Plateforme Bioinformatique Midi-Pyrénées

F12d
RNA-Seq data analysis

Delphine Labourdette Get-Biopuce / Christophe Klopp Bioinfo Genotoul

Session organisation

Morning (9h30 -12h30) :

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

Afternoon (14h-17h) :

- mRNA calling
  - Theory + exercises
- expression measurement
  - Theory + exercises

Material

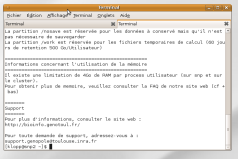
Slides
- pdf : one per page
- pdf : three per page with comment lines

Exercise leaflet

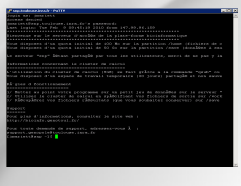
Data, results files and a unix command reminder
http://bioinfo.genotoul.fr/index.php?id=119

What you should know

How to connect to snp.toulouse.inra.fr?
How to use unix commands?

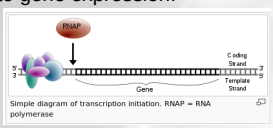


wget URL
qlogin -pe parrallel_smp 4

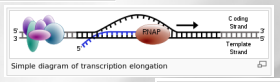


Transcription

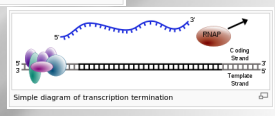
Transcription is the process of creating a complementary RNA copy of a sequence of DNA. Transcription is the first step leading to gene expression.



Simple diagram of transcription initiation. RNAP = RNA polymerase



Simple diagram of transcription elongation



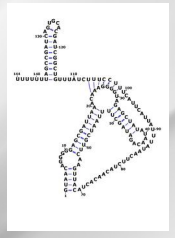
Simple diagram of transcription termination

http://en.wikipedia.org/wiki/Transcription_(genetics)

Transcription products

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

- tRNA,
- rRNA,
- snoRNAs,
- microRNAs,
- siRNAs,
- piRNAs....

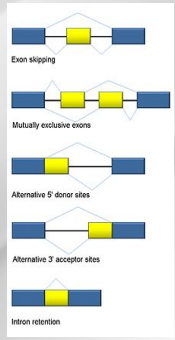


http://en.wikipedia.org/wiki/User:Amarchais/RsaOG_RNA

Alternative splicing

Alternative splicing (or differential splicing) is a process by which the exons of the RNA produced by transcription of a gene (a primary gene transcript or pre-mRNA) are reconnected in multiple ways during RNA splicing. The resulting different mRNAs may be translated into different protein isoforms; thus, a single gene may code for multiple proteins.

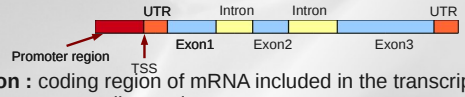
Post-transcriptional modification is a process in cell biology by which, in eukaryotic cells, primary transcript RNA is converted into mature RNA. A notable example is the conversion of precursor messenger RNA into mature messenger RNA (mRNA), which includes splicing and occurs prior to protein synthesis.



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

Vocabulary

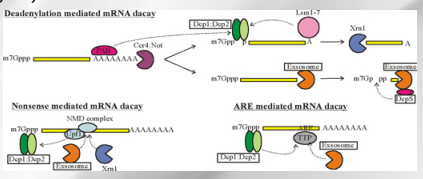
Gene : functional units of DNA that contain the instructions for generating a functional product.



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

Transcript degradation

After export to the cytoplasm, mRNA is protected from degradation by a 5' cap structure and a 3' poly adenine tail. In the deadenylation dependent mRNA decay pathway, the polyA tail is gradually shortened by exonucleases. This ultimately attracts the degradation machinery that rapidly degrades the mRNA in both in the 5' to 3' direction and in the 3' to 5' direction. Additional mechanisms, including the nonsense mediated decay pathway, bypass the need for deadenylation and can remove the mRNA from the transcriptional pool independently. Interestingly, the same enzymes are responsible for the actual degradation of the mRNA independent of the pathway taken (see figure).

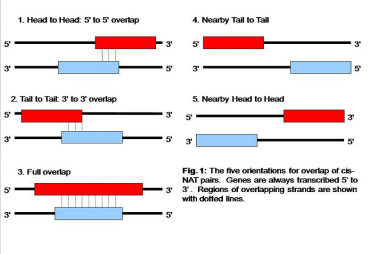


http://www.eb.tuebingen.mpg.de/research-groups/remco-sprangers

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.

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



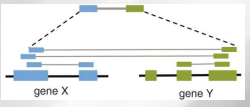
Fusion genes

- 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.
- 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.

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.

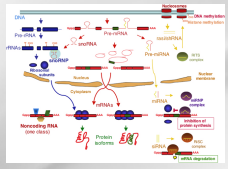


http://en.wikipedia.org/wiki/Fusion_gene

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

Transcriptome variability

- Number of transcripts
  - possible variation factor between transcripts: 10^6 or more,
  - expression variation between samples.
- 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



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

How can we study the transcriptome?

Lots of techniques

- EST (Expressed sequence tags)
- PCR (polymerase chain reaction)
- SAGE (serial analysis of gene expression)
- Micro-Arrays
  - Different types: spotting, synthesis
  - Different densities : few thousands up 4M probes / slide

Techniques classification ?

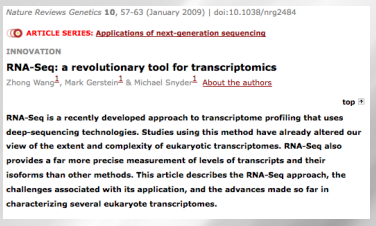
EST	PCR/RT-QPCR	SAGE	MicroArrays
No quantification quantification	quantification		indirect
low throughput	low throughput (up to hundreds)	High (up to thousands)	High throughput (up to millions)
Discovery (Yes)	no		no (except tiling)

→ Need transcript sequence partially known

→ Difficulties in discovering novels splice events

What is RNA-Seq ?

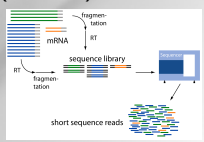
- use of **high-throughput sequencing technologies** to sequence cDNA in order to get information about a sample's RNA content
- Thanks to the deep coverage and base level resolution provided by next-generation sequencing instruments, RNA-seq provides researchers with efficient ways to measure transcriptome data experimentally



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, ...

