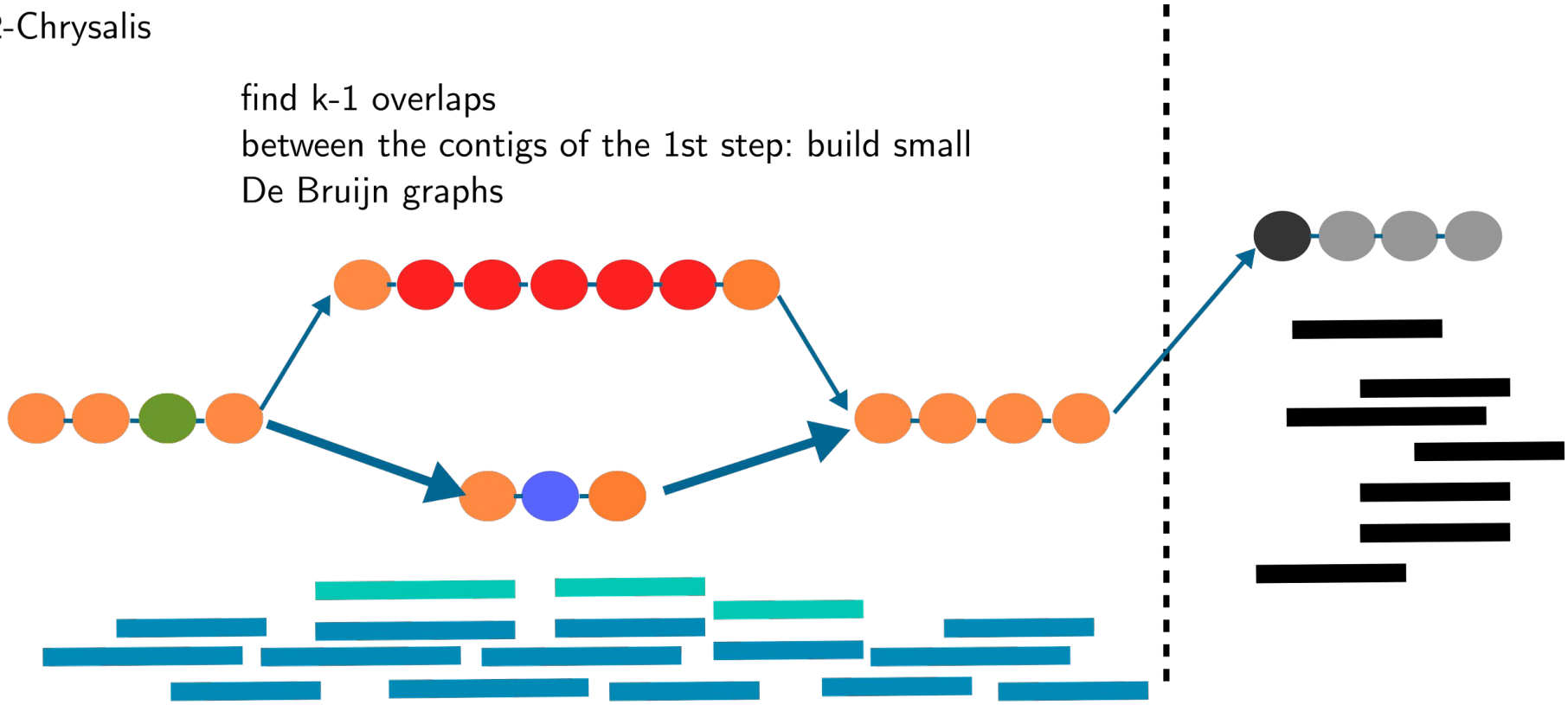
1-Inchworm



Trinity: detail

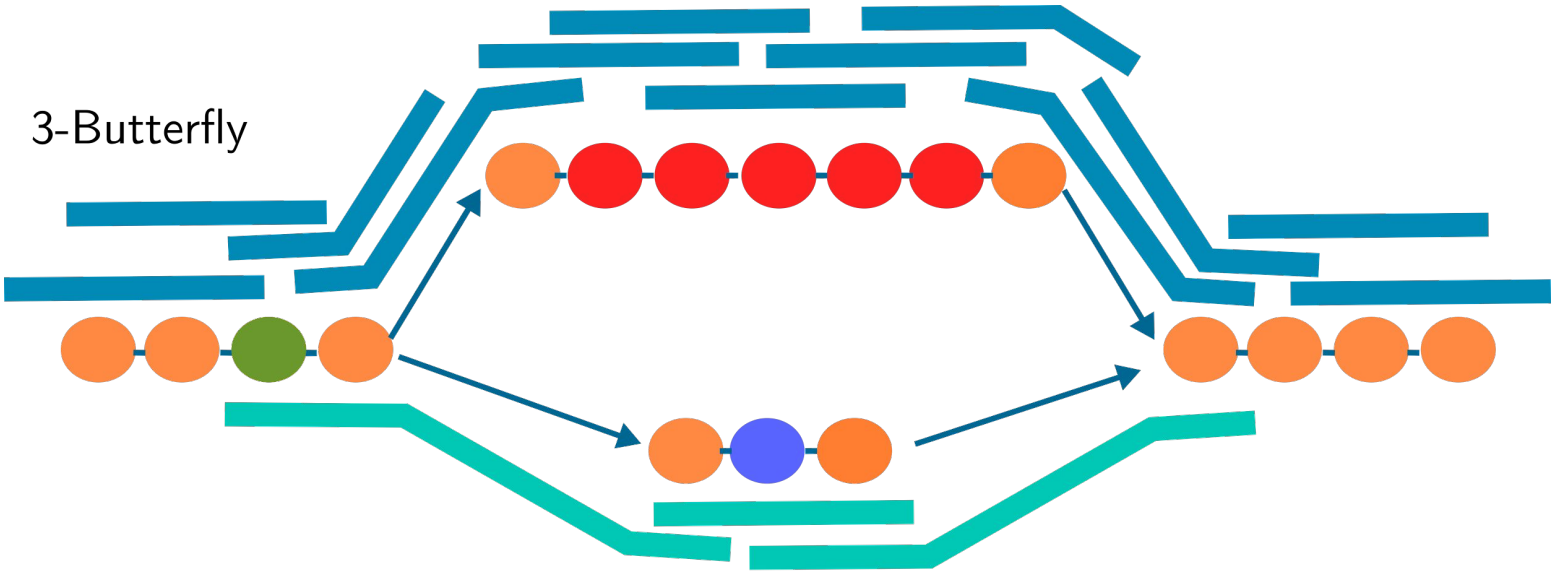
2-Chrysalis

find k-1 overlaps
between the contigs of the 1st step: build small
De Bruijn graphs

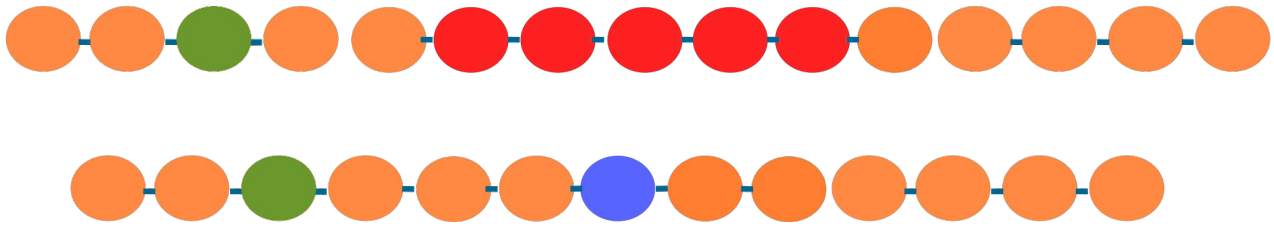


use read mapping information to separate clusters

Trinity: detail



output read-coherent isoforms



Trinity output

```
>TRINITY_DN1000_c115_g5_i1 len=247 path=[31015:0-148 23018:149-246]
 AATCTTTTTTGGTATTGGCAGTACTGTGCTCTGGGTAGTGATTAGGGCAAAAGAAGACAC
 ACAATAAAGAACCAGGTGTTAGACGTCAGCAAGTCAAGGCCTTGGTTCTCAGCAGACAGA
 AGACAGCCCTTCTCAATCCTCATCCCTTCCCTGAACAGACATGTCTTCTGCAAGCTTCTC
 CAAGTCAGTTGTTCACAGGAACATCATCAGAATAAATTTGAAATTATGATTAGTATCTGA
 TAAAGCA
```

