

RNA-Seq data analysis



Material

- **Slides:** <http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/doc/>
 - pdf : one per page
 - pdf : three per page with comment lines
- **Memento:**
 - <http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/doc/MementoUNIX.pdf>
 - <http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/doc/MementoCluster.pdf>
- **Hands on:**
 - Data files: <http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/>
 - Results files:
http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/correction_star_rsem/



The speakers

Sarah Maman



Christine Gaspin



Céline Noirot



Matthias Zytnicki



Claire Hoede



Cédric Cabau



Nathalie Villa-Vialaneix (pour la partie biostat)



Session organisation

Day 1

Morning (9h00 -12h30) :

- Prerequisite unix/format
Exercises
- Biological reminders

Afternoon (14h-17h) :

- Sequence quality
Theory + exercises
- Spliced read mapping
Theory

Day 2

Morning (9h00 -12h30) :

- Spliced read mapping
Exercises and Visualisation
- Expression quantification
Theory + exercises

Afternoon (14h-17h) :

- mRNA calling
Theory + exercises
- Models comparison
Theory + exercises



Prerequisite unix

Summary - Unix reminders

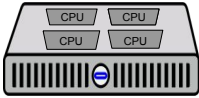
- Genotoul infrastructure organisation
- How to connect to genotoul
- How to transfer data
 - From the web to genotoul
 - From genotoul to your computer
- How to launch jobs on the cluster
- ...

Vocabulary : Cluster / Node

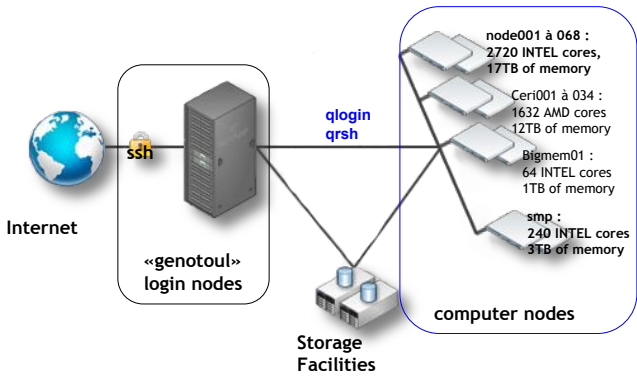
- Cluster : set of nodes



- Node : Huge computer (with several CPUs)



« genotoul » cluster



Genotoul Bioinfo

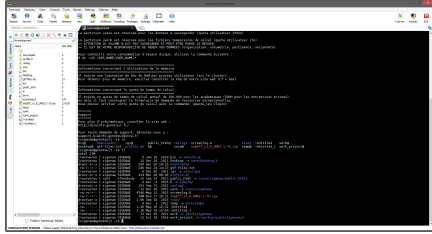
Disk spaces

/usr/local/bioinfo/src	Bioinformatics Software
/bank/	International genomics Databanks
/home/	User configuration files (ONLY) (100 MB user quota)
/work/	HPC computational disk space (TEMPORARY) (1 TB user quota)
/save/	User disk space (with BACKUP) (250 GB user quota)

How to connect to genotoul ?

- Xming (Windows graphic) + Putty (Connection)

- MobaXterm (Executable file)



- command line: in a Terminal windows
`ssh username@genotoul.toulouse.inra.fr`

Linux

Linux account

Access to a work environment

- Login + password
- Share resources (Cpu, memory, disk)
- Usage of software installed
- Free access to computational cluster
- Own space disk (/save & /work directory)

Linux

Which are the main unix/linux commands you know ?

Very Important Tips

- **Copy / Paste with the mouse**
 - Select a text (it is automatically copied)
 - Click on the mouse wheel (the text is pasted where the cursor is located)
- **Command and path completion :**
 - Use the TAB key
- **Back to the previously used commands :**
 - Use the « up » and « down » keys

OGE (Open Grid Engine)

Queues availables for users

Queue	Access	Priority	Max time	Max slots
workq (default)	everyone	300	96H	4120
unlimitq	everyone	100	unlimited	680
smpq	on demand	0	unlimited	240
hypermemq	on demand	0	unlimited	96
Interq (qlogin)	everyone	100	48H	40
galaxyq	galaxy users	No node shared	unlimited	120

OGE (Open Grid Engine)

Characteristics of "work" working space

- Workq
 - 1 core
 - 8 GB memory maximum
 - Write only /work directory (temporary disk space)
- Work space
 - 1 TB quota disk per user (on /work directory)
 - 120 days files without access automatic purged
- Time resource constraint
 - 100 000H annually computing time (more on demand)

OGE (Open Grid Engine)

qlogin (with display) / qrsh or qrsh -X

```
Connected → [laborie@genotoul2 ~]$ qlogin
Your job 2470388 ("QLOGIN") has been submitted
waiting for interactive job to be scheduled ...
Your interactive job 2470388 has been successfully scheduled.
Establishing /SGE/ogs/inra/tools/qlogin_wrapper.sh session to host
node001 ...
[laborie@node001 ~]$

Disconnected → [laborie@node001 ~]$ exit
logout
/SGE/ogs/inra/tools/qlogin_wrapper.sh exited with exit code 0
[laborie@genotoul2 ~]$
```

OGE (Open Grid Engine)

Job Submission : some examples

- Default (workq, 1 core, 8 GB memory max)

```
$qsub myscript.sh
Your job 15660
("mon_script.sh") has been submitted
```

- More memory (workq, 1 core, 32 / 36 GB memory)

```
$qsub -l mem=32G -l h_vmem=36G myscript.sh
Your job 15661
("mon_script.sh") has been submitted
```

- More cores (workq, 8 core, 8*8 GB memory)

```
$qsub -l parallel smp 8 myscript.sh
Your job 15662
("mon_script.sh") has been submitted
```

OGE (Open Grid Engine)

Job Submission : some examples

```
Script edition → $nedit myscript.sh

### head of myscript.sh ###
# !/bin/bash
# $ -m a
# $ -l mem=32G
# $ -l h_vmem=36G

# Mon programme commence ici
ls
### end of myscript.sh ###

Submission → $qsub myscript.sh
Your job 15660
("mon_script.sh") has been submitted
```

OGE (Open Grid Engine)

qsub : batch Submission

1 - First write a script (ex: *myscript.sh*) with the command line as following:

```
#$ -N job_name           to give a name to the job
#$ -o /work/.../output_file_name  to redirect output standard
#$ -e /work/.../error_file_name   error_file_name : to redirect error file
#$ -q workq             queue_name : to specify the batch queue
#$ -m bea               mail sending : (b:begin, a:abort, e:end)
#$ -l mem=8G           to ask for 8GB of mem (minimum reservation)
#$ -l h_vmem=10G       to fix the maximum consumption of memory
# My command lines I want to run on the cluster
blastall -d swissprot -p blastx -i /save/.../z72882.fa
```

2 - Then submit the job with the qsub command line as following:

```
$qsub myscript.sh
Your job 15660
("mon_script.sh") has been submitted
```

OGE (Open Grid Engine)

Monitoring jobs : qstat

```
$qstat
```

```
job-ID prior name user state submit/start queue slots ja-task-ID
```

Job-ID : job identifier
prior : priority of job
name : job name
user : user name
state : actual state of job (see follow)
submit/start at : submit/start date
Queue : batch queue name
slots : number of slots asked for the job
ja-task-ID : job array task identifier (see follow)

OGE (Open Grid Engine)

Monitoring jobs : qstat

- **state** : actually state of job
 - d(letion) : job is deleting
 - E(rror) : job is in error state
 - h(old), w(waiting) : job is pending
 - t(ransferring) : job is about to be executed
 - r(unning) : job is running
- **man qstat** : to see all options of qstat command

OGE (Open Grid Engine)

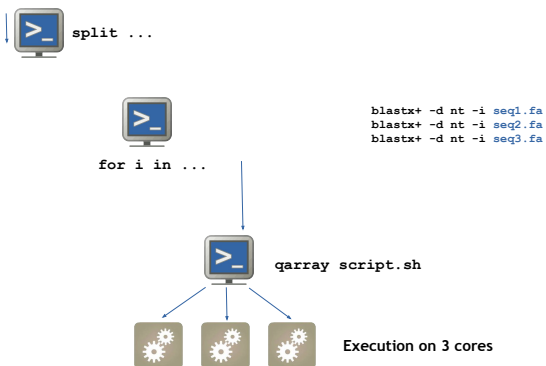
Deleting a job : qdel

```
$qstat -u laborie
job-ID prior name user state submit/start at queue
slots ja-task-ID
-----
3629151 512.54885 sleep laborie r 02/25/2015 16:23:03
workq@node002 1
$ qdel 3629151
laborie has registered the job 3629151 for deletion
```

Array of jobs concept

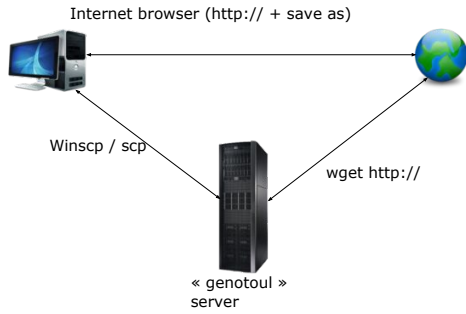
- Concept : segment a job into smaller atomic jobs
- Improve the processing time very significantly (the calculation is performed on multiple processing cores)

blast in job array mode



Downloading / transferring

Several possible cases



Downloading / transferring

File download from Internet to « genotoul server »:

- Copy the URL of the file to download

```
wget http://url.a.telecharger/nom_fichier
```

Downloading / transferring

Transfer between genotoul and desktop computer

We recommend to use « scp » command (secure copy)

```
scp [user@host1:]file1 [user@host2:]file2
```

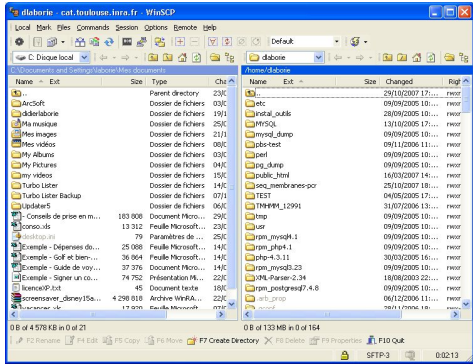
copy file from the network

Example copy from desktop to "genotoul":

```
scp source_name bleuet@genotoul:destination_name
```

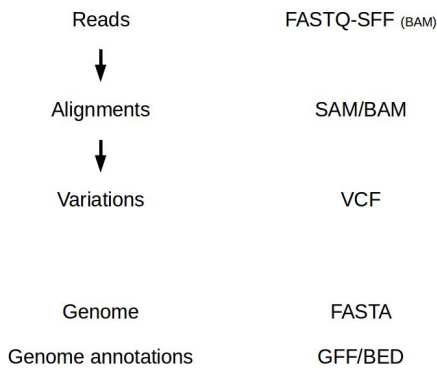
Downloading / transferring

WinSCP / FileZilla : copy via graphical interface



Introduction to NGS formats

Summary - Format remind



fastq format

- The standard for storing outputs of HTS
- A text-based format for storing a read and its corresponding quality scores.
- 1 read \leftrightarrow 4 lines

```
@HWI-ST218:596:C90JYANXX:8:1101:1293:2188 1:N:0:ATTCAGAAATACCTTA
NCTAAGTGTAGAGGGGTTTCCGCCCTTACTGTCTCAGCTACCACTAAGCACTCCCGCGGAGTACGGTGCAGACTGAAAA
+
!<3?BFGGGGGEGGGGGGGGGGGGG#F1FGGGGGDDGGL1F#</9FB#EGGGGGGG>GGGGGGGG<<C/BDGGGGGC=GG
```

1. Begins with '@' character and is followed by a sequence identifier
2. The raw sequence
3. Begins with a '+' character and is optionally followed by the same sequence identifier
4. Encodes the quality values for the read, contains the same number of symbols as letters in the read

fastq format

Published online 16 December 2009

Nucleic Acids Research, 2010, Vol. 38, No. 6, 1767-1771
doi:10.1093/nar/gkq115

SURVEY AND SUMMARY

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

Peter J. A. Cock¹*, Christopher J. Fields², Naohisa Goto³, Michael L. Heuer⁴ and Peter M. Rice⁵

@EASS4_6_R1_2_1_413_324
CCCTTCTGTCTCAGGTTTCTCC

+
!;3;111111111111?11111180

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

```
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....  
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....  
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....  
.....LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....  
!*$%&'()*+,-./0123456789:<=>?@ABCDEFGHIJKLMNPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~  
33 59 64 73 104 126  
0.....26...31.....40  
-5...0.....9.....40  
0.....9.....40  
0.2.....26...31.....41
```

S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

Sequence Alignment/Map (SAM) format

- Data sharing was a major issue with the 1000 genomes
- Capture all of the critical information about NGS data in a single indexed and compressed file
- Generic alignment format
- Supports short and long reads (454 – Solexa – Solid)
- Flexible in style, compact in size, efficient in random access

Website :

<http://samtools.sourceforge.net>

Paper :

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]

Sequence Alignment/Map (SAM) format

- 11 mandatory fields
- Variable number of optional fields
- Fields are tab delimited

Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSITION/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ("-" if same as RNAME)
8	MPOS	1-based Mate POSition
9	TLEN	inferred Template LENGTH (insert size)
10	SEQ	query SEquence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Sequence Alignment/Map (SAM) format

Header

```

ERR000017_2.sam
@SQ   SN:ref   LI:4633380
1    16   ref    740202   0   18M   *   0   0   TTTTTTTTTTTTTTTTTT   >>>>????????????   XT:R:R   NM:i:2   MD:Z:59G56
2    16   ref    740202   0   18M   *   0   0   TTTTTTTTTTTTTTTTTT   <<<<<<<<<<<<<<<<<<<<   XT:R:R   NM:i:2   MD:Z:59G56
3    16   ref    740202   0   18M   *   0   0   TTTTTTTTTTTTTTTTTT   >>>>????????????   XT:R:R   NM:i:2   MD:Z:59G56
4    16   ref    740202   0   18M   *   0   0   TTTTTTTTTTTTTTTTTT   >>>>????????????   XT:R:R   NM:i:2   MD:Z:59G56
5    16   ref    740202   0   18M   *   0   0   TTTTTTTTTTTTTTTTTT   >>>>????????????   XT:R:R   NM:i:2   MD:Z:59G56
6    16   ref    740202   0   18M   *   0   0   TTTTTTTTTTTTTTTTTT   >>>>????????????   XT:R:R   NM:i:2   MD:Z:59G56
7    16   ref    740202   0   18M   *   0   0   TTTTTTTTTTTTTTTTTT   >>>>????????????   XT:R:R   NM:i:2   MD:Z:59G56
8    16   ref    740202   0   18M   *   0   0   TTTTTTTTTTTTTTTTTT   >>>>????????????   XT:R:R   NM:i:2   MD:Z:59G56
9    16   ref    740202   0   18M   *   0   0   TTTTTTTTTTTTTTTTTT   >>>>>>>>>>>>>>   XT:R:R   NM:i:2   MD:Z:59G56
10   0    ref    2702037   25  18M   *   0   0   CTTGCGACTGCTGCTGCTT   >>>>???:?77?><   XT:R:U   NM:i:2   MD:Z:3611G2
11   16   ref    2195895   37  18M   *   0   0   GCGTCTGCTGCTGCTGCTT   >>>>>>>>>>>>><   XT:R:U   NM:i:0   MD:Z:18
12   0    ref    2866664   37  18M   *   0   0   GTTTGTTGTTGTTGTTGTT   !>>>70?>387748(7   XT:R:U   NM:i:0   MD:Z:18
13   16   ref    511005   37  18M   *   0   0   GTTCTGCTGCTGCTGCTGCT   </?>+789+>>>>>>   XT:R:U   NM:i:0   MD:Z:18
14   16   ref    740202   0   18M   *   0   0   TTTTTTTTTTTTTTTTTT   >>>>>>>>>>>>>>   XT:R:R   NM:i:2   MD:Z:59G56
15   4    *      0         0   *     0   0   GTGCGACTGCTGCTGCTGCT   >8!1!>8-9,77(+88
16   16   ref    740202   0   18M   *   0   0   TTTTTTTTTTTTTTTTTT   >>>>????????????   XT:R:R   NM:i:2   MD:Z:59G56
17   0    ref    1847349   37  8M118M *   0   0   GATDGDGRTGRTGRTGRTT   >48!1;9,77+!+6+!/?
    
```

Alignment

<QNAME> <FLAG> <RNAME> <POS> <MAPQ> <CIGAR> <MRNM> <MPOS> <ISIZE> <SEQ> <QUAL>

[<TAG>:<VTYPE>:<VALUE> [...]]

