

# 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/](http://genoweb.toulouse.inra.fr/~formation/LigneCmd/RNAseq/data/correction_star_rsem/)

# The speakers

**Sarah Maman**



**Céline Noirot**



**Claire Hoede**



**Nathalie Villa-Vialaneix**



**Christine Gaspin**



**Matthias Zytnicki**



**Cédric Cabau**

**(pour la partie biostat)**

# Session organisation

## Day 1

### Morning (9h00 -12h30) :

- Prerequisite unix/format  
Exercises
- Biological reminds

### 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

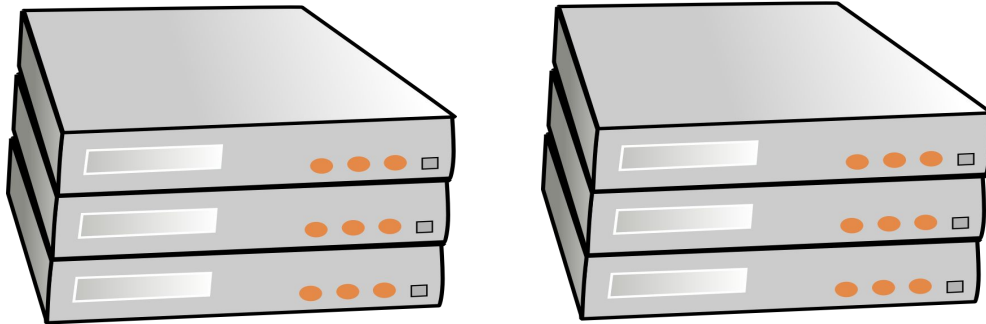
**Prerequisite unix**

# Summary - Unix reminders

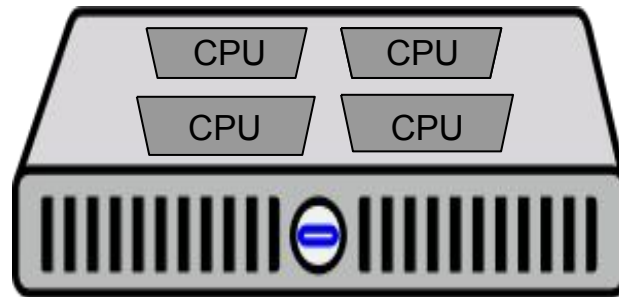
- Genotoul infrastructure organisation
- How to connect to genotoul
- How to transfer data
  - From the web to genotoul
  - From genotoul to your computer
- How to launch jobs on the cluster
- ...

# Vocabulary : Cluster / Node

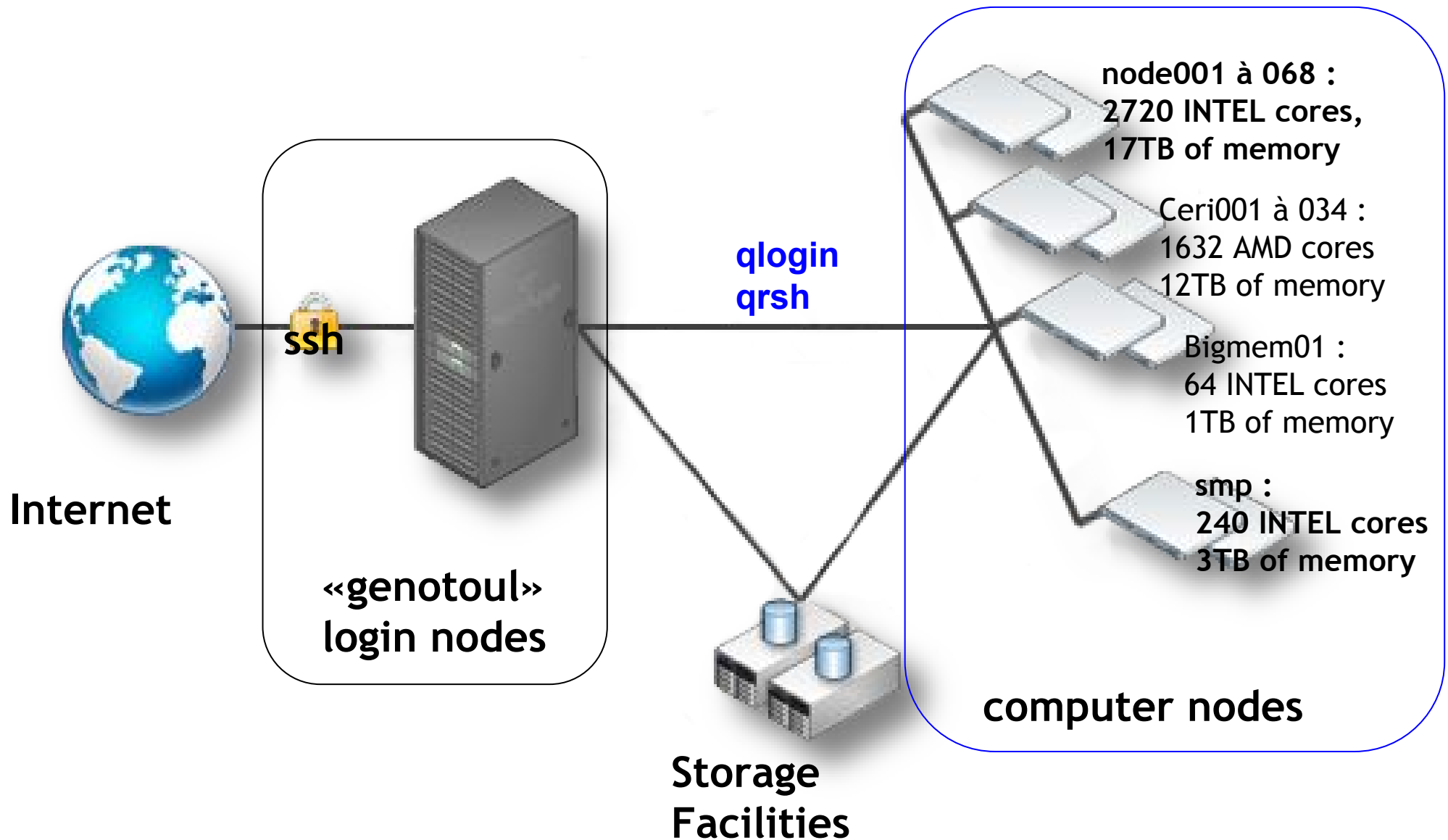
- Cluster : set of nodes



- Node : Huge computer (with several CPUs)



# « genotoul » cluster





# How to connect to genotoul ?

- **Xming** (Windows graphic) + **Putty** (Connection)
- **MobaXterm** (Executable file)



```
sigenae@genotoul
La partition /save est reservee pour les donnees A sauvegarder (quota utilisateur 250Go)
La partition /work est reservee pour les fichiers temporaires de calcul (quota utilisateur 1To)
=> ATTENTION CE VOLUME N EST PAS SAUVEGARDE ET PEUT ETRE PURGE SI BESOIN
=> IL EST DE VOTRE RESPONSABILITE DE GERER VOS DONNEES (organisation, volumetrie, pertinence, anciennete)

Pour connaitre votre consommation d'espace disque, utilisez la commande suivante :
# du -csh /DIR_NAME/USER_NAME/*

=====
Informations concernant l'utilisation de la memoire
=====
Il existe une limitation de 8Go de RAM par process utilisateur (sur le cluster).
Pour obtenir plus de memoire, veuillez consulter la FAQ de notre site web (cf + bas)

=====
Informations concernant le quota de temps de calcul
=====
Il existe un quota de temps de calcul annuel de 100.000 pour les academiques (500H pour les entreprises privees).
Au dela il faut renseigner le formulaire de demande de ressources exceptionnelles.
Vous pouvez verifier votre quota de calcul avec la commande: quota_cpu <login>

=====
Support
=====
Pour plus d'informations, consulter le site web :
http://bioinfo.genotoul.fr/

Pour toute demande de support, adressez-vous a :
support.bioinfo.genotoul@inra.fr

[sigenae@genotoul1 ~]$ ls
bin@  downloads/  igv@  public_html@  rezlog/  screenlog.0  stat/  ~Untitled  work@
Desktop@  gtf-files.txt  profile.d/  R@  save@  snpEff.v3.0_UMD3.1.75.zip  temp@  ~Untitled_1  work_project@
[sigenae@genotoul1 ~]$ ll
total 31M
lrwxrwxrwx 1 sigenae SIGENAE 8 Jan 10 2014 bin -> save/bin/
lrwxrwxrwx 1 sigenae SIGENAE 13 Dec 20 2011 Desktop -> save/Desktop//
drwxr-xr-x 2 sigenae SIGENAE 260 Apr 26 14:18 Downloads/
-rw-r--r-- 1 sigenae SIGENAE 24K Nov 24 14:23 gtf-files.txt
lrwxrwxrwx 1 sigenae SIGENAE 8 Dec 20 2011 igv -> save/igv/
drwxr-xr-x 2 sigenae SIGENAE 410 Mar 28 08:45 profile.d/
lrwxrwxrwx 1 root nfsnobody 25 Sep 25 2013 public_html -> /save/sigenae/public_html/
lrwxrwxrwx 1 sigenae SIGENAE 6 Dec 4 2015 R -> save/R/
drwxr-xr-x 2 sigenae SIGENAE 152 Sep 23 2015 rezlog/
lrwxrwxrwx 1 sigenae SIGENAE 13 Dec 20 2011 save -> /save/sigenae/
-rw-rw-r-- 1 sigenae SIGENAE 974K May 12 2015 screenlog.0
-rw-r--r-- 1 sigenae SIGENAE 24M Nov 17 16:21 snpEff.v3.0_UMD3.1.75.zip
drwxr-xr-x 2 sigenae SIGENAE 1.6K Sep 10 2015 stat/
lrwxrwxrwx 1 sigenae SIGENAE 9 Apr 3 2012 temp -> work/temp/
-rw-r----- 1 sigenae SIGENAE 1.2K May 10 15:32 ~Untitled
-rw-r----- 1 sigenae SIGENAE 2.1K May 10 15:54 ~Untitled_1
lrwxrwxrwx 1 sigenae SIGENAE 13 Dec 20 2011 work -> /work/sigenae/
lrwxrwxrwx 1 sigenae SIGENAE 22 Oct 10 2016 work_project -> /work/project/sigenae//
[sigenae@genotoul1 ~]$
```

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

## *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)

*Which are the main unix/linux  
commands you know ?*

**Three standard fluxes are opened when you launch a command:**

- **Stdin** : standard Input
- **Stdout**: standard output (default: screen)
- **Stderr**: standard error output (default: screen)

## Redirections with specific operators

- **>** : redirection of standard output  
Ex: `ls * > Liste_file`
- **<** : redirection of standard input  
Ex: `RNAfold < file.fa > Result.out`
- **>>** : redirection of standard output with concatenation
- **>&** : redirection of standard error and output
- **| (pipe)** : redirection of standard output on standard input

# Linux & Genotoul Bioinfo file organization

*Tree organization: root represented by “/” (slash)*

*Disk spaces on Genotoul Bioinfo*



`/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)

## Essential commands

- **Help pn commands**

`man cmde`

- **Where I am in the tree ?**

`pwd`

- **Moving in the tree**

`cd dir_name`      move to “dir\_name” child directory

`cd ..`              move to parent directory

- **List directory content**

`ls`                  list the content of current directory

- **Visualize file content**

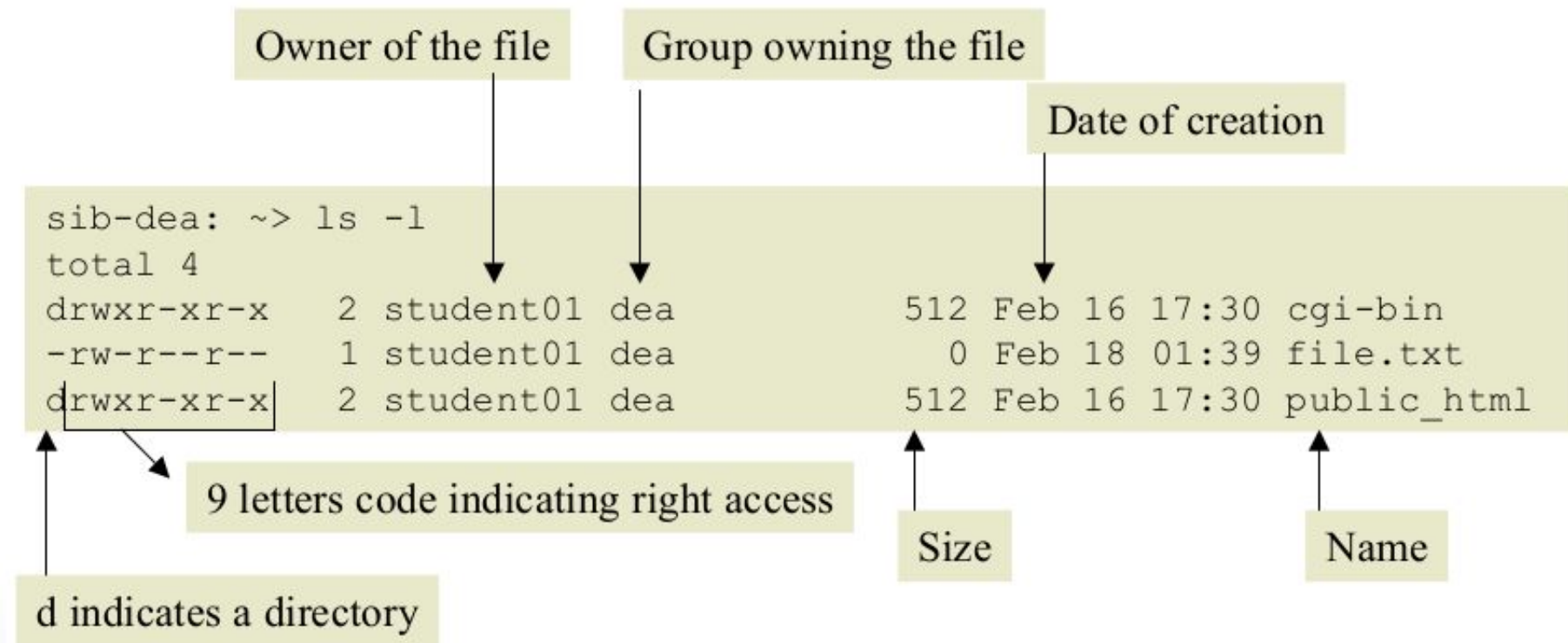
`cat f_name, more f_name, head f_name, tail f_name`

- **Others**

`cp, mv, rm, mkdir, rmdir, which, grep, wc...`

## Access rights

Each file (and directory) has associated access rights, which may be found by typing **ls -l** :



- r (or '-'): indicates the presence (or absence) of permission to read and copy the file
- w (or '-'): indicates write permission (or absence of permission)
- x (or '-'): indicates execution permission (or absence of permission)



# 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

*How to use Genotoul Bioinfo resources ?*

# OGE (Open Grid Engine)

## Queues availables for users

Queue	Access	Priority	Max time	Max slots
<b>workq (default)</b>	everyone	300	96H	4120
<b>unlimitq</b>	everyone	100	unlimited	680
<b>smpq</b>	on demand	0	unlimited	240
<b>hypermemq</b>	on demand	0	unlimited	96
<b>Interq (qlogin)</b>	everyone	100	48H	40
<b>galaxyq</b>	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

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)

## 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)

## Monitoring jobs : qstat

```
$qstat
```

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

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

```
qstat -u my_login | more
```

# OGE (Open Grid Engine)

## Monitoring jobs : qstat

- **state** : actually state of job
  - d(eletion) : 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
```

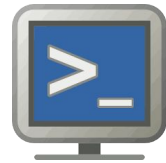
```
$ qdel 3629151
```

```
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



for i in ...

```
blastx+ -d nt -i seq1.fa  
blastx+ -d nt -i seq2.fa  
blastx+ -d nt -i seq3.fa
```



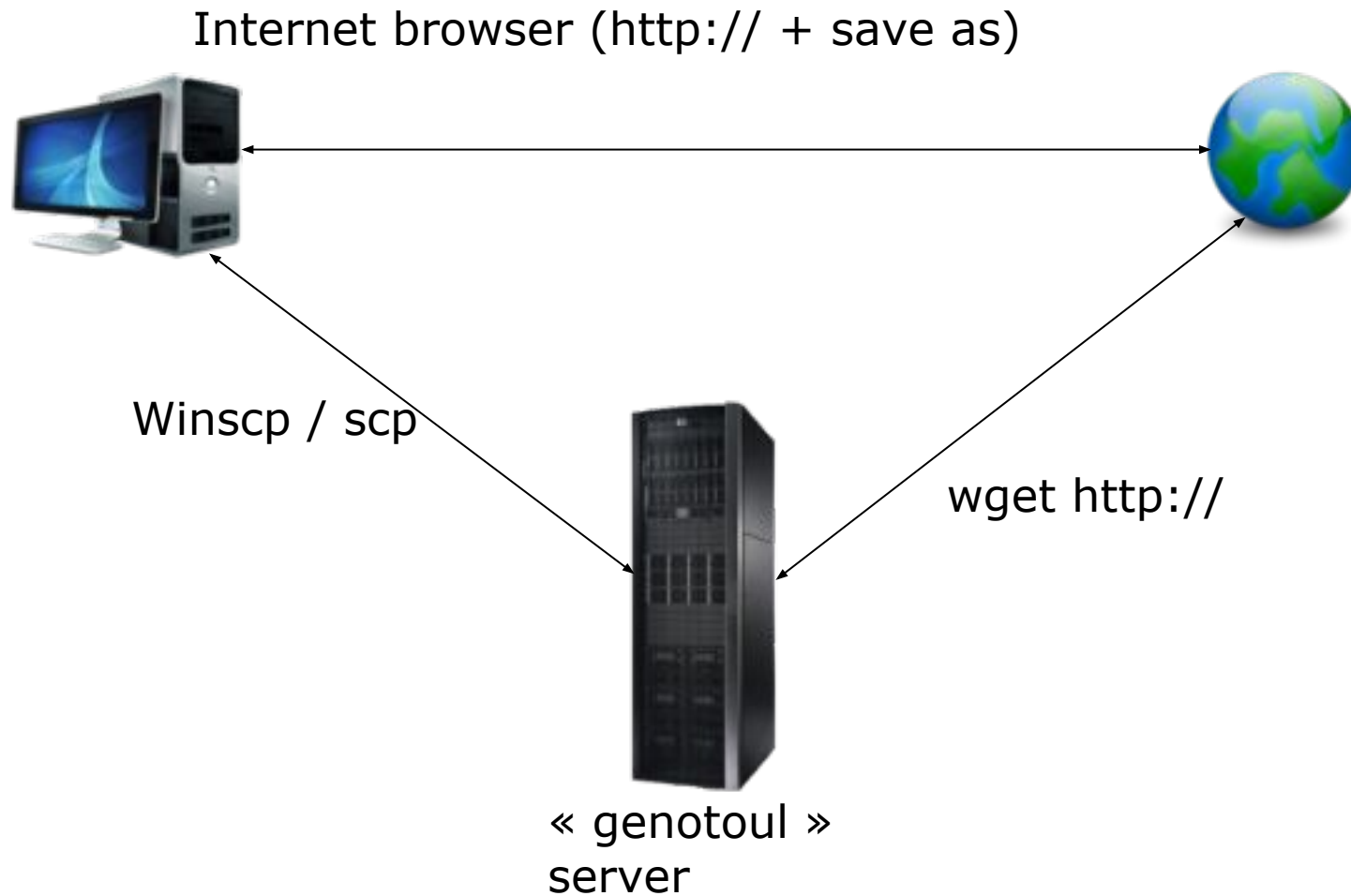
qarray script.sh



Execution on 3 cores

# 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

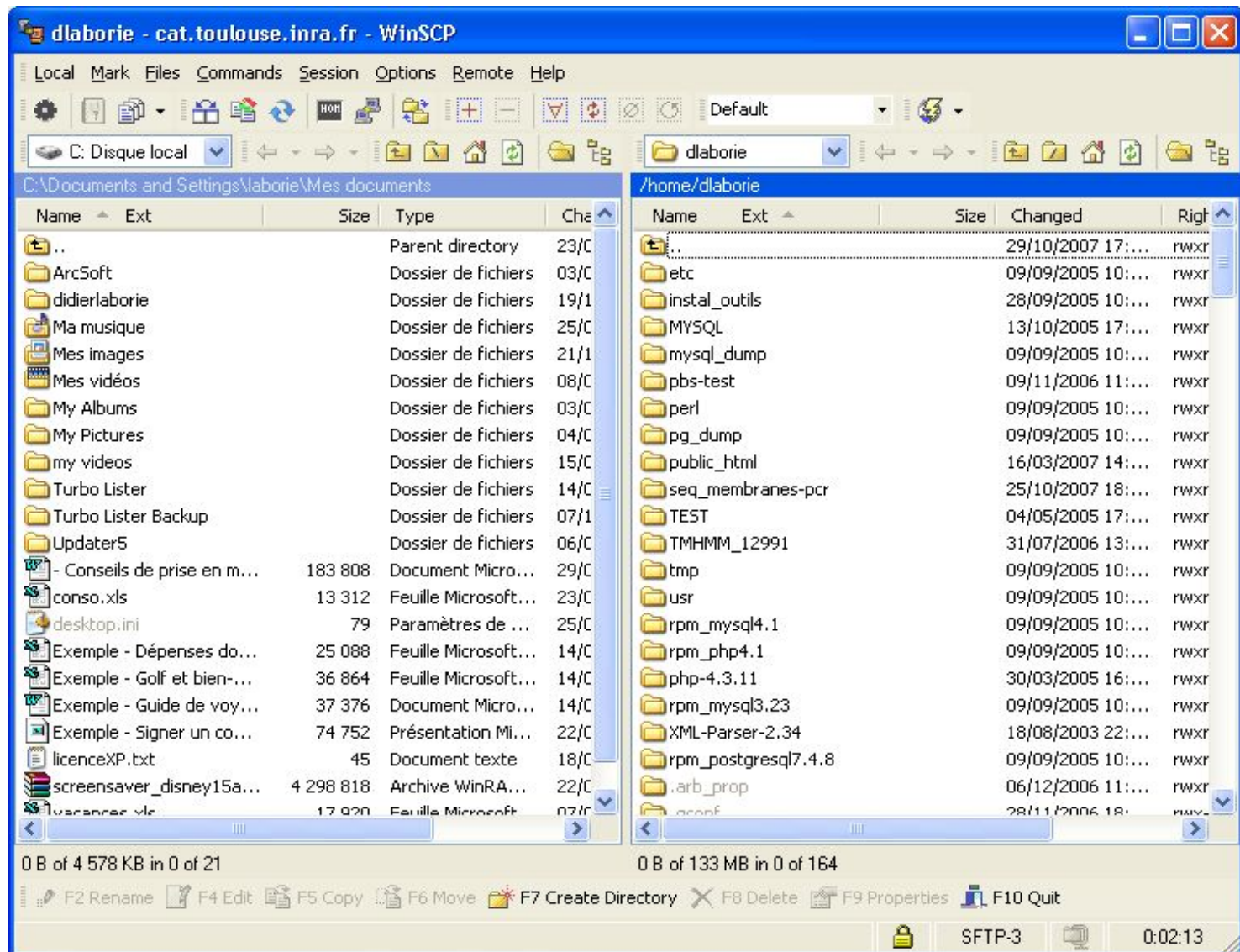
Example copy from desktop to "genotoul":

