Plateforme Bioinformatique Midi-Pyrénées

RNA-Seq data analysis

Cédric Cabau Sigenae / Céline Noirot Bioinfo Genotoul

Material

http://bioinfo.genotoul.fr/index.php?id=119

Slides & Exercise leaflet (doc)

- pdf : one per page
- pdf : three per page with comment lines

Data & results files (data)

http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/

Session organisation

Afternoon (14h-17h) :

- Sequence quality
  - Theory + exercises
- Spliced read mapping
  - Theory + exercises

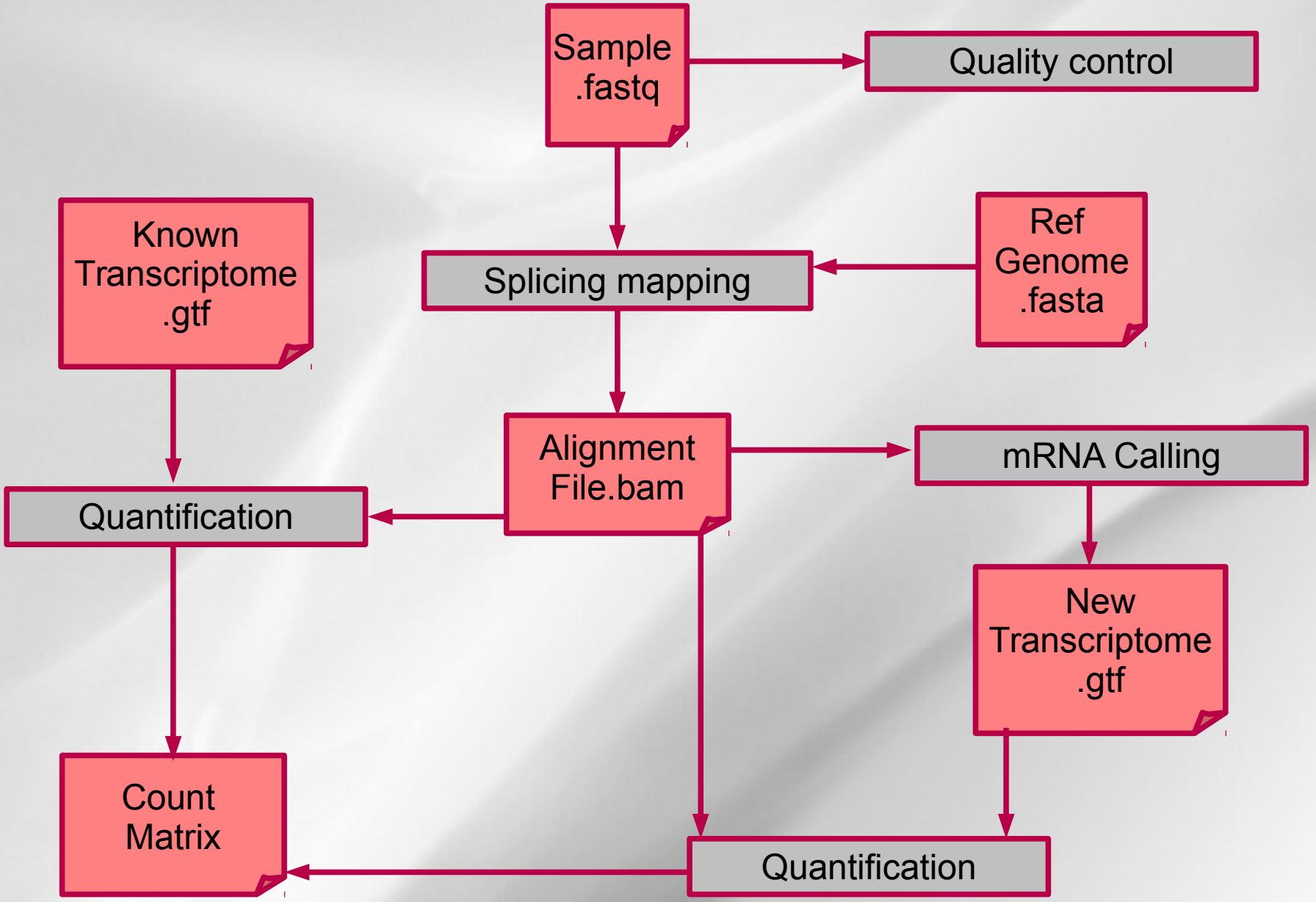
Morning (9h00 -12h30) :

- Visualisation
  - Exercises
- expression measurement
  - Theory + exercises

Afternoon (14h-17h) :

- mRNA calling
  - Theory + exercises
- some statistics ...

Session organisation



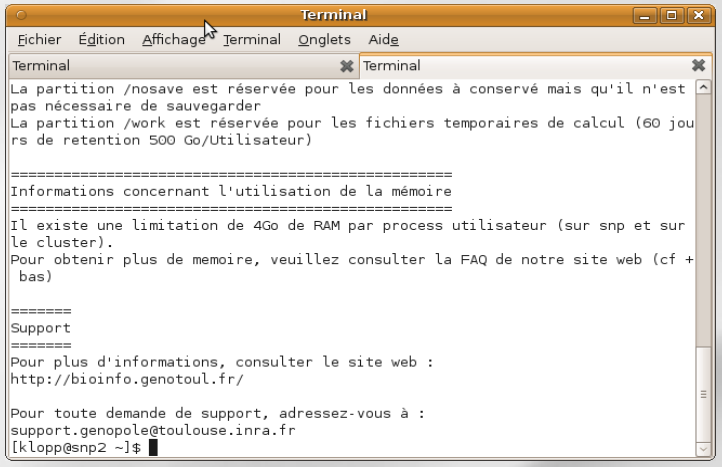
What you should know

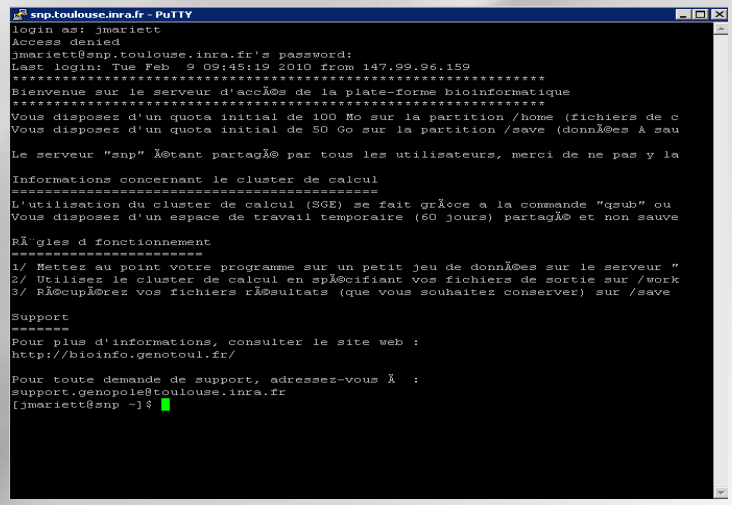
How to connect to genotoul.toulouse.inra.fr?

How to use unix commands?

wget URL

qlogin -pe parallel_smp 4





Summary - Sequence quality

- Context, vacabulary, transcriptome variability ...
- Methods to analyse transcriptoms
- What is RNAseq ?
- High throughput sequencers
- Illumina protocol, paired-end library, directionnal library
- Known biais
- How to check quality ?

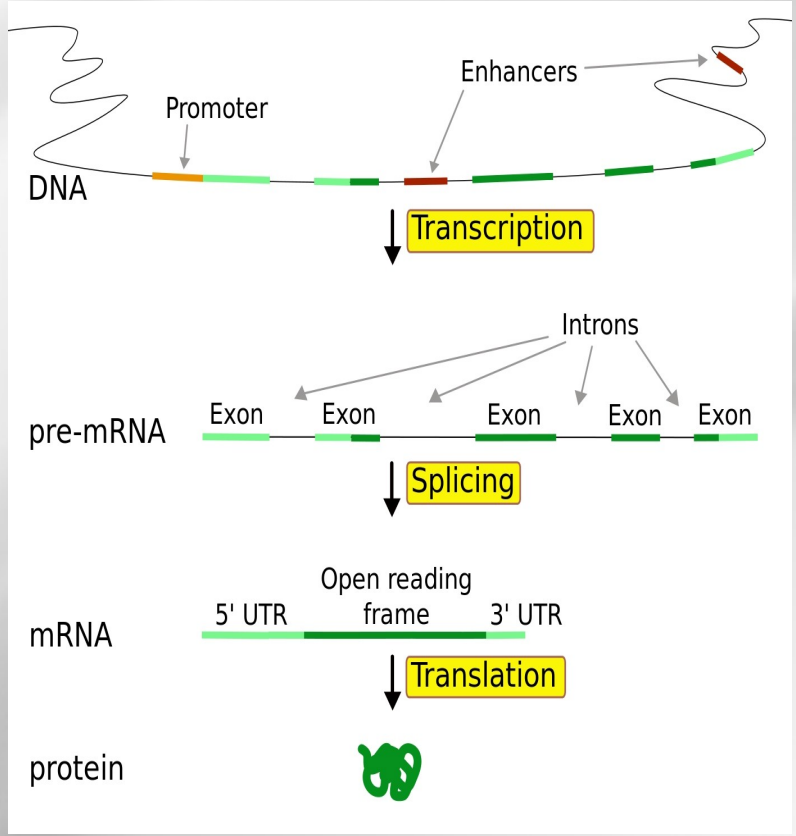
Context

Prerequis :

- Reference genome available
- RNAseq sequencing
  (sequence of transcript)

Try to answer to :

- How to map transcript to the genome ?
- How to discover new transcript ?
- What are the alternative transcript ?

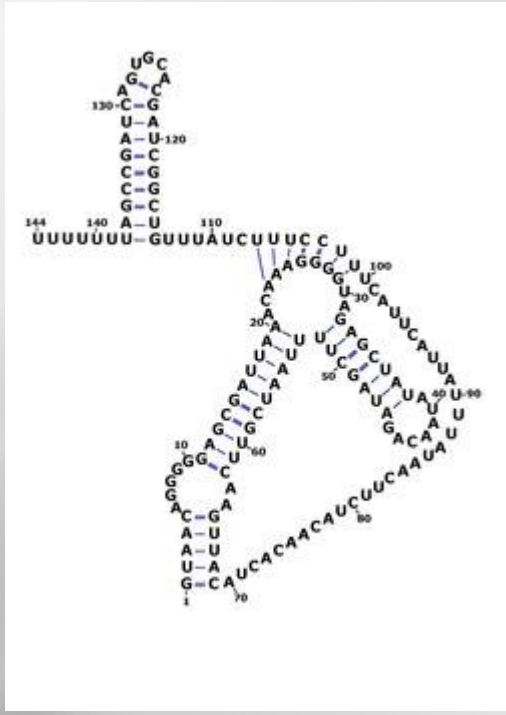


Source : en.wikipedia.org/wiki/User:Forluvoft/sandbox

Transcription products

Protein coding gene: transcribed in mRNA
ncRNA : highly abundant and functionally important RNA

- tRNA,
- rRNA,
- snoRNAs,
- microRNAs,
- siRNAs,
- PiRNAs
- lincRNA



http://en.wikipedia.org/wiki/User:Amarchais/RsaOG_RNA

ENCODE

- Project
- Data
- Statistics
- Participants
- Publications
- RGASP 1/2
- RGASP 3
- Contact us

Statistics about the current GENCODE freeze (version 13)

Statistics of previous Gencode freezes are found archived here.

*The statistics derive from the **gtf files**, which include only the main chromosomes of the human reference genome.

Version 13 (March 2012 freeze, GRCh37)

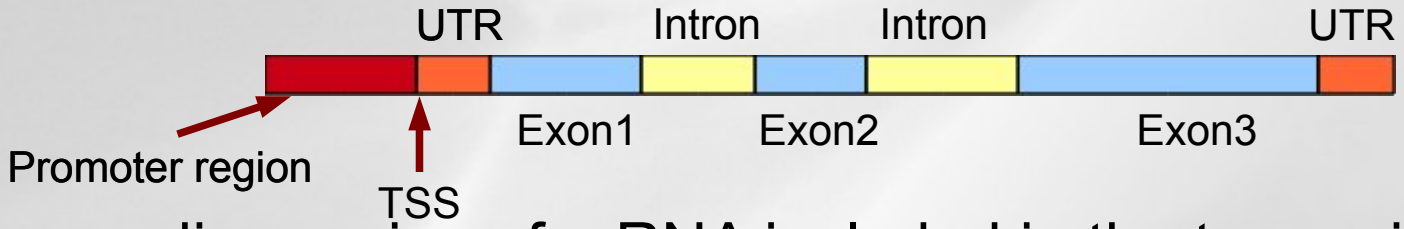
General stats

Total No of Genes	55123	**Total No of Transcripts**	182967
Protein-coding genes	20070	**Protein-coding transcripts**	77901
Long non-coding RNA genes	12393	- full length protein-coding:	55928
Small non-coding RNA genes	9173	- partial length protein-coding:	21973
Pseudogenes	13123	**Nonsense mediated decay transcripts**	11549
- processed pseudogenes:	9895	**Long non-coding RNA loci transcripts**	19835
- unprocessed pseudogenes:	2550		
- unitary pseudogenes:	156		
- polymorphic pseudogenes:	31		
- pseudogenes:	298		
Immunoglobulin/T-cell receptor gene segments		**Total No of distinct translations**	78119
- protein coding segments:	364	**Genes that have more than one distinct translations**	14235
- pseudogenes:	193		

http://www.gencodegenes.org/stats.html

Vocabulary

Gene : functional units of DNA that contain the instructions for generating a functional product.

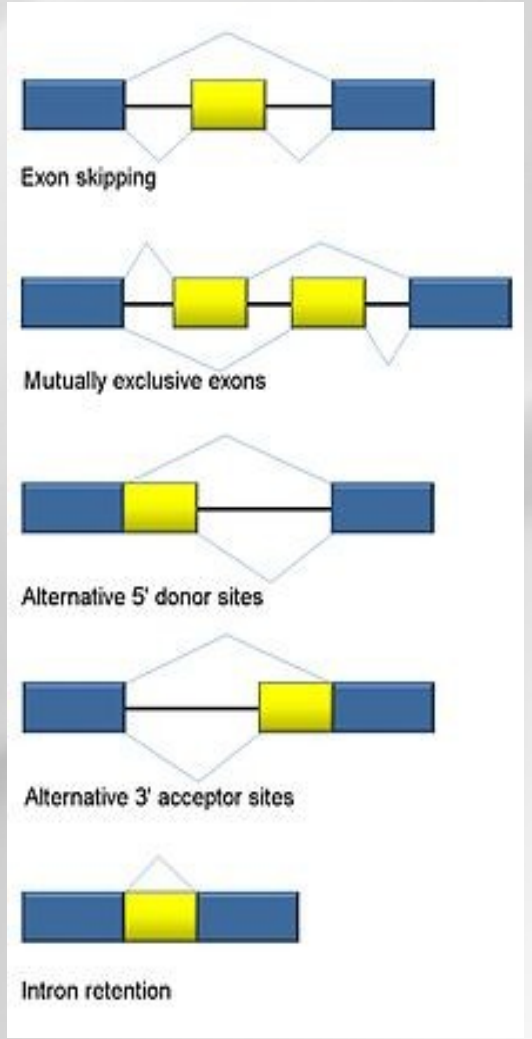


Exon : coding region of mRNA included in the transcript
Intron : non coding region
TSS : Transcription Start Site ≠ 1st amino acid
Transcript : stretch of DNA transcribed into an RNA molecule

Alternative splicing

Alternative splicing (or differential splicing)

- the exons are reconnected in multiple ways during RNA splicing.
- different mRNAs translated into different protein isoforms
- a single gene may code for multiple proteins.



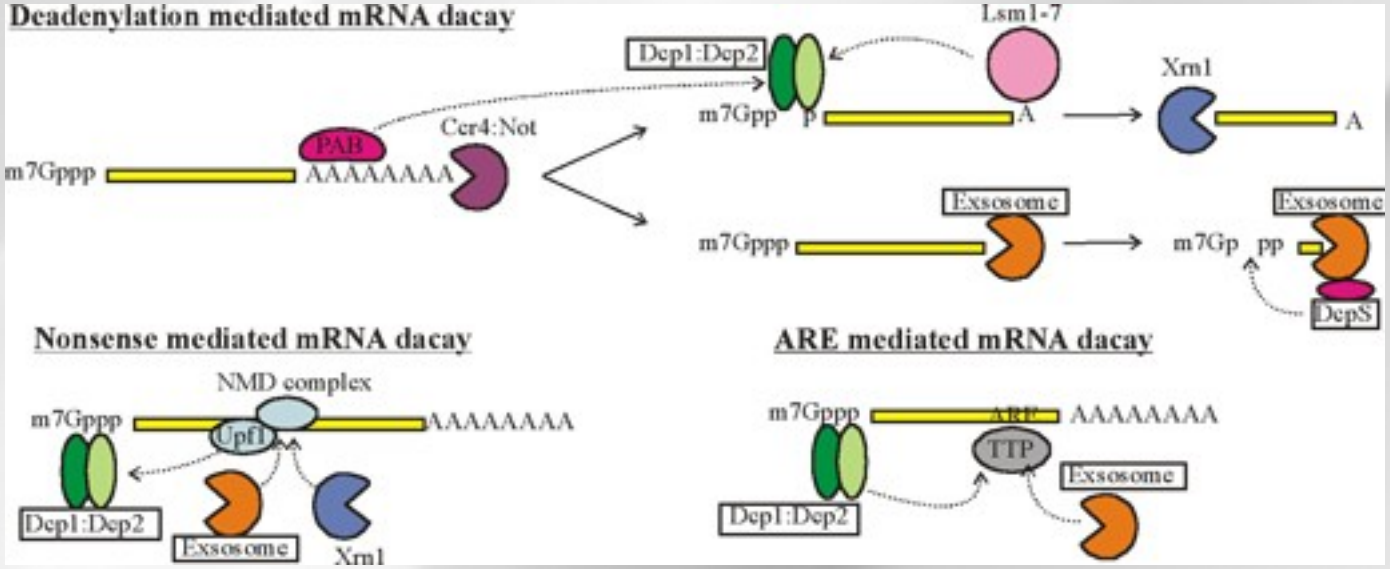
Intron Retention

Post-transcriptional modification (eukaryotic cells) eg: the conversion of precursor messenger RNA into mature mRNA (mRNA), editing...

http://en.wikipedia.org/wiki/Alternative_splicing

Transcript degradation

- mRNA export to the cytoplasm,
- protected from degradation by a 5' cap structure and a 3' polyA tail.
- the polyA tail is gradually shortened by exonucleases
- the degradation machinery rapidly degrades the mRNA in both in directions.
- others mechanisms, bypass the need for deadenylation and can remove the mRNA from the transcriptional pool independently.



http://www.eb.tuebingen.mpg.de/research-groups/remco-sprangers

Cis-natural antisense transcript

- Natural antisense transcripts (NATs) are a group of RNAs encoded within a cell that have transcript complementarity to other RNA transcripts.

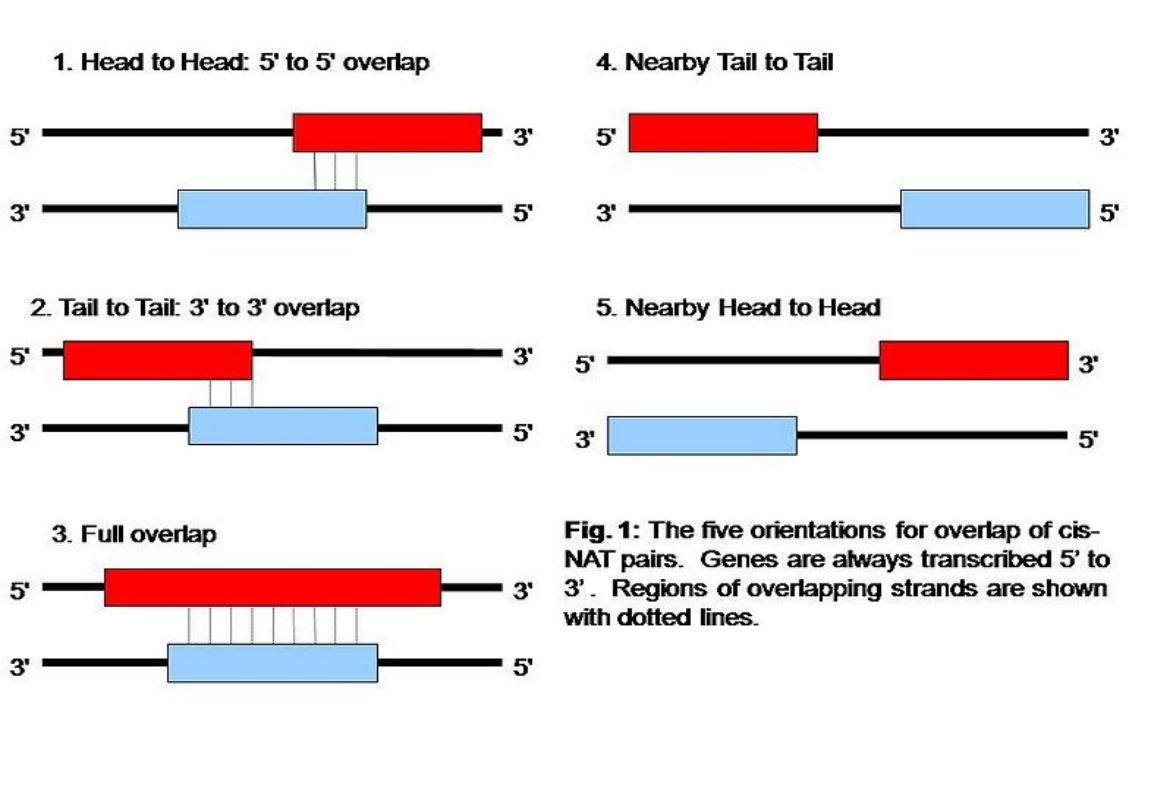


Fig. 1: The five orientations for overlap of cis-NAT pairs. Genes are always transcribed 5' to 3'. Regions of overlapping strands are shown with dotted lines.

http://en.wikipedia.org/wiki/Cis-natural_antisense_transcript

Fusion genes

- A fusion gene is a hybrid gene formed from two previously separate genes. It can occur as the result of a translocation, interstitial deletion, or chromosomal inversion. Often, fusion genes are oncogenes.

Genome Biol. 2011 Jan 19;12(1):R6. [Epub ahead of print]

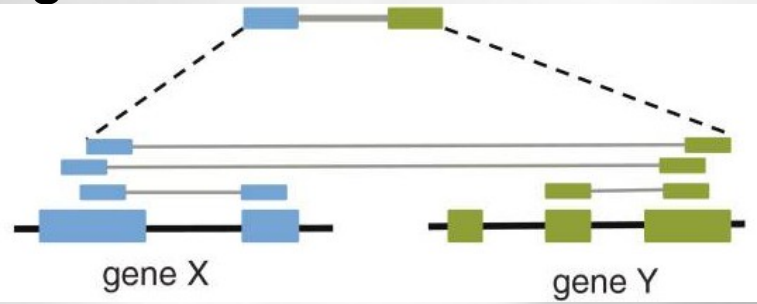
Identification of fusion genes in breast cancer by paired-end RNA-sequencing.

Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O.

Institute for Molecular Medicine Finland (FIMM), Tukholmankatu 8, Helsinki, 00290, Finland. olli.kallioniemi@fimm.fi.

http://en.wikipedia.org/wiki/Fusion_gene

- They often come from trans-splicing : Trans-splicing is a special form of RNA processing in eukaryotes where exons from two different primary RNA transcripts are joined end to end and ligated.



http://en.wikipedia.org/wiki/Trans-splicing

Transcriptome variability

- Many types of transcripts (mRNA, ncRNA ...)
- Many isoform (non canonical splice sites, intron retention ...)
- Number of transcripts
  - possible variation factor between transcripts: 10^6 or more,
  - expression variation between samples.
- Allele specific expression

How can we study the transcriptome?