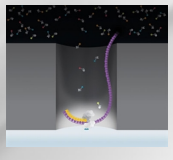


RNA-Seq platforms comparison

Platform	**454 Roche Titanium**	**HiSeq2000 Illumina**	**Solid 3+ Life Technologies**
Caracteristics	-Titanium chemistry -Pyrosequencing -PCR amplification	- Polymerase-based sequence-by-synthesis -PCR amplification -Multiplexing	-ligation-base-sequencing -PCR amplification
Applications	-De novo sequencing -Small genomes -Transcriptome	-Resequencing -Transcriptome -Epigenomic -Small RNA -Allele specific sequencing	-De novo sequencing -Resequencing -Transcriptome -Epigenomic -Small RNA
Paired end separation	Not used	200bp	200bp
Mb / run	800Mb	600Gb	60Gb
Read length	800 bp	100bp	50bp
Known Biases	- Long homopolymer - makes signal saturation - read duplication	- Rich GC or AT regions: under-representation during amplification - Most error in end of cycle	- read duplication ?

Third Generation RNA-Seq

- No more amplification
- Single Molecule Sequencing Technology (tSMS)
- Single Molecule Real Time (SMRT) sequencing technology (PacBio RS)



http://www.genengnews.com/gen-articles/third-generation-sequencing-debuts/3257/

What are we looking for?

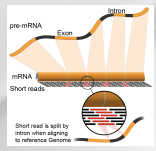
Identify genes
- List new genes

Identify transcripts
- List new alternative splice forms

Quantify these elements → differential expression

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



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 ?

ENCODE answers

- 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 $\mathbf{R^2}$ (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 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

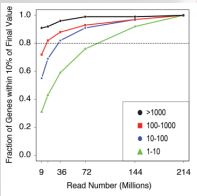
Producing a full transcriptome

RNA-sequence analysis of human B-cells

Jonathan M. Toung, Michael Morley, Mingyao Li, and Vivian G. Cheung

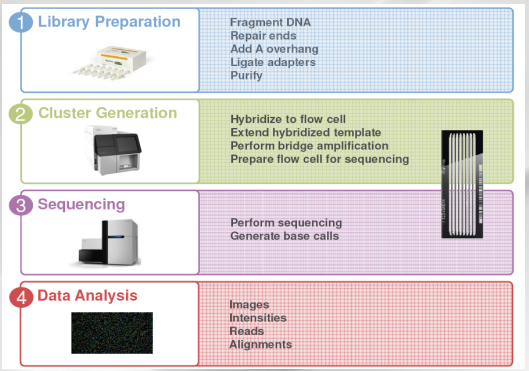
Genome Res. June 2011 21: 991-998; Published in Advance May 2, 2011, doi:10.1101/gr.116335.110

RNA-sequencing (RNA-seq) allows quantitative measurement of expression levels of genes and their transcripts. In this study, we sequenced complementary DNA fragments of cultured human B-cells and obtained 879 million 50-bp reads comprising 44 Gb of sequence. The results allowed us to study the gene expression profile of B-cells and to determine experimental parameters for sequencing-based expression studies. We identified 20,766 genes and 67,453 of their alternatively spliced transcripts. More than 90% of the genes with multiple exons are alternatively spliced; for most genes, one isoform is predominantly expressed. We found that while chromosomes differ in gene density, the percentage of transcribed genes in each chromosome is less variable. In addition, genes involved in related biological processes are expressed at more similar levels than genes with different functions. Besides characterizing gene expression, we also used the data to investigate the effect of sequencing depth on gene expression measurements. While 100 million reads are sufficient to detect most expressed genes and transcripts, about 500 million reads are needed to measure accurately their expression levels. We provide examples in which deep sequencing is needed to determine the relative abundance of genes and their isoforms. With data from 20 individuals and about 40 million sequence reads per sample, we uncovered only 21 alternatively spliced, multi-exon genes that are not in databases; this result suggests that at this sequence coverage, we can detect most of the known genes. Results from this project are available on the UCSC Genome Browser to allow readers to study the expression and structure of genes in human B-cells.

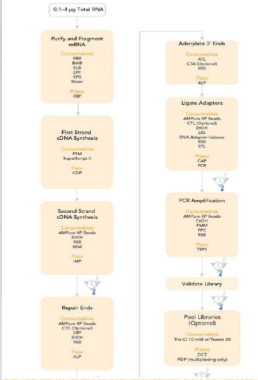


http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046

Illumina RNA-Seq protocol



RNA-Seq library preparation

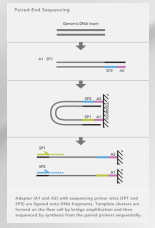


- Isolate poly-A containing mRNA
- capture mRNA with oligoT beads
- Randomly fragment RNA
- Random prime mRNA → cDNA
- Make 2nd strand cDNA
- Repair-Ends and 3' Ends Adenylate
- Ligate sequencing adapters
- Enrich up to 15 cycles of PCR
- gel purify
- validate library w/ Bioanalyzer

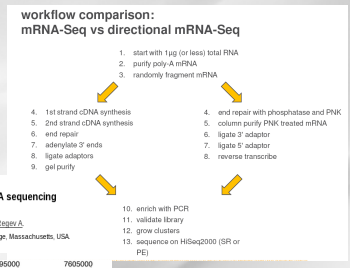
Library prep takes <2 days

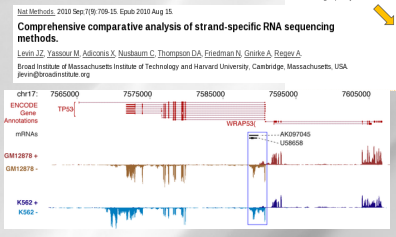
Paired-end sequencing

- modification to 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



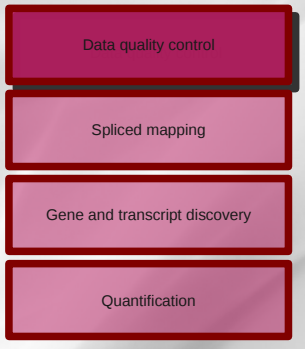
Strand specific RNA-Seq protocol





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

Analysis workflow



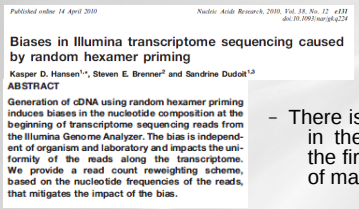
RNAseq specific bias

- Influence of the library preparation
- Random hexamer priming
- Coverage across the transcript may not be random
- Transcript length bias
- Positional bias and sequence specificity bias.

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

- BWA can force incorrect alignments to meet pairing criteria
- Some reads map to multiple locations

Hexamer random priming bias



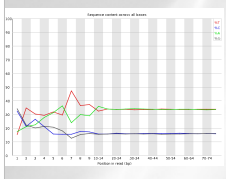
Published online 14 April 2010 Nucleic Acids Research, 2010, Vol. 38, No. 12 e131
doi:10.1093/nar/gkq224

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.



- There is a strong distinctive pattern in the nucleotide frequencies of the first 13 positions at the 5′-end of mapped RNA-Seq reads:
  - 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

Transcript length bias

Biol Direct. 2009 Apr 16;4:14.

