EUR NEG VEC

European Network for Neglected Vectors and Vector-Borne Infections

cost
EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

bio-informatic analysis of RNASeq data

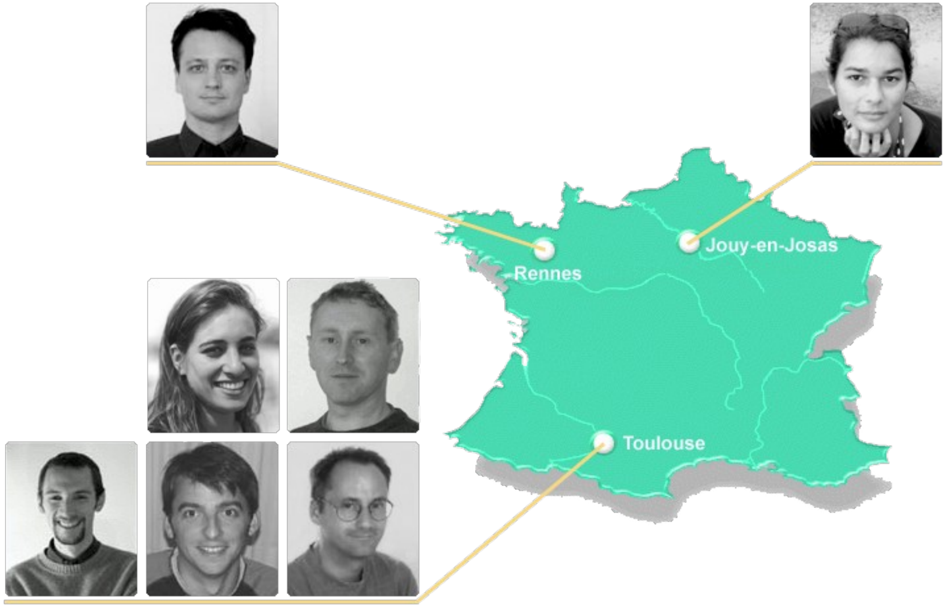
Christophe Klopp / www.sigenae.org, bioinfo.genotoul.fr

Overview

- Transcriptome and transcription variability
- Sequencing techniques
- Usual questions
- Data quality control
- Read spliced alignment
- Expression quantification
- Novel gene and transcript identification

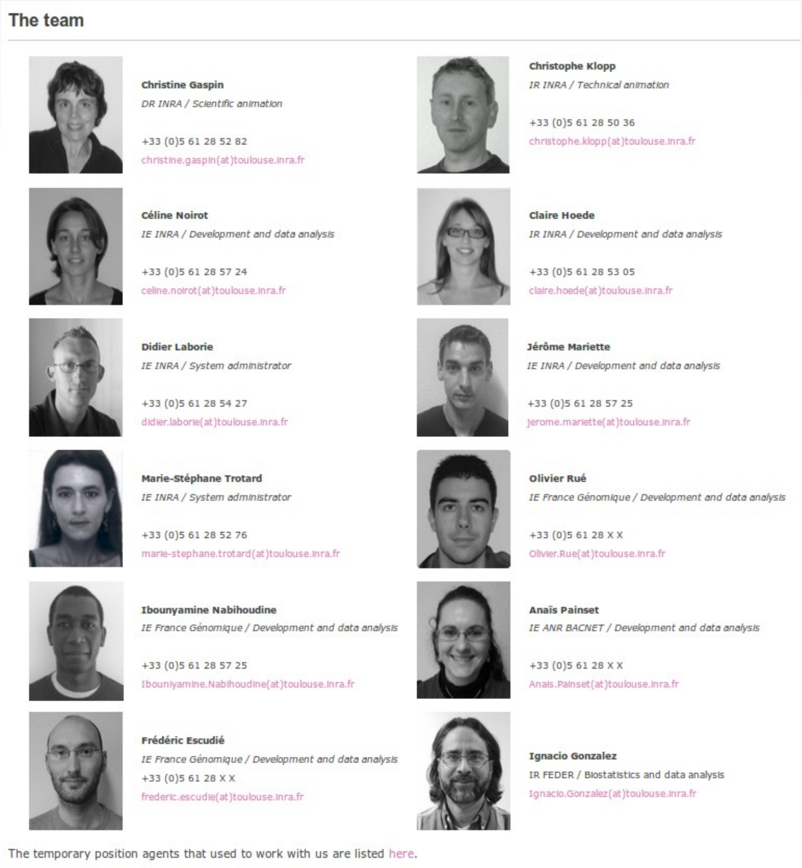
Sigenae

- 7 engineers work in farm animal genomics
- 30 running projects
- > 400 publications (citing the team or having a team member in the authors)



Bioinfo Genotoul

- 12 engineers
- > 4,000 cpus, 1Pb disk space
- 10 training sessions
- > 20 running projects



The team

Christine Gaspin
DR INRA / Scientific animation
+33 (0)5 61 28 52 82
christine.gaspin(at)toulouse.inra.fr

Christophe Klopp
IR INRA / Technical animation
+33 (0)5 61 28 50 36
christophe.klopp(at)toulouse.inra.fr

Céline Noirot
IE INRA / Development and data analysis
+33 (0)5 61 28 57 24
celine.noirot(at)toulouse.inra.fr

Claire Hoede
IR INRA / Development and data analysis
+33 (0)5 61 28 53 05
claire.hoede(at)toulouse.inra.fr

Didier Laborie
IE INRA / System administrator
+33 (0)5 61 28 54 27
didier.laborie(at)toulouse.inra.fr

Jérôme Mariette
IE INRA / Development and data analysis
+33 (0)5 61 28 57 25
jerome.mariette(at)toulouse.inra.fr

Marie-Stéphane Trotard
IE INRA / System administrator
+33 (0)5 61 28 52 76
marie-stephane.trotard(at)toulouse.inra.fr

Olivier Rué
IE France Génomique / Development and data analysis
+33 (0)5 61 28 X X
Olivier.Rue(at)toulouse.inra.fr

Ibounyamine Nabihoudine
IE France Génomique / Development and data analysis
+33 (0)5 61 28 57 25
Ibouniyamine.Nabihoudine(at)toulouse.inra.fr

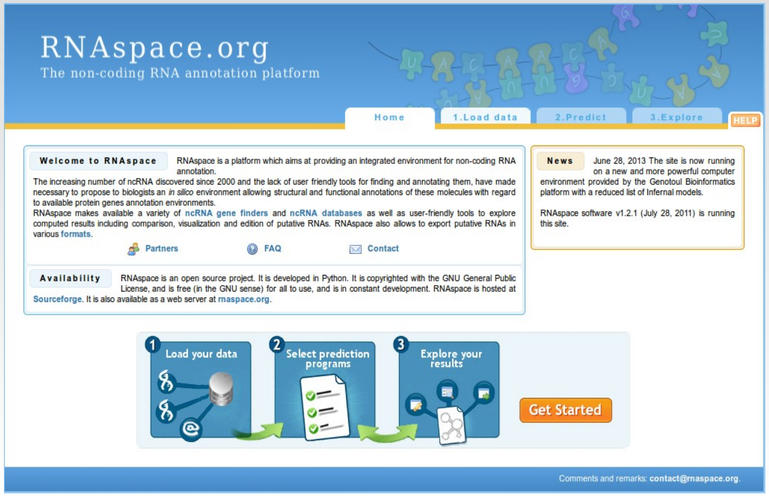
Anaïs Painset
IE ANR BACNET / Development and data analysis
+33 (0)5 61 28 X X
Anais.Painset(at)toulouse.inra.fr

Frédéric Escudié
IE France Génomique / Development and data analysis
+33 (0)5 61 28 X X
frederic.escudie(at)toulouse.inra.fr

Ignacio Gonzalez
IR FEDER / Biostatistics and data analysis
Ignacio.Gonzalez(at)toulouse.inra.fr

The temporary position agents that used to work with us are listed here.

Common software developments



RNAspace.org
The non-coding RNA annotation platform

Welcome to RNAspace RNAspace is a platform which aims at providing an integrated environment for non-coding RNA annotation.
The increasing number of ncRNA discovered since 2000 and the lack of user friendly tools for finding and annotating them, have made necessary to propose to biologists an *in silico* environment allowing structural and functional annotations of these molecules with regard to available protein genes annotation environments.
RNAspace makes available a variety of ncRNA gene finders and ncRNA databases as well as user-friendly tools to explore computed results including comparison, visualization and edition of putative RNAs. RNAspace also allows to export putative RNAs in various formats.

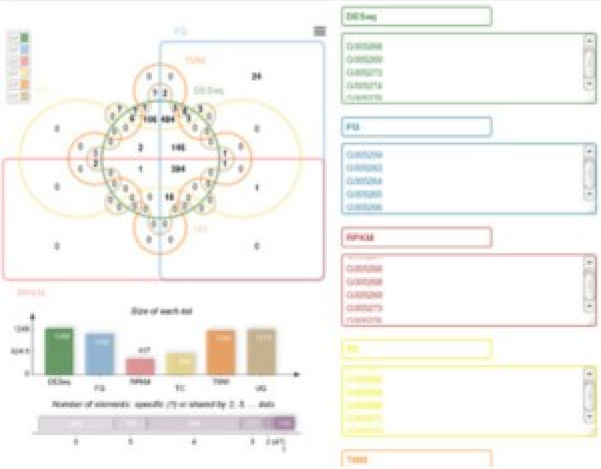
Availability RNAspace is an open source project. It is developed in Python. It is copyrighted with the GNU General Public License, and is free (in the GNU sense) for all to use, and is in constant development. RNAspace is hosted at Sourceforge. It is also available as a web server at rnaspace.org.

News June 28, 2013 The site is now running on a new and more powerful computer environment provided by the Genotoul Bioinformatics platform with a reduced list of Infernal models.

RNAspace software v1.2.1 (July 28, 2011) is running this site.

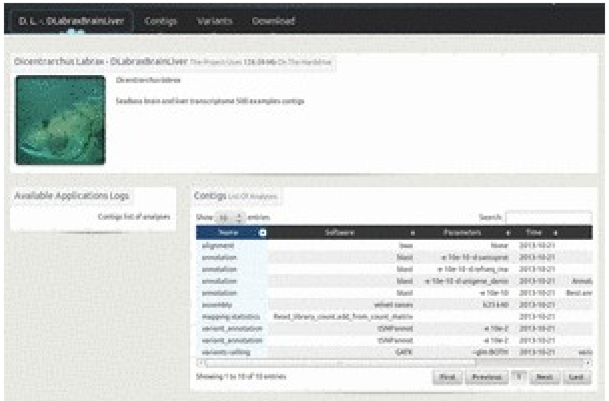
Comments and remarks: contact@rnaspace.org.

http://rnaspace.org/



DESeq
FG
RPKM
TC
TMM
UQ
Size of each list
DESeq
FG
RPKM
TC
TMM
UQ
Number of elements: specific (1) or shared by 2, 3, ... lists

http://bioinfo.genotoul.fr/jvenn/example.html



Dicentrarchus Labrax - DLabraxBrainLiver [illegible]

Dicentrarchus labrax

Seabass brain and liver transcriptome 500 examples contigs

Available Applications Logs

Contigs list of analyses

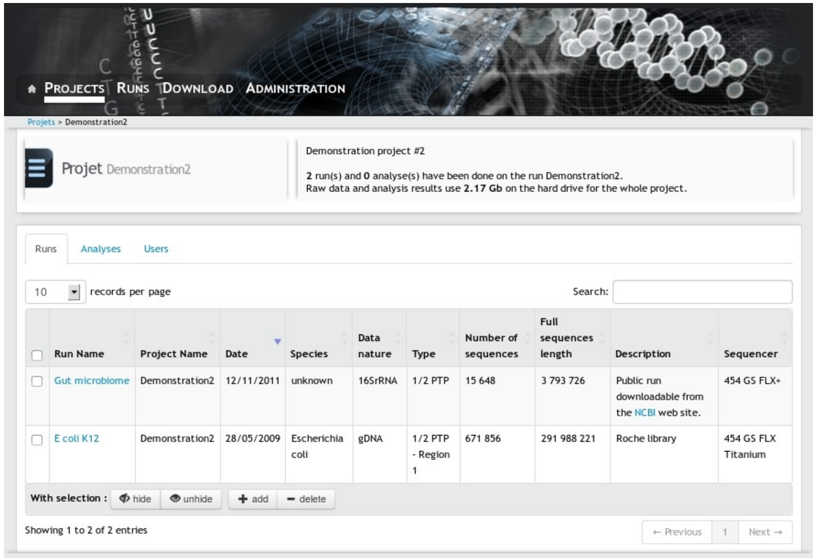
Contigs List Of Analyses

Show 10 entries Search:

Name	Software	Parameters	Time	
alignment	bwa	None	2013-10-21	
annotation	blast	-e 10e-10 -d swissprot	2013-10-21	
annotation	blast	-e 10e-10 -d refseq_rna	2013-10-21	
annotation	blast	-e 10e-10 -d unigene_danio	2013-10-21	Annot
annotation	blast	-e 10e-10	2013-10-21	Best an
assembly	velvet oases	k25 k40	2013-10-21	
mapping statistics	Read_library_count.add_from_count_matrix		2013-10-21	
variant_annotation	[illegible]	-e 10e-2	2013-10-21	
variant_annotation	[illegible]	-e 10e-2	2013-10-21	
variants calling	GATK	-glm BOTH	2013-10-21	vario

Showing 1 to 10 of 10 entries

http://ngspipelines.toulouse.inra.fr:9019/



Projets > Demonstration2

Projet Demonstration2

Demonstration project #2

2 run(s) and **0** analyse(s) have been done on the run Demonstration2.
Raw data and analysis results use **2.17 Gb** on the hard drive for the whole project.

10 records per page Search:

	Run Name	Project Name	Date	Species	Data nature	Type	Number of sequences	Full sequences length	Description	Sequencer
□	Gut microbiome	Demonstration2	12/11/2011	unknown	16SrRNA	1/2 PTP	15 648	3 793 726	Public run downloadable from the NCBI web site.	454 GS FLX+
□	E coli K12	Demonstration2	28/05/2009	Escherichia coli	gDNA	1/2 PTP - Region 1	671 856	291 988 221	Roche library	454 GS FLX Titanium

With selection : hide unhide add delete

Showing 1 to 2 of 2 entries

http://ng6.toulouse.inra.fr/

Definitions

- RNA-Seq :

  RNA-seq (RNA Sequencing), also called Whole Transcriptome Shotgun Sequencing (WTSS), is a technology that uses the capabilities of next-generation sequencing to reveal a snapshot of RNA presence and quantity from a genome at a given moment in time.

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

RNA-Seq aims

- Find the structures and functions of expressed genes and transcripts (possible splice forms),
- Measure the expression levels usually to find differentially expressed transcripts (explaining the phenotype),
- Find polymorphisms in the transcripts :
    - SSR (short sequence repeat),
    - SNP (Single nucleotide polymorphism),
    - INDEL (Insertion / Deletion).

RNA-Seq limitations

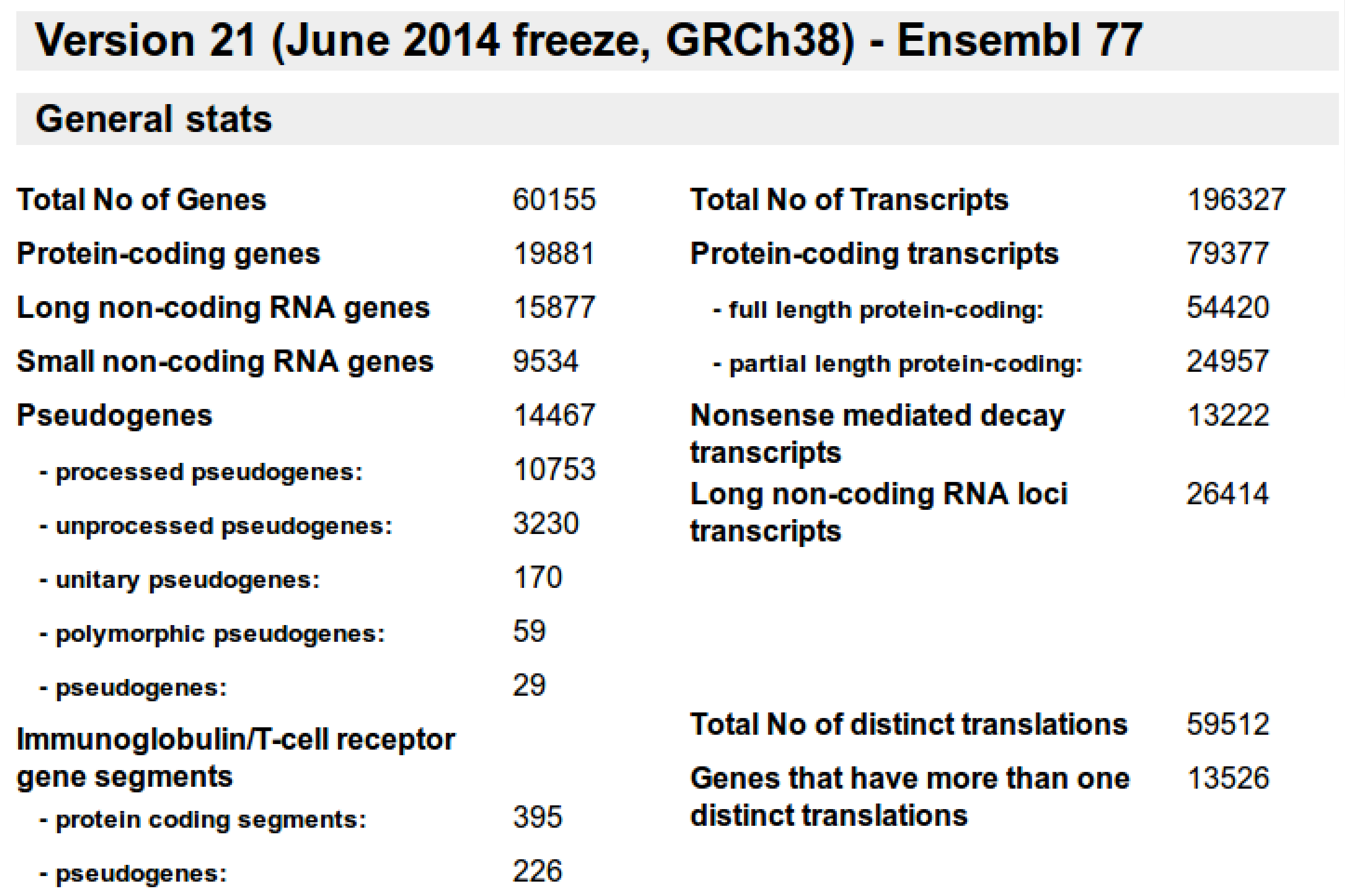
- No sequencer is able, today, to produce large quantities of reliable sequences corresponding to full length transcripts :
    - HiSeq produces short reads
    - MiSeq, PGM, proton produce lower read numbers (quantities)
    - PacBio reads have an high error rate and low through-put

Hands-on

- In small groups, define transcription and the different products produced by transcription.
- Group the products depending on their features.

- List the different forms of variability found in transcription products and discuss their impact.

GENCODE view



Version 21 (June 2014 freeze, GRCh38) - Ensembl 77

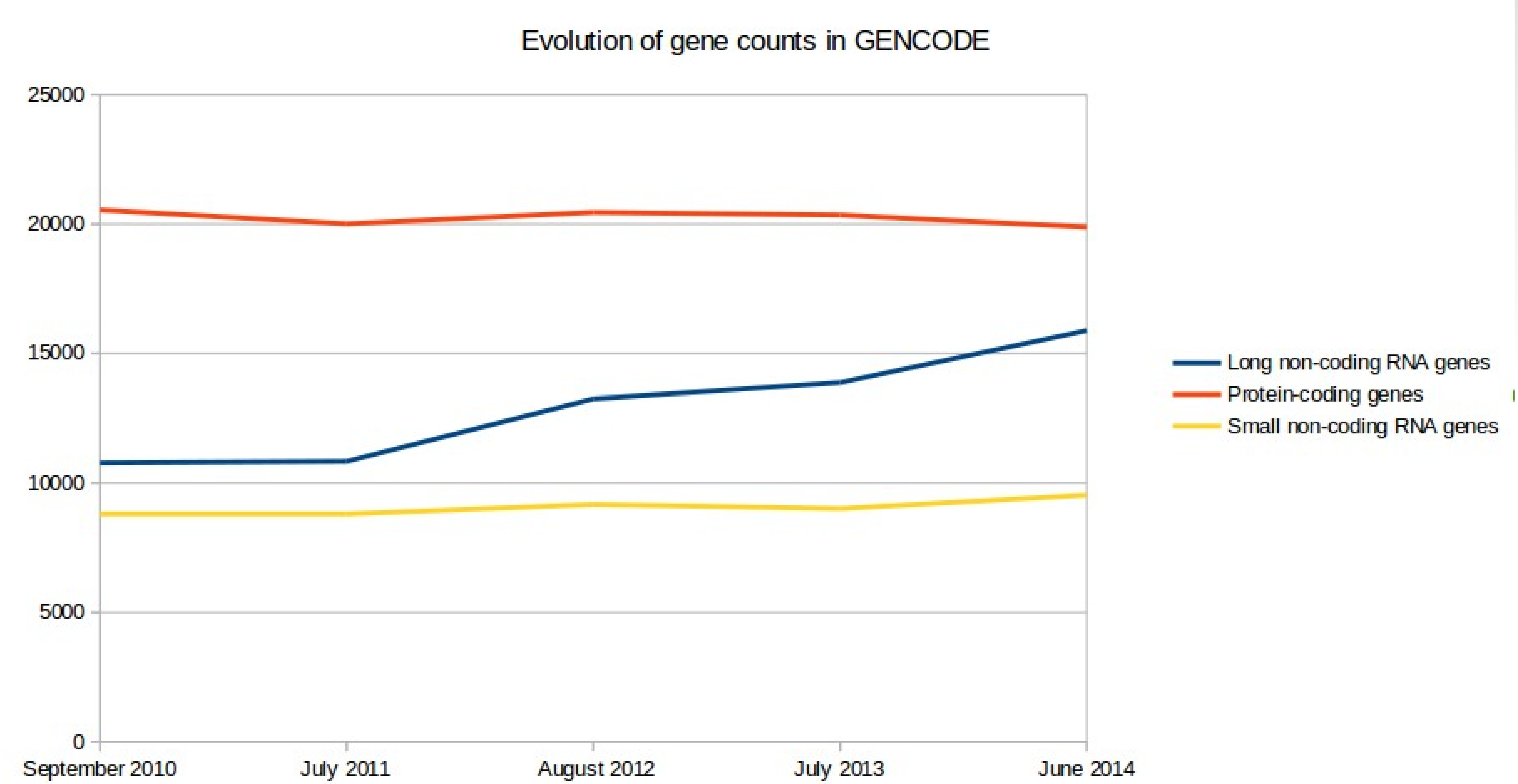
General stats

Total No of Genes	60155
Protein-coding genes	19881
Long non-coding RNA genes	15877
Small non-coding RNA genes	9534
Pseudogenes	14467
- processed pseudogenes:	10753
- unprocessed pseudogenes:	3230
- unitary pseudogenes:	170
- polymorphic pseudogenes:	59
- pseudogenes:	29
Immunoglobulin/T-cell receptor gene segments	
- protein coding segments:	395
- pseudogenes:	226

Total No of Transcripts	196327
Protein-coding transcripts	79377
- full length protein-coding:	54420
- partial length protein-coding:	24957
Nonsense mediated decay transcripts	13222
Long non-coding RNA loci transcripts	26414

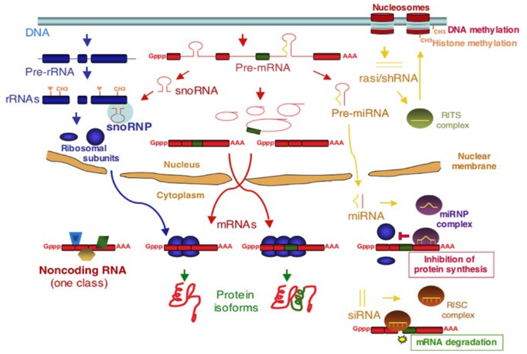
Total No of distinct translations	59512
Genes that have more than one distinct translations	13526