Techniques classification

EST	PCR/RT-QPCR	SAGE	MicroArrays
No quantification	Quantification	Quantification	Indirect quantification
Low throughput	Low throughput (up to hundreds)	Low throughput (up to thousands)	High throughput (up to millions)
Discovery (Yes)	No	No	Discovery (Yes)

⟶ Need transcript sequence partially known

⟶ Difficulties in discovering novels splice events

What is RNA-Seq ?

- use of **high-throughput sequencing technologies** to sequence cDNA in order to get information about a sample's RNA content
- the deep coverage and base level resolution => measure transcriptome data experimentally

Nature Reviews Genetics **10**, 57-63 (January 2009) | doi:10.1038/nrg2484

ARTICLE SERIES: Applications of next-generation sequencing

INNOVATION

RNA-Seq: a revolutionary tool for transcriptomics

Zhong Wang[1], Mark Gerstein[1] & Michael Snyder[1] About the authors

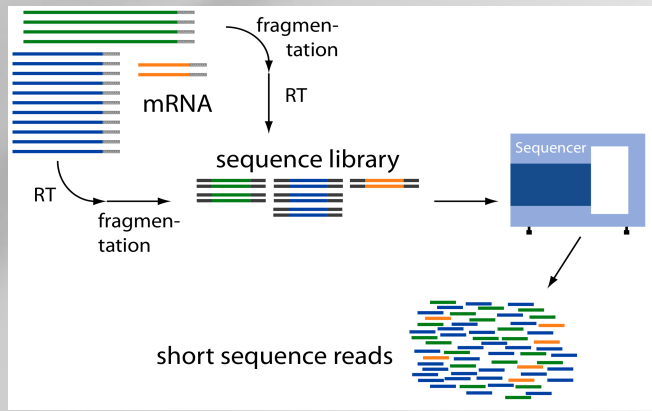
top

RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptomes. RNA-Seq also provides a far more precise measurement of levels of transcripts and their isoforms than other methods. This article describes the RNA-Seq approach, the challenges associated with its application, and the advances made so far in characterizing several eukaryote transcriptomes.

http://en.wikipedia.org/wiki/RNA-Seq

What is different with RNA-Seq ?

- No prior knowledge of sequence needed
- Specificity of what is measured
- Increased dynamic range of measure, more sensitive detection
- Direct quantification
- Good reproducibility
- Different levels : genes, transcripts, allele specificity, structure variations
- New feature discovery: transcripts, isoforms, ncRNA, structures (fusion...)
- Possible detection of SNPs, ...



RNA-Seq platforms comparison

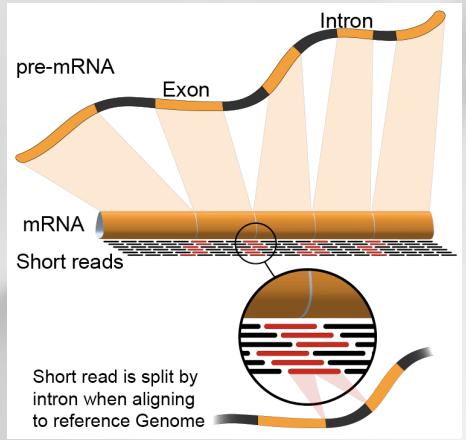
Séquenceurs 2ème génération (2013)

Société	Roche		Illumina				Life technologies						
Plateforme													
	GS Junior	454	MiSeq	GAIIx	HiSeq 1000/1500	HiSeq 2000/2500	Ion PGM			Ion Proton		5500xl SOLiD	5500 SOLiD
Technologie	Titanium	GS FLX+					Chip 314 v2	Chip 316 v2	Chip 318 v2	Chip PI	Chip PII		
Génome humain	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✓	✓
Exome	✗	✗	✓	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓
Petit génome (Bactéries, levures)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Régions ciblées	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Transcriptome	✗	✗	✗	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓
Chip-Seq	✗	✗	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓
Métagénomique	✗	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓

Different approaches :

Alignment to

- De novo
  - No reference genome, no transcriptome available
  - Very expensive computationally
  - Lots of variation in results depending on the software used
- Reference transcriptome
  - Most are incomplete
  - Computationally inexpensive
- Reference genome
  - When available
  - Allow reads to align to unannotated sites
  - Computationally expensive
  - Need a spliced aligner



What are we looking for?

Identify genes

- List new genes

Identify transcripts

- List new alternative splice forms

Quantify these elements ⟶ differential expression

Usual questions on RNA-Seq !

- How many replicates ?
    - Technical or/and biological replicates ?
- How many reads for each sample?
- How many conditions for a full transcriptome ?
- How long should my reads be ?
- Single-end or paired-end ?

Depth VS Replicates

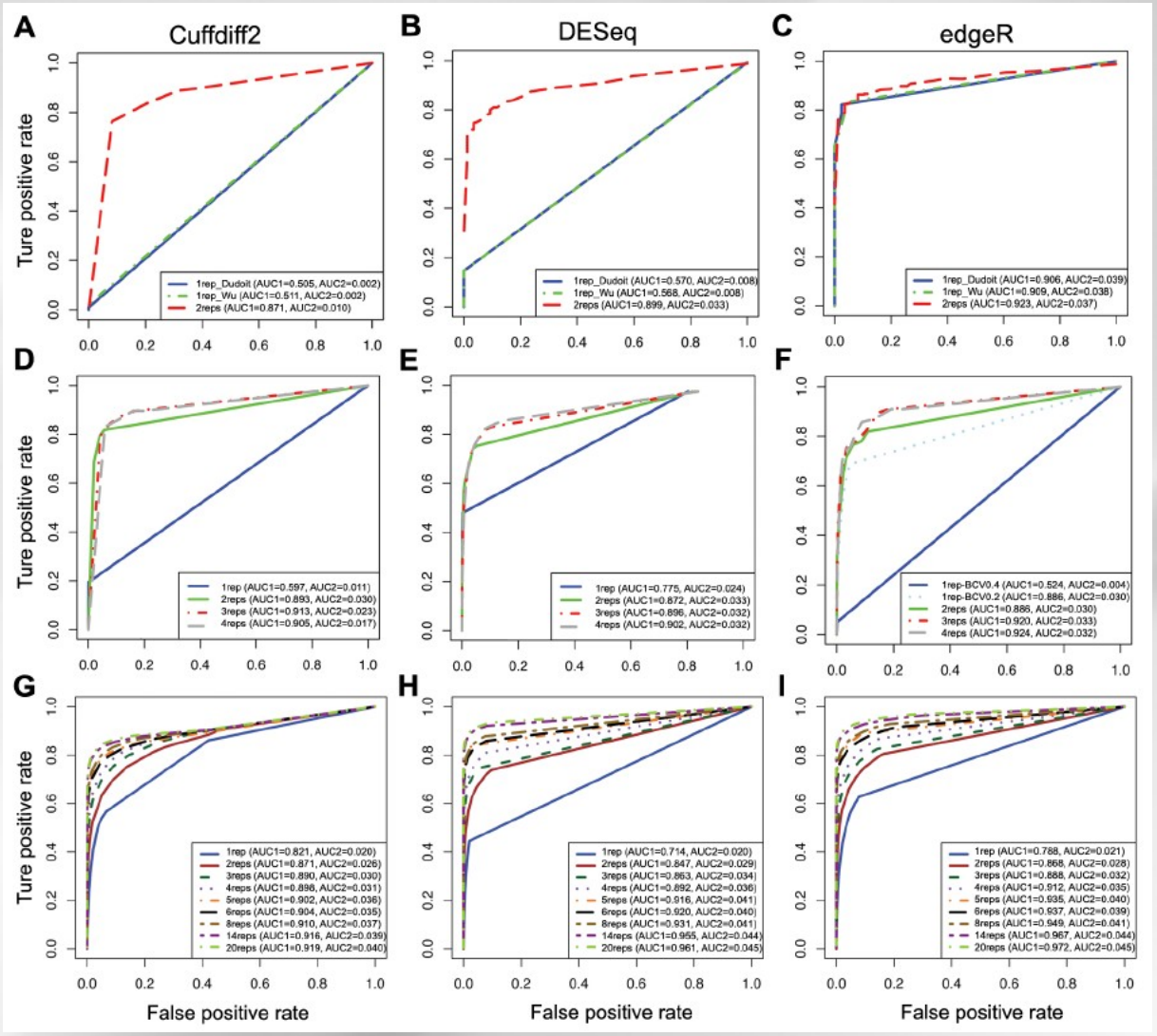
- Encode (2011) : https://www.encodeproject.org/data-standards/
  - Experiments should be performed with **two or more biological replicates**, unless there is a compelling reason why this is impractical or wasteful
  - A typical $\mathbf{R^2}$ (Pearson) correlation of gene expression (RPKM) between two biological replicates, for RNAs that are detected in both samples using RPKM or read counts, should be [0.92 - 0.98] Experiments with biological correlations < 0.9 should either be repeated or explained.
- Between **30M and 100M reads** per sample depending on the study.
- On Human 100M reads are enough to detect 90% of transcript from 81% af genes.
- Zhang et al. 2014 : From 3 replicates improve DE detection and control false positive rate.

Depth VS Replicates

Effect of number of replicates on true positive rate and false positive rate.

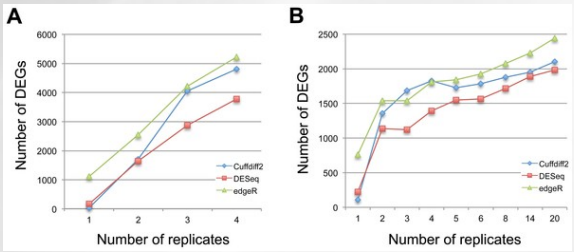
MAQC

K_N



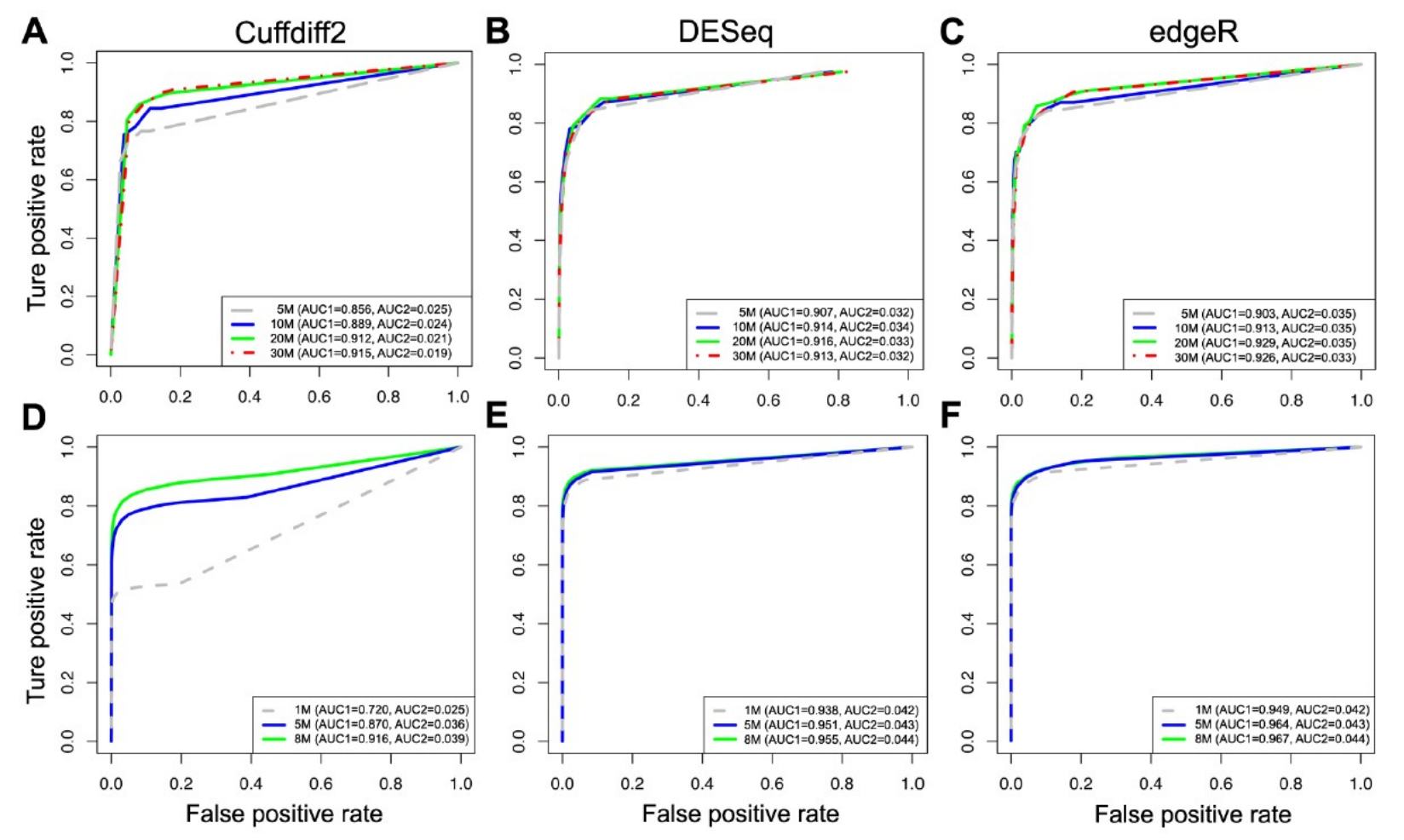
K_N

LCL2

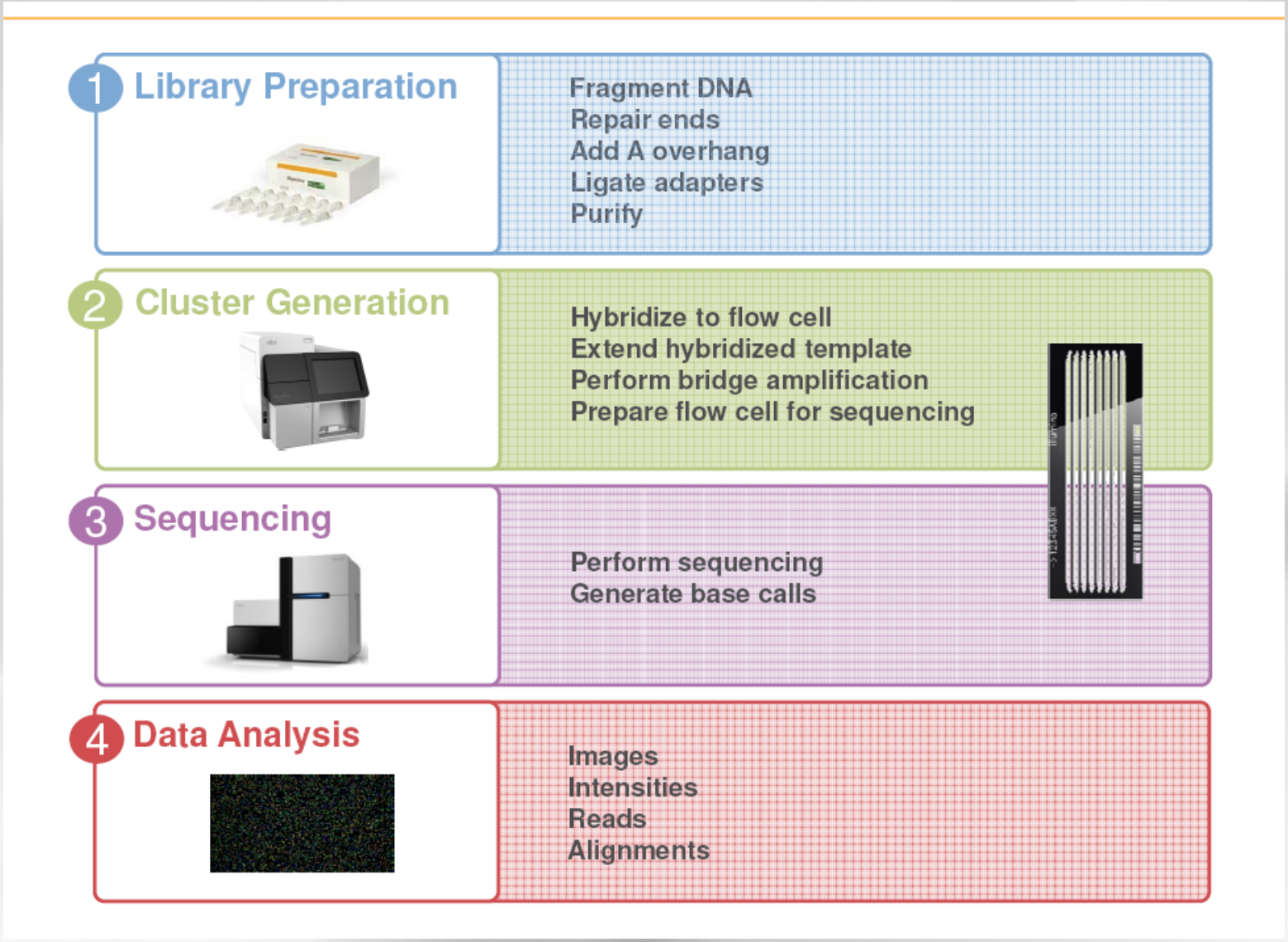


Depth VS Replicates

Depth effect



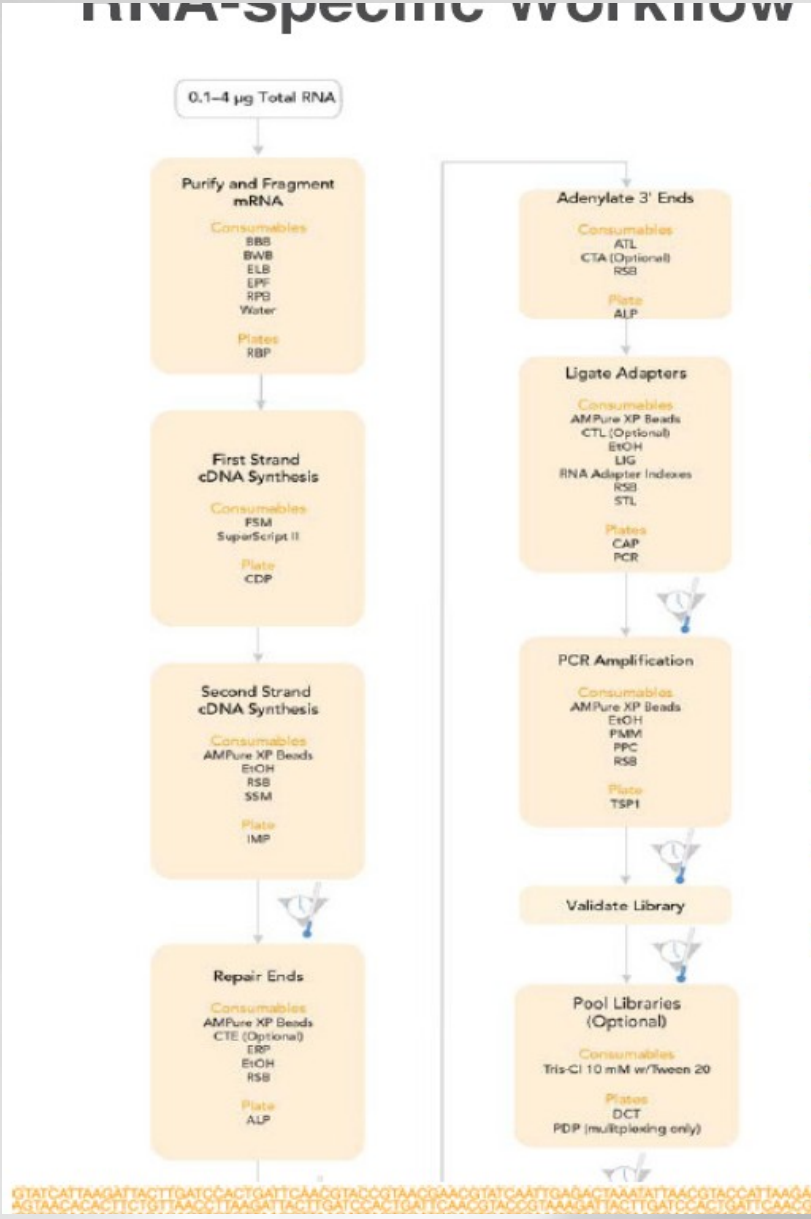
Illumina RNA-Seq protocol



1 Flowcell:

- in general 1 run
- equivalent to 8 Lane
- Hiseq 2500: 2 Billion reads single or 4 Billion paired reads.