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



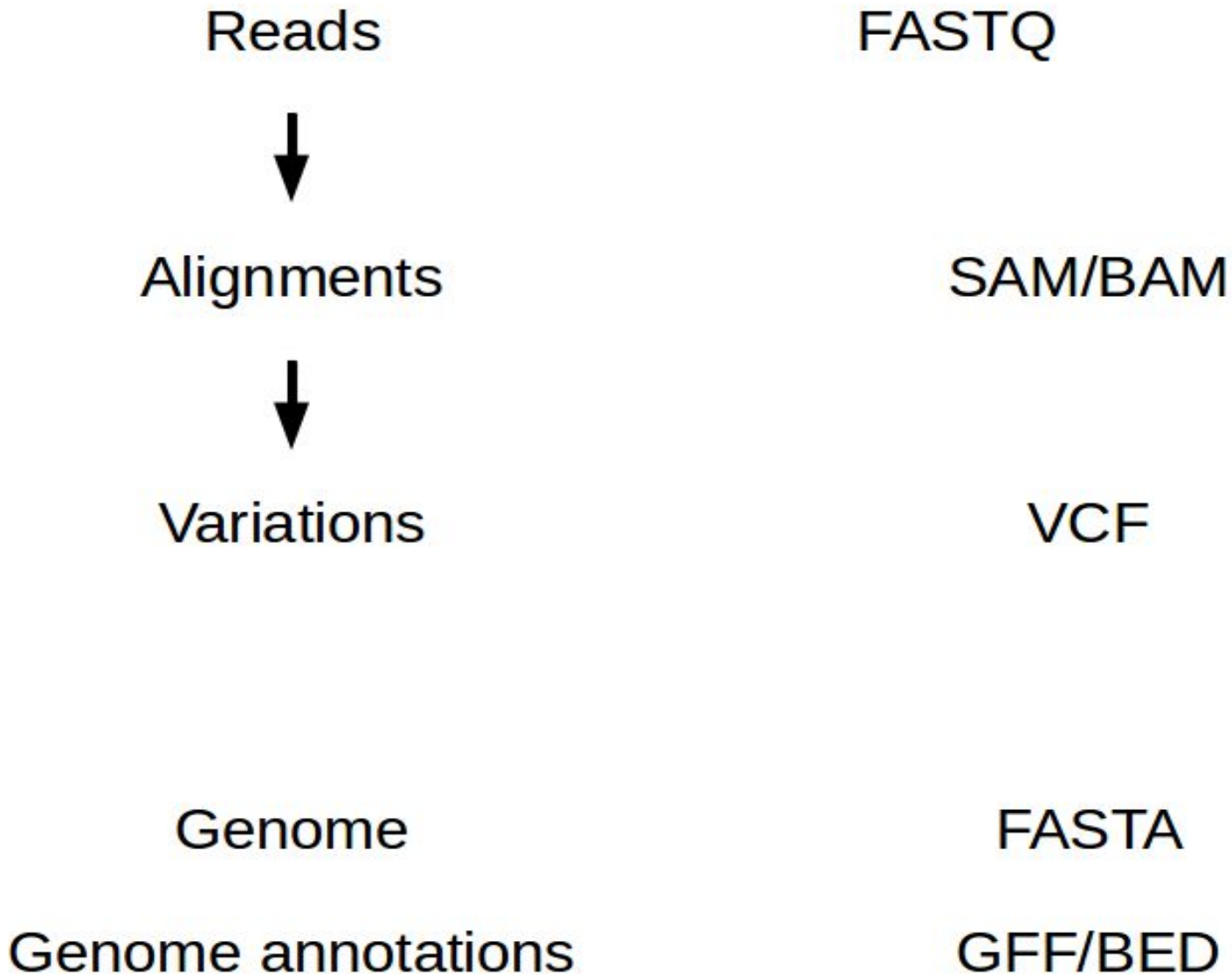
# Downloading / transferring

*WinSCP / FileZilla : copy via graphical interface*



# Introduction to NGS formats

# Summary - Format remind



# fastq format

- Standard for storing outputs of HTS
- A text-based format for storing a read and its corresponding quality scores
- 1 read <-> 4 lines

```
@HWI-ST218:596:C90JYANXX:8:1101:1293:2188 1:N:0:ATTCAGAATAATCTTA  
NCTAAGTGTTAGGGGGTTTCCGCCCTTAGTGCTGCAGCTAACGCATTAAGCACTCCGCCTGGGGAGTACGGTCGCAAGACTGAAAA  
+  
#<3?BFGGGGGGEGGGGGGGEGGGGGG@F1FGGGGGGDDGG1FB</9FE=EGGGGGGGG>GGGGBGGGGG<<C/BDGGGGGC=GGG
```

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

- Sequence identifier

@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG

1. Begins with '@' character and is followed by a sequence identifier

<b>EAS139</b>	the unique instrument name
<b>136</b>	the run id
<b>FC706VJ</b>	the flowcell id
<b>2</b>	flowcell lane
<b>2104</b>	tile number within the flowcell lane
<b>15343</b>	'x'-coordinate of the cluster within the tile
<b>197393</b>	'y'-coordinate of the cluster within the tile
<b>1</b>	the member of a pair, 1 or 2 ( <i>paired-end or mate-pair reads only</i> )
<b>Y</b>	Y if the read is filtered, N otherwise
<b>18</b>	0 when none of the control bits are on, otherwise it is an even number
<b>ATCACG</b>	index sequence

# fastq format

- Base quality (Sanger standard)

```
@HWI-ST218:596:C90JYANXX:8:1101:1293:2188 1:N:0:ATTCAGAATAATCTTA  
NCTAAGTGTAGGGGGTTTCCGCCCTTAGTGCTGCAGCTAACGCATTAAGCACTCCGCCTGGGGAGTACGGTCGCAAGACTGAAAA  
+  
#<3?BFGGGGGGEGGGGGGGEGGGGGG@F1FGGGGGGDDGG1FB</9FE=EGGGGGGGG>GGGGBGGGGG<<C/BDGGGGGGC=GGG
```

ASCII-encoded version of the PHRED quality given by  $Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$

SANGER=PHRED+33 : H=ASCII(40+33)       $Q = -10 \log_{10} P \Leftrightarrow P = 10^{\frac{-Q}{10}}$

Score de qualité phred	Probabilité d'une identification incorrecte	Précision de l'identification d'une base
10	1 pour 10	90 %
20	1 pour 100	99 %
30	1 pour 1000	99.9 %
40	1 pour 10000	99.99 %
50	1 pour 100000	99.999 %

## SURVEY AND SUMMARY

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

Peter J. A. Cock<sup>1,\*</sup>, Christopher J. Fields<sup>2</sup>, Naohisa Goto<sup>3</sup>, Michael L. Heuer<sup>4</sup> and Peter M. Rice<sup>5</sup>

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

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



- 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 (bam)
- 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

- Header file : generic information

```
@HD VN:1.0 GO:none SO:coordinate
@SQ SN:chrM LN:16571
@SQ SN:chr1 LN:247249719
@SQ SN:chr2 LN:242951149
[cut for clarity]
@SQ SN:chr9 LN:140273252
@SQ SN:chr10 LN:135374737
@SQ SN:chr11 LN:134452384
[cut for clarity]
@SQ SN:chr22 LN:49691432
@SQ SN:chrX LN:154913754
@SQ SN:chrY LN:57772954
@RG ID:20FUK.1 PL:illumina PU:20FUKAAXX100202.1 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.2 PL:illumina PU:20FUKAAXX100202.2 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.3 PL:illumina PU:20FUKAAXX100202.3 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.4 PL:illumina PU:20FUKAAXX100202.4 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.5 PL:illumina PU:20FUKAAXX100202.5 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.6 PL:illumina PU:20FUKAAXX100202.6 LB:Solexa-18484 SM:NA12878 CN:BI
@RG ID:20FUK.7 PL:illumina PU:20FUKAAXX100202.7 LB:Solexa-18483 SM:NA12878 CN:BI
@RG ID:20FUK.8 PL:illumina PU:20FUKAAXX100202.8 LB:Solexa-18484 SM:NA12878 CN:BI
@PG ID:BWA VN:0.5.7 CL:tk
@PG ID:GATK TableRecalibration VN:1.0.2864
```

**Required:** Standard header

**Essential:** contigs of aligned reference sequence. Should be in karotypic order.

**Essential:** read groups. Carries platform (PL), library (LB), and sample (SM) information. Each read is associated with a read group

**Useful:** Data processing tools applied to the reads

# Sequence Alignment/Map (SAM) format

- Header file: generic information
- Body file (alignment description)
  - 11 mandatory fields
  - Variable number of optional fields
  - Tab delimited fields

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	LN:4833080											
1	16	ref	740202	0	18M	*	0	0	TTTTTTTTTTTTTTTTTTTT	>>>>>??????????????	XT:A:R	NM:i:2	MD:Z:5A5A6
2	16	ref	740202	0	18M	*	0	0	TTTTTTTTTTTTTTTTTTTT	<>>>>>>>>>>>>>>	XT:A:R	NM:i:2	MD:Z:5A5A6
3	16	ref	740202	0	18M	*	0	0	TTTTTTTTTTTTTTTTTTTT	>>>>>??????????????	XT:A:R	NM:i:2	MD:Z:5A5A6
4	16	ref	740202	0	18M	*	0	0	TTTTTTTTTTTTTTTTTTTT	>>>>>??????????????	XT:A:R	NM:i:2	MD:Z:5A5A6
5	16	ref	740202	0	18M	*	0	0	TTTTTTTTTTTTTTTTTTTT	>>>>>??????????????	XT:A:R	NM:i:2	MD:Z:5A5A6
6	16	ref	740202	0	18M	*	0	0	TTTTTTTTTTTTTTTTTTTT	>>>>>??????????????	XT:A:R	NM:i:2	MD:Z:5A5A6
7	16	ref	740202	0	18M	*	0	0	TTTTTTTTTTTTTTTTTTTT	>>>>>??????????????	XT:A:R	NM:i:2	MD:Z:5A5A6
8	16	ref	740202	0	18M	*	0	0	TTTTTTTTTTTTTTTTTTTT	>>>>>??????????????	XT:A:R	NM:i:2	MD:Z:5A5A6
9	16	ref	740202	0	18M	*	0	0	TTTTTTTTTTTTTTTTTTTT	>>>>>>>>>>>>>>>	XT:A:R	NM:i:2	MD:Z:5A5A6
10	0	ref	4702037	25	18M	*	0	0	CTATGCAGCTATATGTTT	>>>>;>7>>>>>7>7>>><	XT:A:U	NM:i:2	MD:Z:3C11G2
11	16	ref	2919865	37	18M	*	0	0	GGTGGTTATGCTATTC	:>>>>>>>>>>>><>>><	XT:A:U	NM:i:0	MD:Z:18
12	0	ref	2995664	37	18M	*	0	0	GTTTTGTTATGTGAATAT	; '>>70>>7>3977+%(7	XT:A:U	NM:i:0	MD:Z:18
13	16	ref	510805	37	18M	*	0	0	ATTCTCTATGGAGTGGGT	<///>+>799+>>>>>>>	XT:A:U	NM:i:0	MD:Z:18
14	16	ref	740202	0	18M	*	0	0	TTTTTTTTTTTTTTTTTTTT	>>>>>>>>>>>>>>>	XT:A:R	NM:i:2	MD:Z:5A5A6
15	4	*	0	0	*	*	0	0	GTGDDACTCCTGGTCTTG	>8;>1;>>9-9.7/(+5\$			
16	16	ref	740202	0	18M	*	0	0	TTTTTTTTTTTTTTTTTTTT	>>>>>??????????????	XT:A:R	NM:i:2	MD:Z:5A5A6
17	0	ref	1847349	37	9M1I8M	*	0	0	CATCACATATATCATCATT	>49;;>9,77>;+>6+ '/'	XT:A:U	NM:i:2	MD:Z:5T11

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
- RG : 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

## Examples

1 = 00000001 → paired end read  
2 = 00000010 → mapped as proper pair  
4 = 00000100 → unmappable read  
8 = 00001000 → read mate unmapped  
16 = 00010000 → read mapped on reverse strand

The flag **11** → **1 + 2 + 8 = 0001011** (conditions 1, 2 and 8 satisfied)

## Other examples

0=00000000 ???  
99=01100011 ???  
147=10010011 ???

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>

# SAM format - cigar line

M: match/mismatch

I: insertion

D: deletion

S: softclip

H: hardclip

P: padding

N: skip

Diagram illustrating SAM format - cigar line with sequence alignment and annotations.

Coordinates: 12345678901234 5678901234567890123456789012345

Reference: AGCATGTTAGATAA\*\*GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

Paired-end annotations: r001+, r002+, r003+, r004+, r003-, r001-

Multipart annotations: r003+, r003-, r001-

Sequence alignment:

```

TTAGATAAAGGATA*CTG
aaaAGATAA*GGATA
gcctaAGCTAA
ATAGCT.....TCAGC
ttagctTAGGC
CAGCGCCAT
    
```

Annotations:

- Ins & padding: r001 163 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTA \*
- Soft clipping: r002 0 ref 9 30 3S6M1P1I4M \* 0 0 AAAAGATAAGGATA \*
- Splicing: r003 0 ref 9 30 5H6M \* 0 0 AGCTAA \* NM:i:1
- Hard clipping: r004 0 ref 16 30 6M14N5M \* 0 0 ATAGCTTCAGC \*
- Hard clipping: r003 16 ref 29 30 6H5M \* 0 0 TAGGC \* NM:i:0
- Hard clipping: r001 83 ref 37 30 9M = 7 -39 CAGCGCCAT \*

# 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

***Exercise 1 : using basic unix commands***

***Exercise 2 : format manipulation***



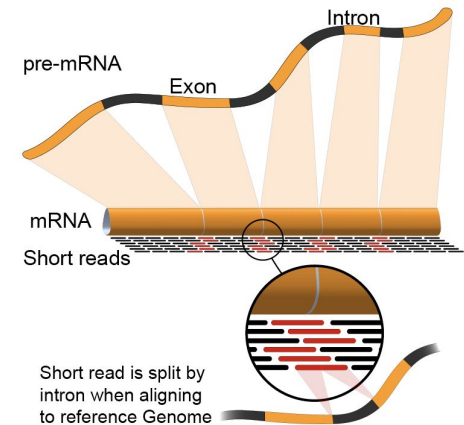
# Summary - Biological reminders

- 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



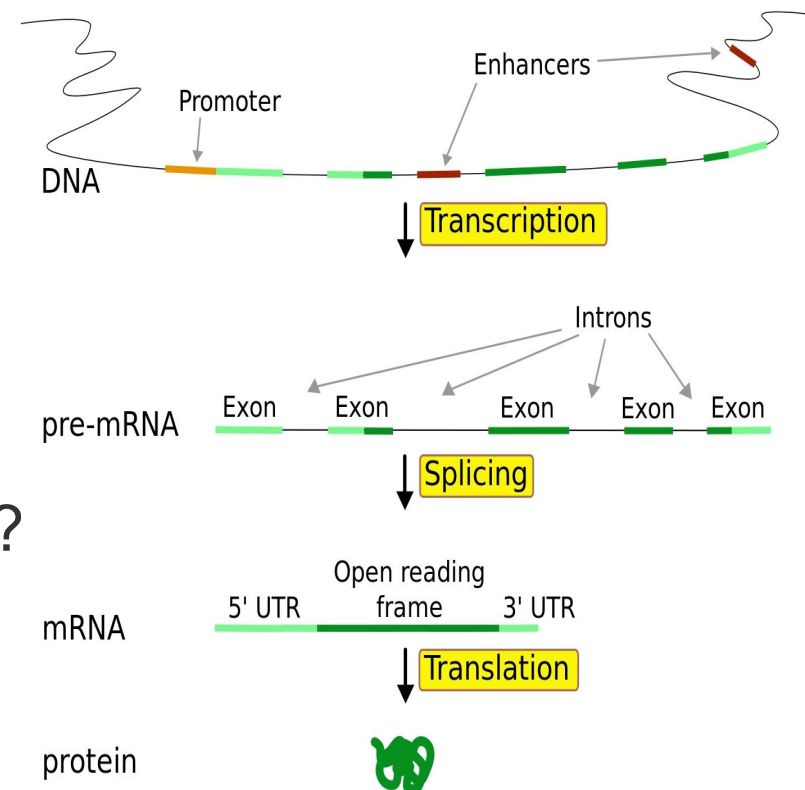
# Context

## Prerequis :

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

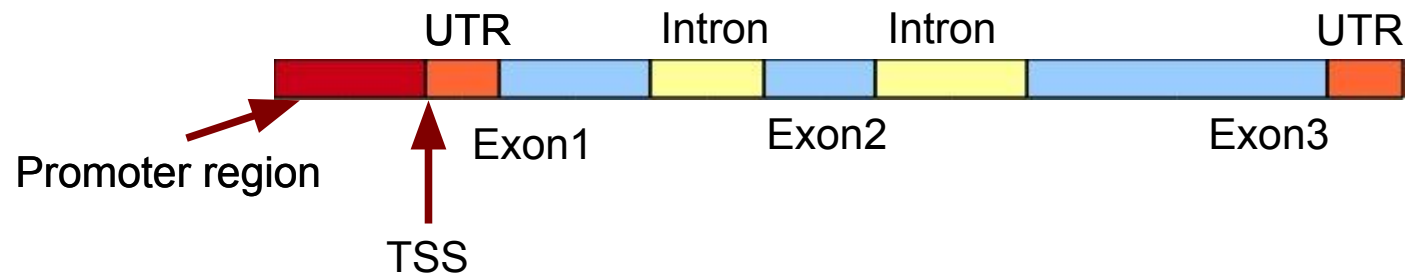
## Try to answer to :

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



# 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  $\neq$  1<sup>st</sup> amino acid

**Transcript** : stretch of DNA transcribed into an RNA molecule

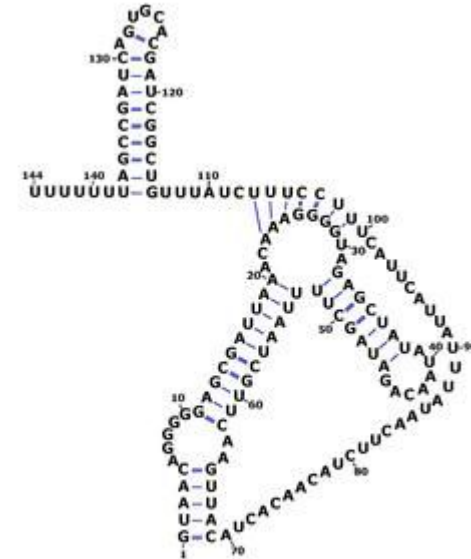


# Transcription products

**Protein coding gene:** transcribed in mRNA

**ncRNA :** highly abundant and functionally important RNA

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





## Statistics about the current GENCODE freeze (version 21)

Statistics of previous GENCODE freezes are found archived [here](#).