-Trinity read cluster 'TRINITY_DN1000_c115'

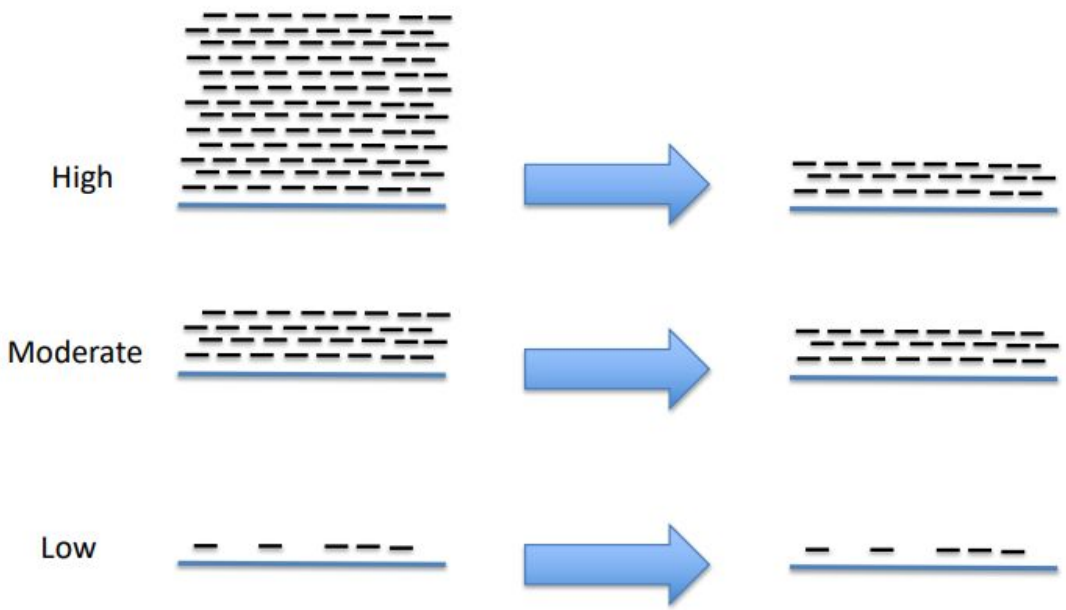
- gene 'g5'

- isoform 'i1'

-path=[31015:0-148 23018:149-246]") indicates the path traversed in the Trinity de Bruijn graph to construct that transcript

Normalization effects on assembly (example of Trinity)

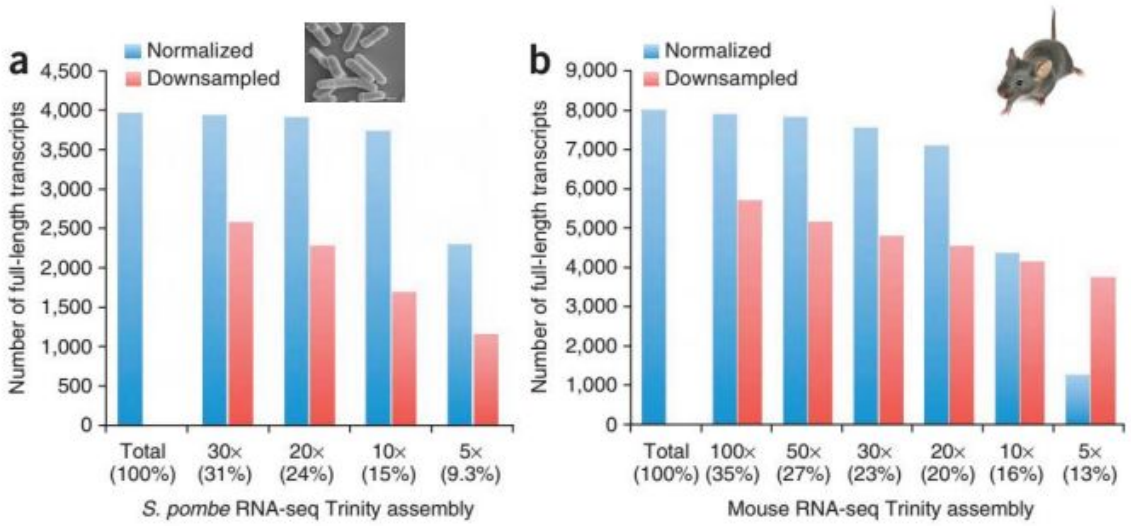
From Brian Haas



Normalization effects on assembly (example of Trinity)

From Brian Haas

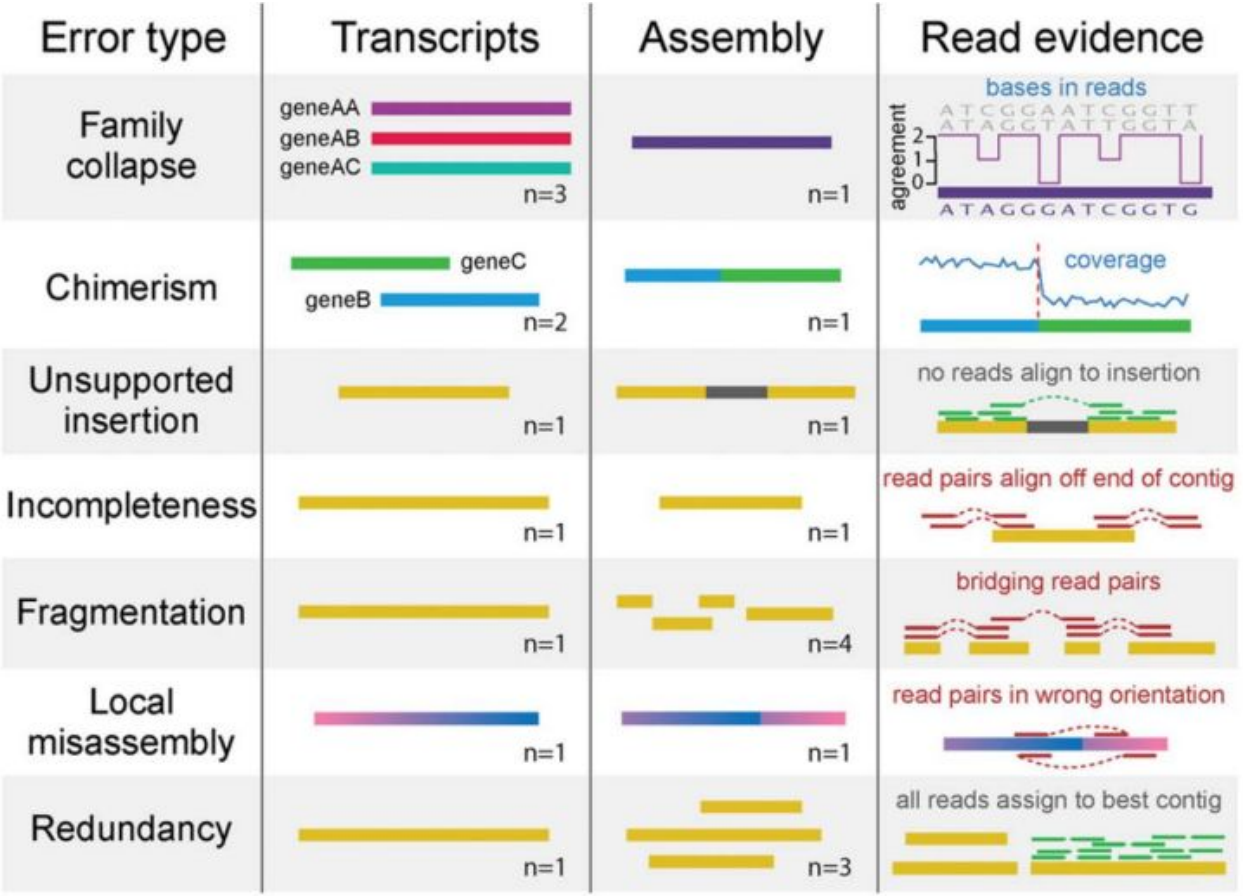
Impact of Normalization on *De novo* Full-length Transcript Reconstruction



Largely retain full-length reconstruction, but use less RAM and assemble much faster.