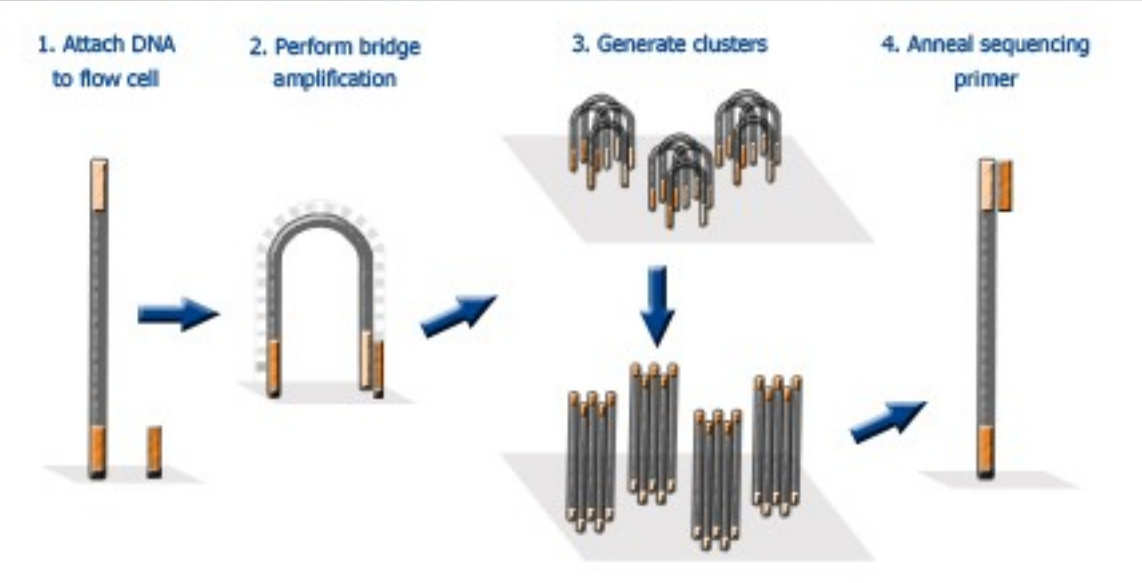
RNA-Seq library preparation

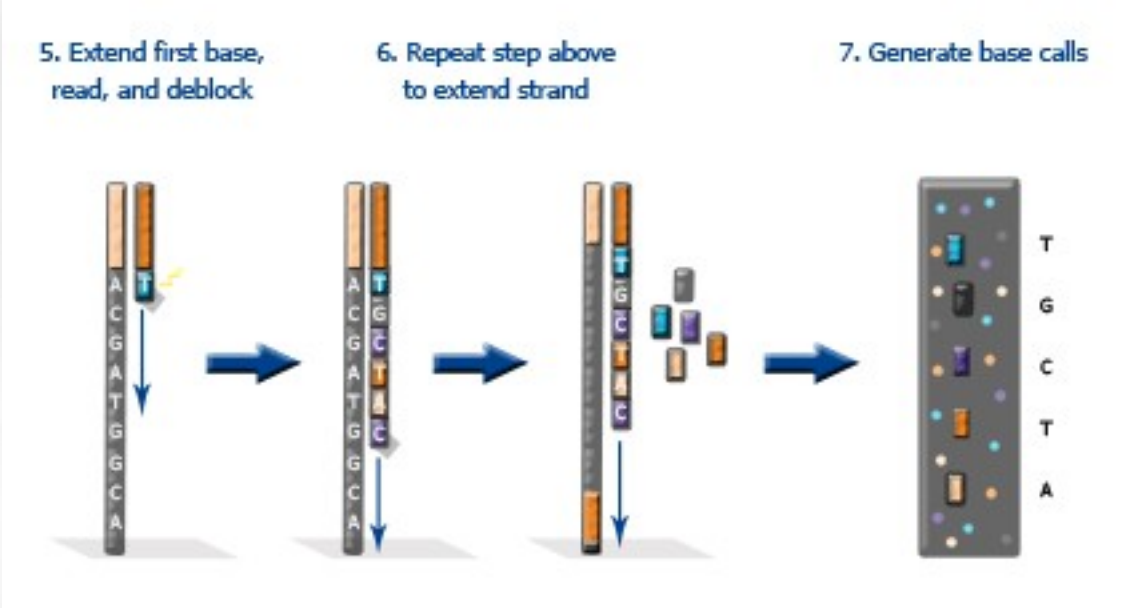


- Isolate poly-A containing mRNA
- capture mRNA with oligoT beads
- Randomly fragment RNA
- Random prime mRNA → cDNA
- Make 2nd strand cDNA
- Repair-Ends and 3' Ends Adenylate
- Ligate sequencing adapters
- Enrich up to 15 cycles of PCR
- gel purify
- validate library w/ Bioanalyzer

Library prep takes <2 days

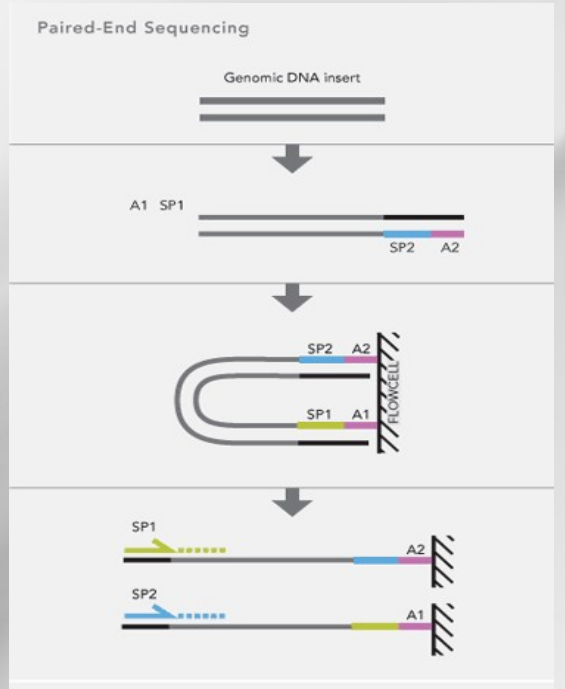
Clusters generation / Sequencing





Paired-end sequencing

- Modification of the standard single-read DNA library preparation facilitates reading both ends of each fragment
- Improvement of mapping
- Help to detect structural variations in the genome like insertions or deletions, copy number variations, and genome rearrangements



Adapter (A1 and A2) with sequencing primer sites (SP1 and SP2) are ligated onto DNA fragments. Template clusters are formed on the flow cell by bridge amplification and then sequenced by synthesis from the paired primers sequentially.

Strand specific RNA-Seq protocol

workflow comparison: mRNA-Seq vs directional mRNA-Seq

1. start with 1µg (or less) total RNA
2. purify poly-A mRNA
3. randomly fragment mRNA

4. 1st strand cDNA synthesis
5. 2nd strand cDNA synthesis
6. end repair
7. adenylate 3' ends
8. ligate adaptors
9. gel purify

4. end repair with phosphatase and PNK
5. column purify PNK treated mRNA
6. ligate 3' adaptor
7. ligate 5' adaptor
8. reverse transcribe

10. enrich with PCR
11. validate library
12. grow clusters
13. sequence on HiSeq2000 (SR or PE)

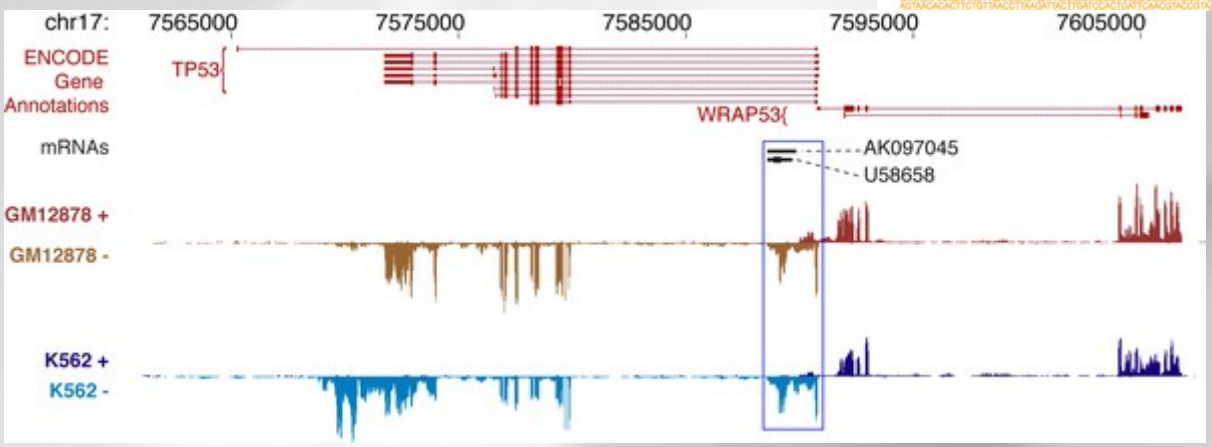
Nat Methods. 2010 Sep;7(9):709-15. Epub 2010 Aug 15.

Comprehensive comparative analysis of strand-specific RNA sequencing methods.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A.

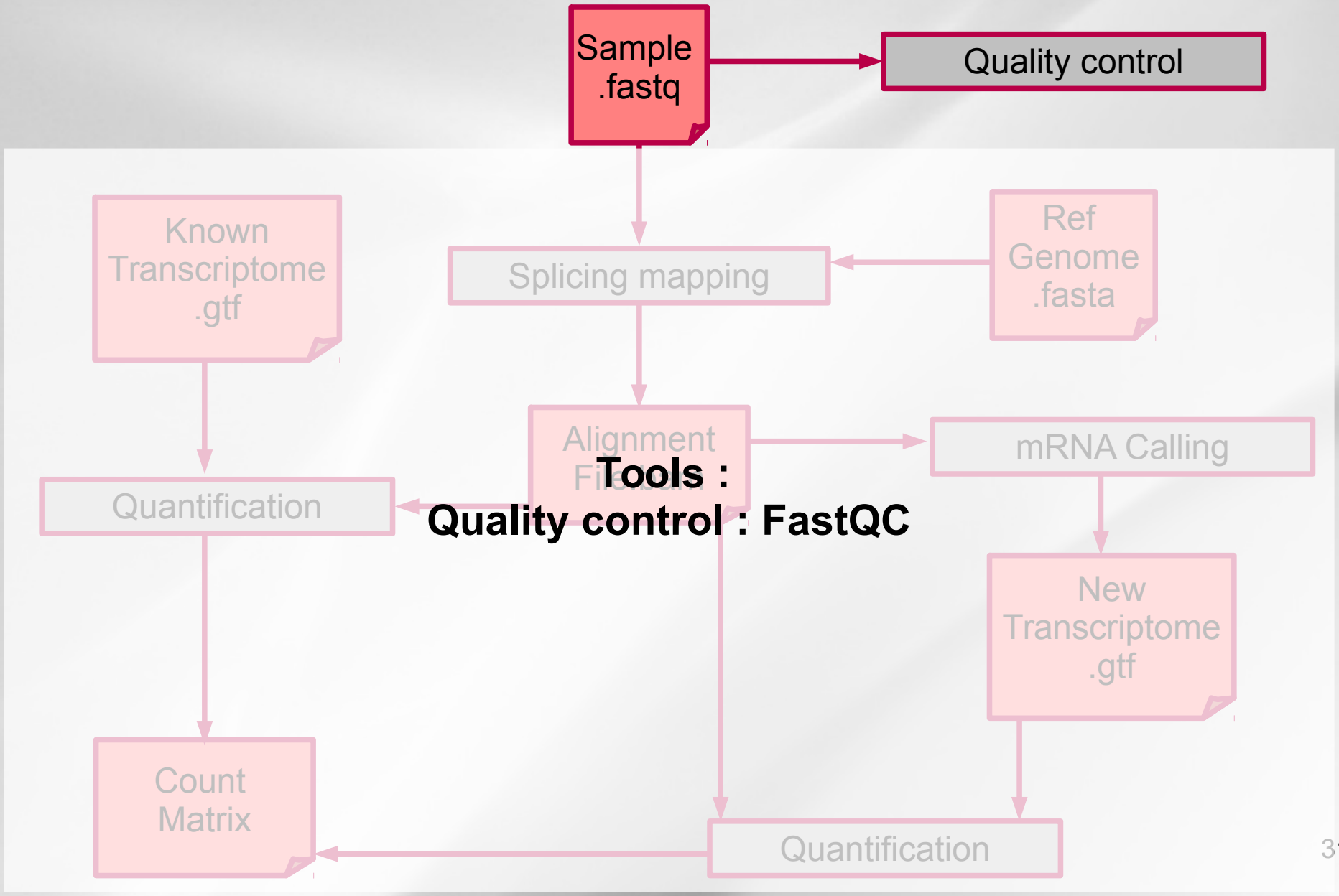
Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA.
jlevin@broadinstitute.org

Abstract



http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046

Analysis workflow



Tools :
Quality control : FastQC

fastq file formats

Published online 16 December 2009 *Nucleic Acids Research, 2010, Vol. 38, No. 6* **1767–1771**
doi:10.1093/nar/gkp1137

SURVEY AND SUMMARY

The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants

Peter J. A. Cock[1,*], Christopher J. Fields[2], Naohisa Goto[3], Michael L. Heuer[4] and Peter M. Rice[5]