\* The statistics derive from the [gtf file](#) <sup>Ⓜ</sup> that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the [README\\_stats.txt](#) <sup>Ⓜ</sup> 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	Total No of distinct translations	59512
Immunoglobulin/T-cell receptor gene segments		Genes that have more than one distinct translations	13526
- protein coding segments:	395		
- pseudogenes:	226		

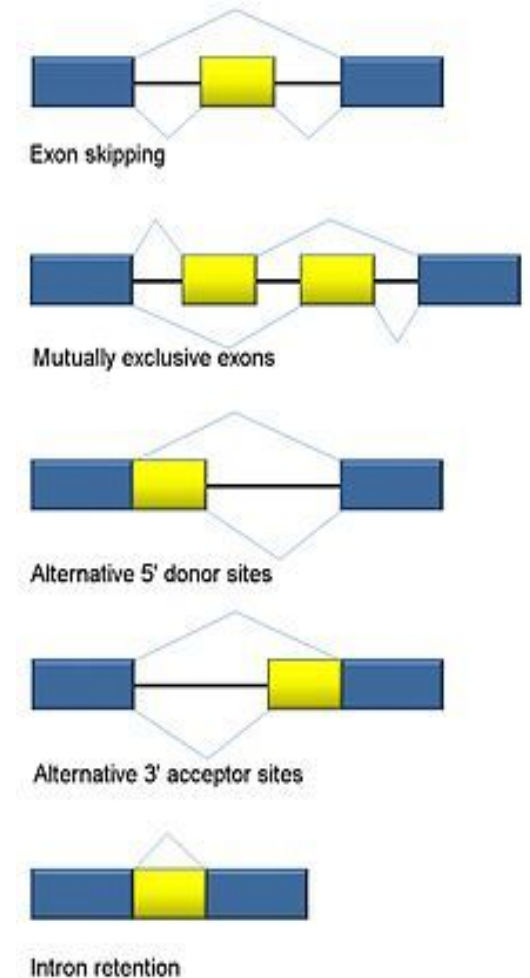
# Alternative splicing

## 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

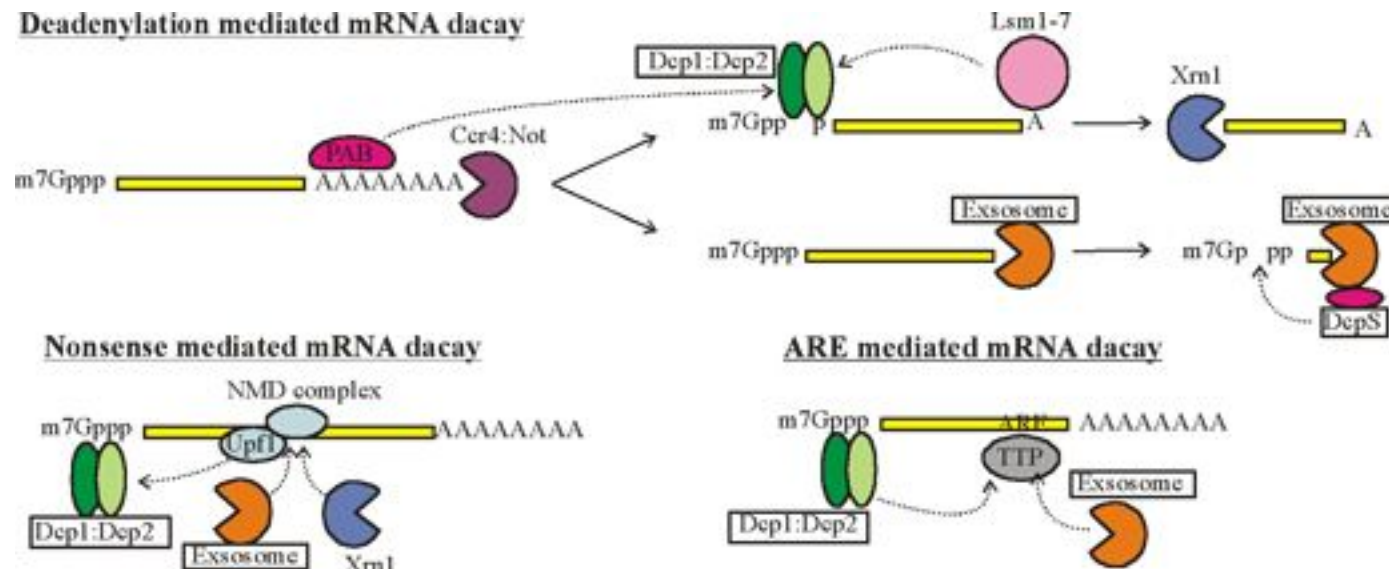
**Post-transcriptional modification** (eukaryotic cells) eg: the conversion of precursor messenger RNA into mature mRNA (mRNA), editing...



[http://en.wikipedia.org/wiki/Alternative\\_splicing](http://en.wikipedia.org/wiki/Alternative_splicing)

# Transcript degradation

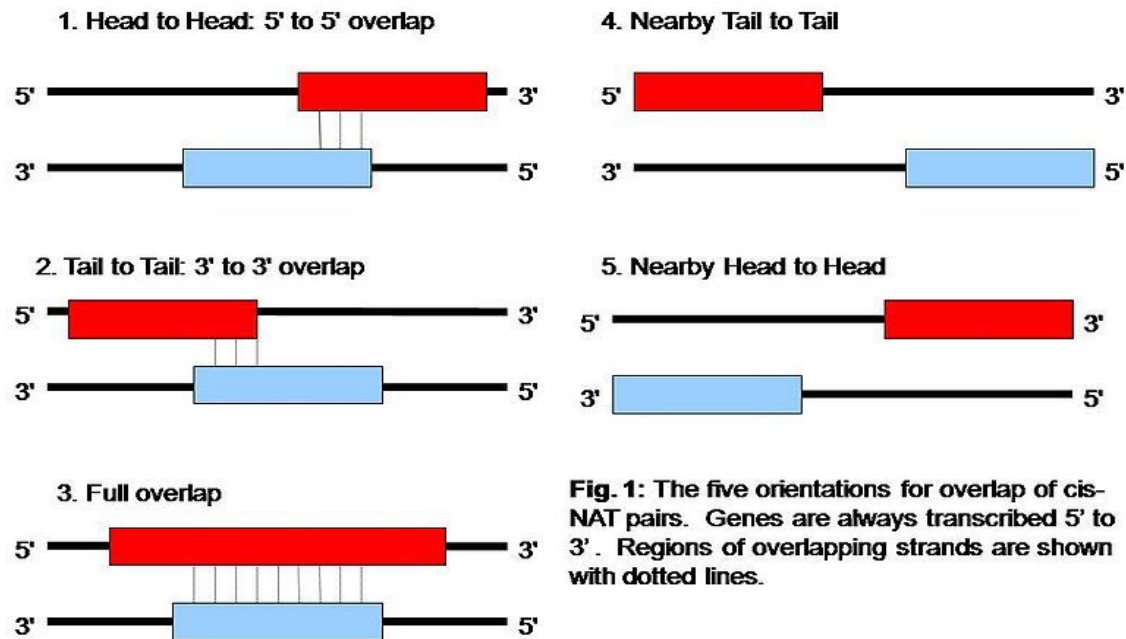
- mRNA export to the cytoplasm,
- protected from degradation by a 5' cap structure and a 3' polyA tail.
- the polyA tail is gradually shortened by exonucleases
- the degradation machinery rapidly degrades the mRNA in both in directions.
- others mechanisms, bypass the need for deadenylation and can remove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





# Cis-natural antisense transcript

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



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

# Fusion genes

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

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

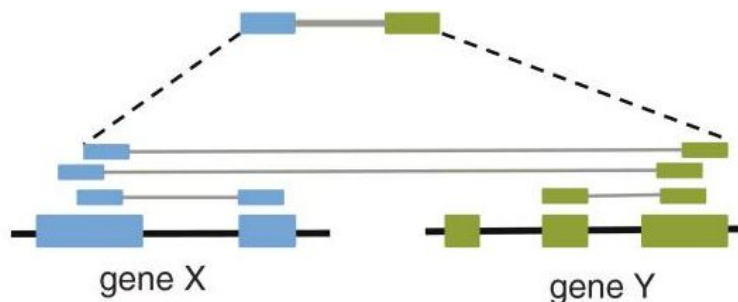
## Identification of fusion genes in breast cancer by paired-end RNA-sequencing.

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

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

[http://en.wikipedia.org/wiki/Fusion\\_gene](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

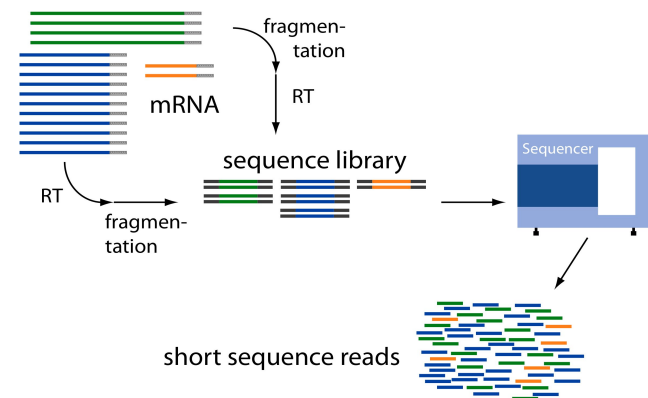
EST	PCR/RT-QPCR	SAGE	MicroArrays
No quantification	Quantification	Quantification	Indirect quantification
Low throughput	Low throughput (up to hundreds)	Low throughput (up to thousands)	High throughput (up to millions)
Discovery (Yes)	No	No	Discovery (Yes)

→ Need transcript sequence partially known

→ Difficulties in discovering novel 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, ...



# SGS platforms

## Séquenceurs 2<sup>ème</sup> génération (2013)

Société

Roche

Illumina

Life technologies

Plateforme



GS Junior

454

MiSeq

GAIIx

HiSeq  
1000/1500

HiSeq  
2000/2500

Ion PGM

Ion Proton

5500xl  
SOLiD

5500  
SOLiD

Technologie

Titanium

GS FLX+

Génome humain



Exome



Petit génome  
(Bactéries, levures)



Régions ciblées



Transcriptome



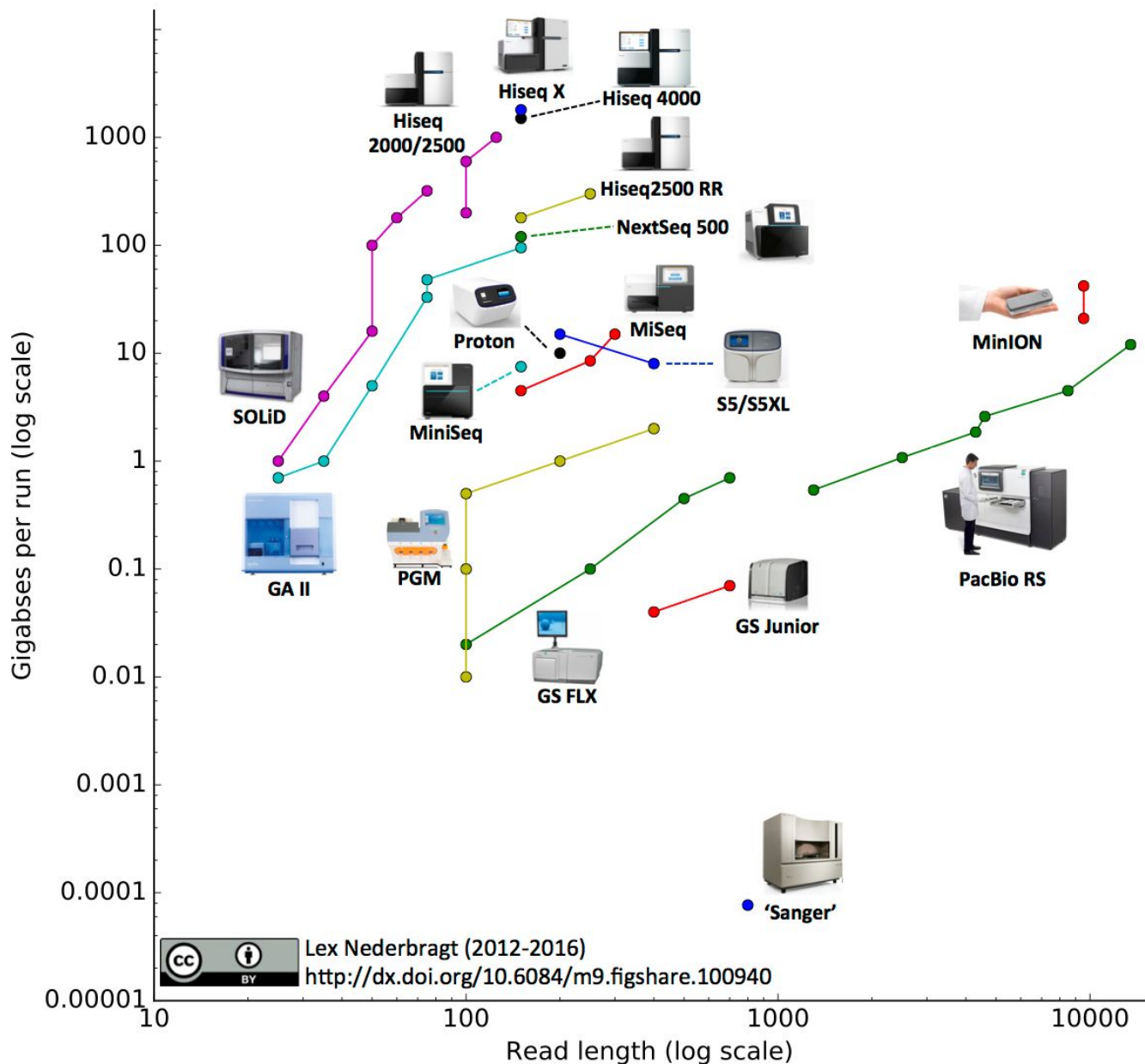
Chip-Seq



Métagénomique



# SGS platforms



Lex Nederbragt (2012-2016)  
<http://dx.doi.org/10.6084/m9.figshare.100940>

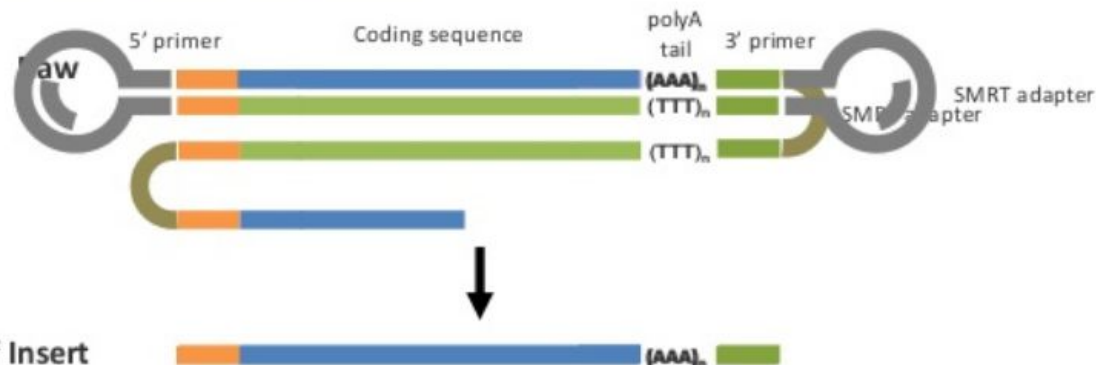
# PacBio : ISOseq

- Produce full-length transcripts without assembly (up to 10 kb in length)
- Discover isoform
- Can not be used for differential expression analysis

## Experimental Pipeline



## SampleNet: Iso-Seq Method with Clontech cDNA Synthesis Kit



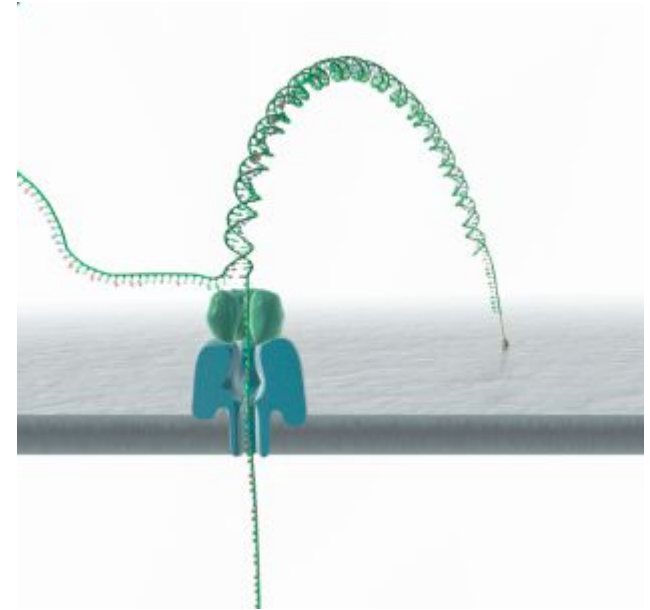


# MinION

Available now for sequencing cDNAs

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

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



Coming next: direct analysis of RNA

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

# 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

<https://www.encodeproject.org/documents/cede0cbe-d324-4ce7-ace4-f0c3eddf5972>

- Encode (2016) : [/@@@download/attachment/ENCODE%20Best%20Practices%20for%20RNA\\_v2.pdf](https://www.encodeproject.org/documents/cede0cbe-d324-4ce7-ace4-f0c3eddf5972)
  - 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.