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

Oshlack A, Wakefield MJ.

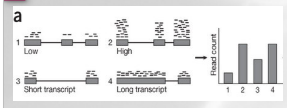
Abstract

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**

Gene expression

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

Liyan Gao[1,†], Zhide Fang[2,†], Kui Zhang[1], Degui Zhi[3] and Xiangqin Cui[1,4]

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]

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
```


Verifying RNA-Seq raw data

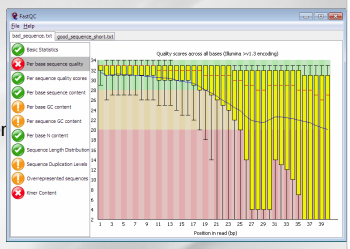
FastQC :

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*

Verifying RNASeq raw data

Per base sequence Q :



The higher the score the better the base call:

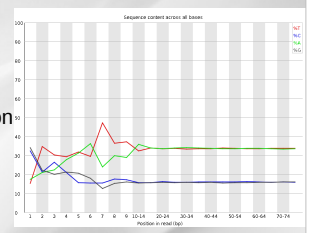
- very good quality calls(green)
- reasonable quality (orange)
- poor quality (red)

The quality of calls on most platforms will degrade as the run progresses

- The lower quartile for any base is less than 10 or the median for any base is less than 25
- The lower quartile for any base is less than 5 or the median for any base is less than 20

Hexamer random priming bias

Per base sequence content :

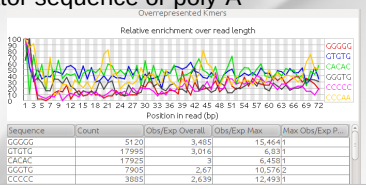


Proportion of each base at each position
In a random library : no differences, parallel lines

- difference between A and T, or G and C is greater than 10% in any position
- difference between A and T, or G and C is greater than 20% in any position

Kmer content

- A kmer is a subsequence of length k
- look for the enrichment of Kmer sequences. This should spot overrepresented sequences. Looking at 5-7mers should give a good impression of any contamination.
- Kmers showing a rise towards the end of the library indicate progressive contamination with adapters.
- Check for adaptator sequence or poly-A sequence



Sequence	Count	Obs/Exp Overall	Obs/Exp Max	Max Obs/Exp P...
GGGGG	5120	3,485	15,464	1
GTGTG	17995	3,016	6,83	1
CACAC	17925	3	6,458	1
GGGTG	7905	2,67	10,576	2
CCCCC	3885	2,639	12,493	1

Take home message on quality analysis

Elements to check:

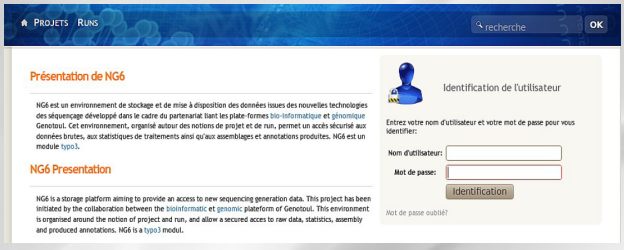
- Random priming effect
- K-mer (polyA, polyT)

Alignment on reference for the second quality check and filtering

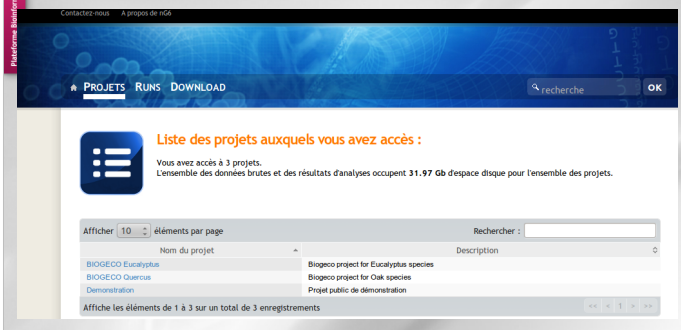
A good run?:

- Expected number of reads producted (2x500millions / flowwcell)
- Length of the reads expected (100pb)
- Quality
- Good alignment: very few unmapped reads, paires mapped in opposite sens

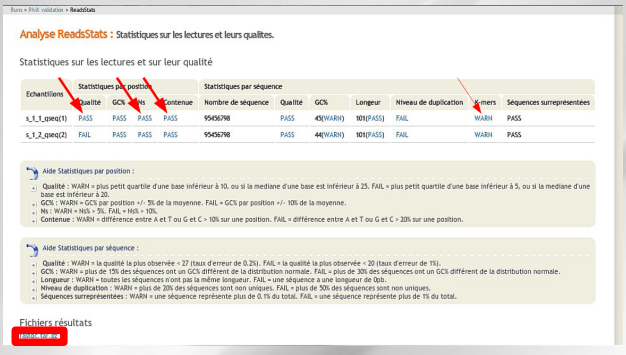
NG6 exemple



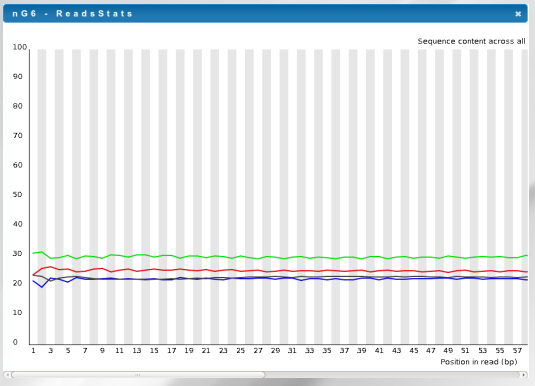
NG6 exemple



NG6 exemple



NG6 exemple



Exercises 1 : data quality

Connection snp:

ssh -X nom@snp

Pour se positionner sur un noeud du cluster:

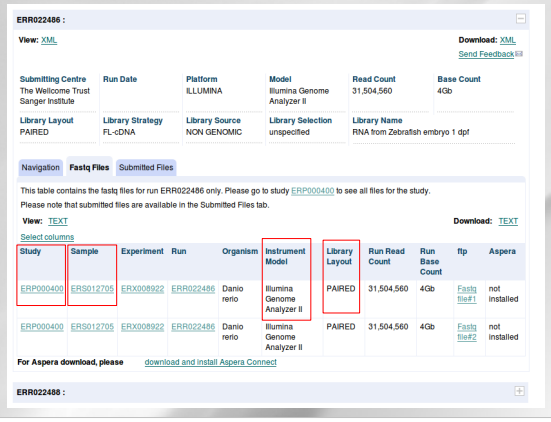
qlogin

Comptes formation: *anemone* *aster*
bleuet *iris*
muguet *narcisse*
pensee *rose*
tulipe *violette*

FastQC:

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

Exercises 1 : data quality


ERR022486 :

View: XML — Download: XML — Send Feedback

Submitting Centre	Run Date	Platform	Model	Read Count	Base Count
The Wellcome Trust Sanger Institute		ILLUMINA	Illumina Genome Analyzer II	31,504,560	4Gb

Library Layout	Library Strategy	Library Source	Library Selection	Library Name
PAIRED	FL-cDNA	NON GENOMIC	unspecified	RNA from Zebrafish embryo 1 dpf

Navigation | Fastq Files | Submitted Files

This table contains the fastq files for run ERR022486 only. Please go to study ERP000400 to see all files for the study.
Please note that submitted files are available in the Submitted Files tab.

View: TEXT — Download: TEXT

Select columns

Study	Sample	Experiment	Run	Organism	Instrument Model	Library Layout	Run Read Count	Run Base Count	ftp	Aspera