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
```

la proba d'une erreur : $Q_{\mathrm{PHRED}} = -10 \times \log_{10}(P_e)$

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

RNAseq specific bias

- Influence of the library preparation
- Random hexamer priming
- Positional bias and sequence specificity bias.

  Robert et al. Genome Biology, 2011,12:R22

- Transcript length bias

- Some reads map to multiple locations

Hexamer random priming bias

Published online 14 April 2010 *Nucleic Acids Research, 2010, Vol. 38, No. 12 e131*
doi:10.1093/nar/gkq224

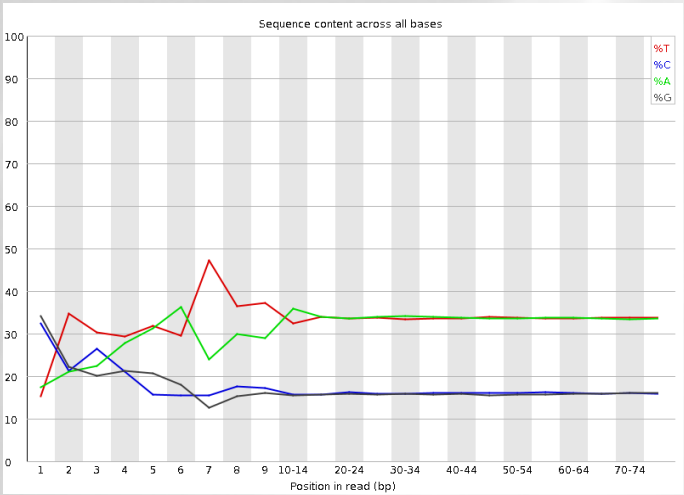
Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

ABSTRACT

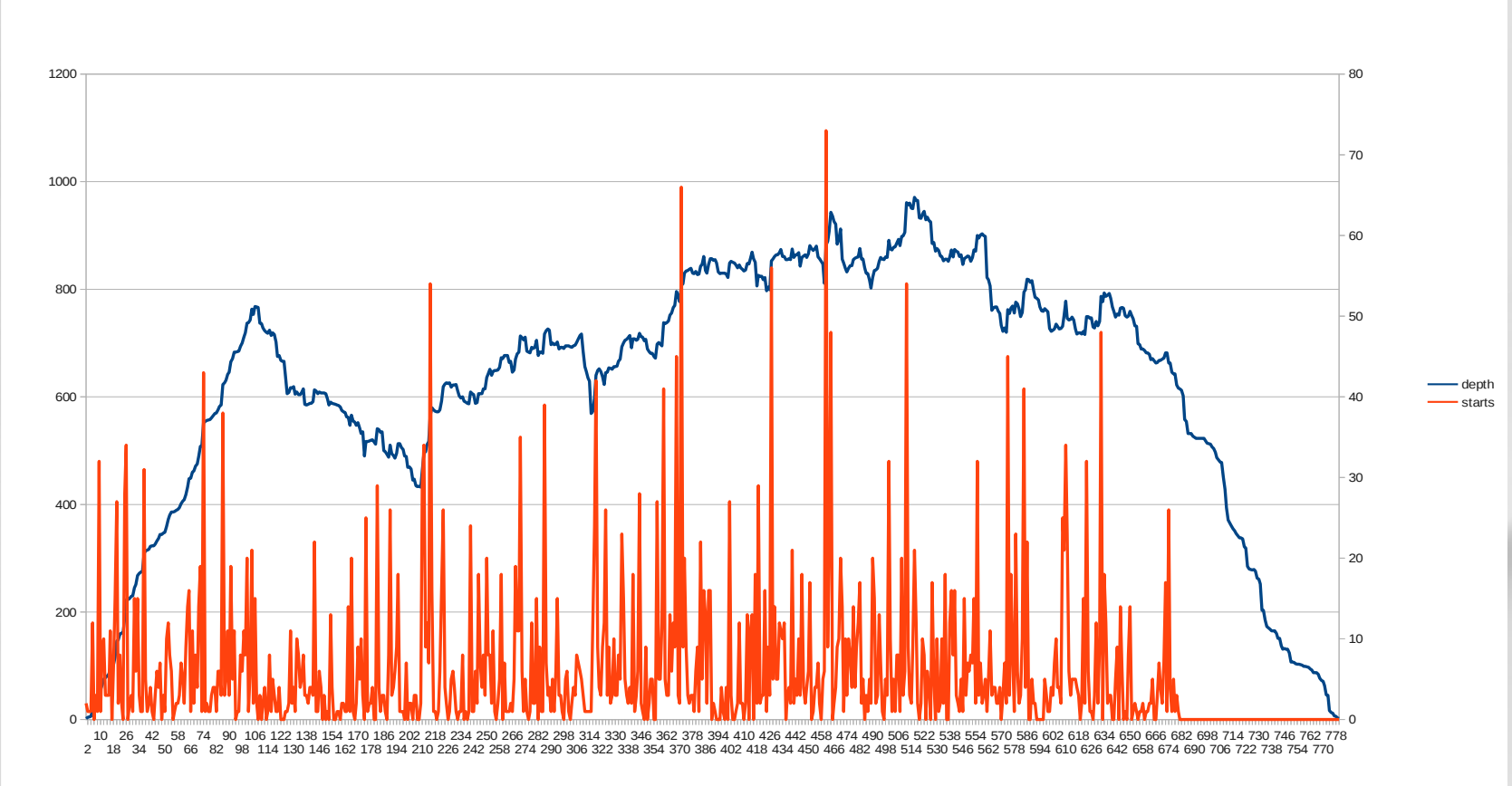
Generation of cDNA using random hexamer priming induces biases in the nucleotide composition at the beginning of transcriptome sequencing reads from the Illumina Genome Analyzer. The bias is independent of organism and laboratory and impacts the uniformity of the reads along the transcriptome. We provide a read count reweighting scheme, based on the nucleotide frequencies of the reads, that mitigates the impact of the bias.

- A strong distinctive pattern in the nucleotide frequencies of the first 13 positions at the 5'-end :
  - sequence specificity of the polymerase
  - due to the end repair performed



- Reads beginning with a hexamer over-represented in the hexamer distribution at the beginning relative to the end are down-weighted

Hexamer random effect



- Orange = reads start sites
- Blue = coverage

Transcript length bias

Biol Direct. 2009 Apr 16;4:14.

Transcript length bias in RNA-seq data confounds systems biology.

Oshlack A, Wakefield MJ.

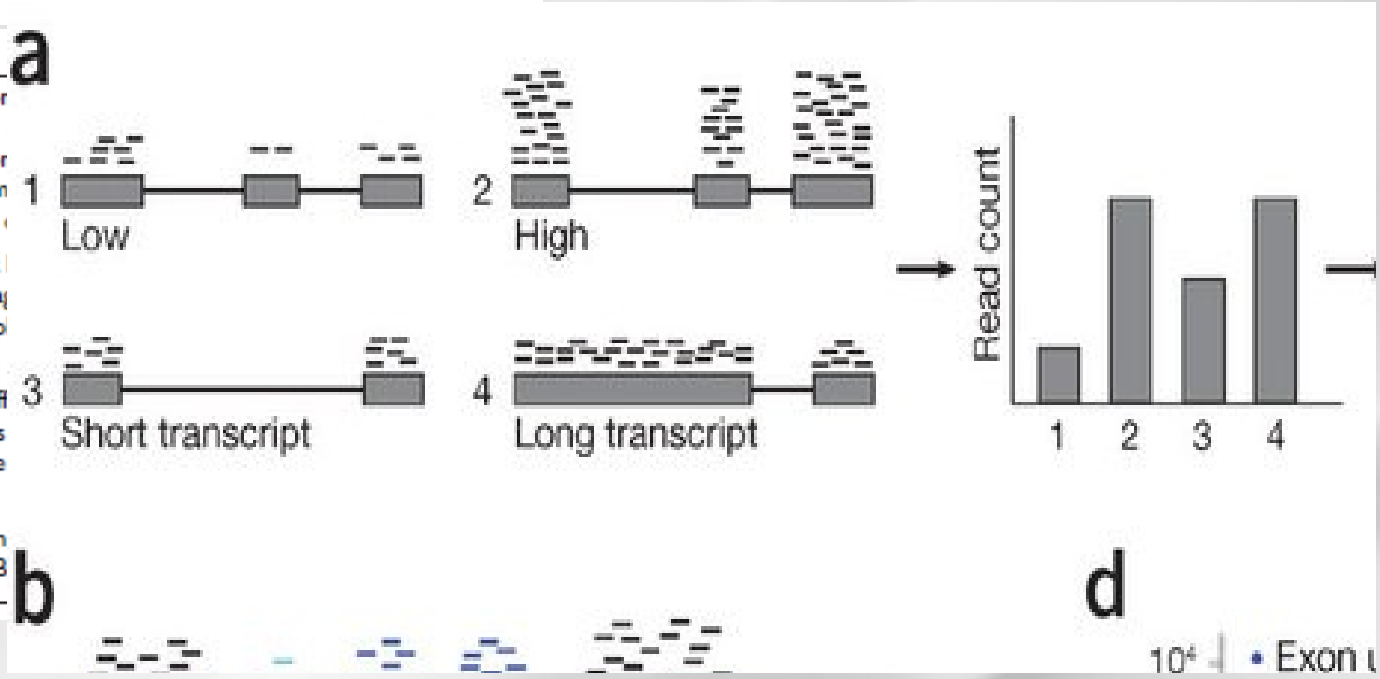
Abstract

Background: Several recent studies have demor
transcriptome analysis (RNA-seq) in mammals.
genome transcriptional profiling is likely to becor
genomic sequences. As yet, a rigorous analysis m
still in the stages of exploring the features of the

Results: We investigated the effect of transcript
published data sets. For standard analyses using a
call differentially expressed genes between samp
transcript.

Conclusion: Transcript length bias for calling diff
current protocols for RNA-seq technology. This
expressed genes, and in particular may introduce
other multi-gene systems biology analyses.

Reviewers: This article was reviewed by Rohan
Cloonan (nominated by Mark Ragan) and James B



- *the differential expression of longer transcripts is more likely to be identified than that of shorter transcripts*

BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 5 2011, pages 662–669
doi:10.1093/bioinformatics/btr005
Gene expression
Advance Access publication January 19, 2011

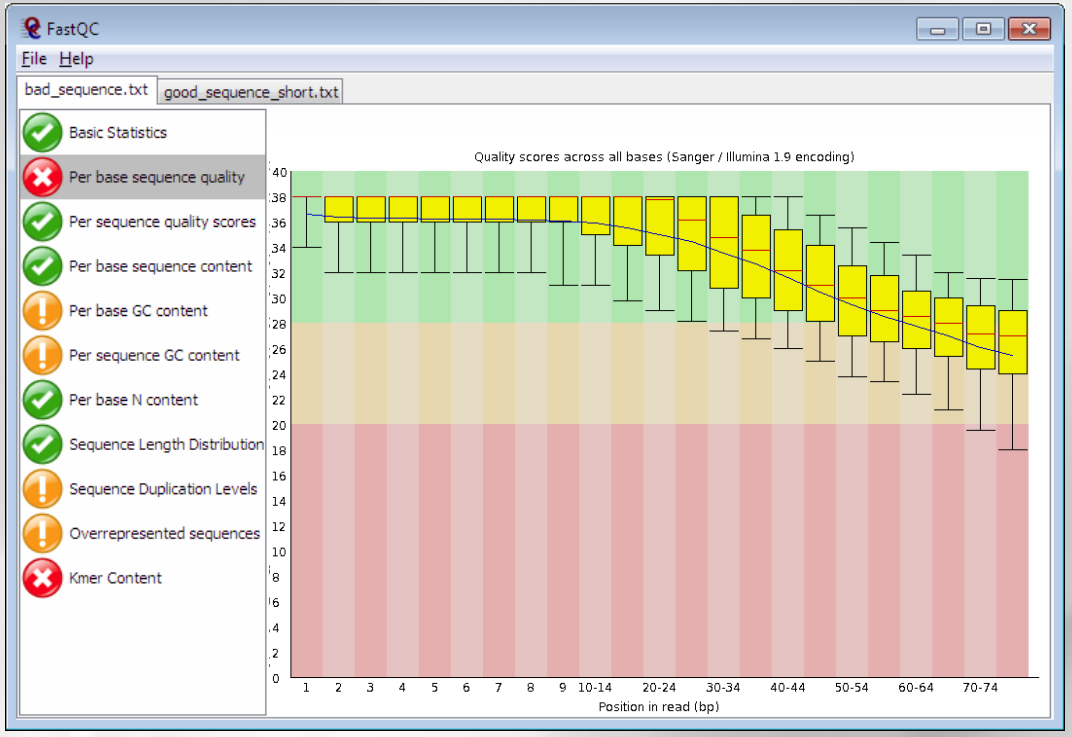
Length bias correction for RNA-seq data in gene set analyses

Liyan Gao[1,†], Zhide Fang[2,†], Kui Zhang[1], Degui Zhi[1] and Xiangqin Cui[1,*]

Verifying RNA-Seq raw data

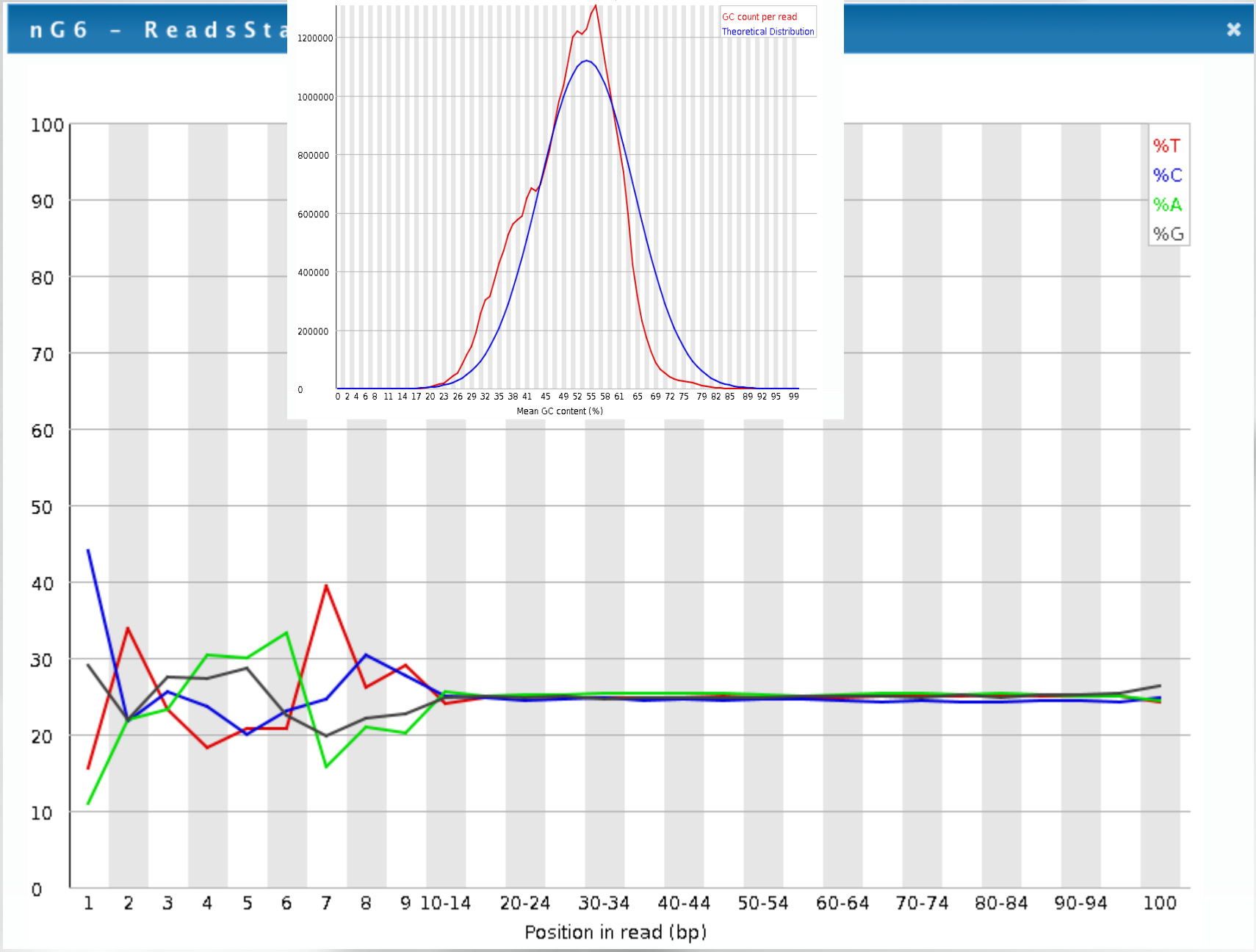
FastQC :

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/



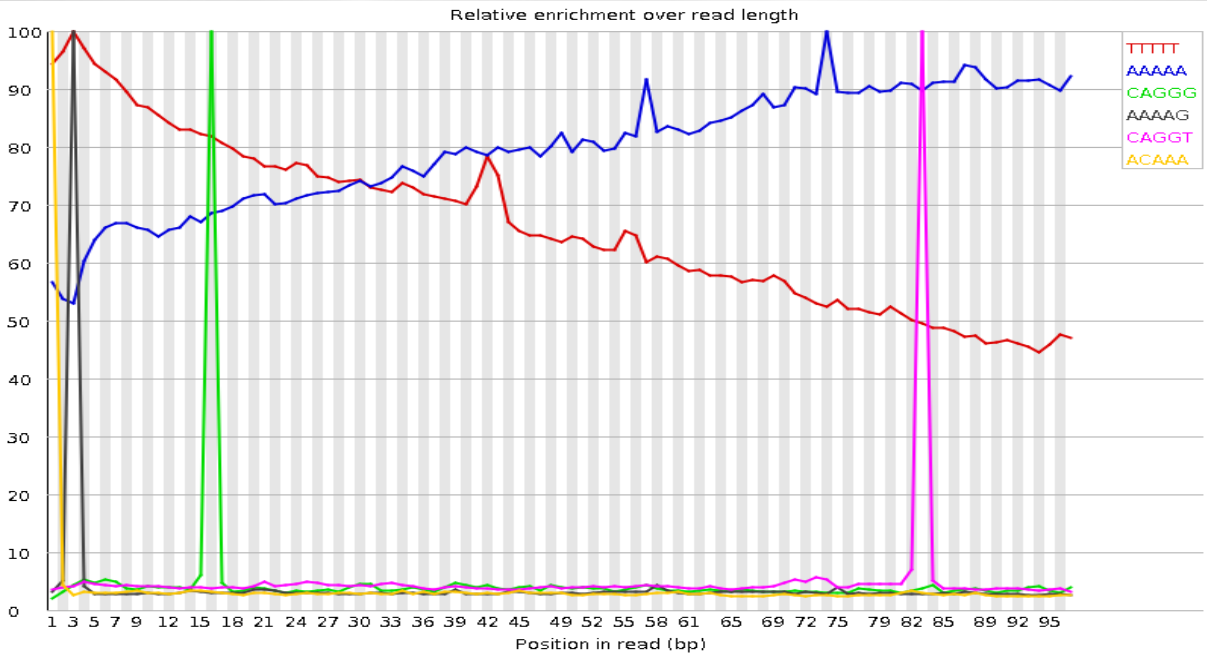
– ***Has been developed for genomic data***

FastQC graphics



FastQC graphics : Kmer content

- A kmer is a subsequence of length k
- Should spot overrepresented sequences, give a good impression of any contamination.
- Kmers showing a rise towards the end of the library indicate progressive contamination with adapters.
- Check for adaptor sequence or poly-A sequence



Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp Position
TTTTT	47499960	4.84021	7.2762637	3
AAAAA	18101385	4.2297845	5.3006034	74
CAGGG	12486915	2.3769662	49.03375	16
AAAAG	10728075	2.3667703	56.233307	3

Take home message on quality analysis

Elements to be checked :

- Random priming effect
- K-mer (polyA, polyT)

Alignment on reference for the second quality check and filtering.

A good run?:

- Expected number of reads produced (2x 2 billions / flowcell),
- Length of the reads expected (150pb),
- Random selection of the nucleotides and the GC%,
- Good alignment: very few unmapped reads, pairs mapped on opposite strands.

Hands-on : data quality

Connection genotoul : *ssh -X nom@genotoul*

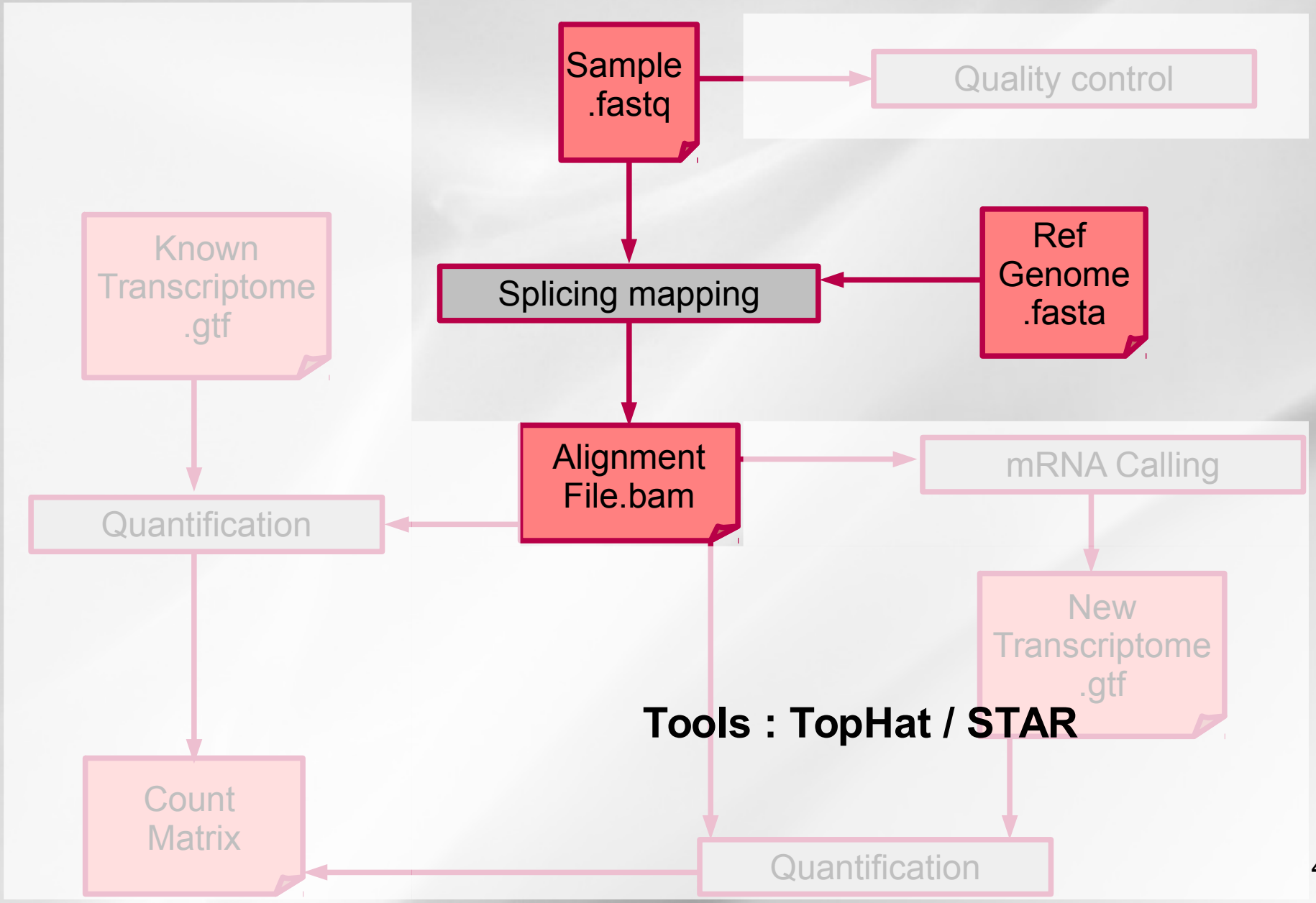
To connect to the processing node : *qlogin*

Training accounts : *anemone* *aster*

bleuet	*iris*
muguet	*narcisse*
pensee	*rose*
tulipe	*violette*

FastQC location : /usr/local/bioinfo/src/FastQC/current/fastqc

Analysis workflow



Aim -

Spliced read mapping & Visualisation

- Discover the true location (origin) of each read with respect to the reference
- Obviously features of the reference (repetitive regions, assembly errors, missing information) will render this objective impossible for a subset of the reads
- Because sequencing library was constructed from transcribed RNA, account for reads that may be split by potentially thousands of bases of intronic sequence
- Take advantage of intron/exon boundary annotations and be able to split reads across exons from no additional information (de novo spliced alignment)
- Do it in/with reasonable time/resources

Summary -

Spliced read mapping & Visualisation

- Reference genome & Reference transcriptome files formats
- What is a spliced aligner ?
- Tophat principle and usage
- BAM & Bed files formats
- STAR usage
- Visualisation with IGV

Where to find a reference genome?

Retrieving the genome file (fasta):

- The Genome Reference Consortium

http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

- ! NCBI chromosome naming with « | » not well supported by mapping software

- Prefer EMBL

http://www.ensembl.org/info/data/ftp/index.html

Reference transcriptome file

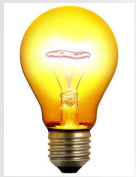
What is a GTF file ?

- derived from GFF (General Feature Format, for description of genes and other features)
- Gene Transfer Format: http://genome.ucsc.edu/FAQ/FAQformat.html#format4

<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes] [comments]

The [attribute] list must begin with:

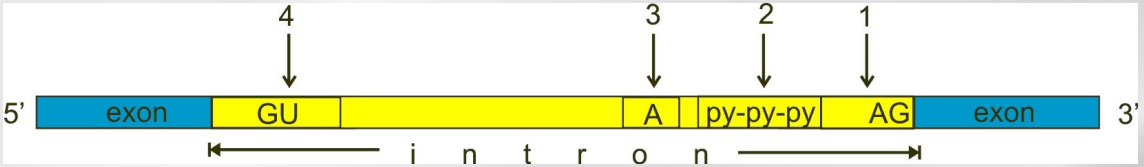
- gene_id value : unique identifier for the genomic source of the sequence.
- transcript_id value : unique identifier for the predicted transcript.



The chromosome name should be the same in the gtf file and fasta file

Splice sites

- Canonical splice site:
  which accounts for more than 99% of splicing
  GT and AG for donor and acceptor sites



http://en.wikipedia.org/wiki/RNA_splicing

- Non-canonical site:
  GC-AG splice site pairs, AT-AC pairs

Nucleic Acids Res. 2000 Nov 1;28(21):4364-75.

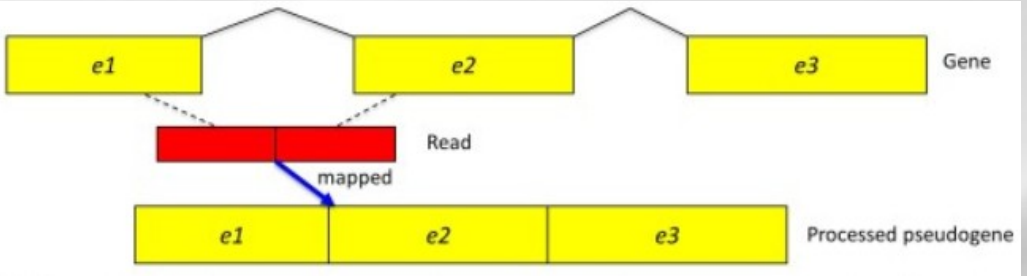
Analysis of canonical and non-canonical splice sites in mammalian genomes.

Burset M, Seledtsov IA, Solovyev VV.

- Trans-splicing:
  splicing that joins two exons that are not within the same RNA transcript

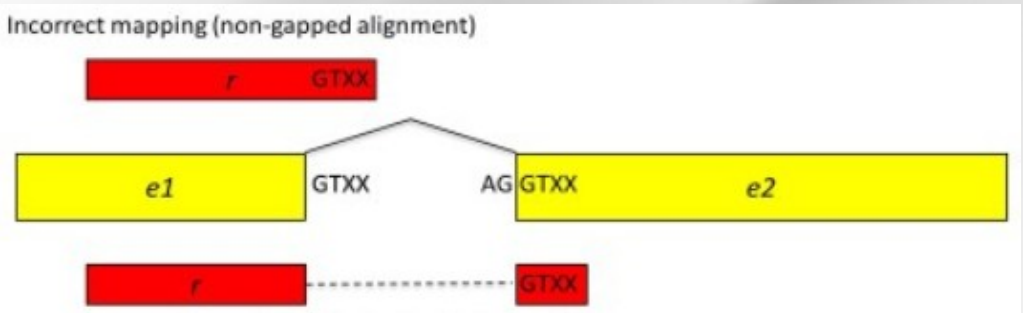
Hard case

- Lot of variations (sequencing errors, mutations)
- Repeats
- Reads spanning 3+ exons
- Gene or pseudogene



Kim et al, Genome Biology, 2013

- Small end “anchor”



- Unknown junction inside poorly rarely expressed gene

Alignment Tools

Tools for splice-mapping:

- Tophat:

BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 9 2009, pages 1105–1111
doi:10.1093/bioinformatics/btp120

Sequence analysis

TopHat: discovering splice junctions with RNA-Seq

Cole Trapnell[1,*], Lior Pachter[2] and Steven L. Salzberg[1]

Genome Biol. 2013 Apr 25;14(4):R36. doi: 10.1186/gb-2013-14-4-r36.

TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL.

- STAR:

STAR: ultrafast universal RNA-seq aligner

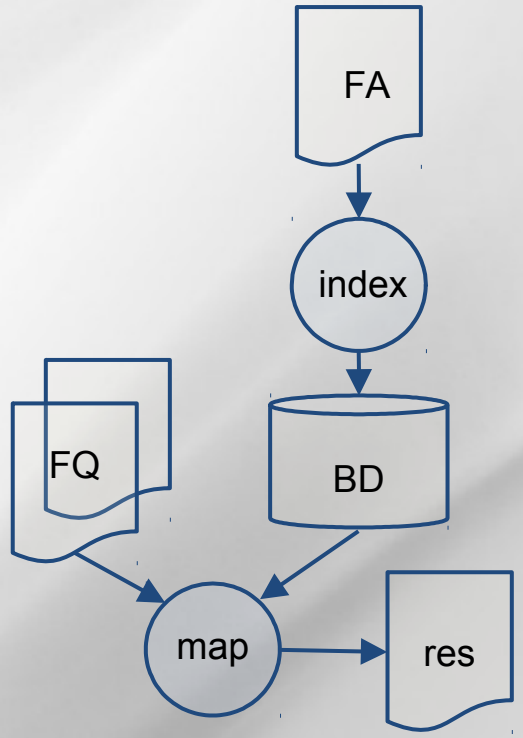
Alexander Dobin[1*], Carrie A. Davis[1], Felix Schlesinger[1], Jorg Drenkow[1], Chris Zaleski[1], Sonali Jha[1], Philippe Batut[1], Mark Chaisson[2] and Thomas R. Gingeras[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA.
[2]Pacific Biosciences, Menlo Park, California, USA.

Associate Editor: Dr. Inanc Birol

Mapping steps

- Indexing reference (once only)
- Mapping reads using index

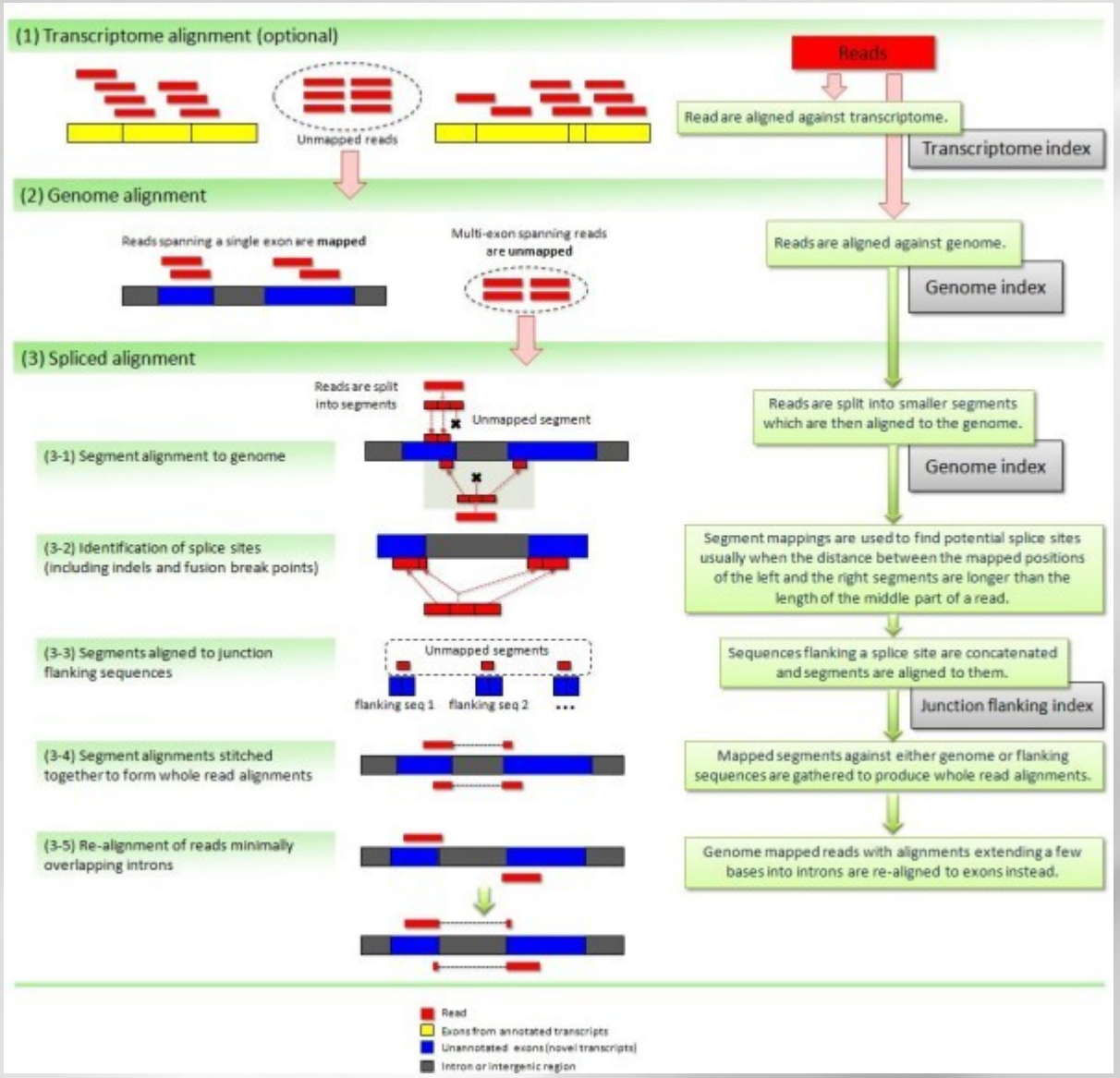


TopHat

http://ccb.jhu.edu/software/tophat

- *Aligns RNA-Seq reads to a reference genome with Bowtie2*
- *splice junction mapper for reads without knowledges*
- *identify splice junctions between exons*

TopHat pipeline



Kim et al, Genome Biology, 2013

- Numerous steps to resolve hard cases
- Each step makes use of heuristics with parameters users have to define a value

TopHat Inputs

Inputs :

- bowtie2 index of the genome

  ftp://ftp.cbcb.umd.edu/pub/data/bowtie_indexes/

  http://bowtie-bio.sourceforge.net/index.shtml

- Fasta file (.fa) of the reference => build index with bowtie
- Fastq file(s) of reads

! the GTF file and the Bowtie index should have same name of chromosome or contig

Command lines :

bowtie2-build <reference.fasta> <index_base>

tophat2 [options] <index_base> <reads1[,reads2,...]> [reads1[,reads2,...]]

TopHat Options

Some useful options (command line) :

-h/--help

-v/--version

--bowtie1 (instead of bowtie2)

-o/--output-dir

-r/--mate-inner-dist [50]

--mate-std-dev [20]

-i/--min-intron-length [50]

-I/--max-intron-length [500000]

-N/--read-mismatches [2]

--read-edit-dist [2]

-p/--num-threads [1]

Special note on the website

Please Note TopHat has a number of parameters and options, and their default values are tuned for processing mammalian RNA-Seq reads.

If you would like to use TopHat for another class of organism, we recommend setting some of the parameters with more strict, conservative values than their defaults.

Usually, setting the maximum intron size to 4 or 5 Kb is sufficient to discover most junctions while keeping the number of false positives low.

More topHat options

Your own junctions :

-G/--GTF *<GTF2.2file>*

-j/--raw-juncs *<.juncs file>*

--no-novel-juncs (ignored without -G/-j)

Your own insertions/deletions:

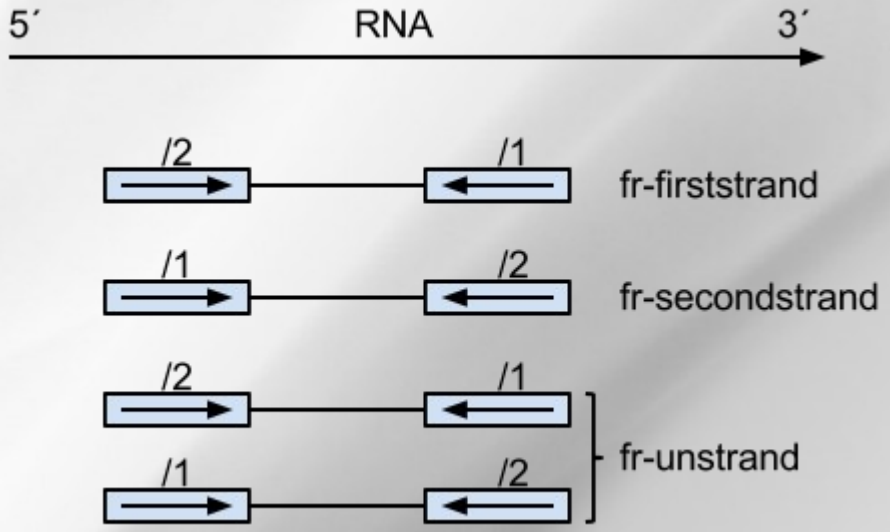
--insertions/--deletions <.juncs file>

--no-novel-indels

Library types

`--library-type` TopHat will treat the reads as strand specific. Every read alignment will have an XS attribute tag. Consider supplying library type options below to select the correct RNA-seq protocol.

Library Type	Examples	Description
fr-unstranded	`Standard Illumina`	Reads from the left-most end of the fragment (in transcript coordinates) map to the transcript strand, and the right-most end maps to the opposite strand.
fr-firststrand	`dUTP, NSR, NNSR`	Same as above except we enforce the rule that the right-most end of the fragment (in transcript coordinates) is the first sequenced (or only sequenced for single-end reads). Equivalently, it is assumed that only the strand generated during first strand synthesis is sequenced.
fr-secondstrand	`Ligation, Standard SOLiD`	Same as above except we enforce the rule that the left-most end of the fragment (in transcript coordinates) is the first sequenced (or only sequenced for single-end reads). Equivalently, it is assumed that only the strand generated during second strand synthesis is sequenced.



TopHat Outputs

Outputs :

- ***accepted_hits.bam***: list of read alignments in SAM format compressed
- ***junctions.bed***: track of junctions, scores : number of alignments spanning the junction
- ***insertions.bed*** and ***deletions.bed*** : tracks of insertions and deletions
- **logs:** directory files
- **unmapped.bam**: unmapped or multi-mapped (over the threshold) reads
- **prep_reads.info**: number of reads and read length for input and output
- **align_summary.txt**

Spliced cigar line

- Extend CIGAR strings

op	Description
M	Alignment match (can be a sequence match or mismatch
I	Insertion to the reference
D	Deletion from the reference
N	Skipped region from the reference
S	Soft clip on the read (clipped sequence present in <seq>)
H	Hard clip on the read (clipped sequence NOT present in <seq>)
P	Padding (silent deletion from the padded reference sequence)

- Example: intron de 81 bases

flag chr pos pair

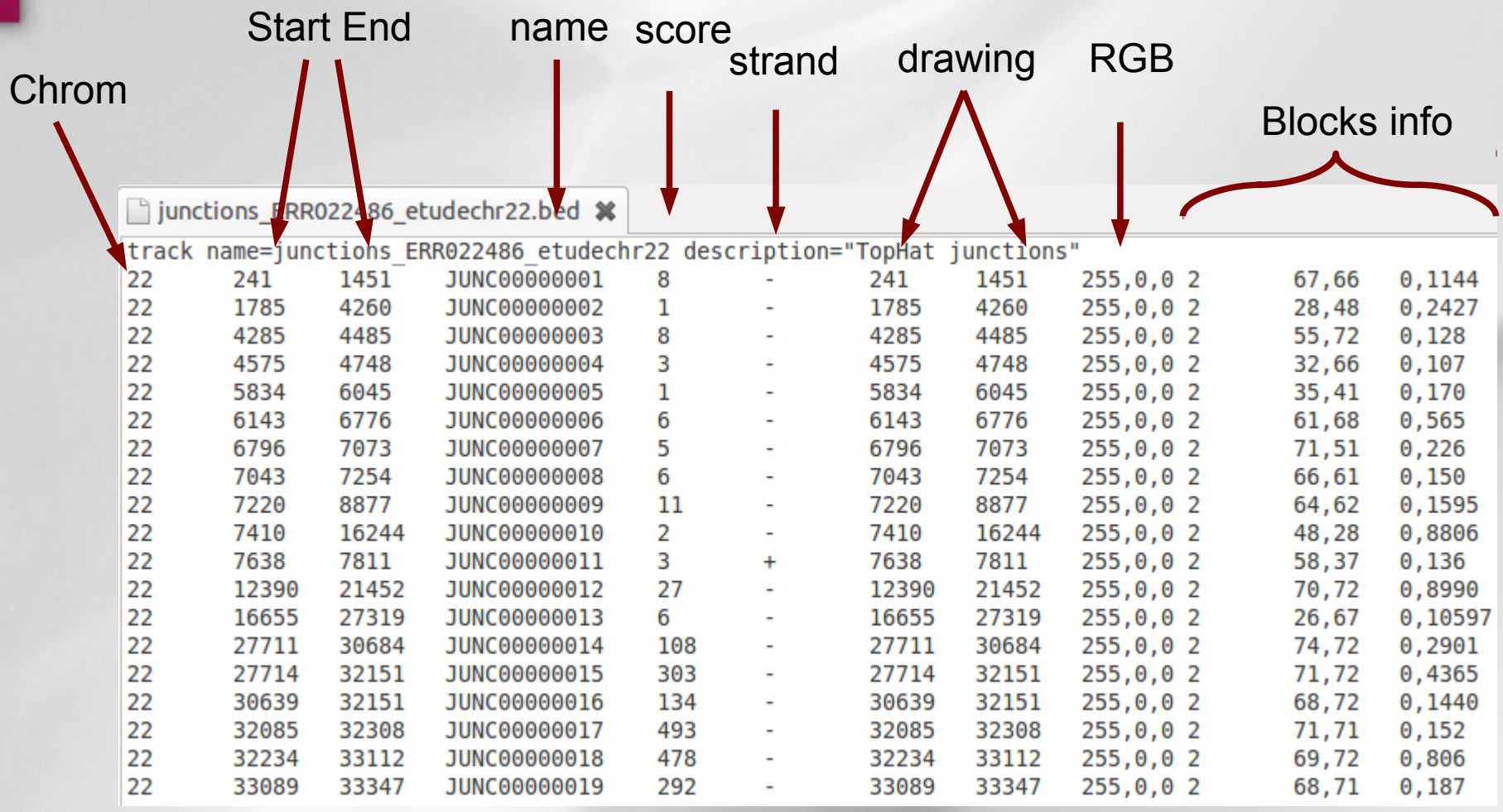
ERR022486.8388510 81 22 32099 255 **58M81N18M** = 27484 -4772
CCTTGGTCTTGCCGAAGTAGATCTCATTGAGAGTGGAGCGGATCTTGTTCTCCATTTCCTCCACC
AGGCGTCCGAT :9=<==;<<><=><?>>?<?==>>?>><?>>??<AA?
@AFADDD;GDGAG@GGCBE@GG?GG>GGGG?GGGGGGGG NM:i:0 XS:A:- NH:i:1

Bam & Bed

- BAM (Binary Alignment/Map) format:
  - Compressed binary representation of SAM
  - Greatly reduces storage space requirements to about 27% of original SAM
  - Bamtools: reading, writing, and manipulating BAM files

- Bed (Browser Extensible Data) format:
  - tab-delimited text file that defines a feature track
    http://genome.ucsc.edu/FAQ/FAQformat.html#format1
  - The first three required BED fields are:
    <chromosome> <start> <end>
  - 9 additional optional BED fields

Bed exemple