Total (100%) 30× (31%) 20× (24%) 10× (15%) 5× (9.3%)
S. pombe RNA-seq Trinity assembly
Total (100%) 100× (35%) 50× (27%) 30× (23%) 20× (20%) 10× (16%) 5× (13%)
Mouse RNA-seq Trinity assembly

Errors made by assemblers



Smith-Unna et al. Genome Research, 2016

Assembly quality assessment

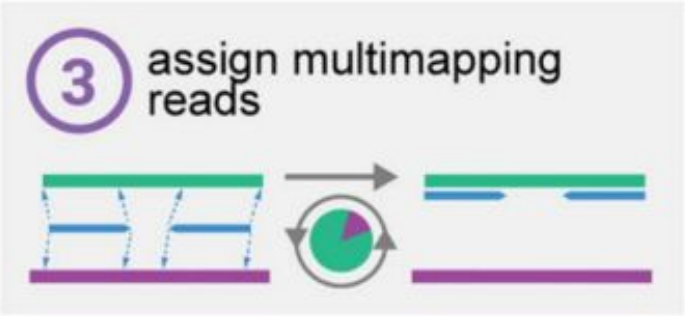
In transcriptome assemblies

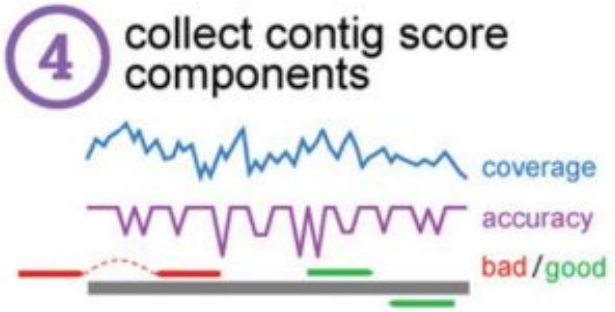
- N50 is not very useful.
  - unreasonable isoform annotation for long transcripts drives higher N50
  - very sensitive reconstruction for short lowly expressed transcripts leads to lower N50

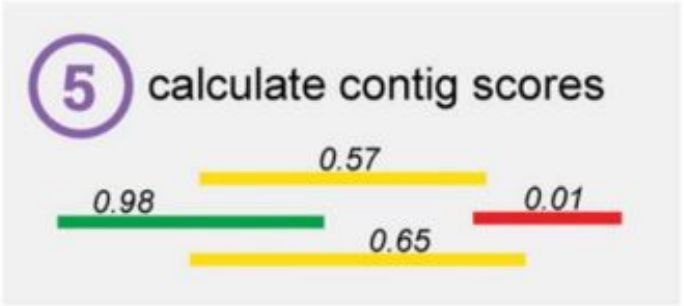
95%-assembled isoforms statistics
reference-free evaluation must be preferred
read remapping

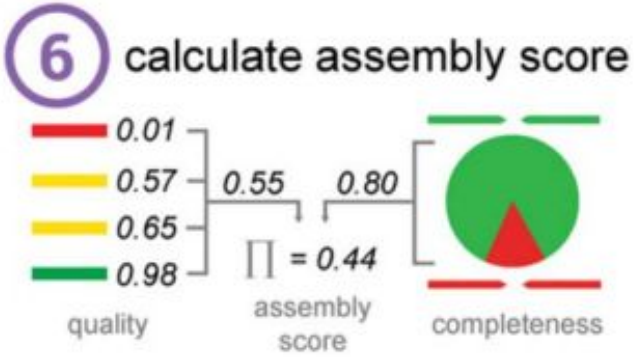
Main tools:
- rnaQuast http://cab.spbu.ru/software/rnaquast/
- Transrate http://hibberdlab.com/transrate/