X? : Reserved for end users
 NM : Number of nuc. Difference
 MD : String for mismatching positions
 RC : Read group
 [...]

A : Printable character
 i : Signed 32-bit integer
 f : Single-precision float number
 Z : Printable string
 H : Hex string (high nybble first)

SAM format - Flag field

- Decimal values in sam lines

Flag	Chr	Description
0x0001	p	the read is paired in sequencing
0x0002	P	the read is mapped in a proper pair
0x0004	u	the query sequence itself is unmapped
0x0008	U	the mate is unmapped
0x0010	r	strand of the query (1 for reverse)
0x0020	R	strand of the mate
0x0040	1	the read is the first read in a pair
0x0080	2	the read is the second read in a pair
0x0100	s	the alignment is not primary
0x0200	f	the read fails platform/vendor quality checks
0x0400	d	the read is either a PCR or an optical duplicate

- Picard tools
<https://broadinstitute.github.io/picard/explain-flags.html>

How to manipulate them ?

- Samtools
<http://samtools.sourceforge.net/>
- Picard tools
<https://broadinstitute.github.io/picard/>
- Bedtools
<http://bedtools.readthedocs.io/en/latest/>

Hands-on : unix & formats

Training accounts :

anemone	arome
aster	bleuet
camelia	capucine
chardon	clematite
cobee	coquelicot
cosmos	cyclamen
dahlia	digitale
geranium	gerbera

Exercice 1 : using basic unix commands

Exercice 2 : format manipulation

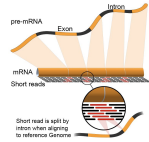
Summary - Biological reminders

- Context, vocabulary, transcriptome variability ...
- Methods to analyse transcriptomes
- What is RNAseq ?
- High throughput sequencers
- Illumina protocol, paired-end library, directional library
- Retrieve public data and presentation of data for practical work

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



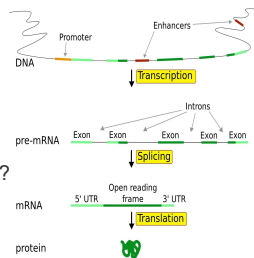
Context

Prerequis :

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

Try to answer to :

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



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

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

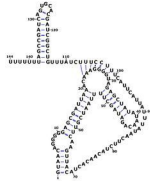


Transcription products

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



GENCODE Data Stats Browser Blog

Statistics about the current GENCODE freeze (version 21)

Statistics of previous GENCODE freezes are found archived [here](#).
 * The statistics derive from the [gtf file](#) that contains only the annotation of the main chromosomes.
 For details about the calculation of these statistics please see the [README_stats.txt](#) file.

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

General stats

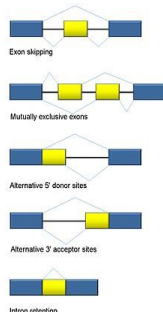
Total No of Genes	60155	Total No of Transcripts	196327
Protein-coding genes	19881	Protein-coding transcripts	79377
Long non-coding RNA genes	15877	- full length protein-coding:	54420
Small non-coding RNA genes	9534	- partial length protein-coding:	24957
Pseudogenes	14467	Nonsense mediated decay transcripts	13222
- processed pseudogenes:	10753	Long non-coding RNA loci transcripts	26414
- unprocessed pseudogenes:	3230		
- unitary pseudogenes:	170		
- polymorphic pseudogenes:	59		
- pseudogenes:	29		
Immunoglobulin/T-cell receptor gene segments		Total No of distinct translations	59512
- protein coding segments:	395	Genes that have more than one distinct translations	13526
- pseudogenes:	226		

<http://www.encodegenes.org/stats.html>

Alternative splicing

Alternative splicing (or differential splicing)

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



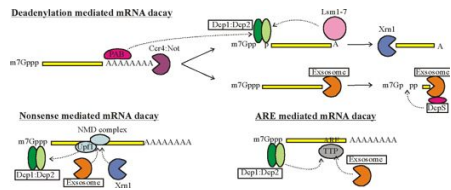
Intron Retention

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

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

Transcript degradation

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



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

Cis-natural antisense transcript

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

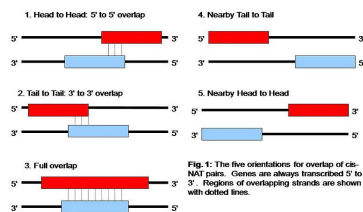


Fig. 1: The five orientations for overlap of cis-NAT pairs. Cisense are always transcribed 5' to 3'. Regions of overlapping strands are shown with dotted lines.

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

Fusion genes

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

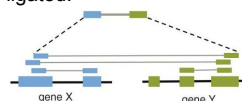
Genome Biol. 2011; 12(10):1186. Epub ahead of print.

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

Edgren H, Murumali A, Kandaswami S, Hicinski D, Monaghan V, Kozlowski K, Rye H, Nishida S, Hoff M, Borresen-Dale AL, Kallioniemi O, Institute for Molecular Medicine Finland (FIMM), Tukholmankatu 8, Helsinki, 00290, Finland. olli.kallioniemi@fimm.fi

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

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



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

Transcriptome variability

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

How can we study the transcriptome?

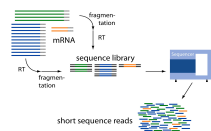
Techniques classification

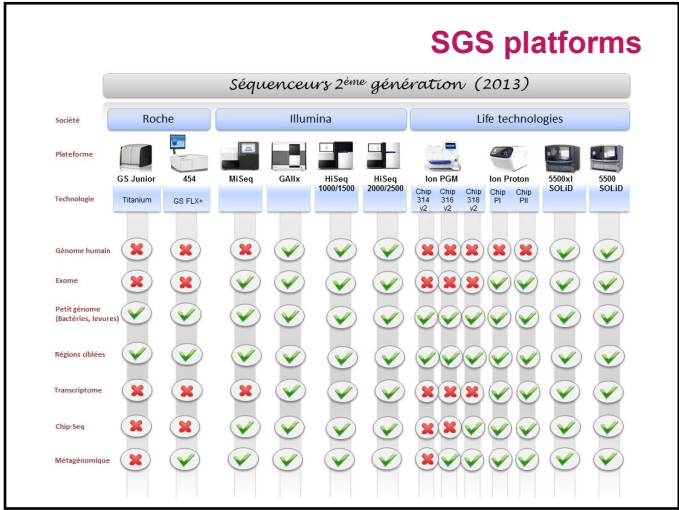
EST	PCR/RT-QPCR	SAGE	MicroArrays
No quantification	Quantification	Quantification	Indirect quantification
Low throughput	Low throughput (up to hundreds)	Low throughput (up to thousands)	High throughput (up to millions)
Discovery (Yes)	No	No	Discovery (Yes)

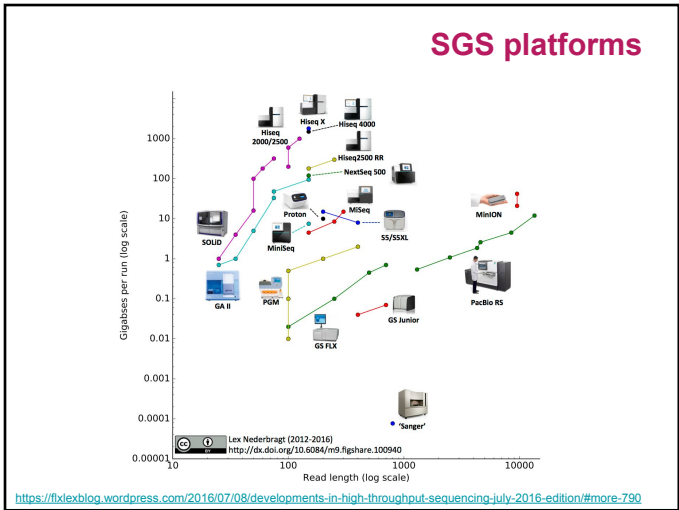
- Need transcript sequence partially known
- Difficulties in discovering novels splice events

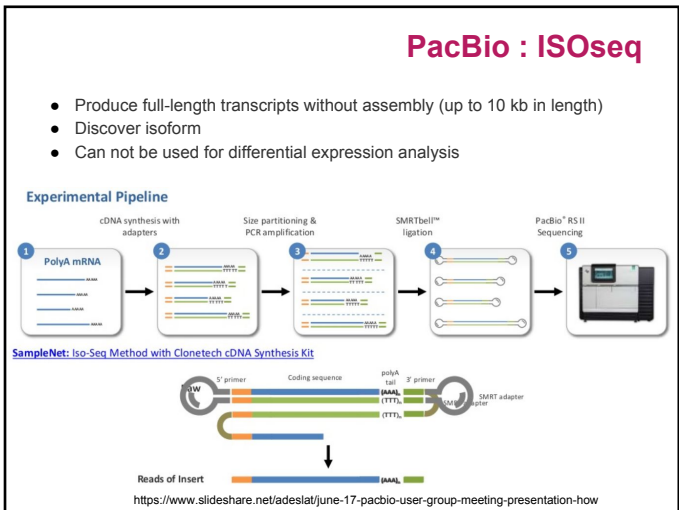
What is different with RNA-Seq ?

- 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, ...







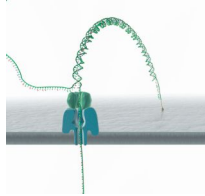


MinION

Available now for sequencing cDNAs

- Longest read length: 98kb
- Median read length: 1kb
- Mean read length: 2kb

<http://dx.doi.org/10.1016/j.bbq.2015.02.001>



Coming next: direct analysis of RNA

- RNA modifications
- PCR-free protocols
- Increased accuracy compared to using reverse transcriptases

<https://www.slideshare.net/adeslat/june-17-pacbio-user-group-meeting-presentation-how>

What are we looking for?

Identify genes

- List new genes

Identify transcripts

- List new alternative splice forms

Quantify these elements → differential expression



Usual questions on RNA-Seq !

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

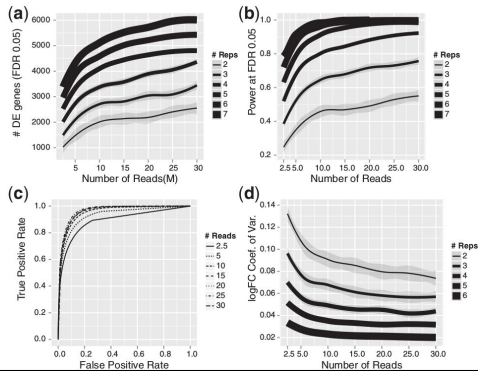
Depth VS Replicates

- Encode (2016) : https://www.encodeproject.org/documents/cede0cbe-d324-4ce7-ace4-f0c3eddf5972/@@download/attachment/ENCODE%20Best%20Practices%20for%20RNA_v2.pdf
 - Experiments should be performed with **two or more biological replicates**, unless there is a compelling reason why this is impractical or wasteful
 - Replicate concordance: the gene level quantification should have a Spearman correlation of >0.9 between isogenic (same donor) replicates and >0.8 between anisogenic (different donor) replicates.
- Between **30M and 100M reads** per sample depending on the study.
 - evaluate the similarity between the transcriptional profiles of two polyA+ samples ==> modest depths of sequencing.
 - discovery of novel transcribed elements and strong quantification of known transcript isoforms ==> more extensive sequencing.
- Zhang et al. 2014 : From 3 replicates improve DE detection and control false positive rate.