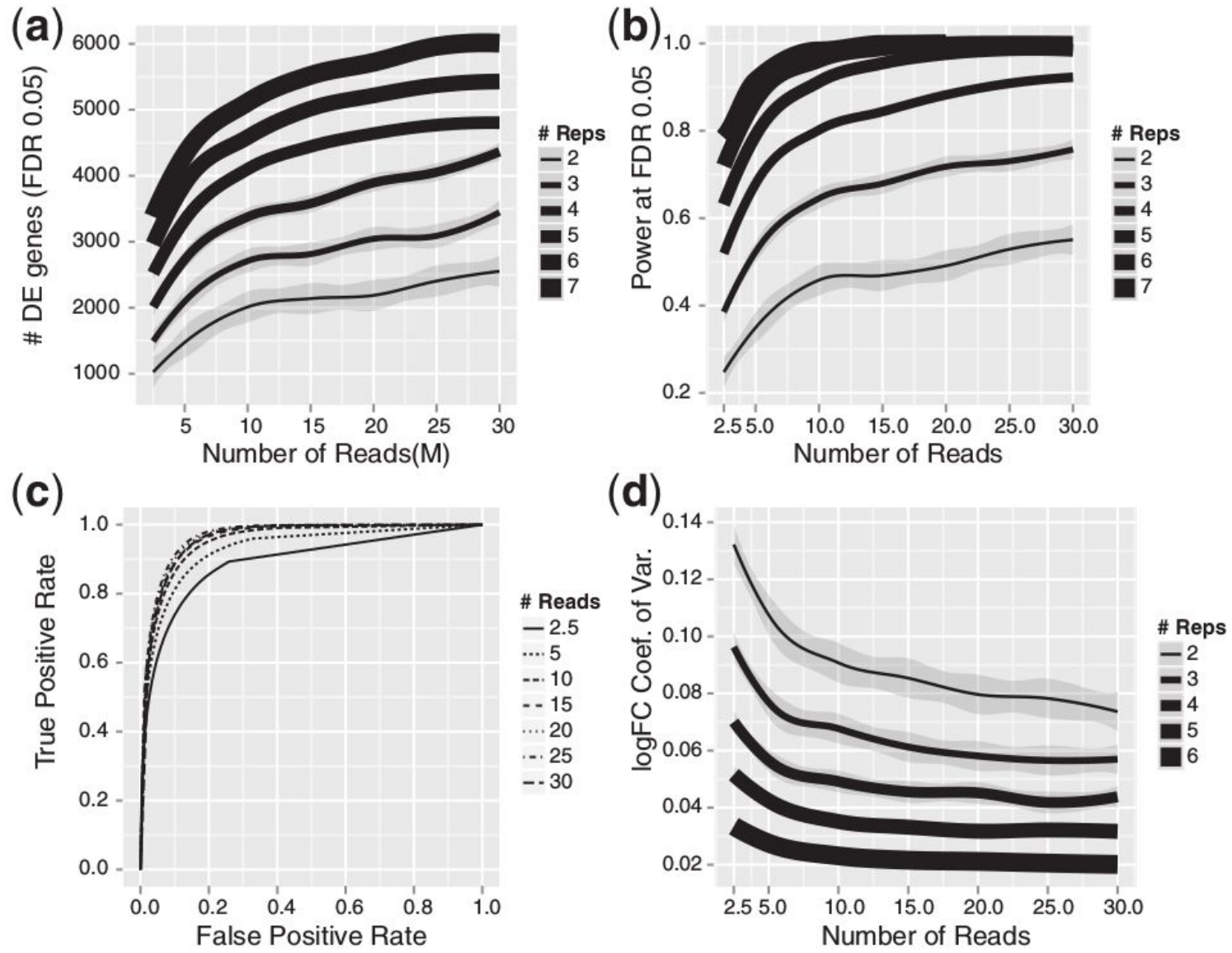
# Depth VS Replicates

Gene expression

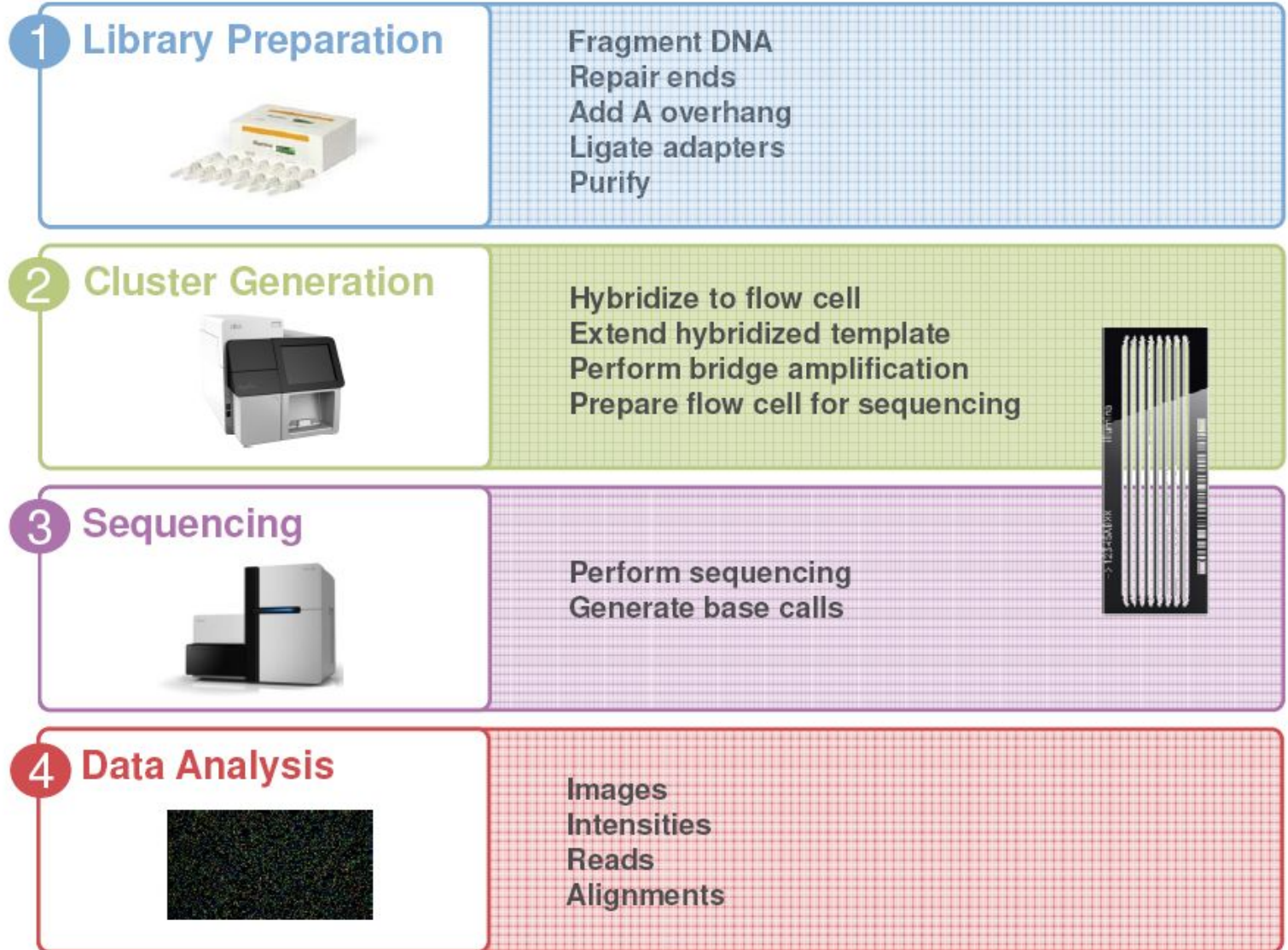
Advance Access publication December 6, 2013

## RNA-seq differential expression studies: more sequence or more replication?

Yuwen Liu<sup>1,2</sup>, Jie Zhou<sup>1,3</sup> and Kevin P. White<sup>1,2,3,\*</sup>



# 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

Préparation des Echantillons biologiques pour le RNAseq

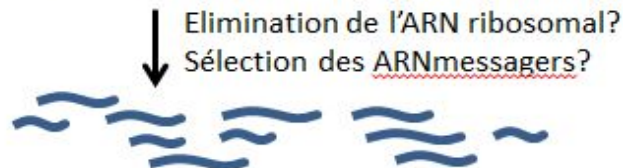
1. ARN messenger ou ARN total



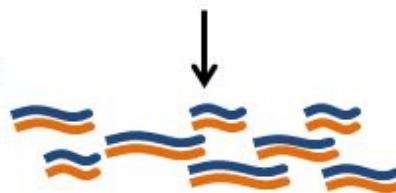
2. Elimination de l'ADN contaminant



3. Fragmentation de l'ARN



4. Retro-transcription de l'ARN en cDNA, hybride d'ADN/ARN



5. Synthèse du second brin d'ADN et ligation d'adaptateurs



6. Sélection des fragments par la taille



7. Séquençage des extrémités et production de « reads »



# Clusters generation / Sequencing

1. Attach DNA to flow cell

2. Perform bridge amplification

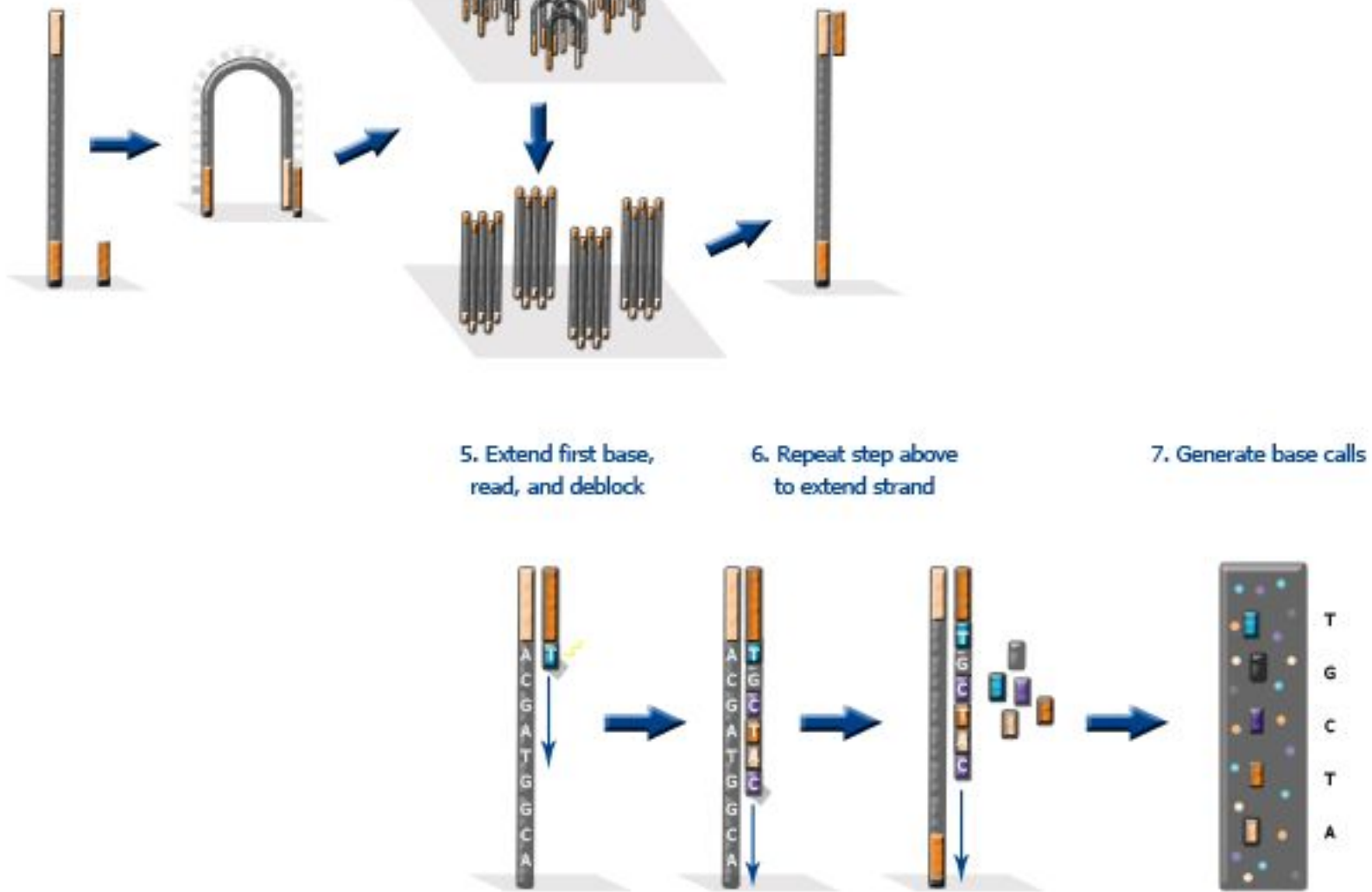
3. Generate clusters

4. Anneal sequencing primer

5. Extend first base, read, and deblock

6. Repeat step above to extend strand

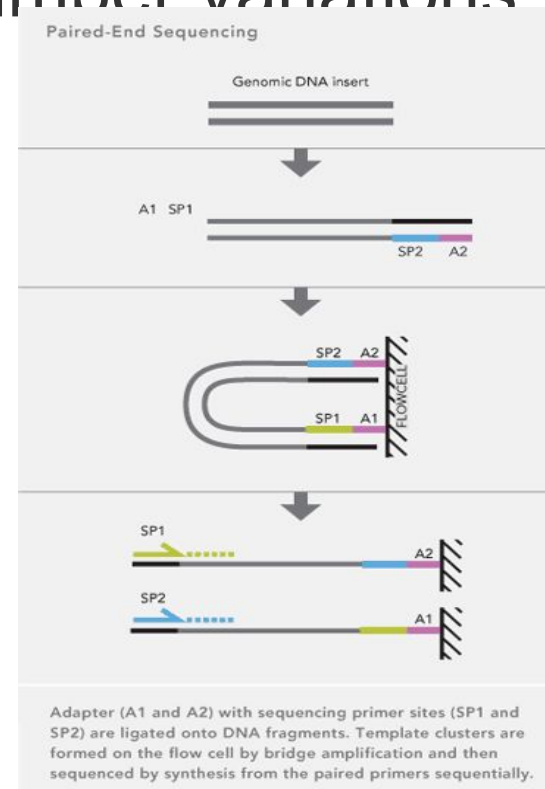
7. Generate base calls





# Paired-end sequencing

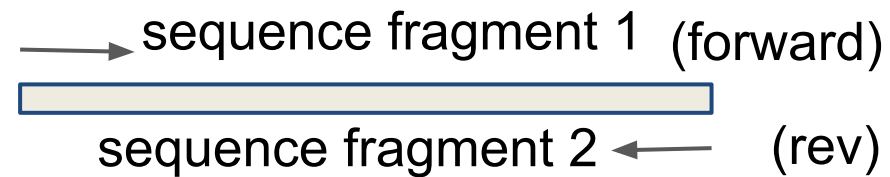
- Modification of the standard single-read DNA library preparation facilitates reading both ends of each fragment
- Improvement of mapping
- Help to detect structural variations in the genome like insertions or deletions, copy number variations and genome rearrangements



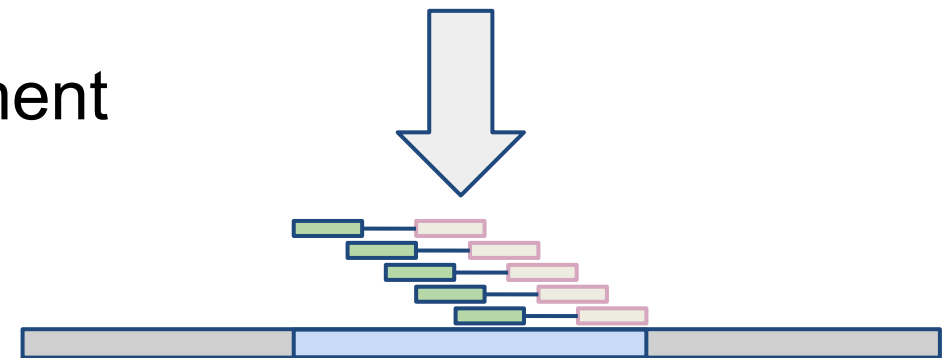
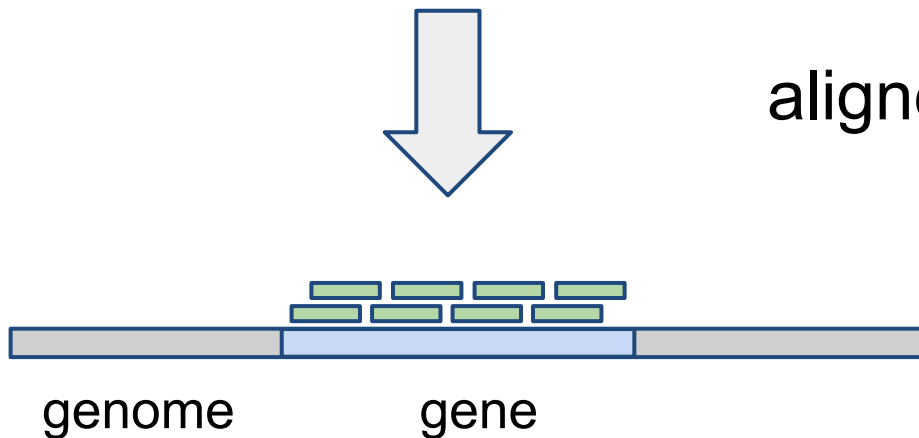
# Paired-end VS single-end

## Single-end

## Paired-end



alignement



- The cDNA size give 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.

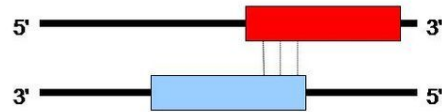
## Comprehensive comparative analysis of strand-specific RNA sequencing methods.

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

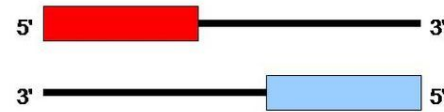
Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA.  
jlevin@broadinstitute.org

**A**

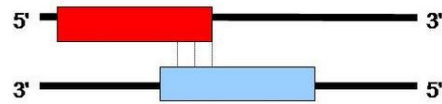
1. Head to Head: 5' to 5' overlap



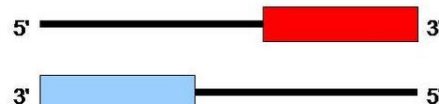
4. Nearby Tail to Tail



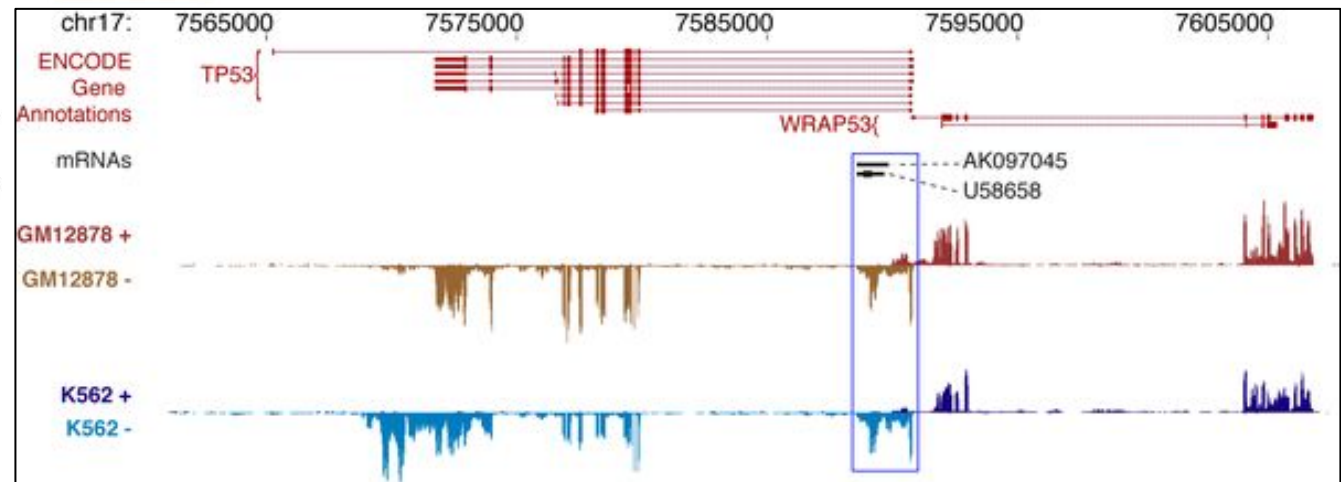
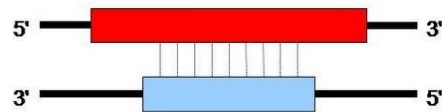
2. Tail to Tail: 3' to 3' overlap



5. Nearby Head to Head



3. Full overlap



# 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

Search

Examples: [BN000065](#), [histone](#)

[Advanced](#)  
[Sequence](#)

[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 EMBL-EBI 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

[Contact Helpdesk](#) 

Experiment: ERX1604042

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

View: [XML](#)

Download: [XML](#)

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

<b>Library Layout</b>	<b>Library Strategy</b>	<b>Library Source</b>	<b>Library Selection</b>	<b>Library Name</b>
PAIRED	RNA-Seq	TRANSCRIPTOMIC	cDNA	Sample 1_p

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

This table contains the files for experiment ERX1604042

[Bulk Download Files](#)

Download:  -  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 files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galaxy)
<a href="#">PRJEB14805</a>	<a href="#">SAMEA4079218</a>	<a href="#">ERS1250328</a>	<a href="#">ERX1604042</a>	<a href="#">ERR1533150</a>	62890	<a href="#">Solanum habrochaites</a>	ILLUMINA	PAIRED	19,975,820	<a href="#">File 1</a> <a href="#">File 2</a>	<a href="#">File 1</a> <a href="#">File 2</a>	<a href="#">Fastq file 1</a> <a href="#">Fastq file 2</a>	<a href="#">Fastq file 1</a> <a href="#">Fastq file 2</a>	<a href="#">File 1</a>	<a href="#">File 1</a>		

# Retrieve public data

  
Examples: [BN000065](#), [histone](#)  
[Search](#)  
[Advanced Sequence](#)

- 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 ERA007448 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 ERR000916 ),

<dir2> does not exist if the run accession has six digits. For example, fastq files for run ERR000916 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 Site map All databases Search

**Sequence Read Archive**

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

**Overview**

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

SRA is NIH's primary archive of high-throughput sequencing data and is part of the international partnership of archives (INSDC) 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

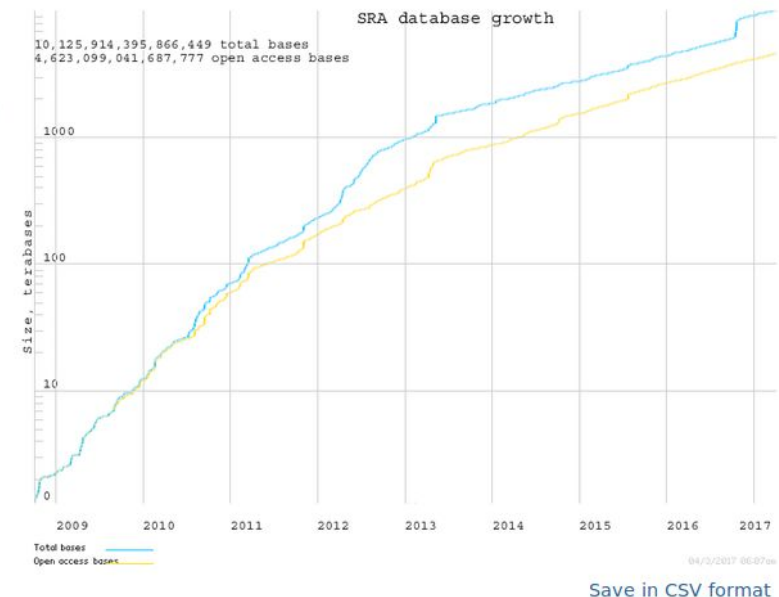
Making data available to the research community enhances reproducibility and allows for new discovery by comparing data sets.