TransRate





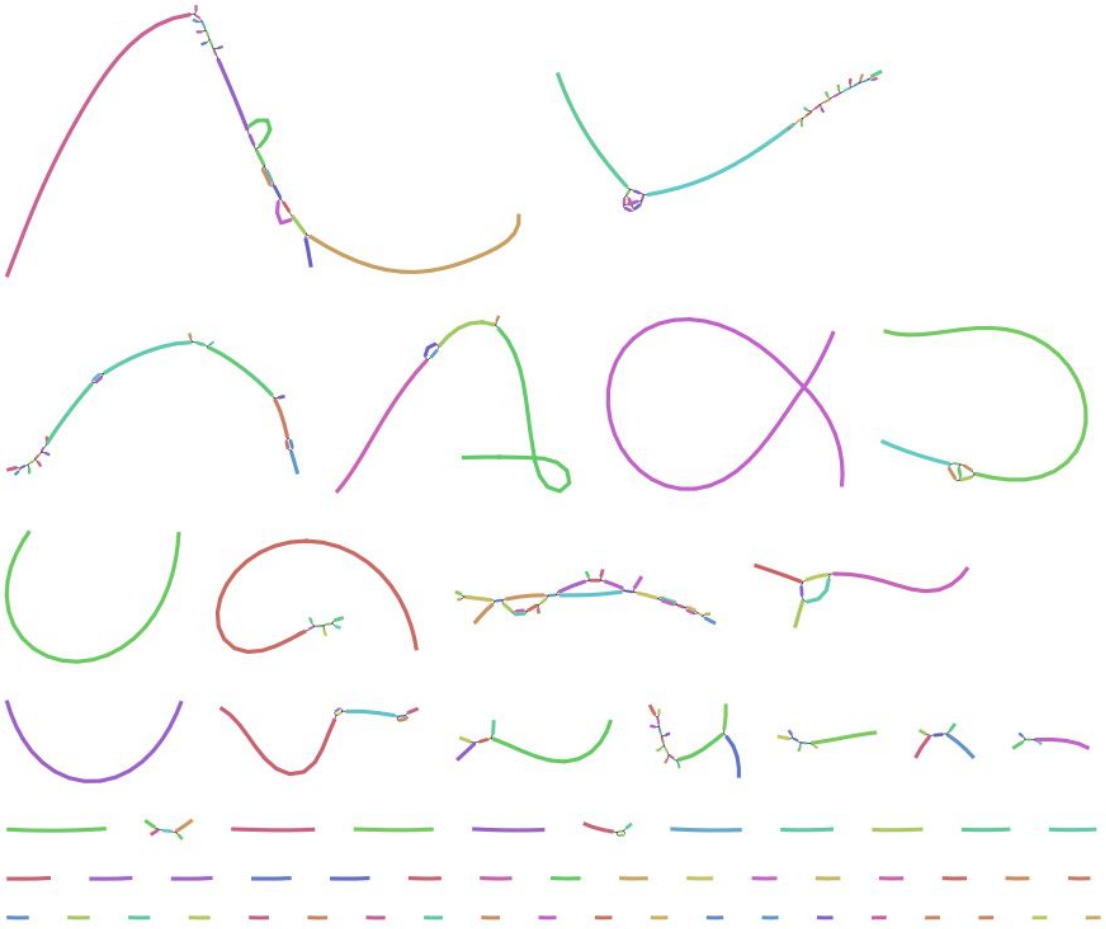




Smith-Unna et al. Genome Research, 2016

Visualization: Bandage

https://rrwick.github.io/Bandage/



Meta-practices

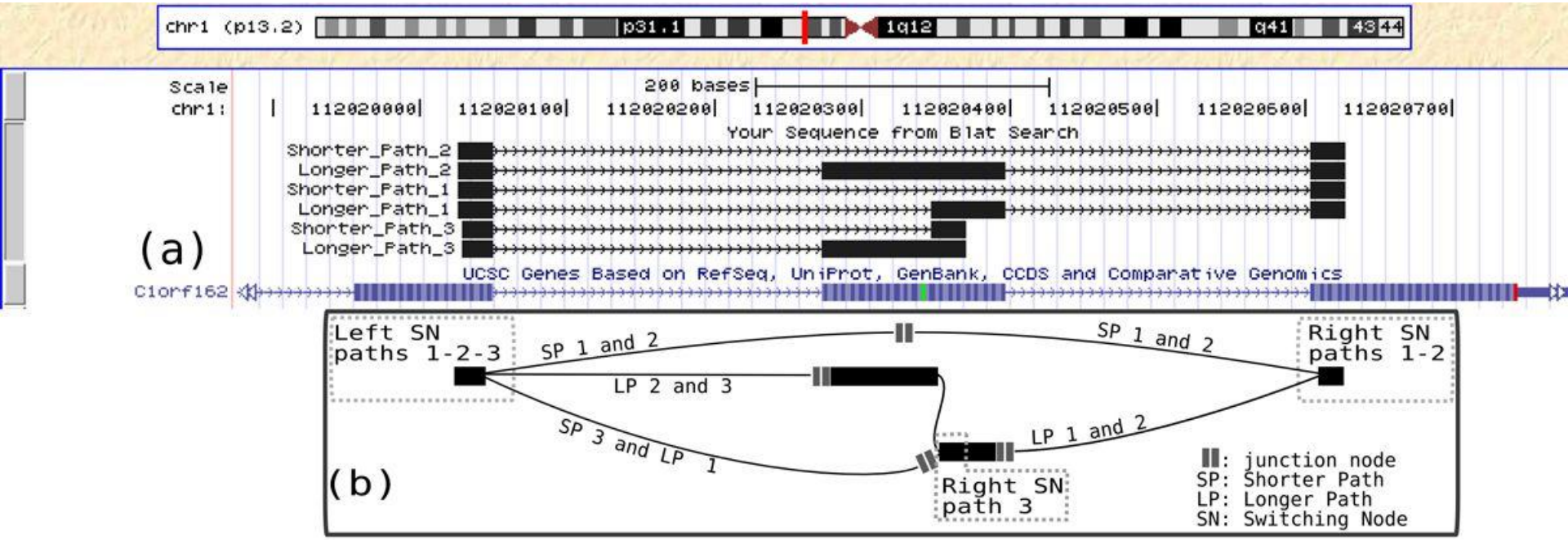
1- Read surveys, Twitter, blogs
2. Pick two assemblers
3. Run each assembler at least two times (different parameters)
4. Compare assemblies
5. If possible, visualize them

An assembly is not the absolute truth, it is a mostly complete, generally fragmented and mostly accurate hypothesis

Currently, Trinity, RNASpades and TransAbyss could be pointed as the most trustworthy/qualitative
(for known species. Not one tool for all issues).

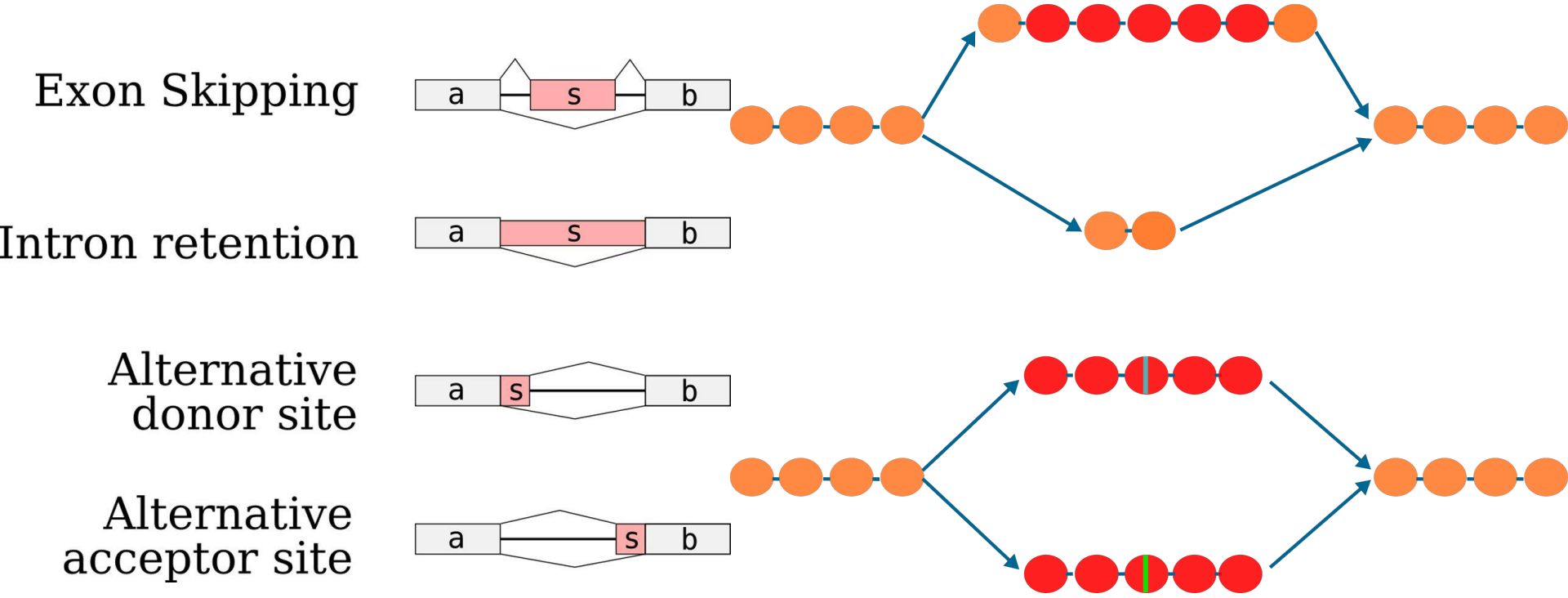
Practical: Trinity assembly

Assembly does not output all variants

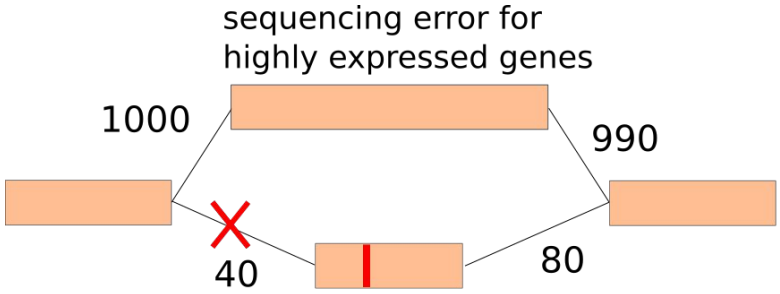


KISSPLICE

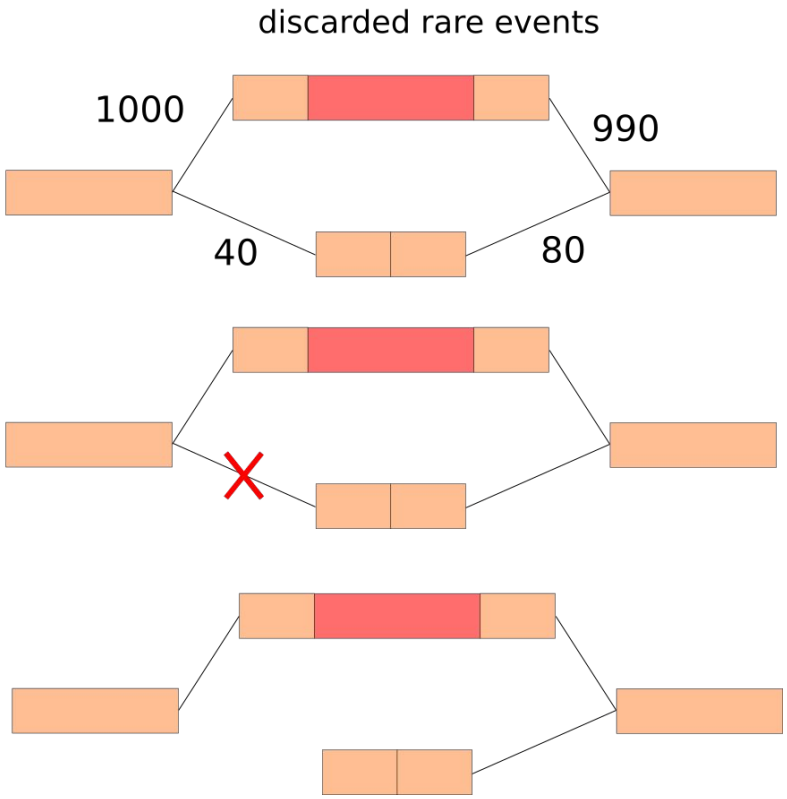
Goal: instead of assembling full-length transcripts, KISSPLICE (Sacomoto et al. 2012) focuses on assembling ONLY the **bubbles** that contain events and **enumerate** the maximum of them