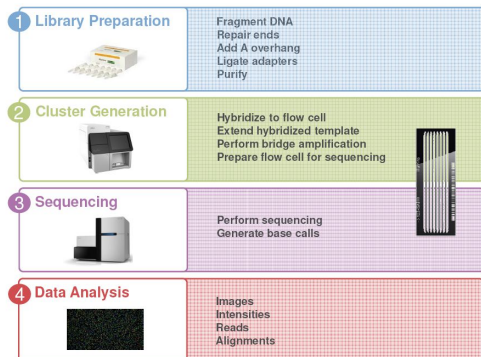
Depth VS Replicates

Gene expression Advance Access publication December 6, 2013
RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu^{1,2}, Jie Zhou^{1,3} and Kevin P. White^{1,2,3,*}



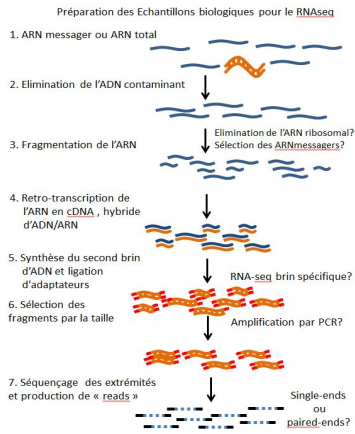
Illumina RNA-Seq protocol



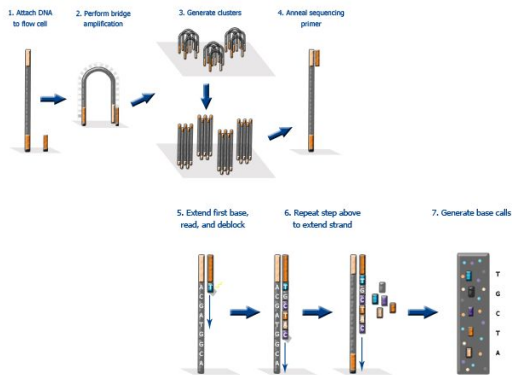
1 Flowcell:

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

RNA-Seq library preparation

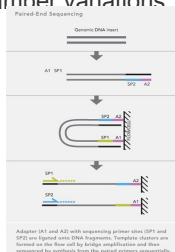


Clusters generation / Sequencing

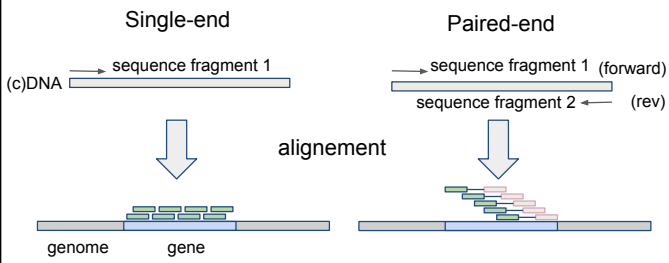


Paired-end sequencing

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



Paired-end VS single-end



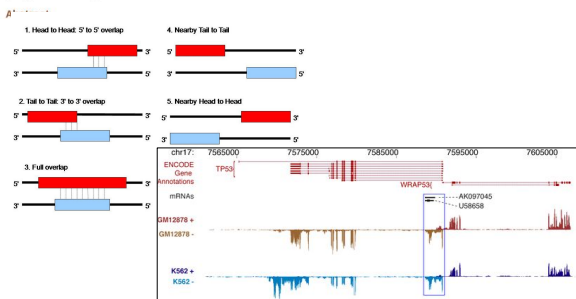
- The cDNA size have the insert size (ex. 200-500 pb).
- The fragment are usually forward-reverse.

Strand specific RNA-Seq protocol

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

Comprehensive comparative analysis of strand-specific RNA sequencing methods.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gmirik A, Regev A. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA. jlevin@broadinstitute.org



Retrieve public data

Why ?

- Because there's a lot of public data that would be sufficient for your analysis
- The authors often use only part of the data to answer their own problems
- Perhaps you don't need your own data

Retrieve public data

EMBL-EBI | ENA European Nucleotide Archive | Services | Research | Training | About us | Aiming: cell coordination breakdown

Home | Search & Browse | Submit & Update | Software | About ENA | Support

ENA > Search and browse

Searching ENA

ENA data can be searched and retrieved interactively and programmatically and visualized using the ENA Browser. Please refer to the following sections for more information about the ENA data access functionality with links to more detailed documentation.

Free text search
Free text search is provided from the search box in the header of all ENA web pages and through the search available at the top of all ENA web pages. Advanced search options are available from the ENA Advanced Search page.

Sequence similarity search
Sequence similarity search is provided from the ENA home page. Advanced search options are available from the ENA Sequence Search page.

Programmatic data access
The main programmatic interface for accessing ENA data is through the ENA Browser. The ENA Browser is designed to be accessed through REST URLs for easy programmatic access to retrieve data and metadata in a variety of formats.

Bulk data download
Most ENA data can be downloaded in bulk through FTP and Aspera protocols ... more information.

Search & Browse

- Data formats
 - Genome assemblies
 - Marker portal
 - Taxon portal
 - Programmatic access
- Data retrieval
 - Taxon portal
 - Marker portal
 - Search
 - File reports
 - XREF service
- Genome assembly database
- Taxonomy Service
- Translation tables
- Download

Retrieve public data

Experiment: ERX1604042 [CONTACT HELP PAGE](#) 41

Illumina HiSeq 2500 paired end sequencing; Root transcriptome profiling in chilling-sensitive tomato (*S. lycopersicon* cv. MoneyMaker) and the more cold-tolerant wild tomato *S. lycopersicon* LA1777 compared at optimal and suboptimal temperature.

View: [XML](#) Download: [XML](#)

Submitting Centre: University of Groningen, Genomics Research in Ecology & Evolution in Nature (GREEN) - Plant Physiology, Groningen Institute for Evolutionary Life Sciences (GELIFES) Platform: ILLUMINA Model: Illumina HiSeq 2500

Library Layout: PAIRED Library Strategy: RNA-Seq Library Source: TRANSCRIPTOMIC Library Selection: cDNA Library Name: Sample 1.p

Navigation: [Read Files](#) [Attributes](#)

This table contains the files for experiment ERX1604042

[Bulk Download Files](#)

Download: 1 - 1 of 1 results in [TEXT](#)

Select columns

Showing results 1 - 1 of 1 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument platform	Library layout	Read count	FASTQ file (FFP)	FASTQ file (Galaxy)	Submitted file (FFP)	Submitted file (Galaxy)	NCBI file (FFP)	NCBI file (Galaxy)	CRAH file (FFP)	CRAH file (Galaxy)
ERR014805	SAMEA079219	ERS1220328	ERX1604042	ERS1533150	62990	Solanum tuberosum	ILLUMINA	PAIRED	19,975,820	File 1 File 2	Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2	File 1	File 1		

Retrieve public data

EMBL-EBI | ENA European Nucleotide Archive | Services | Research | Training | About us | Aiming: cell coordination breakdown

Home | Search & Browse | Submit & Update | Software | About ENA | Support

ENA > Search & Browse > Download > Downloading read data

Downloading read data

Sequencing reads are available for download through FTP and Aspera protocols in their original format and in an archive generated fastq formats described here.

- Submitted data files
- Archive generated fastq files
- Downloading files using FTP
- Downloading files using Globus GridFTP
- Downloading files using ENA Browser
- Downloading files using Aspera

Submitted data files
Submitted data files are organised by submission accession number under vol1/ directory in ftp.sra.ebi.ac.uk: <ftp://ftp.sra.ebi.ac.uk/vol1/<submission accession prefix>/<submission accession>> where <submission accession prefix> contains the first 6 letters and numbers of the SRA Submission accession. For example, the files submitted in the SRA Submission CRA007448 are available at: <ftp://ftp.sra.ebi.ac.uk/vol1/ERA007/ERA007448/>.

Archive generated fastq files
Archive generated fastq files are organised by run accession number under vol1/fastq directory in ftp.sra.ebi.ac.uk: <ftp://ftp.sra.ebi.ac.uk/vol1/fastq/<dir1>/<dir2>/<run accession>> <dir1> is the first 6 letters and numbers of the run accession (e.g. ERR000 for ERR0000916), <dir2> does not exist if the run accession has six digits. For example, fastq files for run ERR0000916 are in

Search & Browse

- Data formats
 - Genome assemblies
 - Marker portal
 - Taxon portal
 - Programmatic access
- Data retrieval
 - Taxon portal
 - Marker portal
 - Search
 - File reports
 - XREF service
- Genome assembly database
- Taxonomy Service
- Translation tables
- Download
 - Sequences
 - Feature level products
 - Reads
 - Taxonomy
 - Sequence search

Retrieve public data

NCBI Sequence Read Archive

Home | Sequence Search | Overview | Documentation | Software | Track Archive | Track Assembly | Track BLAST

Overview

The Sequence Read Archive (SRA) stores raw sequence data from "next-generation" sequencing technologies including Illumina, 454, IonTorrent, Complete Genomics, PacBio and Oxford Nanopore. In addition to raw sequence data, SRA now stores alignment information in the form of read placements as a reference sequence.

SRA is NIH's primary archive of high-throughput sequencing data and is part of the international partnership of archives (HGSC) at the NCBI, the European Bioinformatics Institute and the DNA Database of Japan. Data submitted to any of the three organizations are shared among them.

Please check [SRA Overview](#) for more information.

Submitting to SRA

Making data available to the research community enhances reproducibility and allows for new discovery by comparing data sets.

- Submit your data
- Submit your data
- Submit your data

Using SRA data with SRA ToolKit

Use SRA data to validate experimental results, increase sample sizes, determine variance and open up new avenues of research.

- Documentation
- Usage Guide
- Download
- Get sources code on GitHub (for developers using SRA)



Retrieve public data

NCBI Sequence Read Archive

Home | Sequence Search | Overview | Documentation | Software | Track Archive | Track Assembly | Track BLAST

Studies

Search: tomato

What can be entered in this field?

List of Studies. 421 records found.

#	Accession ID	Title	Project	Center
1	DRS000124	Solanum lycopersicum strain Micro-Tom Genome sequencing and assembly	23722	KAZUSA
2	DRS001059	Resequencing data for tomato 'Ailsa Craig'	231413	KAZUSA
3	DRS001060	Resequencing data for tomato 'Turkoma'	231413	KAZUSA
4	DRS001061	Resequencing data for tomato 'WEE'	231413	KAZUSA
5	DRS001062	Resequencing data for tomato 'Tomato Chukanbohon No. 11'	231413	KAZUSA
6	DRS001063	Resequencing data for tomato 'Yondrusa'	231413	KAZUSA
7	DRS001064	Resequencing data for tomato 'Wagyu'	231413	KAZUSA
8	DRS001065	Tomato genome sequence	259811	TSUKUBA
9	DRS002116	Whole-genome sequencing of tomato mutants	232911	KAZUSA
10	DRS002117	RNA-seq in a sunlight-type plant factory	232911	OSAKA_PREF
11	DRS002118	continuous light tomato RNA-seq	283366	OSAKA_PREF
12	DRS002205	Whole genome shotgun sequencing for 96 tomato cultivars	313263	KAZUSA
13	DRS002206	RNA-seq for tomato	313261	KAZUSA
14	DRS002207	Time course transcriptome data of 5ly-Summer in sunlight-type plant factory	313884	OSAKA_PREF
15	DRS002208	Strategic Innovation Promotion Program	324438	RIKEN_BRC
16	DRS002209	Transcriptome profiling comparison during AM development between L. japonicus and tomato	302053	SHRSHI
17	DRS002210	Carbon nanotubes as fertilizers: effects on tomato growth, reproductive system and soil microbial community	254239	NCTR
18	DRS002211	Defining root colonization strategies in cucumber, tomato, maize and wheat plant species	226214	ARIZOLCAN
19	DRS002212	Bacterial communities associated with the surfaces of fresh fruits and vegetables	205412	CCME-COLORADO
20	DRS002213	Resequencing Solanum (Peato and Tomato) 18th century samples	249937	MPITUEBINGEN
21	DRS002214	Resequencing Physalis peruviana strains	249936	MPITUEBINGEN
22	DRS002215	Using a pericardial chimera to determine layer-specific gene expression	275560	ICL-CFR



Retrieve public data

NCBI SRA Run Selector

Search: DRP002631

Facets: Run, BioSample, Sample name, MReads, MBytes, Equipment, sample name, sample title

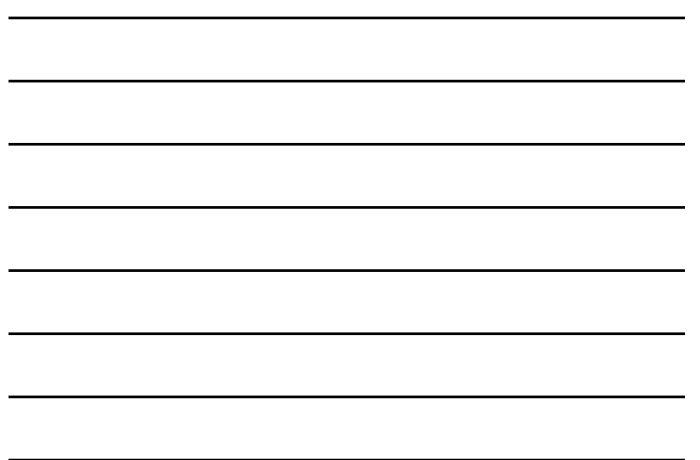
Assay Type: RNA-Seq
 AvgSpotLet: 49
 BioProject: PRJEB3882
 Center Name: OSAKA_PREF
 Consent: public
 InsertSize: 0
 Instrument: Illumina HiSeq 2000
 LibraryLayout: SINGLE
 LibrarySelection: Hybrid Selection
 LibrarySource: TRANSCRIPTOMIC
 LoadDate: 2015-05-01
 Organism: Solanum lycopersicum
 Platform: ILLUMINA
 ReleaseDate: 2015-05-01
 SRA Study: DRP002631
 Subproject ID: PRJEB3882
 Issue type: leaf

Runs: 50 | Bytes: 1.58 Gb | Bases: 2.81 G

Download: RunInfo Table | Accession List

50 Runs found

Run	BioSample	Sample name	Mbases	Mbytes	Experiment	sample name	sample title
DRR034293	SAMD00029631	DRS019544	53	30	DRX030928	SunB30	Sunlight tomato Bset Time30
DRR034296	SAMD00029632	DRS019545	59	34	DRX030927	SunB32	Sunlight tomato Bset Time32
DRR034295	SAMD00029633	DRS019546	76	44	DRX030928	SunB34	Sunlight tomato Bset Time34
DRR034298	SAMD00029634	DRS019547	56	32	DRX030929	SunB36	Sunlight tomato Bset Time36
DRR034296	SAMD00029636	DRS019549	55	32	DRX030931	SunB4	Sunlight tomato Bset Time4
DRR034296	SAMD00029637	DRS019550	70	40	DRX030932	SunB40	Sunlight tomato Bset Time40
DRR034300	SAMD00029638	DRS019551	56	32	DRX030933	SunB42	Sunlight tomato Bset Time42
DRR034301	SAMD00029639	DRS019552	50	29	DRX030934	SunB44	Sunlight tomato Bset Time44
DRR034287	SAMD00029625	DRS019538	61	35	DRX030926	SunB2	Sunlight tomato Bset Time2
DRR034302	SAMD00029640	DRS019553	78	45	DRX030936	SunB46	Sunlight tomato Bset Time46



Retrieve public data

Run	BioSample	Sample name	MBases	MBytes	Experiment	sample name	sample title
DRR034293	SAM00002631	DRS019544	53	30	DRX030926	SunB30	Sunlight tomato Beet Time30
DRR034294	SAM00002632	DRS019545	59	34	DRX030927	SunB32	Sunlight tomato Beet Time32
DRR034295	SAM00002633	DRS019546	76	44	DRX030928	SunB34	Sunlight tomato Beet Time34
DRR034296	SAM00002634	DRS019547	56	32	DRX030929	SunB36	Sunlight tomato Beet Time36
DRR034297	SAM00002635	DRS019548	58	32	DRX030930	SunB40	Sunlight tomato Beet Time40
DRR034298	SAM00002636	DRS019549	70	40	DRX030931	SunB42	Sunlight tomato Beet Time42
DRR034299	SAM00002637	DRS019550	56	32	DRX030932	SunB44	Sunlight tomato Beet Time44
DRR034300	SAM00002638	DRS019551	50	29	DRX030933	SunB46	Sunlight tomato Beet Time46
DRR034301	SAM00002639	DRS019552	81	35	DRX030934	SunB62	Sunlight tomato Beet Time62
DRR034302	SAM00002640	DRS019553	78	40	DRX030935	SunB68	Sunlight tomato Beet Time68

Retrieve public data

NCBI Site map All databases Search

Sequence Read Archive

Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace BLAST

Download Toolkit Documentation XML Schema

SRA Toolkit Documentation

SRA Toolkit Installation and Configuration Guide
Protected Data Usage Guide

Frequently Used Tools:

- fastq-dump:** Convert SRA data into fastq format
- prefetch:** Allows command-line downloading of SRA, dbGaP, and ADSP data
- sam-dump:** Convert SRA data to sam format
- sra-pileup:** Generate pileup statistics on aligned SRA data
- vdb-config:** Display and modify VDB configuration information
- vdb-decrypt:** Decrypt non-SRA dbGaP data ("phenotype data")

Additional Tools:

- abi-dump:** Convert SRA data into ABI format (csfasta / qual)
- illumina-dump:** Convert SRA data into Illumina native formats (qseq, etc.)
- sff-dump:** Convert SRA data to sff format
- sra-stat:** Generate statistics about SRA data (quality distribution, etc.)
- vdb-dump:** Output the native VDB format of SRA data.
- vdb-encrypt:** Encrypt non-SRA dbGaP data ("phenotype data")
- vdb-validate:** Validate the integrity of downloaded SRA data

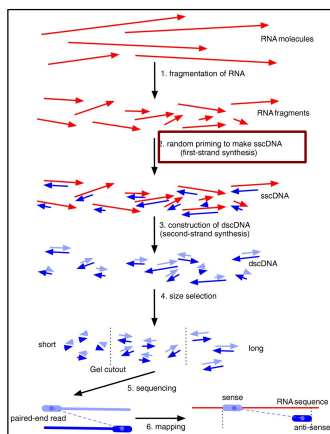
Summary - Sequence quality

- Known RNAseq biases
- How to check the quality ?
- How to clean the data ?