- [Submission Quick Start](#)
- [Frequently Asked Questions](#)
- [Submitter Login](#)

## 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 [Site map](#) [All databases](#) [Search](#)

**Sequence Read Archive**

[Main](#) [Browse](#) [Search](#) [Download](#) [Submit](#) [Documentation](#) [Software](#) [Trace Archive](#) [Trace Assembly](#) [Trace BLAST](#)

**Studies** [Samples](#) [Analyses](#) [Run Browser](#) [Run Selector](#) [Provisional SRA](#)

Search:

[What can be entered in this field?](#)

## List of Studies. 421 records found.

<< < Page 1 / 17 > >>

#	<a href="#">Accession ↑</a>	<a href="#">Title</a>	<a href="#">Project</a>	<a href="#">Center</a>
1.	<a href="#">DRP000312</a>	Solanum lycopersicum strain:Micro-Tom Genome sequencing and assembly	<a href="#">59759</a>	KAZUSA
2.	<a href="#">DRP001059</a>	Resequencing data for tomato 'Ailsa Craig'	<a href="#">231443</a>	KAZUSA
3.	<a href="#">DRP001060</a>	Resequencing data for tomato 'Furikoma'	<a href="#">231443</a>	KAZUSA
4.	<a href="#">DRP001061</a>	Resequencing data for tomato 'M82'	<a href="#">231443</a>	KAZUSA
5.	<a href="#">DRP001062</a>	Resequencing data for tomato 'Tomato Chuukanbohon Nou 11'	<a href="#">231443</a>	KAZUSA
6.	<a href="#">DRP001063</a>	Resequencing data for tomato 'Ponderosa'	<a href="#">231443</a>	KAZUSA
7.	<a href="#">DRP001064</a>	Resequencing data for tomato 'Regina'	<a href="#">231443</a>	KAZUSA
8.	<a href="#">DRP001954</a>	Tomato genome sequence	<a href="#">259841</a>	TSUKUBA
9.	<a href="#">DRP002514</a>	Whole-genome sequencing of tomato mutants	<a href="#">275947</a>	KAZUSA
10.	<a href="#">DRP002631</a>	RNAseq in a sunlight-type plant factory	<a href="#">283367</a>	OSAKA_PREF
11.	<a href="#">DRP002638</a>	continuous light tomato RNAseq	<a href="#">283366</a>	OSAKA_PREF
12.	<a href="#">DRP002905</a>	Whole genome shotgun sequencing for 96 tomato cultivars	<a href="#">313365</a>	KAZUSA
13.	<a href="#">DRP003058</a>	RAD-Seq for tomato	<a href="#">315247</a>	KAZUSA
14.	<a href="#">DRP003091</a>	Time-course transcriptome data of Sly-Summer in sunlight-type plant factory	<a href="#">318884</a>	OSAKA_PREF
15.	<a href="#">DRP003147</a>	Strategic Innovation Promotion Program	<a href="#">324478</a>	RIKEN_BRC
16.	<a href="#">DRP003540</a>	Transcriptional profiling comparison during AM development between L. japonicus and tomato	<a href="#">380093</a>	SHINSHU
17.	<a href="#">ERP001270</a>	Carbon nanotubes as fertilizers: effects on tomato growth, reproductive system and soil microbial community	<a href="#">204399</a>	NCTR
18.	<a href="#">ERP001999</a>	Defining root colonization strategies in cucumber, tomato, maize and wheat plant species	<a href="#">204914</a>	ARO-VOLCANI
19.	<a href="#">ERP002018</a>	Bacterial communities associated with the surfaces of fresh fruits and vegetables	<a href="#">205672</a>	CCME-COLORADO
20.	<a href="#">ERP002550</a>	Resequencing Solanaceae (Potato and Tomato) 19th century samples	<a href="#">204997</a>	MPI-TUEBINGEN
21.	<a href="#">ERP002552</a>	Resequencing Phytophthora strains	<a href="#">204996</a>	MPI-TUEBINGEN
22.	<a href="#">ERP002648</a>	Using a periclinal chimera to determine layer-specific gene expression	<a href="#">225680</a>	ICL-CFB

# Retrieve public data

NCBI SRA Run Selector Help Permalink

Search:

Facets

- Run
- BioSample
- Sample name
- MBases
- MBytes
- Experiment
- sample name
- sample title


Hide common fields

Assay Type: RNA-Seq  
AvgSpotLen: 49  
BioProject: [PRJDB3892](#)  
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](#)  
bioproject id: PRJDB3892  
cultivar: Taian-kichijitsu  
tissue type: leaf

	Runs	Bytes	Bases	Download
Total:	50	1.58 Gb	2.81 G	<a href="#">RunInfo Table</a> <a href="#">Accession List</a>
Selected:				<a href="#">RunInfo Table</a> <a href="#">Accession List</a>

50 Runs found

	Run	BioSample	Sample name	MBases	MBytes	Experiment	sample name	sample title
<input type="checkbox"/>	<a href="#">DRR034293</a>	<a href="#">SAMD00029631</a>	DRS019544	53	30	<a href="#">DRX030926</a>	SunB30	Sunlight tomato Bset Time30
<input type="checkbox"/>	<a href="#">DRR034294</a>	<a href="#">SAMD00029632</a>	DRS019545	59	34	<a href="#">DRX030927</a>	SunB32	Sunlight tomato Bset Time32
<input type="checkbox"/>	<a href="#">DRR034295</a>	<a href="#">SAMD00029633</a>	DRS019546	76	44	<a href="#">DRX030928</a>	SunB34	Sunlight tomato Bset Time34
<input type="checkbox"/>	<a href="#">DRR034296</a>	<a href="#">SAMD00029634</a>	DRS019547	56	32	<a href="#">DRX030929</a>	SunB36	Sunlight tomato Bset Time36
<input type="checkbox"/>	<a href="#">DRR034298</a>	<a href="#">SAMD00029636</a>	DRS019549	55	32	<a href="#">DRX030931</a>	SunB4	Sunlight tomato Bset Time4
<input type="checkbox"/>	<a href="#">DRR034299</a>	<a href="#">SAMD00029637</a>	DRS019550	70	40	<a href="#">DRX030932</a>	SunB40	Sunlight tomato Bset Time40
<input type="checkbox"/>	<a href="#">DRR034300</a>	<a href="#">SAMD00029638</a>	DRS019551	56	32	<a href="#">DRX030933</a>	SunB42	Sunlight tomato Bset Time42
<input type="checkbox"/>	<a href="#">DRR034301</a>	<a href="#">SAMD00029639</a>	DRS019552	50	29	<a href="#">DRX030934</a>	SunB44	Sunlight tomato Bset Time44
<input type="checkbox"/>	<a href="#">DRR034287</a>	<a href="#">SAMD00029625</a>	DRS019538	61	35	<a href="#">DRX030920</a>	SunB2	Sunlight tomato Bset Time2
<input type="checkbox"/>	<a href="#">DRR034302</a>	<a href="#">SAMD00029640</a>	DRS019553	78	45	<a href="#">DRX030935</a>	SunB46	Sunlight tomato Bset Time46



# Retrieve public data

NCBI SRA Run Selector [Help](#) [Permalink](#)

Search:

- Facets
- Run
  - BioSample
  - Sample name
  - MBases
  - MBytes
  - Experiment
  - sample name
  - sample title

Hide common fields

Assay Type: RNA-Seq  
 AvgSpotLen: 49  
 BioProject: [PRJDB3892](#)  
 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](#)  
 bioproject id: PRJDB3892  
 cultivar: Taian-kichijitsu  
 tissue type: leaf

	Runs	Bytes	Bas
Total:	50	1.58 Gb	2
Selected:			

## 50 Runs found

<input type="checkbox"/>	Run	BioSample	Sample name	MBases	MBytes	Experiment	sample name	sample title
<input type="checkbox"/>	<a href="#">DRR034293</a>	<a href="#">SAMD00029631</a>	DRS019544	53	30	<a href="#">DRX030926</a>	SunB30	Sunlight tomato Bset Time30
<input type="checkbox"/>	<a href="#">DRR034294</a>	<a href="#">SAMD00029632</a>	DRS019545	59	34	<a href="#">DRX030927</a>	SunB32	Sunlight tomato Bset Time32
<input type="checkbox"/>	<a href="#">DRR034295</a>	<a href="#">SAMD00029633</a>	DRS019546	76	44	<a href="#">DRX030928</a>	SunB34	Sunlight tomato Bset Time34
<input type="checkbox"/>	<a href="#">DRR034296</a>	<a href="#">SAMD00029634</a>	DRS019547	56	32	<a href="#">DRX030929</a>	SunB36	Sunlight tomato Bset Time36
<input type="checkbox"/>	<a href="#">DRR034298</a>	<a href="#">SAMD00029636</a>	DRS019549	55	32	<a href="#">DRX030931</a>	SunB4	Sunlight tomato Bset Time4
<input type="checkbox"/>	<a href="#">DRR034299</a>	<a href="#">SAMD00029637</a>	DRS019550	70	40	<a href="#">DRX030932</a>	SunB40	Sunlight tomato Bset Time40
<input type="checkbox"/>	<a href="#">DRR034300</a>	<a href="#">SAMD00029638</a>	DRS019551	56	32	<a href="#">DRX030933</a>	SunB42	Sunlight tomato Bset Time42
<input type="checkbox"/>	<a href="#">DRR034301</a>	<a href="#">SAMD00029639</a>	DRS019552	50	29	<a href="#">DRX030934</a>	SunB44	Sunlight tomato Bset Time44
<input type="checkbox"/>	<a href="#">DRR034287</a>	<a href="#">SAMD00029625</a>	DRS019538	61	35	<a href="#">DRX030920</a>	SunB2	Sunlight tomato Bset Time2
<input type="checkbox"/>	<a href="#">DRR034302</a>	<a href="#">SAMD00029640</a>	DRS019553	78	45	<a href="#">DRX030935</a>	SunB46	Sunlight tomato Bset Time46

SRR\_Acc\_List.txt (/tmp/mozilla\_choedeO) - gedit

Fichier Édition Affichage Rechercher Outils Documents

Ouvrir Enregistrer Annuler

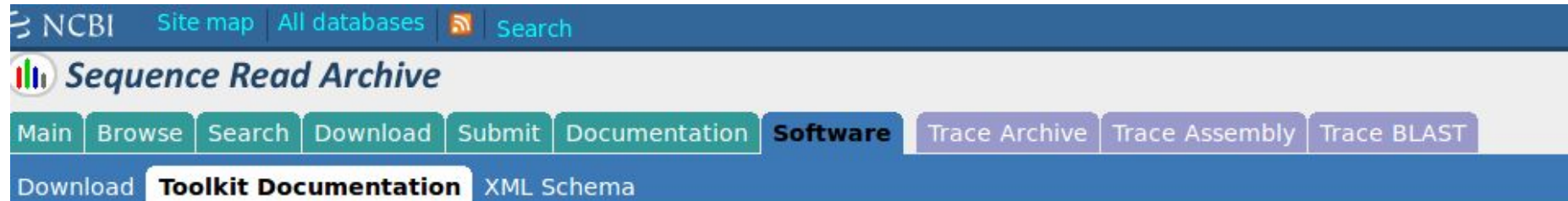
SRR\_Acc\_List.txt

```

DRR034293
DRR034294
DRR034295
DRR034296
DRR034298
DRR034299
DRR034300
DRR034301
DRR034287
DRR034302
DRR034291
DRR034305
DRR034290
DRR034292
DRR034303
  
```

Texte brut Largeur des tabulations : 8 Lig 1, Col 1 INS

# Retrieve public data



## 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

```
prefetch <sra_accession> --max-size  
(20G by default)
```

```
fastq-dump sra_file.sra
```

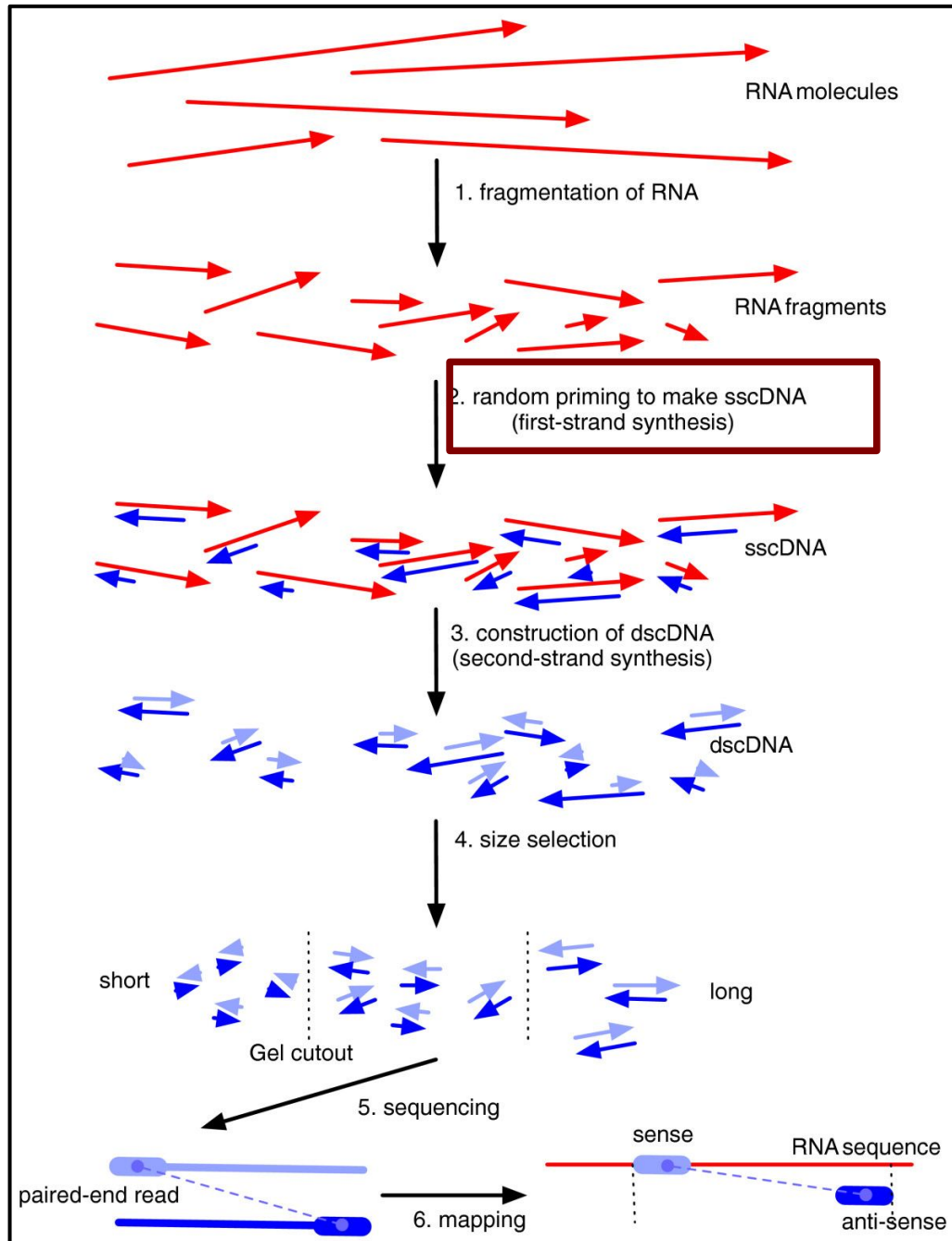
# Summary - Sequence quality

- 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 14 April 2010

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

## Biases in Illumina transcriptome sequencing caused by random hexamer priming

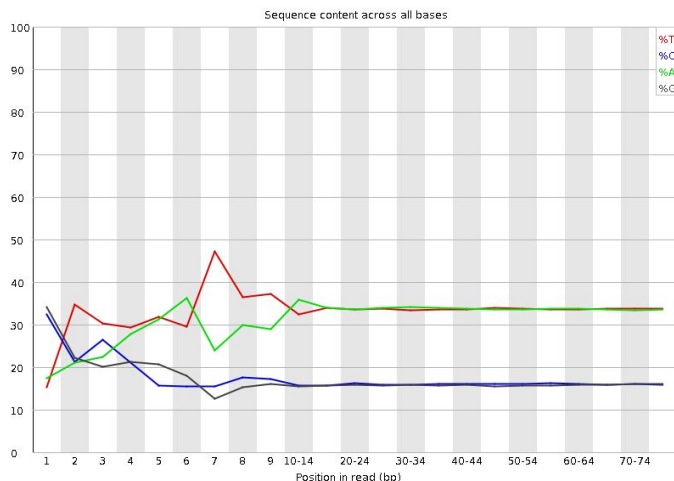
Kasper D. Hansen<sup>1,\*</sup>, Steven E. Brenner<sup>2</sup> and Sandrine Dudoit<sup>1,3</sup>

### ABSTRACT

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

-A strong distinctive pattern in the nucleotide frequencies of the first 13 positions at the 5'-end :

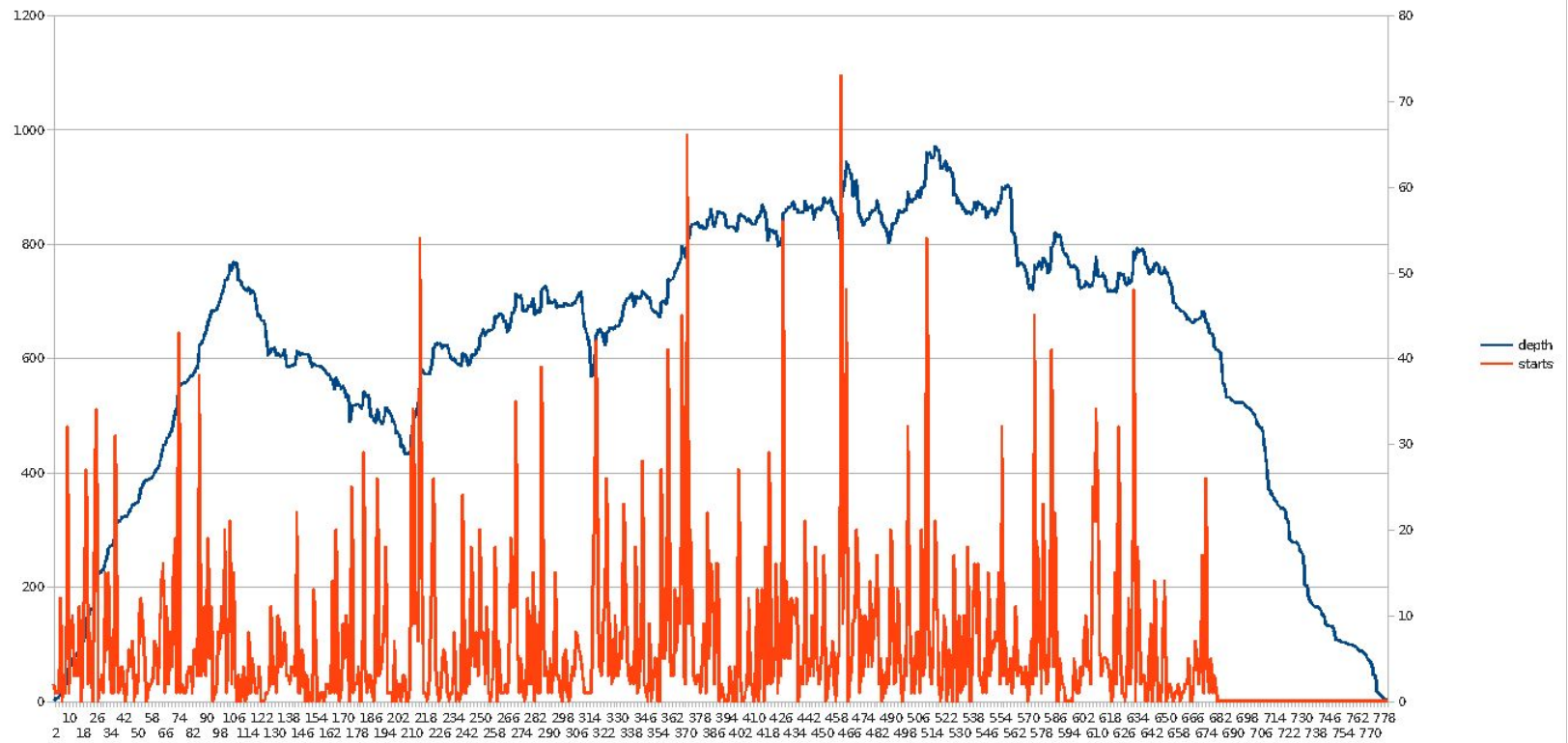
- sequence specificity of the polymerase
- due to the end repair performed



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



# Hexamer random effect



- Orange = reads start sites
- Blue = coverage

# Transcript length bias

Biol Direct. 2009 Apr 16;4:14.

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

Oshlack A, Wakefield MJ.