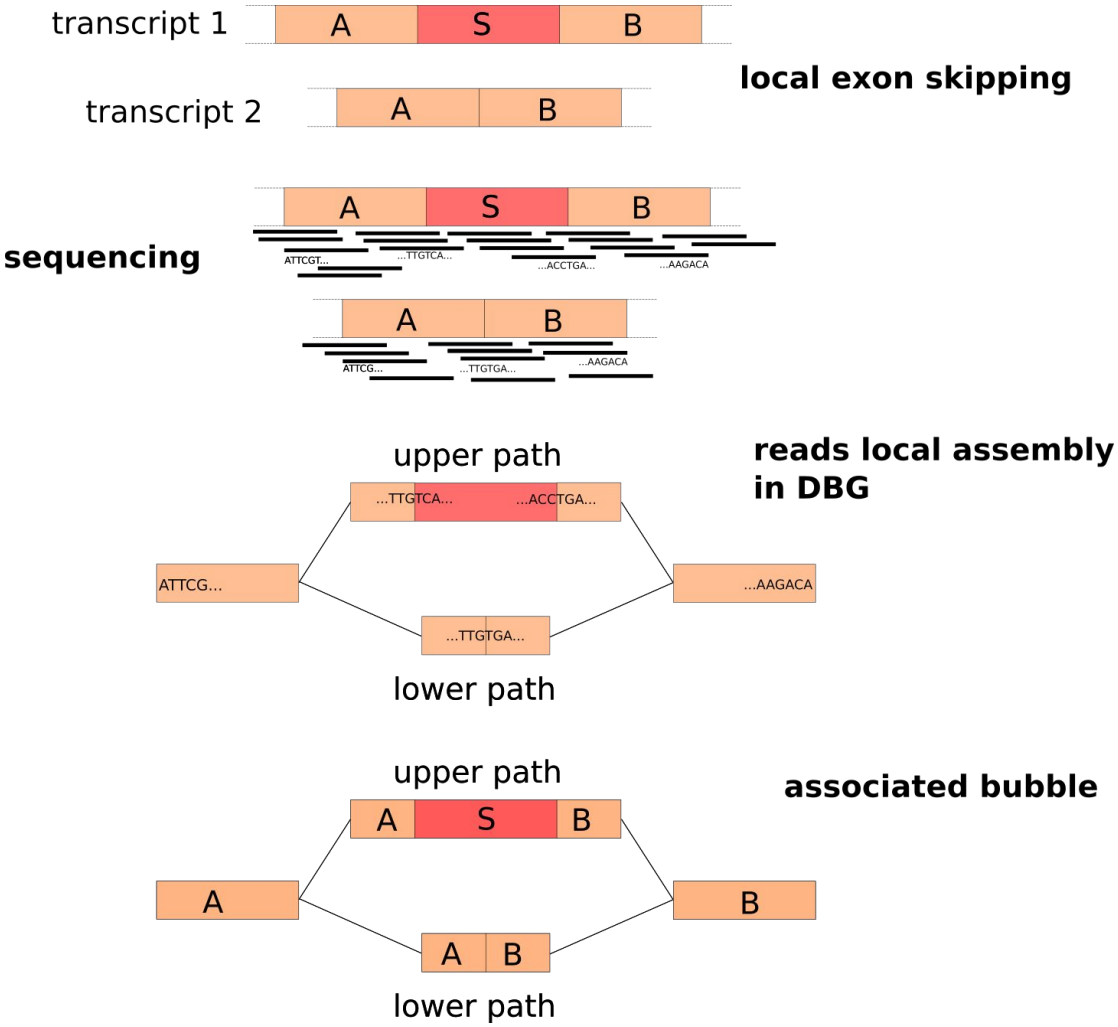
KISSPLICE: graph cleaning + local assembly



example: discard if ratio is <0.05

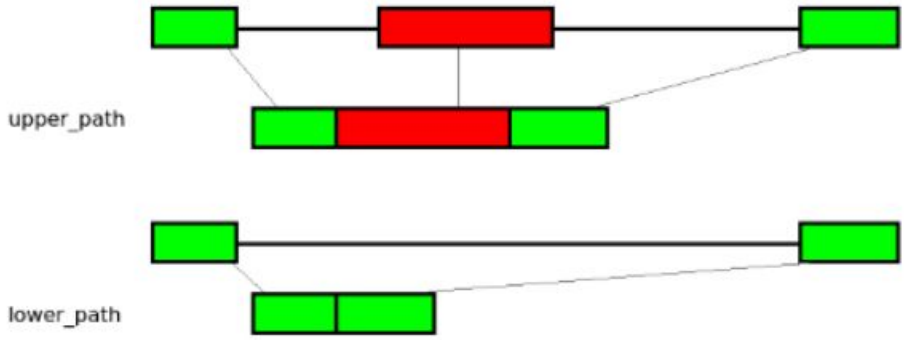


Variants in local assembly



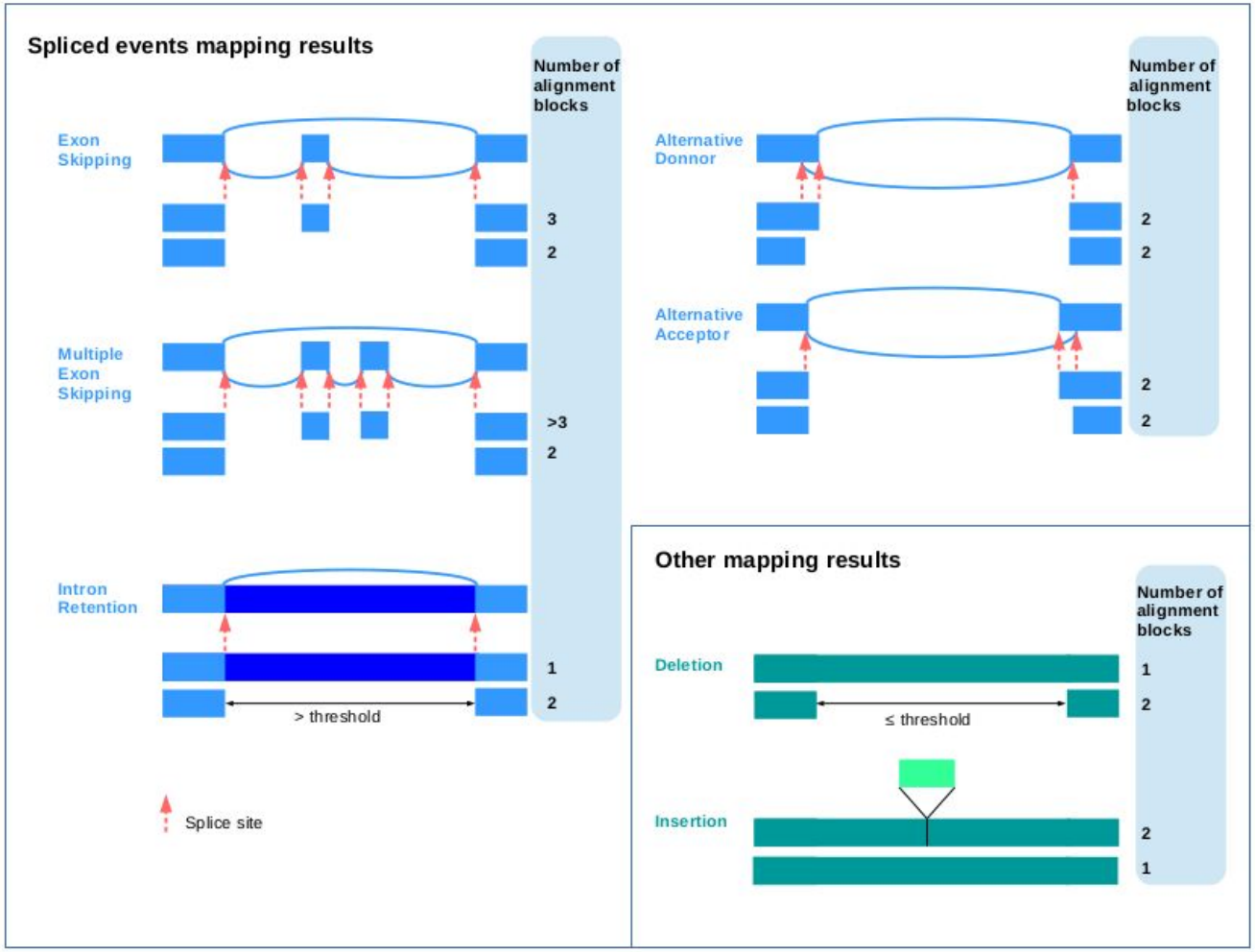
KISSPLICE's output

```
>bcc_89|Cycle_0|Type_1|upper_path_length_122|C1_0|C2_1|C3_2|C4_1|rank_0.55097
CCCTGATGGCCTCAGAGGAGGAGTAAATGTGGGGACCTAGAGGAGGAGCTGAAAATTGTTACCAACAACTTGAAATCCCTGGAGGCCCAGGCGGACAAGTATTCCACCAAAGAAGATAAATA
>bcc_89|Cycle_0|Type_1|lower_path_length_46|C1_0|C2_0|C3_2|C4_6|rank_0.55097
CCCTGATGGCCTCAGAGGAGGAGTATTCCACCAAAGAAGATAAATA
```