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



Hexamer random priming bias

Published online 11 April 2010
Nucleic Acids Research, 2010, Vol. 38, No. 12, e111
 doi:10.1093/nar/gkq1254

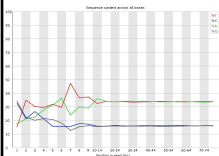
Biases in Illumina transcriptome sequencing caused by random hexamer priming

Kasper D. Hansen^{1*}, Steven E. Brenner² and Sandrine Dudot^{1,3}

ABSTRACT

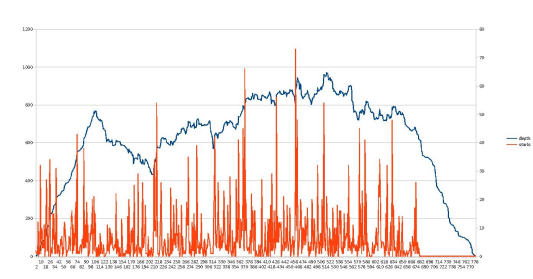
Generation of cDNA using random hexamer priming induces biases in the nucleotide composition at the beginning of transcriptome sequencing reads from the Illumina Genome Analyzer. The bias is independent of organism and laboratory and impacts the uniformity of the reads along the transcriptome. We provide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

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



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

Hexamer random effect



- Orange = reads start sites
- Blue = coverage

Transcript length bias

BMC Bioinformatics 2009, 10:414

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

Qin et al., [Wiley InterScience](#)

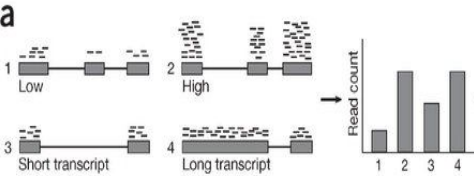
Abstract

Background: Several recent transcriptome analysis (RNA-seq) methods profile genomic sequences. As yet, it is still in the stages of exploring the full potential of these methods.

Results: We investigated the published data sets. For some call differentially expressed genes.

Conclusions: Transcript length current protocols for RNA-seq expressed genes, and in particular other multi-gene systems bio.

Reviewers: This article was reviewed by Mark



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

BIOINFORMATICS ORIGINAL PAPER Vol. 10, No. 4, 2009, pages 414-419
doi:10.1093/bioinformatics/btp414
Gene expression Advances in genome biology January 19, 2011
Length bias correction for RNA-seq data in gene set analyses
Liyun Gao^{1,2}, Zhide Fang^{1,2}, Kui Zhang¹, Degui Zhu¹ and Xiangjin Cui^{1,2*}

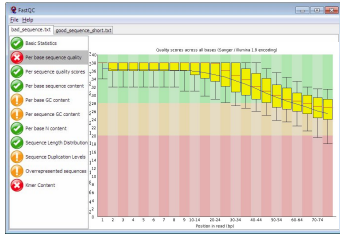
Bias “mappability”

- Quality of the reference genome influence results
 - assembly
 - finishing
- Sequence composition
- Repeated sequences
- Annotation quality

Verifying RNA-Seq quality

FastQC :

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



Has been developed for genomic data

fastQC Report

Summary

- ✔ Basic Statistics
- ✘ Per base sequence quality
- ✔ Per sequence quality scores
- ✘ Per base sequence content
- ✘ Per base GC content
- ✘ Per sequence GC content
- ! Per base N content
- ✔ Sequence Length Distribution
- ✘ Sequence Duplication Levels
- ✘ Overrepresented sequences
- ✘ Kmer Content

The analysis in FastQC is performed by a series of analysis modules.

Quick evaluation of whether the results of the module seem :

- entirely normal (green tick),
- slightly abnormal (orange triangle)
- or very unusual (red cross).

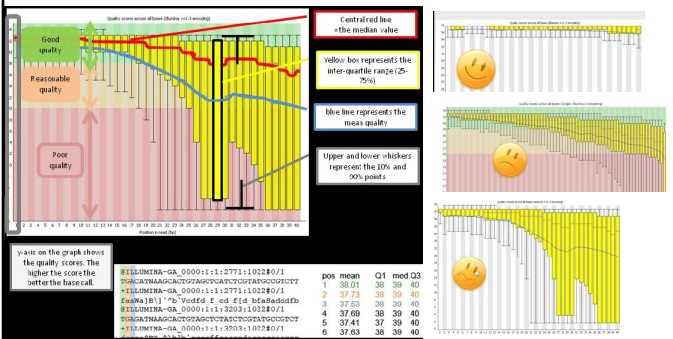
These evaluations must be taken in the context of what you expect from your library. A 'normal' sample as far as FastQC is concerned is random and diverse.

fastQC Report

Statistics per Base Sequence Quality

This view shows an overview of the range of quality values across all bases at each position in the FastQ file.

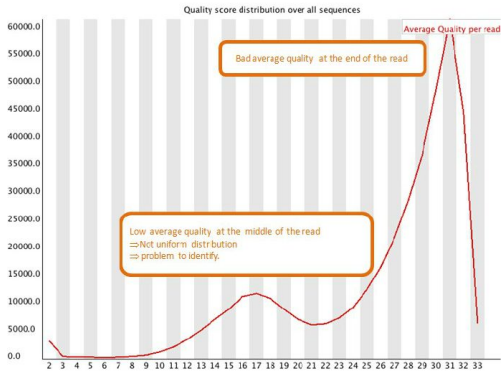
Common to see base calls falling into the orange area towards the end of a read.



fastqQC Report

Statistics per Sequence Quality Score

See if a subset of your sequences have universally low quality values.

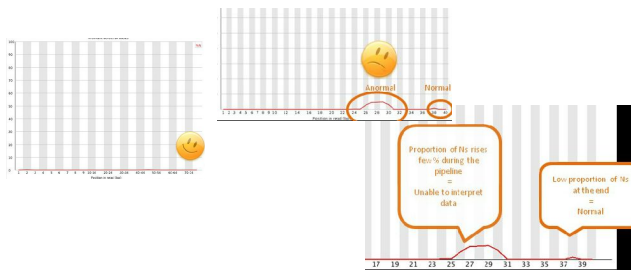


fastqQC Report

Statistics per Base N Content

This module plots out the percentage of base calls at each position for which an N was called.

Usual to see a very low proportion of Ns appearing nearer the end of a sequence.



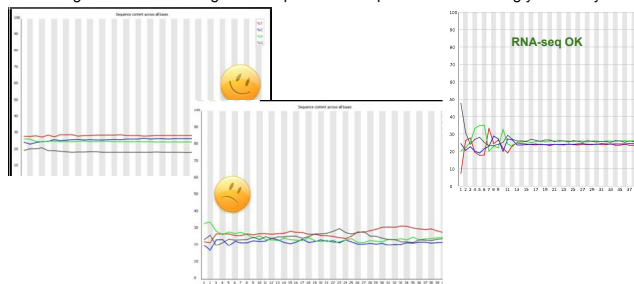
fastqQC Report

Statistics Per Base Sequence Content

Per Base Sequence Content plots out the proportion of each base position in a file for which each of the four normal DNA bases has been called.

In a random library : little/no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other.

If strong biases which change : overrepresented sequence contaminating your library.



fastqQC Report

Statistics per Base GC Distribution

Per Base GC Content plots out the GC content of each base position in a file.

Random library : little/no difference between the different bases of a sequence run
=> plot horizontally.

The overall GC content should reflect the GC content of the underlying genome.

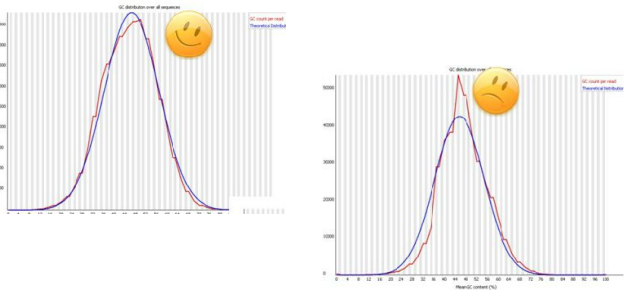
GC bias: changes in different bases, overrepresented sequence contaminating your library.
=> plot not horizontally.



fastqQC Report

Statistics per Sequence GC Content

This module measures the GC content across the whole length of each sequence in a file and compares it to a modeled normal distribution of GC content.

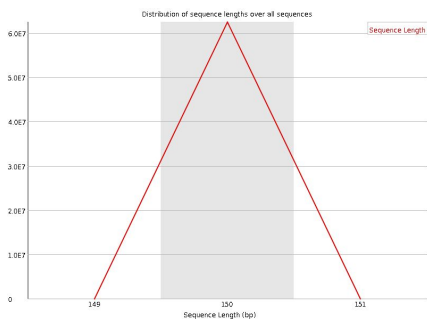


fastqQC Report

Statistics per Sequence Length Distribution

Some sequence fragments contain reads of wildly varying lengths.

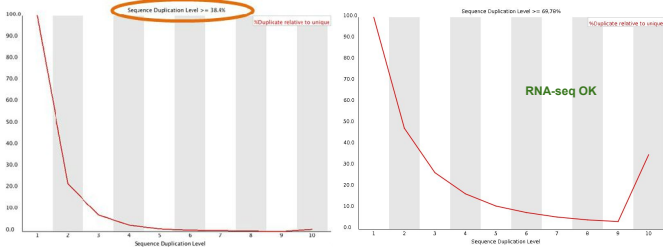
Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end.



fastqQC Report

Statistics per Duplicate Sequences

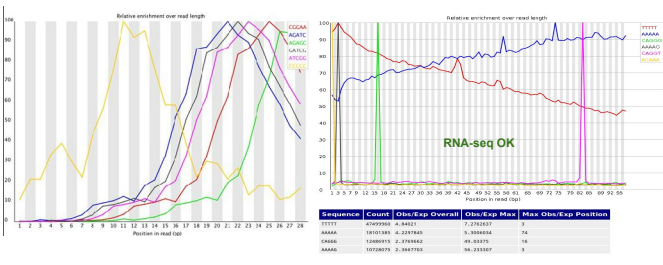
High level of duplication indicate an enrichment bias.



fastqQC Report

Overrepresented Kmers

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



Take home message on quality analysis

Elements to be checked :

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

Alignment on reference for the second quality check and filtering.

A good run?:

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

Cleaning analysis

- Cleaning :
 - Low quality bases
 - Adaptors
- Software :
 - Trim_galore
 - Cutadapt
 - Trimmomatic
 - Sickle
 - PRINSEQ
 - ...

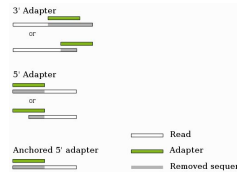
Cutadapt

- Searches & removes adapter & tag in all reads.
- Trim quality
- Filter too short or untrimmed reads (in a separate output file).

cutadapt -a ADAPTER [options] [-o output.fastq] input.fastq

Ex.: cutadapt -a AACCGGTT -o output.fastq input.fastq

Input file : fasta, fastq or compressed (gz, bz2, xz).



Source : <http://cutadapt.readthedocs.io/en/stable/guide.html>

Cutadapt

Cutadapt supports trimming of paired-end reads, trimming both reads in a pair at the same time.

Processing both files at the same time is highly recommended.

```
cutadapt -a ADAPTER_FWD -A ADAPTER_REV -o out.1.fastq -p out.2.fastq reads.1.fastq reads.2.fastq
```

```
Paired-end options:
The -A/-G/-B/-U options work like their -a/-b/-g/-u counterparts.
-A ADAPTER      3' adapter to be removed from the second read in a
                pair.
-G ADAPTER      5' adapter to be removed from the second read in a
                pair.
-B ADAPTER      5'/3' adapter to be removed from the second read in a
                pair.
-U LENGTH       Remove LENGTH bases from the beginning or end of each
                read (see --cut).
-p FILE, --paired-output=FILE
                Write second read in a pair to FILE.
--untrimmed-paired-output=FILE
                Write the second read in a pair to this FILE when no
                adapter was found in the first read. Use this option
                together with --untrimmed-output when trimming paired-
                end reads. (Default: output to same file as trimmed
                reads.)
```

Hands-on: quality control

Data for the exercises:

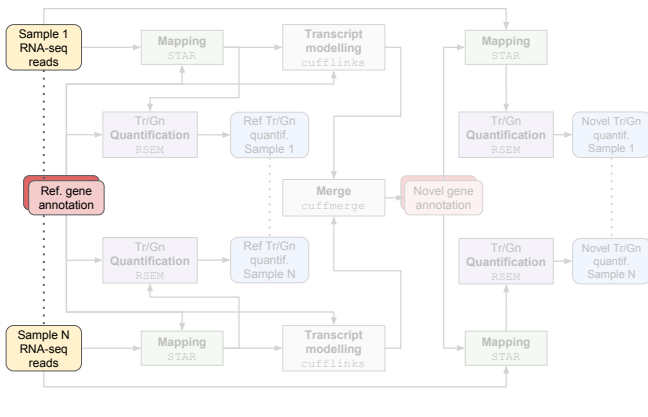
- from Mohammed Zouine (ENSAT)
- tomato wild type and mutant type (without seeds) with the transcription factor SI-ARF8 (auxine response factor 8) overexpressed
- clonal lineage
- paired, 100 pb non stranded
- triplicated
- in the publication process
- subsampled on chromosome 6 for faster analysis

Use FastQC and cutadapt

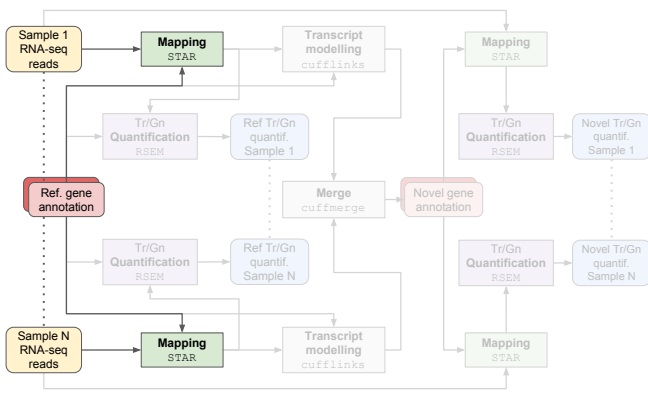
Exercise 3 : quality control of used datasets

Exercise 4: cleaning used datasets

Analysis workflow



Analysis workflow



Summary -

Spliced read mapping & Visualisation

1. What is a spliced aligner?
2. Reference genome & transcriptome files formats
3. Tophat principle
4. STAR principle and usage
5. BAM & Bed files formats
6. Visualisation with IGV

Aim -

Spliced read mapping & Visualisation

Aim: Discover the true location (origin) of each read on the reference.