Lnc-RNA counts in GENCODE



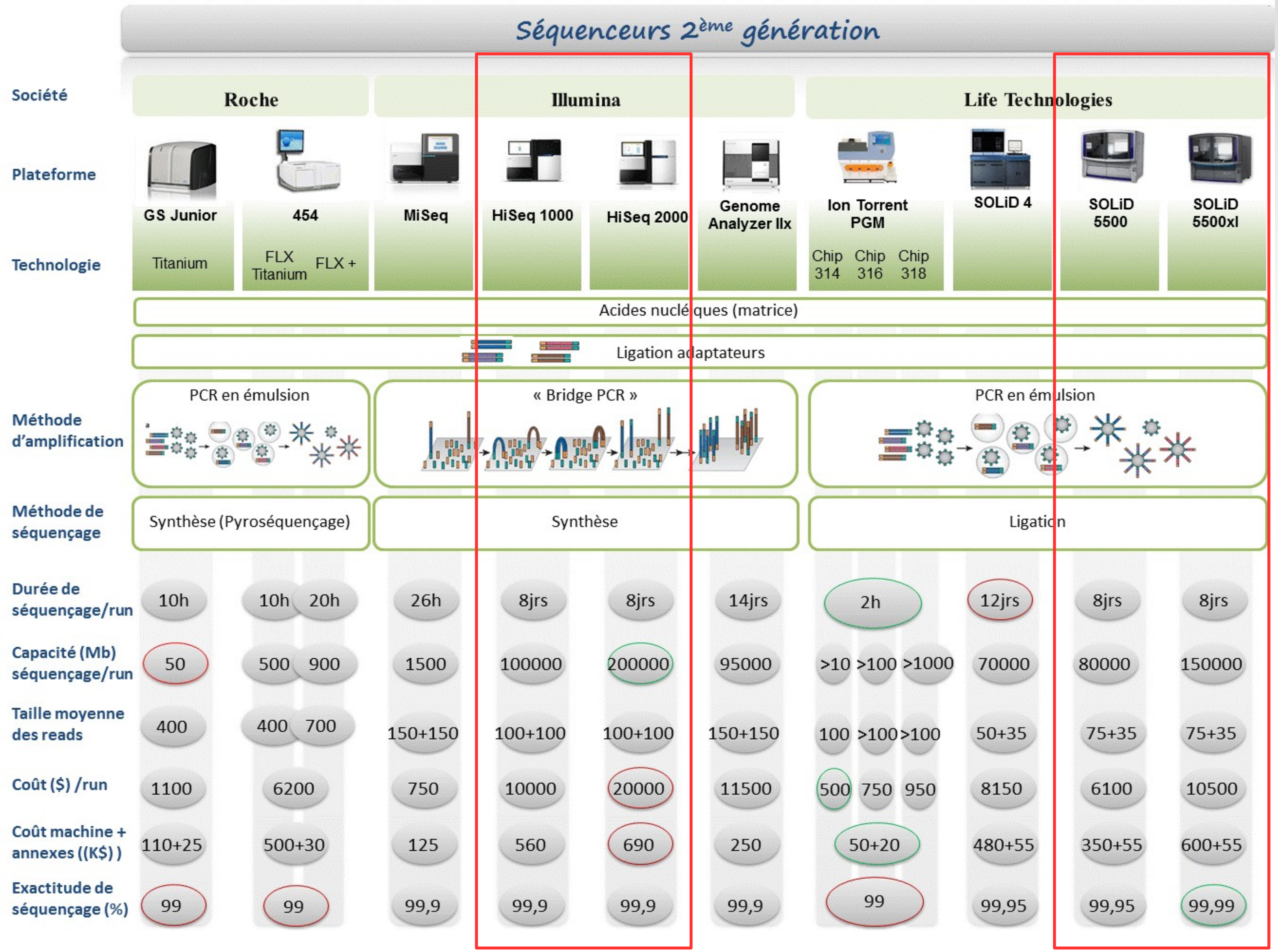
Transcription variability

- Number of transcripts
  - possible variation factor between transcripts: 10^6 or more,
  - expression variation between samples (biological repeats, technical repeats).
- Many types of transcripts
  - mRNA, ncRNA,...
- Isoforms (with non canonical splice sites)
- Intron retention
  - The splicing is not always completed
  - Is a new isoform or a transcription error
- Transcript decay (degradation)
- Allele specific expression
- Gene fusion (found in cancer cells)



http://www.nature.com/emboj/journal/v25/n5/fig_tab/7601023a_F2.html

Sequencers



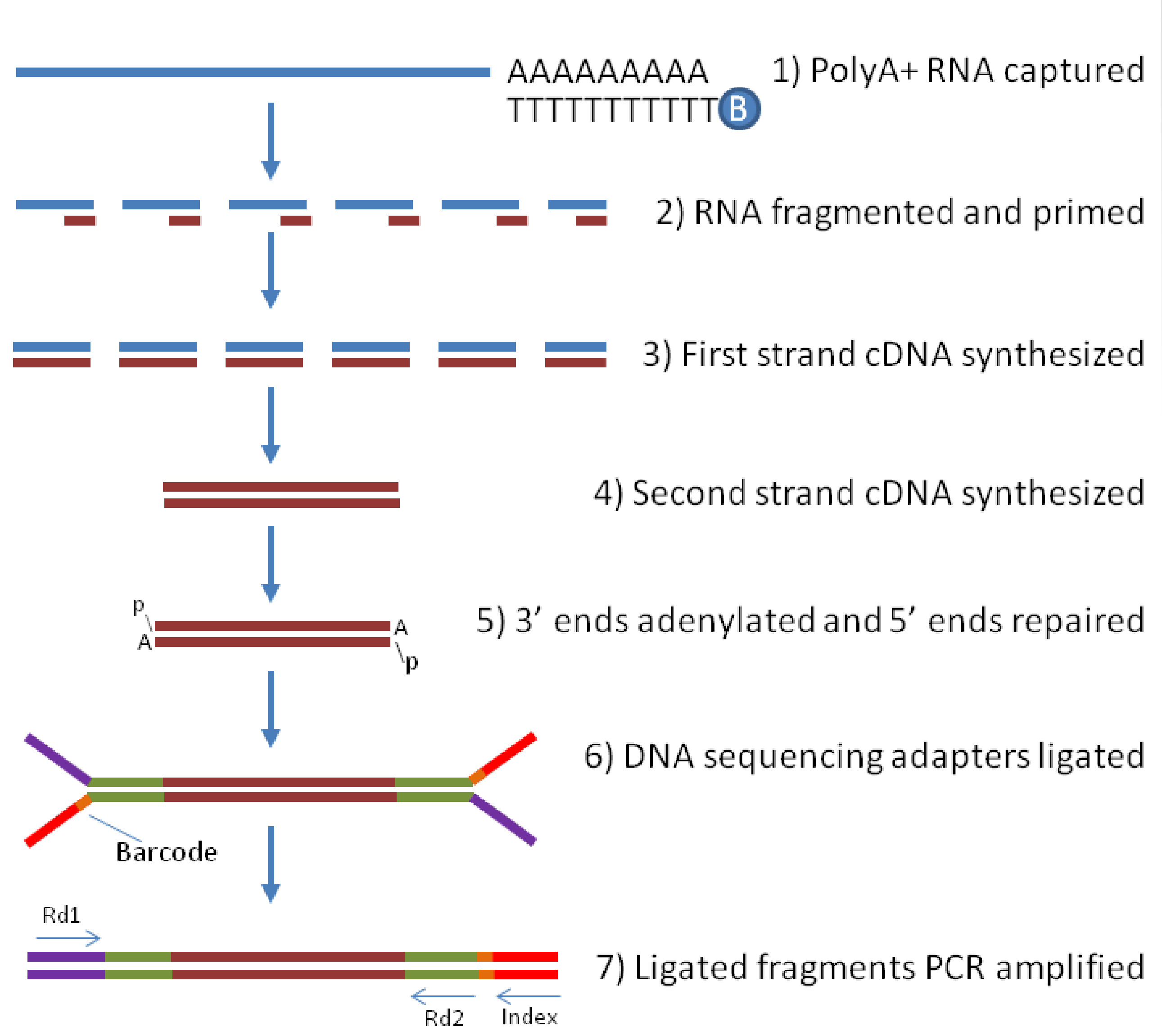
Séquenceurs 2ème génération

Société	Roche			Illumina				Life Technologies					
Plateforme	GS Junior	454		MiSeq	HiSeq 1000	HiSeq 2000	Genome Analyzer IIx	Ion Torrent PGM			SOLiD 4	SOLiD 5500	SOLiD 5500xl
Technologie	Titanium	FLX Titanium	FLX +					Chip 314	Chip 316	Chip 318			
	Acides nucléiques (matrice)												
	Ligation adaptateurs												
Méthode d'amplification	PCR en émulsion			« Bridge PCR »				PCR en émulsion					
Méthode de séquençage	Synthèse (Pyroséquençage)			Synthèse				Ligation					
Durée de séquençage/run	10h	10h	20h	26h	8jrs	8jrs	14jrs	2h			12jrs	8jrs	8jrs
Capacité (Mb) séquençage/run	50	500	900	1500	100000	200000	95000	>10	>100	>1000	70000	80000	150000
Taille moyenne des reads	400	400	700	150+150	100+100	100+100	150+150	100	>100	>100	50+35	75+35	75+35
Coût ($) /run	1100	6200		750	10000	20000	11500	500	750	950	8150	6100	10500
Coût machine + annexes ((K$))	110+25	500+30		125	560	690	250	50+20			480+55	350+55	600+55
Exactitude de séquençage (%)	99	99		99,9	99,9	99,9	99,9	99			99,95	99,95	99,99

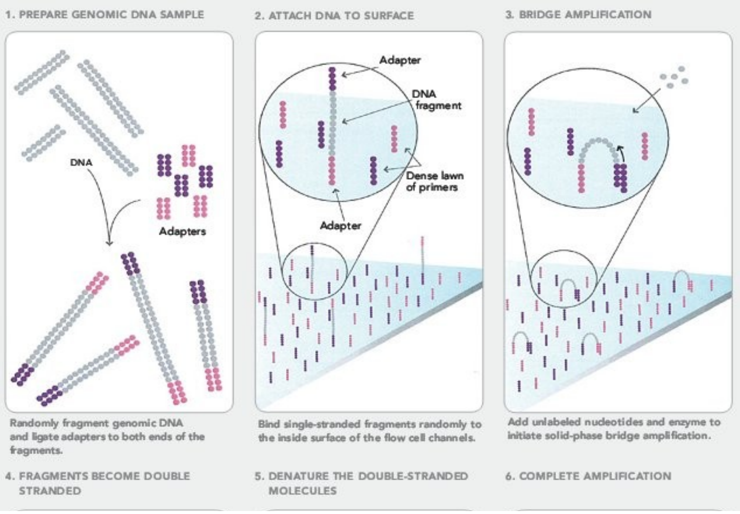
Technological variability

- Types of reads
  - Long (> 200 bp … 40kb)
  - Short (16 bp ...200 bp)
- Number of reads
  - millions … billions
- Strand specific or not
- Paired or not
- Different biases

Illumina library preparation


AAAAAAAAA
TTTTTTTTTTTB
1) PolyA+ RNA captured
2) RNA fragmented and primed
3) First strand cDNA synthesized
4) Second strand cDNA synthesized
p
A
A
p
5) 3’ ends adenylated and 5’ ends repaired
6) DNA sequencing adapters ligated
Barcode
Rd1
7) Ligated fragments PCR amplified
Rd2
Index

Illumina Sequencing protocol



1. PREPARE GENOMIC DNA SAMPLE

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

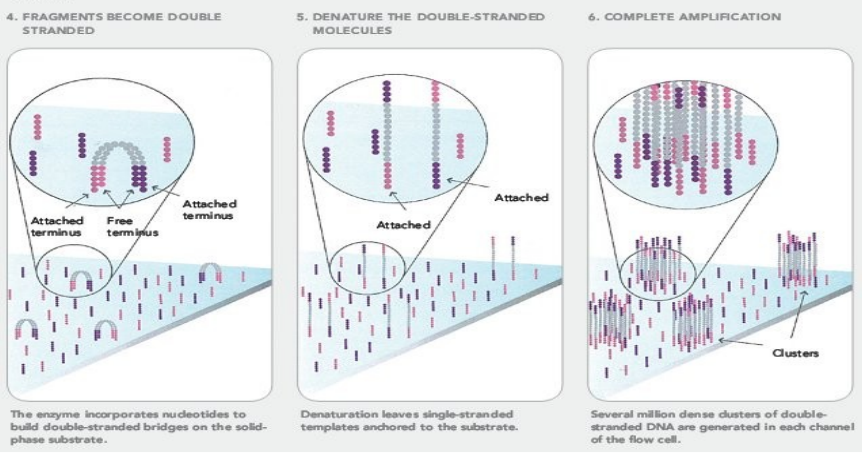
3. BRIDGE AMPLIFICATION

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED

5. DENATURE THE DOUBLE-STRANDED MOLECULES

6. COMPLETE AMPLIFICATION



4. FRAGMENTS BECOME DOUBLE STRANDED

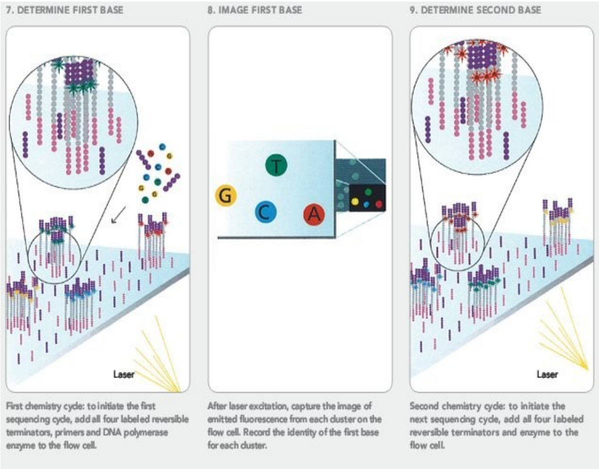
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES

Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION

Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.



7. DETERMINE FIRST BASE

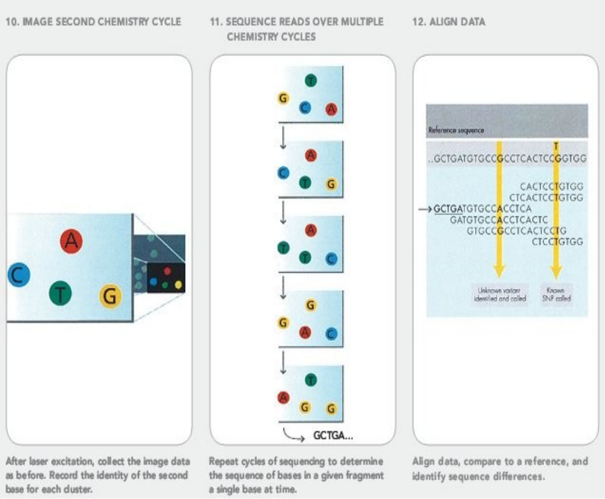
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

9. DETERMINE SECOND BASE

Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.



10. IMAGE SECOND CHEMISTRY CYCLE

After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

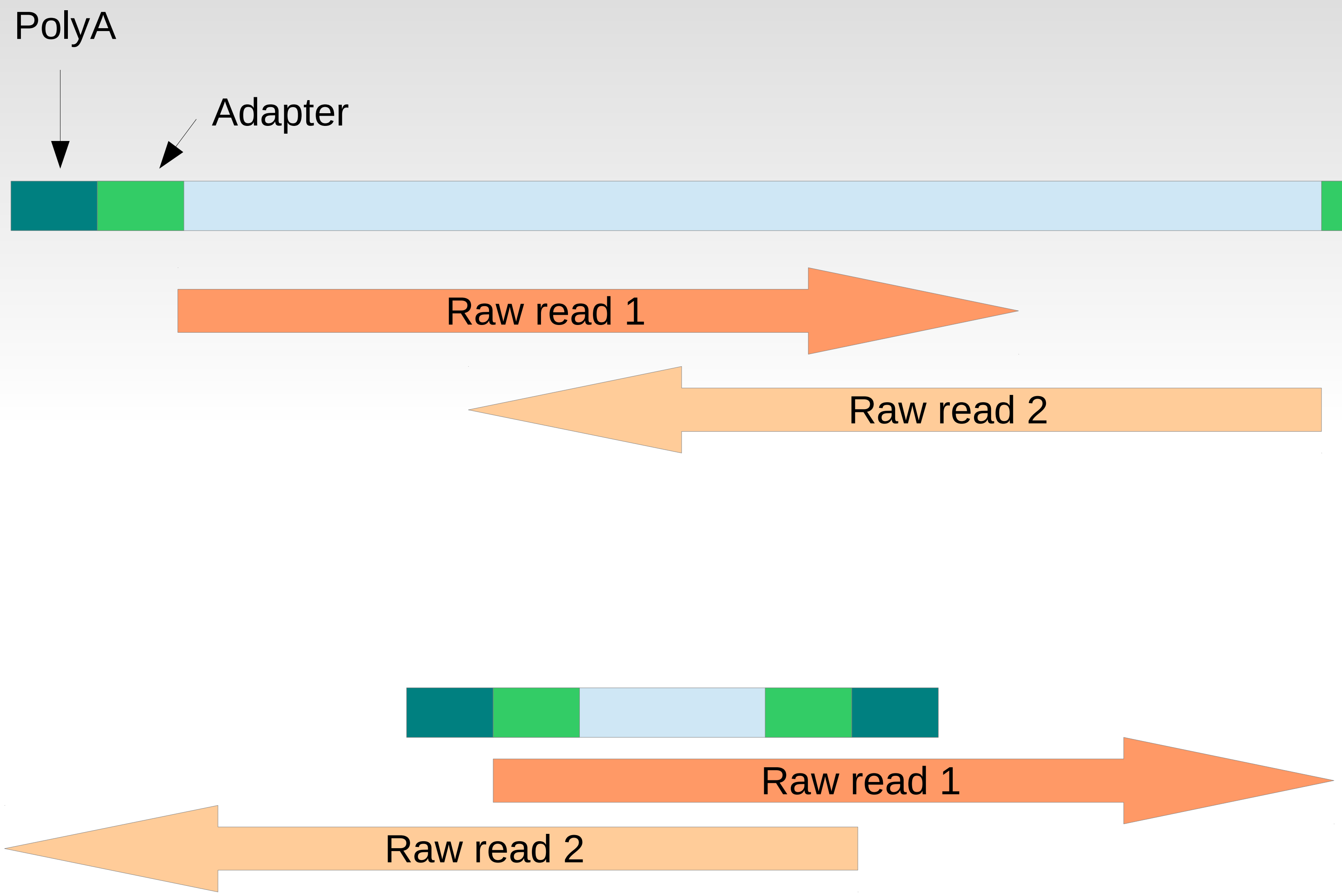
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

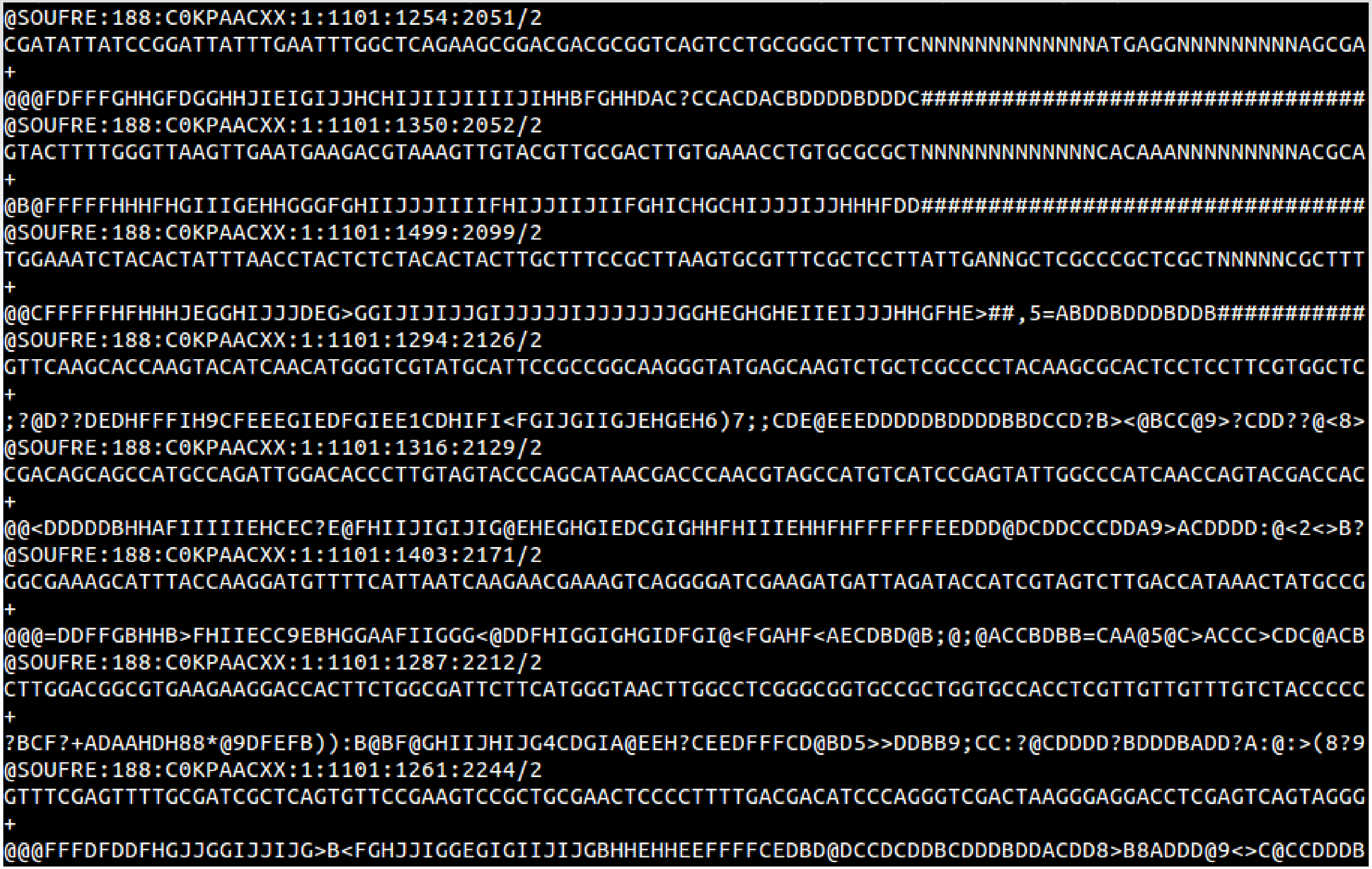
12. ALIGN DATA

Align data, compare to a reference, and identify sequence differences.

