Plateforme Bioinformatique Midi-Pyrénées

RNA-Seq data analysis

http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/

Delphine Labourdette Get-Biopuce / Céline Noirot Bioinfo Genotoul

Material

http://genoweb.toulouse.inra.fr/~formation/4_Galaxy_RNAseq/

Slides & Exercise leaflet (doc)

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

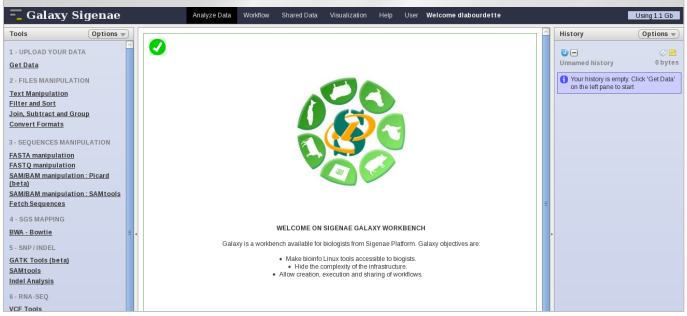
Data & results files (data)

Session organisation

- Sequence quality
  - Theory + exercises
- Spliced read mapping Visualisation
  - Theory + exercises
- expression measurement
  - Theory + exercises
- mRNA calling
  - Theory + exercises

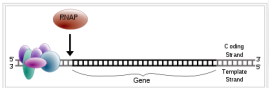
What you should know

How to connect to Sigenar galaxy workbench?
http://sigenae-workbench.toulouse.inra.fr/galaxy/

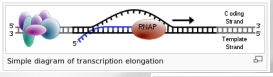


Transcription

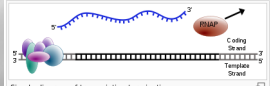
Transcription is the process of creating a complementary RNA copy of a sequence of DNA. Transcription is the first step leading to gene expression.



Simple diagram of transcription initiation. RNAP = RNA polymerase



Simple diagram of transcription elongation



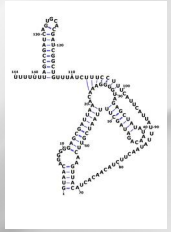
Simple diagram of transcription termination

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

Transcription products

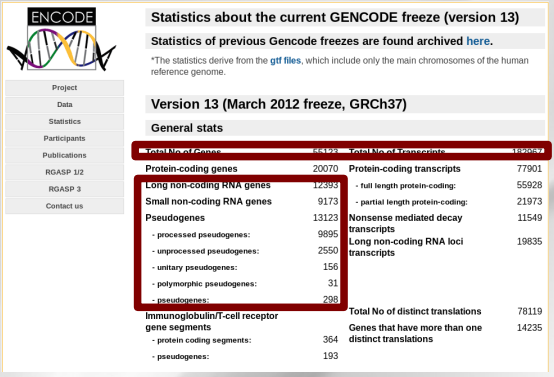
Protein coding gene: transcribed in mRNA
ncRNA : highly abundant and functionally important RNA

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



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

ENCODE



ENCODE

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

Statistics about the current GENCODE freeze (version 13)

Statistics of previous Gencode freezes are found archived here.

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

Version 13 (March 2012 freeze, GRCh37)

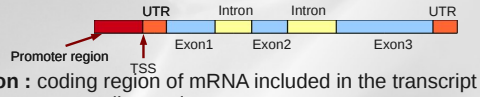
General stats

Total No of Genes	55123	**Total No of Transcripts**	185854
Protein-coding genes	20070	**Protein-coding transcripts**	77901
Long non-coding RNA genes	12393	**- full length protein-coding:**	55928
Small non-coding RNA genes	9173	**- partial length protein-coding:**	21973
Pseudogenes	13123	**Nonsense mediated decay transcripts**	11549
- processed pseudogenes:	9895	**Long non-coding RNA loci transcripts**	19835
- unprocessed pseudogenes:	2550		
- unitary pseudogenes:	156		
- polymorphic pseudogenes:	31		
- pseudogenes:	298		
Immunoglobulin/T-cell receptor gene segments		**Total No of distinct translations**	78119
- protein coding segments:	364	**Genes that have more than one distinct translations**	14235
- pseudogenes:	193		

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

Vocabulary

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

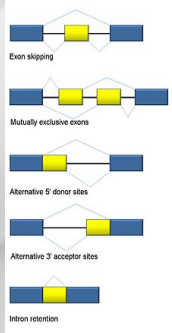


Exon : coding region of mRNA included in the transcript
Intron : non coding region
TSS : Transcription Start Site ≠ 1st amino acid
Transcript : stretch of DNA transcribed into an RNA molecule

Alternative splicing

Alternative splicing (or differential splicing) 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.

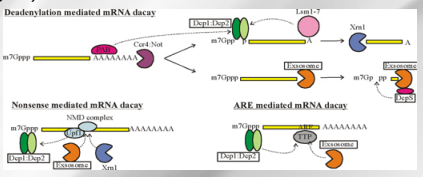


Intron Retention

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

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

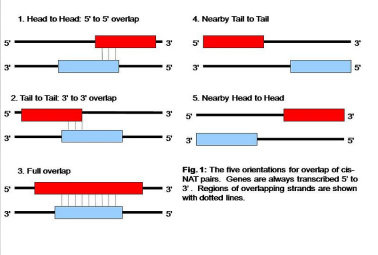


Fig. 1: The five orientations for overlap of cis-NAT pairs. Genes are always transcribed 5' to 3'. Regions of overlapping strands are shown with dotted lines.