Problems:

- Some features (repetitive regions, assembly errors, missing information) make it impossible for some reads.
- Reads may be split by potentially thousands of bases of intronic sequence.

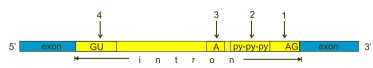


And:

Do it in/with reasonable time/resources.

Splice sites

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



- Non-canonical site:
 - GC-AG splice site pairs, AT-AC pairs
 - Trans-splicing: Nucleic Acids Res. 2000 Nov 12;28(21):4364-75. Burset M, Solodov IA, Solovnev VV. splicing that joins two exons that are not within the same RNA transcript

Hard case

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



- Small end "anchor"



- Unknown junction inside gene



Most used tools

Tools for splice-mapping:

- Tophat:

BIOINFORMATICS ORIGINAL PAPER Vol. 12, No. 8, 2008, pages 1165-1171
doi:10.1093/bioinformatics/btn184

Sequence analysis

TopHat: discovering splice junctions with RNA-Seq

Cole Trapnell¹*, Lior Pachter² and Steven L. Salzberg¹

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

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

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

- STAR:

STAR: ultrafast universal RNA-seq aligner

Alexander Dobin¹, Carrie A. Davis¹, Felix Schlesinger¹, Jorg Drenkow¹, Chris Zaleski¹, Sonali Jha¹, Philippe Batut¹, Mark Chaisson² and Thomas R. Gingeras¹

¹Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA.

²Genetic Bioinformatics, Merck Park, California, USA.

Associate Editor: Dr. Huan Tang

Comparing tools

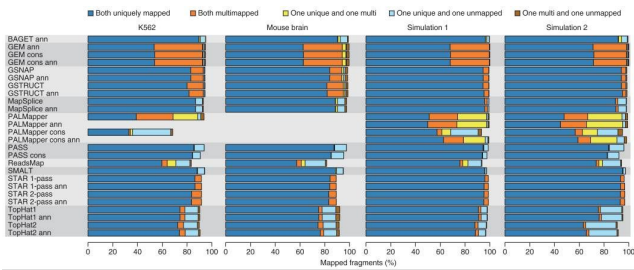
How to compare tools ?

- sensibility (maximize #mapped reads)
 - specificity (assign reads to the correct position)
- for reads and for junctions
- processing time
 - memory requirement

All of these are conflicting criteria ...

RGASP3

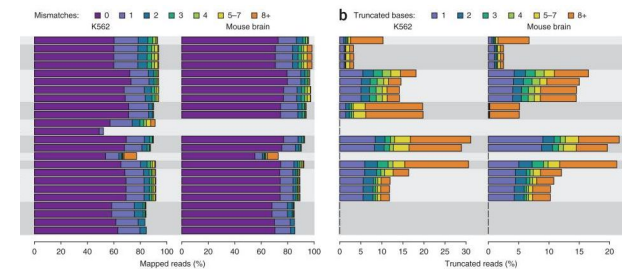
The RNA-seq Genome Annotation Assessment Project



Engström et al., Nature Methods, 2013

RGASP3

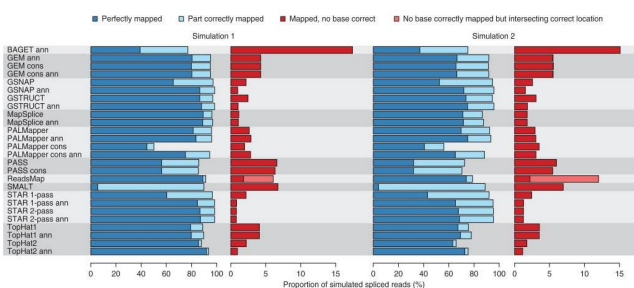
The RNA-seq Genome Annotation Assessment Project



Engström et al., Nature Methods, 2013

RGASP3

The RNA-seq Genome Annotation Assessment Project



Engström et al., Nature Methods, 2013

Other benchmark

Basically similar conclusions...

NATURE METHODS | ANALYSIS



Simulation-based comprehensive benchmarking of RNA-seq aligners

Giacomo Baruzzo, Katharina E Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A FitzGerald & Gregory R Grant

Affiliations | Contributions | Corresponding author

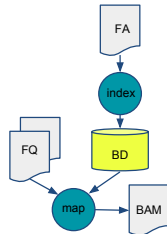
Nature Methods 14, 135–139 (2017) | doi:10.1038/nmeth.4106

Received 18 April 2016 | Accepted 15 November 2016 | Published online 12 December 2016

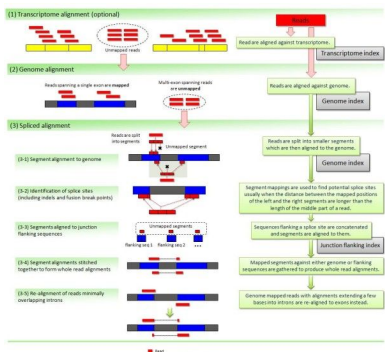
Corrected online 22 December 2016

Mapping steps

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



TopHat pipeline



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

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

Kim et al, Genome Biology, 2013

An other aligner : STAR



Bioinformatics, 2013, Jan, 29(1): 15-21
Published online 2012 Oct 25, doi: 10.1093/bioinformatics/btq535

PMCID: PMC3530905

STAR: ultrafast universal RNA-seq aligner

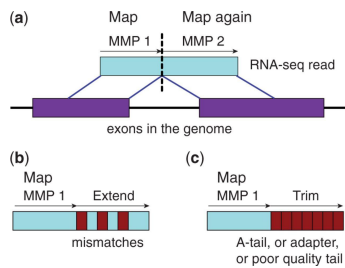
Alexander Dobin,^{1,2} Carrie A. Davis,¹ Felix Schlesinger,¹ Jorg Drenkow,¹ Chris Zaleski,¹ Sonali Jha,¹ Philippe Batut,¹ Mark Chaisson,² and Thomas R. Gingeras¹

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

STAR

Another strategy:

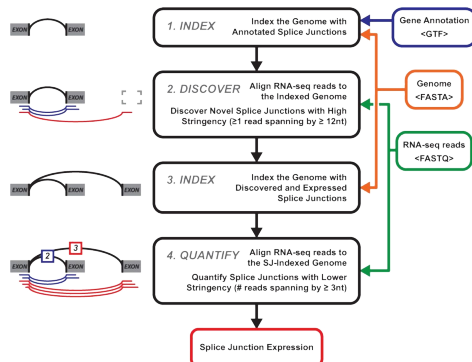
- search for a MMP from the 1st base
- MMP search repeated for the unmapped portion next to the junction
- do it in both fwd and rev directions
- cluster seeds from the mates of paired-end RNA-seq reads



Soft-clipping is the main difference between Tophat and STAR

Dobin et al, Bioinformatics, 2011

Two passes strategy



Veeneman et al, Bioinformatics, 2016

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

STAR indexing

Hands-on: Type STAR and count the number of options.

“Core” command:

```
STAR --runMode genomeGenerate --genomeDir  
genome_dir --genomeFastaFiles genome.fasta
```

To use *N* CPUs, add: `--runThreadN N`

If you have an annotation: `--sjdbGTFfile annot.gtf`

Some precomputed indices are already available:

<http://labshare.cshl.edu/shares/gingeraslab/www-data/bin/STAR/STARgenomes>


or on your preferred platform: `/bank/STARdb`

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 ?

- An annotation file: loci of coding genes (transcripts, CDS, UTRs), non-coding genes, etc.
- Gene Transfer Format (derived from GFF):


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

```
chr source feature start end score strand frame [attributes]
1 ENSEMBL exon 3000 2000 . + . gene_id "ENSG01"; transcript_id "ENST01.1"; gene_name "ABC";
1 ENSEMBL exon 3000 4000 . + . gene_id "ENSG01"; transcript_id "ENST01.1"; gene_name "ABC";
1 ENSEMBL exon 3000 4000 . + . gene_id "ENSG01"; transcript_id "ENST01.2"; gene_name "ABC";
1 ENSEMBL exon 5000 6000 . + . gene_id "ENSG02"; transcript_id "ENST02.1"; gene_name "DEF";
```



- `gene_id value` : unique identifier for the gene.

- `transcript_id value` : unique identifier for the transcript.

 **The chromosome names should be the same in the gtf file and fasta files (e.g. chr1 vs Chr1 vs 1).**

Hands-on : STAR

Exercice n°2 A/

Create a directory for the genome and annotation files.

Get the FASTA and GTF files from:

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

Create the STAR index.

Tip: you can allocate *N* CPUs with the qsub/qrsh option
-pe parallel_smp *N*

STAR mapping

“Core” command:

```
STAR --genomeDir genome_dir --readFilesIn  
reads1.fastq reads2.fastq [--sjdbGTFfile  
annot.gtf --runThreadN n]
```

If the read files are gzipped (*reads1.fq.gz*):

```
--readFilesCommand zcat
```

Intron options: genomic gap is considered intron if

```
--alignIntronMin [21]
```

```
--alignIntronMax [500000]
```

Max. number of mismatches:

```
--outFilterMismatchNmax [10]
```

Default options are probably tuned for mammalian genomes.

SAM / BAM formats

Sequence Alignment/Map format:

- Each line stores an alignment/map

```
Coord 12345678901234 5678901234567890123456789012345  
ref AGCATGTTAGATAA*GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
```

```
+r001/1 TTAGATAAAGGATA*CTG  
+r002 aaAGATAA*GGATA  
+r003 gcctaAGCTAA  
+r004 ATAGCT.....TCAGC  
-r003 ttagctTAGGC  
-r001/2 CAGCGGCAT
```

name	flag	chr	start	mapQ	cigar	nNext	sNext	tlen	seq	qual	tags
r001	99	ref	7	30	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*	
r002	0	ref	9	30	35G4M1P114M	*	0	0	AAAGATAAGGATA	*	
r003	0	ref	9	30	55GM	*	0	0	GCCTAAGCTAA	*	SA:Z:ref,29,-,6HSM,17,0;
r004	0	ref	16	30	6M14NSM	*	0	0	ATAGCTTCAGC	*	
r003	2064	ref	29	17	6HSM	*	0	0	TAGGC	*	SA:Z:ref,9,+,556M,30,1;
r001	147	ref	37	30	9M	=	7	-39	CAGCGGCAT	*	NM:i:1

- Header stores genome information

```
@HD VN:1.5 SO:coordinate  
@SQ SN:ref LN:45
```

Fields

```
Coord 12345678901234 5678901234567890123456789012345
ref AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
```

```
+r001/1 TTAGATAAAGGATA*CTG
+r002 aaaAGATAA*GGATA
+r003 gcctaAGCTAA
+r004 ATAGCT.....TCAGC
-r003 ttagctTAGGC
-r001/2 CAGCGGCAT
```

name	flag	chr	start	mapQ	cigar	nNext	sNext	tlen	seq	qual	tags
r001	99	ref	7	30	8N2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*	
r002	0	ref	9	30	356M1P114M	*	0	0	AAAAGATAAGGATA	*	
r003	0	ref	9	30	556M	*	0	0	GCCTAAGCTAA	*	SA:Z:ref,29,-,6H5M,17,0;
r004	0	ref	16	30	6M14N5M	*	0	0	ATAGCTTCAGC	*	
r003	2064	ref	29	17	6H5M	*	0	0	TAGGC	*	SA:Z:ref,9,+,556M,30,1;
r001	147	ref	37	30	9M	=	7	-39	CAGCGGCAT	*	NM:i:1

- Flags: <https://broadinstitute.github.io/picard/explain-flags.html>
- MapQ: similar to a phred score
- nNext: = means same chr
- In general, * means NA

CIGAR

```
Coord 12345678901234 5678901234567890123456789012345
ref AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
```

```
+r001/1 TTAGATAAAGGATA*CTG
+r002 aaaAGATAA*GGATA
+r003 gcctaAGCTAA
+r004 ATAGCT.....TCAGC
-r003 ttagctTAGGC
-r001/2 CAGCGGCAT
```

name	flag	chr	start	mapQ	cigar	nNext	sNext	tlen	seq	qual	tags
r001	99	ref	7	30	8N2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*	
r002	0	ref	9	30	356M1P114M	*	0	0	AAAAGATAAGGATA	*	
r003	0	ref	9	30	556M	*	0	0	GCCTAAGCTAA	*	SA:Z:ref,29,-,6H5M,17,0;
r004	0	ref	16	30	6M14N5M	*	0	0	ATAGCTTCAGC	*	
r003	2064	ref	29	17	6H5M	*	0	0	TAGGC	*	SA:Z:ref,9,+,556M,30,1;
r001	147	ref	37	30	9M	=	7	-39	CAGCGGCAT	*	NM:i:1

- 30M means 30 matches or mismatches
- I and D: insertion/deletion
- S and H: soft/hard clipping

Tags

```
Coord 12345678901234 5678901234567890123456789012345
ref AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT
```

```
+r001/1 TTAGATAAAGGATA*CTG
+r002 aaaAGATAA*GGATA
+r003 gcctaAGCTAA
+r004 ATAGCT.....TCAGC
-r003 ttagctTAGGC
-r001/2 CAGCGGCAT
```

name	flag	chr	start	mapQ	cigar	nNext	sNext	tlen	seq	qual	tags
r001	99	ref	7	30	8N2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*	
r002	0	ref	9	30	356M1P114M	*	0	0	AAAAGATAAGGATA	*	
r003	0	ref	9	30	556M	*	0	0	GCCTAAGCTAA	*	SA:Z:ref,29,-,6H5M,17,0;
r004	0	ref	16	30	6M14N5M	*	0	0	ATAGCTTCAGC	*	
r003	2064	ref	29	17	6H5M	*	0	0	TAGGC	*	SA:Z:ref,9,+,556M,30,1;
r001	147	ref	37	30	9M	=	7	-39	CAGCGGCAT	*	NM:i:1

- Format: *2-Letter name:format:value* (many different)
- NM: # mismatches
- SA: chimeric reads
- NH, HI: # hits for this sequence, hit index
- AS: alignment score
- nM: # mismatches per fragment

SAM / BAM

BAM (Binary Alignment/Map) format:

- Compressed binary representation of SAM
- Greatly reduces storage space requirements to about 27% of original SAM
- samtools: reading, writing, and manipulating BAM files
- Most tools require a sorted and indexed BAM file.