For small or empty inserts

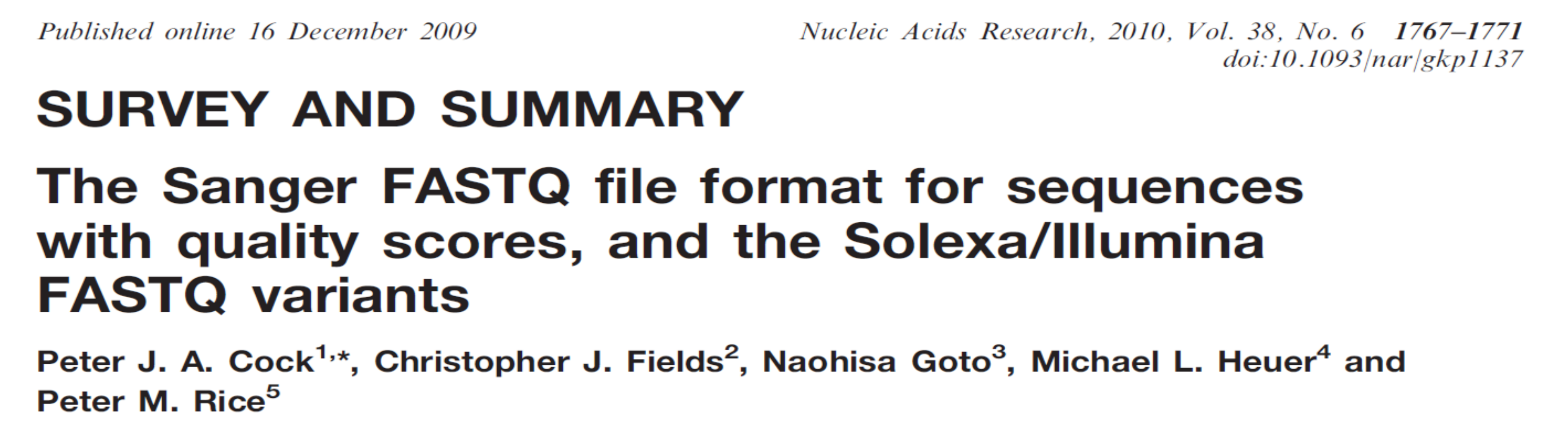


The output : fastq file



```
@SOUFRE:188:C0KPAACXX:1:1101:1254:2051/2
CGATATTATCCGGATTATTTGAATTTGGCTCAGAAGCGGACGACGCGGTCAGTCCTGCGGGCTTCTTCNNNNNNNNNNNNNATGAGGNNNNNNNNNAGCGA
+
@@@FDFFFGHHGFDGGHHJIEIGIJJHCHIJIIJIIIIJIHHBFGHHDAC?CCACDACBDDDDBDDDC#################################
@SOUFRE:188:C0KPAACXX:1:1101:1350:2052/2
GTACTTTTGGGTTAAGTTGAATGAAGACGTAAAGTTGTACGTTGCGACTTGTGAAACCTGTGCGCGCTNNNNNNNNNNNNNCACAAANNNNNNNNNACGCA
+
@B@FFFFFHHHFHGIIIGEHHGGGFGHIIJJJIIIIFHIJJIIJIIFGHICHGCHIJJJIJJHHHFDD#################################
@SOUFRE:188:C0KPAACXX:1:1101:1499:2099/2
TGGAAATCTACACTATTTAACCTACTCTCTACACTACTTGCTTTCCGCTTAAGTGCGTTTCGCTCCTTATTGANNGCTCGCCCGCTCGCTNNNNNCGCTTT
+
@@CFFFFFHFHHHJEGGHIJJJDEG>GGIJIJIJJGIJJJJJIJJJJJJJGGHEGHGHEIIEIJJJHHGFHE>##,5=ABDDBDDDBDDB###########
@SOUFRE:188:C0KPAACXX:1:1101:1294:2126/2
GTTCAAGCACCAAGTACATCAACATGGGTCGTATGCATTCCGCCGGCAAGGGTATGAGCAAGTCTGCTCGCCCCTACAAGCGCACTCCTCCTTCGTGGCTC
+
;?@D??DEDHFFFIH9CFEEEGIEDFGIEE1CDHIFI<FGIJGIIGJEHGEH6)7;;CDE@EEEDDDDDBDDDDBBDCCD?B><@BCC@9>?CDD??@<8>
@SOUFRE:188:C0KPAACXX:1:1101:1316:2129/2
CGACAGCAGCCATGCCAGATTGGACACCCTTGTAGTACCCAGCATAACGACCCAACGTAGCCATGTCATCCGAGTATTGGCCCATCAACCAGTACGACCAC
+
@@<DDDDDBHHAFIIIIIEHCEC?E@FHIIJIGIJIG@EHEGHGIEDCGIGHHFHIIIEHHFHFFFFFFEEDDD@DCDDCCCDDA9>ACDDDD:@<2<>B?
@SOUFRE:188:C0KPAACXX:1:1101:1403:2171/2
GGCGAAAGCATTTACCAAGGATGTTTTCATTAATCAAGAACGAAAGTCAGGGGATCGAAGATGATTAGATACCATCGTAGTCTTGACCATAAACTATGCCG
+
@@@=DDFFGBHHB>FHIIECC9EBHGGAAFIIGGG<@DDFHIGGIGHGIDFGI@<FGAHF<AECDBD@B;@;@ACCBDBB=CAA@5@C>ACCC>CDC@ACB
@SOUFRE:188:C0KPAACXX:1:1101:1287:2212/2
CTTGGACGGCGTGAAGAAGGACCACTTCTGGCGATTCTTCATGGGTAACTTGGCCTCGGGCGGTGCCGCTGGTGCCACCTCGTTGTTGTTTGTCTACCCCC
+
?BCF?+ADAAHDH88*@9DFEFB)):B@BF@GHIIJHIJG4CDGIA@EEH?CEEDFFFCD@BD5>>DDBB9;CC:?@CDDDD?BDDDBADD?A:@:>(8?9
@SOUFRE:188:C0KPAACXX:1:1101:1261:2244/2
GTTTCGAGTTTTGCGATCGCTCAGTGTTCCGAAGTCCGCTGCGAACTCCCCTTTTGACGACATCCCAGGGTCGACTAAGGGAGGACCTCGAGTCAGTAGGG
+
@@@FFFDFDDFHGJJGGIJJIJG>B<FGHJJIGGEGIGIIJIJGBHHEHHEEFFFFCEDBD@DCCDCDDBCDDDBDDACDD8>B8ADDD@9<>C@CCDDDB
```

Fastq file format



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]

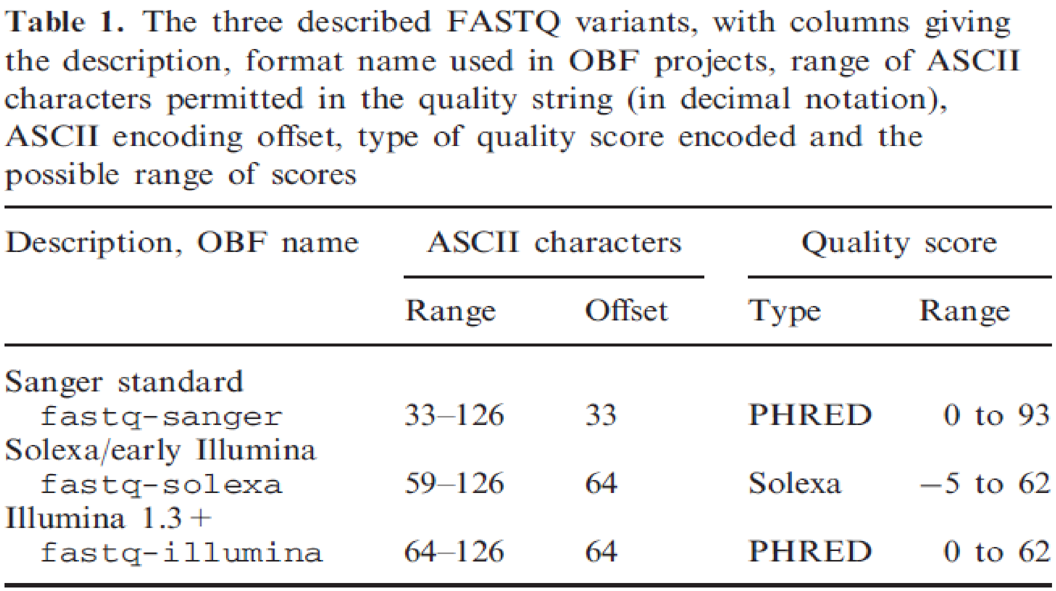
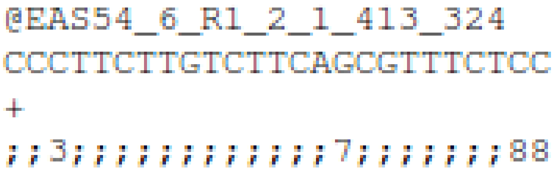


Table 1. The three described FASTQ variants, with columns giving the description, format name used in OBF projects, range of ASCII characters permitted in the quality string (in decimal notation), ASCII encoding offset, type of quality score encoded and the possible range of scores

Description, OBF name	ASCII characters		Quality score	
	Range	Offset	Type	Range
Sanger standard `fastq-sanger`	33–126	33	PHRED	0 to 93
Solexa/early Illumina `fastq-solexa`	59–126	64	Solexa	−5 to 62
Illumina 1.3+ `fastq-illumina`	64–126	64	PHRED	0 to 62

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

$$Q_{\text{Solexa}} = -10 \times \log_{10}\left(\frac{P_e}{1 - P_e}\right)$$



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

Usual questions

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

ENCODE answers in 2011

- RNA-Seq is not a mature technology.
- Experiments should be performed with two or more biological replicates, unless there is a compelling reason why this is impractical or wasteful
- A typical R2 (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 between 0.92 to 0.98. Experiments with biological correlations that fall below 0.9 should be either be repeated or explained.
- Between 30M and 100M reads per sample depending on the study.
- NB. Guidelines for the information to publish with the data.

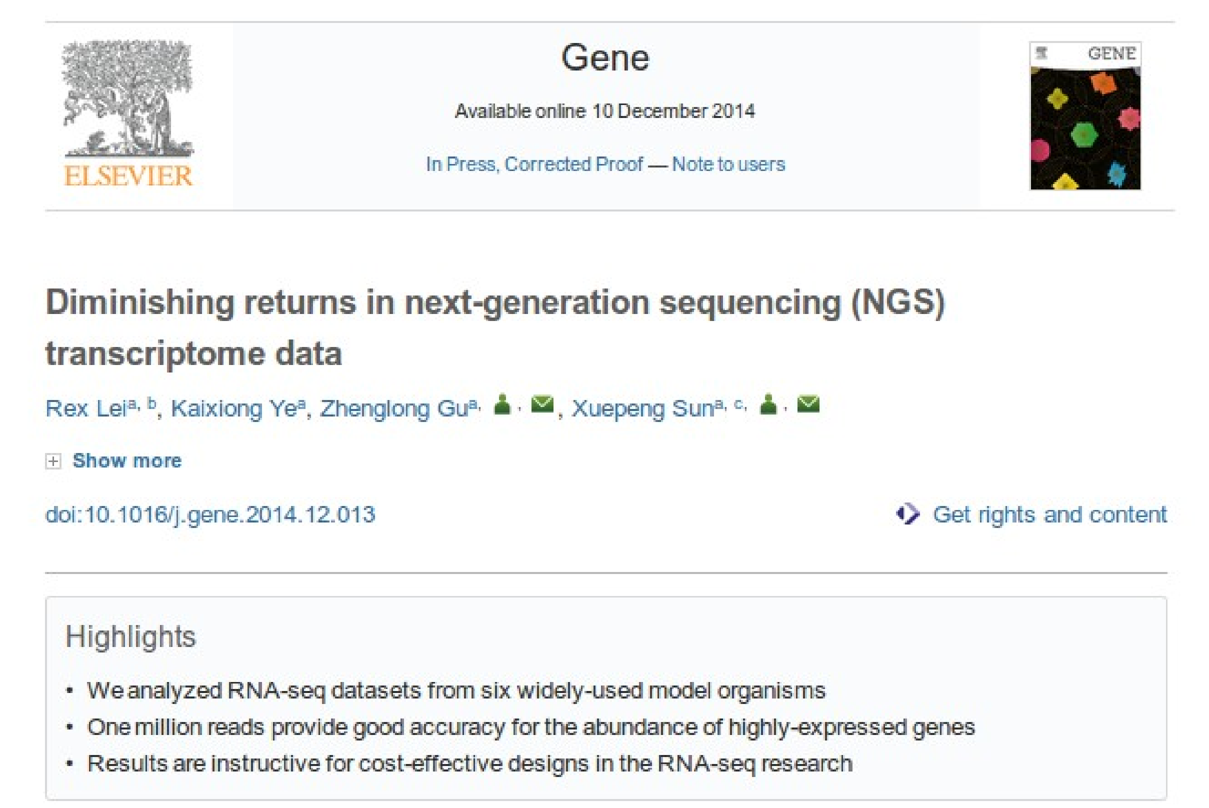
ENCODE

Encyclopedia of DNA Elements

http://encodeproject.org/ENCODE/dataStandards.html

Statistician answers

- Less reads
- More samples



Gene

Available online 10 December 2014

In Press, Corrected Proof — Note to users

Diminishing returns in next-generation sequencing (NGS) transcriptome data

Rex Lei[a, b], Kaixiong Ye[a], Zhenglong Gu[a], Xuepeng Sun[a, c]

Show more

doi:10.1016/j.gene.2014.12.013

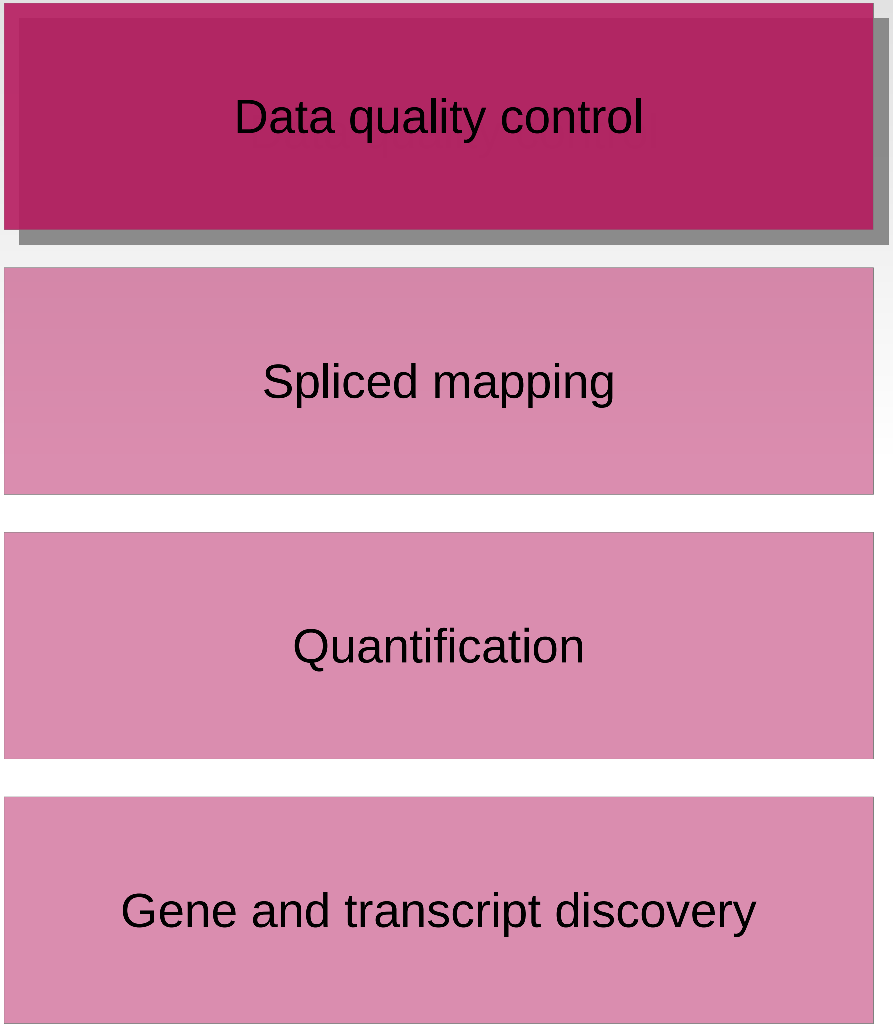
Get rights and content

Highlights

- We analyzed RNA-seq datasets from six widely-used model organisms
- One million reads provide good accuracy for the abundance of highly-expressed genes
- Results are instructive for cost-effective designs in the RNA-seq research

http://www.sciencedirect.com/science/article/pii/S0378111914013869

Analysis workflow



Data quality control

Spliced mapping

Quantification

Gene and transcript discovery

Verifying RNASeq raw data

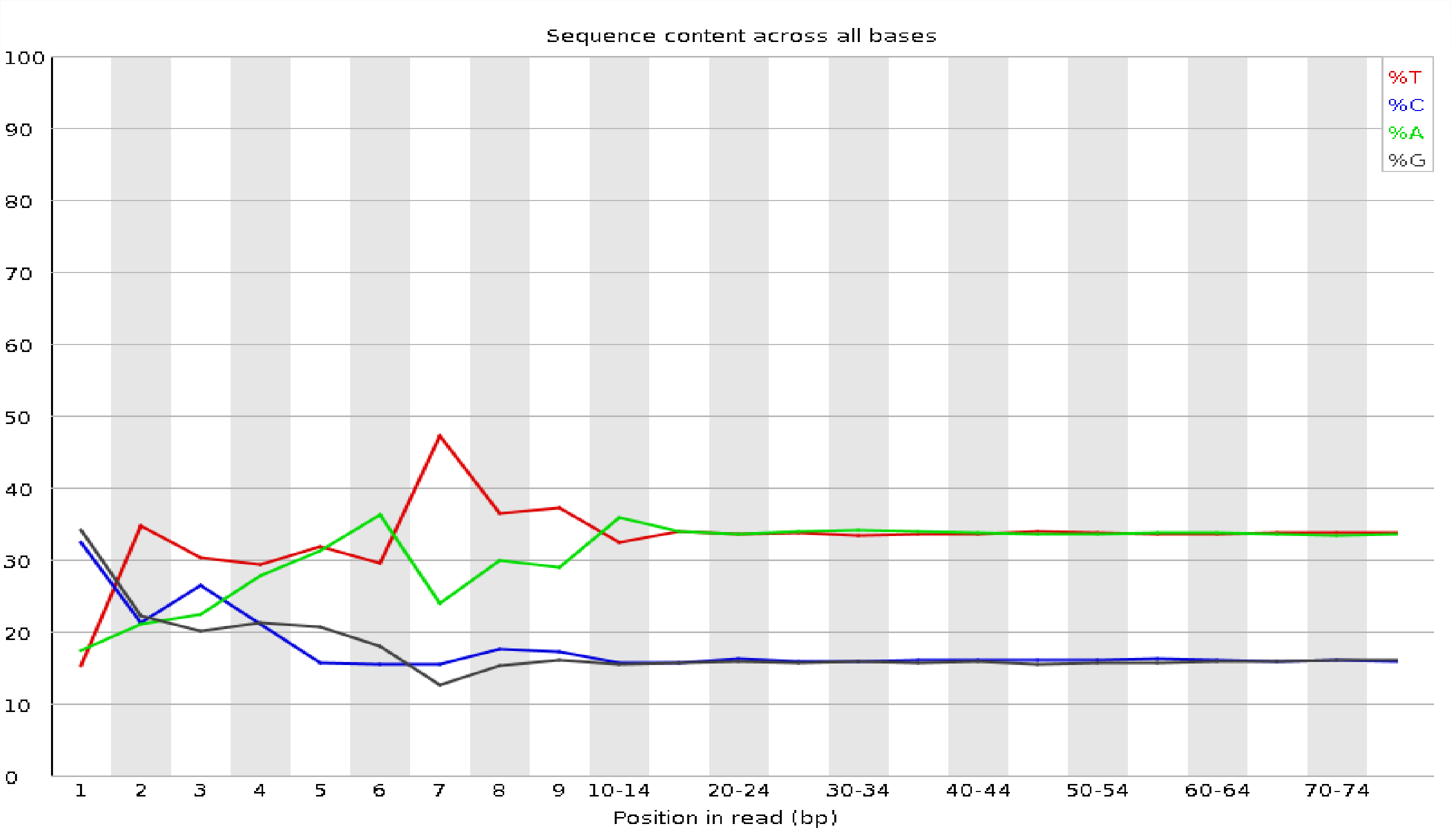
FastQC :

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

- *Import of data from BAM, SAM or FastQ files*
- *quick overview*
- *Summary graphs and tables to quickly assess your data*
- *Export of results to an HTML report*
- *Offline operation to allow automated generation of reports*
- *Color code to check quickly the quality*

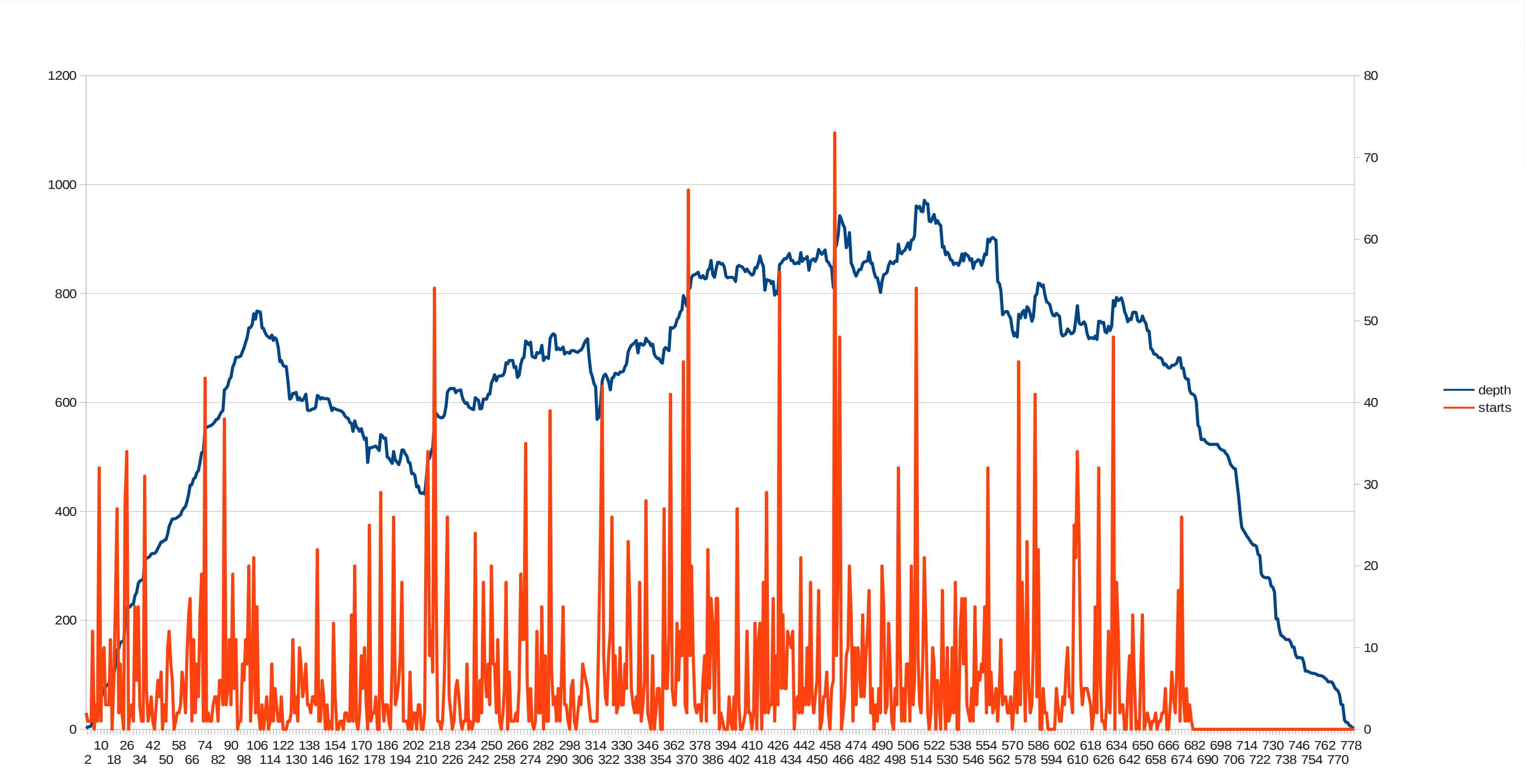
Quality control

- Technical characteristics conformity
- Contamination search
- Classical RNA-Seq biases
  - Example : hexamer random priming

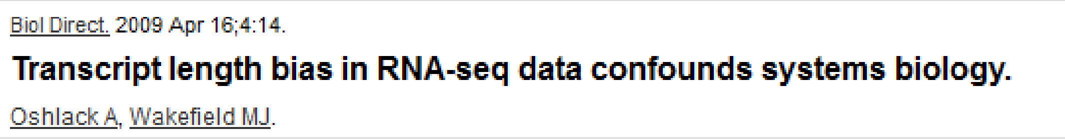


Bias impact on alignment

- Orange = reads start sites
- Blue = coverage



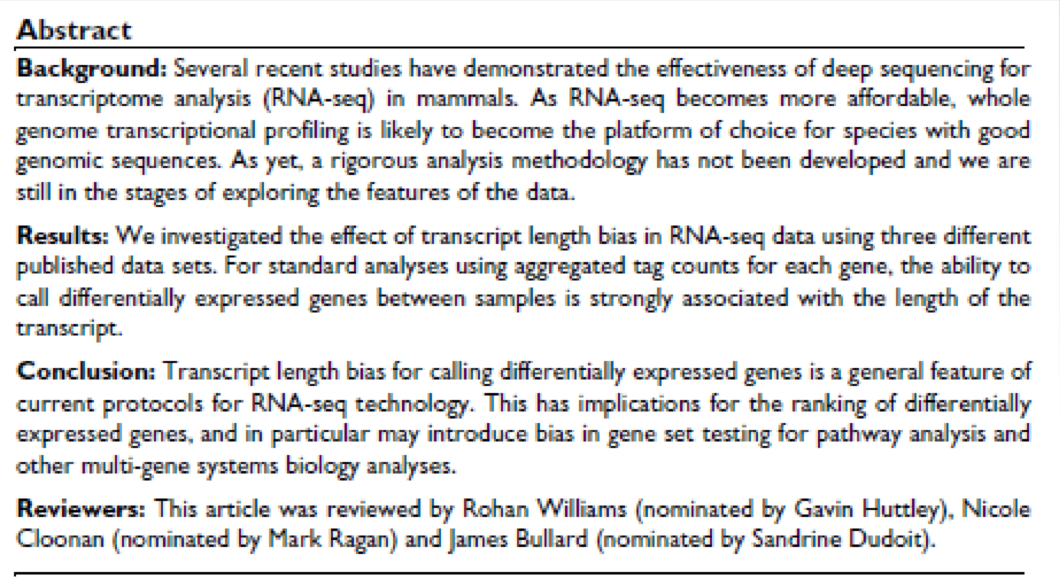
Transcript length bias



Biol Direct. 2009 Apr 16;4:14.

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

Oshlack A, Wakefield MJ.



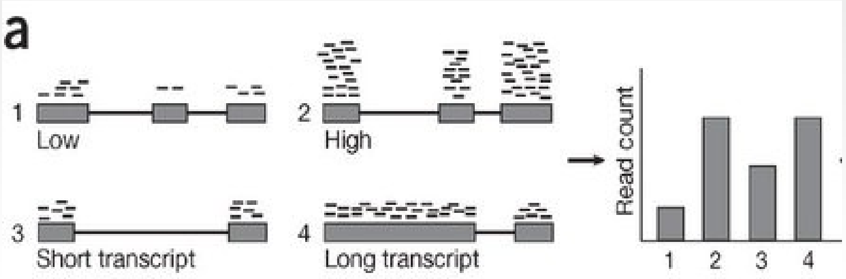
Abstract

Background: Several recent studies have demonstrated the effectiveness of deep sequencing for transcriptome analysis (RNA-seq) in mammals. As RNA-seq becomes more affordable, whole genome transcriptional profiling is likely to become the platform of choice for species with good genomic sequences. As yet, a rigorous analysis methodology has not been developed and we are still in the stages of exploring the features of the data.

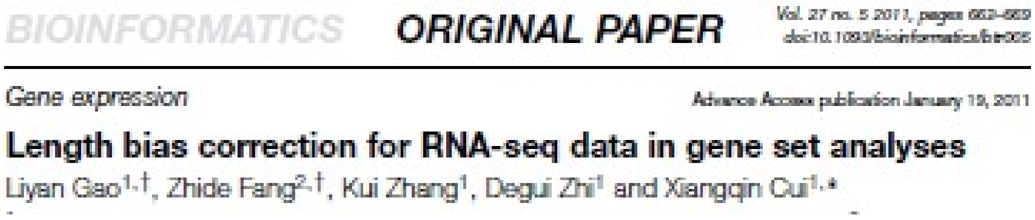
Results: We investigated the effect of transcript length bias in RNA-seq data using three different published data sets. For standard analyses using aggregated tag counts for each gene, the ability to call differentially expressed genes between samples is strongly associated with the length of the transcript.