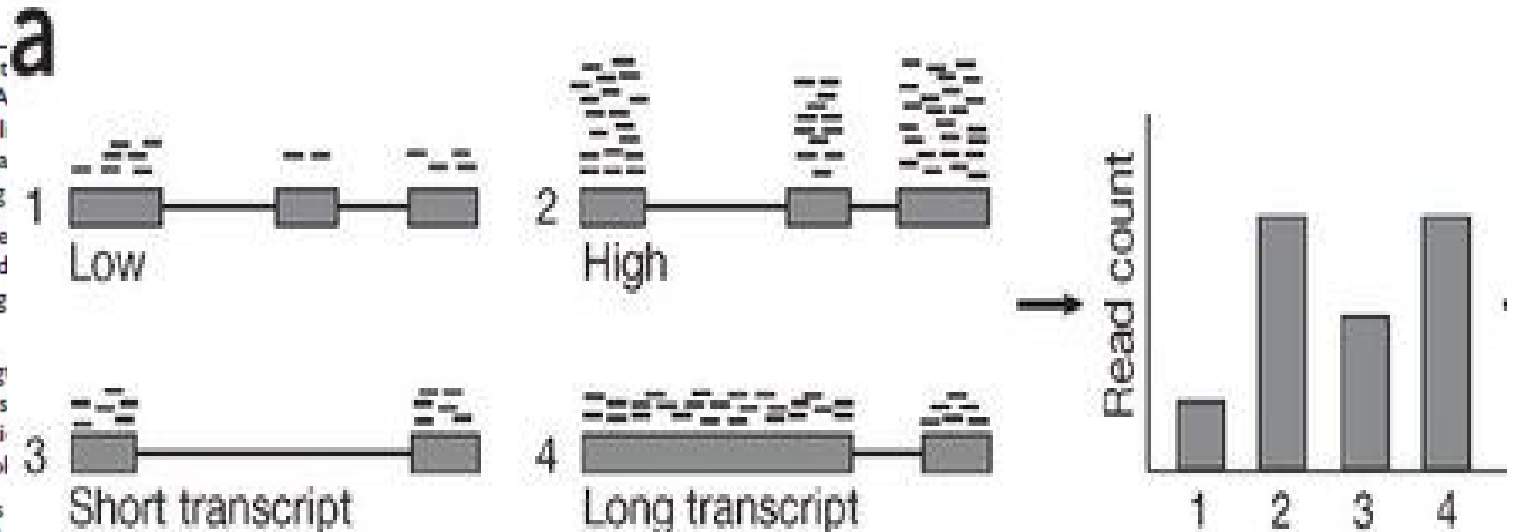
### Abstract

**Background:** Several recent transcriptome analysis (RNA genome transcriptional profile genomic sequences. As yet, a still in the stages of exploring

**Results:** We investigated the published data sets. For stand call differentially expressed g transcript.

**Conclusion:** Transcript leng current protocols for RNA-s expressed genes, and in parti other multi-gene systems biol

**Reviewers:** This article was Cloonan (nominated by Mark



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

BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 5 2011, pages 682-689  
doi:10.1093/bioinformatics/btr005

Gene expression

Advance Access publication January 19, 2011

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

Liyen Gao<sup>1,\*</sup>, Zhide Fang<sup>2,\*</sup>, Kui Zhang<sup>1</sup>, Degui Zhi<sup>1</sup> and Xiangqin Cui<sup>1,\*</sup>

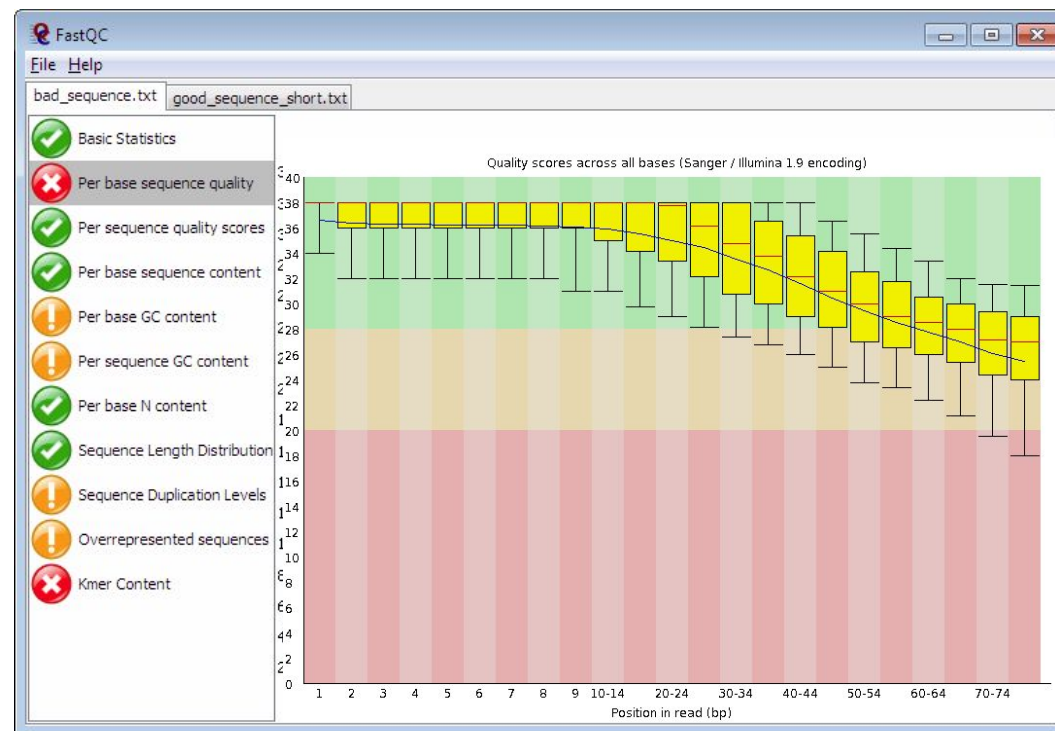
# Bias “mappability”

- 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*

# fastqQC 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).

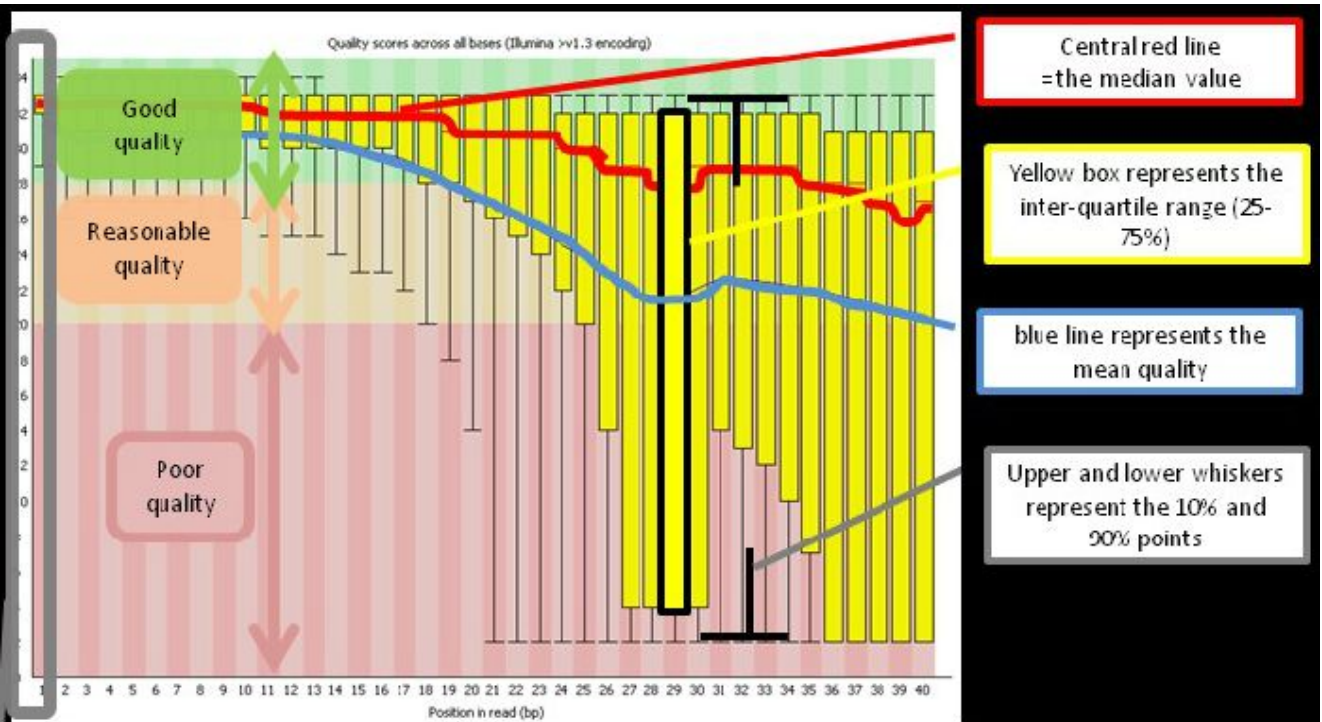
These evaluations must be taken in the context of what you expect from your library. A 'normal' sample as far as FastQC is concerned is random and diverse.

# fastqQC 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.

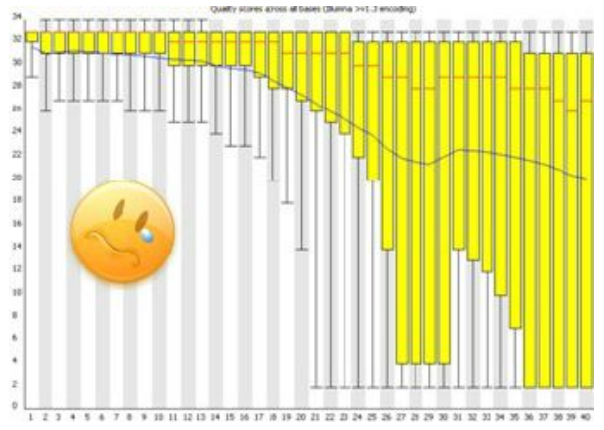
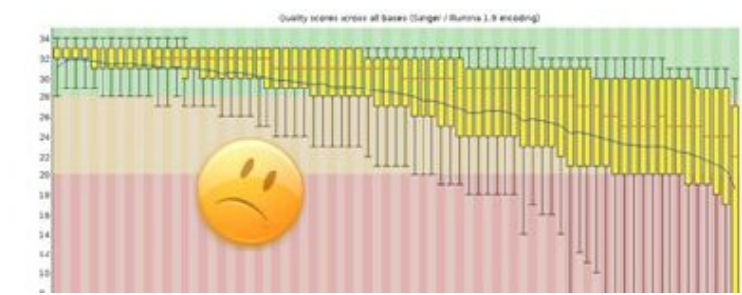


- Central red line = the median value
- Yellow box represents the inter-quartile range (25-75%)
- blue line represents the mean quality
- Upper and lower whiskers represent the 10% and 90% points

y-axis on the graph shows the quality scores. The higher the score the better the base call.

```
@ILLUMINA-GA_0000:1:1:2771:1022#0/1
TGACATNAAGCACTGTAGCTCATCTCGTATGCCGCTCTT
+ILLUMINA-GA_0000:1:1:2771:1022#0/1
faaWa]B\]^b`vcdfd_f_cd_f[d_bfaSadddfb
@ILLUMINA-GA_0000:1:1:3203:1022#0/1
TGAGATNAAGCACTGTAGCTCATCTCGTATGCCGCTCT
+ILLUMINA-GA_0000:1:1:3203:1022#0/1
janeABY_A)hjh`nncff6aenndeaxxxxxxxx
```

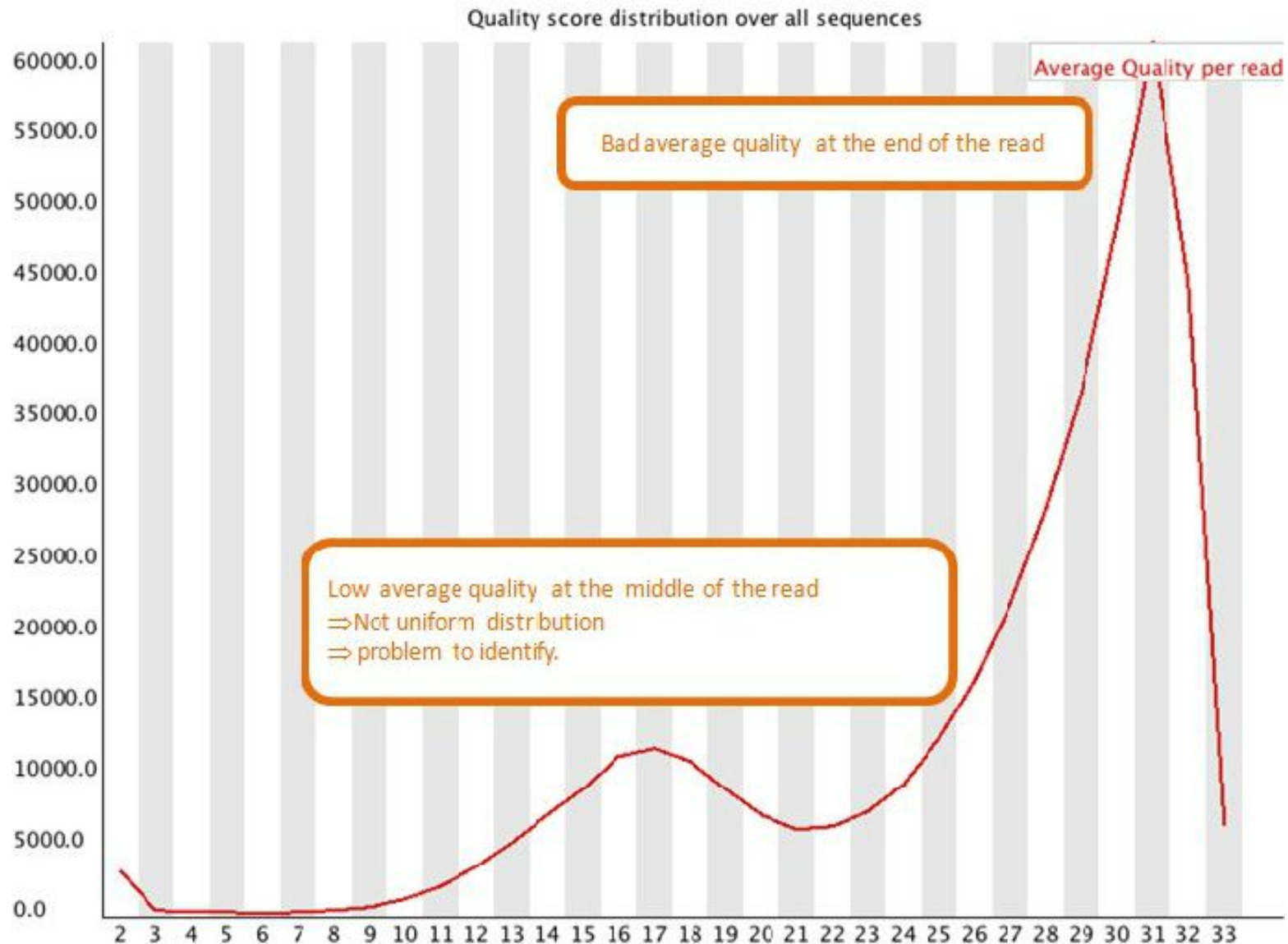
pos	mean	Q1	med	Q3
1	38.01	38	39	40
2	37.73	38	39	40
3	37.53	38	39	40
4	37.69	38	39	40
5	37.41	37	39	40
6	37.63	38	39	40



# fastqQC Report

## Statistics per Sequence Quality Score

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

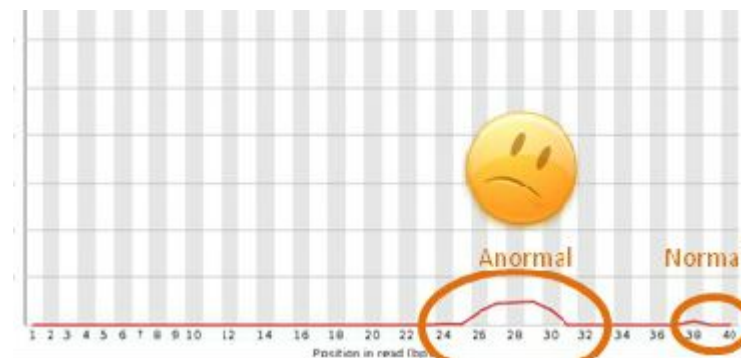
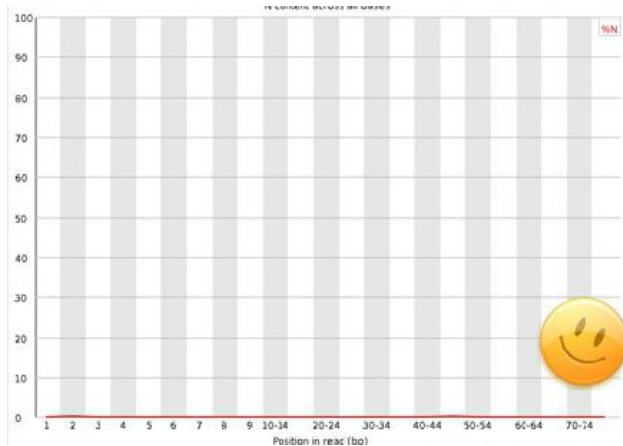


# fastqQC Report

## 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.



Proportion of Ns rises  
few % during the  
pipeline  
= Unable to interpret  
data

Low proportion of Ns  
at the end  
= Normal

17 19 21 23 25 27 29 31 33 35 37 39



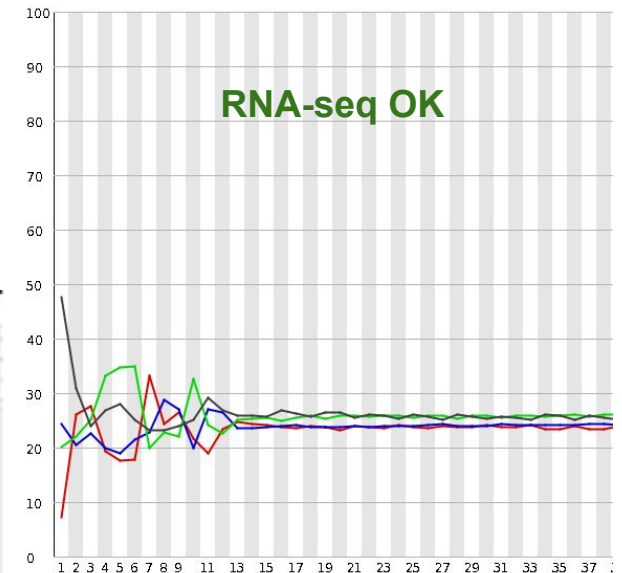
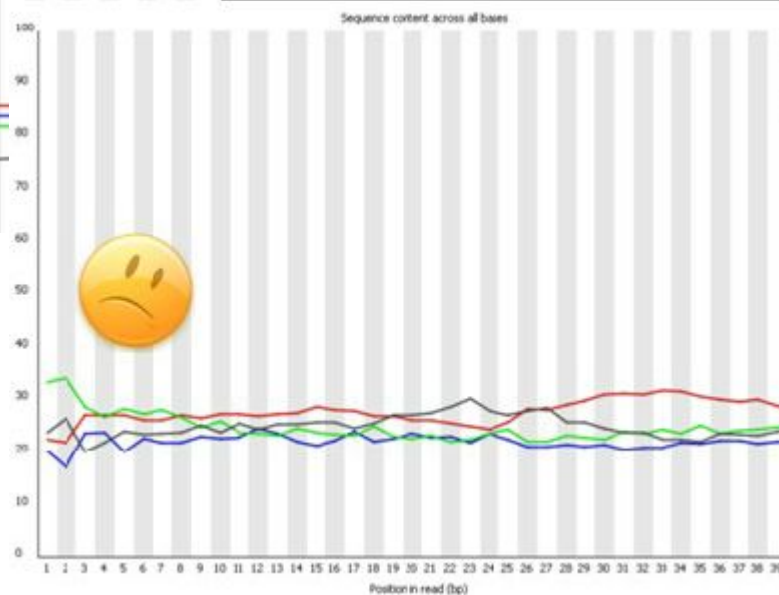
# fastqQC Report

## 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

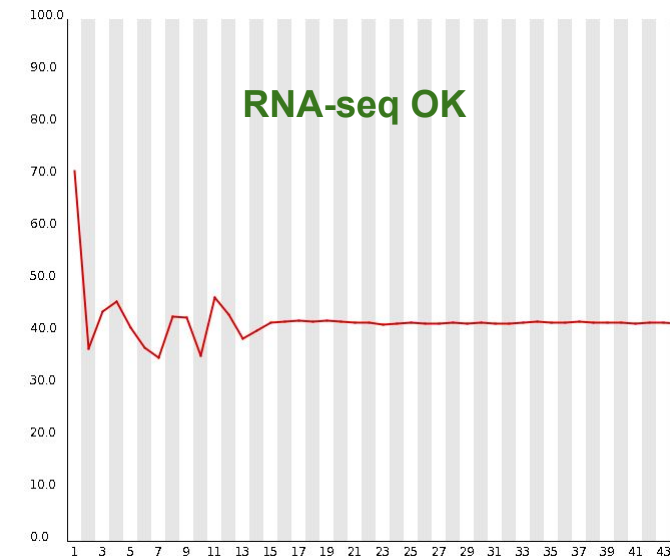
## Statistics per Base GC Distribution

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.