STAR output options

Output format:

`--outSAMtype BAM SortedByCoordinate [SAM]`

Add more tags:

`--outSAMattributes All`

Default output file name: `Aligned.bam` Modify prefix:

`--outFileNamePrefix prefix`

Infer strand using intron motifs (for Cufflinks)

`--outSAMstrandField intronMotif [None]`

Start IH at `--outSAMattrIHstart 0 [1]` (for Cufflinks)

STAR other options

Remove reads that did not pass the junction filter:

`--outFilterType BySJOut [Normal]`

Filter out alignments with non-canonical intron motifs

`--outFilterIntronMotifs RemoveNoncanonical`

Output SAM/BAM alignments to transcriptome into a separate file (for RSEM)

`--quantMode TranscriptomeSAM`

Two passes mode:

- STAR is run once and discover new junctions.
- STAR is run again, knowing the new junctions. (Probably most useful for poorly annotated genomes.)

STAR Outputs

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

- `Aligned.sortedByCoord.out.bam`: list of read alignments in SAM format compressed
- `Log.out`: main log file with a lot of detailed information about the run (for troubleshooting)
- `Log.progress.out`: reports job progress statistics
- `Log.final.out`: summary mapping statistics after mapping job is complete, very useful for quality control.
- `SJ.out.tab`: contains high confidence collapsed splice junctions in tab-delimited format
(chr, intron start, end, strand, intron motif, in database, # uniquely mapping reads, # multi, max. overhang)

STAR technical issues

- Temporary disk space:
 - Indexing the mouse genome requires 128GB and 1 hour on 6 slots.
 - Mapping a 16M paired-end reads requires 110GB and 4 mins on 6 slots.
- 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 bandwidth)

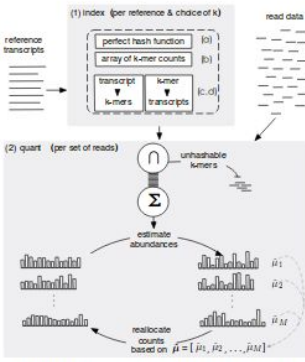
Hands-on : STAR

Exercice n°2 B/
Map the 2 FASTQ files.
Do not forget to provide a different output file name for each set.

Index the output BAM files with:
`samtools index file.bam`

Get some stats with:
`samtools flagstat file.bam`

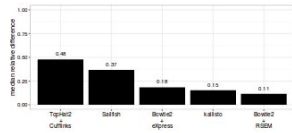
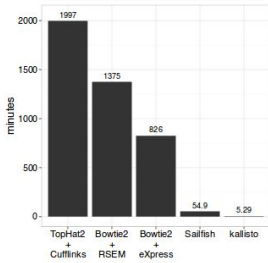
Quasi-mapping: Sailfish



Patro *et al*, Nat. Biotech., 2014

- Reads are *not* mapped.
- Transcriptome is cut into small chunks of small k -mers.
- Same for reads.
- Take a k -mer from a transcript, counts how many times you find it in reads.
- “Average” the counts over a transcript.
- Resolve ambiguous counts.

Quasi-mapping: why?



Bray *et al*, Nat. Biotech., 2016

Other (most used) tool: kallisto, salmon

Quasi-mapping: limitations

Heavily relies on a good annotation:

- Unannotated genes will not be counted and may bias other genes counts.

Does not align reads:

- Cannot find variation (SNP) in the reads.

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

Step 1: set the genome

- Exercice n°2 C/
- Open the Genomes menu
- Choose Load Genome from File...
- Provide your FASTA file.

Some updated fields:

- Genome
- Chromosome
- Locus

Tips:

- Some chromosomes are bundled with IGV (but they should have the same chromosome names).
- You can fetch some others through the server.

Step 2: add the tracks

- Open the File menu
- Choose Load from File...
- Provide your GTF file.
- Provide your BAM files (the BAI file should be also present).

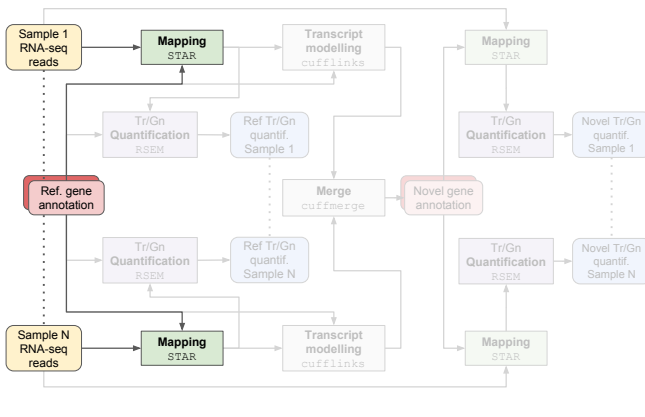
Some interesting loci:

- Go to locus: SL2.40ch06:34,298,666-34,306,292
- Thin lines indicate introns. Notice that gene introns match with read introns. Notice that the first and last exons seems longer than annotation. It's probably not annotated UTR.

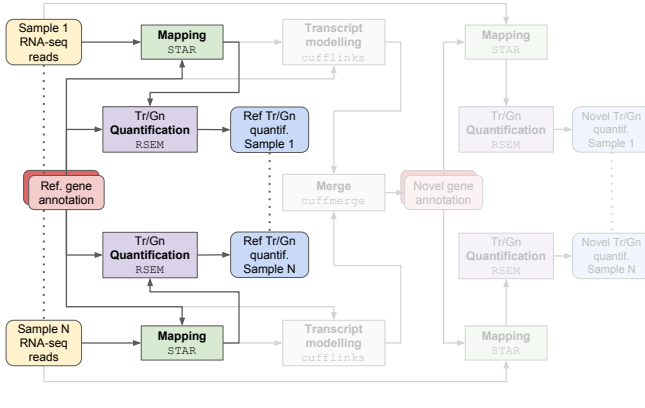
Explore IGV

- Zoom in/out
 - Go right/left
 - Hover over the reads and get some info.
 - Notice (colored) genome variations.
 - Change panel height.
 - Go to next TSS with Ctrl+F (Ctrl+B for previous TSS)
-
- Go to SL2.40ch06:34,209,900-34,260,000
 - Look at the strand of the gene.
 - Expand the gene track.
 - Do you think the annotation is complete here?
 - Which condition is more expressed?

Analysis workflow

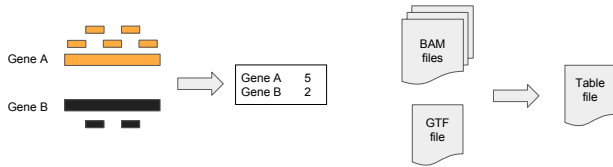


Analysis workflow



Quantification

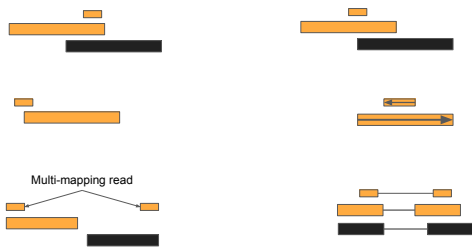
Quantification: estimation of expression based on a read count.



Estimation of:

- gene expression
- transcript expression
- exon expression

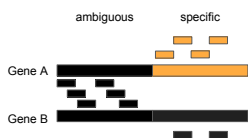
Difficult cases



Every quantification tools uses its own rules!

Raw counts vs estimation

Raw count vs estimation: what to do with ambiguous reads?



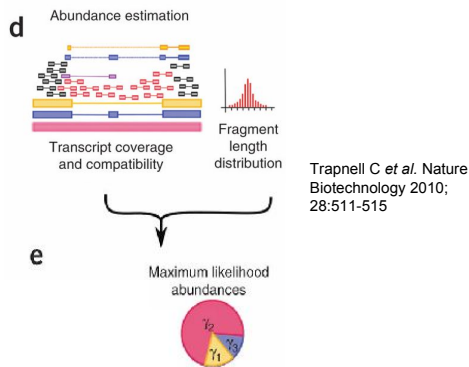
Pros estimation:

- Use more reads.
- More accurate?

Cons estimation:

- Underlying model inaccurate.
- Raw counts for differential expression does not matter much.

Transcript expression



Raw counts tool: featureCounts

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

Yang Liao^{1,2}, Gordon K. Smyth^{1,3} and Wei Shi^{1,2,*}

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

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

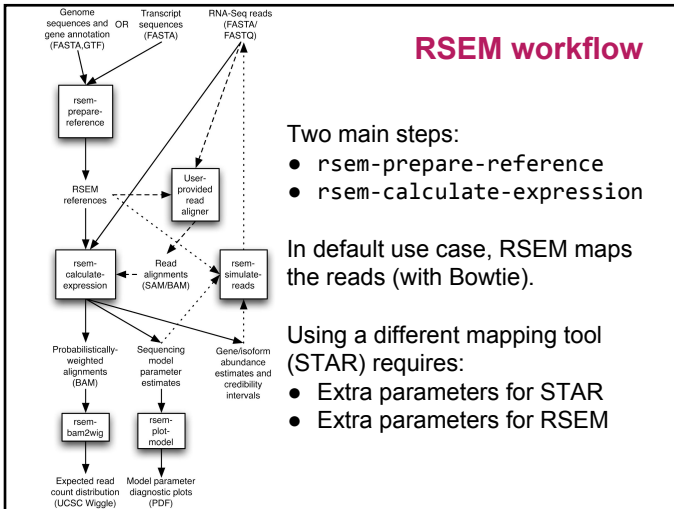
Estimation tool: RSEM

RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome

Bo Li and Colin N Dewey

BMC Bioinformatics 2011 12:323 | DOI: 10.1186/1471-2105-12-323 | © Li and Dewey; licensee BioMed Central Ltd. 2011
Received: 10 May 2011 | Accepted: 4 August 2011 | Published: 4 August 2011

- Exhaustive tool
- Levels : transcript, gene
- May be used without reference genome (RNA-Seq *de novo*)
- Also exists: cufflinks



Hands-in: prepare reference

Exercise n°3
 Command line:
`/usr/local/bioinfo/src/RSEM/RSEM-1.3.0/rsem-p
 prepare-reference --gtf annot.gtf genome.fasta
rsem_lib`

Output files:

- *rsem_lib.grp*, *rsem_lib.ti*, *rsem_lib.seq*, and *rsem_lib.chrlist* are for internal use.
- *rsem_lib.idx.fa*: the transcript sequences
- *rsem_lib.n2g.idx.fa*: same, with N→G

Hands-in: calculate expression

Command line:
`/usr/local/bioinfo/src/RSEM/RSEM-1.3.0/rsem-c
 alculate-expression --alignments
alignment.bam rsem_lib quant`

Outputs:

- *quant.isoforms.results*: isoform level expression estimates
- *quant.genes.results*: same for genes
- *quant.stat*: directory with stats on various aspects of this step

Hands-in: calculate expression

Other parameters:

- `--paired-end`: specify paired-end reads
- `-p N`: use N CPUs
- `--seed N`: seed for random number generators
- `--calc-ci`: calculate 95% credibility intervals and posterior mean estimates.
- `--ci-memory 30000`: size in MB of the buffer used for computing CIs
- `--estimate-rspd`: estimate the read start position distribution
- `--no-bam-output`: do not output any BAM file (produced by internal mapper)

Output file format

- `effective_length`: # positions that can generate a fragment
- `expected_count`: read count, with mapping prob. and read qual
- TPM: Transcripts Per Million, relative transcript abundance, see *infra*
- FPKM: Fragments Per Kilobase of transcript per Million mapped reads, see *infra*
- IsoPct: isoform percentage
- `posterior_mean_count`, `posterior_standard_deviation_of_count`, `pme_TPM`, `pme_FPKM`: estimates calculated Gibbs sampler

Output file format

- `IsoPct_from_pme_TPM`: isoform percentage calculated from `pme_TPM` values
- `TPM_ci_lower_bound`, `TPM_ci_upper_bound`, `FPKM_ci_lower_bound`, `FPKM_ci_upper_bound`: bounds of 95% credibility intervals
- `TPM_coefficient_of_quartile_variation`, `RPKM_coefficient_of_quartile_variation`: coefficients of quartile variation, a robust way of measuring the ratio between the standard deviation and the mean

RPKM vs FPKM vs TPM

RPKM: Reads Per Kb of transcript per Million mapped

- r = # reads on a gene
- k = size of the gene (in kb)
- m = # reads in the sample (in millions)
- $RPKM = r / (k m)$

FPKM: Fragments Per Kilobase...

- Same with f = # fragments (2 reads in PE) on a gene

Meaning:

If you sequence at depth 10^6 , you will have x = FPKM fragments of a 1kb-gene.

RPKM vs FPKM vs TPM

TMP:

- r_i = # reads on a gene i
- s_i = size of the gene i
- $cpb_i = r_i / s_i$
- $cpb = \sum cpb_i$
- $TMP_i = cpb_i / cpb \times 10^6$

Remark:

- $TMP_i = FPKM_i / (\sum FPKM_j) \times 10^6$

Meaning:

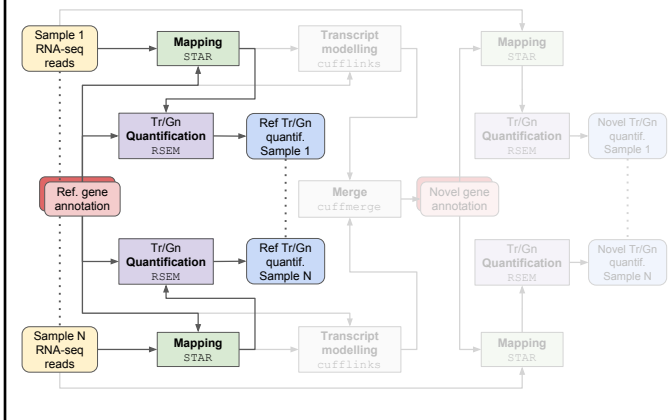
If you have 10^6 transcripts, $x = TMP_i$ will originate from gene i .