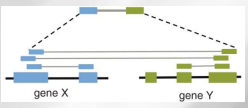
Fusion genes

- 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.kalloniemi@finmi.fi.

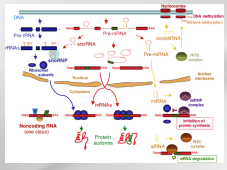


http://en.wikipedia.org/wiki/Fusion_gene

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

Transcriptome variability summary

- 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



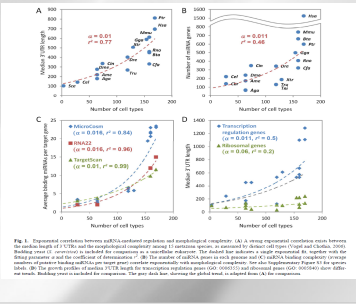
http://www.nature.com/emboj/journal/v25/n5/fig_tab/7601023a_F2.html

UTR length

Lengthening of 3'UTR Increases with Morphological Complexity in Animal Evolution

Cho-Yi Chen[1,2], Shui-Tein Chen[2], Hsueh-Fen Juan[1,*] and Hsuan-Cheng Huang[3,*]
[1]Genome and Systems Biology Degree Program, Department of Life Science, Institute of Molecular and Cellular Biology, Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan,
[2]Institute of Biological Chemistry, Academia Sinica, Taipei, Taiwan
[3]Institute of Biomedical Informatics, Center for Systems and Synthetic Biology, National Yang-Ming University, Taipei, Taiwan

Associate Editor: Martin Bishop



Techniques classification ?

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

→ Need transcript sequence partially known

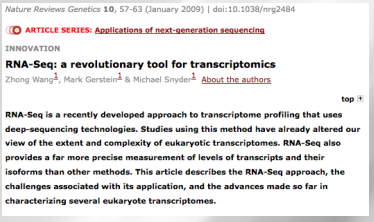
→ Difficulties in discovering novels splice events

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

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


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

ARTICLE SERIES: Applications of next-generation sequencing

INNOVATION

RNA-Seq: a revolutionary tool for transcriptomics

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

top

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

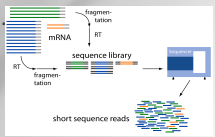
http://en.wikipedia.org/wiki/RNA-Seq

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

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, ...



Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

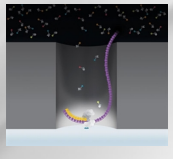
RNA-Seq platforms comparison

Séquenceurs 2ème génération

	Roche			Illumina				Life Technologies			
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	Titanum	FLX Titanum	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

Third Generation RNA-Seq

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



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

Different approaches :

Alignment to

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

What are we looking for?

Identify genes
- List new genes

Identify transcripts
- List new alternative splice forms

Quantify these elements → differential expression

Usual questions on RNA-Seq !

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

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

Statisticians answers

Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing

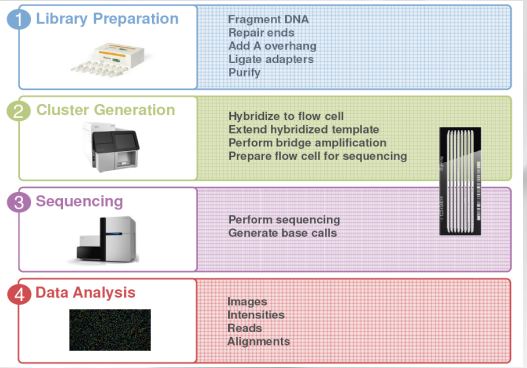
BMC Genomics 2012, **13**:484 doi:10.1186/1471-2164-13-484

Jose A Robles (jose.robles@csiro.au)

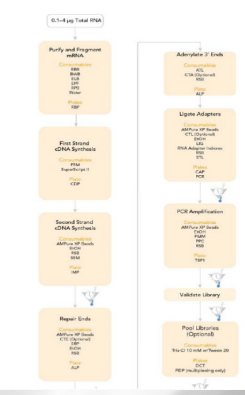
Conclusions

This work quantitatively explores comparisons between contemporary analysis tools and experimental design choices for the detection of differential expression using RNA-Seq. We found that the DESeq algorithm performs more conservatively than edgeR and NBPSeq. With regard to testing of various experimental designs, this work strongly suggests that greater power is gained through the use of biological replicates relative to library (technical) replicates and sequencing depth. Strikingly, sequencing depth could be reduced as low as 15% without substantial impacts on false positive or true positive rates.

Illumina RNA-Seq protocol



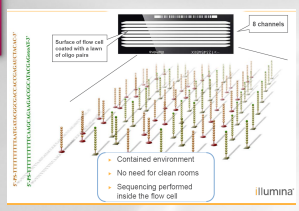
RNA-Seq library preparation

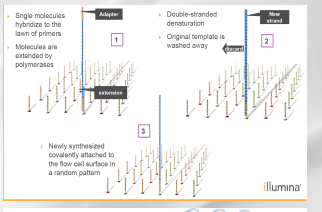


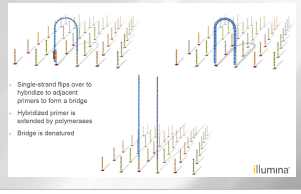
- Isolate poly-A containing mRNA
- capture mRNA with oligoT beads
- Randomly fragment RNA
- Random prime mRNA → cDNA
- Make 2nd strand cDNA
- Repair-Ends and 3' Ends Adenylate
- Ligate sequencing adapters
- Enrich up to 15 cycles of PCR
- gel purify
- validate library w/ Bioanalyzer