Conclusion: Transcript length bias for calling differentially expressed genes is a general feature of current protocols for RNA-seq technology. This has implications for the ranking of differentially expressed genes, and in particular may introduce bias in gene set testing for pathway analysis and other multi-gene systems biology analyses.

Reviewers: This article was reviewed by Rohan Williams (nominated by Gavin Huttley), Nicole Cloonan (nominated by Mark Ragan) and James Bullard (nominated by Sandrine Dudoit).



- *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,*]

Hands-on

- Run fastqc (fastqc)on one of the fastq files found on you USB stick
- In groups explain the different graphics produced by fastqc

Take home messages on quality analysis

Elements to be checked :

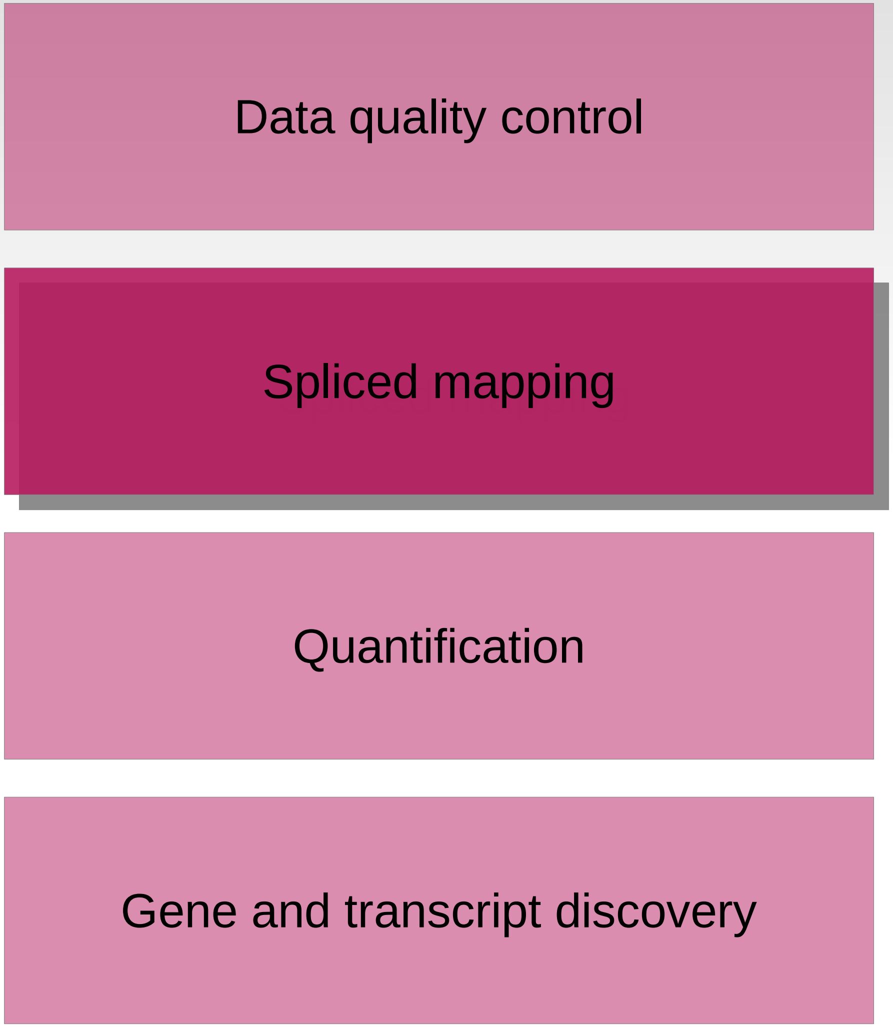
- Random priming effect
- K-mer (polyA, polyT)

Alignment on reference for the second quality check and filtering.

A good run has :

- the expected number of reads (2x500millions / flowcell),
- the expected reads length (100pb),
- a random nucleotides selection and the GC%,
- a high alignment rate : very few unmapped reads, pairs mapped on opposite strands (shown in the next part).

Analyse workflow



Data quality control

Spliced mapping

Quantification

Gene and transcript discovery

Where to find a reference genome?

- Fasta file
- Retrieving the genome file:
  - 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

⚠ The chromosome names should be the same in the gtf file and fasta file.

Reference transcriptome file

What is a GTF file ?

- Tab delimited text 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

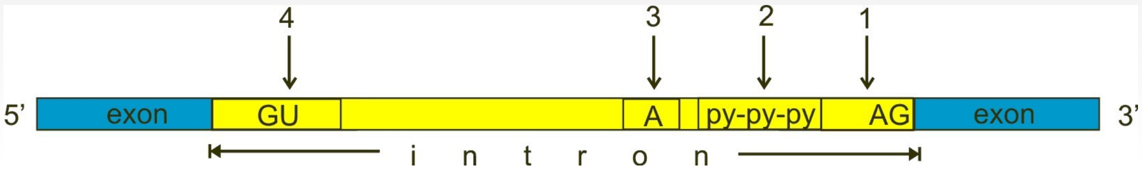
<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes] [comments]

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.

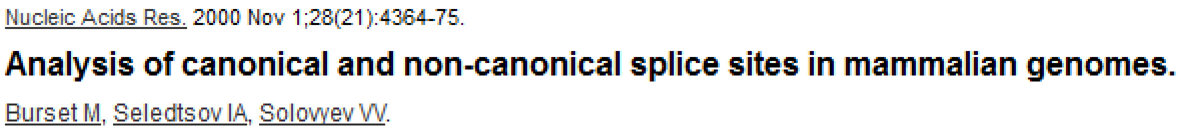
Splice sites

- 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.

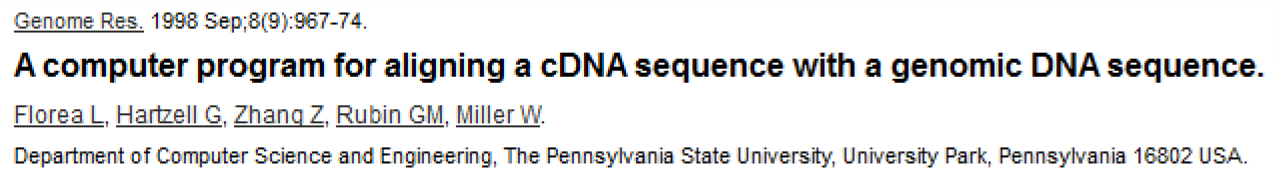
- Trans-splicing : splicing that joins two exons that are not within the same RNA transcript

Spliced alignment

- The recognition of exon/intron junctions can be inferred from the reads that overlap the splicing sites. The resulting spliced reads can produce very short alignments, part of the read will not map contiguously to the reference.

  → therefore this approach requires a dedicated algorithm

- Generation :
  - Sim4



Genome Res. 1998 Sep;8(9):967-74.

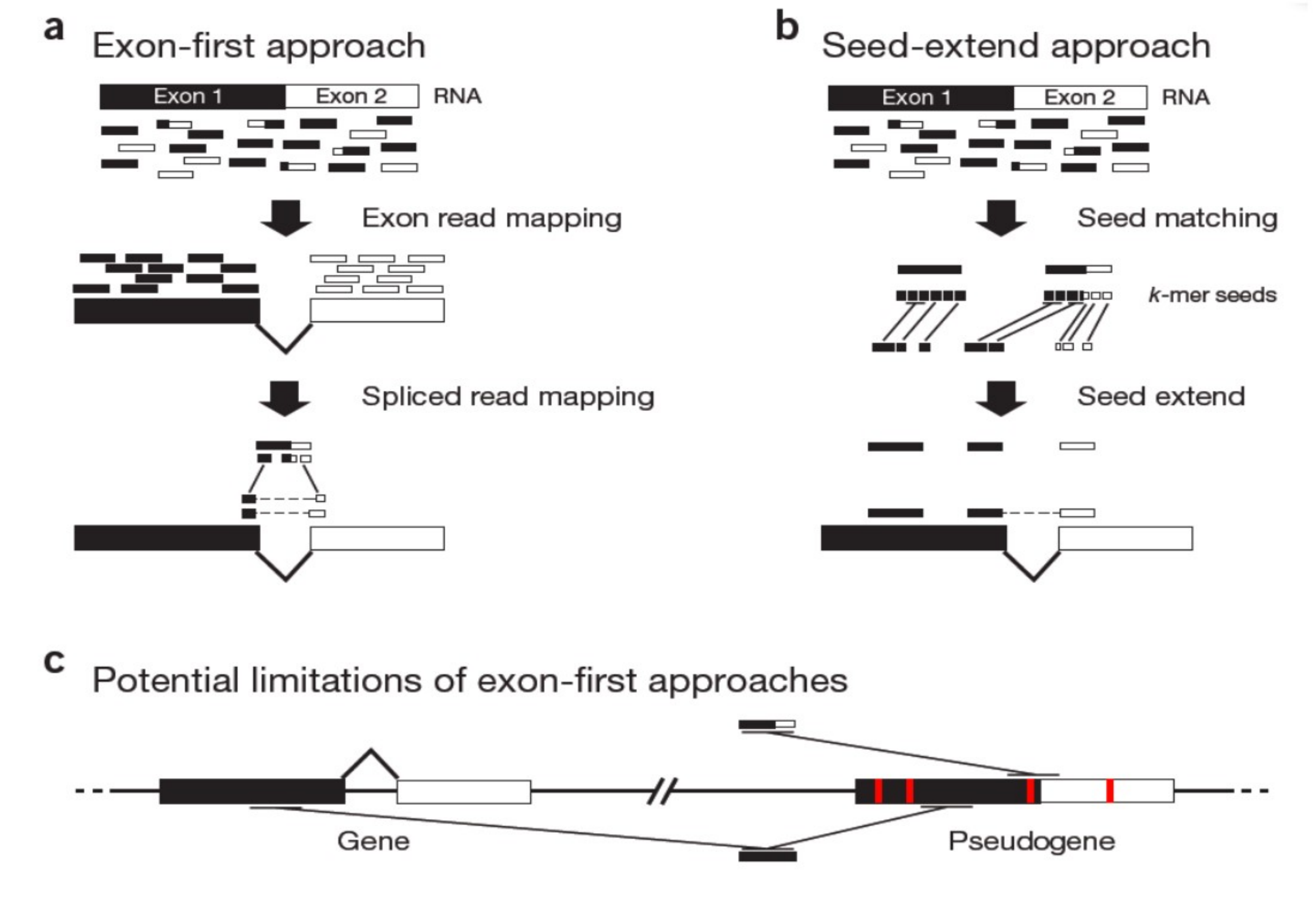
A computer program for aligning a cDNA sequence with a genomic DNA sequence.

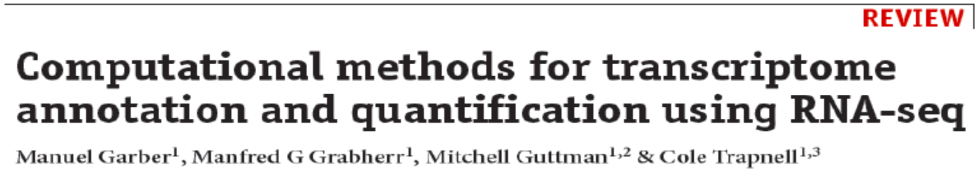
Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W.

Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802 USA.

  - Seqanswer : http://seqanswers.com/wiki/Software/list
- Idea :
  - Database of potential splice junction sequences (known)
  - splice canonical / non canonical site search (seed then mapping)

Exon first vs seed extend

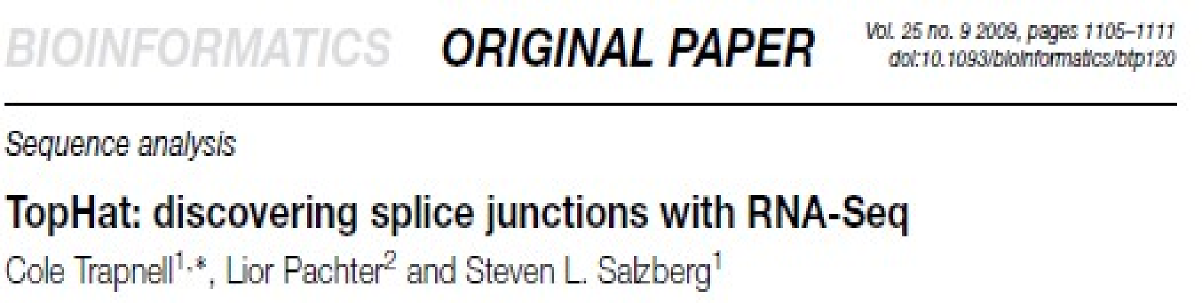



REVIEW

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]

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

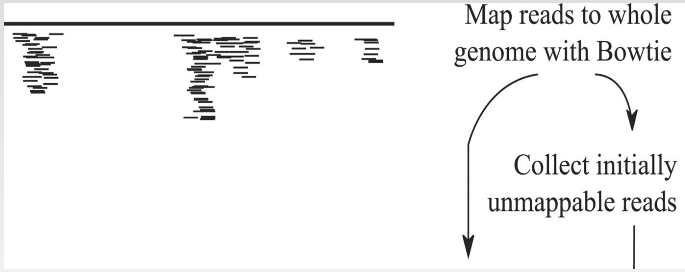
Cole Trapnell[1,*], Lior Pachter[2] and Steven L. Salzberg[1]

http://tophat.cbcb.umd.edu/

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

http://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools#Spliced_aligners

TopHat initial algorithm : first step

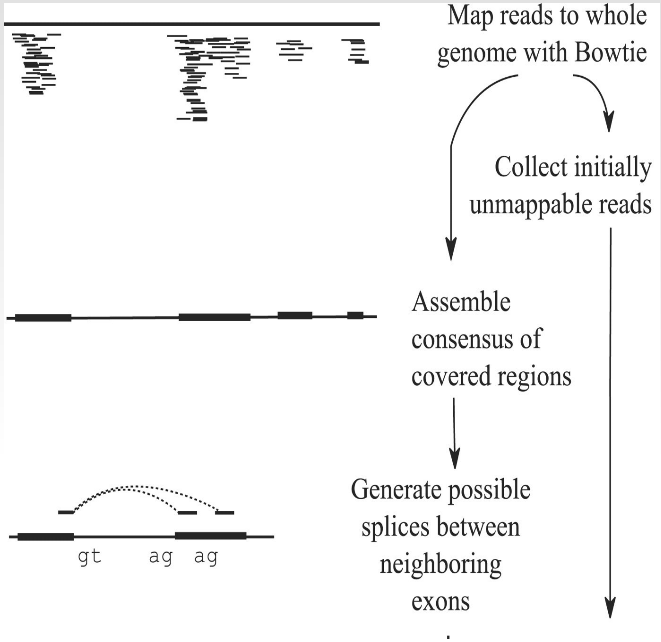

Map reads to whole genome with Bowtie
Collect initially unmappable reads

- TopHat finds junctions by mapping reads to the reference:
  - all reads are mapped to the reference genome using Bowtie
  - reads not mapped to the genome are set aside as IUM (initially unmapped)
  - low complexity reads are discarded
  - for each read : allow until 20 alignments

Trapnell C et al. Bioinformatics 2009;25:1105-1111

Exon assembly process

- TopHat then assembles mapped reads
- Define island: aggregates mapped reads in islands of candidate exons
  - Generate potential donor/acceptor splice sites using neighbouring exons
- Extend islands to cover eventually splice junctions
  - +/- 45 bp from reference on either side of island


Map reads to whole genome with Bowtie
Collect initially unmappable reads
Assemble consensus of covered regions
Generate possible splices between neighboring exons
gt
ag
ag

Trapnell C et al. Bioinformatics 2009;25:1105-1111

Spice junction reference

To map reads to splice junction :

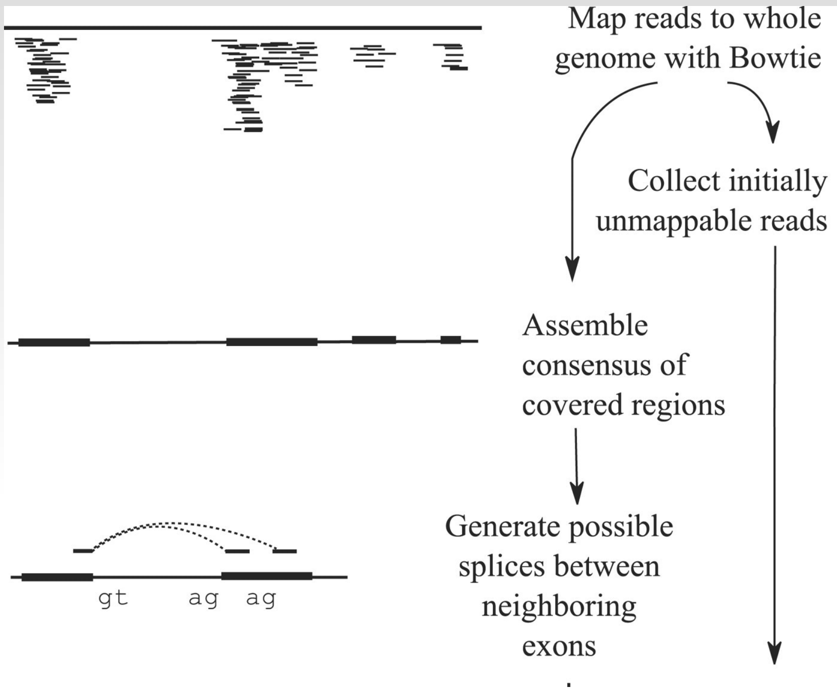
- Enumerate all canonical donor and acceptor sites in islands
    - long (>= 75 bp) reads:

"GT-AG","GC-AG" and "AT-AC" introns

- Shorter reads:

only "GT-AG" introns

- Find all pairings which produce GT-AG introns between islands
    - 70 bp < Intron size < 20,000 bp

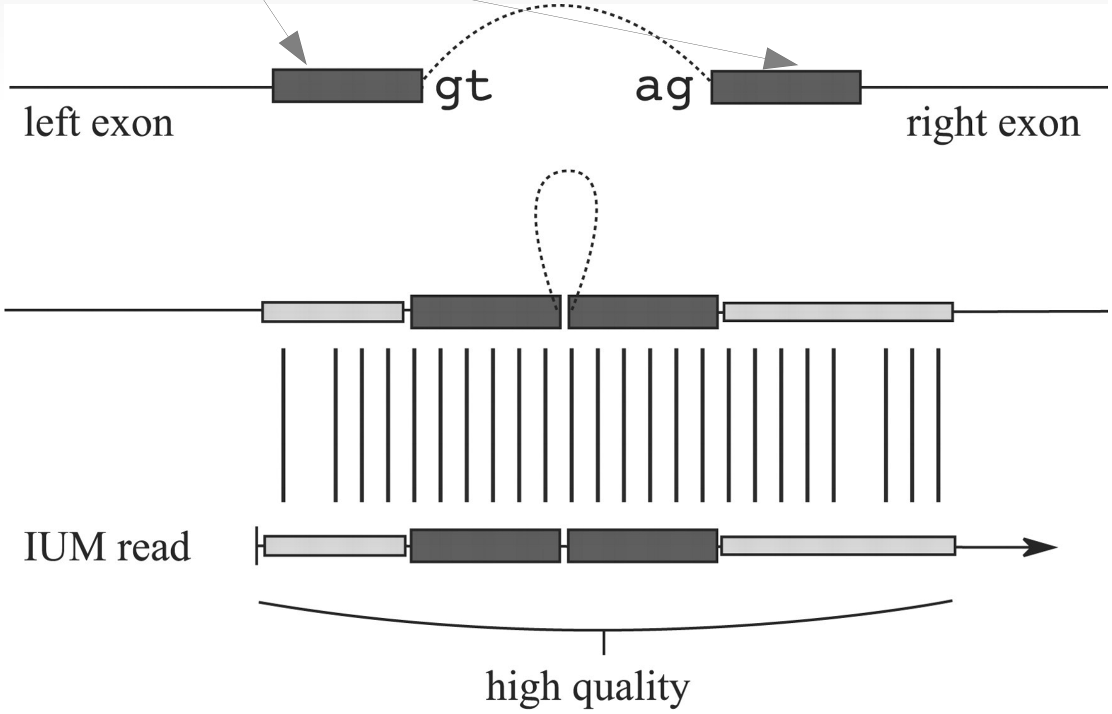

Map reads to whole genome with Bowtie
Collect initially unmappable reads
Assemble consensus of covered regions
Generate possible splices between neighboring exons
gt
ag
ag

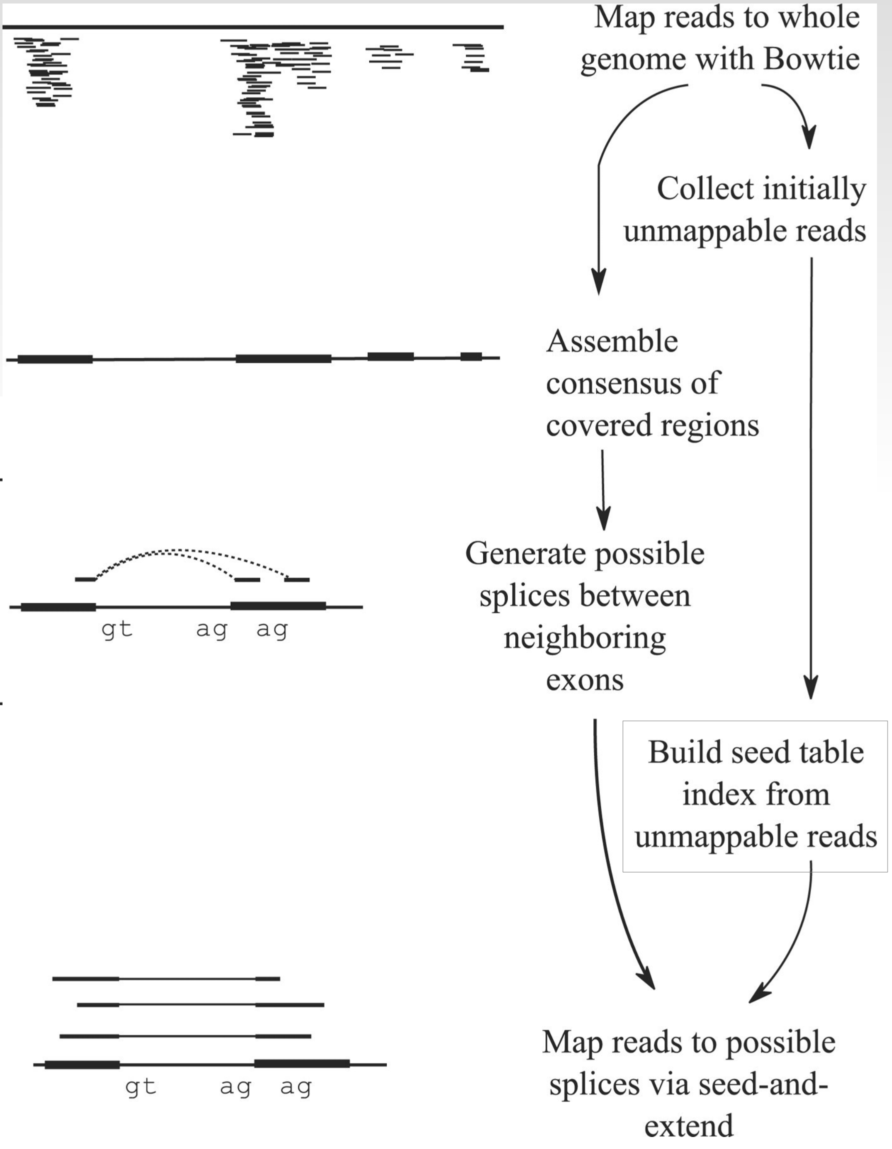
Trapnell C et al. Bioinformatics 2009;25:1105-1111

IUM alignment

- Each possible intron is checked against the IUM

→ seed and extend alignment





Map reads to whole genome with Bowtie

Collect initially unmappable reads

Assemble consensus of covered regions

Generate possible splices between neighboring exons

Build seed table index from unmappable reads

Map reads to possible splices via seed-and-extend

Trapnell C et al. Bioinformatics 2009;25:1105-1111

TopHat Inputs

Inputs :