ERP000400	ERS012705	ERX008922	ERR022486	Danio rerio	Illumina Genome Analyzer II	PAIRED	31,504,560	4Gb	Fastq file#1	not installed
ERP000400	ERS012705	ERX008922	ERR022486	Danio rerio	Illumina Genome Analyzer II	PAIRED	31,504,560	4Gb	Fastq file#2	not installed

For Aspera download, please download and install Aspera Connect

ERR022486 :

Exercises 1 : data quality

Connection snp:

ssh -X snp

Pour se positionner sur un noeud du cluster:

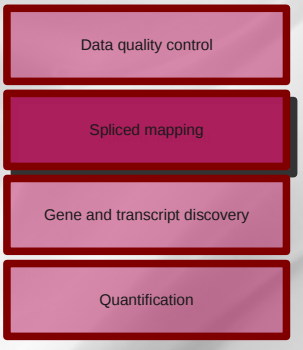
qlogin

Comptes formation: *anemone* *aster*
bleuet *iris*
muguet *narcisse*
pensee *rose*
tulipe *violette*

FastQC:

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

Analyse workflow



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

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.

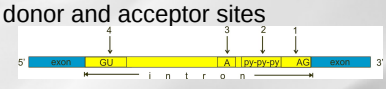
transcript_id value : unique identifier for the predicted transcript.

Splice sites

- Canonical splice site:

  which accounts for more than 99% of splicing

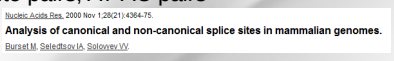
  GT and AG for donor and acceptor sites



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

- Non-canonical site:

  GC-AG splice site pairs, AT-AC pairs



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

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.

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]

- Mapsplice:

Published online 28 August 2010 Nucleic Acids Research, 2010, Vol. 38, No. 18 e178 doi:10.1093/nar/gkq622

MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery

Kai Wang[1], Darshan Singh[2], Zheng Zeng[1], Stephen J. Coleman[3], Yan Huang[1], Gleb L. Savich[4], Xiaping He[3], Piotr Mieczkowski[4], Sara A. Grimm[4], Charles M. Perou[4], James N. MacLeod[3], Derek Y. Chiang[1], Jan F. Prins[3] and Jinze Liu[1,*]

http://www.netlab.uky.edu/p/bioinfo/MapSplice

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.*

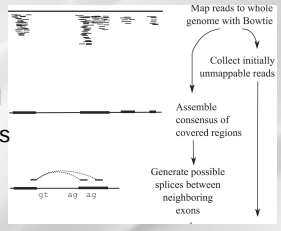
TopHat

- 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

TopHat

- TopHat then assembles the 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

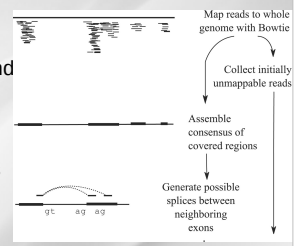


Trapnell C et al. Bioinformatics 2009;25:1105-1111

TopHat

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

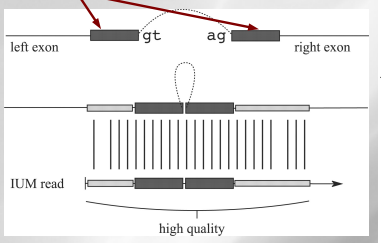


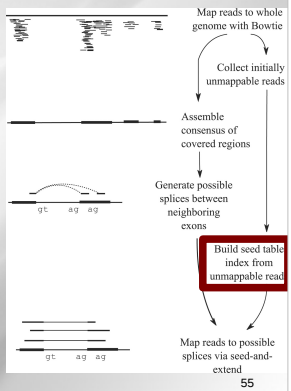
Trapnell C et al. Bioinformatics 2009;25:1105-1111

TopHat

• Each possible intron is checked against the IUM

_→ seed and extend alignment





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

TopHat Inputs

Inputs :

- Bowtie 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

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

Command line :

tophat [options] <index_base> <reads1_1[,...,readsN_1]> <[reads1_2,...readsN_2]>

TopHat Options

Some useful options :

-h/--help
-v/--version
-o/--output-dir
-r/--mate-inner-dist : no default value
-m/--splice-mismatches : default 0
-i/--min-intron-length : default 70
-I/--max-intron-length : default 500000
--max-insertion-length : default 3
--max-deletion-length : default 3
-p/--num-threads

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

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 files
- Unmapped reads

TopHat Alignment Strategies

- Default options :
  - No reference genome
- Alignment with reference genome without discovery:
  - Options -G and –no-novel-juncs

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
CCTTGGTCTTGCCGAAGTAGATCTCATTGAGAGTGGAG CGGATCTTGTTCTCCATTTCCTCCACC
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

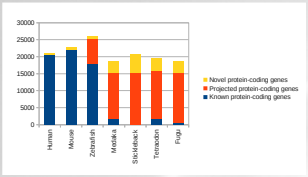
Bed exemple

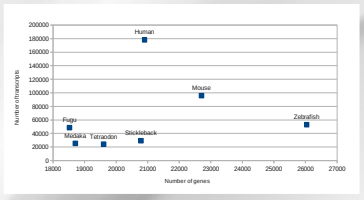
Chrom | Start | End | name | score | strand | drawing | RGB | Blocks info

junctions_ERR022486_etudechr22.bed