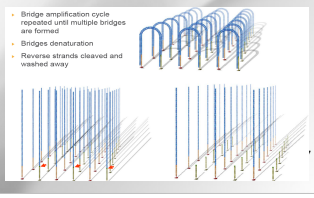
Library prep takes <2 days

Clusters generation









Sequencing

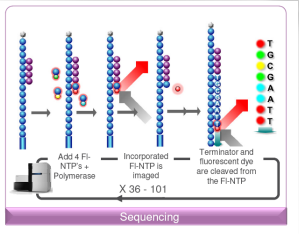
Cluster generation:

- 35 amplification cycles
- 1 cluster → 2 000 identical molecules
- 500 000 clusters / flocell

Sequencing:

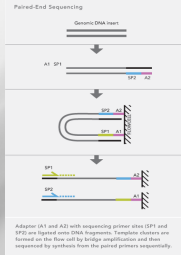
- Image acquisition:
  - 50 min / cycle

Ex: 2x100bp → 2x100x50 min

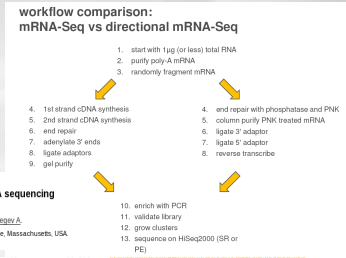


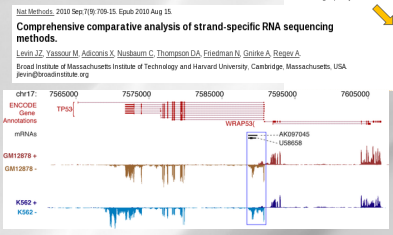
Paired-end sequencing

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



Strand specific RNA-Seq protocol





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

Analysis workflow

Data quality control

Spliced mapping

Quantification

Gene and transcript discovery

fastq file formats

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

SURVEY AND SUMMARY

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

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

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

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

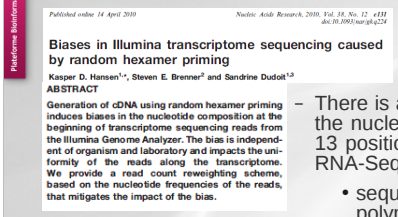
RNAseq specific bias

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

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

- Transcript length bias
- Some reads map to multiple locations

Hexamer random priming bias

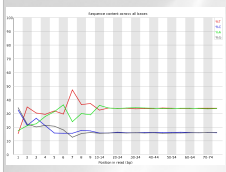

Published online 14 April 2010 Nucleic Acids Research, 2010, Vol. 38, No. 12 e131
doi:10.1093/nar/gkq224

Biases in Illumina transcriptome sequencing caused by random hexamer priming

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

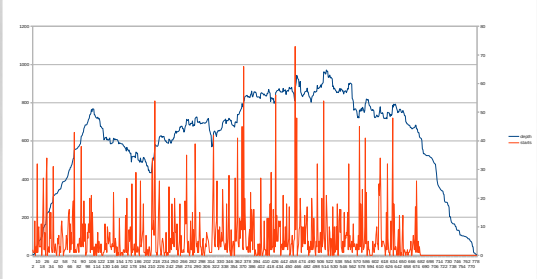
ABSTRACT

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



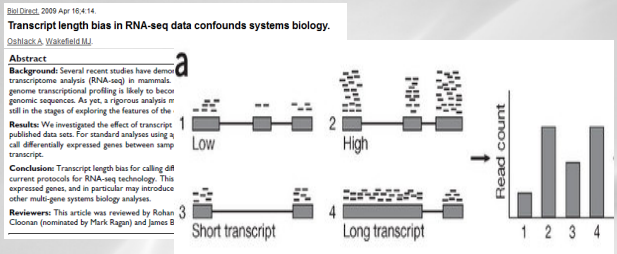
- There is a strong distinctive pattern in the nucleotide frequencies of the first 13 positions at the 5'-end of mapped RNA-Seq reads:
  - sequence specificity of the polymerase
  - due to the end repair performed
- Reads beginning with a hexamer over-represented in the hexamer distribution at the beginning relative to the end are down-weighted

Hexamer random effect



- Orange = reads start sites
- Blue = coverage

Transcript length bias


Biol Direct. 2009 Apr 16;4:14.

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

Oshlack A, Wakefield MJ.

Abstract

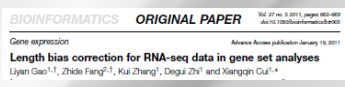
Background: Several recent studies have demon
transcriptome analysis (RNA-seq) in mammals.
genome transcriptional profiling is likely to becom
genomic sequences. As yet, a rigorous analysis m
still in the stages of exploring the features of the

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

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

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

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


BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 8 2011, pages 662–669
doi:10.1093/bioinformatics/btr000

Gene expression Advance Access publication January 19, 2011

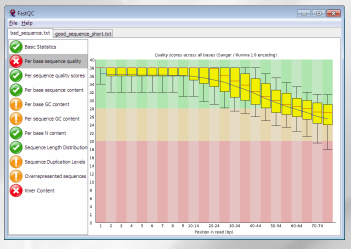
Length bias correction for RNA-seq data in gene set analyses