- bowtie2 index of the genome

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

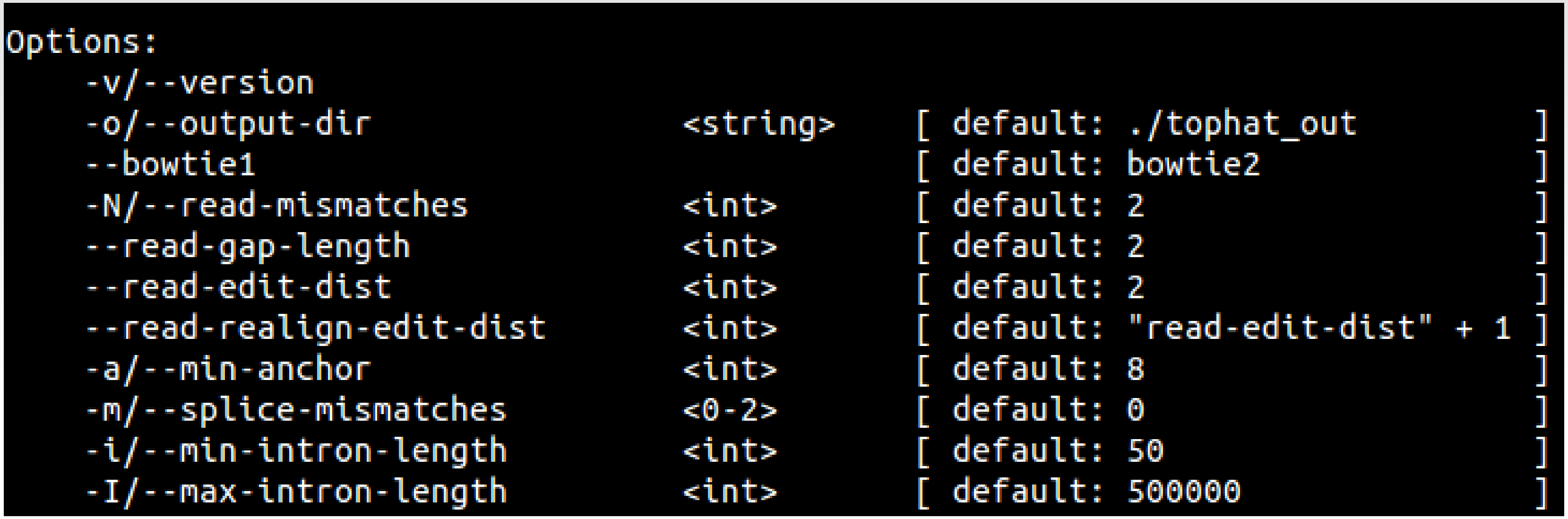
  http://bowtie-bio.sourceforge.net/index.shtml

- file fasta (.fa) of the reference or will be build by bowtie, in the index directory
- File fastq of the reads

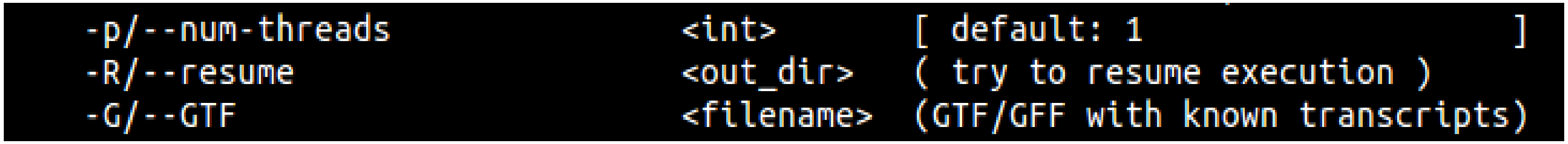
Command lines :

```
bowtie2-build <reference.fasta> <index_base>
tophat [options] <index_base> <reads1_1[,...,readsN_1]><[reads1_2,...readsN_2]>
```

TopHat Options



```
Options:
    -v/--version
    -o/--output-dir                  <string>    [ default: ./tophat_out         ]
    --bowtie1                                    [ default: bowtie2              ]
    -N/--read-mismatches             <int>       [ default: 2                    ]
    --read-gap-length                <int>       [ default: 2                    ]
    --read-edit-dist                 <int>       [ default: 2                    ]
    --read-realign-edit-dist         <int>       [ default: "read-edit-dist" + 1 ]
    -a/--min-anchor                  <int>       [ default: 8                    ]
    -m/--splice-mismatches           <0-2>       [ default: 0                    ]
    -i/--min-intron-length           <int>       [ default: 50                   ]
    -I/--max-intron-length           <int>       [ default: 500000               ]
```



```
    -p/--num-threads                 <int>       [ default: 1                  ]
    -R/--resume                      <out_dir>   ( try to resume execution )
    -G/--GTF                         <filename>  (GTF/GFF with known transcripts)
```

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>*

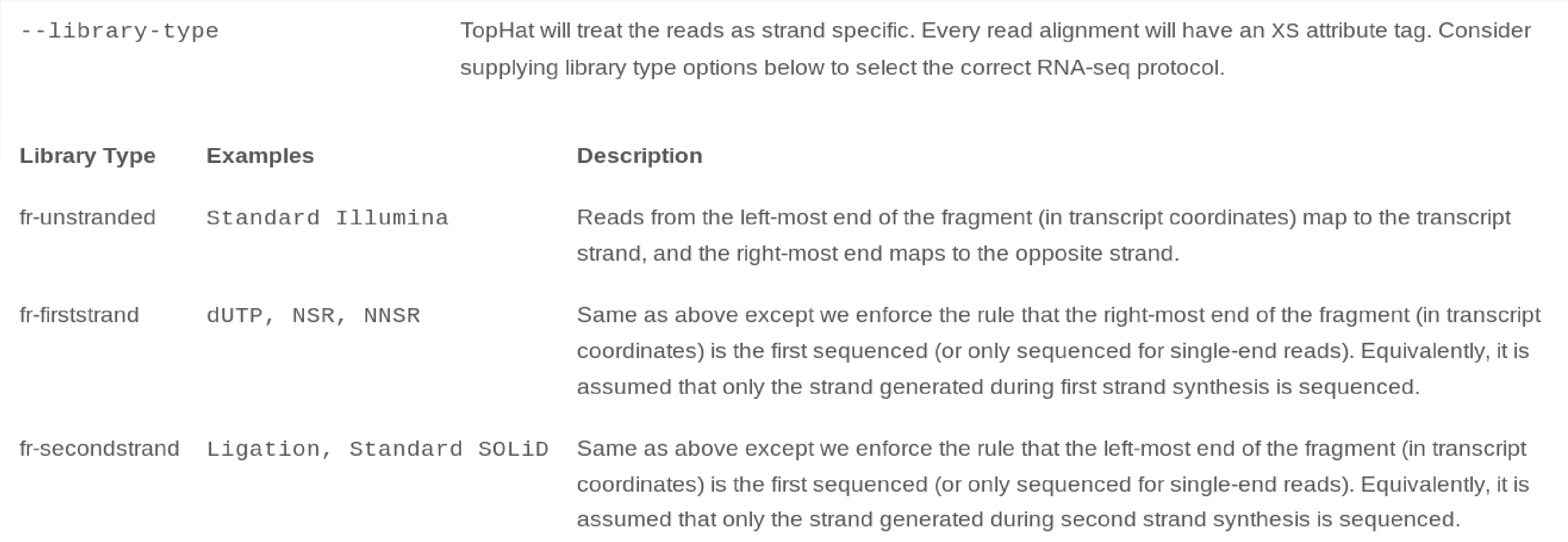
--no-novel-juncs (ignored without -G/-j)

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

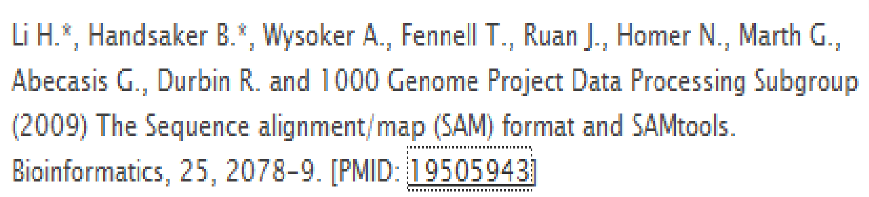
Sequence alignment and map

- SAM (Sequence Alignment/Map) format:
  - Capture all of the critical information about NGS data in a single indexed and compressed file
  - Sharing : data across and tools
  - Generic alignment format
  - SAMTOOLS: provide various

    utilities for manipulating alignments in the SAM format:sorting, merging, indexing...

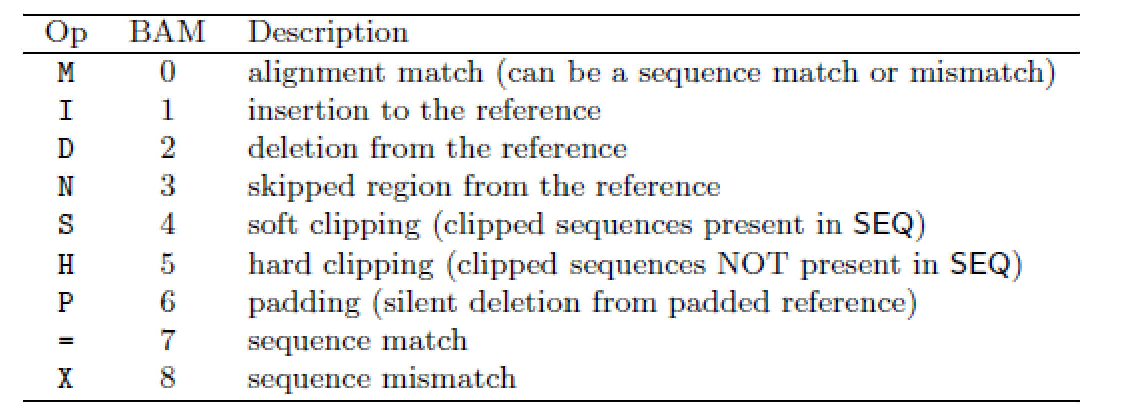
    http://samtools.sourceforge.net/

    http://picard.sourceforge.net/explain-flags.html


Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9. [PMID: 19505943]

Spliced cigar line

- Extend CIGAR strings



Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

- Example: intron de 81 bases

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

Bam & Bed

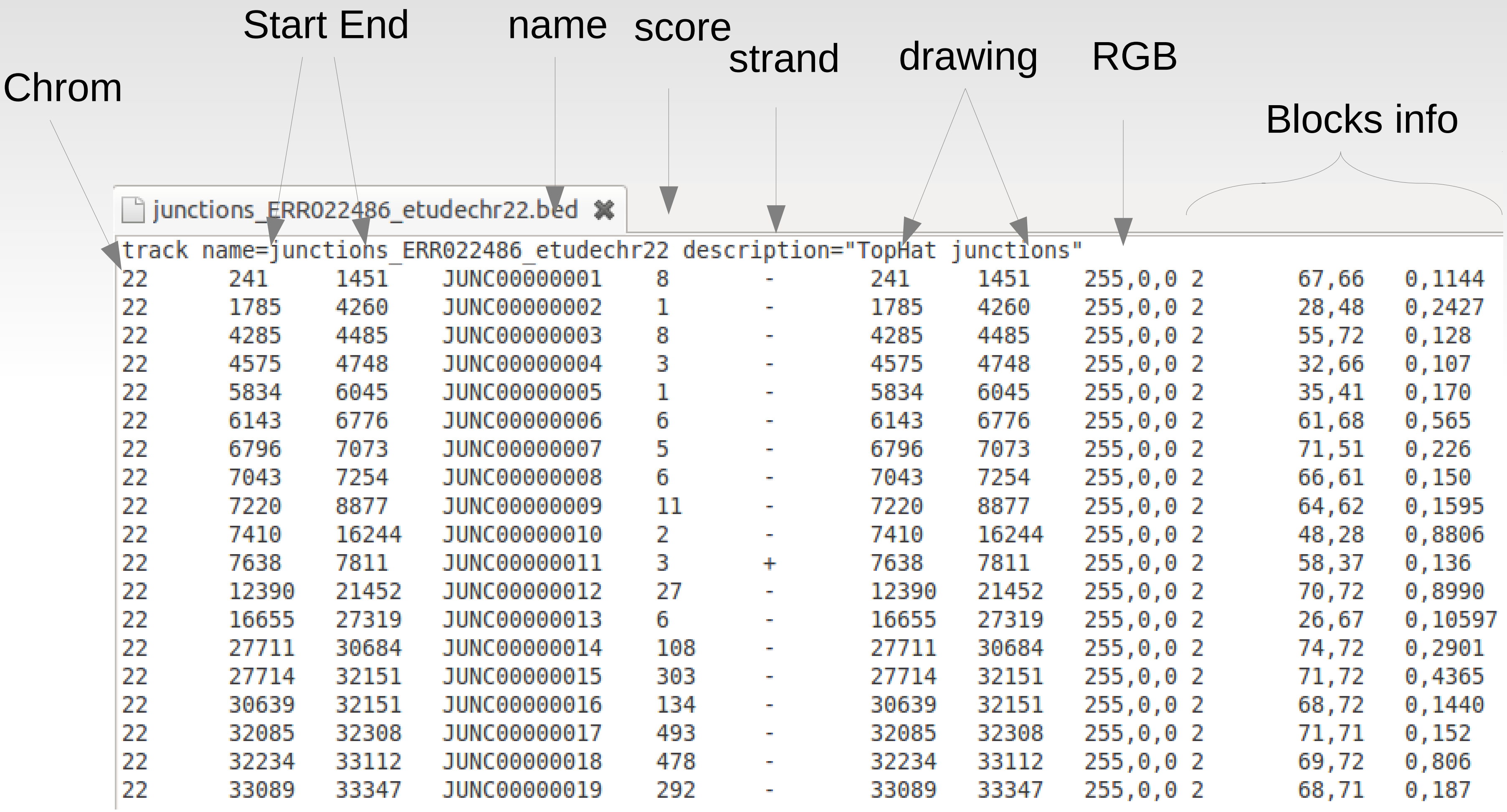
- BAM (Binary Alignment/Map) format:
  - Compressed binary representation of SAM
  - Greatly reduces storage space requirements to about 27% of original SAM
  - Bamtools: reading, writing, and manipulating BAM files

- 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:

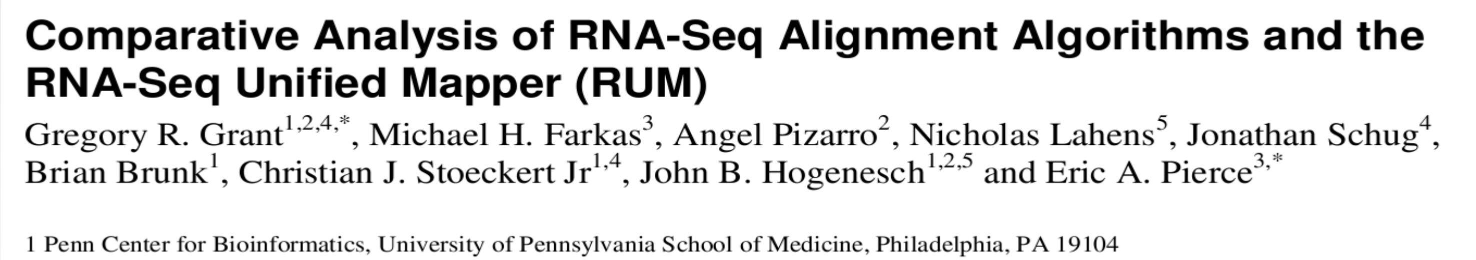
    <chromosome> <start> <end>
  - 9 additional optional BED fields

Bed example



```
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
```

Mapper comparisons



Comparative Analysis of RNA-Seq Alignment Algorithms and the RNA-Seq Unified Mapper (RUM)

Gregory R. Grant[1,2,4,*], Michael H. Farkas[3], Angel Pizarro[2], Nicholas Lahens[5], Jonathan Schug[4], Brian Brunk[1], Christian J. Stoeckert Jr[1,4], John B. Hogenesch[1,2,5] and Eric A. Pierce[3,*]

1 Penn Center for Bioinformatics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104

2 Institute for Translational Medicine and Therapeutics, University

3 F.M. Kirby Center for Molecular Ophthalmology, University of F

4 Department of Genetics, University of Pennsylvania School of M

5Department of Pharmacology, University of Pennsylvania School

Associate Editor: Prof. Ivo Hofacker

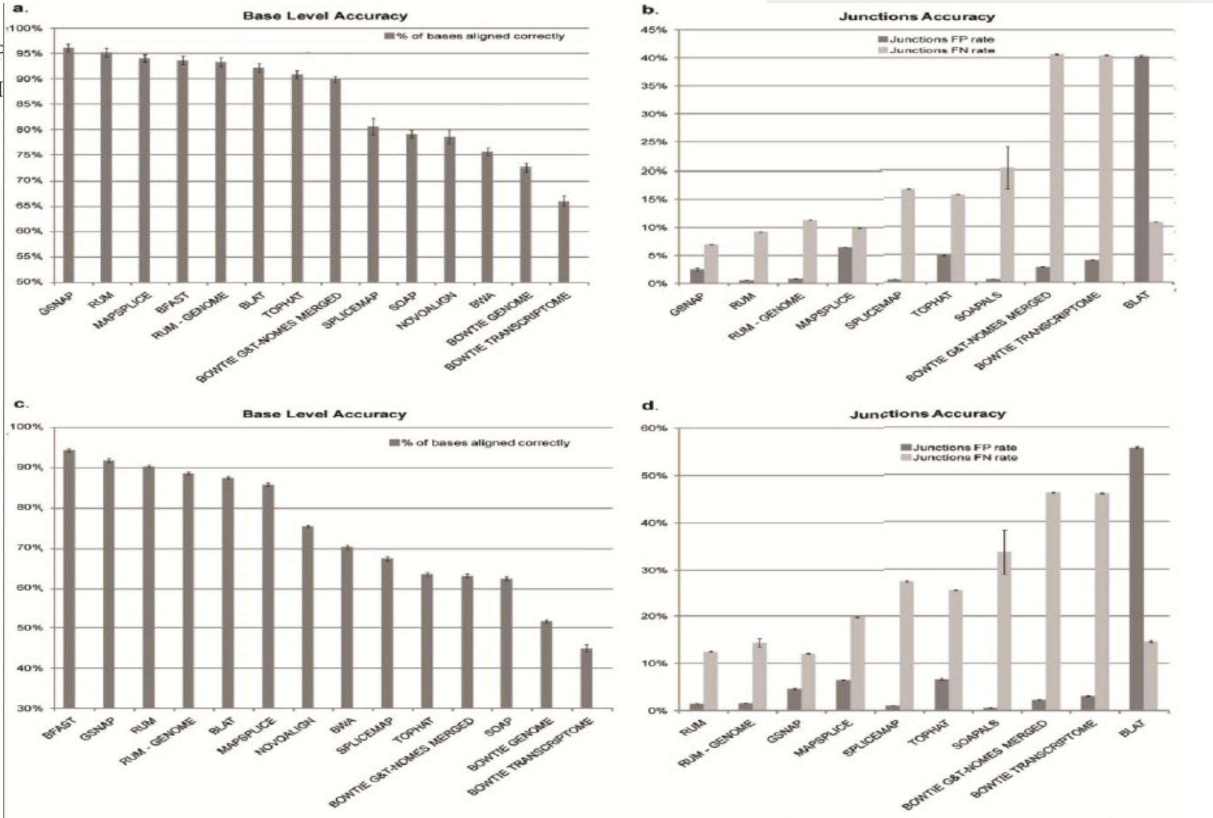


Fig. 6. Accuracy statistics for analyses of simulated data sets. A, B. Simulated data set 1. C,D. Simulated data set 2. Test 1 has low polymorphism and error rates, while Test 2 has moderate polymorphism and error rates. In A and C The dark bars show the base-wise accuracy (the percent of bases that aligned and to the right location); the light bars give the coverage plot accuracy. B and D show the accuracy of the junction calls, dark bars show the false positive (FP) rate and light bars show the false negative (FN) rate. The algorithms are sorted in A and C by accuracy and in B and D by the sum of the FP and FN rates. Results are mean +/- SEM over the three replicate simulated data sets for each test. There is a considerable dropoff in accuracy seen in Test 2 for the algorithms that do not align across indels (SpliceMap, TopHat, and Bowtie). The base-wise accuracy and the FP and FN rates on junction calls are taken in conjunction to determine the overall effectiveness of an algorithm. Based on these results, we conclude that GSPAN, MapSplice and RUM are the ones that are most viable for RNA-Seq alignment.

Hands-in : spliced alignment

- *Index the genome file Danio_rerio.Zv9.62.dna.chromosome.22.fa with bowtie2*
- *Align both reads paired files to the genome using tophat2*
    - *ERR022486_chr22_read1.fastq.gz ERR022486_chr22_read2.fastq.gz*
    - *ERR022488_chr22_read1.fastq.gz ERR022488_chr22_read2.fastq.gz*
    - *Parameters :*
        - *Max intron size : 5kb*
        - *Number of threads : 4*
        - *Use the name of the file ERR022486 ERR022488 as output directory name*
- *Index the accepted_hits.bam file*
- *Count the number of alignments with samtools flagstat for ERR022486*

Hands-in : commands

```
bowtie2-build Danio_rerio.Zv9.62.dna.chromosome.22.fa  Danio_rerio.Zv9.62_chr22
```

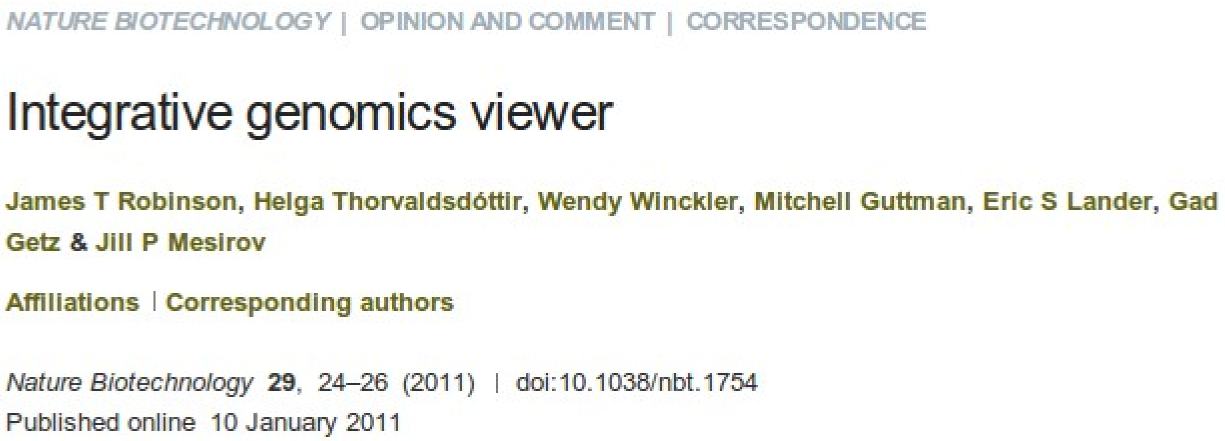
```
tophat -p 4 –output-dir=tophat_ERR022486 -I 5000 Danio_rerio.Zv9.62_chr22
ERR022486_read1.fastq,ERR022486_read2.fastq
```

```
samtools index ERR022486/accepted_hits.bam
```

```
samtools flagstat ERR022486/accepted_hits.bam
```

Visualizing alignments on IGV

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



NATURE BIOTECHNOLOGY | OPINION AND COMMENT | CORRESPONDENCE

Integrative genomics viewer

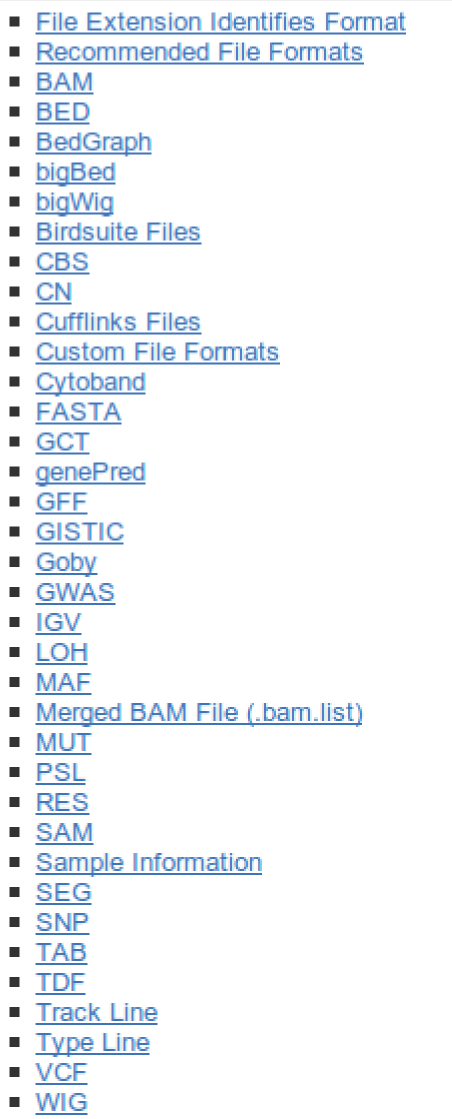
James T Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S Lander, Gad Getz & Jill P Mesirov