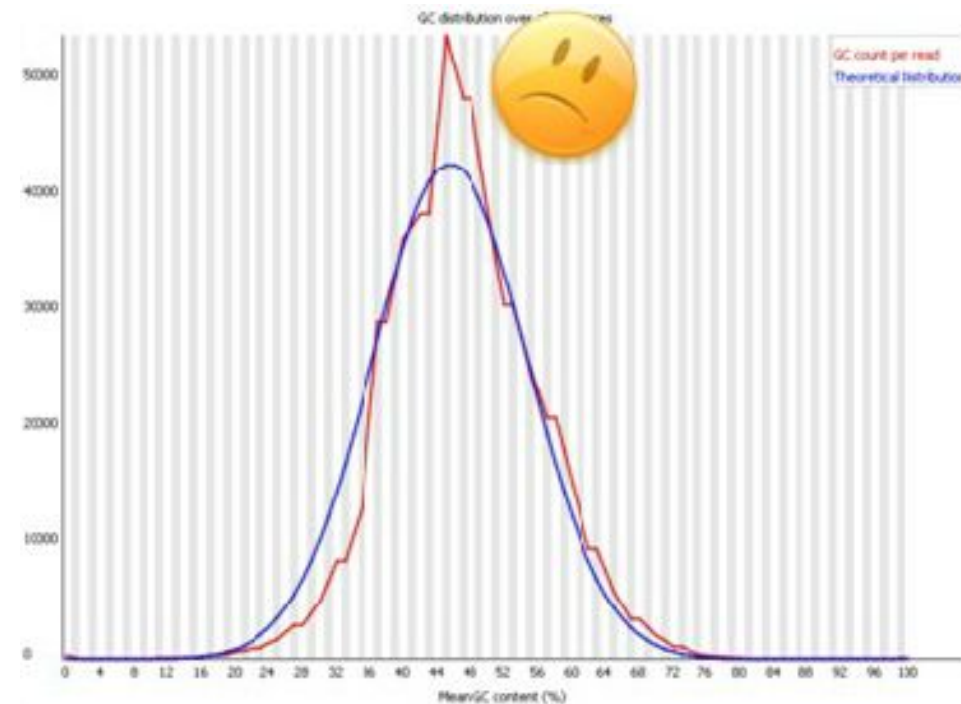
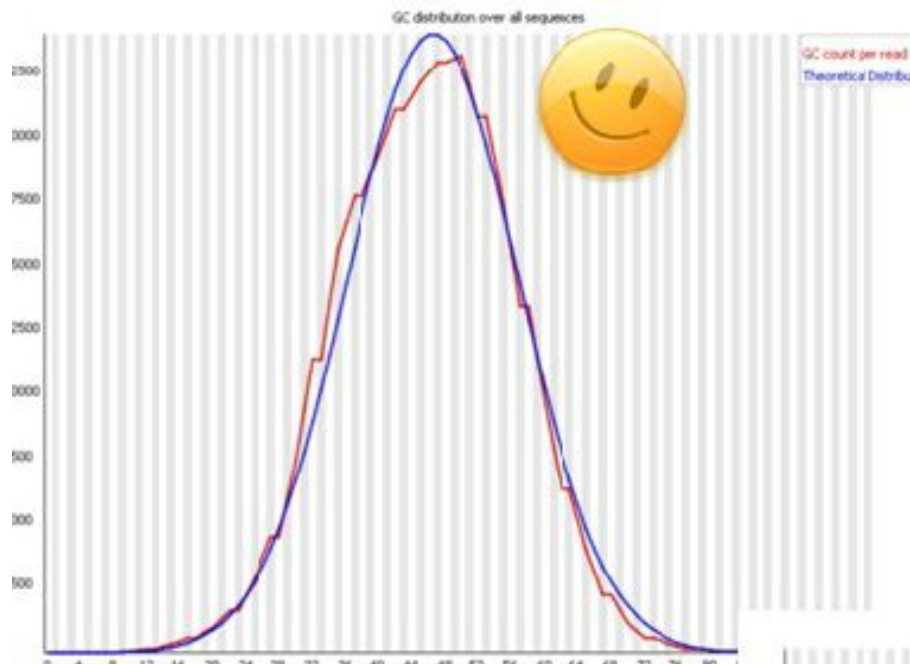
GC bias: changes in different bases, overrepresented sequence contaminating your library.  
=> plot not horizontally.



# 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.

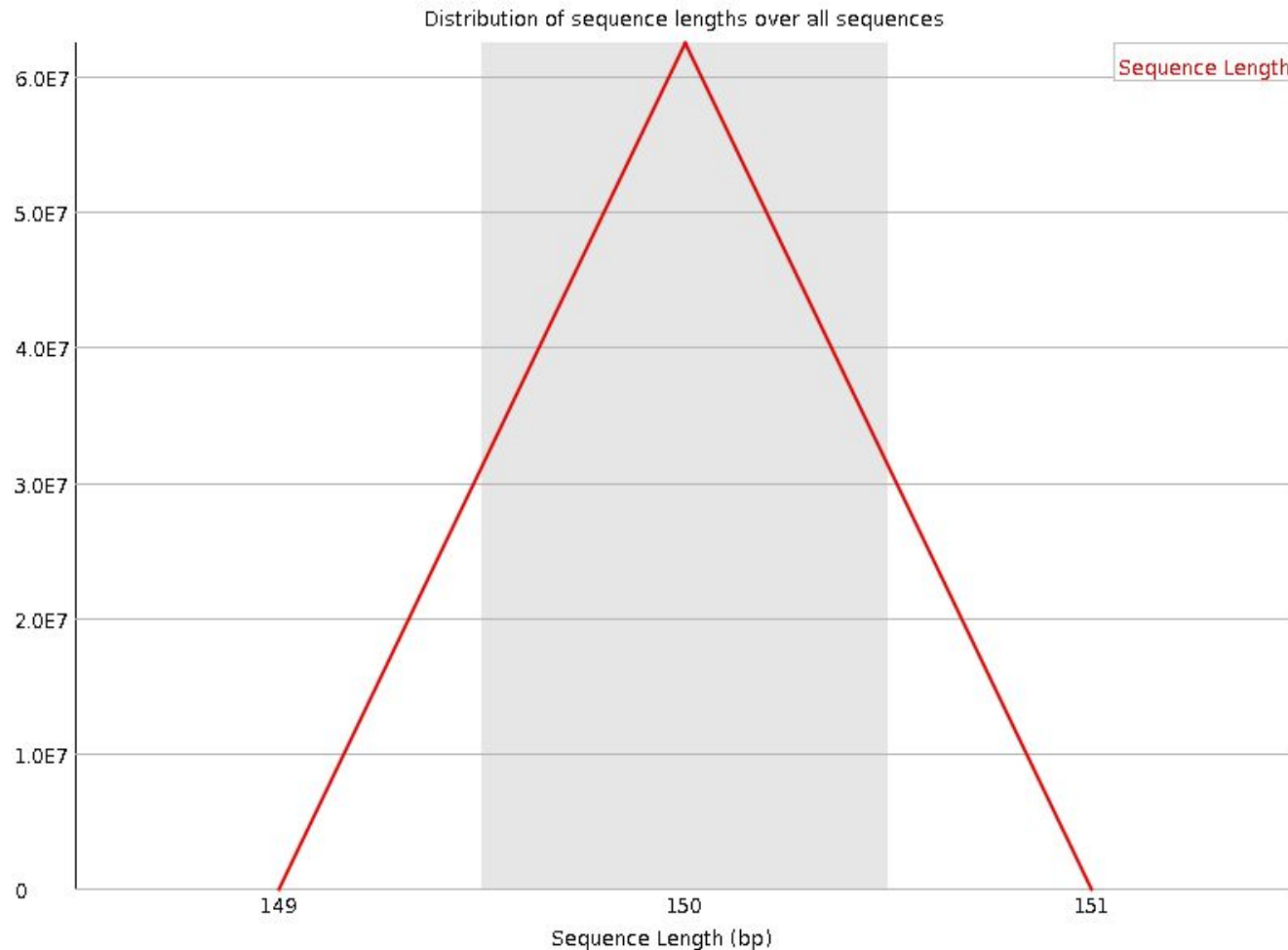


# fastqQC Report

## Statistics per Sequence Length Distribution

Some sequence fragments contain reads of wildly varying lengths.

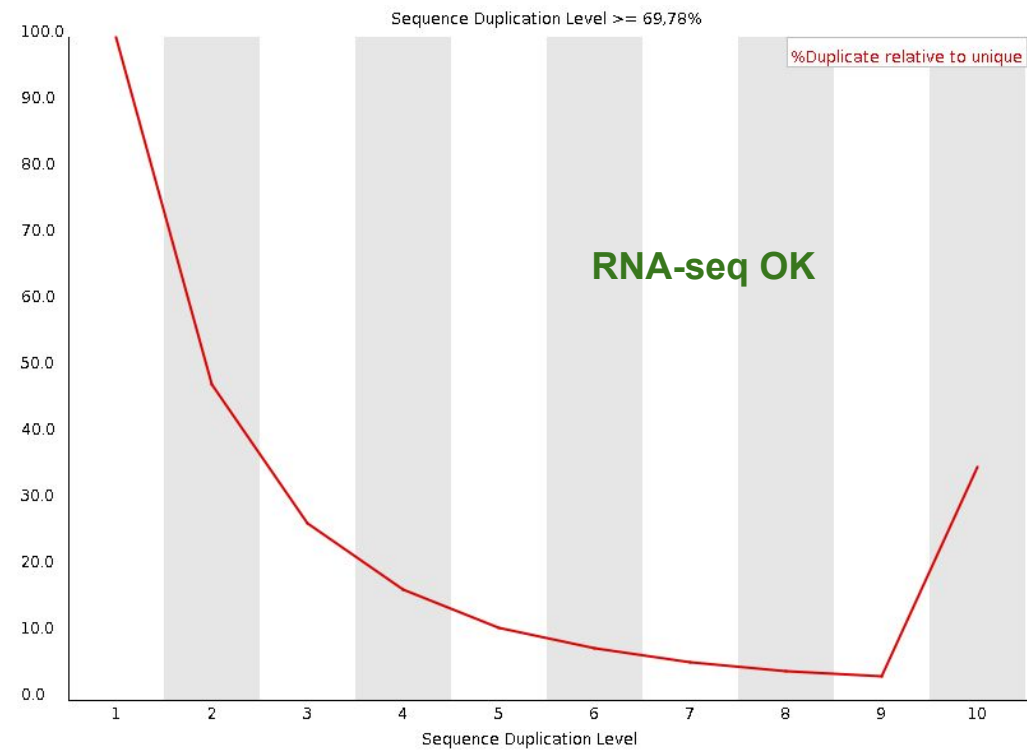
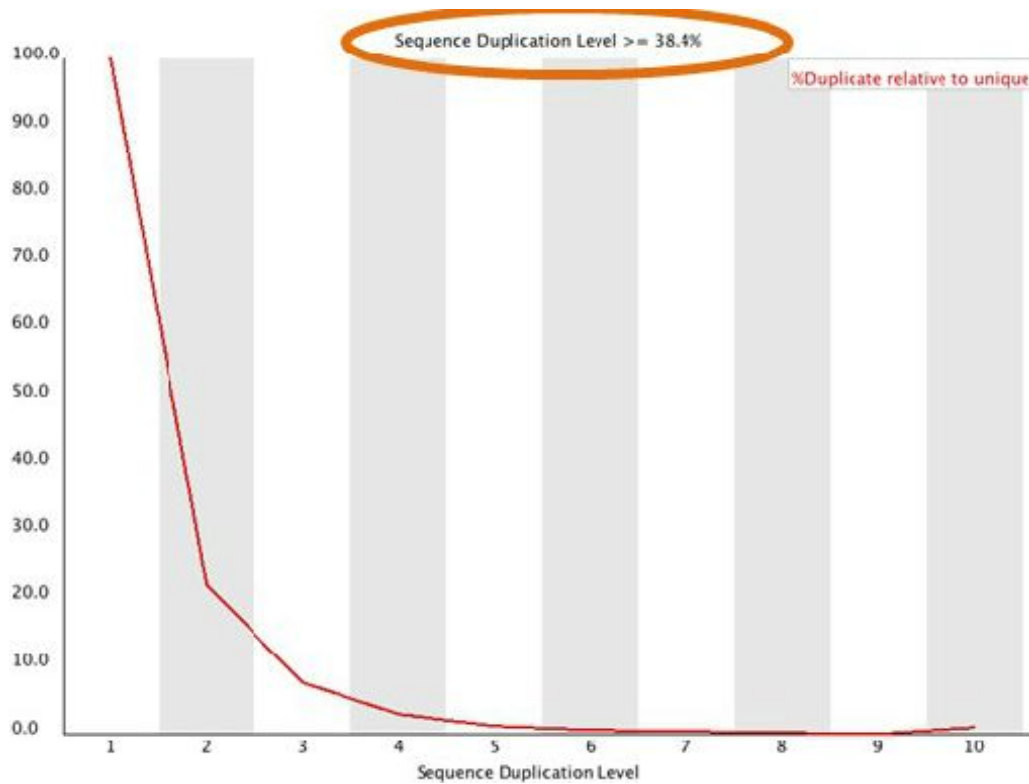
Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end.



# fastqQC Report

## Statistics per Duplicate Sequences

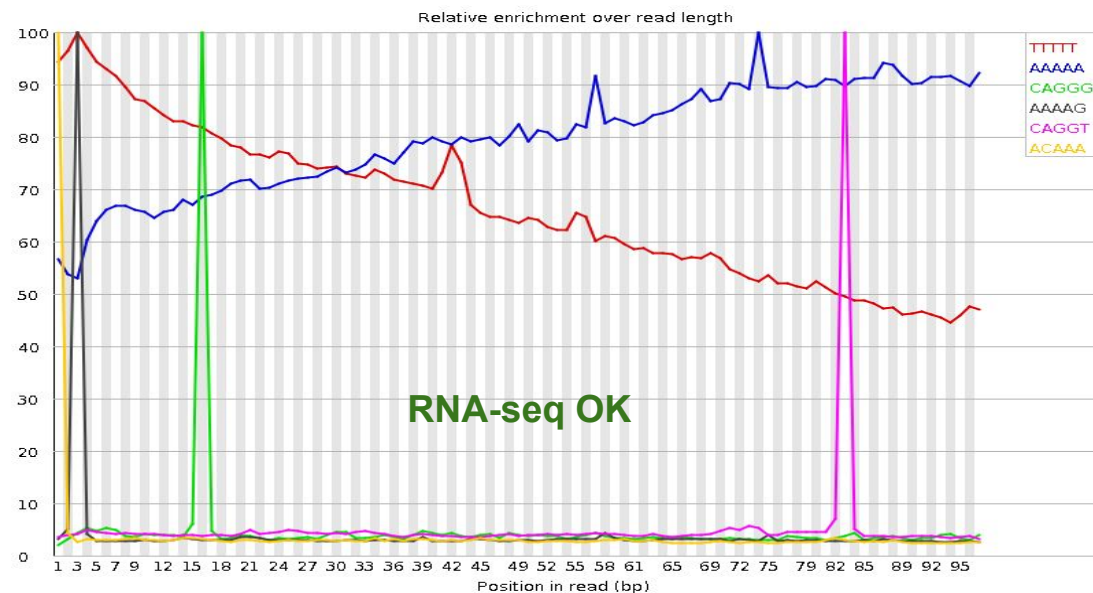
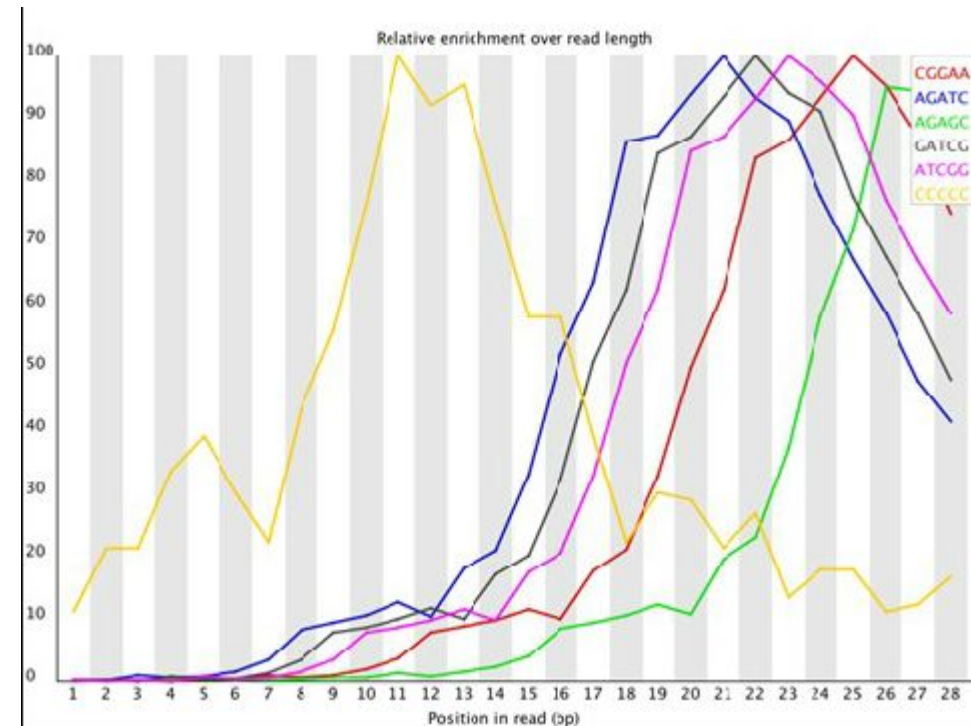
High level of duplication indicate an enrichment bias.



# 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 adapters.
- Check for adaptor sequence or poly-A sequence



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

# Take home message on quality analysis

Elements to be checked :

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

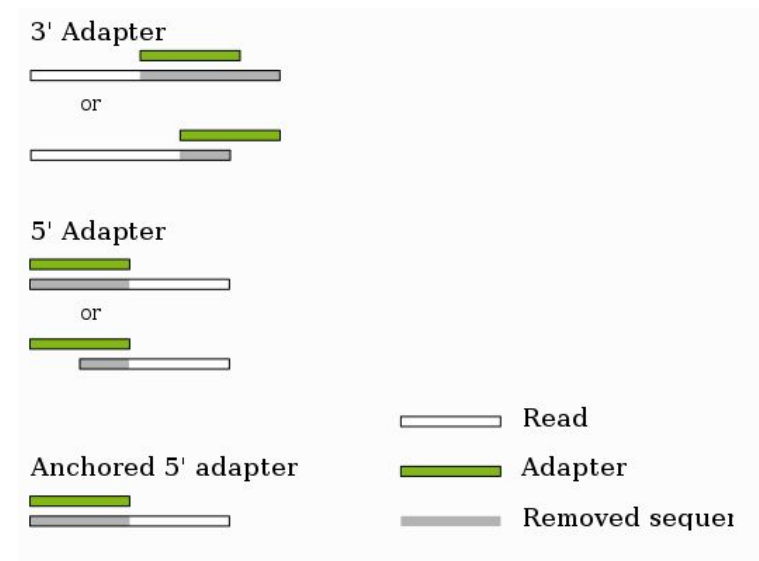
- Searches & removes adapter & tag in all reads.
- Trim quality
- Filter too short or untrimmed reads (in a separate output file).

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

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

(3' adapter, single read)

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.)

- Sickle trims the ends of the reads having poor quality.
- It's using sliding window instead of brutal threshold like cutadapt.
- Window length is 0.1 times the length of the read.
- The window slides along the quality values until the average quality in the window rises above the threshold, at which point the algorithm determines where within the window the rise occurs and cuts the read and quality there for end cut.

# 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, cutadapt and sickle*

***Exercise 3 : quality control of used datasets***

***Exercise 4: cleaning used datasets***





# 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

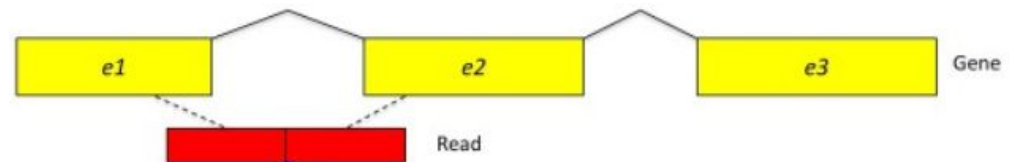
# Aim -

## 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.



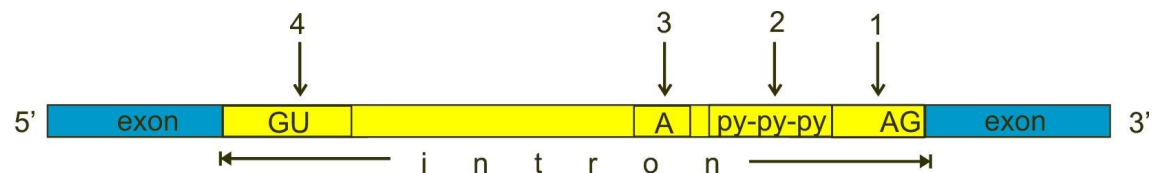
### And:

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



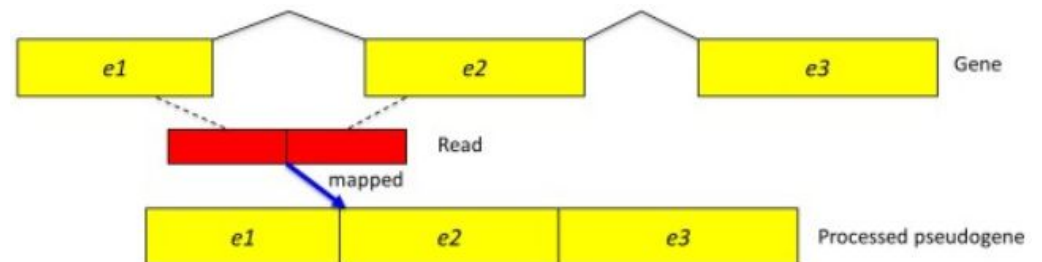
[http://en.wikipedia.org/wiki/RNA\\_splicing](http://en.wikipedia.org/wiki/RNA_splicing)

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

- Trans-splicing: Nucleic Acids Res. 2000 Nov 1;28(21):4364-75. Analysis of canonical and non-canonical splice sites in mammalian genomes. Burset M, Seledtsov IA, Solovjev VV. splicing that joins two exons that are not within the same RNA transcript

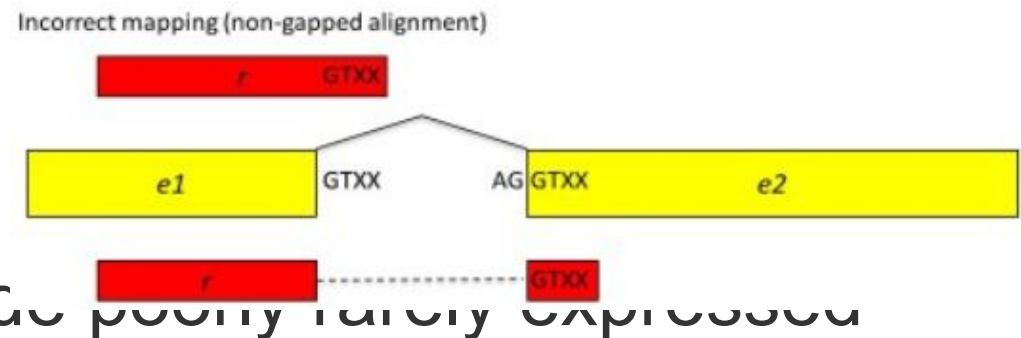
# Hard case

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



Kim et al, Genome Biology, 2013

- Small end “anchor”



- Unknown junction inside gene poorly rarely expressed

# Most used tools

## Tools for splice-mapping:

### - Tophat:

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

*Sequence analysis*

#### **TopHat: discovering splice junctions with RNA-Seq**

Cole Trapnell<sup>1,\*</sup>, Lior Pachter<sup>2</sup> and Steven L. Salzberg<sup>1</sup>

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

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

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

### - STAR:

#### **STAR: ultrafast universal RNA-seq aligner**

Alexander Dobin<sup>1\*</sup>, Carrie A. Davis<sup>1</sup>, Felix Schlesinger<sup>1</sup>, Jorg Drenkow<sup>1</sup>, Chris Zaleski<sup>1</sup>, Sonali Jha<sup>1</sup>, Philippe Batut<sup>1</sup>, Mark Chaisson<sup>2</sup> and Thomas R. Gingeras<sup>1</sup>

<sup>1</sup>Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA.