```
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
```

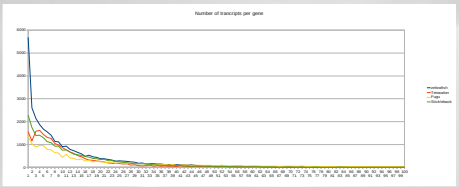

Know your transcriptome!



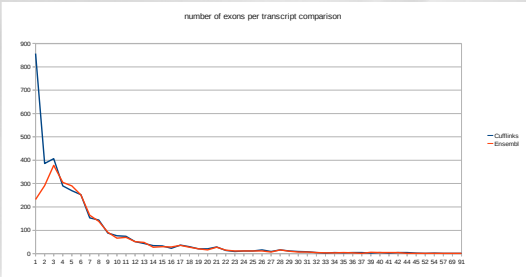


Ensembl 64
Www.ensembl.org

Numbers of exons



Number of exons



Tophat technical issues

- Temporary disk space
  - 100 000 000 pair-ends = 0,5Gb 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)

TopHat-Fusion

- an enhanced version of TopHat with the ability to align reads across fusion points
- identify fusions due to chromosomal rearrangements whether inter- or intra-chromosomal
- suggest that reads are at least 50-bp long, where a read is split into two segments (25-bp each)
- Both single and paired-end reads can be used and the output alignments are given in a modified SAM format with a new CIGAR* operator 'F' to indicate fusion points

MapSplice

MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery
Kai Wang; Darshan Singh; Zheng Zeng; Stephen J. Coleman; Yan Huang; Gleb L. Savich; Xiaping He; Piotr Mieczkowski; Sara A. Grimm; Charles M. Perou; James N. MacLeod; Derek Y. Chiang; Jan F. Prins; Jinze Liu
Nucleic Acids Research 2010; doi: 10.1093/nar/gkq622

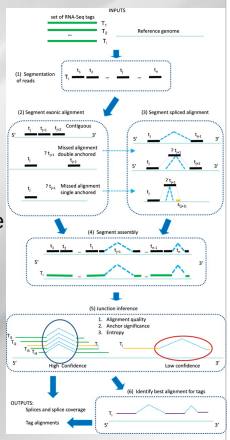
Features:

- splice junction discovery
- both CPU and memory efficiency.
- detection of small exons.
- discovery of canonical, semi-canonical and non-canonical junctions.
- splice inference based on the alignment quality and diversity of reads mapped to a junction.
- identification of chimeric events (intra-chromosomes and inter-chromosomes, inter-strands)
- support paired-end reads and single-end reads

MapSplice

Algorithm in two phases:

- tag alignment (Step 1–Step 4): give a set of candidate alignments for each tag
- splice inference (Step 5–Step 6): determine a splice significance score based on the quality and diversity of alignments that include the splice



MapSplice

Command line:

- *python bin/mapsplice_segments.py [inputs|options] MapSplice.cfg*
  or
- *python bin/mapsplice_segments.py [inputs|options]*

Options:

-Q/--reads-format: fa or fq

-o/--output-di

-c/--chromosome-files-dirr: directory containing the sequence files of the reference genome (in FASTA format)

-u/--reads-file: for paired-end reads, the order should be: reads1_end1,reads1_end2,reads2_end1,read2_end2

-B/--Bowtieidx: path and basename of index

--pairend

-L/--seglen: Length of read segments in range of [18,25]

.....

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

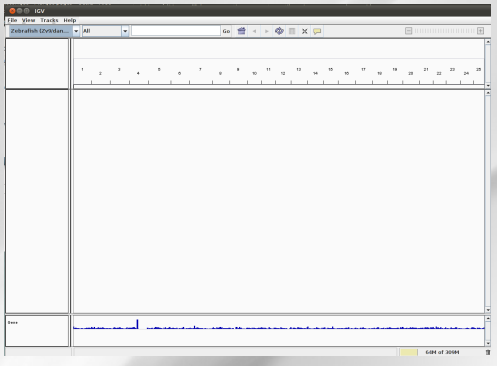
Nature Biotechnology **29**, 24–26 (2011) | doi:10.1038/nbt.1754
Published online 10 January 2011

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

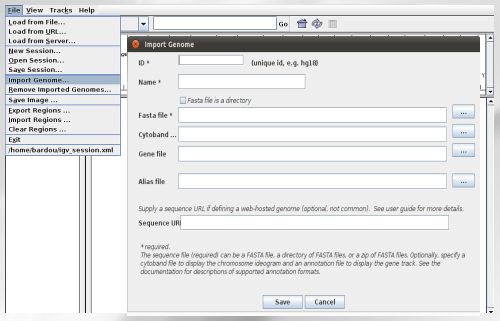
- File Extension Identifies Format
- Recommended File Formats
- BAM
- BED
- BedGraph
- bigBed
- bigWig
- Birdsuite Files
- CBS
- CN
- Cufflinks Files
- Custom File Formats
- Cytoband
- FASTA
- GCT
- genePred
- GFF
- GISTIC
- Goby
- GWAS
- IGV
- LOH
- MAF
- Merged BAM File (.bam.list)
- MUT
- PSL
- RES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

Visualizing alignments on IGV



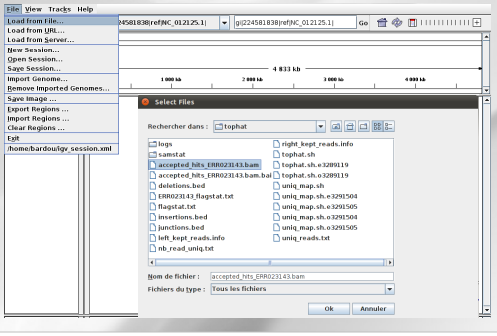
Visualizing alignments on IGV

- Import a reference genome



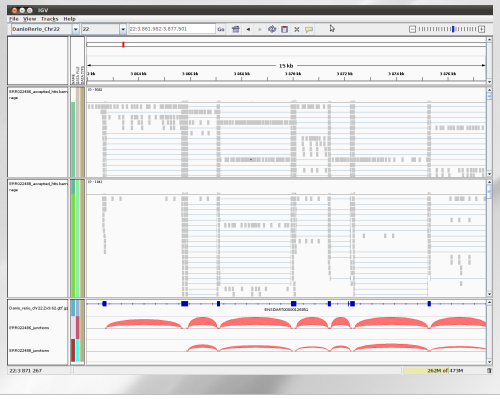
Visualizing alignments on IGV

- Import your BAM Files



Visualizing alignments on IGV

- Exemple of visualisation of bam and bed files

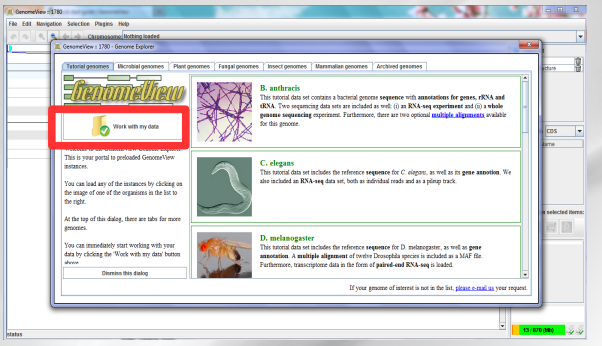