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

Verifying RNA-Seq raw data

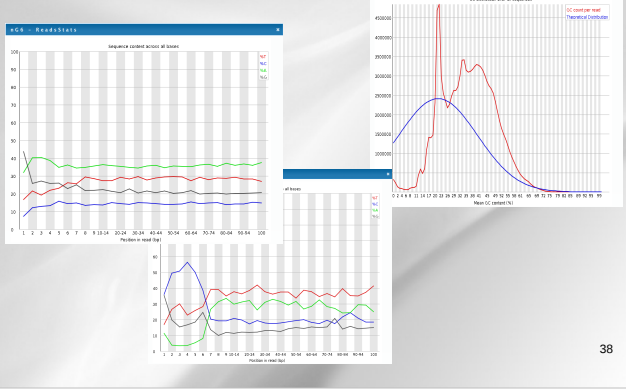
FastQC :

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



- ***Has been developed for genomic data***

Fourmiz !!!



Take home message on quality analysis

Elements to be checked :

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

Alignment on reference for the second quality check and filtering.

A good run?:

- Expected number of reads produced (2x500millions / flowcell),
- Length of the reads expected (100pb),
- Random selection of the nucleotides and the GC%,
- Good alignment: very few unmapped reads, pairs mapped on opposite strands.

Analysis workflow

Data quality control

Spliced mapping

Quantification

Gene and transcript discovery

Where to find a reference genome?

Retrieving the genome file (fasta):

- The Genome Reference Consortium

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

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

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

Reference transcriptome file

What is a GTF file ?:

- derived from GFF (General Feature Format, for description of genes and other features)
- Gene Transfer Format:

http://genome.ucsc.edu/FAQ/FAQformat.html#format4

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

The [attribute] list must begin with:

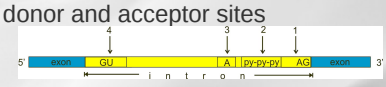
gene_id value : unique identifier for the genomic source of the sequence.

transcript_id value : unique identifier for the predicted transcript.

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

Splice sites

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



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

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

Nucleic Acids Res. 2000 Nov 1;28(21):4364-75.
Analysis of canonical and non-canonical splice sites in mammalian genomes.
Burset M, Seledtsov IA, Solovyev VV.

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

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

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.

- Idea :
  - Database of potential splice junction sequences (known)
  - splice canonical / non canonical site search (seed then mapping)

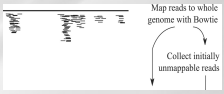
TopHat

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]

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.*

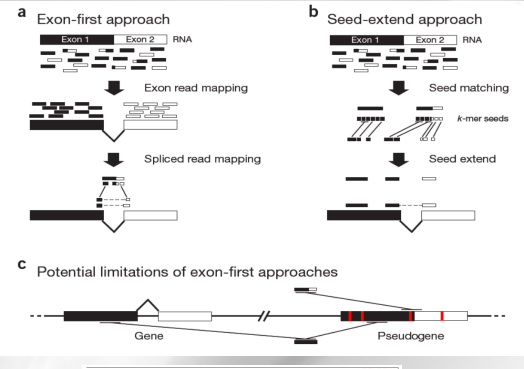
TopHat algorithm : first step



- 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 first approach limitation

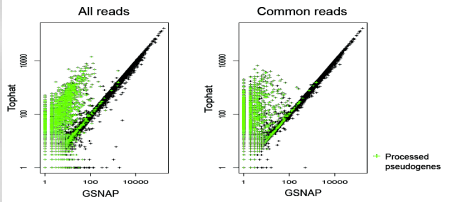


REVIEW

Computational methods for transcriptome annotation and quantification using RNA-seq

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

TopHat and pseudogenes



Gene-level quantification : HTSeq intersection non-empty

`--read-realign-edit-dist`	Some of the reads spanning multiple exons may be mapped incorrectly as a contiguous alignment to the genome even though the correct alignment should be a spliced one - this can happen in the presence of processed pseudogenes that are rarely (if at all) transcribed or expressed. This option can direct TopHat to re-align reads for which the edit distance of an alignment obtained in a previous mapping step is above or equal to this option value. If you set this option to 0, TopHat will map every read in all the mapping steps (transcriptome if you provided gene annotations, genome, and finally splice variants detected by TopHat), reporting the best possible alignment found in any of these mapping steps. This may greatly increase the mapping accuracy at the expense of an increase in running time. The default value for this option is set such that TopHat will not try to realign reads already mapped in earlier steps.

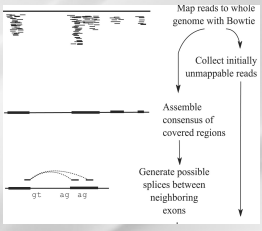
Why does it find small exons?

- In the last tophat versions :

  Short read sequencing machines can currently produce reads 100bp or longer but many exons are shorter than this so they would be missed in the initial mapping. **TopHat solves this problem mainly by splitting all input reads into smaller segments which are then mapped independently**. The segment alignments are put back together in a final step of the program to produce the end-to-end read alignments.