```
track name=junctions_ERR022486_etudechr22 description="TopHat junctions"
22      241     1451    JUNC00000001    8       -       241     1451    255,0,0 2       67,66   0,1144
22      1785    4260    JUNC00000002    1       -       1785    4260    255,0,0 2       28,48   0,2427
22      4285    4485    JUNC00000003    8       -       4285    4485    255,0,0 2       55,72   0,128
22      4575    4748    JUNC00000004    3       -       4575    4748    255,0,0 2       32,66   0,107
22      5834    6045    JUNC00000005    1       -       5834    6045    255,0,0 2       35,41   0,170
22      6143    6776    JUNC00000006    6       -       6143    6776    255,0,0 2       61,68   0,565
22      6796    7073    JUNC00000007    5       -       6796    7073    255,0,0 2       71,51   0,226
22      7043    7254    JUNC00000008    6       -       7043    7254    255,0,0 2       66,61   0,150
22      7220    8877    JUNC00000009    11      -       7220    8877    255,0,0 2       64,62   0,1595
22      7410    16244   JUNC00000010    2       -       7410    16244   255,0,0 2       48,28   0,8806
22      7638    7811    JUNC00000011    3       +       7638    7811    255,0,0 2       58,37   0,136
22      12390   21452   JUNC00000012    27      -       12390   21452   255,0,0 2       70,72   0,8990
22      16655   27319   JUNC00000013    6       -       16655   27319   255,0,0 2       26,67   0,10597
22      27711   30684   JUNC00000014    108     -       27711   30684   255,0,0 2       74,72   0,2901
22      27714   32151   JUNC00000015    303     -       27714   32151   255,0,0 2       71,72   0,4365
22      30639   32151   JUNC00000016    134     -       30639   32151   255,0,0 2       68,72   0,1440
22      32085   32308   JUNC00000017    493     -       32085   32308   255,0,0 2       71,71   0,152
22      32234   33112   JUNC00000018    478     -       32234   33112   255,0,0 2       69,72   0,806
22      33089   33347   JUNC00000019    292     -       33089   33347   255,0,0 2       68,71   0,187
```

Tophat technical issues

- Temporary disk space
  - 100 000 000 pair-ends = 0,5 To of temporary disk space
- Number of cpus
  - 100 000 000 pair-ends = 5-7 cpu days on the local cluster
- New platform cluster:
  - 34 cluster nodes with 4*12 cores and 384 GB of ram per node: 1632 cores
  - 1 hypermem node (32 cores and 1024 GB of ram)
  - A scratch file system (157 To available, 6 Gbps bandwith)

An other aligner : STAR

OXFORD JOURNALS Bioinformatics (Oxford, England)

Bioinformatics. 2013 Jan; 29(1): 15–21. PMCID: PMC3530905
Published online 2012 Oct 25. doi: 10.1093/bioinformatics/bts635

STAR: ultrafast universal RNA-seq aligner

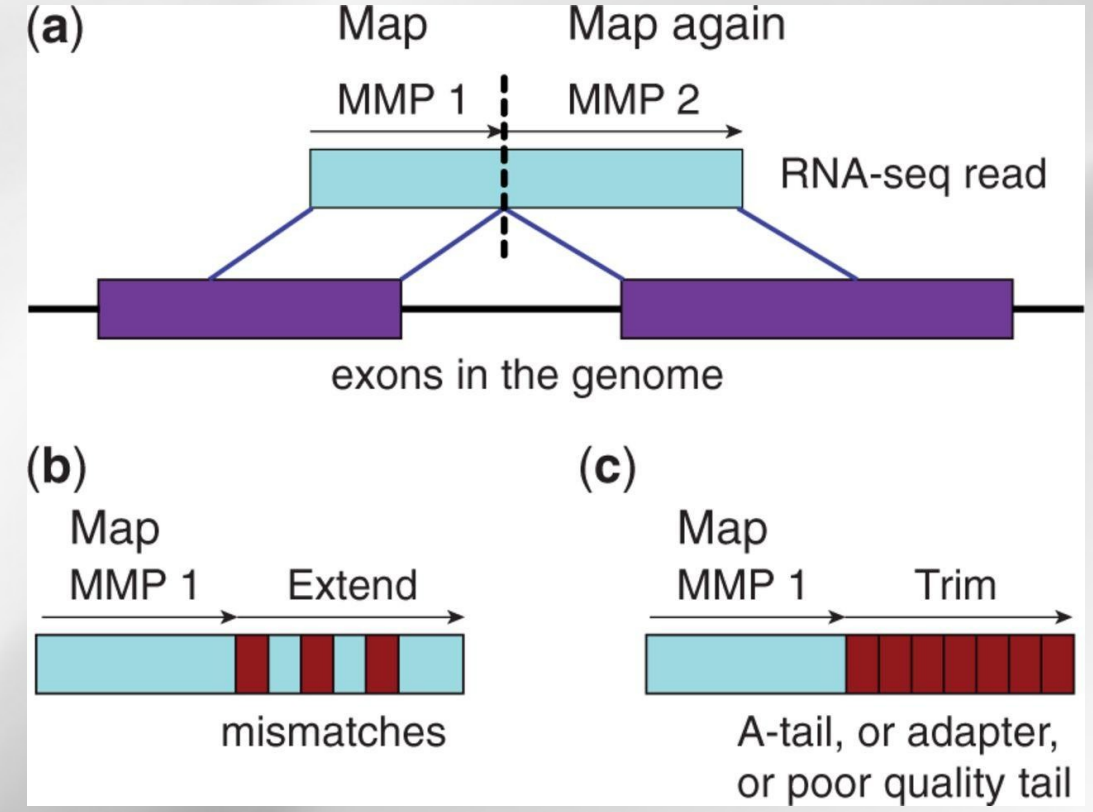
Alexander Dobin,[1,*] Carrie A. Davis,[1] Felix Schlesinger,[1] Jorg Drenkow,[1] Chris Zaleski,[1] Sonali Jha,[1] Philippe Batut,[1] Mark Chaisson,[2] and Thomas R. Gingeras[1]

- Spliced Transcripts Alignment to a Reference
- Outperforms other aligners by more than a factor of 50 in mapping speed

STAR

Another strategy:

- search for a MMP from the 1rst base
- MMP search repeated for the unmapped portion next to the junction
- do it in both fwd and rev directions
- cluster seeds from the mates of paired-end RNA-seq reads



Soft-clipping is the main difference between Tophat and STAR

Dobin et al, Bioinformatics, 2011

STAR Inputs

Inputs :

- STAR index of the genome

  ftp://ftp2.cshl.edu/gingeraslab/tracks/STARrelease/STARgenomes/

- Fasta file (.fa) of the reference to index

- Fastq file(s) of reads

Command lines :

STAR --runMode genomeGenerate --genomeDir /path/to/GenomeDir
--genomeFastaFiles /path/to/genome/fasta1 /path/to/genome/fasta2
--runThreadN <n>

STAR --genomeDir /path/to/GenomeDir --readFilesIn /path/to/read1 [/path/to/read2]
--runThreadN <n> --<inputParameterName> <input parameter value(s)>

STAR Options

Input options:

--readFilesCommand zcat

Intron options: genomic gap is considered intron if

--alignIntronMin [21]

--alignIntronMax [500000]

Filter output options :

--outFilterMismatchNmax [10]: has fewer mismatches than this value

Output format options :

--outSAMattributes All

--outFileNamePrefix

--outSAMstrandField intronMotif [None] Required for cufflinks

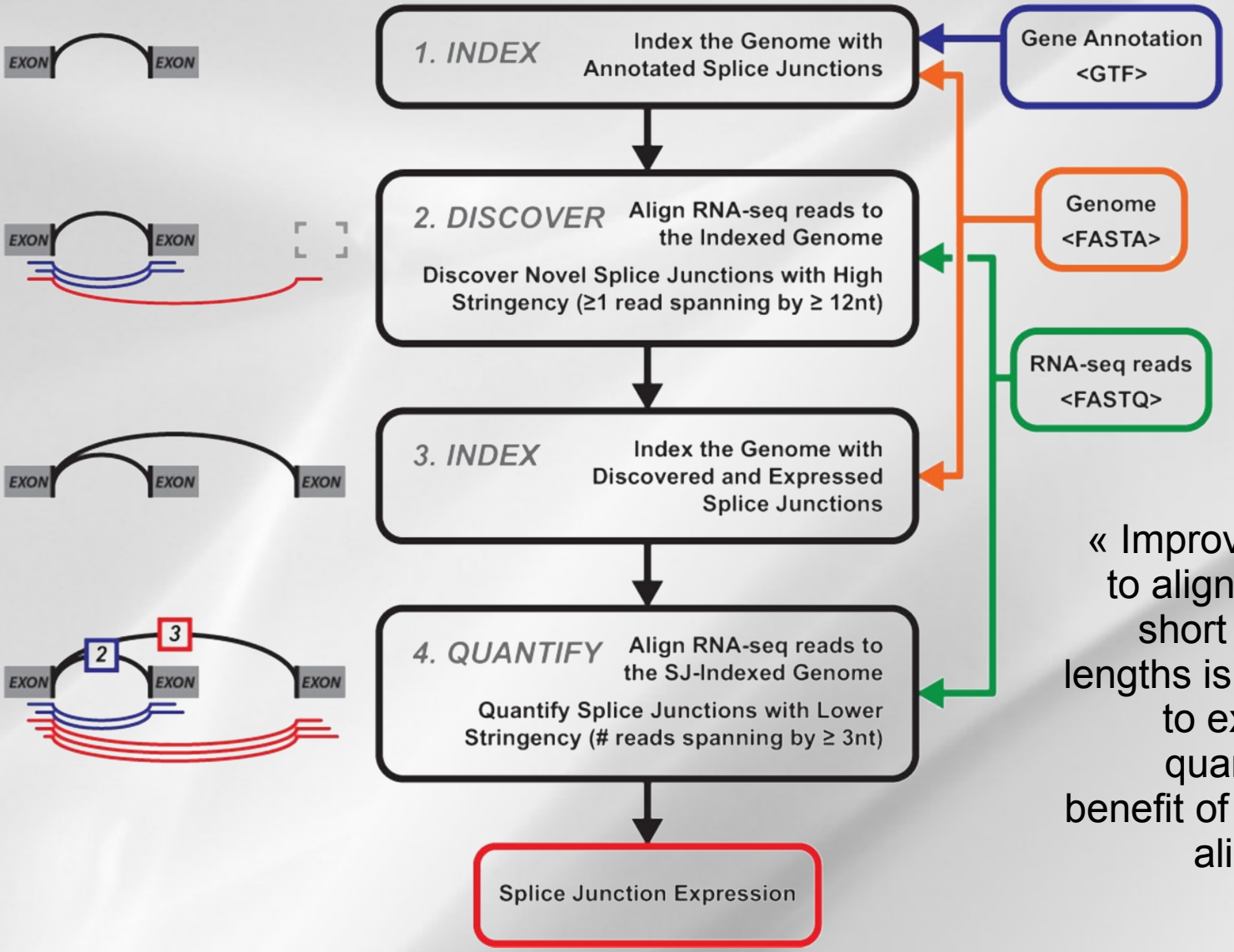
--outSAMtype BAM SortedByCoordinate [SAM]

STAR Outputs

Outputs (w/o specific options except BAM SortedByCoordinate):

- **Aligned.sortedByCoord.out.bam**: list of read alignments in SAM format compressed
- **Log.out**: main log file with a lot of detailed information about the run (for troubleshooting)
- **Log.progress.out**: reports job progress statistics
- **Log.final.out**: summary mapping statistics after mapping job is complete, very useful for quality control.
- **SJ.out.tab**: contains high confidence collapsed splice junctions in tab-delimited format

Two passes strategy



« Improved ability to align reads by short spanning lengths is sufficient to explain the quantification benefit of two-pass alignment »

Veeneman et al, Bioinformatics, 2016

Existing tools

How to compare tools ?

- sensibility (maximize #mapped reads)
- specificity (assign reads to the correct position)

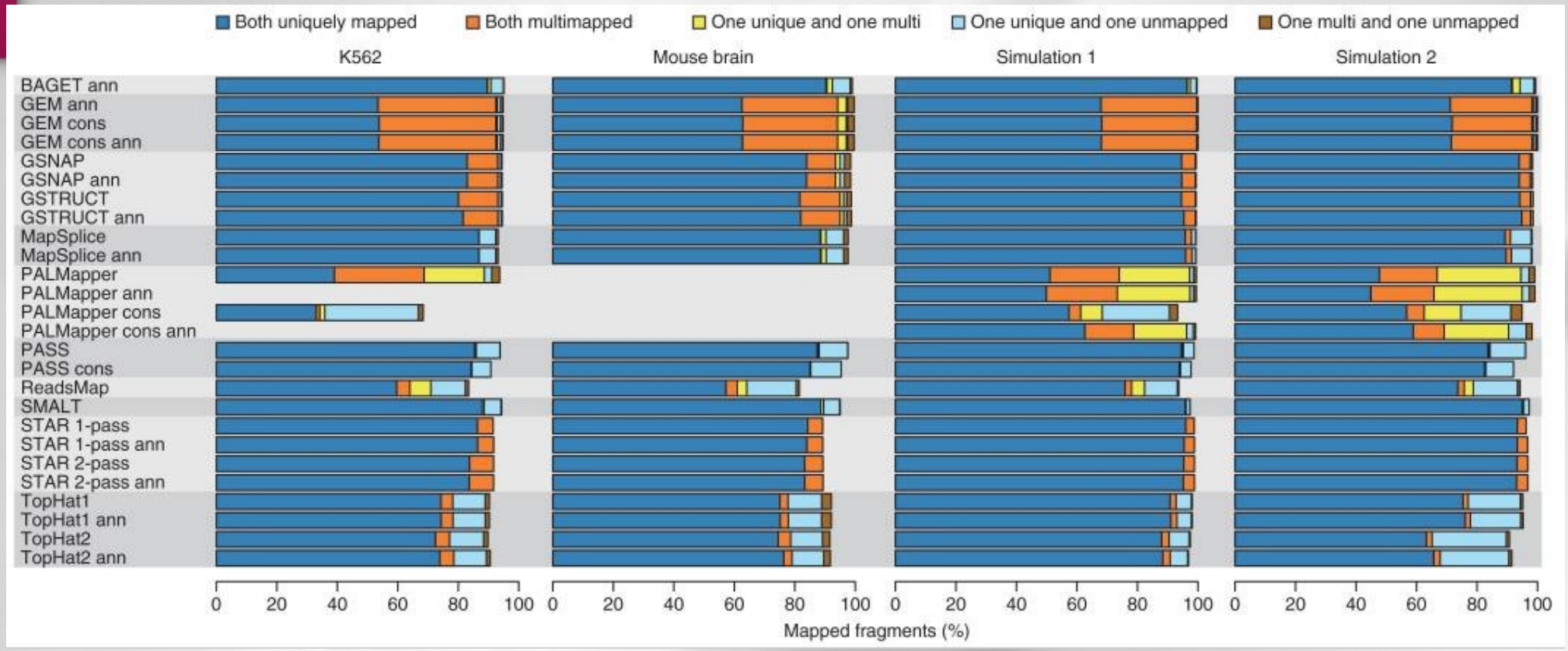
→ for reads and for junctions

- processing time
- memory requirement

All of these are conflicting criteria ...

RGASP3

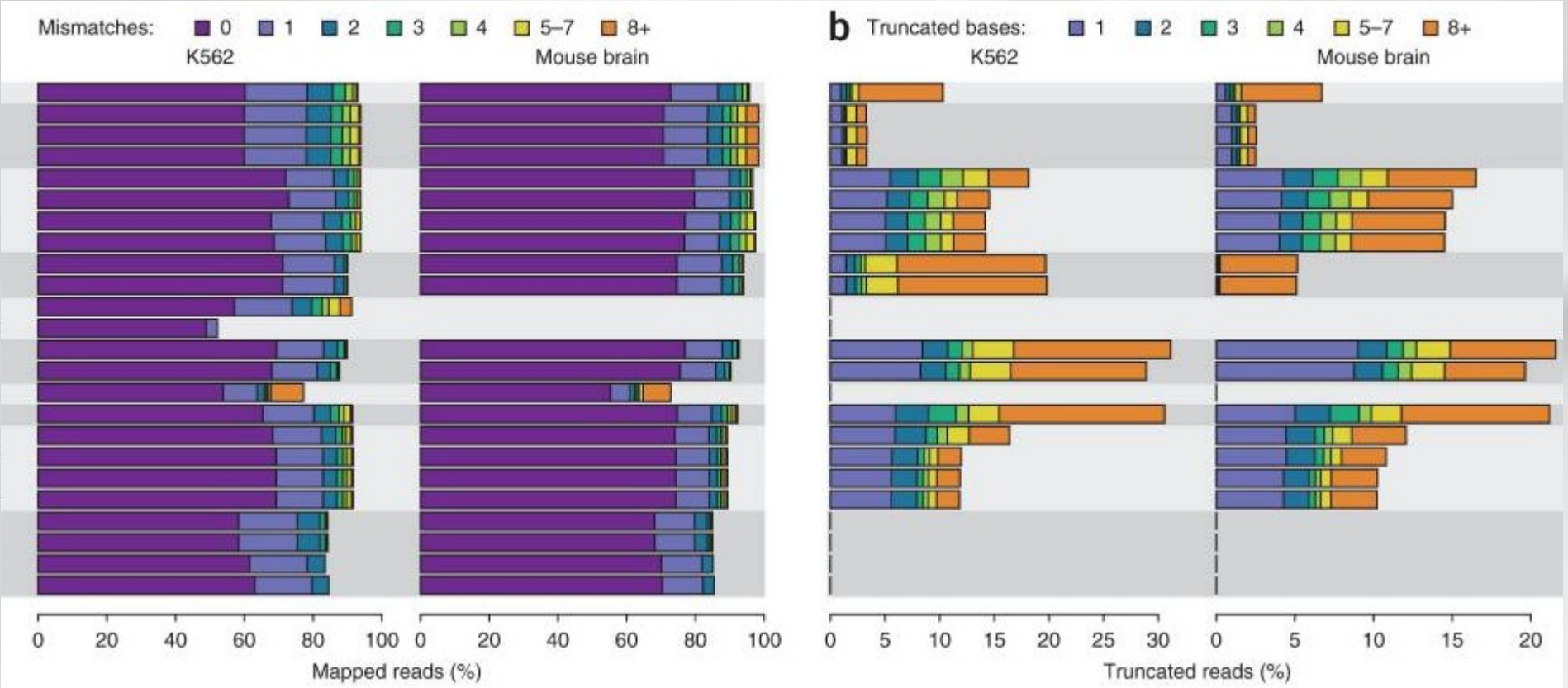
The RNA-seq Genome Annotation Assessment Project



Engström et al., Nature Methods, 2013

RGASP3

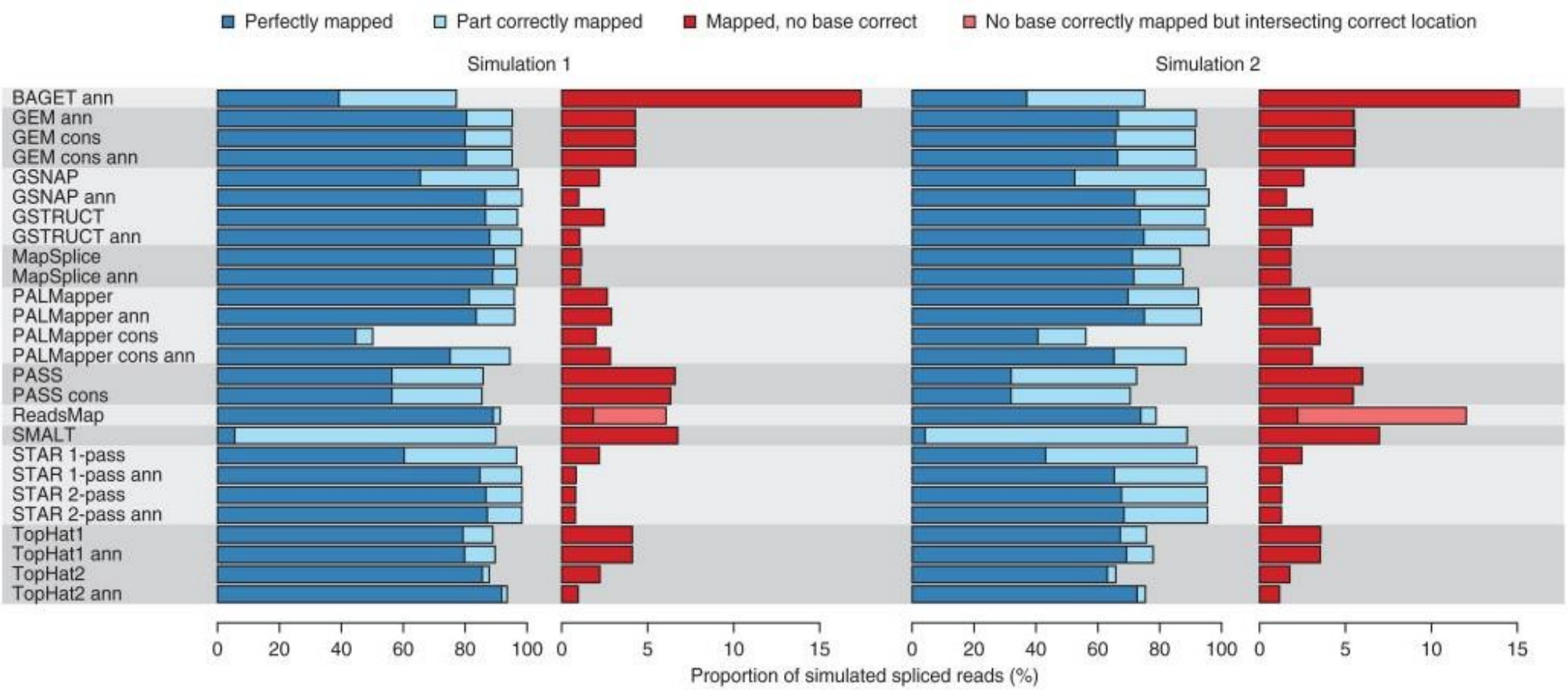
The RNA-seq Genome Annotation Assessment Project



Engström et al., Nature Methods, 2013

RGASP3

The RNA-seq Genome Annotation Assessment Project



Engström et al., Nature Methods, 2013

TopHat vs STAR

The RNA-seq Genome Annotation Assessment Project

STAR	vs	TopHat2
+	# lectures alignées	-
-	# lectures correctement alignées	+
-	Sensibilité aux variations	+
-	Sensibilité aux annotations	+

Visualizing alignments on IGV

http://www.broadinstitute.org/igv/home

NATURE BIOTECHNOLOGY | OPINION AND COMMENT | CORRESPONDENCE

Integrative genomics viewer

James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz & Jill P Mesirov

Affiliations | Corresponding authors

Nature Biotechnology **29**, 24–26 (2011) | doi:10.1038/nbt.1754
Published online 10 January 2011

hands-on : tophat

Example of used commands:

bowtie2-build ITAG2.3_genomic_Ch6.fasta index-bowtie2/tomato_chr6

qsub -N tophat_wt -pe parallel_smp 4 -b Y 'tophat2 -o aln_tophat_wt --max-intron-length 5000 --mate-inner-dist 200 bowtie2-index/tomato_chr6 WT_rep1_1_Ch6.fastq.gz WT_rep1_2_Ch6.fastq.gz '

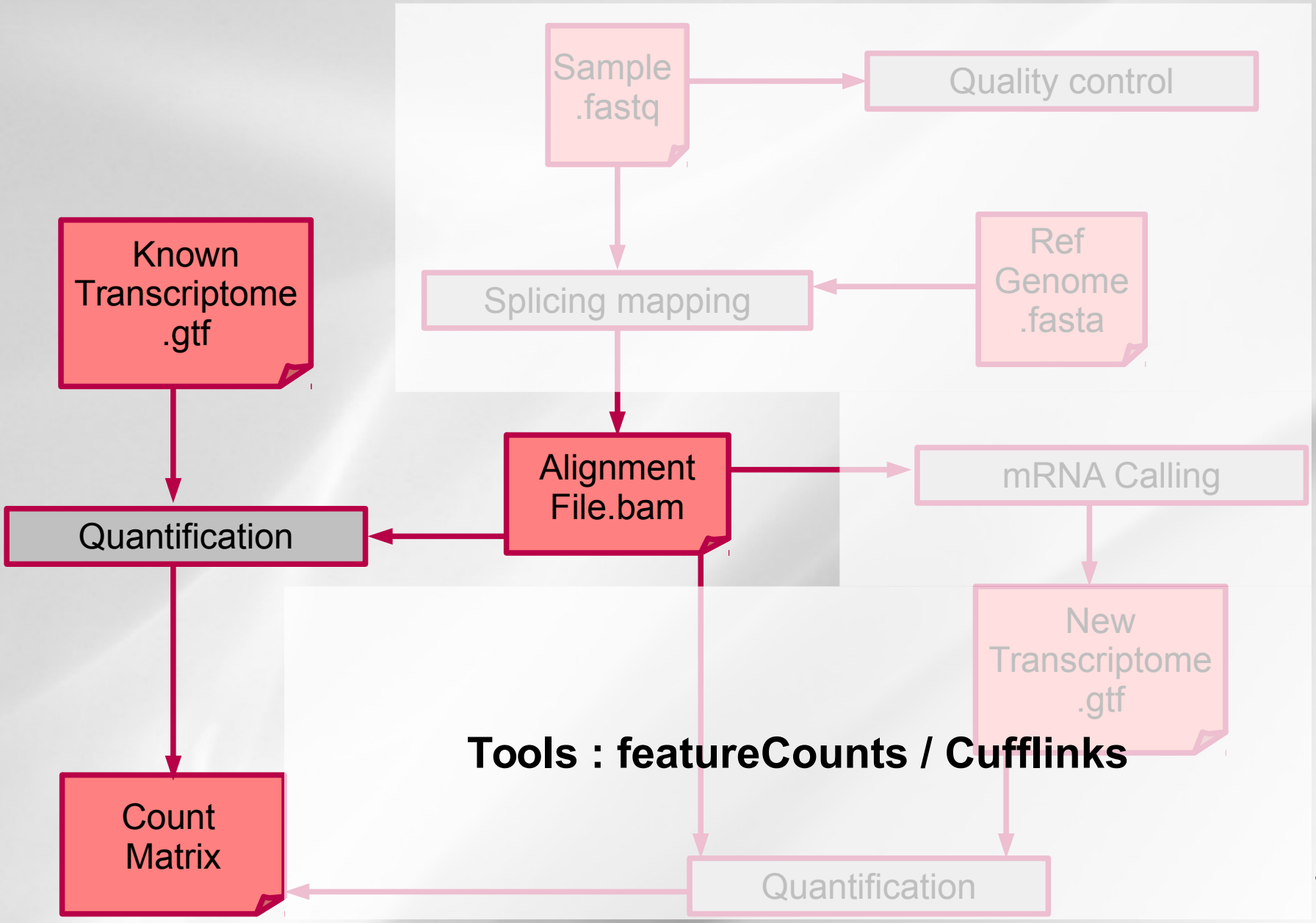
samtools index file.bam

samtools view file.bam | cut -f 1 | sort | uniq -c | cut -c 1-7 | sort -n | uniq -c

Or

samtools flagstat file.bam

Analysis workflow



Summary - Quantification

- What do we want ?

- Raw count :
  - FeatureCounts (Gene/Transcript level) usage
- Abundance estimation
  - Cufflinks (Gene/Transcript level) principle and usage

What do we want to build?

The gene / transcript description file (and corresponding fasta)