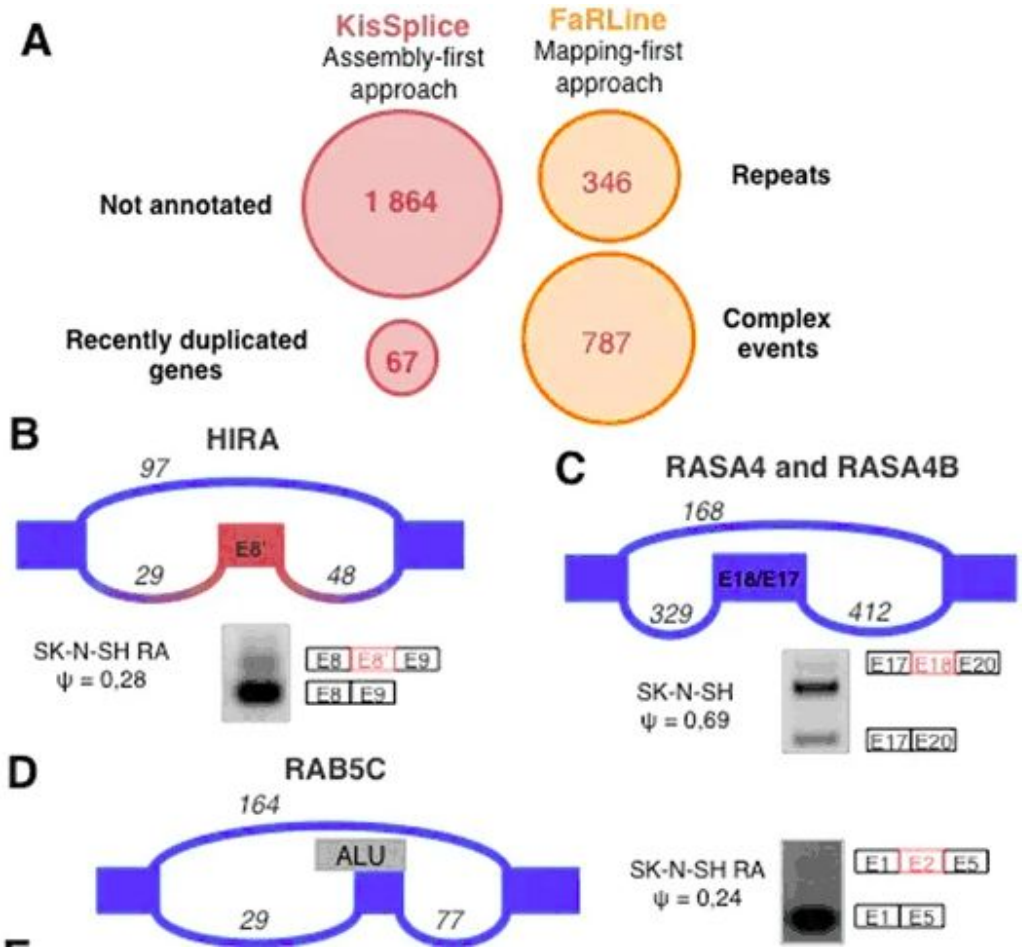
Post-processings

What do I have?	What I can use	
I have a reference genome	KisSplice2refgenome	differential analysis: kissDE
I have no reference genome	KisSplice2refTranscriptome	



for quantification only
see de-Kupl
Audoux et al. 2017

KISSPLICE case studies

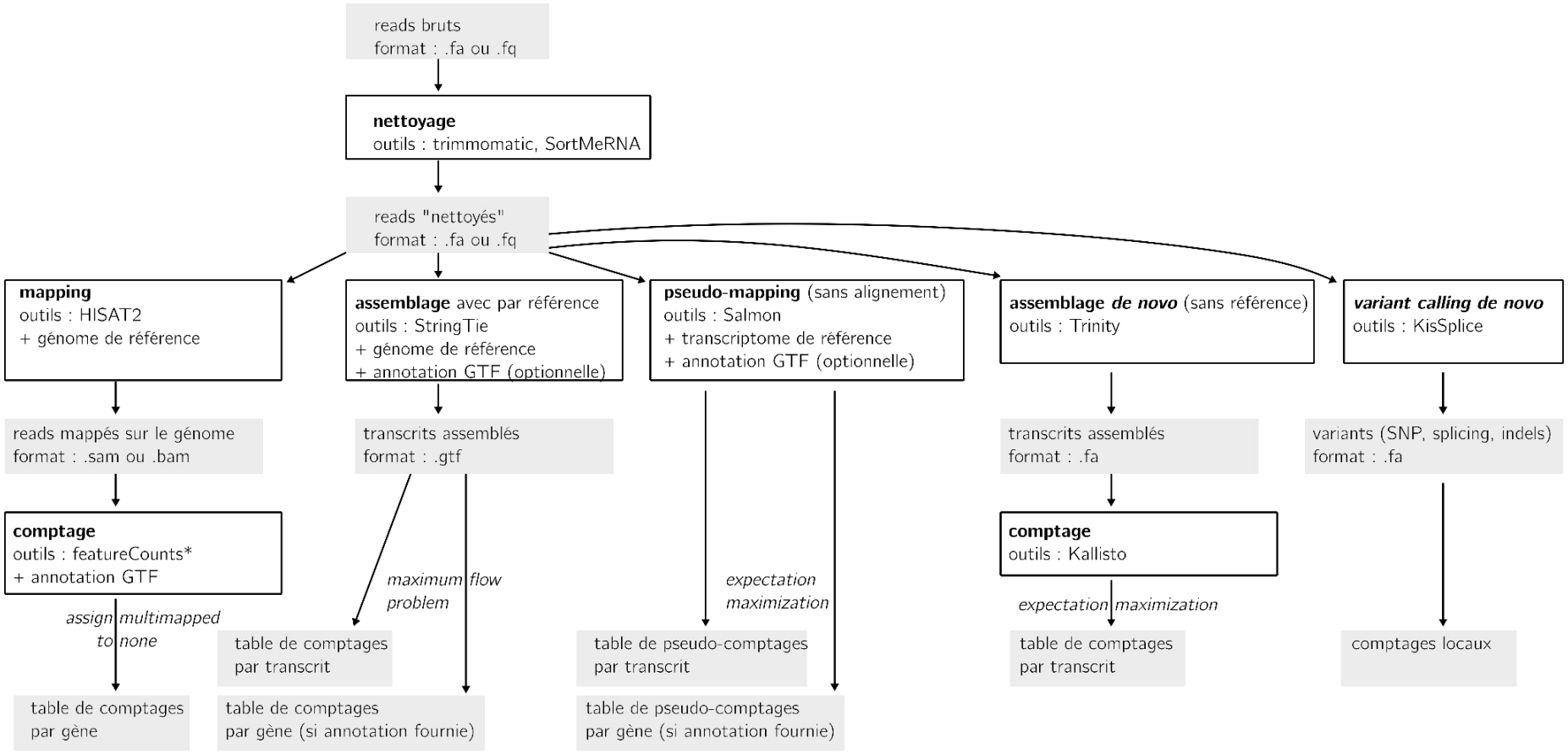


Discover splicing events:
Benoit Pilven et al. 2018

Farline: mapping
B found only by Kissplice (not annotated)
C found only by Kissplice (paralog)
D found only by mapping (Alu repeat)

Discover SNPs in pooled RNA-seq: Lopez-Maestre et al. 2016

Practical: Kissplice



* à vocation pédagogique pour le TP, déconseillé par ailleurs.