Affiliations | Corresponding authors

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

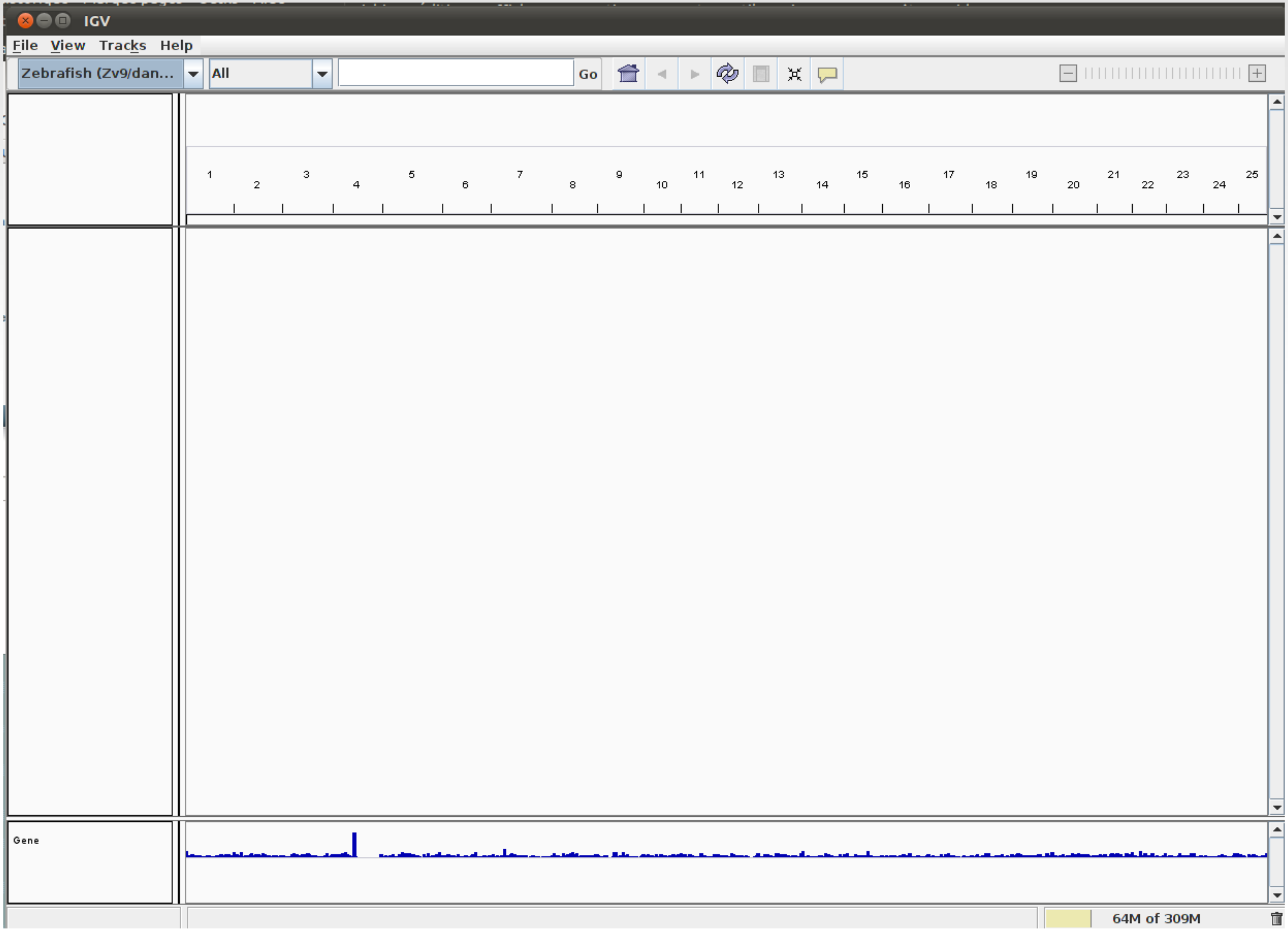
Visualizing alignments on IGV

- High-performance visualization tool
- Interactive exploration of large datasets
- Supports a wide variety of data types
- Documentations available
- Developed at the Broad Institute of MIT and Harvard



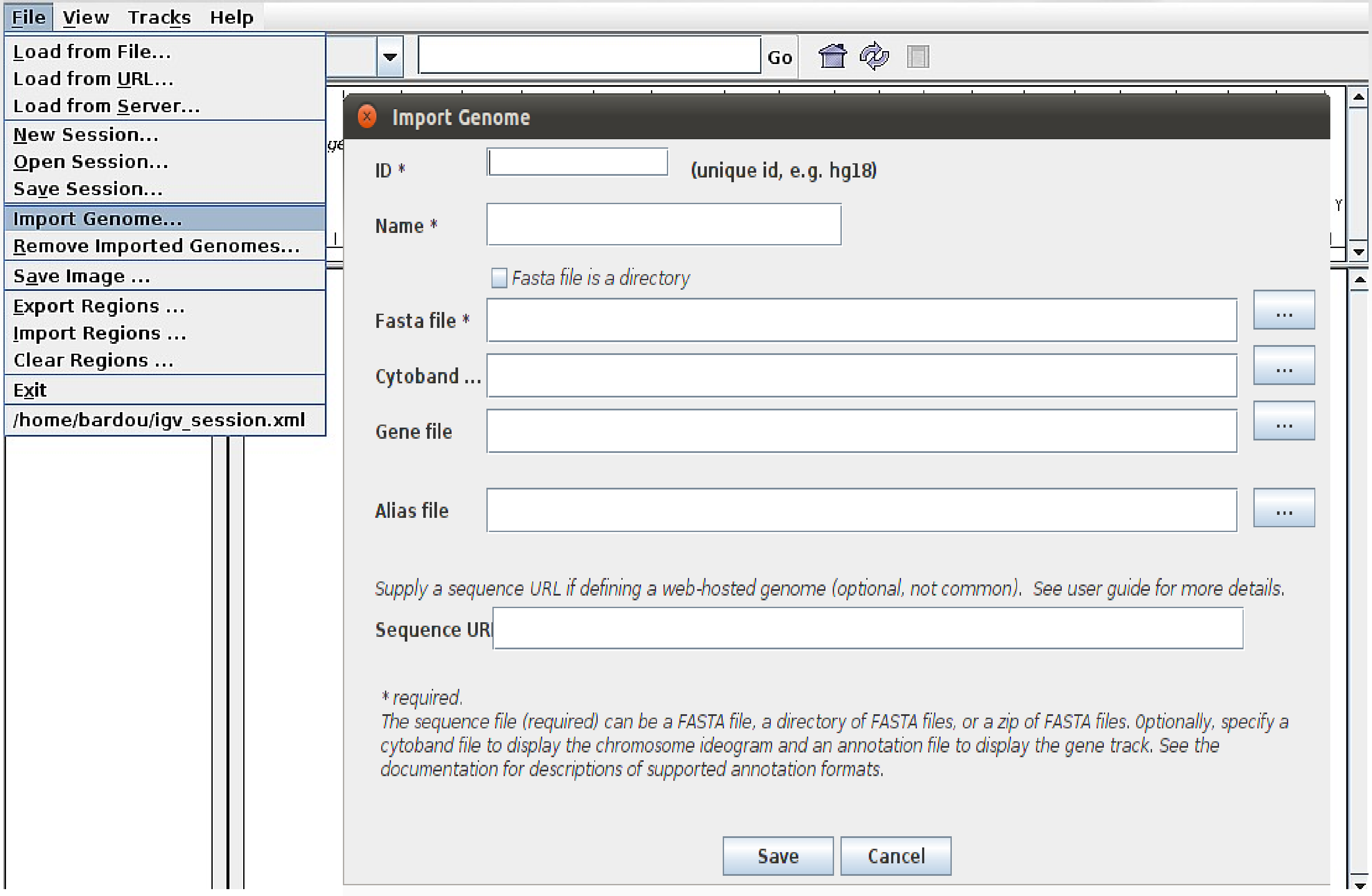
- File Extension Identifies Format
- Recommended File Formats
- BAM
- BED
- BedGraph
- bigBed
- bigWig
- Birdsuite Files
- CBS
- CN
- Cufflinks Files
- Custom File Formats
- Cytoband
- FASTA
- GCT
- genePred
- GFF
- GISTIC
- Goby
- GWAS
- IGV
- LOH
- MAF
- Merged BAM File (.bam.list)
- MUT
- PSL
- RES
- SAM
- Sample Information
- SEG
- SNP
- TAB
- TDF
- Track Line
- Type Line
- VCF
- WIG

Visualizing alignments on IGV



Visualizing alignments on IGV

Import a reference genome



ID * (unique id, e.g. hg18)

Name *

□ Fasta file is a directory

Fasta file *

Cytoband ...

Gene file

Alias file

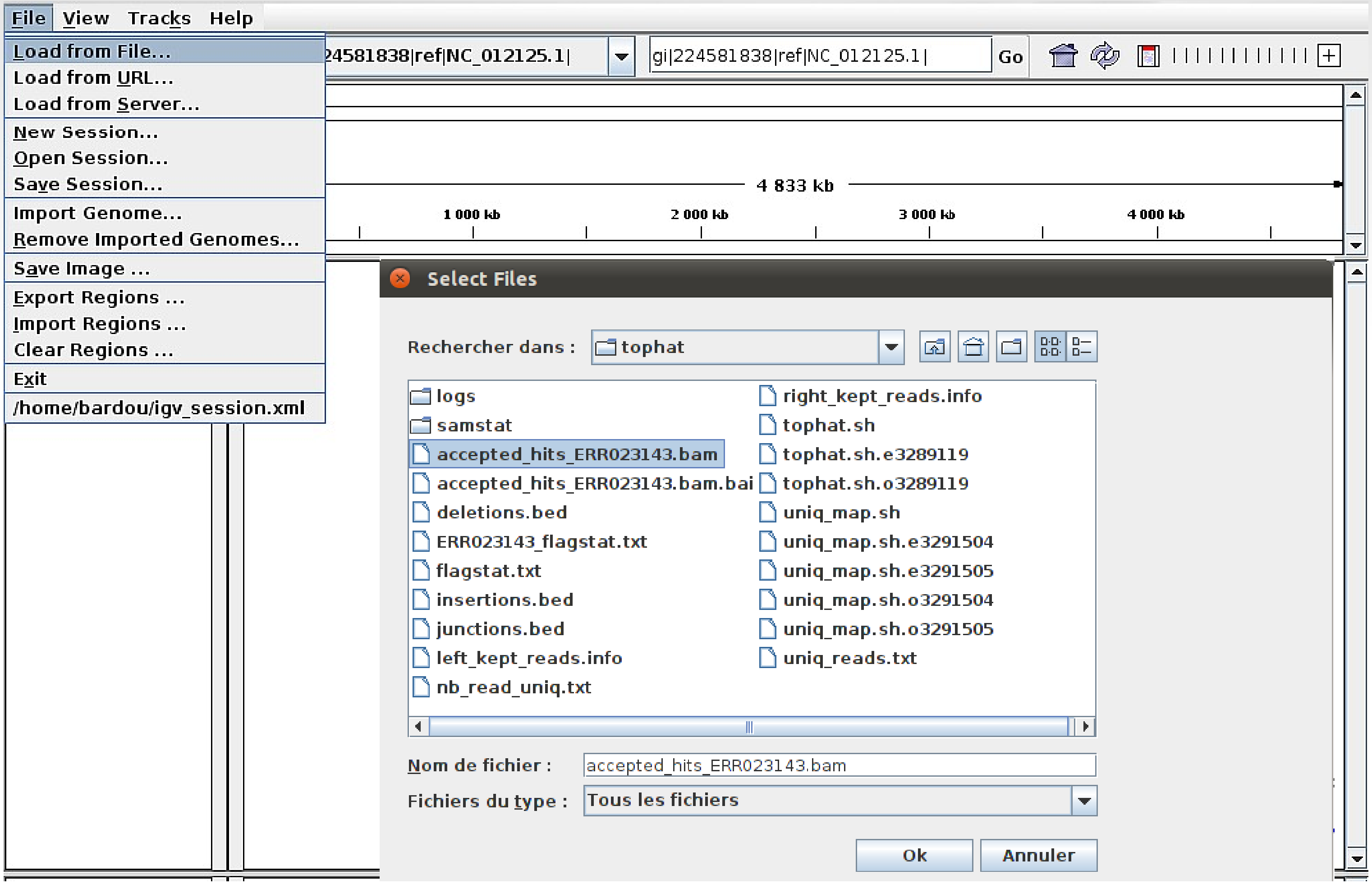
Supply a sequence URL if defining a web-hosted genome (optional, not common). See user guide for more details.

Sequence URl

** required.*
The sequence file (required) can be a FASTA file, a directory of FASTA files, or a zip of FASTA files. Optionally, specify a cytoband file to display the chromosome ideogram and an annotation file to display the gene track. See the documentation for descriptions of supported annotation formats.

Visualizing alignments on IGV

Import your BAM Files



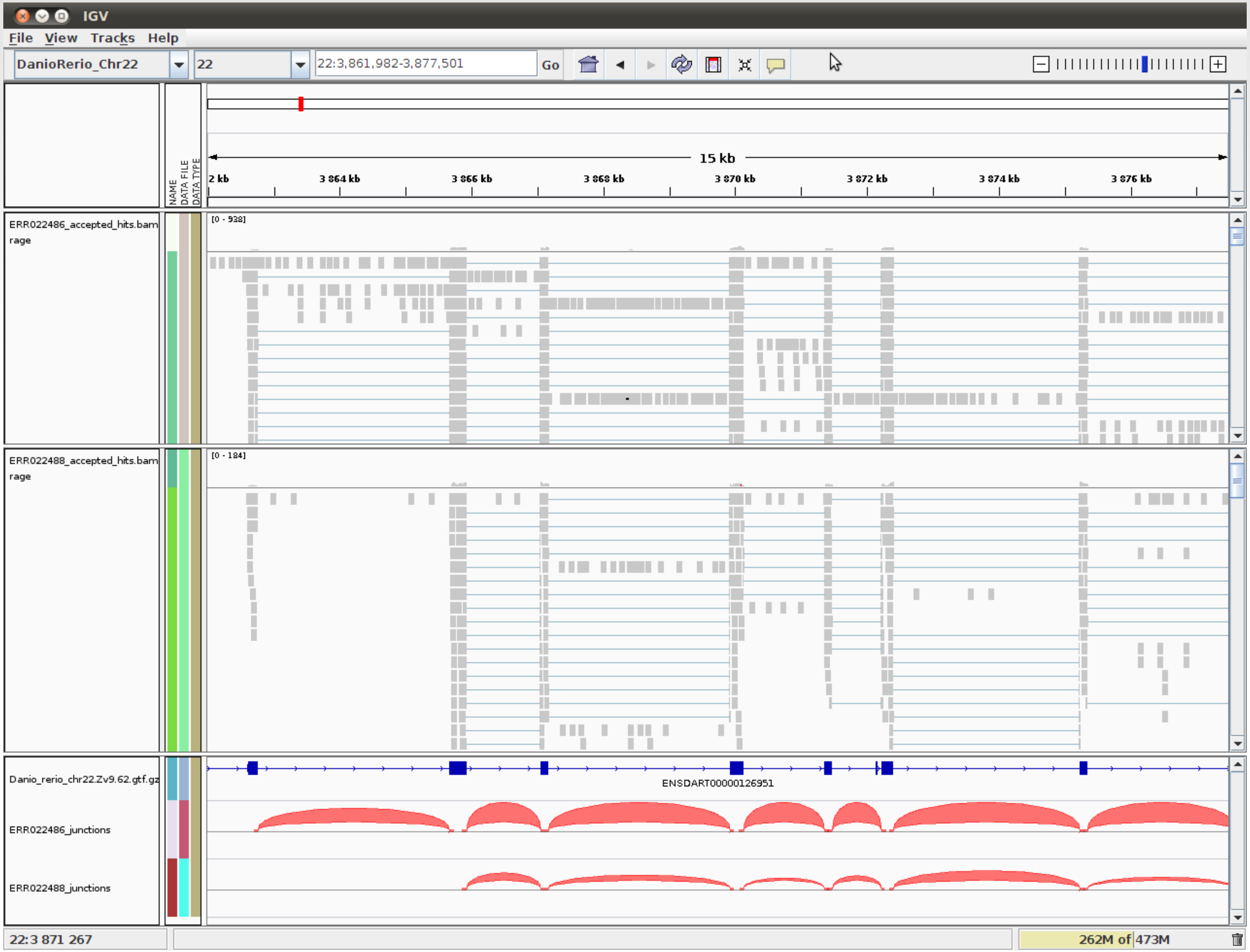
Rechercher dans : tophat

Nom de fichier : accepted_hits_ERR023143.bam

Fichiers du type : Tous les fichiers

Visualizing alignments on IGV

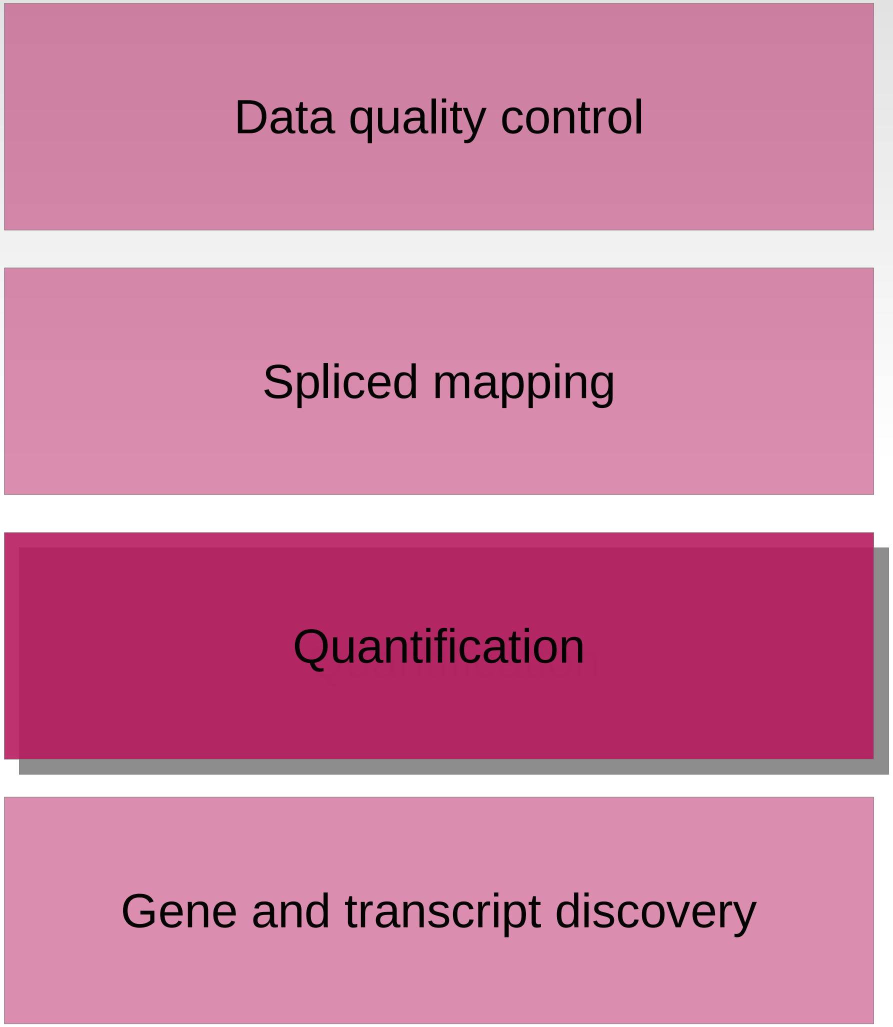
- Exemple of bam and bed files visualisation



hands-on : IGV

- *Create the genome dr22 in IGV using Danio_rerio.Zv9.62.dna.chromosome.22.fa*
- *Load the gtf file : Danio_rerio_chr22.Zv9.62.gtf*
- *Load the bam file : ERR022486/accepted_hits.bam*

Analyse workflow



Data quality control

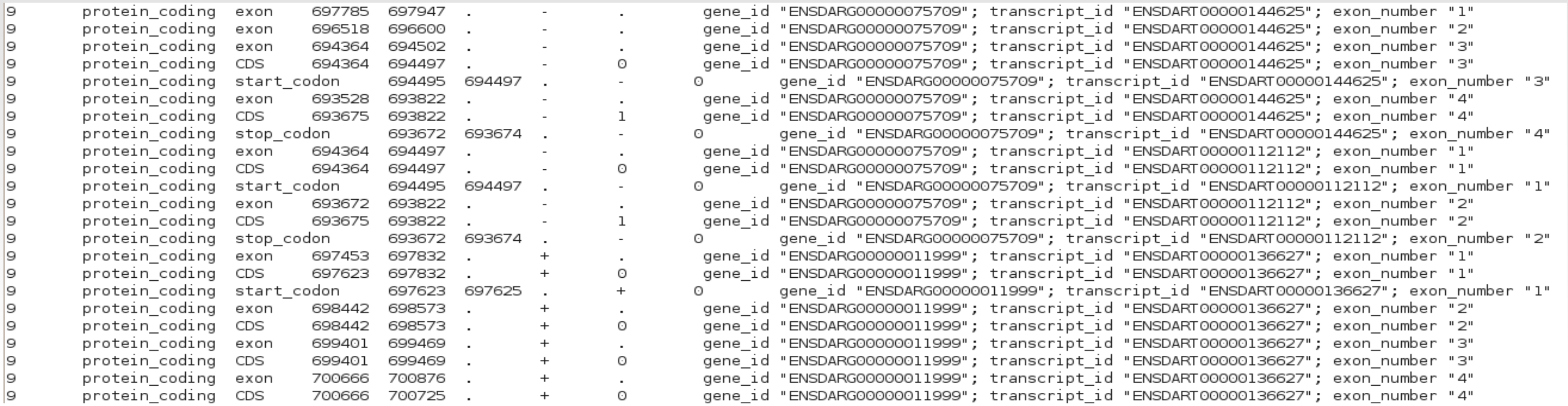
Spliced mapping

Quantification

Gene and transcript discovery

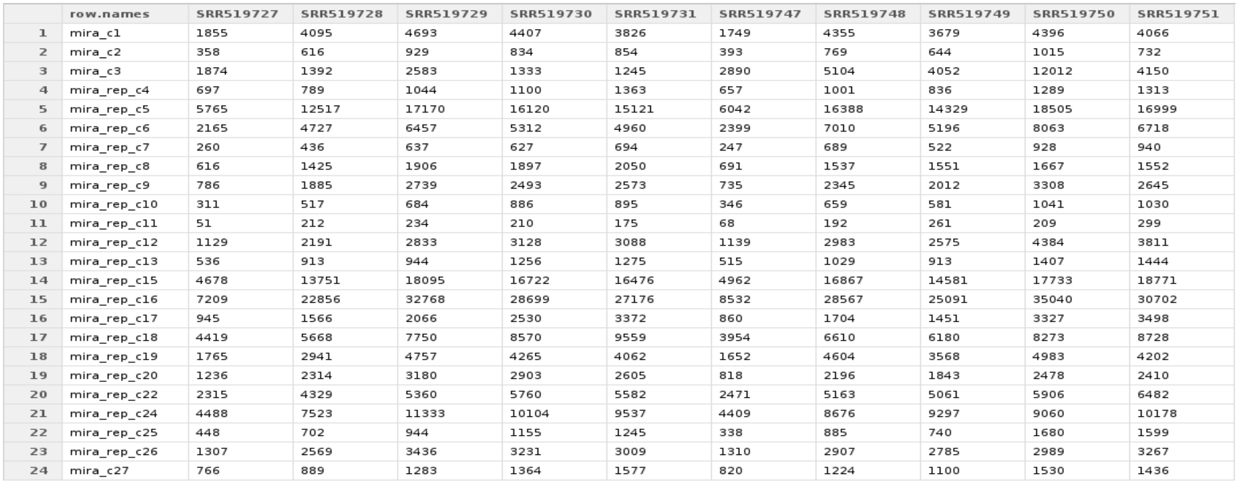
What do we want to build?

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

Tools for each approach

- htseq-count
- Cufflinks or FeatureCounts

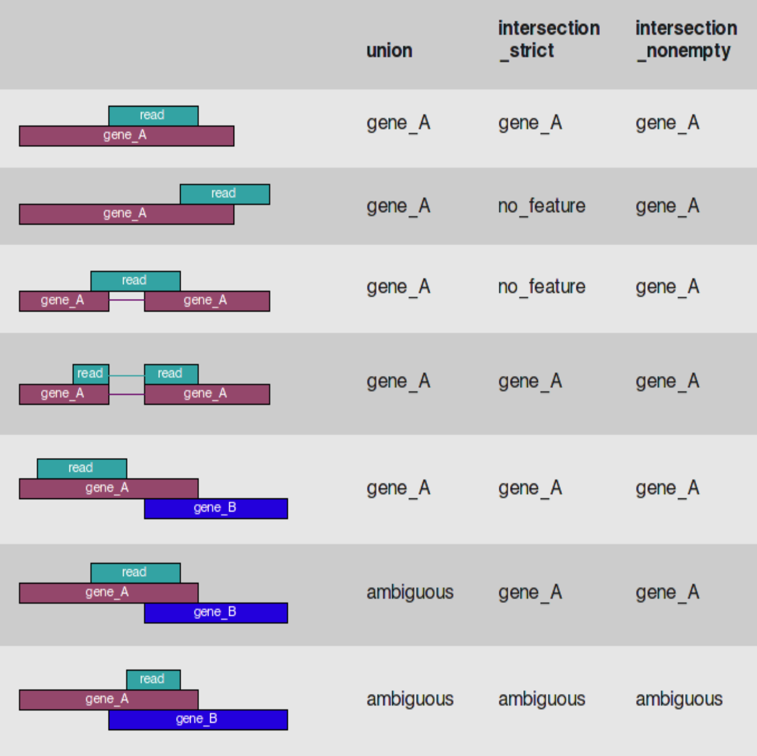
HTSeq-count

http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html

- Process the output from short read aligners in various formats
- Count how many reads map to each feature (in RNA-Seq, the features are typically genes)
  - counting reads by genes
  - or consider each exon as a feature to check for alternative splicing
- Inputs:
  - file with aligned sequencing reads: bam (or sam) file
  - list of genomic feature: gtf file

HTSeq-count parameters

- Command line :
  - *htseq-count [options] <sam_file> <gtf_file>*
  - *samtools view accepted_hits.bam | htseq-count --stranded=no -m intersection-nonempty - file.gtf -q > output.htseq-count.txt &*



	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	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

read
gene_A
gene_B

Some options:

-m <mode> : intersection-strict or intersection-nonempty (default union)

--stranded =<yes, no, or reverse> (default yes)
-t <feature type> : 3rd column in GTF file
-q : quiet
-h : help

HTSeq-count output

- Output: a table with counts for each feature and a summary of reads not counted for any feature:
    - *no_feature*: reads which couldn't be assigned to any feature
    - *ambiguous*: reads which could have been assigned to more than one feature and hence were not counted for any of these
    - *not_aligned*: reads in the SAM file without alignment
    - *alignment_not_unique*: reads with more than one reported alignment. These reads are recognized from the NH optional SAM field tag. (If the aligner does not set this field, multiply aligned reads will be counted multiple times.)