Visualizing on GenomeView

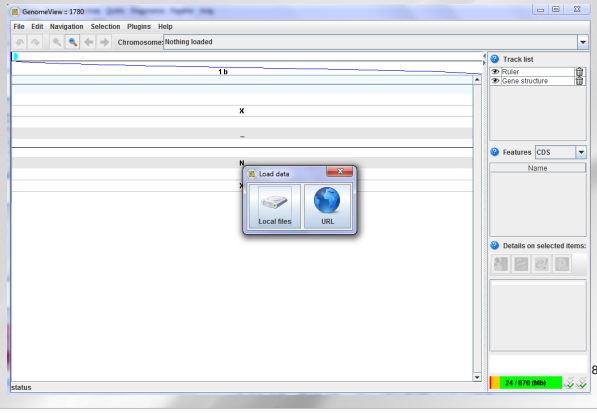
http://genomeview.org/

GenomeView is a next-generation stand-alone genome browser and editor initiated in the BSB group at VIB and currently developed at Broad Institute. It provides interactive visualization of sequences, annotation, multiple alignments, syntenic mappings, short read alignments and more. Many standard file formats are supported and new functionality can be added using a plugin system.

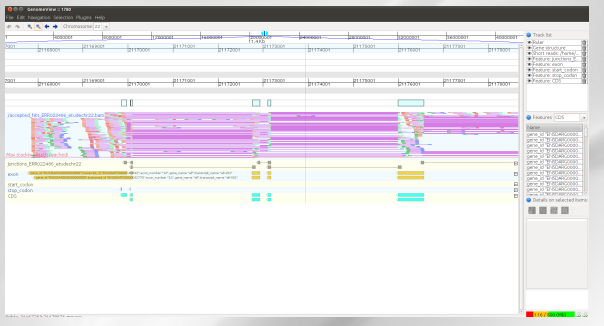
Starting GenomeView



Loading data on GenomeView



Exemple on GenomeView



Examples et TP : tophat

Commands guide :

bowtie-build sequence.fa index_name

tophat [options] <index_name> <reads1_1,reads1_2>

Options: *-h/--help*

-o/--output-dir

-r/--mate-inner-dist : no default value

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

Indexation:

samtools index accepted_hits.bam

Samtools flagstat:

samtools flagstat accepted_hits_ERR022486_chr22.bam

Examples et TP : tophat

Example of used commands:

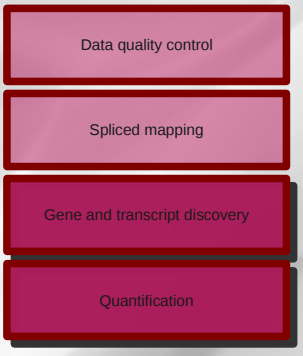
bowtie-build Danio_rerio.Zv9.62.dna.chromosome.22.fa Danio_rerio.Zv9.62_chr22

qsub -b y tophat –output-dir=tophat_ERR022486 -r 200 -G Danio_rerio_chr22.Zv9.62.gtf /work/.../bowtie-index/Danio_rerio.Zv9.62_chr22 ERR022486_read1.fastq ERR022486_read2.fastq

samtools index accepted_hits_ERR022486_chr22.bam

samtools view accepted_hits_ERR022486.bam | cut -f 1 | sort | uniq -c | cut -c 1-7 | sort -n | uniq -c

Analyse workflow



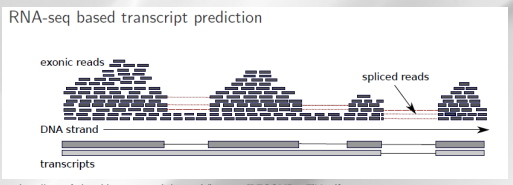
Transcript reconstruction

What mapping give us?

- Bam file with splicing localisation and pairing reads

What can we do then?

- Transcript assembly : structure determination
- Novel transcript discovery
- quantification



http://www.fml.tuebingen.mpg.de/raetsch/lectures/RECOMB-mTiM.pdf

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 tools

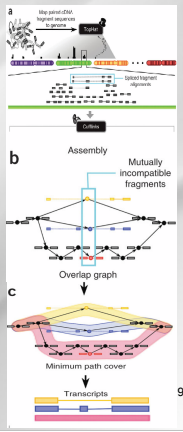
- **Cufflinks (1 bam at the time)**
  - **Assembling : gene and transcript**
- Cuffdiff (2 bam together)
  - Differential expression of genes
  - Differential promoter usage, splicing
- **Cuffcompare: (n cufflinks results)**
  - **Comparing assembled transcripts to reference annotation**
  - **Comparing transcripts across time points**
- Cuffmerge : (n cufflinks results)
  - Merge together several cufflinks assemblies
  - Run Cuffcompare

Cufflinks tools

- **Cufflinks (1 bam at the time)**
  - **Assembling : gene and transcript**
- Cuffdiff (2 bam together)
  - Differential expression of genes
  - Differential promoter usage, splicing
- **Cuffcompare: (n cufflinks results)**
  - **Comparing assembled transcripts to reference annotation**
  - **Comparing transcripts across time points**
- Cuffmerge : (n cufflinks results)
  - Merge together several cufflinks assemblies
  - Run Cuffcompare

Cufflinks transcript assembly

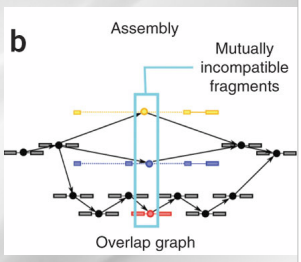
- Transcripts assembly :
  - Fragments are divided into non-overlapping loci
  - each locus is assembled independently :
- Cufflinks assembler
  - find the mini nb of transcripts that explain the reads
  - find a minimum path cover (Dilworth's theorem) :
    - nb incompatible read = mini nb of transcripts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

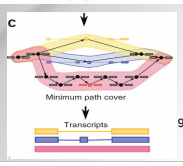


Trapnell C et al. Nature Biotechnology 2010;28:511-515

Cufflinks transcript assembly

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