RPKM vs FPKM vs TPM

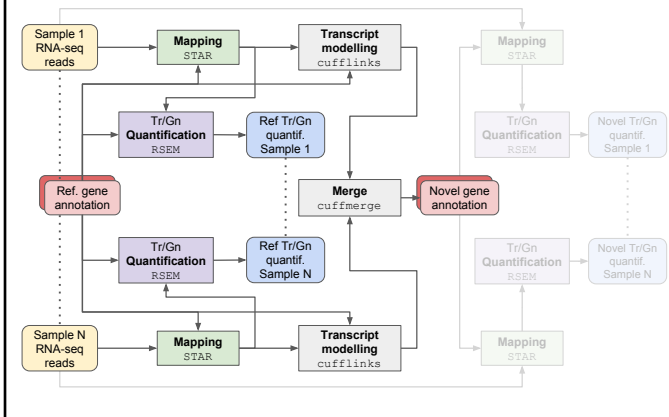
- These are refinement of library size normalization, with gene length effect.
- RPKM should not be used for PE reads.
- TPM tend to be favored now w.r.t. R/FPKM.
- None of them should be used for differential expression: only raw counts.

Ask your questions to the stats guys.

Analysis workflow



Analysis workflow



New transcriptome: why?

Ensembl Release 88 (March 2017)

Homo sapiens

Coding genes	20,310 (incl 656 readthrough)
Non coding genes	22,529
Pseudogenes	14,589 (incl 6 readthrough)
Gene transcripts	199,234

Mus musculus

Coding genes	22,615 (incl 226 readthrough)
Non coding genes	14,299
Pseudogenes	10,937 (incl 6 readthrough)
Gene transcripts	125,665

Rattus norvegicus

Coding genes	22,250 (incl 12 readthrough)
Non coding genes	8,934
Pseudogenes	1,668
Gene transcripts	41,078

Bos taurus

Coding genes	19,994
Non coding genes	3,825
Pseudogenes	797
Gene transcripts	26,740

Oryctolagus cuniculus

Coding genes	19,293
Non coding genes	3,375
Pseudogenes	1,001
Gene transcripts	24,984

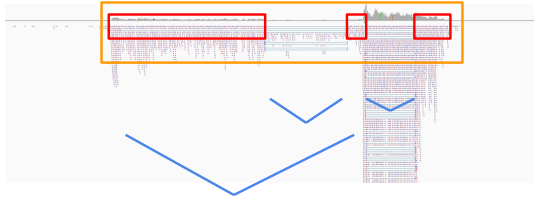
Sus scrofa

Coding genes	21,630 (incl 10 readthrough)
Non coding genes	3,124
Pseudogenes	568
Gene transcripts	30,585

Gallus gallus

Coding genes	18,346
Non coding genes	6,492
Pseudogenes	43
Gene transcripts	38,118

Transcript reconstruction



Gene location



Exon location



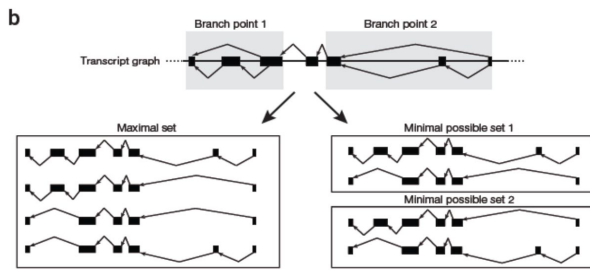
Junctions :

- between read pair junction



- within read junction

Model building strategies



Computational methods for transcriptome annotation and quantification using RNA-seq

Mannuel Garber¹, Manfred G Grabherr¹, Mitchell Guttman^{1,2} & Cole Trapnell^{1,2}

REVIEW

Cufflinks

NATURE BIOTECHNOLOGY | RESEARCH | LETTER

日本語要約

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

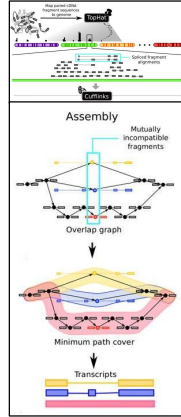
Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

<http://cole-trapnell-lab.github.io/cufflinks/>

- assembles transcripts
- estimates their abundances: based on how many reads support each one
- last version: cufflinks 2.2.1, released May 05, 2014

Cufflinks transcript assembly

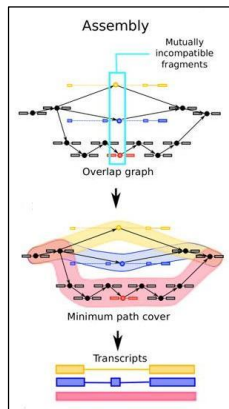
- Transcripts assembly:
 - fragments are divided into non-overlapping loci
 - each locus is assembled independently
- Cufflinks assembler
 - find the mini nb of transcripts that explain the reads
 - find a minimum path cover (Dilworth's theorem):
 - nb incompatible read = mini nb of transcripts needed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



Trapnell C et al. Nature Biotechnology 2010

Cufflinks transcript assembly

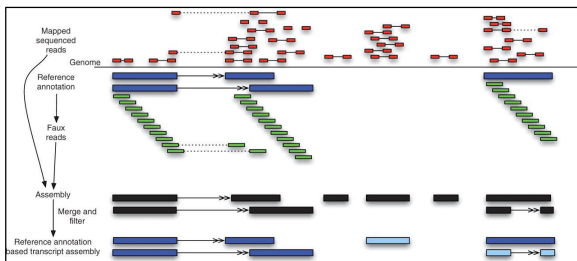
- Transcripts assembly:
- identification of incompatible fragments originated from distinct isoforms
 - connection of compatible fragments in an overlap graph
 - assembling isoforms from the overlap graph: here minimally 'covered' by three paths, each representing a different isoform



Trapnell C et al. Nature Biotechnology 2010

Cufflinks transcript assembly

Reference Annotation Based Transcripts Assembly



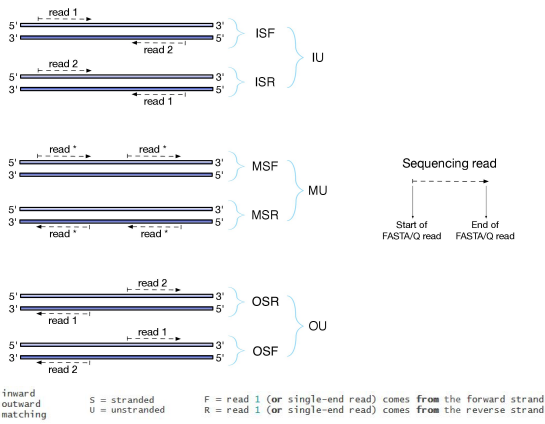
Assembling novel transcripts in the context of an existing annotation

Roberts et al. Bioinformatics 2011

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 novel transcripts
 - -g/--GTF-guide <reference_annotation.(gtf/gff)>
use reference transcript annotation to guide assembly
 - --max-bundle-length [3,500,000]
 - --max-bundle-frags [500,000]
 - --library-type
library prep used for input reads

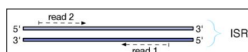
Cufflinks library types



http://salmon.readthedocs.io/en/latest/library_type.html

Cufflinks library types

Library Type	Examples	Description
fr-unstranded (default)	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	Directional Illumina (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	Salmon (and Sailfish)
	Paired-end	Single-end
-fr-unstranded	-1 IU	-1 U
-fr-firststrand	-1 ISK	-1 SK
-fr-secondstrand	-1 ISF	-1 SF

http://salmon.readthedocs.io/en/latest/library_type.html

<https://github.com/cole-trapnell-lab/cufflinks/cuffdiff#library-types>

Library Type

In the analysis of RNA-seq data, both TopHat and Cufflinks can take into account the nature of the sample preparation. Specifically, the analysis can specify that the sequenced fragments are either:

- Unstranded
- Correspond to the first strand
- Correspond to the second strand

For the TruSeq RNA Sample Prep Kit, the appropriate library type is "fr-unstranded". For TruSeq stranded sample prep kits, the library type is specified as "fr-firststrand".

<https://www.illumina.com/documents/products/technotes/RNASeqAnalysisTopHat.pdf>

Cufflinks outputs

- **transcripts.gtf**
contains assembled isoforms (coordinates and abundances)
- **genes.fpkm_tracking**
contains the genes FPKM
- **isoforms.fpkm_tracking**
contains the isoforms FPKM
- **skipped.gtf**
contains skipped loci (too many fragments)



Cufflinks GTF description

transcripts.gtf (coordinates and abundances):

- contains assembled isoforms
- can be visualized with a genome viewer
- attributes: ids, FPKM, confidence interval, read coverage & support

- score: most abundant isoform = 1000
minor isoforms = minor FPKM/major FPKM
- cov: estimate for depth across the transcript

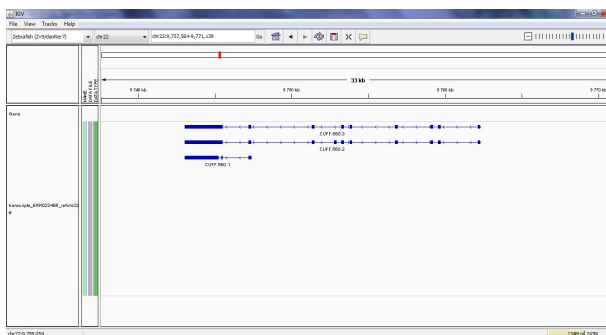
```
1 Cufflinks transcript 459812 468830 1 - -
1 Cufflinks exon 459812 468830 - -
1 Cufflinks transcript 463572 478996 1000 - -
1 Cufflinks exon 463572 463746 1000 - -
1 Cufflinks exon 466228 466405 1000 - -
```

```
gene_id "ENST0000000013841", transcript_id "ENST0000000013837", fpkm "0.000000000", frac "0.000000",
gene_id "ENST0000000013841", transcript_id "ENST0000000013837", exon_number "1", fpkm "0.000000000", frac "0.000000",
gene_id "CUFF 2", transcript_id "ENST0000000015119", fpkm "25.4745974237", frac "1.000000",
gene_id "CUFF 2", transcript_id "ENST0000000015119", exon_number "1", fpkm "25.4745974237", frac "1.000000",
gene_id "CUFF 2", transcript_id "ENST0000000015119", exon_number "2", fpkm "25.4745974237", frac "1.000000",
```

```
conf_lo "0.000000", conf_hi "0.000000", cov "0.000000", full_read_support "no";
conf_lo "0.000000", conf_hi "0.000000", cov "0.000000";
conf_lo "21.387219", conf_hi "29.561976", cov "422.904985", full_read_support "yes";
conf_lo "21.387219", conf_hi "29.561976", cov "422.904985";
conf_lo "21.387219", conf_hi "29.561976", cov "422.904985";
```

Cufflinks GTF description

transcripts.gtf (coordinates and abundances):
visualization in IGV



Cufflinks / Cuffcompare

Compare assemblies between conditions:

- compare your assembled transcripts to a reference annotation
- track Cufflinks transcripts across multiple experiments

Command:












```
cuffcompare [-r <reference.gtf>] [-o <outprefix>] <input1.gtf> ...
```

Outputs:

- <outprefix>.stats - overall summary statistics
- <outprefix>.combined.gtf - “union” of all transfrags
- <cuff_in>.refmap - transfrags matching to reference transcript
- <cuff_in>.tmap - best reference transcript for each transfrag
- <outprefix>.tracking - tracking transfrags across samples

Cuffcompare

Class code de cuffcompare

=	complete match	
c	contained	
j	novel isoform	
e	single exon	
i	within intron	
o	exonic overlap	
p	polymerase run-on	
r	repeat	
u	unknown, intergenic	
x	exonic overlap on the opposite strand	
s	intronic overlap on the opposite strand	

<http://cole-tranpell-lab.github.io/cufflinks/cuffcompare/index.html#transfrags-class-codes>

Cufflinks / Cuffmerge

Merge together several assemblies:

- merge novel isoforms and known isoforms
- filters a number of transfrags that are probably artifacts
- build a new gene model describing all conditions

Command:

```
cuffmerge [options] -o <assembly_GTF_list>
```

Options:

- -o/--output-dir
- -g/--ref-gtf
- -s/--ref-sequence
- --min-isoform-fraction
discard isoforms with abundance below this [0.05]
- -p/--num-threads

Cufflinks / Cuffmerge

merged.gtf (coordinates and legacy):

- contains merged input assemblies
- can be visualized with a genome viewer
- attributes: ids, name, old, nearest_ref, class_code, tss_id, p_id

```
1 Cufflinks exon 34627 35558 + .
1 Cufflinks exon 242394 242946 + .
1 Cufflinks exon 275223 275681 + .
1 Cufflinks exon 242402 242946 + .
1 Cufflinks exon 254559 254693 + .
1 Cufflinks exon 247340 249673 + .
1 Cufflinks exon 321346 321374 + .
1 Cufflinks exon 355264 355237 + .
1 Cufflinks exon 327793 327822 + .
1 Cufflinks exon 361144 362915 + .
```

```
gene_id "XLOC_000001"; transcript_id "TC0NS_00000001"; exon_number "1"; gene_name "ENST000000006850";
gene_id "XLOC_000002"; transcript_id "TC0NS_00000002"; exon_number "1"; gene_name "CBX3";
gene_id "XLOC_000002"; transcript_id "TC0NS_00000002"; exon_number "2"; gene_name "CBX3";
gene_id "XLOC_000002"; transcript_id "TC0NS_00000003"; exon_number "1";
gene_id "XLOC_000002"; transcript_id "TC0NS_00000003"; exon_number "2";
gene_id "XLOC_000003"; transcript_id "TC0NS_00000004"; exon_number "1";
gene_id "XLOC_000004"; transcript_id "TC0NS_00000005"; exon_number "1"; gene_name "RCAN1";
gene_id "XLOC_000004"; transcript_id "TC0NS_00000005"; exon_number "2"; gene_name "RCAN1";
gene_id "XLOC_000004"; transcript_id "TC0NS_00000005"; exon_number "3"; gene_name "RCAN1";
gene_id "XLOC_000004"; transcript_id "TC0NS_00000005"; exon_number "4"; gene_name "RCAN1";
```

```
old "ENST00000000004"; nearest_ref "ENST00000000904"; class_code "*"; tss_id "TSS1";
old "CUFF 1.1"; nearest_ref "ENST00000007283"; class_code "c"; tss_id "TSS2";
old "CUFF 1.1"; nearest_ref "ENST00000007283"; class_code "r"; tss_id "TSS2";
old "CUFF 1.2"; class_code "u"; tss_id "TSS2";
old "CUFF 1.2"; class_code "u"; tss_id "TSS2";
old "CUFF 2.1"; class_code "u"; tss_id "TSS3";
old "CUFF 3.1"; nearest_ref "ENST00000007243"; class_code "j"; tss_id "TSS4";
old "CUFF 3.1"; nearest_ref "ENST00000007243"; class_code "j"; tss_id "TSS4";
old "CUFF 3.1"; nearest_ref "ENST00000007243"; class_code "j"; tss_id "TSS4";
old "CUFF 3.1"; nearest_ref "ENST00000007243"; class_code "j"; tss_id "TSS4";
```