```
9   protein_coding  exon    697785  697947  .   -   .   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000144625"; exon_number "1"
9   protein_coding  exon    696518  696600  .   -   .   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000144625"; exon_number "2"
9   protein_coding  exon    694364  694502  .   -   .   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000144625"; exon_number "3"
9   protein_coding  CDS     694364  694497  .   -   0   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000144625"; exon_number "3"
9   protein_coding  start_codon    694495  694497  .   -   0   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000144625"; exon_number "3"
9   protein_coding  exon    693528  693822  .   -   .   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000144625"; exon_number "4"
9   protein_coding  CDS     693675  693822  .   -   1   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000144625"; exon_number "4"
9   protein_coding  stop_codon    693672  693674  .   -   0   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000144625"; exon_number "4"
9   protein_coding  exon    694364  694497  .   -   .   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000112112"; exon_number "1"
9   protein_coding  CDS     694364  694497  .   -   0   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000112112"; exon_number "1"
9   protein_coding  start_codon    694495  694497  .   -   0   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000112112"; exon_number "1"
9   protein_coding  exon    693672  693822  .   -   .   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000112112"; exon_number "2"
9   protein_coding  CDS     693675  693822  .   -   1   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000112112"; exon_number "2"
9   protein_coding  stop_codon    693672  693674  .   -   0   gene_id "ENSDARG00000075709"; transcript_id "ENSDART00000112112"; exon_number "2"
9   protein_coding  exon    697453  697832  .   +   .   gene_id "ENSDARG00000011999"; transcript_id "ENSDART00000136627"; exon_number "1"
9   protein_coding  CDS     697623  697832  .   +   0   gene_id "ENSDARG00000011999"; transcript_id "ENSDART00000136627"; exon_number "1"
9   protein_coding  start_codon    697623  697625  .   +   0   gene_id "ENSDARG00000011999"; transcript_id "ENSDART00000136627"; exon_number "1"
9   protein_coding  exon    698442  698573  .   +   .   gene_id "ENSDARG00000011999"; transcript_id "ENSDART00000136627"; exon_number "2"
9   protein_coding  CDS     698442  698573  .   +   0   gene_id "ENSDARG00000011999"; transcript_id "ENSDART00000136627"; exon_number "2"
9   protein_coding  exon    699401  699469  .   +   .   gene_id "ENSDARG00000011999"; transcript_id "ENSDART00000136627"; exon_number "3"
9   protein_coding  CDS     699401  699469  .   +   0   gene_id "ENSDARG00000011999"; transcript_id "ENSDART00000136627"; exon_number "3"
9   protein_coding  exon    700666  700876  .   +   .   gene_id "ENSDARG00000011999"; transcript_id "ENSDART00000136627"; exon_number "4"
9   protein_coding  CDS     700666  700725  .   +   0   gene_id "ENSDARG00000011999"; transcript_id "ENSDART00000136627"; exon_number "4"
```

The count file

	row.names	SRR519727	SRR519728	SRR519729	SRR519730	SRR519731	SRR519747	SRR519748	SRR519749	SRR519750	SRR519751
1	mira_c1	1855	4095	4693	4407	3826	1749	4355	3679	4396	4066
2	mira_c2	358	616	929	834	854	393	769	644	1015	732
3	mira_c3	1874	1392	2583	1333	1245	2890	5104	4052	12012	4150
4	mira_rep_c4	697	789	1044	1100	1363	657	1001	836	1289	1313
5	mira_rep_c5	5765	12517	17170	16120	15121	6042	16388	14329	18505	16999
6	mira_rep_c6	2165	4727	6457	5312	4960	2399	7010	5196	8063	6718
7	mira_rep_c7	260	436	637	627	694	247	689	522	928	940
8	mira_rep_c8	616	1425	1906	1897	2050	691	1537	1551	1667	1552
9	mira_rep_c9	786	1885	2739	2493	2573	735	2345	2012	3308	2645
10	mira_rep_c10	311	517	684	886	895	346	659	581	1041	1030
11	mira_rep_c11	51	212	234	210	175	68	192	261	209	299
12	mira_rep_c12	1129	2191	2833	3128	3088	1139	2983	2575	4384	3811
13	mira_rep_c13	536	913	944	1256	1275	515	1029	913	1407	1444
14	mira_rep_c15	4678	13751	18095	16722	16476	4962	16867	14581	17733	18771
15	mira_rep_c16	7209	22856	32768	28699	27176	8532	28567	25091	35040	30702
16	mira_rep_c17	945	1566	2066	2530	3372	860	1704	1451	3327	3498
17	mira_rep_c18	4419	5668	7750	8570	9559	3954	6610	6180	8273	8728
18	mira_rep_c19	1765	2941	4757	4265	4062	1652	4604	3568	4983	4202
19	mira_rep_c20	1236	2314	3180	2903	2605	818	2196	1843	2478	2410
20	mira_rep_c22	2315	4329	5360	5760	5582	2471	5163	5061	5906	6482
21	mira_rep_c24	4488	7523	11333	10104	9537	4409	8676	9297	9060	10178
22	mira_rep_c25	448	702	944	1155	1245	338	885	740	1680	1599
23	mira_rep_c26	1307	2569	3436	3231	3009	1310	2907	2785	2989	3267
24	mira_c27	766	889	1283	1364	1577	820	1224	1100	1530	1436

If you have the model file

The model is presented in the GTF file (Gene Transfer Format)

Two approaches

- Gene level
- Transcript level