Long reads : the ~~future~~ present of transcriptomics

Long reads overview

Possibilities & pipelines

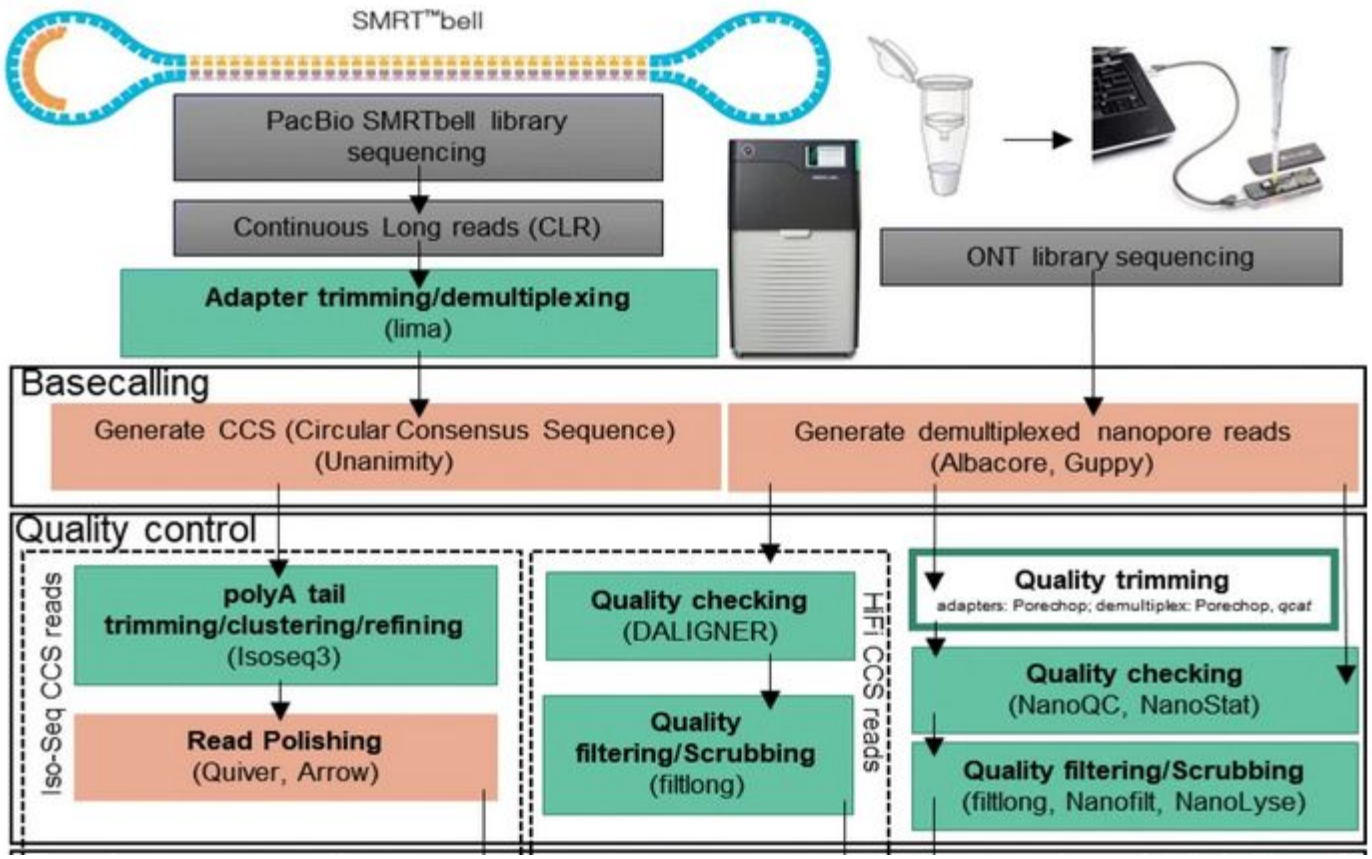
Limitations of short reads

- recent studies suggest that our reference transcriptomes **miss isoforms**
- in particular in the context of **alternative splicing**
- *de novo* assembly of species with unknown/hardly known transcriptomes is still a challenge
- the mandatory cDNA step in short reads protocols implies **bias**

Long reads technologies

- sequencing of long (>10kb) molecules is possible
  - **full RNAs !**
- with a higher (~1-5% to 14%) **error rate**
- **error profile** is different from SR: indels in **homopolymers**
- some allow to sequence directly RNA (reduced bias, epitranscriptomics)

Long reads technologies



from Shanika L. Amarasinghe et al. Genome Biol. 2020

Pacific Biosciences (Pacbio)

- in the case of RNA, a fragment is **read several times** and a consensus is computed
- read length limited by the longevity of the polymerase
- circular consensus sequence quality = *f*(fragment length, pol longevity)
- 4 passes : 1% error (0.1% reached after 9 passes)
- bias for indels in homopolymers

Pacific Biosciences (Pacbio)

- the protocol is better suited for studying **isoform identification only** (not quantification)
  - initial overrepresentation of shorter molecules lead to size selection which introduces a bias
  - mitigation solutions still in progress

Oxford Nanopore technologies (ONT)

- no limit to read length
- the fragment is read only once in the pore
- read quality depends on the speed of the fragment through the pore
  - **quality decreases in the late stages** of sequencing
- error rate >5%
- bias for **indels in homopolymers**

Oxford Nanopore technologies (ONT)

- 1D sequencing protocol : **single pass** of strands
- ($1D^2$ protocol: sequence the **complementary strand immediately after** the forward strand and compute a consensus)
- accuracy over homopolymers is in progress (from R10 chemistry)

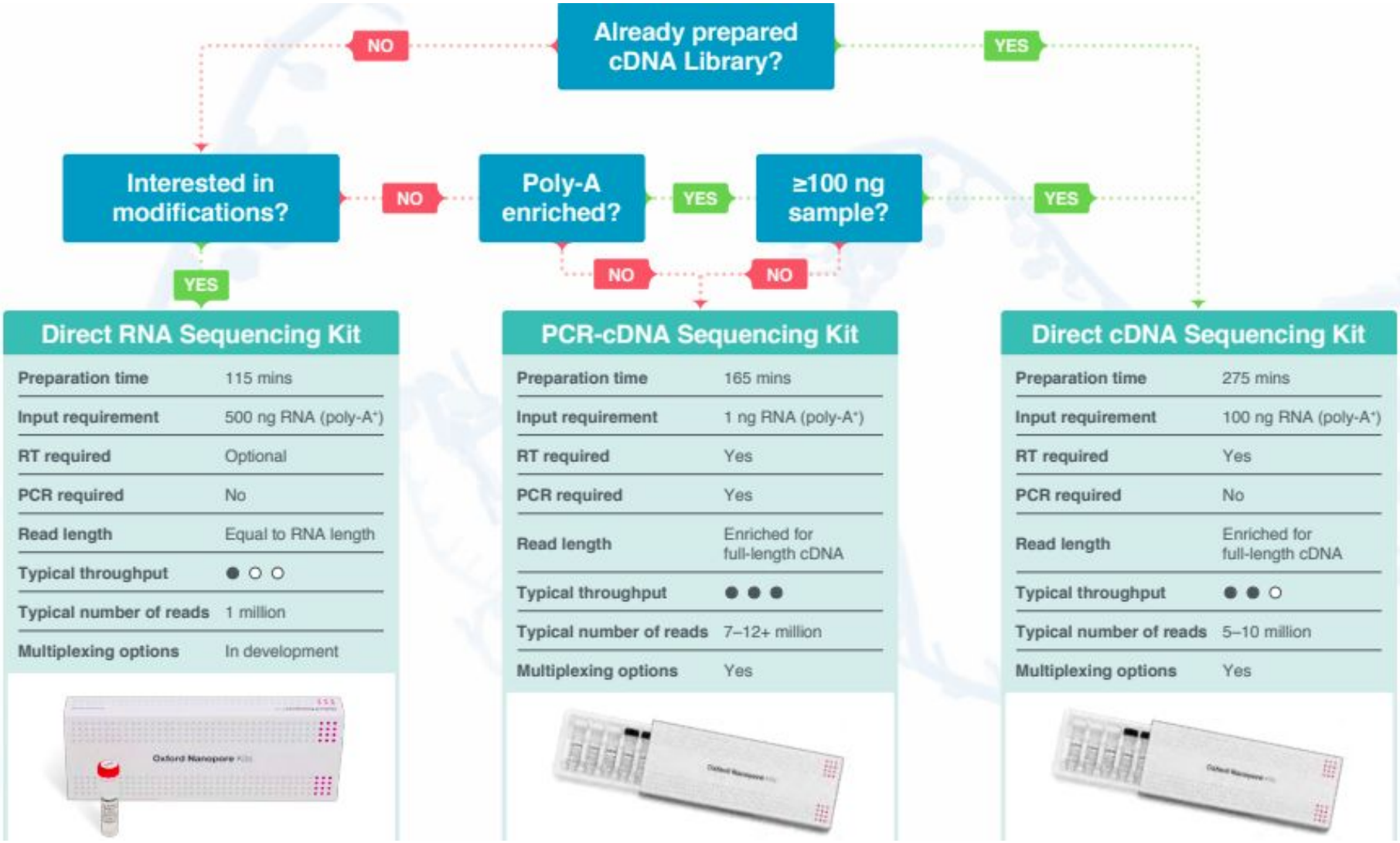
Oxford Nanopore technologies (ONT)'s RNA direct