Exon assembly process

- 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



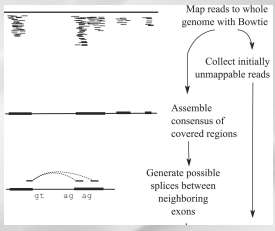
- Extend islands to cover eventually splice junctions
  - +/- 45 bp from reference on either side of island

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

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
  - 50 bp < Intron size < 500,000 bp

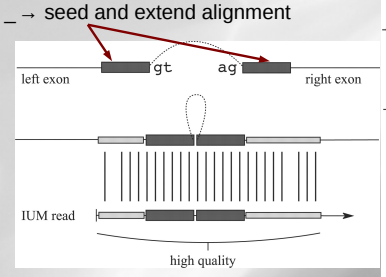


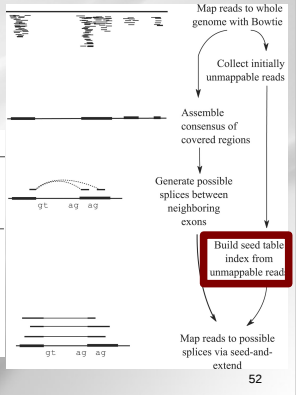
Trapnell C et al. Bioinformatics 2009;25:1105-1111

IUM alignment

- Each possible intron is checked against the IUM

_ → seed and extend alignment





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

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

Command lines :

bowtie2-build <reference.fasta> <index_base>

tophat [options] <index_base> <reads1_1]> <[reads1_2]>

TopHat Options

Some useful options (command line) :

-h/--help
-v/--version
- - bowtie1 (instead of bowtie2)
- o/--output-dir
-r/--mate-inner-dist : no default value
-m/--splice-mismatches : default 0
-i/--min-intron-length : default 50
-I/--max-intron-length : default 500000, prefer 25000 for non human
--max-insertion-length : default 3
--max-deletion-length : default 3
-p/--num-threads