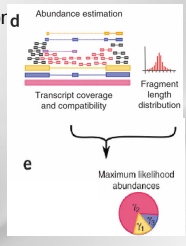


Trapnell C et al. Nature Biotechnology 2010;28:511-515

Cufflinks expression measurement

- Transcripts abundances:
  - statistical model to derive a likelihood function for the abundances
  - bayesian inference for isoform with lower abundances, or nb of read very small

  → can estimate the abundances of isoforms

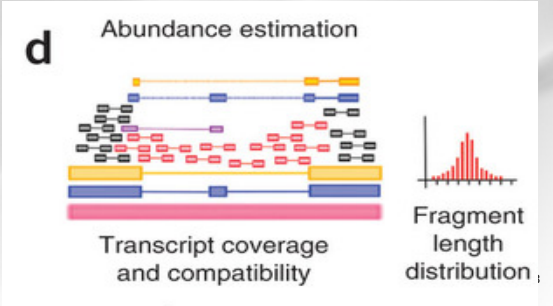


- Fragment length distribution :
  - single-end read : no way, use approximate Gaussian distribution
  - paired-end reads +assembly : learn the distribution from reads mapped to single isoform genes
  - paired-end – no assembly : genomic length of de paired-end reads (without splice) or Gaussian approximation

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

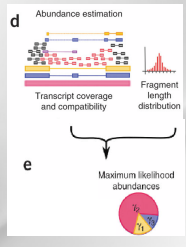
Cufflinks read attribution

- 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
  - 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

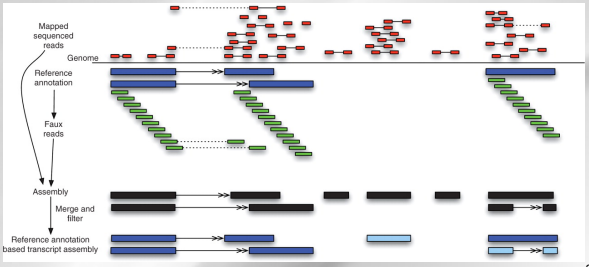
Cufflinks inputs and options

- 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
  - *-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



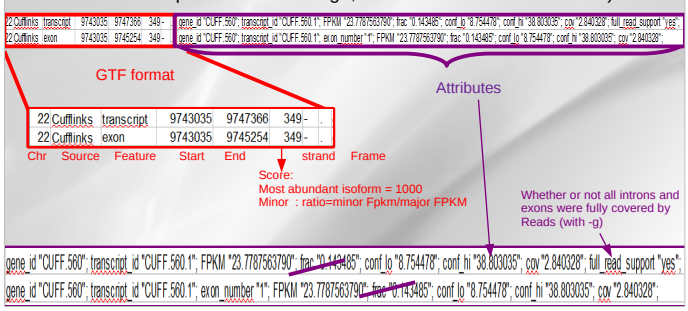
Roberts A et al. Bioinformatics 2011;27:2325-2329

Cufflinks outputs

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

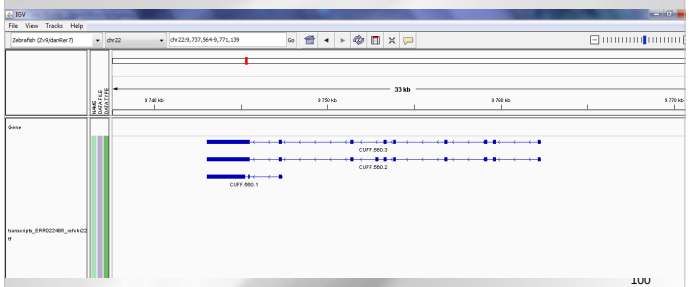
Cufflinks GTF description

- **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)



Cufflinks GTF description

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



Cufflinks tracking description

- **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

Using Cufflinks software pieces

Tophat

Identification of differential expression

Discovery of novels features : Cufflinks, Cuffmerge, cuffcompare

With discovery : Cufflinks, cuffmerge, cuffdiff

Without discovery : cuffdiff

Cuffcompare

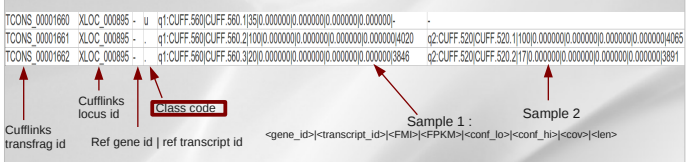
- Help to analyse the transcript assembly
- Compare your assembled transcripts to a reference annotation
- Track Cufflinks transcripts across multiple experiments
- Inputs:
  - Cufflinks' GTF output
  - optionally can take a "reference" annotation

Cuffcompare command line / options

- Command line :
  - *cuffcompare [options]* <cuff1.gtf> [cuff2.gtf] ... [cuffN.gtf]*
- Somes options:
  - *-o <outprefix>*
  - *-r* : optional "reference" annotation GFF file
  - *-R*: (If -r specified) ignore reference transcripts that are not overlapped by any transcript in one of cuff1.gtf,...,cuffN.gtf

Cuffcompare output

- Main output:
  - **<outprefix>.tracking**: matches transcripts up between samples



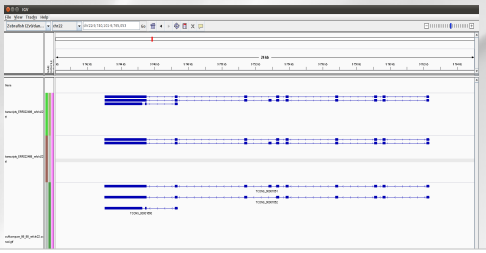
Cufcompare class codes

Priority	Code	Description
1	=	Complete match of intron chain
2	c	Contained
3	j	Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript
4	e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment.
5	i	A transfrag falling entirely within a reference intron
6	o	Generic exonic overlap with a reference transcript
7	p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
8	r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
9	u	Unknown, intergenic transcript
10	x	Exonic overlap with reference on the opposite strand
11	s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)
12	.	(.tracking file only, indicates multiple classifications)

Cuffcompare outputs

- Outputs:
  - **<outprefix>.combined.gtf**:GTF file containing the "union" of all transfrags in each sample