Methods based on reverse transcription:

- Template switching and artifactual splicing
- Loss of strandedness information
- Loss of base modifications
- Propagation of error due to PCR

Direct RNA

- no bias due to PCR
- possible to study some RNA modifications
- as of today not adequate for quantification (too much material is required)



material from Oxford Nanopore

What has been studied with long reads so far

Near mature:

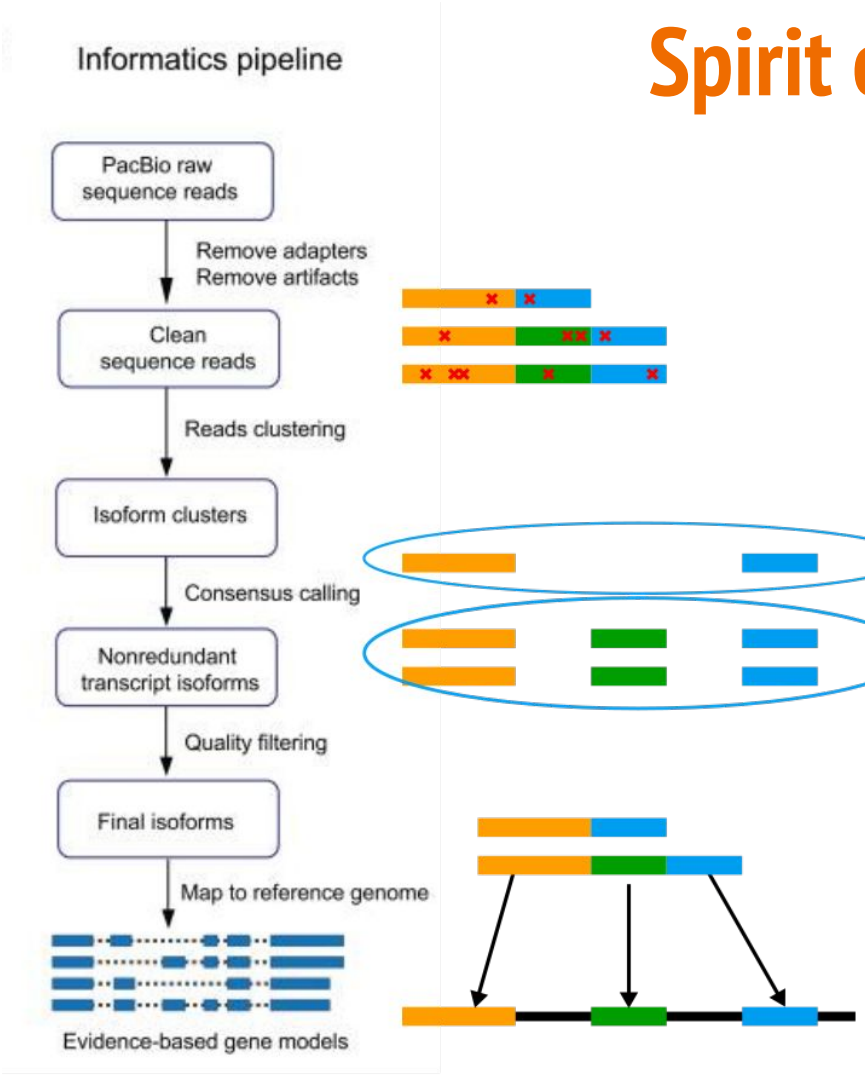
- **Quantification** of already **known genes** and isoforms
- **Quantification** of **novel isoforms** from known genes
- Transcript reconstruction (assembly) **based on a reference**

What has been studied with long reads so far

Exploratory:

- RNA of paralogous genes (Dougherty et al., 2018, Chen et al., 2017)
- Fusion transcripts (Nattestad et al., 2018).
- Allele-specific expression (Tilgner et al., 2014), avelier et al., 2015).

Spirit of most analysis pipelines



reads
comparison all vs all

clusters:
isoform detection
compute consensus

report non redundant
polished transcript sequences

alignment to genome
(Minimap2, GraphMap2, GMAP...)

report genes/isoforms
quantify

adapted from Gordon et al. 2015

Isoform detection: PacBio's Iso-Seq3 + ToFU/Cupcake

https://github.com/ylipacbio/IsoSeq3/

- will tend to **merge alternative transcripts** (heavily depends on the reference quality)
- computationally expensive
- tailored to **Pacbio reads only**
- scripts for exon-junction description and quantification

Alternative isoforms detection pipelines

Specialized for Pacbio

- SQANTI (reference genome, gff)
- ToFu (reference genome & limited *de novo*)
- TAPIS (reference genome)
- IsoCon (*de novo* correction and detection of different transcripts at the base level, targeted data)

Specialized for Nanopore

- FLAIR (reference genome)

Technology agnostic

- TALON (input = alignments to ref)
- MANDALORION
- TrackCluster (*de novo*)

Pipelines focused on quantification

- Developed by Nanopore (based on alignment + Salmon) https://github.com/nanoporetech/pipeline-transcriptome-de
- LIQA (truncated reads treated using an EM algorithm)

Application example

Front Genet. 2021; 12: 683408.
Published online 2021 Jul 15. doi: 10.3389/fgene.2021.683408
PMCID: PMC8321248
PMID: 34335690

PacBio Iso-Seq Improves the Rainbow Trout Genome Annotation and Identifies Alternative Splicing Associated With Economically Important Phenotypes

Ali Ali,[1] Gary H. Thorgaard,[2] and Mohamed Salem[1,*]

Long-read cDNA sequencing identifies functional pseudogenes in the human transcriptome

Robin-Lee Troskie, Yohaann Jafrani, Tim R. Mercer, Adam D. Ewing ✉, Geoffrey J. Faulkner ✉ & Seth W. Cheetham ✉

Genome Biology **22**, Article number: 146 (2021) | Cite this article

Long reads miscellaneous

- Specific spliced alignment tools start to emerge (uLTRA, Sahlin et al. 2021)
- Cleaning for spliced sites (with ref) TranscriptClean , FLAIR
- Reference-free correction might become a standard in the years to come (isONcorrect, Sahlin et al. 2021) (/!\ generally, do not use reference free correction methods tailored for genomic long reads)
- Assembly using short+long reads+ref: StringTie2
- De novo assembly (RATTLE, de La Rubia et al. 2022)

- A website that lists long reads tools: https://long-read-tools.org/table.html

Next challenges with long reads

- guarantee full-length RNA or cDNA libraries
- sequence all different RNAs (not only poly-A)
- allele-specific transcripts
- acquire knowledge about 3' and 5' ends, polyA tails (homopolymers)
- new steps toward full de novo pipelines

What was not viewed during this session

-bacterial RNA

-genome-guided assembly

-metatranscriptomics

-single cell RNA

-tools specialized for ncRNAs, smallRNAs

-tools specialized for fusion transcripts

-transcript annotation (https://busco.ezlab.org/ for instance)

- ...

-**up next**: differential study (statistics for RNA-seq)