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

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

flag chr pos pair

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

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

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
```


Tophat technical issues

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

Trans-splicing with 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

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

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 Penn
2 Institute for Translational Medicine and Therapeuti
3 F.M. Kirby Center for Molecular Ophthalmology, U
4 Department of Genetics, University of Pennsylvani
5Department of Pharmacology, University of Pennsy

Associate Editor: Prof. Ivo Hofacker

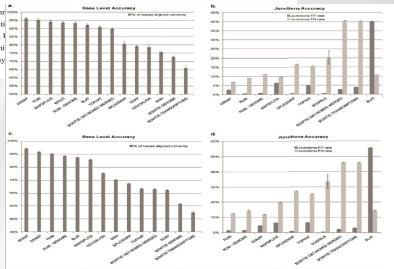


Fig 6. Accuracy statistics for analyses of simulated data sets. A, B. Simulated data set 1. C,D. Simulated data set 2. Test 1 has low polymorphism and error rates, while Test 2 has moderate polymorphism and error rates. In A and C The dark bars show the base wise accuracy (the percent of bases that aligned 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 FP rate is the number of false junctions called divided by the total, and 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 GSNAP, MapSplice and RUM are the ones that are most viable for RNA Seq alignment.

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

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

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

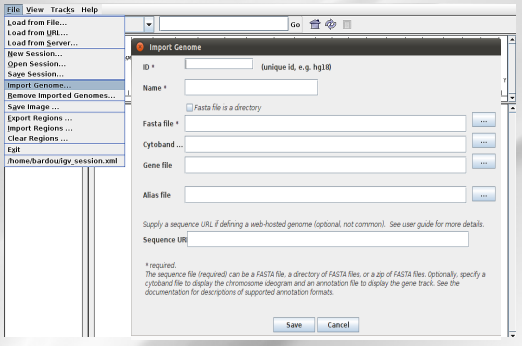
Visualizing alignments on IGV

- 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

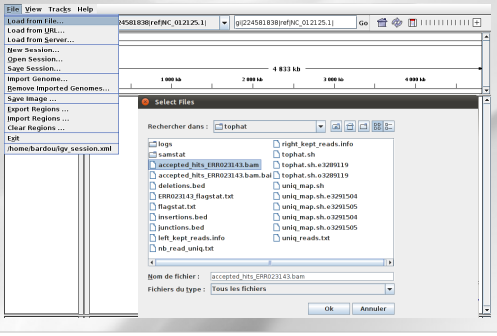
Visualizing alignments on IGV

Import a reference genome



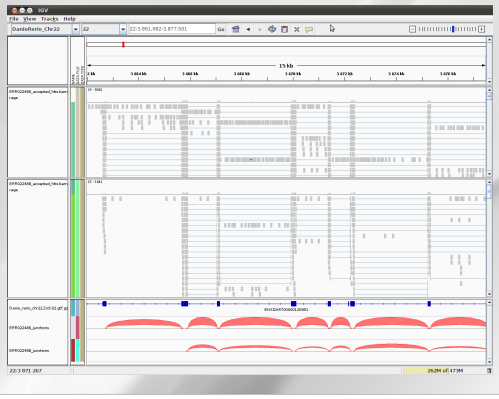
Visualizing alignments on IGV

- Import your BAM Files



Visualizing alignments on IGV

- Exemple of bam and bed files visualisation



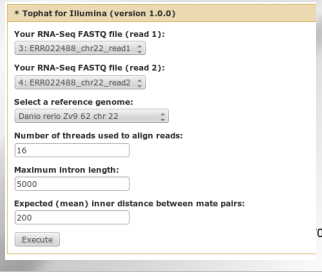
Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

hands-on : tophat

Tophat location: 8 – Trainings

- *RNA-Seq*
  - *Step 2 : Alignement and statistics*
    - ** Tophat for Illumina Find splice junctions using RNA-seq data*



Indexation: ** Samtools index*

Samtools flagstat

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

Analysis workflow

Data quality control

Spliced mapping

Quantification

Gene and transcript discovery

Plateforme Bioinformatique Midi-Pyrénées

geno toul bioinfo

What do we want to build?

The gene / transcript description file (and corresponding fasta)

9	protein_coding	exon	697785	697947	.	-	.	gene_id "ENSDARG00000075706"; transcript_id "ENSDART00000144625"; exon_number "1"
9	protein_coding	exon	698518	698690	.	-	.	gene_id "ENSDARG00000075706"; transcript_id "ENSDART00000144625"; exon_number "2"
9	protein_coding	exon	694364	694502	.	-	.	gene_id "ENSDARG00000075706"; transcript_id "ENSDART00000144625"; exon_number "3"
9	protein_coding	CDS	694364	694497	.	-	0	gene_id "ENSDARG00000075706"; transcript_id "ENSDART00000144625"; exon_number "4"
9	protein_coding	start_codon	694495	694497	.	-	.	gene_id "ENSDARG00000075706"; transcript_id "ENSDART00000144625"; exon_number "3"
9	protein_coding	exon	693526	693675	.	-	.	gene_id "ENSDARG00000075706"; transcript_id "ENSDART00000144625"; exon_number "4"
9	protein_coding	CDS	693675	693822	.	-	1	gene_id "ENSDARG00000075706"; transcript_id "ENSDART00000144625"; exon_number "4"
9	protein_coding	stop_codon	693672	693674	.	-	0	gene_id "ENSDARG00000075706"; 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 "ENSDARG00000011996"; transcript_id "ENSDART00000136627"; exon_number "1"
9	protein_coding	CDS	697639	697832	.	+	0	gene_id "ENSDARG00000011996"; transcript_id "ENSDART00000136627"; exon_number "1"
9	protein_coding	stop_codon	697832	697835	.	+	0	gene_id "ENSDARG00000011996"; transcript_id "ENSDART00000136627"; exon_number "1"
9	protein_coding	exon	698442	698573	.	+	.	gene_id "ENSDARG00000011996"; transcript_id "ENSDART00000136627"; exon_number "2"
9	protein_coding	CDS	698442	698573	.	+	0	gene_id "ENSDARG00000011996"; transcript_id "ENSDART00000136627"; exon_number "2"
9	protein_coding	exon	694401	694469	.	+	.	gene_id "ENSDARG00000011996"; transcript_id "ENSDART00000136627"; exon_number "3"
9	protein_coding	CDS	694401	694469	.	+	0	gene_id "ENSDARG00000011996"; transcript_id "ENSDART00000136627"; exon_number "3"
9	protein_coding	exon	700596	700876	.	+	.	gene_id "ENSDARG00000011996"; transcript_id "ENSDART00000136627"; exon_number "4"
9	protein_coding	CDS	700666	700725	.	+	0	gene_id "ENSDARG00000011996"; transcript_id "ENSDART00000136627"; exon_number "4"

The count file

	rsem.names	SRR519722	SRR519728	SRR519729	SRR519730	SRR519731	SRR519747	SRR519748	SRR519749	SRR519750	SRR519751
1	mrs_c1	1855	4895	4483	4407	3826	1740	4355	3979	4396	4506
2	mrs_c2	358	614	628	834	854	393	769	444	1015	732
3	mrs_c3	1874	1392	2583	1355	1245	2090	5194	4952	12012	4150
4	mrs_rep_c4	897	789	1044	1100	1362	957	1931	836	1289	1313
5	mrs_rep_c5	5785	12617	17276	16120	15121	6562	16388	12329	18525	16999
6	mrs_rep_c6	2165	4727	6457	5312	4960	2399	7910	5194	8663	6716
7	mrs_rep_c7	293	436	637	627	694	247	839	522	928	840
8	mrs_rep_c8	814	1425	1806	1867	2050	691	1937	1551	1667	1552
9	mrs_rep_c9	786	1885	2739	2489	2573	735	2345	2012	3508	2645
10	mrs_rep_c10	311	517	684	886	895	346	659	581	1041	1030
11	mrs_rep_c11	91	212	294	210	175	88	192	291	259	299
12	mrs_rep_c12	1129	2181	2833	3128	3088	1139	2983	2675	4384	3811
13	mrs_rep_c13	536	913	544	1254	1275	515	1929	913	1407	1444
14	mrs_rep_c14	4678	13751	18595	16722	16476	4962	16867	14581	17733	18771
15	mrs_rep_c15	7209	22854	32748	28889	27176	8532	28967	26591	35040	30702
16	mrs_rep_c17	645	1568	2068	2530	3072	860	1764	1451	3527	3438
17	mrs_rep_c18	4413	1669	7758	8570	9559	3954	6610	6180	8273	8728
18	mrs_rep_c19	1765	2941	4757	4265	4062	1852	4804	3968	4983	4202
19	mrs_rep_c20	1236	2914	3188	2903	2605	818	2196	1843	2478	2410
20	mrs_rep_c22	2315	4329	5968	5760	5582	2471	5153	5451	5906	6482
21	mrs_rep_c24	4488	7523	11333	10104	9537	4409	8676	9297	9040	10178
22	mrs_rep_c25	618	742	961	1155	1265	938	895	715	1481	1506
23	mrs_rep_c26	1397	2569	3436	3251	3009	1310	2947	2785	2989	3267
24	mrs_c27	766	889	1283	1264	1577	820	1224	1139	1539	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 (sigcufflinks)

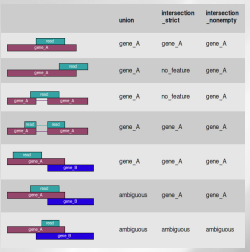
HTSeq-count

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 &*



	union	intersection _strict	intersection _nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A read	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read gene_A read gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous	gene_A	gene_A
read gene_A gene_B	ambiguous	ambiguous	ambiguous

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: a table with counts for each feature and a summary of reads not counted for any feature:

```
ENSDARG00000095643      967
ENSDARG00000095659      4
ENSDARG00000095667      36
ENSDARG00000095677      98
ENSDARG00000095760      5
no_feature      362748
ambiguous       9937
too_low_aQual   0
not_aligned     0
alignment_not_unique    239465
```

- *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.)

Cufflinks in general

NATURE BIOTECHNOLOGY | RESEARCH | LETTER

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

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

Affiliations | Contributions | Corresponding author

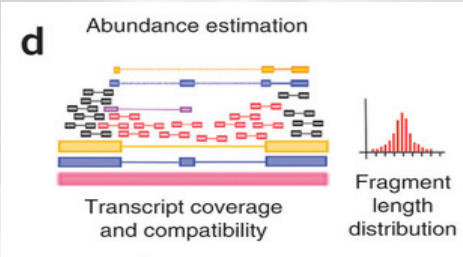
Nature Biotechnology **28**, 511–515 (2010) | doi:10.1038/nbt.1621
Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

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

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

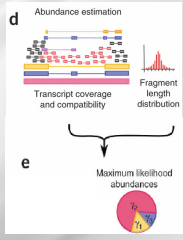
Cufflinks read attribution

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



Cufflinks expression measurement

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



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

RPKM / FPKM

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

Sigcufflinks

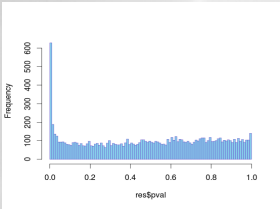
- Cufflinks code has been modified by the Sigenae Team of Toulouse in order to obtain raw count of reads: use **sigcufflinks** on **genotoul**
- Run cufflinks, cufflinks outputs + raw_transcripts.tsv:

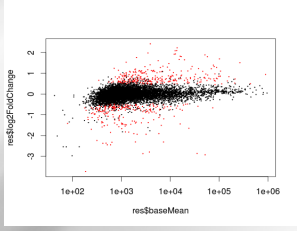
gene_id	transcript_id	pairs	forward	reverse
CUFF.6	CUFF.6.1	4873	4873	3431
CUFF.6	CUFF.6.2	5222	5222	3769
CUFF.6	ENSDART00000067635	4819	4819	3580

In R with DEseq

```
> 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.

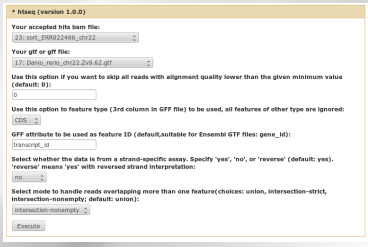
2/ Quantify the genes and transcripts of chromosome 22 using sigcufflinks and the Ensembl GTF file for both samples.

3/ In each case merge the files to produce the count tables.

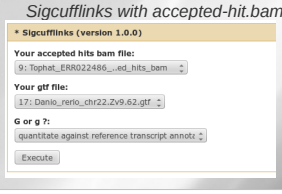
Hands-on : hints

samtools sort by read names

htseq-count on sorted bam file and strand-specific assay specify 'no', select mode to handle reads overlapping more than one feature(choice:intersection-nonempty)

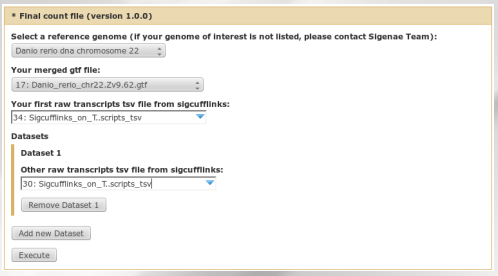


Sigcufflinks with accepted-hit.bam



Hands-on : file merging

Merge sigcufflinks



Analysis workflow

- Data quality control
- Spliced mapping
- Quantification
- Gene and transcript discovery

Transcript reconstruction

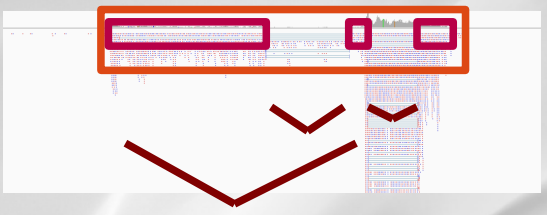
The different paths :

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

Defining the model building strategy

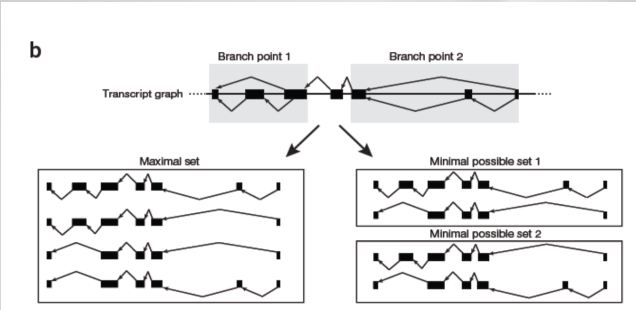
- Number of built models
- Intronic reads