```
22 Cufflinks exon 9743015 9745528 . - . gene_id "XLOC_000895" transcript_id "TCONS_00001662" exon_number "1" oId "CUFF.520.2" class_code "." tss_id "TSS1300"
22 Cufflinks exon 9743015 9745528 . - . gene_id "XLOC_000895" transcript_id "TCONS_00001661" exon_number "1" oId "CUFF.520.1" class_code "." tss_id "TSS1300"
22 Cufflinks exon 9743035 9745254 . - . gene_id "XLOC_000895" transcript_id "TCONS_00001660" exon_number "1" oId "CUFF.560.1" class_code "u" tss_id "TSS1301"
```

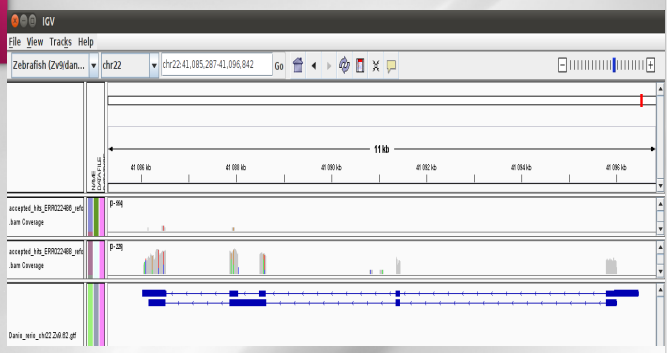


Cuffcompare outputs

- Outputs
  - **.refmap**: tab delimited file lists, for each reference transcript, which cufflinks transcripts either fully or partially match it

ref_gene_id	ref_id	class_code	cuff_id list
sdhb	ENSDART00000062264	=	CUFF.15|ENSDART00000062264_ext
mfap2	ENSDART00000082471	=	CUFF.18|ENSDART00000082471
fbxo42	ENSDART00000131067	=	CUFF.26|ENSDART00000131067

Cuffcompare gtf outputs



Cuffdiff

- Find significant changes in transcript expression, splicing, and promoter use.
- tracks changes in the relative abundance of transcripts sharing a common transcription start site: see changes in splicing
- tracks changes in the relative abundances of the primary transcripts of each gene: see changes in relative promoter use
- Input :
  - gtf file of transcripts (output of cufflinks or cuffcompare combined.gtf)
  - 2 or more sam alignment files (but analyse 2 by 2)
- Output:
  - **.fpkm_tracking** : calculates the FPKM of each transcript, primary transcript, and gene in each sample.
  - **exp.diff:** differential expression tests for each pair of samples at gene, isoform... level

CummeRbund

http://compbio.mit.edu/cummeRbund/

- R package that is designed to aid and simplify the task of analyzing Cudiff outputs: explore and visualize the data
- Now included as part of R/Bioconductor
- Takes the output files from cuffdiff and creates a SQLite database of the results describing appropriate relationships betweeen genes, transcripts, transcription start sites, and CDS regions
- numerous plotting functions as well for commonly used visualizations


NATURE PROTOCOLS | PROTOCOL

Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter

Affiliations | Contributions | Corresponding author

Nature Protocols 7, 562–578 (2012) | doi:10.1038/nprot.2012.016
Published online 01 March 2012

Sigcufflinks

- Cufflinks code has been modified by the BioInformatic Platform of Toulouse in order to obtain raw count of reads: use **sigcufflinks** on **snp**
- Run cufflinks, cufflinks outputs + raw_transcripts.tsv:

gene_id	transcript_id	paires	forward	reverse
CUFF.6	CUFF.6.1	4873	4873	3431
CUFF.6	CUFF.6.2	5222	5222	3769
CUFF.6	ENSDART00000067635	4819	4819	3580

- Other scripts are in development : exemple

cuffcompare_quant_summary.pl on **snp**

Cuffcompare_quant_summary

- Other scripts are in development : exemple

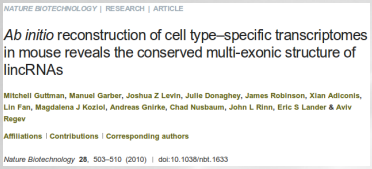
cuffcompare_quant_summary.pl on **snp**

- Parse cuffcompare tracking file and produce summary file with quantification for each transcript in each condition.
- Usage:

quant_summary.pl [-p <prefix>] [-o <outdir>] [-r transcriptome.gtf] --tracking cuffcmp.tracking lib1name_trasncripts.gtf [lib2name_transcripts.gtf …]

- The gtf files are output from cufflinks.
- This script also rename the gene and transcript id for each input cufflinks gtf in order to get shared trans and genes id.

Scripture


NATURE BIOTECHNOLOGY | RESEARCH | ARTICLE

Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs

Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Kozicki, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander & Aviv Regev

Affiliations | Contributions | Corresponding authors

Nature Biotechnology **28**, 503–510 (2010) | doi:10.1038/nbt.1633

- Method to reconstruct the transcriptome using only RNA-seq data and genome sequence (unannotated)
- 6 steps:
  - Use reads uniquely aligned to the genome
  - From the aligned spliced reads: construc a connectivity graph
  - Using all data: segmentation approach to identify significant paths
  - Construct a transcript graph
  - Generate a catalogue of transcripts

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

- 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 &*

Somes 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: 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.)

Exemples et TP : cufflinks

Commands guide:

Cufflinks command:

cufflinks -- output-dir=cufflinks_ERR022486
-g Danio_rerio_chr22.Zv9.62.gtf
accepted_hits_ERR022486.bam

Cuffcompare command:

cuffcompare -o cuffcompare_refVSERR022486
-r Danio_rerio_chr22.Zv9.62.gtf
cufflinks_ERR022486/transcripts_ERR022486.gtf

Cuffdiff command:

cuffdiff Danio_rerio_chr22.Zv9.62.gtf
tophat/accepted_hits_ERR022486.bam
tophat/accepted_hits_ERR022488.bam

Exemples et TP : cufflinks

Examples of used commands :

Nb of genes in GTF:

cat Danio_rerio_chr22.Zv9.62.gtf | cut -f9 | cut -d";" -f1 | sort -u | wc -l

Nb of transcripts in GTF:

cat Danio_rerio_chr22.Zv9.62.gtf | cut -f9 | cut -d";" -f2 | sort -u | wc -l

Cufflinks:

qsub -b y cufflinks –output-dir=/work/.../cufflinks_ERR022486 -g
/work/.../Danio_rerio_chr22.Zv9.62.gtf
/work/../tophat_ERR022486/accepted_hits_ERR022486.bam

Nb of genes in cufflinks results:

cat genes_ERR022486.fpkm_tracking | cut -f1 | sort -u | wc -l

Exemples et TP : cufflinks

Examples of used commands :

Cuffcompare:

cuffcompare -o cuffcompare_refVSERR022486 -r
/work/.../Danio_rerio_chr22.Zv9.62.gtf
/work/.../cufflinks_ERR022486_refchr22/transcripts_ERR022486_refchr22.gtf

Cuffdiff:

qsub -b y cuffdiff /work/.../Danio_rerio_chr22.Zv9.62.gtf
/work/.../tophat_ERR022486/accepted_hits_ERR022486.bam
/work/..../tophat_ERR022488/accepted_hits_ERR022488.bam

Significant differential expression of genes:

cat gene_exp.diff | sort -r -k 14 | less -S

Exemples et TP : cufflinks

Examples of used commands :

Sigcuffdiff:

sigcufflinks --output-dir=/work/.../sigcufflinks_ERR022486 -g
/.../Danio_rerio_chr22.Zv9.62.gtf
/..../tophat_ERR022486/accepted_hits_ERR022486.bam

Cuffcompare_quant_summary:

cuffcompare_quant_summary.pl -p quant_86_88 -o
/work/.../cuffcompare_quantSummary -r /..../Danio_rerio_chr22.Zv9.62.gtf --tracking
/..../cuffcompare_sigcufflinks/cuffcompare_86_88.tracking
/.../sigcufflinks_ERR022486/ERR022486_transcripts.gtf
/.../sigcufflinks_ERR022488/ERR022488_transcripts.gtf

Quality for Bioinfo Plateform!

Exam :
http://bioinfo.genotoul.fr/index.php?id=93

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/