Quantification with 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

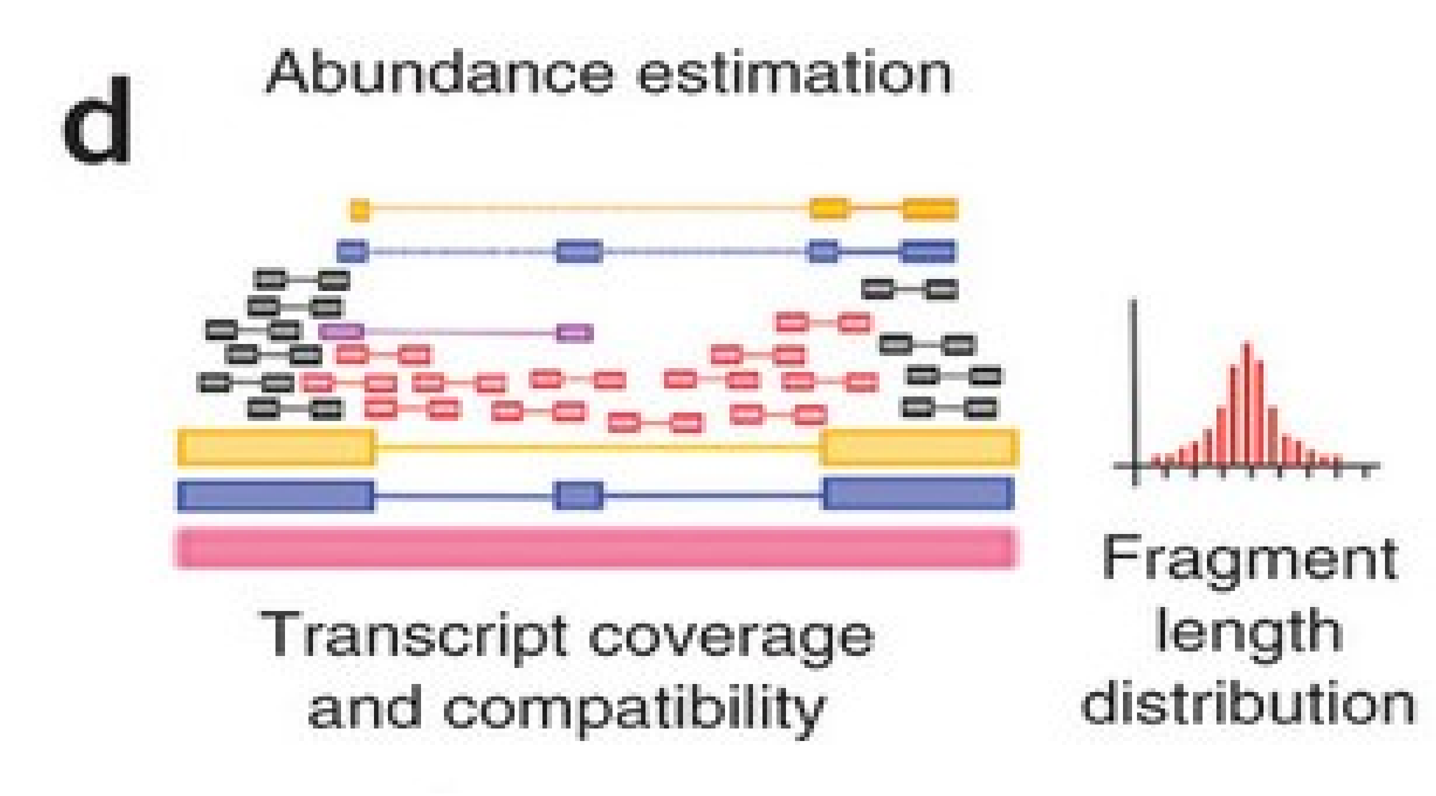
Nature Biotechnology **28**, 511–515 (2010) | doi:10.1038/nbt.1621
Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

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 read attribution

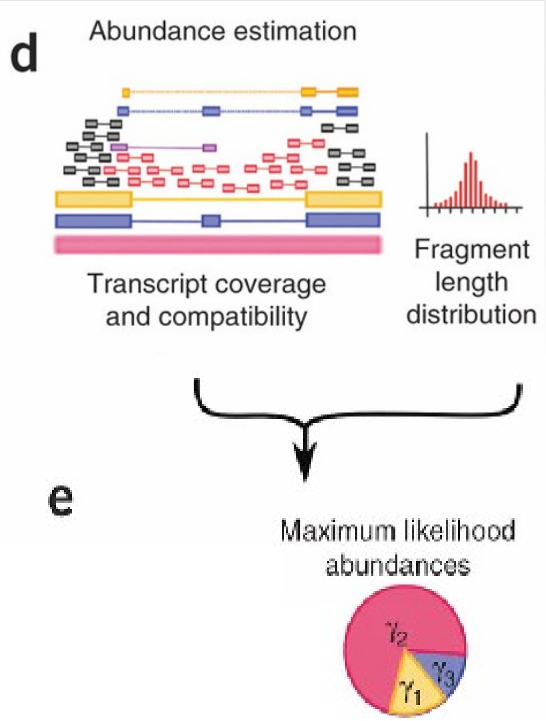
- Violet fragment: from which transcript?
  - Use of Fragment length distribution


d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution

Cufflinks expression measurement

- Fragments attribution

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


d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
γ_2
γ_3
γ_1

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
  - 1kb transcript with 1000 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped) will have:

    RPKM = 1000/(1 * 8) = 125

- 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

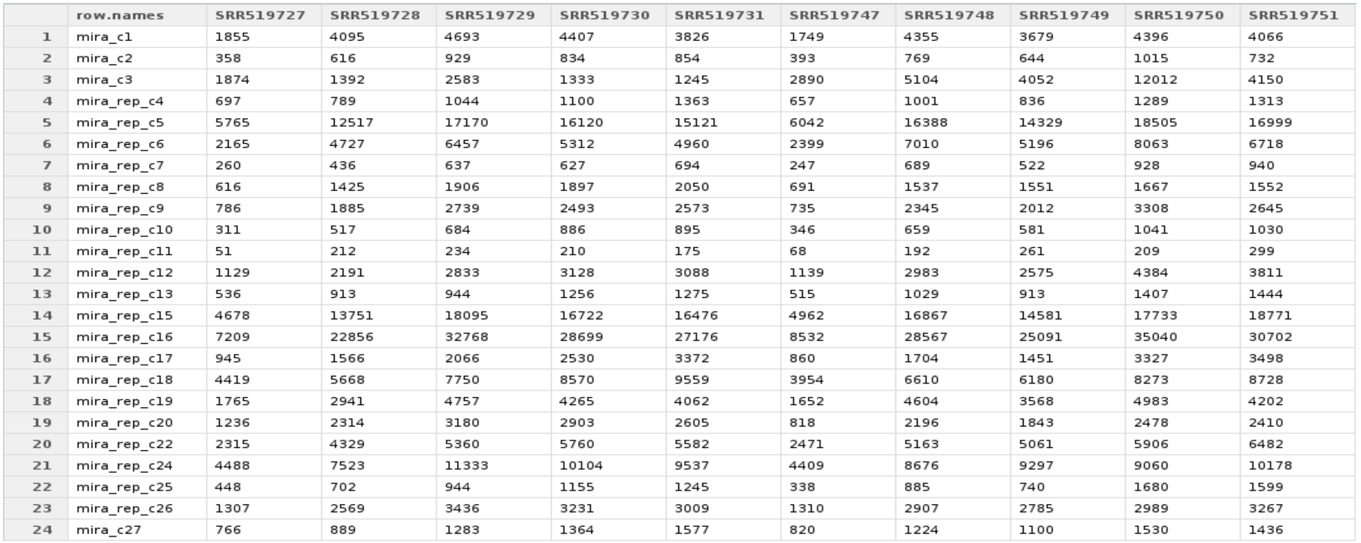
  -o/--output-dir

  -p/--num-threads

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

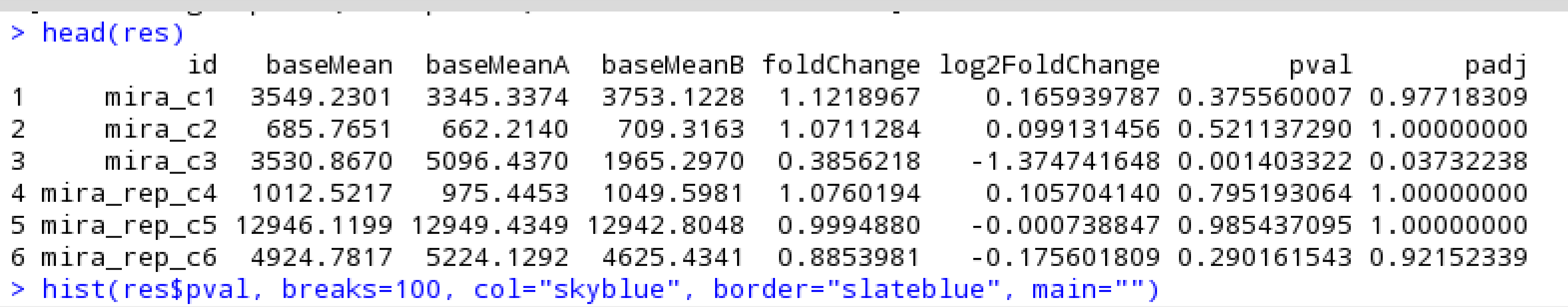
Merging individual count files

- Each quantification is produced from a bam file corresponding to a sample
- The quantification column has to be extracted
- The columns are the joined (paste)
- A header is added to the count file

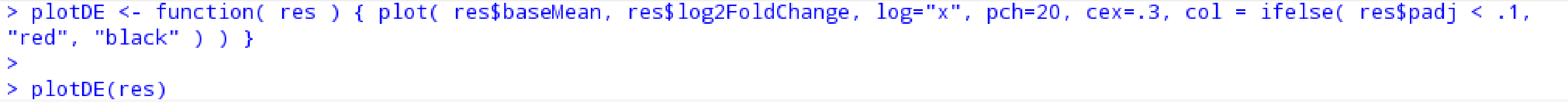


	row.names	SRR519727	SRR519728	SRR519729	SRR519730	SRR519731	SRR519747	SRR519748	SRR519749	SRR519750	SRR519751
1	mira_c1	1855	4095	4693	4407	3826	1749	4355	3679	4396	4066
2	mira_c2	358	616	929	834	854	393	769	644	1015	732
3	mira_c3	1874	1392	2583	1333	1245	2890	5104	4052	12012	4150
4	mira_rep_c4	697	789	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

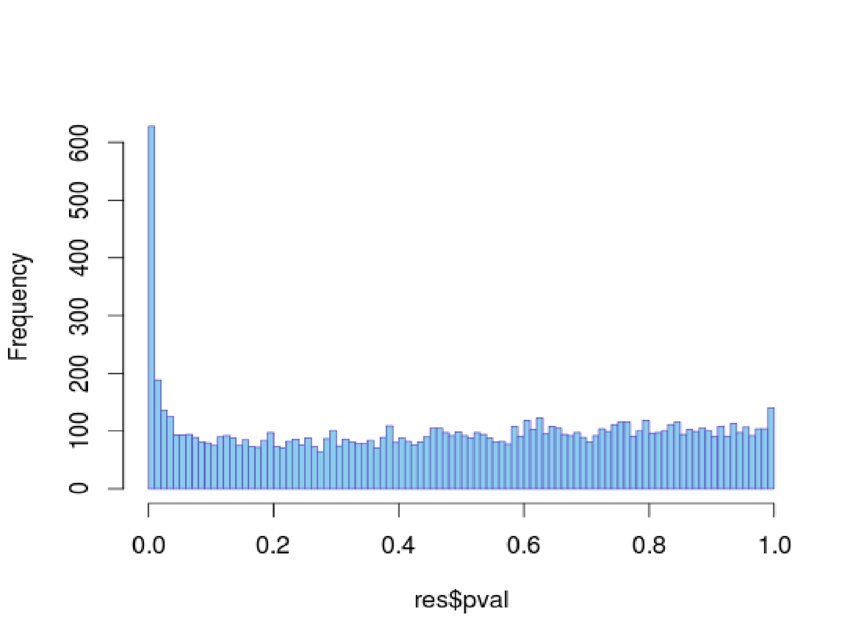
Statistical analysis in R (DESeq2 / edgeR)

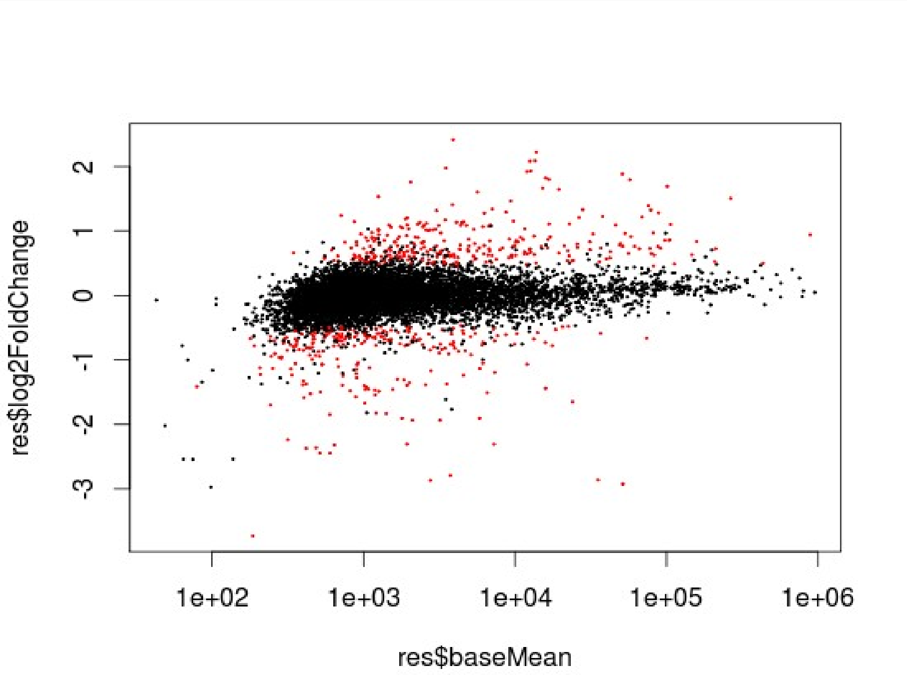


```
> head(res)
           id   baseMean  baseMeanA  baseMeanB foldChange log2FoldChange        pval       padj
1     mira_c1  3549.2301  3345.3374  3753.1228  1.1218967    0.165939787 0.375560007 0.97718309
2     mira_c2   685.7651   662.2140   709.3163  1.0711284    0.099131456 0.521137290 1.00000000
3     mira_c3  3530.8670  5096.4370  1965.2970  0.3856218   -1.374741648 0.001403322 0.03732238
4 mira_rep_c4  1012.5217   975.4453  1049.5981  1.0760194    0.105704140 0.795193064 1.00000000
5 mira_rep_c5 12946.1199 12949.4349 12942.8048  0.9994880   -0.000738847 0.985437095 1.00000000
6 mira_rep_c6  4924.7817  5224.1292  4625.4341  0.8853981   -0.175601809 0.290161543 0.92152339
> hist(res$pval, breaks=100, col="skyblue", border="slateblue", main="")
```



```
> plotDE <- function( res ) { plot( res$baseMean, res$log2FoldChange, log="x", pch=20, cex=.3, col = ifelse( res$padj < .1,
"red", "black" ) ) }
>
> plotDE(res)
```





Hands-on : quantification

1/ Quantify the genes of chromosome 22 using htseq-count and the Ensembl GTF file for both samples.

2/ Merge both files to produce the count tables. Add a header to the count table.

3/ create the count table dotplot

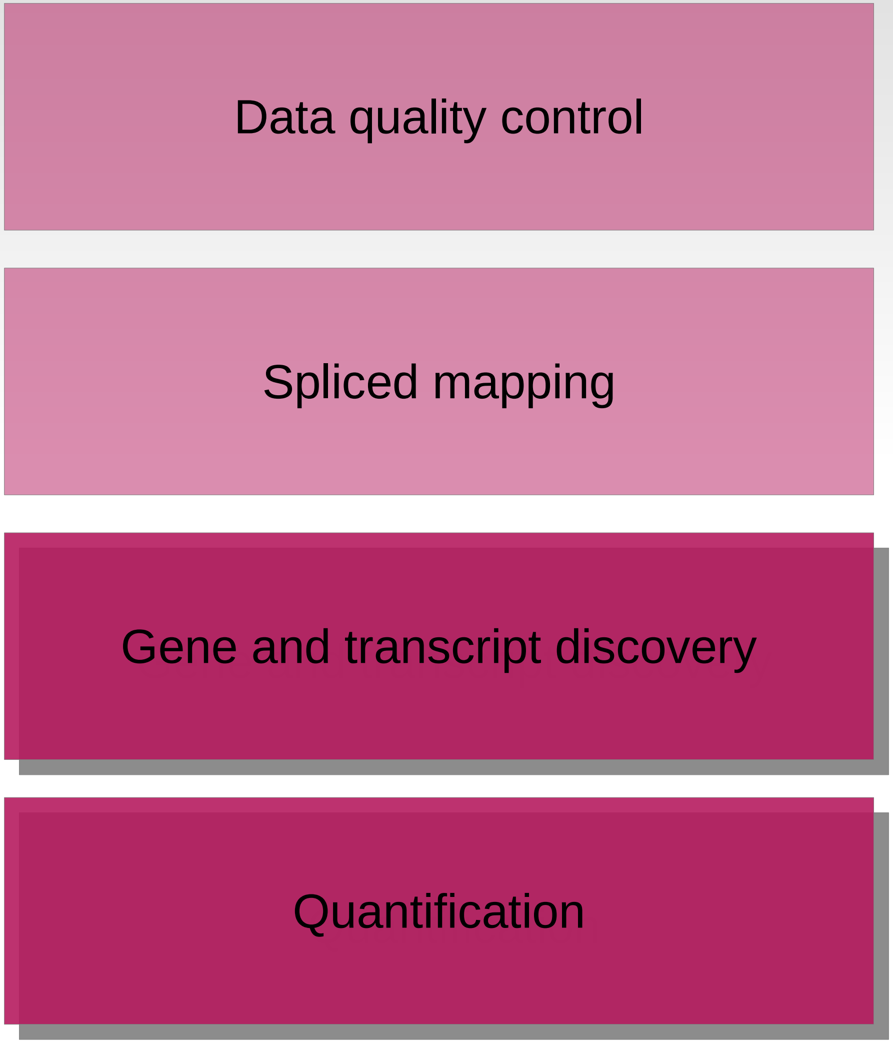
Hands-on : hints

```
samtools view ERR022486/accepted_hits.bam | htseq-count --stranded=no
-m intersection-nonempty - /work/.../Danio_rerio_chr22.Zv9.62.gtf -q >
ERR022486/accepted_hits.bam.htseq-count_nonempty_nostranded  &
```

The same for ERR022488

```
paste ERR022486/accepted_hits.bam.htseq-count_nonempty_nostranded
ERR022488/accepted_hits.bam.htseq-count_nonempty_nostranded | cut
-f1,2,4 > All.htseq-count
```

Analyse workflow



Data quality control

Spliced mapping

Gene and transcript discovery

Quantification

Transcript reconstruction

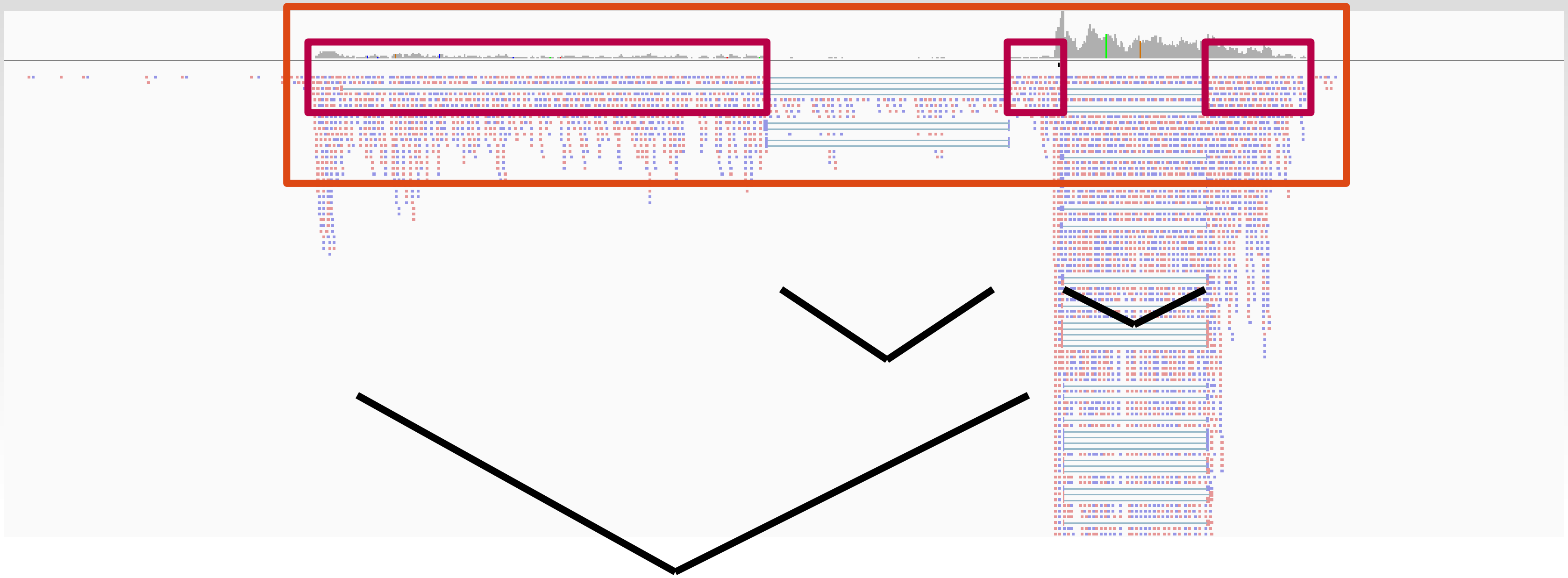
The different ways :

- Finding the gene locations
- Finding the exons
- Finding the junctions :
  - Between pairs junctions
  - Within sequences junction

Defining the model building strategy

- Number of built models
- Intronic reads

The elements of the model



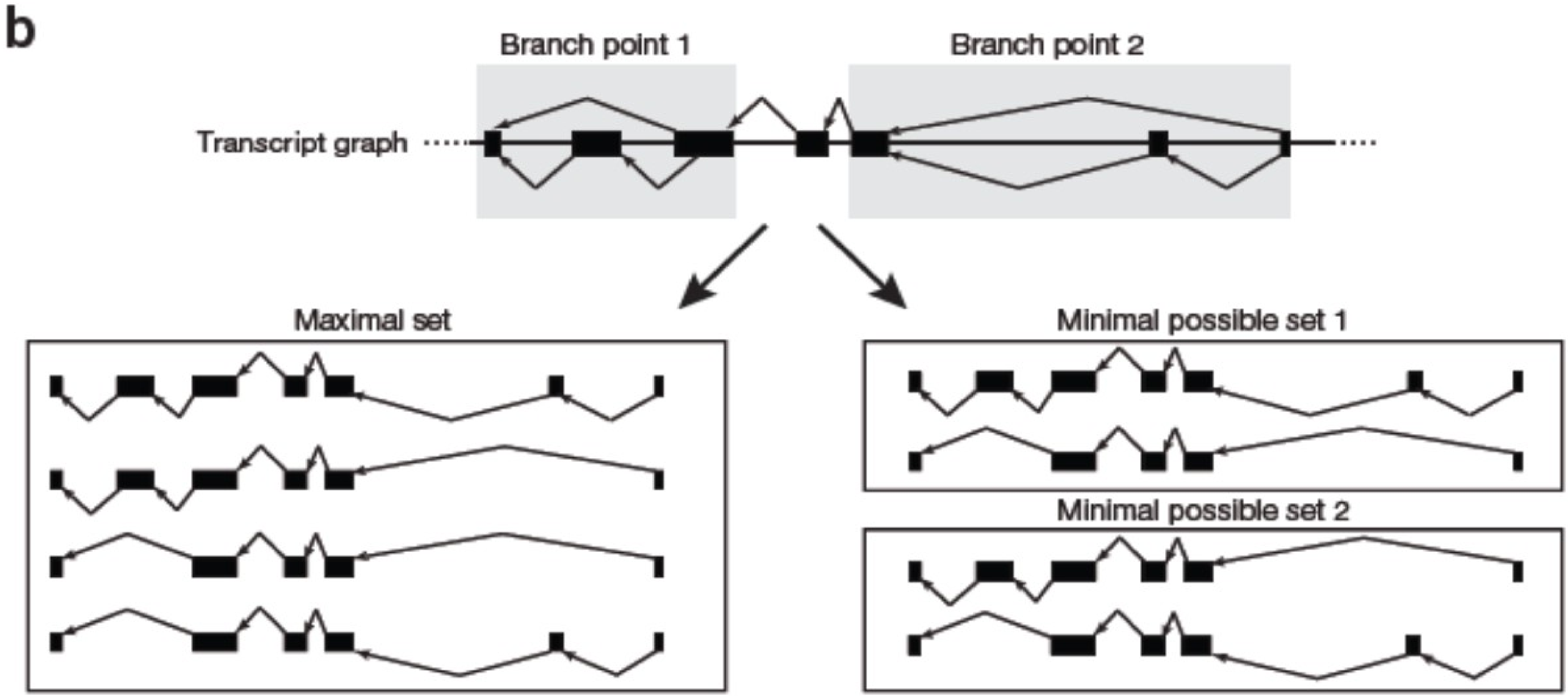
gene location

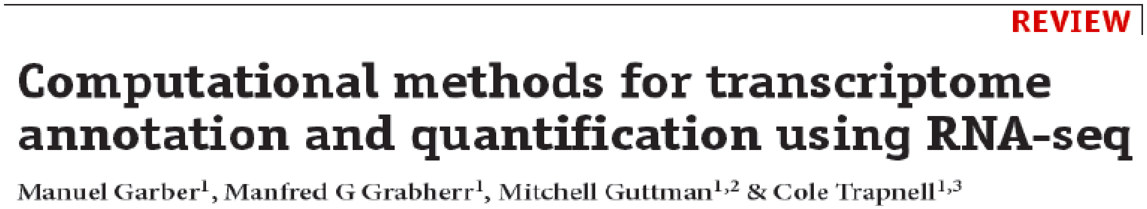
Exon location

Junctions :