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

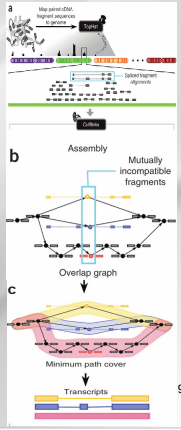
Nature Biotechnology 28, 511–515 (2010) | doi:10.1038/nbt.1621
Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

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

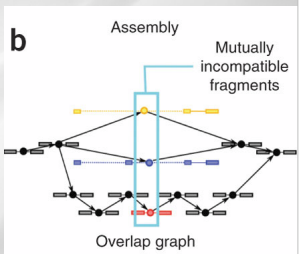
- Transcripts assembly :
  - Fragments are divided into non-overlapping loci
  - each locus is assembled independently :
- Cufflinks assembler
  - find the mini nb of transcripts that explain the reads
  - find a minimum path cover (Dilworth's theorem) :
    - nb incompatible read = mini nb of transcripts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

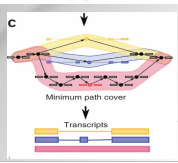


Trapnell C et al. Nature Biotechnology 2010;28:511-515

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

Some videos of examples

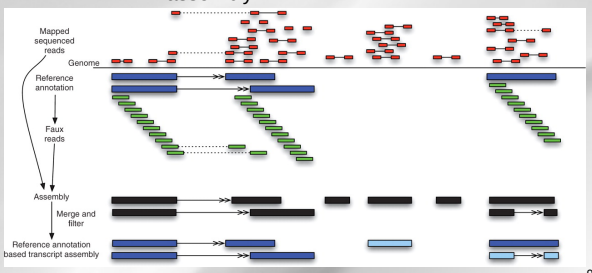
- Chromosome 3 of the bovine genome, UMD3
- 3 locations
- 3 tracks :
  - Ensembl reference gene
  - Cufflinks model
  - Reads alignment

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

Chr	Source	Feature	Start	End		strand	Frame
22	Cufflinks	transcript	9743035	9747366	349	-	.
22	Cufflinks	exon	9743035	9745254	349	-	.

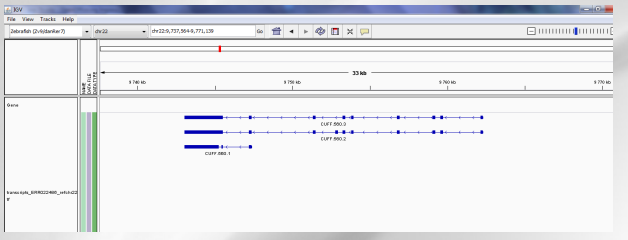
Score:
Most abundant isoform = 1000
Minor : ratio=minor Fpkm/major FPKM

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

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

Cufflinks GTF description

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



Cufflinks tracking description

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

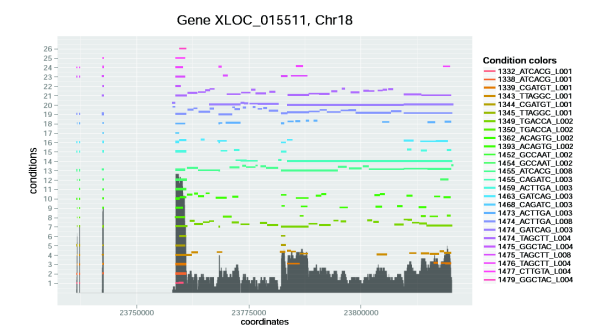
Quantification status

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

Cufflink transcript models



Cufflink transcript models

Gene discovery pipeline

Alignment (Tophat)

↓

Bam merge (samtools)

↓

Discovery of novels features (cufflinks)

↓

Quantification at a transcript level (sigcufflinks)

↓

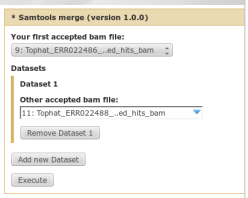
Quantification file merging (shell script)

Quantification strategy

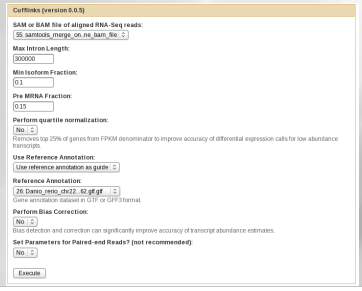
- First set your gene and transcript model = build a reference GTF file
- Then use option -G to quantify the same set of elements on all your samples with sigcufflinks
- Then sort your raw_transcript.tsv files
- cut the second or third column of the sorted file
- Paste all the column in the count file

Hands-on : cufflinks

Merge all bam : Step 5 : RNAseq De Novo



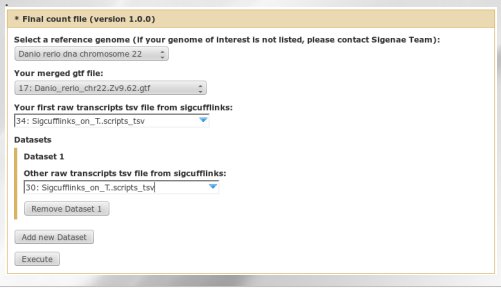
Cufflinks on merge file with -g option (reference annotation as guide) and the Danio gtf file :



Hands-on : file merging

Sigcufflinks with the new gtf file (transcript.gtf of previously step) with -G option

Final count :



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/