Tuxedo protocol

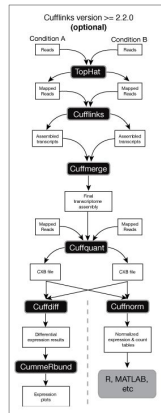
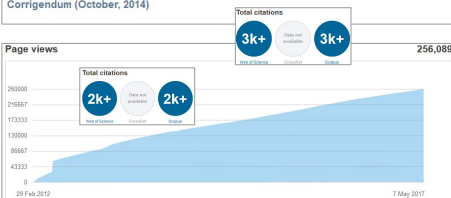
NATURE PROTOCOLS | PROTOCOL

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 Pimental, 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 | Corrected online: 07 August 2014
Corrigendum (October, 2014)



StringTie

NATURE BIOTECHNOLOGY | RESEARCH | LETTER

StringTie enables improved reconstruction of a transcriptome from RNA-seq reads

Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell & Steven L Salzberg

Affiliations | Contributions | Corresponding author

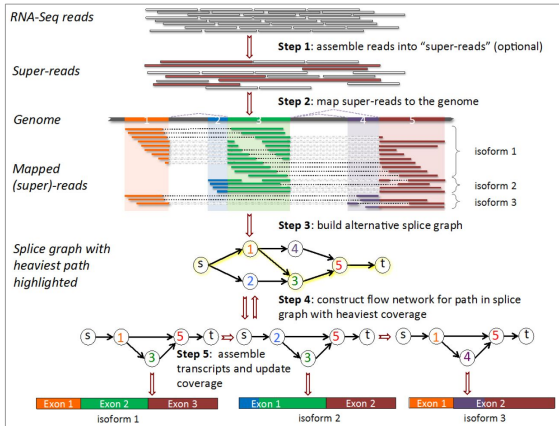
Nature Biotechnology 33, 290–295 (2015) | doi:10.1038/nbt.3122
Received 15 April 2014 | Accepted 09 December 2014 | Published online 18 February 2015

<https://ccb.jhu.edu/software/stringtie/>

- assembles transcripts
- StringTie identified 36-60% more transcripts than the next best assembler (Cufflinks)
- last version: stringtie 1.3.3, released Feb 15, 2017



StringTie transcript assembly



Pertea et al. Nature Biotechnology 2015

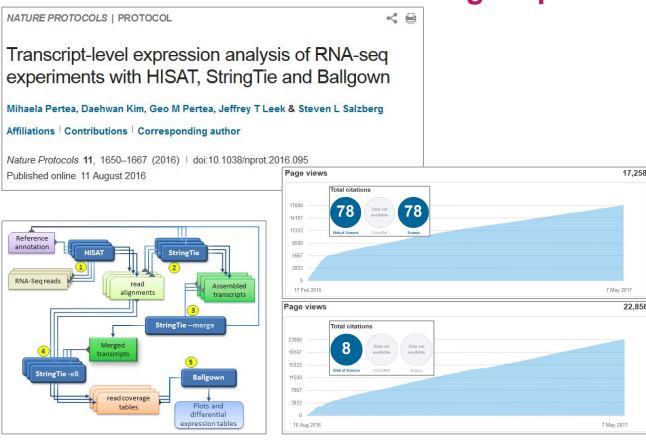
StringTie

Command:
`stringtie <aligned_reads.bam> [options]`

- Some options:
- `-o` [<path/>]<out.gtf>
 - `-G` <ref_ann.gff>
 - `--rf` | `--fr` - stranded library fr-firststrand | fr-secondstrand
 - `-p` <int>
 - `--merge` - transcript merge mode

- Main output:
- GTF file containing the assembled transcripts
 - Gene abundances in tab-delimited format
 - Fully covered transcripts matching the reference annotation
 - Files required as input to Ballgown
 - In merge mode, a merged GTF file from a set of GTF files

StringTie protocol



StringTie / gffcompare

Command:

```
gffcompare [-r <reference.gtf>] [-o <outprefix>] <input1.gtf> ...
```

Some options:

- -R for -r option
consider only the reference transcripts that overlap any of the input transfrags (Sn correction)
- -Q for -r option
consider only the input transcripts that overlap any of the reference transcripts (Precision correction); discard all "novel" loci

Output: cuffcompare like output files

StringTie / gffcompare

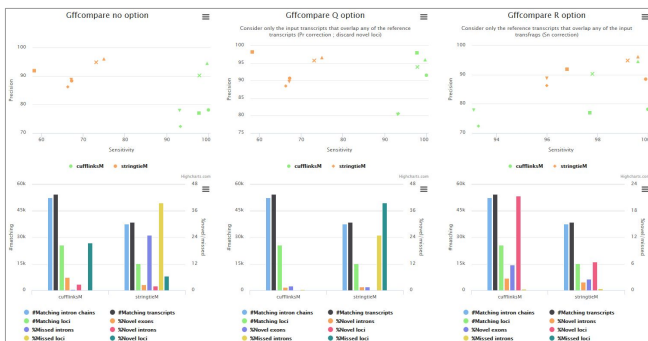
`strtcmp.stats` (transcript assembly accuracy comparison)

```
#= Summary for dataset: stringtie_asm.gtf
# Query mRNAs: 23555 in 17628 loci (17231 multi-exon transcripts)
# (3731 multi-transcript loci, ~1.3 transcripts per locus)
# Reference mRNAs : 16628 in 12062 loci (15850 multi-exon)
# Super-loci w/ reference transcripts: 11552
#-----| Sensitivity | Precision |
Base level: 82.4 | 76.5 |
Exon level: 81.2 | 82.9 |
Intron level: 86.1 | 94.8 |
Intron chain level: 56.9 | 52.4 |
Transcript level: 55.2 | 38.9 |
Locus level: 70.1 | 48.0 |
```

gffcompare2highcharts.pl

Command:

```
gffcompare2highcharts.pl --stats STATS_FILE[...STATS_FILE_n] > output.html
```

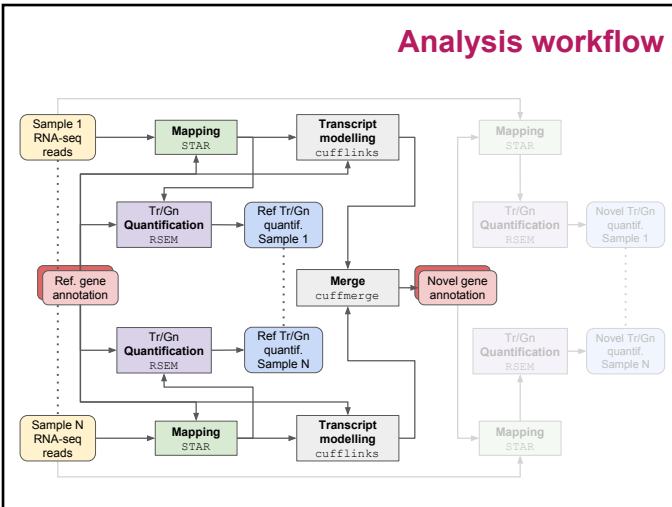


Hands-on: transcripts assembly

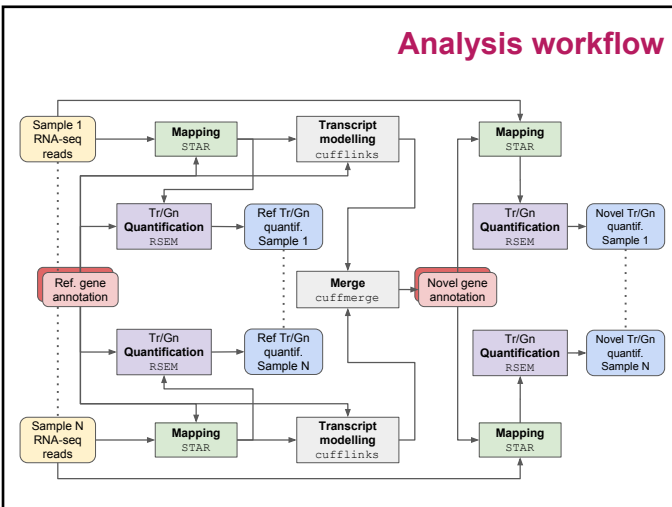
Using cufflinks et al:

Exercise 7: reconstruct known and novel transcripts

Analysis workflow



Analysis workflow



Hands-on : star, RSEM with new gtf

Exercice n°8 (Optional)

Commands :

Star and RSEM: see exercice n°5 and 6

How to choose count matrix ?

- Quality of the annotation :
 - do not forget to check the genes structure with IGV
 - presence of genes of interest
 - too many transcripts
 - quality metrics with gffcompare
- Number of reads mapped
- Number of reads assigned

Jflow

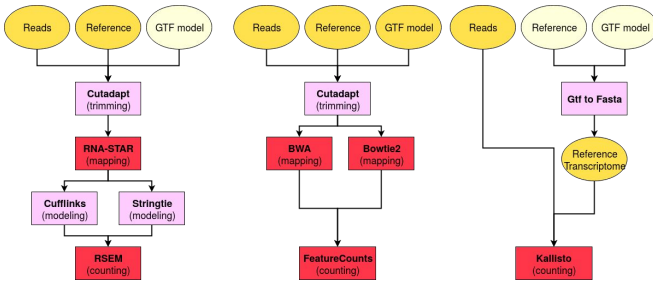
- Workflow management system
- Launch a workflow with one command line
- Available on the Genotoul platform
 - `/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py`
`<workflow_name> <workflow_parameters...>`

Rna-Seq Workflows on Jflow

RNA-Seq for Eucaryotes
(workflow: maseq)

RNA-Seq for Procaryotes
(workflow: maseqproc)

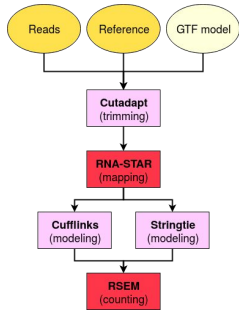
RNA-Seq without alignment
(workflow: maseqnoalign)



Dark colors: required steps / inputs
Light colors: optional steps / inputs

Rna-Seq Workflows on Jflow

RNA-Seq for Eucaryotes
(workflow: maseq)



Launch workflow:

```

/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py maseq
--sample reads-1=myfile_R1.fastq.gz
(reads-2=myfile_R2.fastq.gz)
--reference-genome fasta-file=reference.fasta
(index-directory=/path/to/directory)
--gtf-file model.gtf
--protocol (illumina_stranded, other) default :
illumina_stranded
  
```

Others parameters:

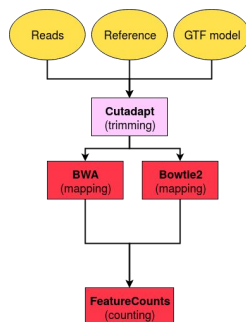
```

--trim-reads : to trim reads before proceeding
default: TruSeq Adapter
--compute-gtf-model : to compute a new gtf
model (gtf-file parameter is optional in this
case)
--modeling-software [cufflinks|stringtie]
(default: cufflinks)
  
```

To list all parameters available for this workflow:
/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py maseq --help

Rna-Seq Workflows on Jflow

RNA-Seq for Procaryotes
(workflow: maseqproc)



Launch workflow:

```

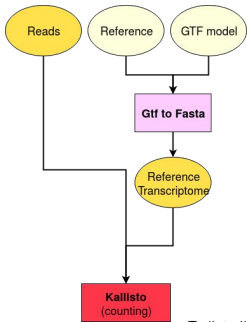
/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py maseqproc
--sample reads-1=myfile_R1.fastq.gz
(reads-2=myfile_R2.fastq.gz)
--reference-genome reference.fasta
(--indexed-genome)
--gtf-file model.gtf
--protocol (illumina_stranded, other) default:
illumina_stranded
  
```

Other parameters:

To list all parameters available for this workflow:
/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py maseqproc --help

Rna-Seq Workflows on Jflow

RNA-Seq without alignment
(workflow: rnaseqnoalign)



Launch workflow:

```
/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py rnaseqnoalign --sample reads-1=myfile_R1.fastq.gz (reads-2=myfile_R2.fastq.gz)
```

Case 1: you have the transcriptome:

```
--transcriptome : transcriptome fasta file
```

Case 2: you don't have the transcriptome:

```
--reference-genome reference.fasta --gtf-file model.gtf
```

To list all parameters available for this workflow:
`/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py rnaseqnoalign --help`

Rna-Seq Workflows on Jflow

- The documentation is here:
`/usr/local/bioinfo/src/Jflow/jflow/workflows/rnaseq/doc` and give the hidden parameters.
- In development:
 - A log file containing: the list of commands launched (to have the parameters) and version of the software.

Useful references

- **Experimental design:**
Liu et al., RNA-seq differential expression studies: more sequence or more replication?, 2014, *Bioinformatics*, Vol. 30 no. 3 2014, pages 301–304.
Schurch et al., How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?, 2016, *RNA* 22:839–851.
- **Pipeline STAR / cufflinks / RSEM:**
Djebali et al., Bioinformatics pipeline for transcriptome sequencing analysis, *Methods in Molecular Biology*, 2017, vol. 1468.
- **Tools / pipelines benchmarks for differentially expressed genes identification:**
Williams et al., Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq, *BMC bioinformatics*, 2017, 18:38.
Baruzzo et al., Simulation-based comprehensive benchmarking of RNA-seq aligners, 2017, *Nature methods*, vol. 14 n°2.

Useful references

- **Best practices from experimental design to differential expression analysis:**

Conesa et al., A survey of best practices for RNA-seq data analysis, 2016, *Genome Biology* 17:13.

- **Pipeline HISAT, Stringtie, Gffcompare, Ballgown:**

Pertea et al., Transcript-level expression analysis of RNA-seq experiments with HISAT, Stringtie and Ballgown, 2016, *Nature Protocols*, vol.11 n°9

- **Alignment-independent quantification:**

<https://cgatoxford.wordpress.com/2016/08/17/why-you-should-stop-using-feature-counts-htseq-or-cufflinks2-and-start-using-kallisto-salmon-or-sailfish/>

- **Transcript-level or gene-level ?**

<http://www.rna-seqblog.com/modern-rna-seq-differential-expression-analyses-transcript-level-or-gene-level-2/>

Quality for Bioinfo Platform!

Satisfaction form :

<https://enquetes.inra.fr/index.php/84236?lang=fr>



Useful links

Seqanswer: <http://seqanswers.com/>

Biostars: <http://www.biostars.org/>

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

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