- Between read pair junction
- Within read junction

Model building strategies



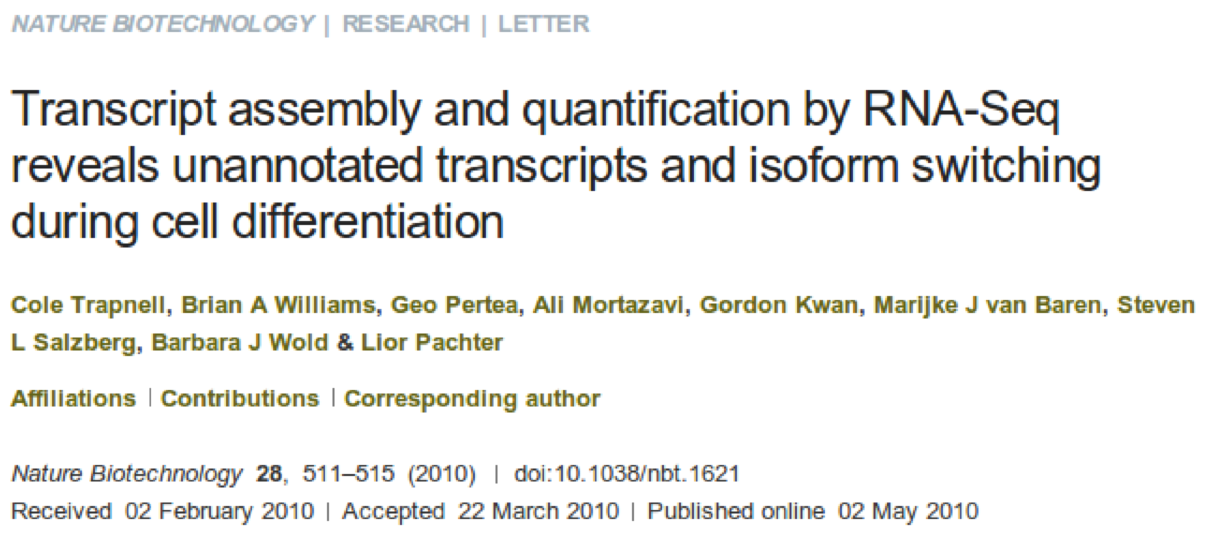


REVIEW

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

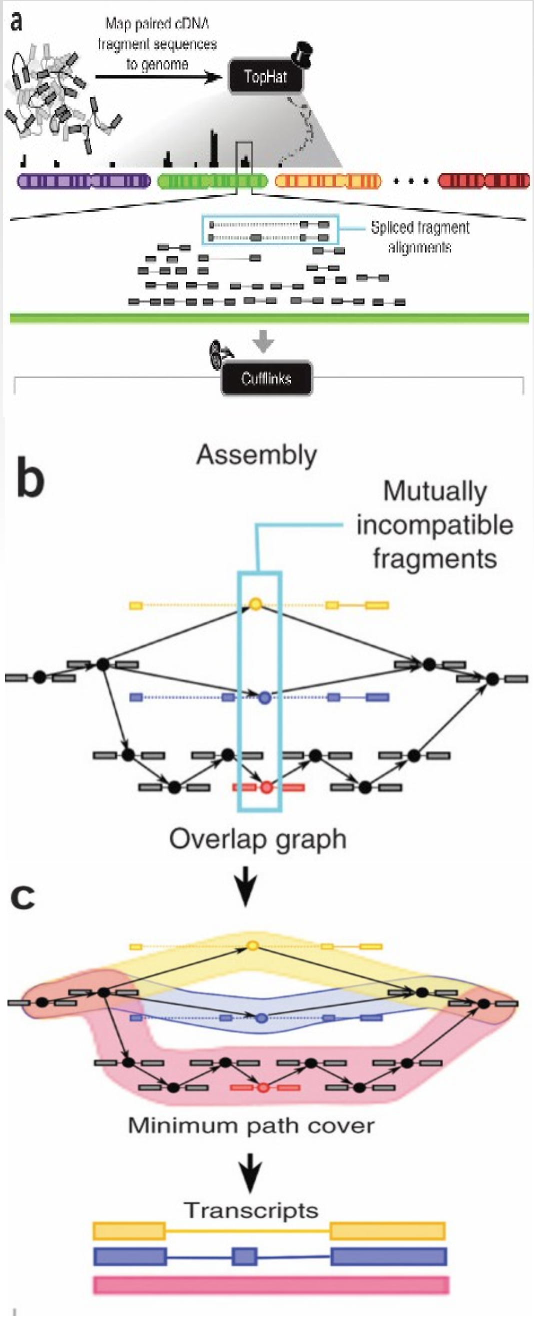
Nature Biotechnology **28**, 511–515 (2010) | doi:10.1038/nbt.1621
Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

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

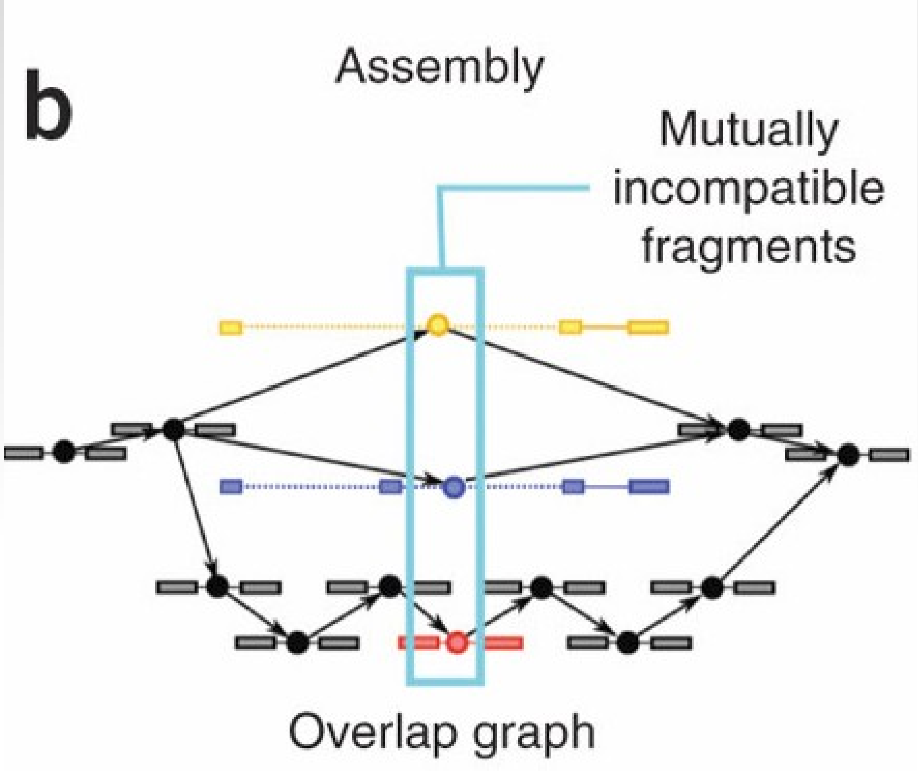
- Transcripts assembly :
  - Fragments are divided into non-overlapping loci
  - each locus is assembled independently :
- Cufflinks assembler
  - find the mini nb of transcripts that explain the reads
  - find a minimum path cover (Dilworth's theorem) :
    - nb incompatible read = mini nb of transcripts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

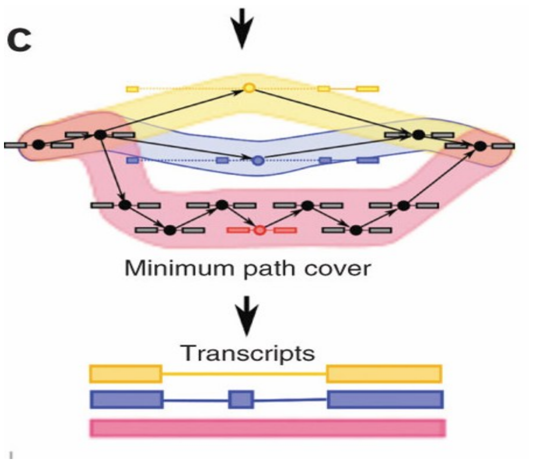

a
Map paired cDNA fragment sequences to genome
TopHat
Spliced fragment alignments
Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts

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

Cufflinks transcript assembly

- Transcripts assembly :
  - Identification incompatibles fragments: distinct isoforms
  - Compatibles fragments are connected: graph 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

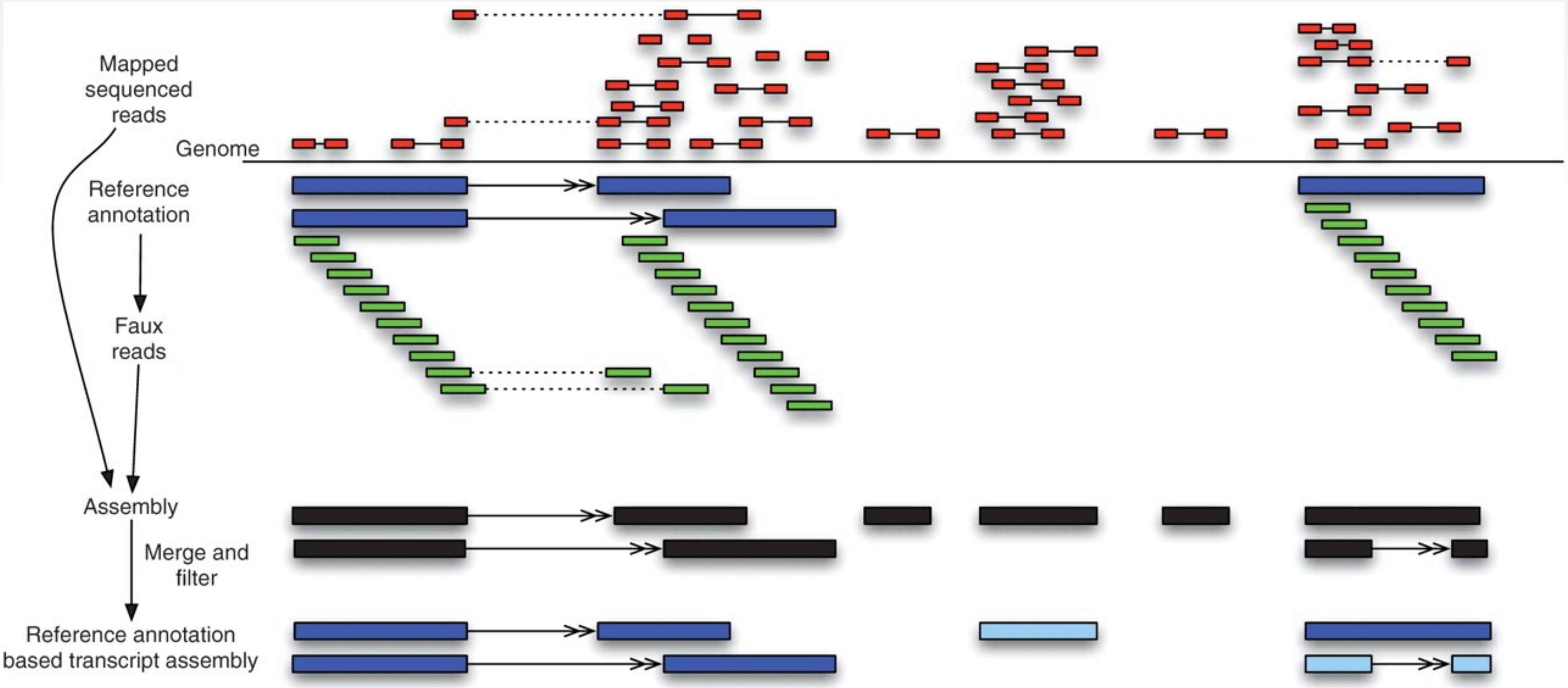
  -G/--GTF <reference_annotation.(gtf/gff)> : estimate isoform expression, no assembly novel transcripts

  *-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



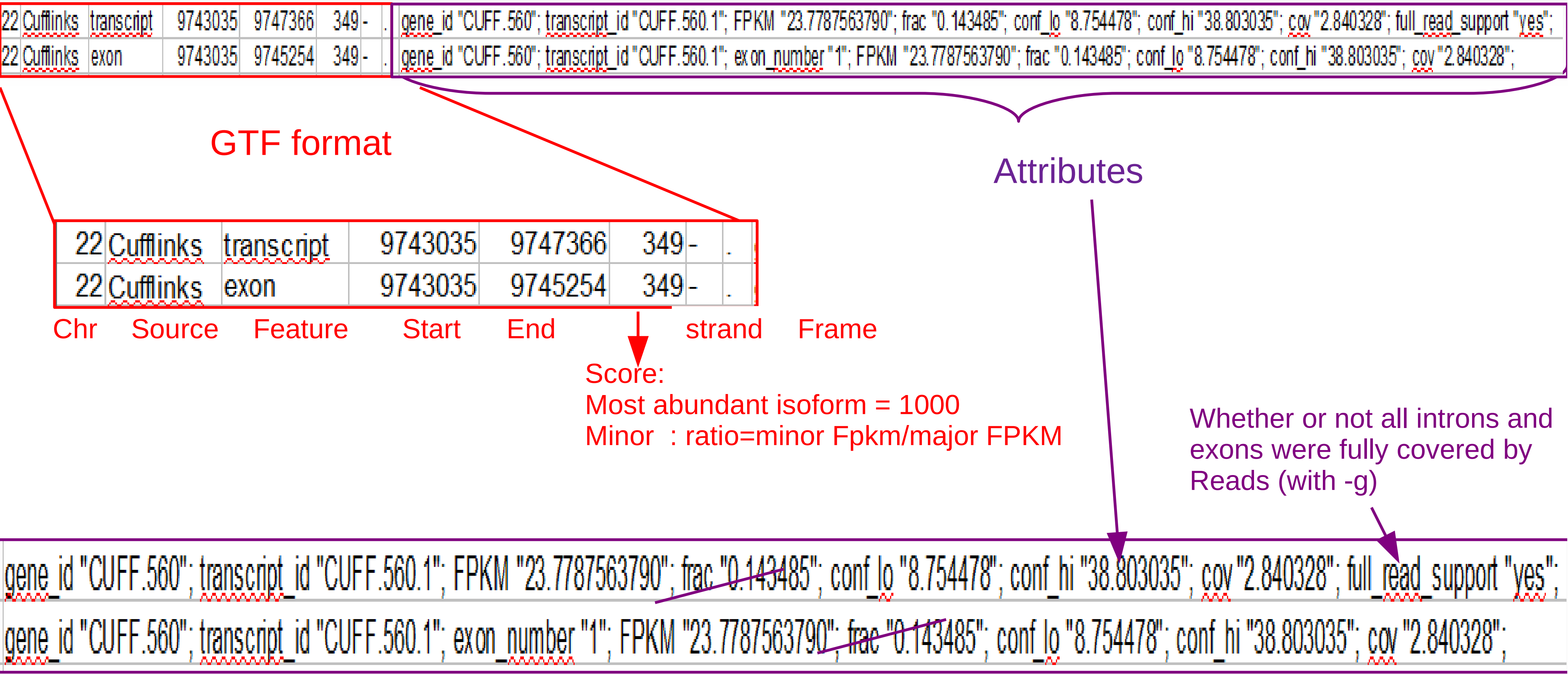
Roberts A et al. Bioinformatics 2011;27:2325-2329

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

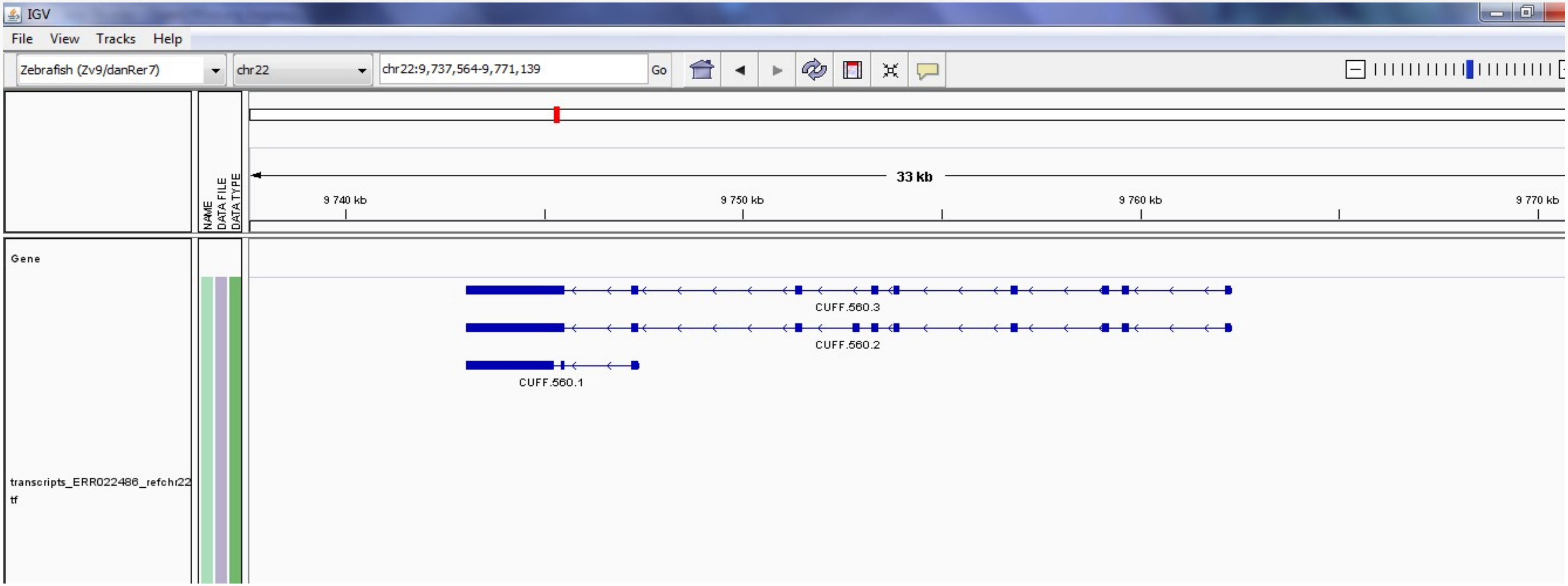
Whether or not all introns and exons were fully covered by Reads (with -g)

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";

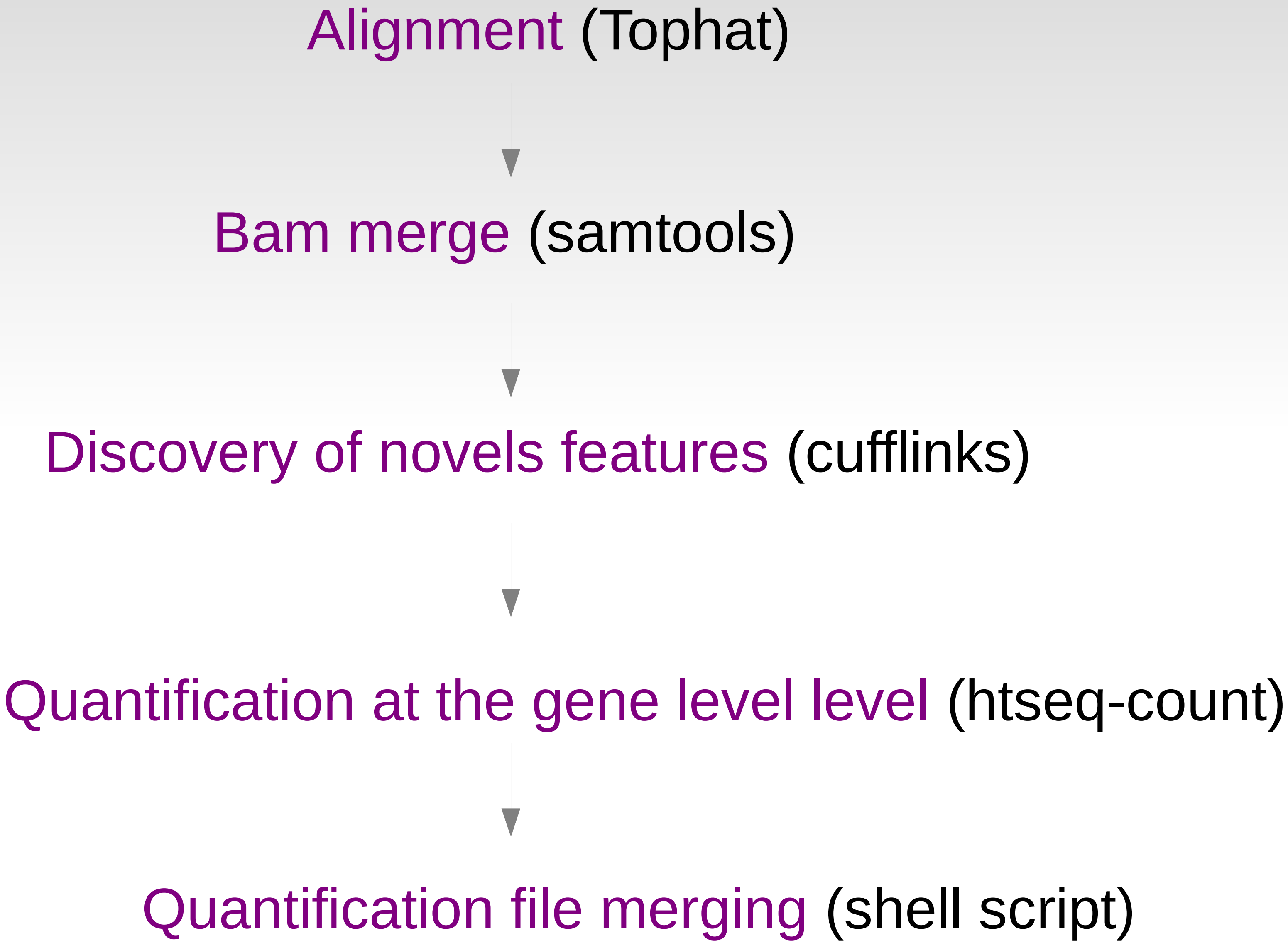
Cufflinks GTF description

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

  - IGV visualization



Gene discovery pipeline



Alignment (Tophat)

Bam merge (samtools)

Discovery of novels features (cufflinks)

Quantification at the gene level level (htseq-count)

Quantification file merging (shell script)

Quantification strategy

- First set your gene and transcript model = build a reference GTF file
- Then use option -G to quantify the same set of elements on all your samples with sigcufflinks
- Then sort your raw_transcript.tsv files
- cut the second or third column of the sorted file
- Paste all the column in the count file

Hands-on : cufflinks

- Merge all bam files using samtools merge.
- Run cufflinks to discover new genes and transcripts using the merged bam file

Hands-on : commands

- Merge all bam files :

```
samtools merge ALL.bam ERR022486/accepted_hits.bam
ERR022488/accepted_hits.bam
```

- Cufflinks command:

```
cufflinks  -- output-dir=CUFFLINKS -g Danio_rerio_chr22.Zv9.62.gtf ALL.bam
```

Conclusions

- RNASeq analysis are performed routinely.
- There are still some questions about the best possible aligner or gene seeker but some tools are now well established as good solutions.
- The number of replicates is particularly important if the expression difference is small between conditions.
- Pay attention to the correspondence between your library type and the program parameters you use.

Questions ?