```
1    ensembl    gene  1735  16308 .    +    .    gene_id "ENSGALG00000009771"; gene_version "4"; gene_source "ensembl"; gen
1    ensembl    transcript  1735  16308 .    +    .    gene_id "ENSGALG00000009771"; gene_version "4"; transcript_id "ENSGA
1    ensembl    exon  1735  2449  .    +    .    gene_id "ENSGALG00000009771"; gene_version "4"; transcript_id "ENSGALT0000
1    ensembl    CDS   2379  2449  .    +    0    gene_id "ENSGALG00000009771"; gene_version "4"; transcript_id "ENSGALT0000
1    ensembl    start_codon 2379  2381  .    +    0    gene_id "ENSGALG00000009771"; gene_version "4"; transcript_id "ENSGA
1    ensembl    exon  9272  9489  .    +    .    gene_id "ENSGALG00000009771"; gene_version "4"; transcript_id "ENSGALT0000
1    ensembl    CDS   9272  9489  .    +    1    gene_id "ENSGALG00000009771"; gene_version "4"; transcript_id "ENSGALT0000
```

If you don't have the model file, you'll need to build it.

featureCounts

featureCounts: an efficient general purpose program for assigning sequence reads to genomic features

Yang Liao[1,2], Gordon K. Smyth[1,3] and Wei Shi[1,2,*]

[1]Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, [2]Department of Computing and Information Systems and [3]Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia

Associate Editor: Martin Bishop

- Levels : exon, transcipt, gene
- Multiple option for :
  - Paired reads
  - Assignation of reads
  - Oriented library
- Also exists HTseq-Count

featureCounts

Command line:

featureCounts [options] -a <annotation_file> -o <output_file> input_file1 [input_file2]

Inputs :

- Gtf : annotation file (-a)
- Bams: input files

Some options :

-t Specify the feature type. Only rows which have the matched matched feature type in the provided GTF annotation file will be included for read counting. `exon' by default.

-g Specify the attribute type used to group features (eg. Exons) into meta-features (eg. genes), when GTF annotation is provided.`gene_id' by default. This argument is useful for the meta-feature level summarization.

featureCounts

-Q The minimum mapping quality score a read must satisfy in order to be counted. For paired-end reads, at least one end should satisfy this criteria. 0 by default.

--primary If specified, only primary alignments will be counted.

--minReadOverlap Specify the minimum number of overlapped bases required to assign a read to a feature. 1 by default.

-p If specified, fragments (or templates) will be counted instead of reads.

-P If specified, paired-end distance will be checked when assigning

-d Minimum fragment/template length, 50 by default.

-D Maximum fragment/template length, 600 by default.

-B If specified, only fragments that have both ends successfully aligned will be considered for summarization.

Cufflinks in general

NATURE BIOTECHNOLOGY | RESEARCH | LETTER

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold & Lior Pachter

Affiliations | Contributions | Corresponding author

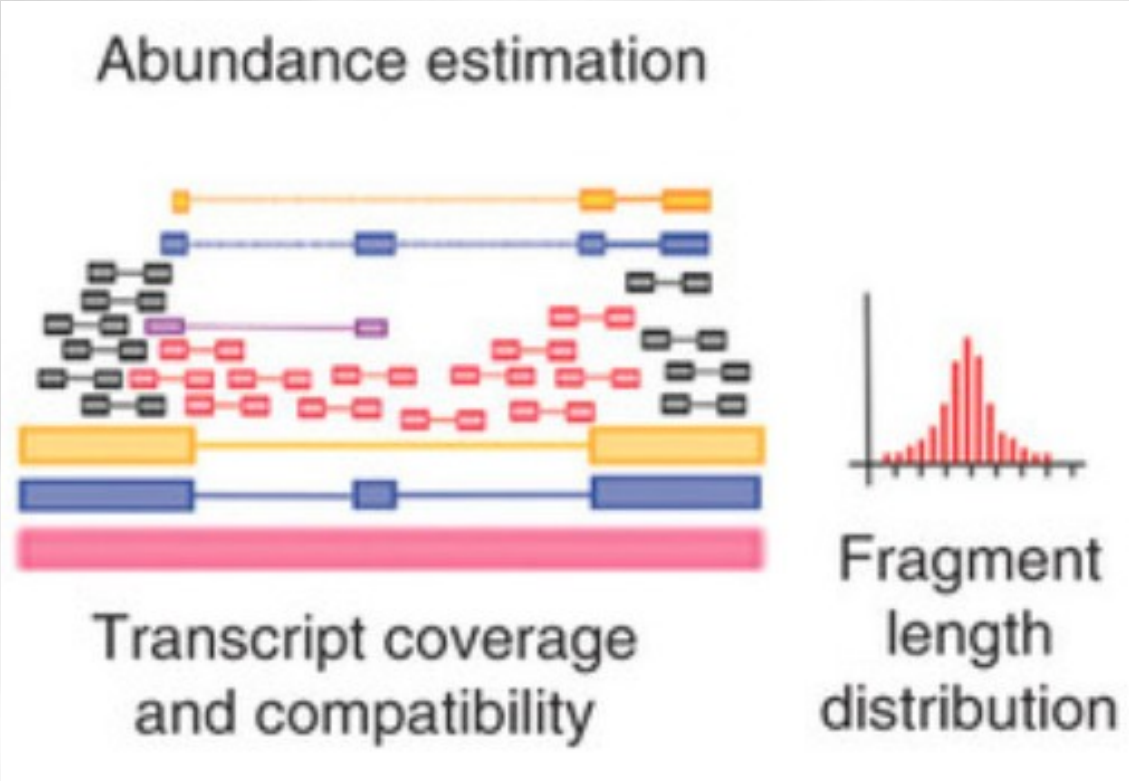
Nature Biotechnology **28**, 511–515 (2010) | doi:10.1038/nbt.1621
Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

http://cufflinks.cbcb.umd.edu/

- *assembles transcripts*
- **estimates their abundances : based on how many reads support each one**
- tests for differential expression in RNA-Seq samples

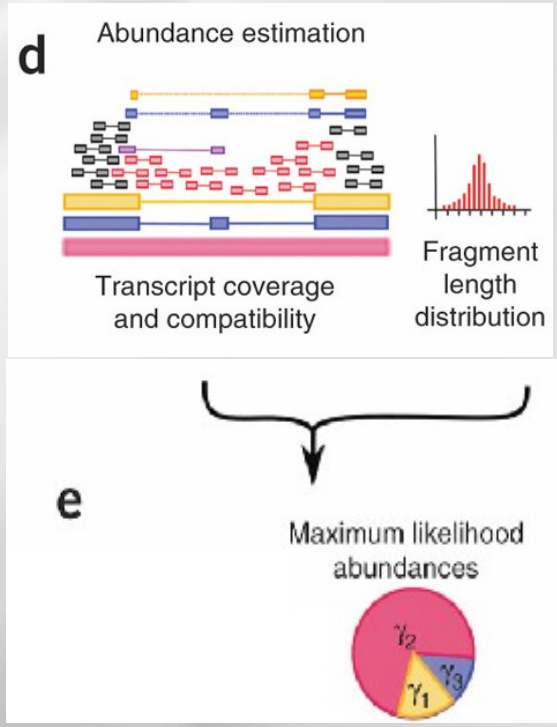
Cufflinks read attribution

- Violet fragment: from which transcript?
  - Use of Fragment length distribution



Cufflinks expression measurement

- Fragments attribution
- Isoforms abundances estimation:
  - RPKM for single reads
  - FPKM for paired-end reads



Trapnell C et al. Nature Biotechnology 2010;28:511-515

RPKM / FPKM

- Transcript length bias
- **RPKM** : Reads per kilobase of exon per million mapped reads

  R = Number of mapped reads

  N = Total mapped reads

  L = Exon gene length in bp

$$RPKM = \frac{10^9 \times R}{N \times L}$$

If my gene length (L) is : 200pb
Number of reads mapped (C) : 400
Total mapped reads (sum for all genes) (N) : 10^8

*RPKM = (10^9 * 400)/(10^8 * 200) = 20*

- the transcript length depends on isoform inference
- **FPKM** : for paired-end sequencing
  - A pair of reads constitute one fragment

Cufflinks inputs and options

- Command line:
  - *cufflinks [options]* <aligned_reads.(sam/bam)>*

- *Some options :*

  -h/--help

  -o/--output-dir

  -p/--num-threads

  ***-G/--GTF <reference_annotation.(gtf/gff)>* : estimate isoform expression, no assembly novel transcripts**

Cufflinks outputs

- **genes.fpkm_tracking**:
  - contains the estimated gene-level expression values in the generic FPKM Tracking Format

Quantification status

tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	status	FPKM	FPKM_conf_lo	FPKM_conf_hi
CUFF.560	-	-	CUFF.560	-	-	22:9743034-9762309	-	-	OK	105.69	77.9404	133.439

- **isoforms.fpkm_tracking**: contains the estimated isoform-level expression values in the generic FPKM Tracking Format

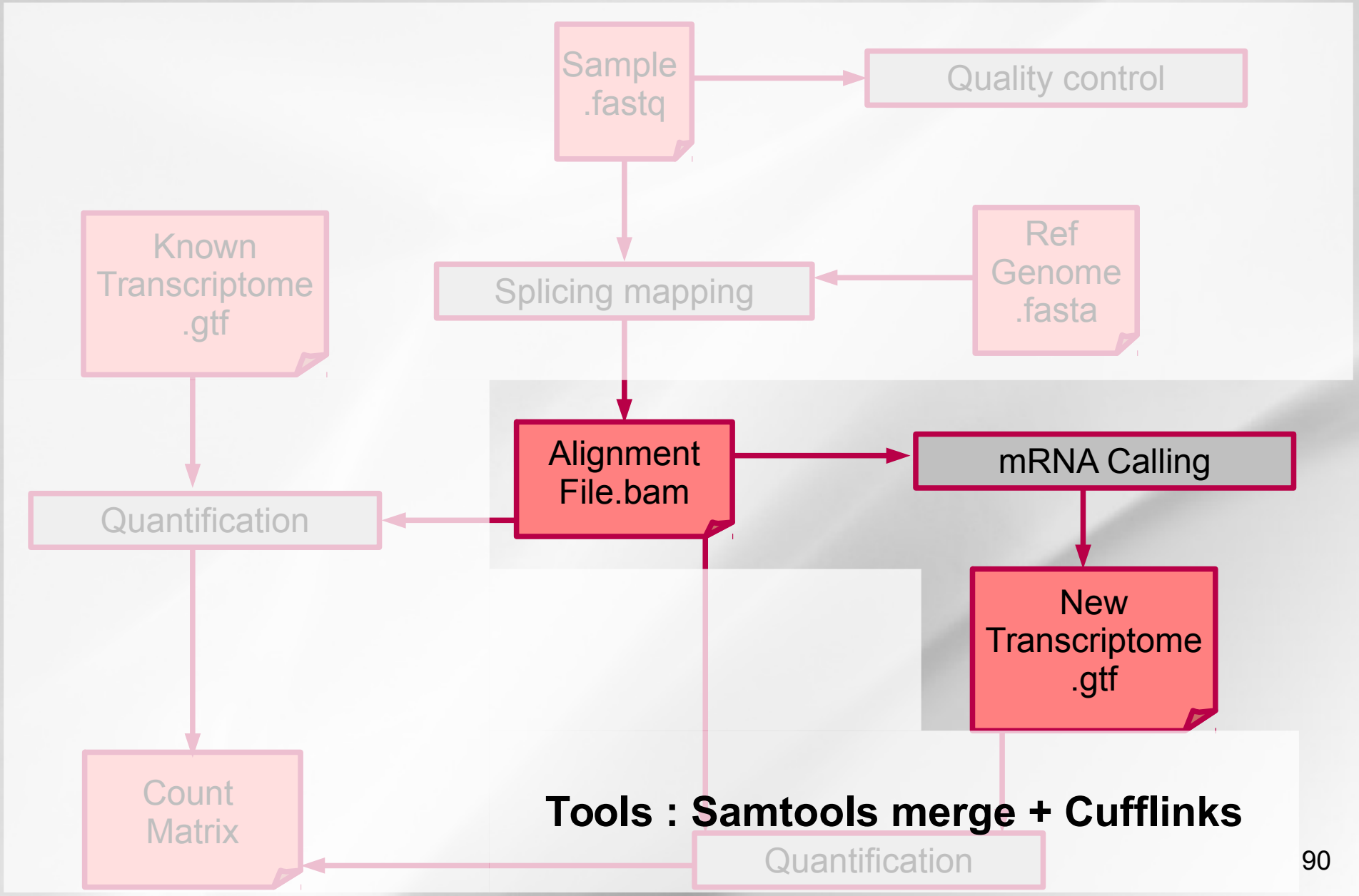
tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	status	FPKM	FPKM_conf_lo	FPKM_conf_hi
CUFF.560.1	-	-	CUFF.560	-	-	22:9743034-9747366	2466	2.84033	OK	23.7788	8.75448	38.803
CUFF.560.2	-	-	CUFF.560	-	-	22:9743034-9762309	4020	8.11967	OK	67.9765	50.3804	85.5727
CUFF.560.3	-	-	CUFF.560	-	-	22:9743034-9762309	3846	1.66444	OK	13.9344	0	29.2533

Hands-on : quantification

Explore GTF file.

Quantify the genes of chromosome 6 using featureCounts for both samples.

Analysis workflow



Tools : Samtools merge + Cufflinks

Transcript reconstruction

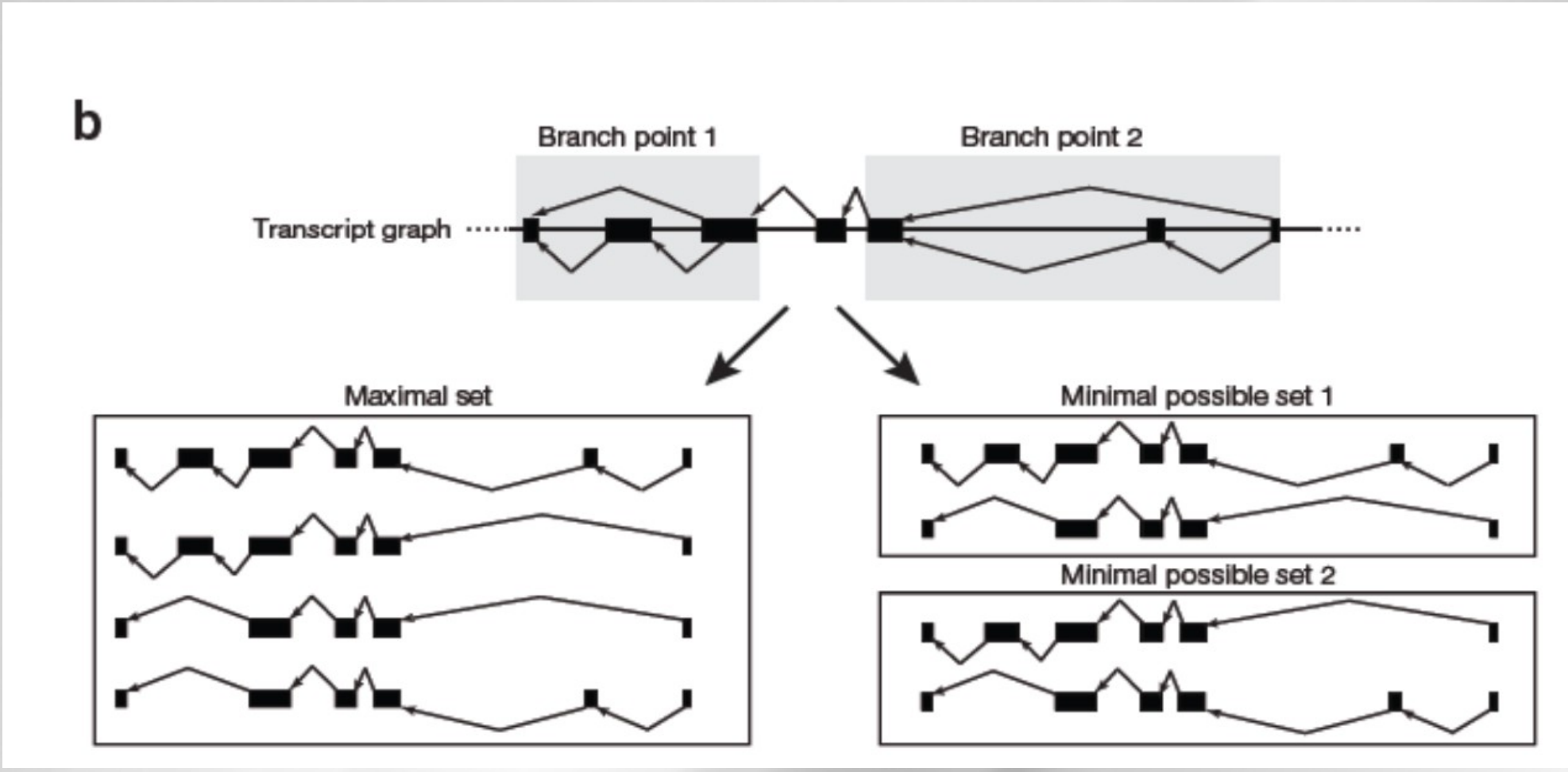
Gene location

Exon location

Junctions :

- Between read pair junction
- Within read junction

Model building strategies



REVIEW

Computational methods for transcriptome annotation and quantification using RNA-seq

Manuel Garber[1], Manfred G Grabherr[1], Mitchell Guttman[1,2] & Cole Trapnell[1,3]

Cufflinks

NATURE BIOTECHNOLOGY | RESEARCH | LETTER

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold & Lior Pachter

Affiliations | Contributions | Corresponding author

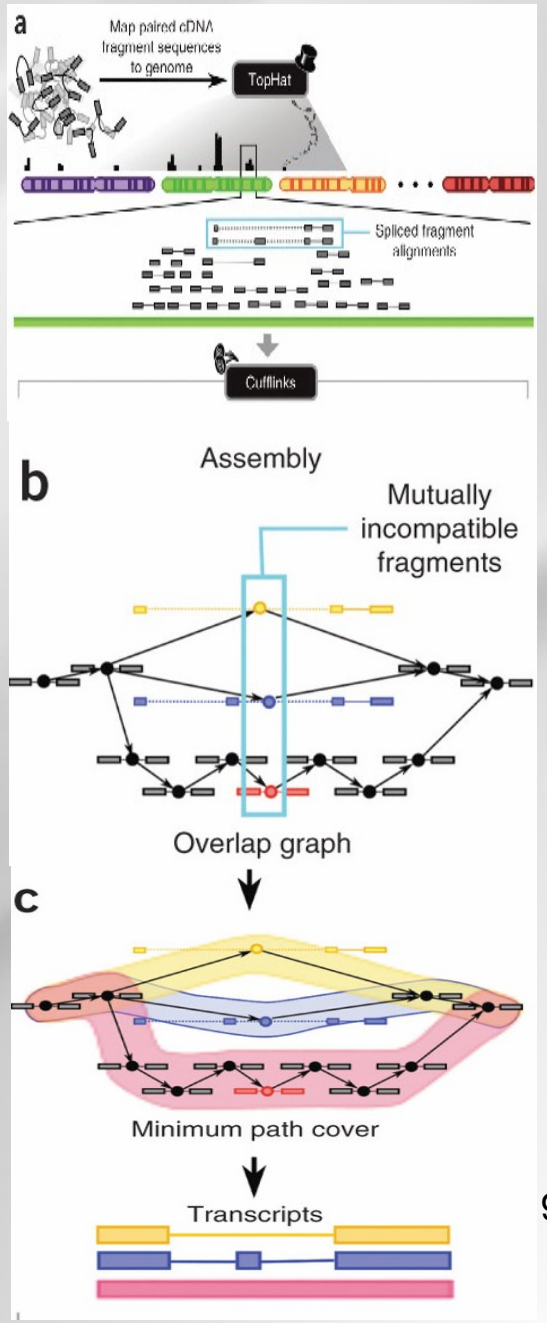
Nature Biotechnology **28**, 511–515 (2010) | doi:10.1038/nbt.1621
Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

http://cufflinks.cbcb.umd.edu/

- ***assembles transcripts***
- estimates their abundances : based on how many reads support each one
- tests for differential expression in RNA-Seq samples

Cufflinks transcript assembly

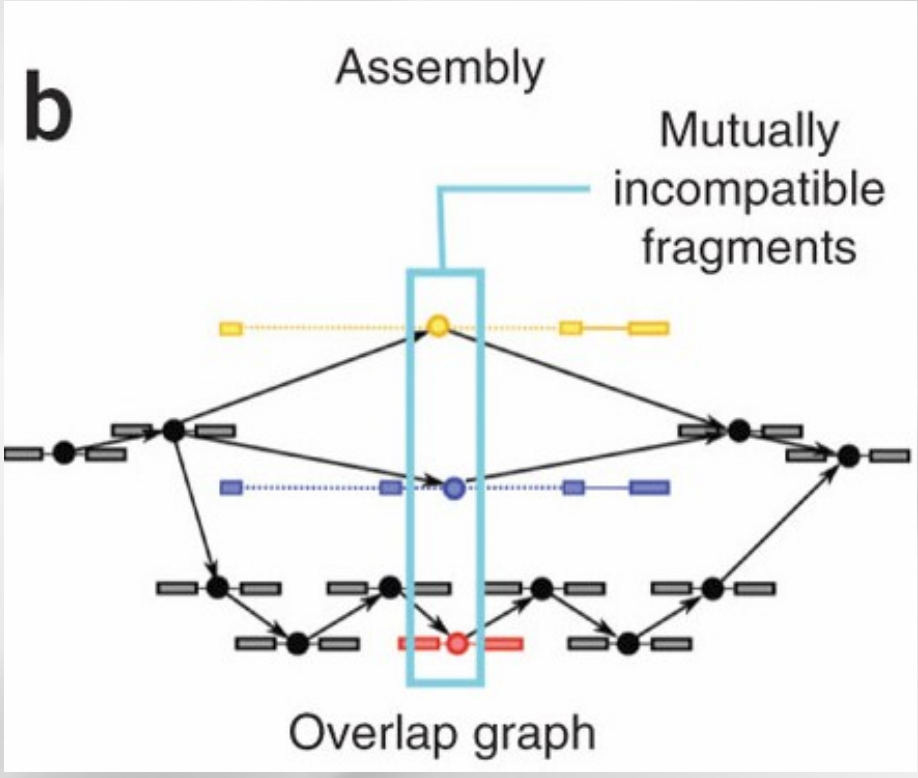
- Transcripts assembly :
  - Fragments are divided into non-overlapping loci
  - each locus is assembled independently :
- Cufflinks assembler
  - find the mini nb of transcripts that explain the reads
  - find a minimum path cover (Dilworth's theorem) :
    - nb incompatible read = mini nb of transcripts needed
    - each path = set of mutually compatible fragments overlapping each other



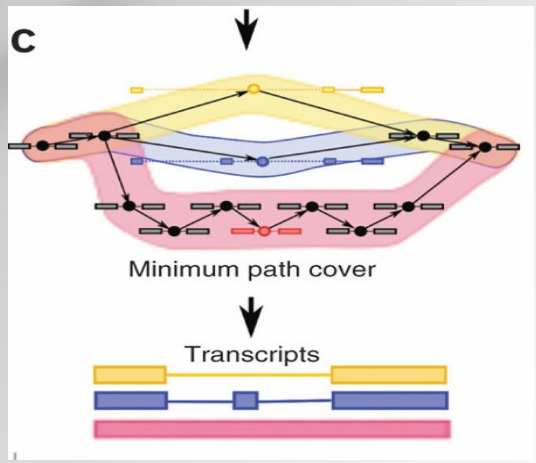
Trapnell C et al. Nature Biotechnology 2010;28:511-515

Cufflinks transcript assembly

- Transcripts assembly :
  - Identification incompatibles fragments: distinct isoforms



  - Compatibles fragments are connected: graphe construction



Trapnell C et al. Nature Biotechnology 2010;28:511-515

Cufflinks inputs and options

- Command line:
  - *cufflinks [options]* <aligned_reads.(sam/bam)>*

- *Some options :*

  -h/--help

  -o/--output-dir

  -p/--num-threads

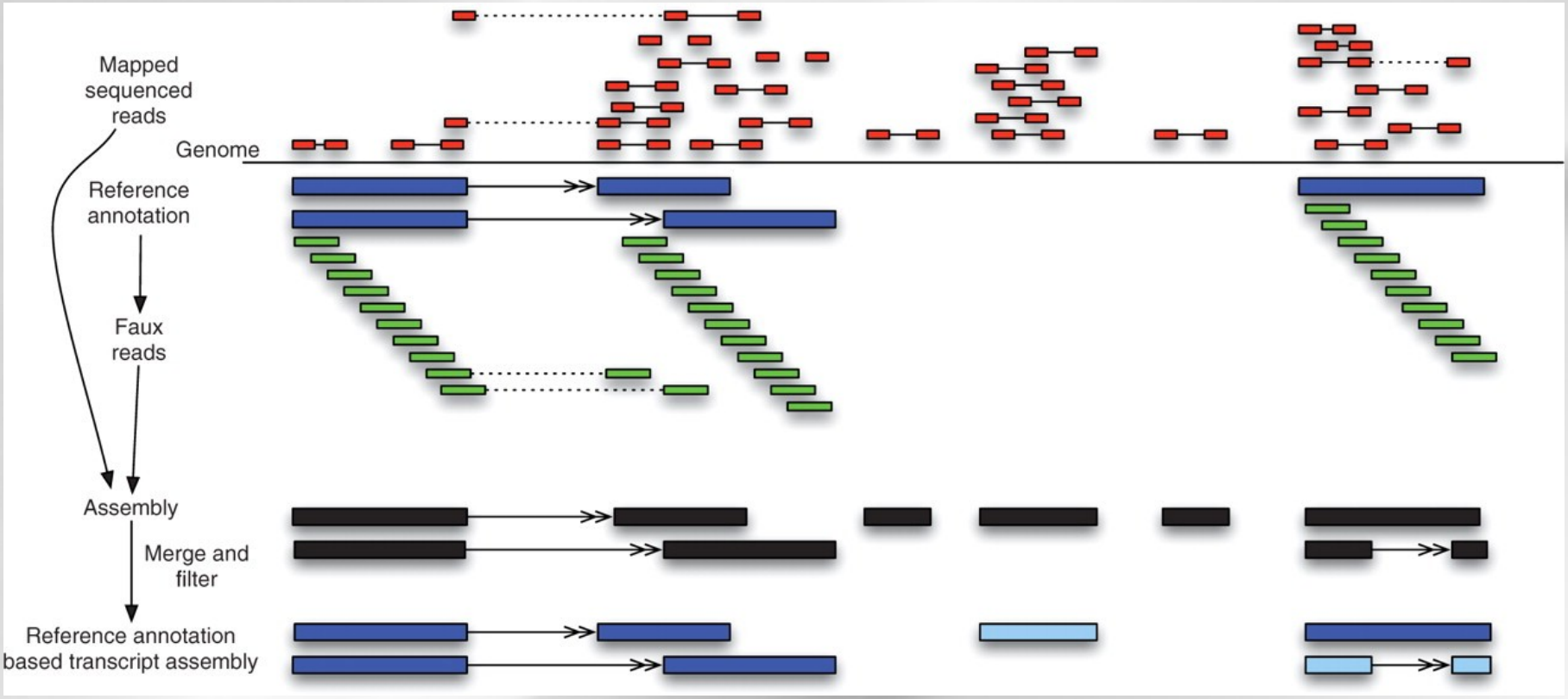
  -G/--GTF <reference_annotation.(gtf/gff)> : estimate isoform expression, no assembly novel transcripts

  -g/--GTF-guide <reference_annotation.(gtf/gff)> **: guide RABT** (**R**eference **A**nnotation **B**ased **T**ranscript) assembly

Cufflinks RABT assembly option

- *Some options :*

  *-**g**/--GTF-guide <reference_annotation.(gtf/gff)>* : guide RABT assembly



Roberts A et al. Bioinformatics 2011;27:2325-2329

Cufflinks outputs

- **transcripts.gtf :** contains assembled isoforms (coordinates and abundances)
- **genes.fpkm_tracking:** contains the genes FPKM
- **isoforms.fpkm_tracking:** contains the isoforms FPKM

Cufflinks GTF description

- **transcripts.gtf** (coordinates and abundances): contains assembled isoforms: can be visualized with a genome viewer

  - GTF format + attributes (ids, FPKM, confidence inteval bounds, depth or read coverage, all introns and exons covered)

22	Cufflinks	transcript	9743035	9747366	349	-	.	gene_id "CUFF.560"; transcript_id "CUFF.560.1"; FPKM "23.7787563790"; frac "0.143485"; conf_lo "8.754478"; conf_hi "38.803035"; cov "2.840328"; full_read_support "yes";
22	Cufflinks	exon	9743035	9745254	349	-	.	gene_id "CUFF.560"; transcript_id "CUFF.560.1"; exon_number "1"; FPKM "23.7787563790"; frac "0.143485"; conf_lo "8.754478"; conf_hi "38.803035"; cov "2.840328";

GTF format

Attributes

22	Cufflinks	transcript	9743035	9747366	349	-	.
22	Cufflinks	exon	9743035	9745254	349	-	.

Chr Source Feature Start End strand Frame

Score:
Most abundant isoform = 1000
Minor : ratio=minor Fpkm/major FPKM

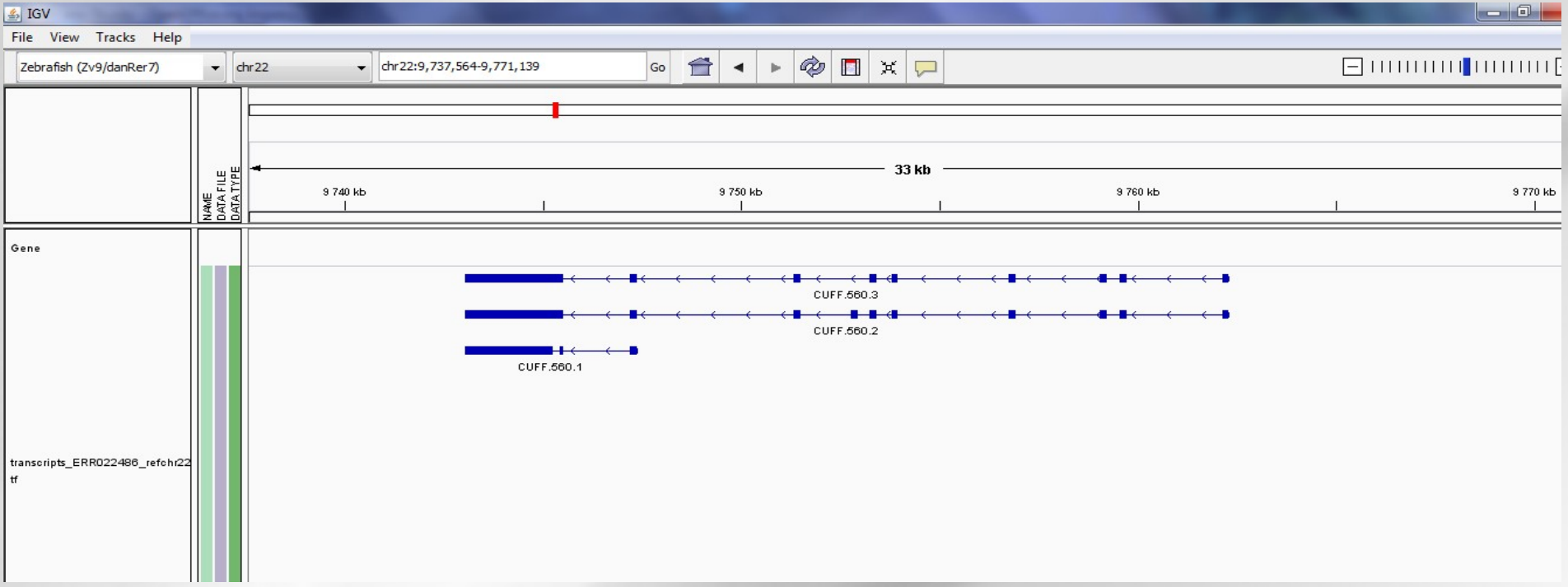
Whether or not all introns and exons were fully covered by Reads (with -g)

gene_id "CUFF.560"; transcript_id "CUFF.560.1"; FPKM "23.7787563790"; frac "0.143485"; conf_lo "8.754478"; conf_hi "38.803035"; cov "2.840328"; full_read_support "yes";
gene_id "CUFF.560"; transcript_id "CUFF.560.1"; exon_number "1"; FPKM "23.7787563790"; frac "0.143485"; conf_lo "8.754478"; conf_hi "38.803035"; cov "2.840328";

Cufflinks GTF description

- **transcripts.gtf** (coordinates and abundances): contains assembled isoforms: can be visualized with a genome viewer

  - Exemple VISUALISATION IGV



Cuffcompare

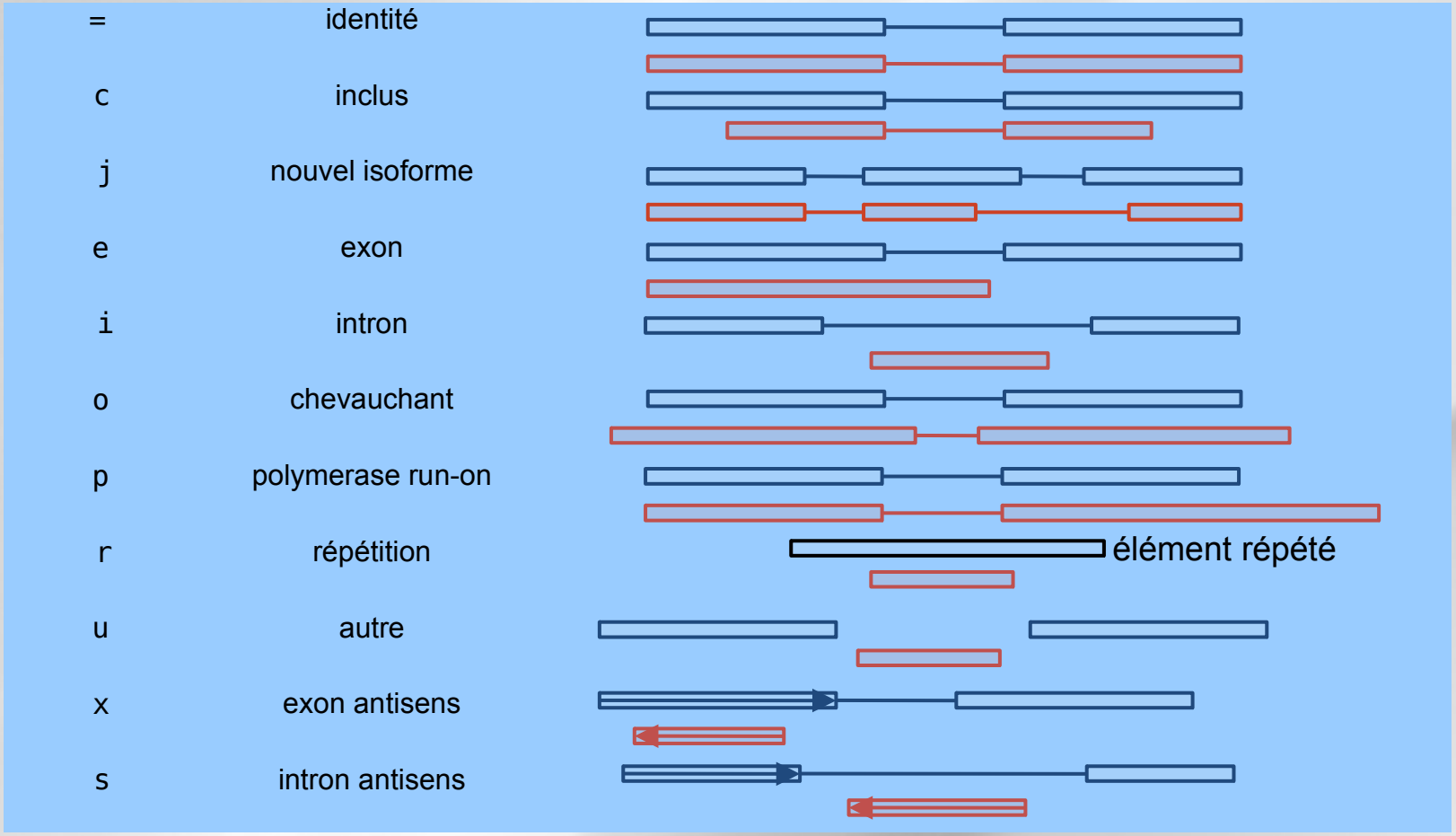
- Comparison of transcriptoms files

- Command:
  cuffcompare -r <reference_mrna.gtf> -o <outprefix> <input1.gtf> …

- Outputs:
  - Overall summary statistics: <outprefix>.stats
    The Sn and Sp columns show specificity and sensitivity values at each level, while the fSn and fSp columns are “fuzzy” variants of these same accuracy calculations, allowing for a very small variation in exon boundaries to still be counted as a “match”.
  - The “union” of all transfrags in all assemblies: <outprefix>.combined.gtf
  - Transfrags matching to each reference transcript: <cuff_in>.refmap
  - Best reference transcript for each transfrag: <cuff_in>.tmap
  - Tracking transfrags through multiple samples: <outprefix>.tracking

Cuffcompare

Class code de cuffcompare



http://cufflinks.cbcb.umd.edu/manual.html#class_codes

Gene discovery pipeline

Alignment (Tophat)

Bam merge (samtools)

Discovery of novels features (cufflinks)

Quantification file with (featureCounts)

Hands-on : cufflinks

Commands :

Merge all bam :

Samtools merge merge_*all.bam file1.bam file2.bam*

Cufflinks command:

cufflinks -p 4 --output-dir=cufflinks

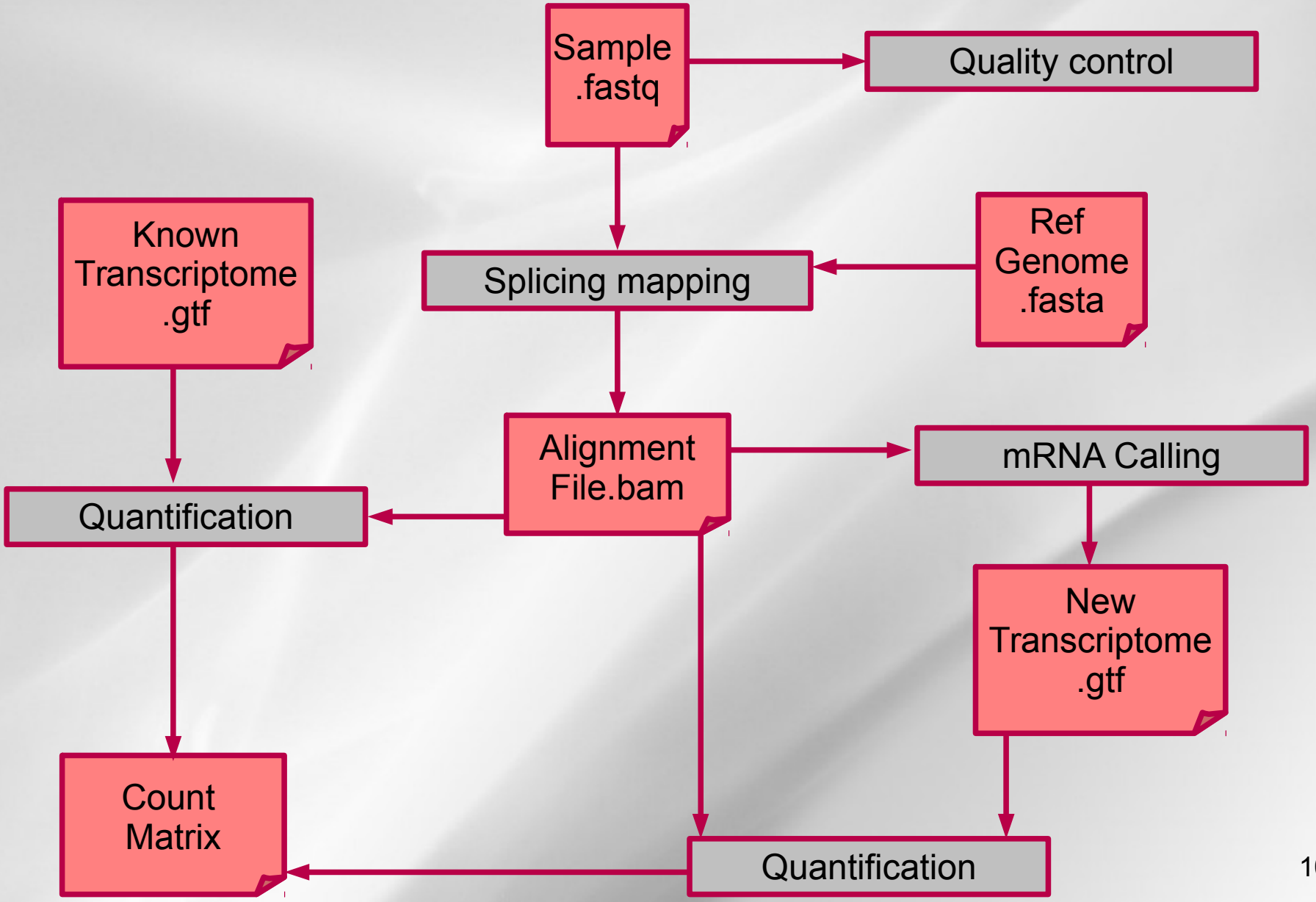
***-g** reference_transcript.gtf*

merge_all.bam

Cuffcompare command :

cuffcompare -f reference_transcript.gtf -o compare cufflink_transcripts.gtf

Analysis workflow



Differential expression

- Biostatistics Genotoul Platform
- Training :
  - http://perso.math.univ-toulouse.fr/biostat/category/formation/
  - Tutotial of RNAseq analysis
    www.nathalievilla.org/teaching/rnaseq.html
- R scripts available on Genotoul cluster
  - See http://bioinfo.genotoul.fr/index.php?id=119

Differential expression

But : trouver les *gènes significativement* différentiellement exprimés entre 2 conditions.

Méthode:

- Normalisation
- Estimation de l'expression
- Test

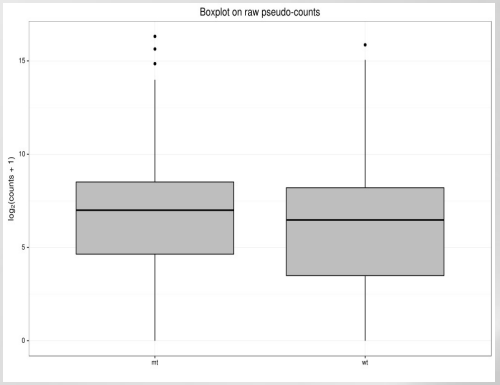
Outils :

- DESeq, **EdgeR**, DESeq2, etc. (en R)
- **CuffDiff** (suite Tuxedo)

Differential expression Normalization

Problème: Le nombre de lectures est différent d'un réplicat à l'autre.

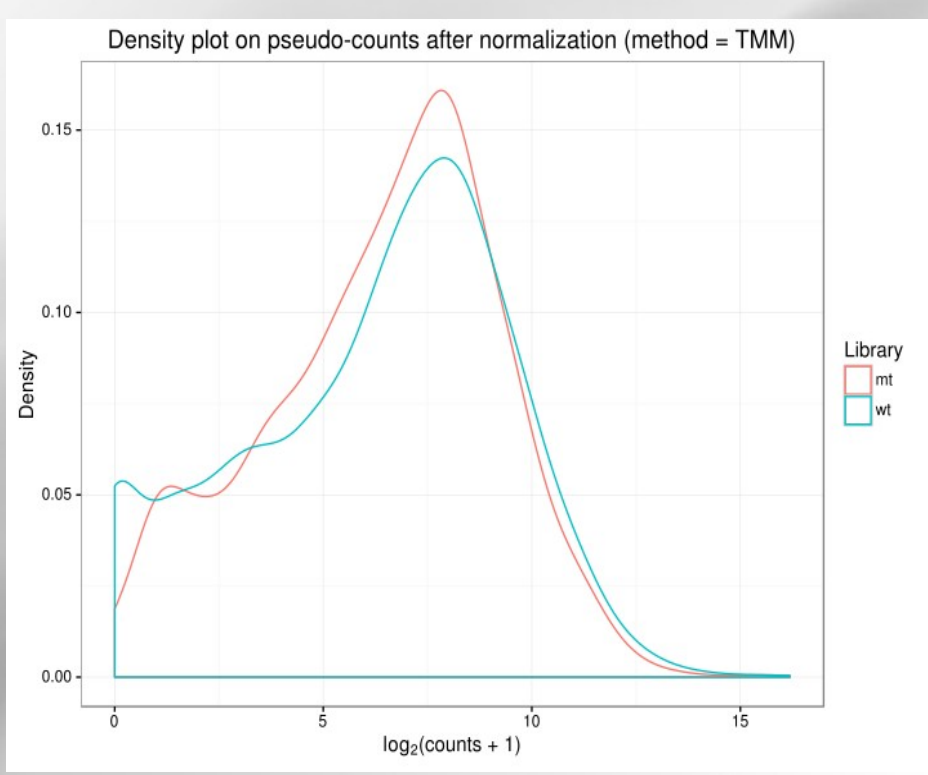
Idée: Appliquer à chaque échantillon un coefficient multiplicatif.



Command :

Rscript /usr/local/bioinfo/Scripts/bin/Normalization.R
-f tomato_count.R.txt
-o norm

Script that perform edgeR normalizations (RLE, TMM upperquartile) and provide graphs



Differential expression Test

But : Comparer les distributions d'expression dans 2 conditions.

Résultats: p-value et q-value (p-value avec correction de tests multiples)

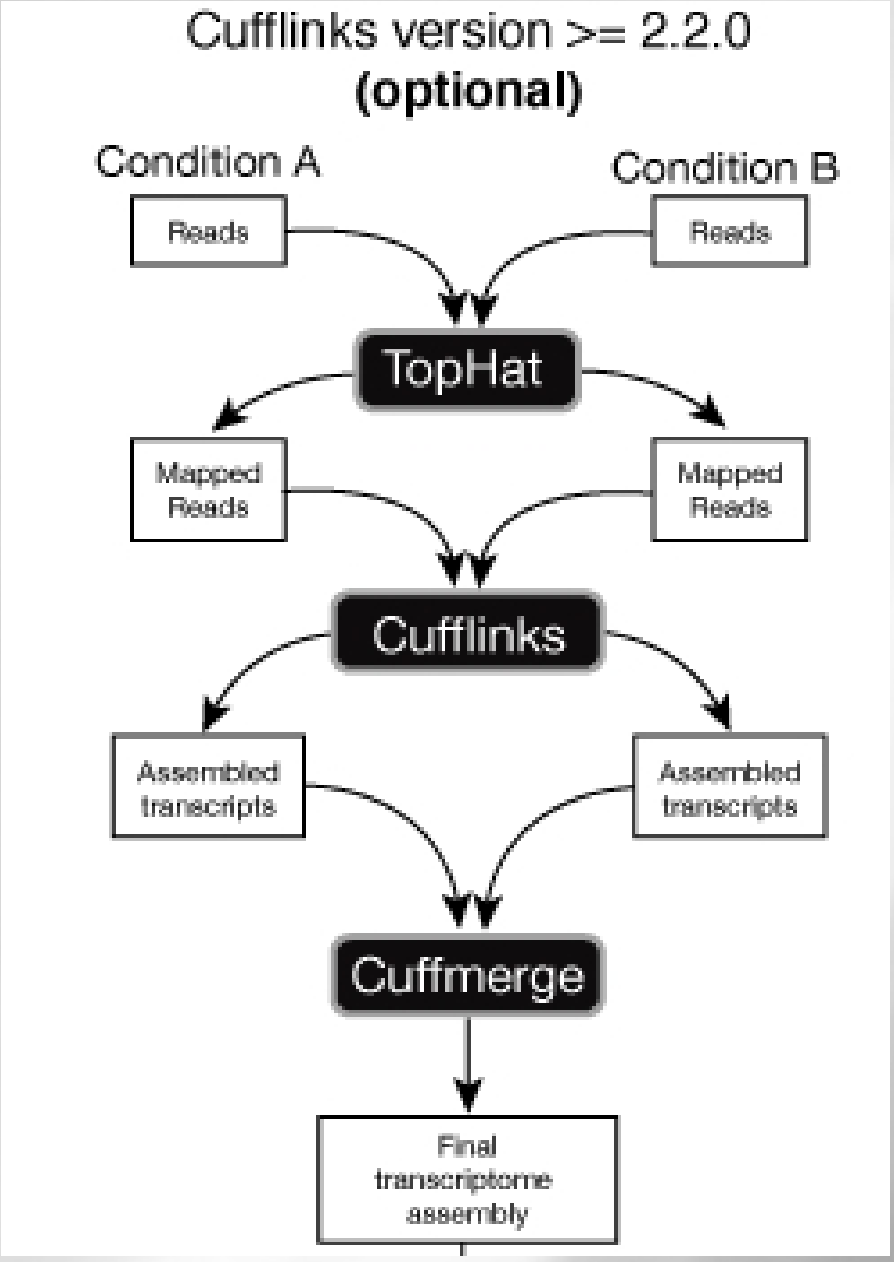
Command :

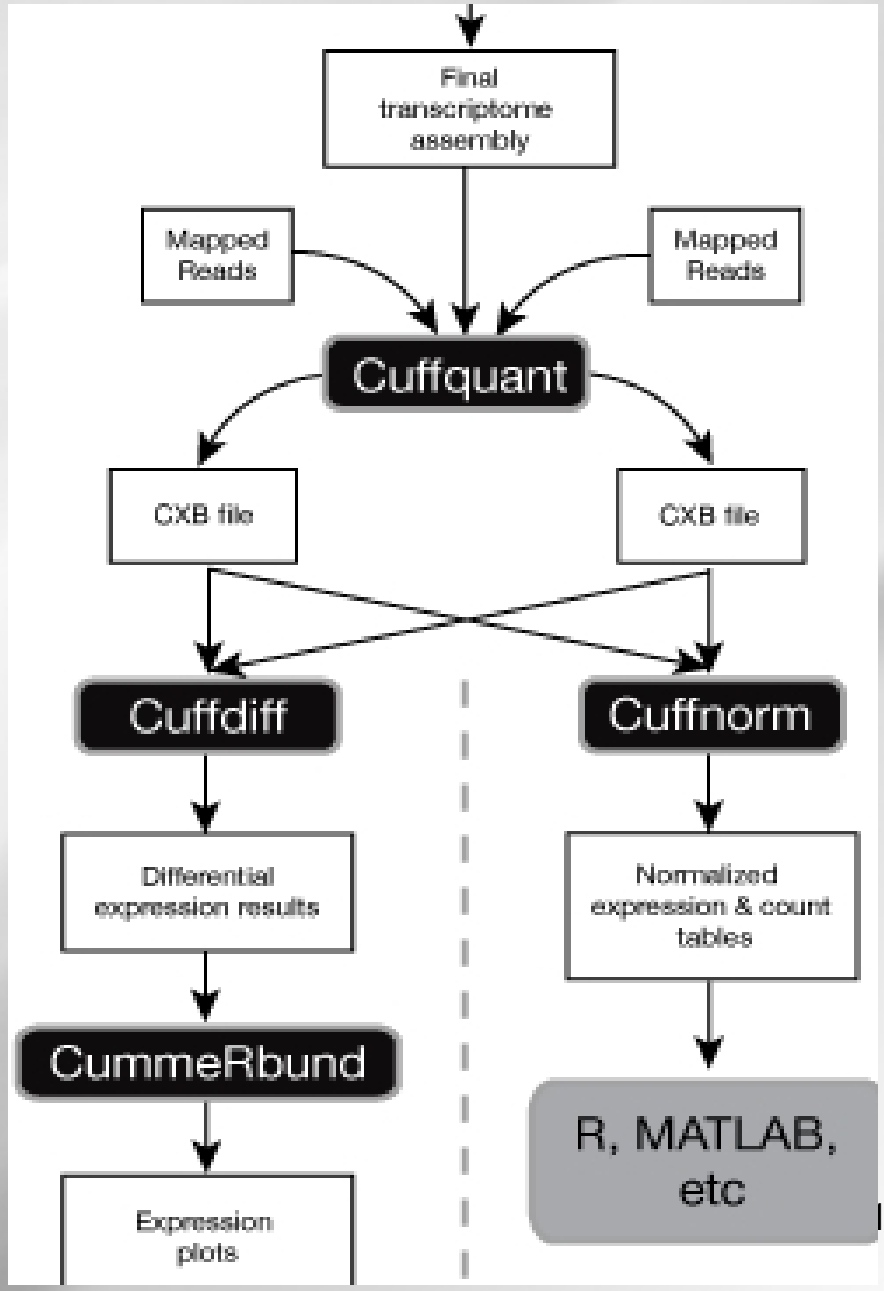
```
Rscript /usr/local/bioinfo/Scripts/bin/DEG.R
   -f tomato_count.R.txt
   --norm norm/RLE_info.txt
   --pool1 mt
   --pool2 wt
   -o DEG
```

Differential expression Test

```
Options:
        -f CHARACTER, --file=CHARACTER
                tabulated raw count matrix with library name as header
                      (e.g. #gene_id     lib1    lib2...)

        -n CHARACTER, --norm=CHARACTER
                file with normalized factors with library name
              (e.g. sample.name  lib.size        norm.factors). This file is obtained
              with script 'Normalization.R'

        -o CHARACTER, --out=CHARACTER
                folder path where results are stored

        --pool1=CHARACTER
                library name in pool 1 separeted by ','

        --pool2=CHARACTER
                library name in pool 2 separeted by ','

        --filter=CHARACTER
                if TRUE low expressed genes are removed [default=TRUE]

        --alpha=CHARACTER
                significance level of the tests (i.e. acceptable rate of
              false-positive in the list of differentially expressed genes)
              [default=0.05]

        --correct=CHARACTER
                method used to adjust p-values for multiple testing
              ('BH', 'BY' or 'fdr') [default=BH]

        --MAplots=CHARACTER
                if TRUE all MA plots are saved [default=FALSE]
```

Cufflinks >= 2.2.0





Quality for Bioinfo Plateform!

Satisfaction form :

http://bioinfo.genotoul.fr/index.php?id=79

Useful links

Seqanswer: http://seqanswers.com/

RNAseq blog: http://rna-seqblog.com/

Illumina: http://www.illumina.com/