<sup>2</sup>Pacific Biosciences, Menlo Park, California, USA.

Associate Editor: Dr. Inanc Birol

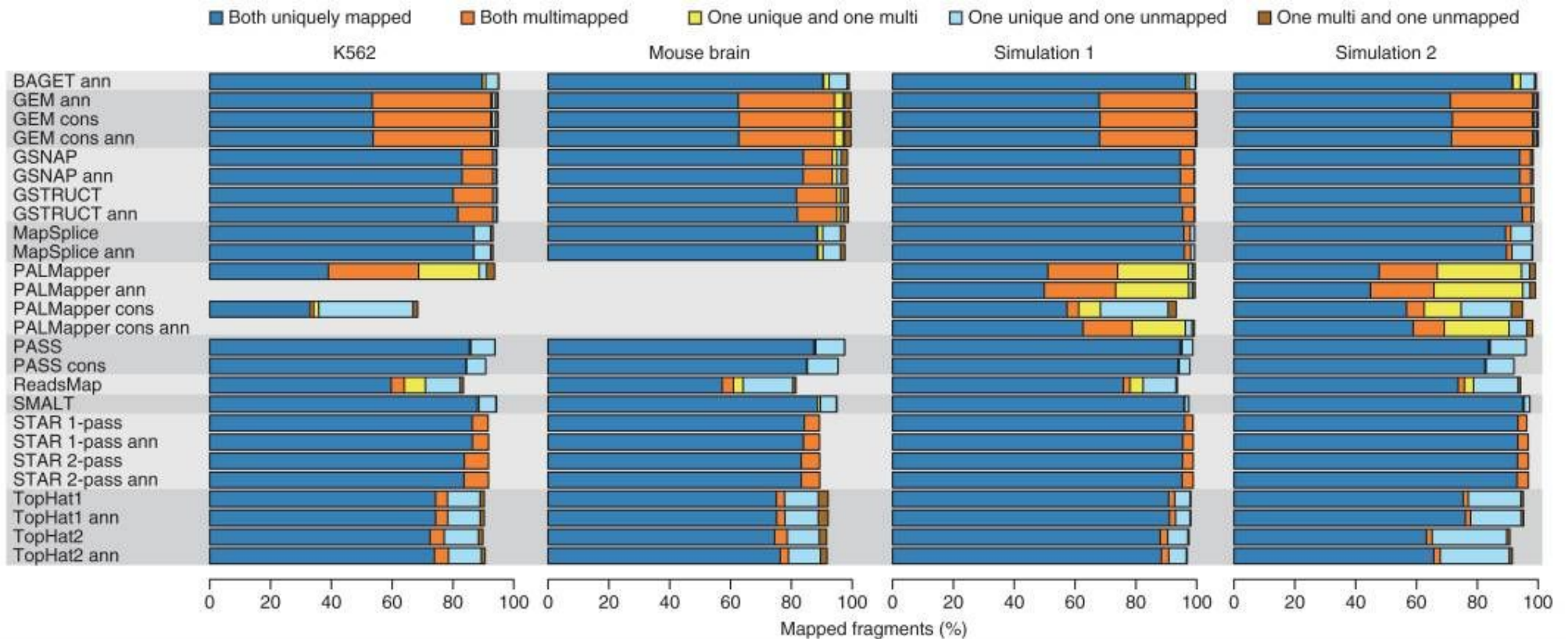
# Comparing tools

How to compare tools ?

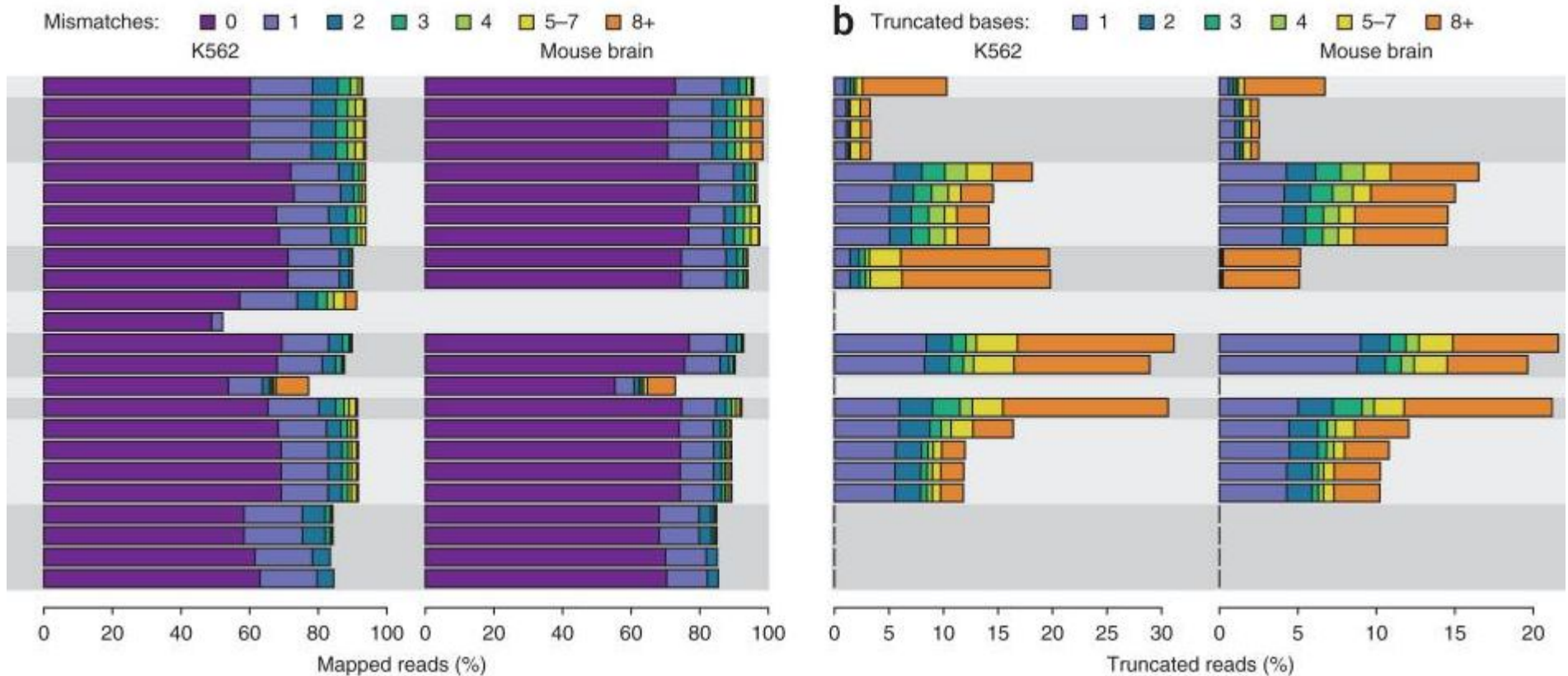
- sensibility (maximize #mapped reads)
- specificity (assign reads to the correct position)  
→ for reads and for junctions
- processing time
- memory requirement

All of these are conflicting criteria ...

## The RNA-seq Genome Annotation Assessment Project

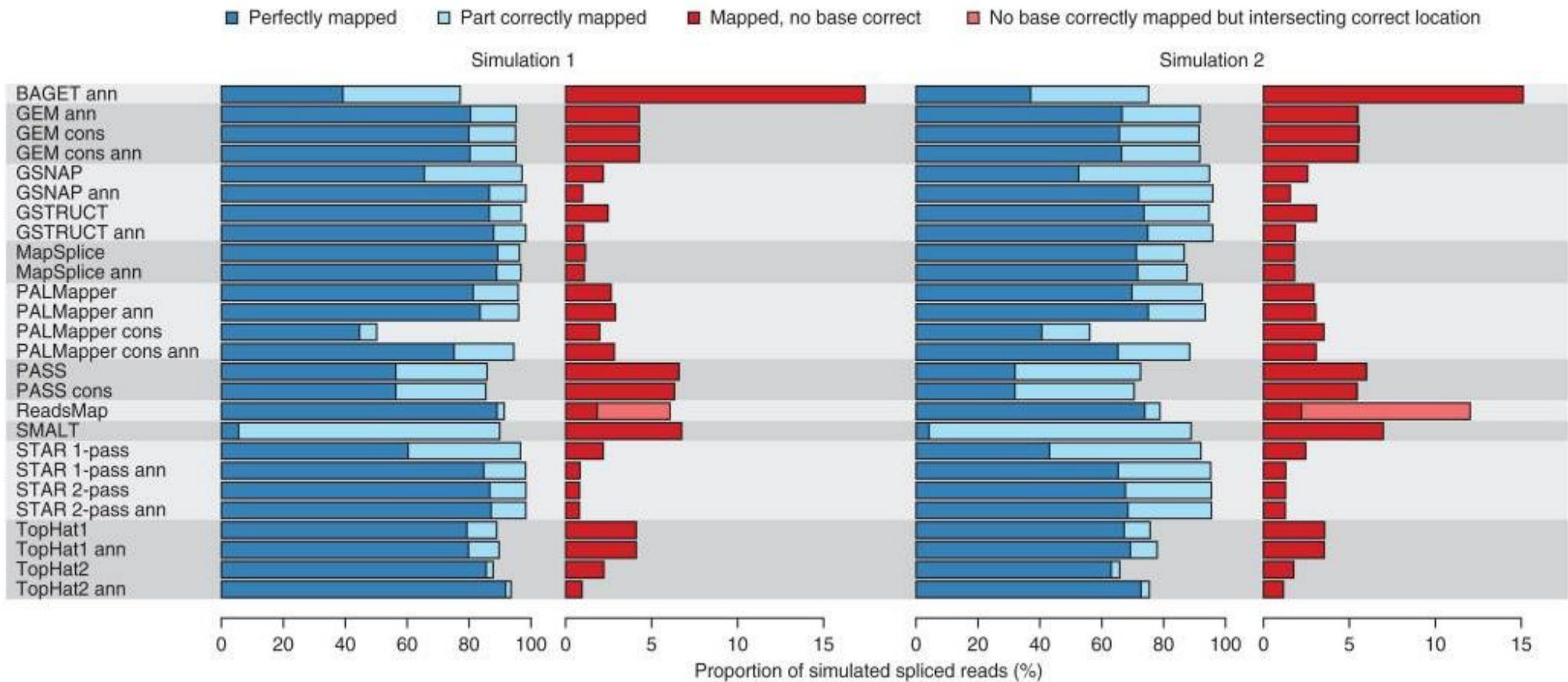


## The RNA-seq Genome Annotation Assessment Project



Engström et al., Nature Methods, 2013

## The RNA-seq Genome Annotation Assessment Project



# 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

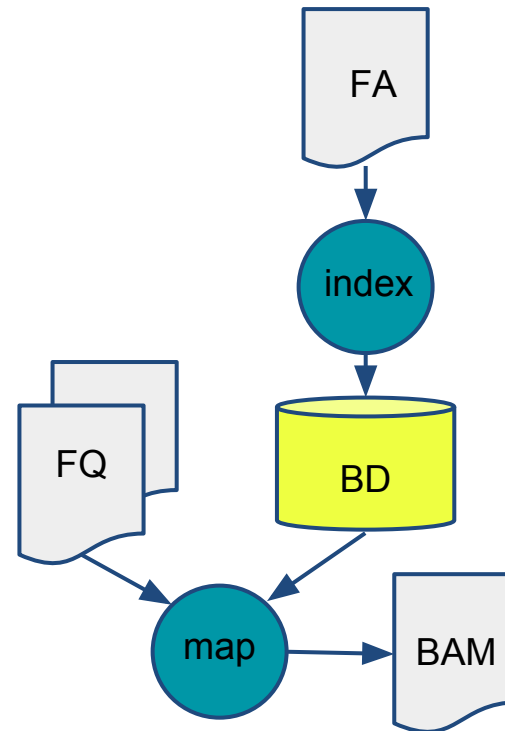
Received 18 April 2016 | Accepted 15 November 2016 | Published online 12 December 2016

| Corrected online **22 December 2016**

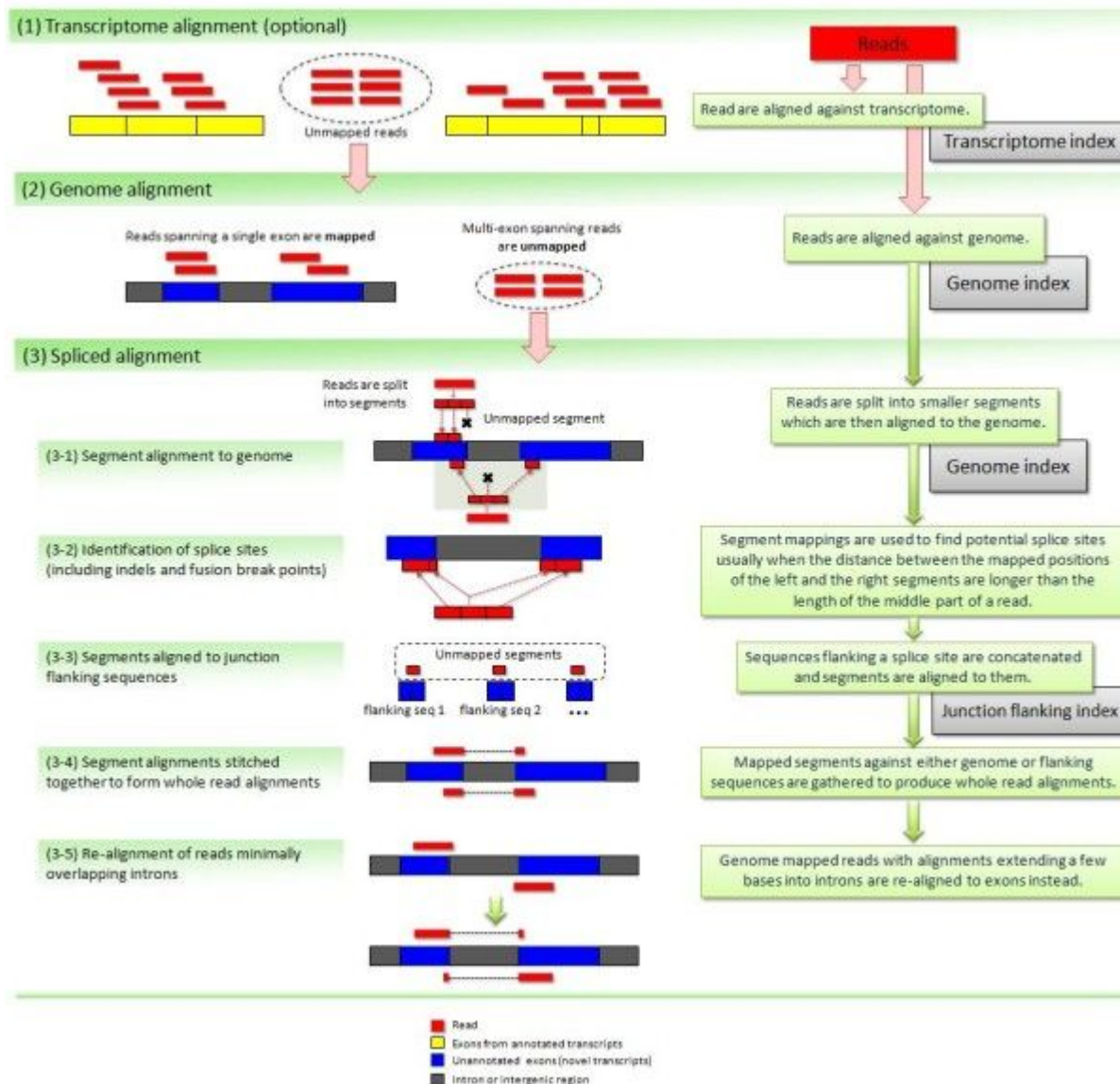


# Mapping steps

- 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>

# An other aligner : STAR



Bioinformatics. 2013 Jan; 29(1): 15–21.

PMCID: PMC3530905

Published online 2012 Oct 25. doi: [10.1093/bioinformatics/bts635](https://doi.org/10.1093/bioinformatics/bts635)

## **STAR: ultrafast universal RNA-seq aligner**

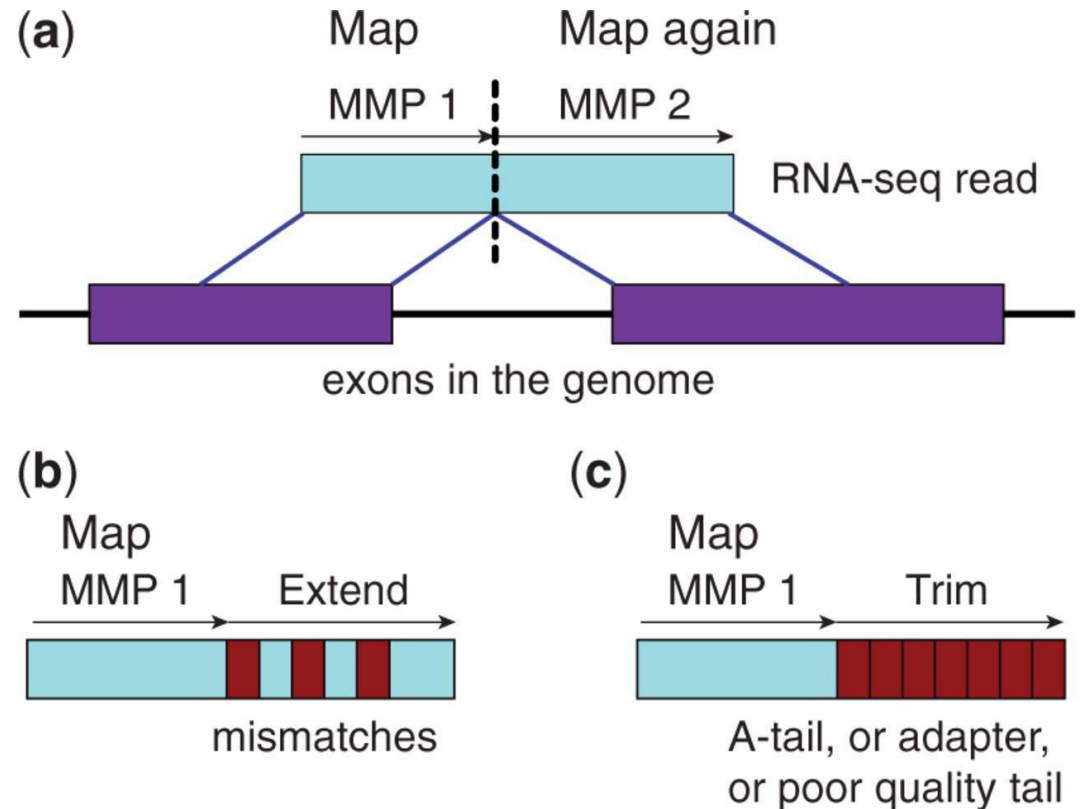
[Alexander Dobin](#),<sup>1,\*</sup> [Carrie A. Davis](#),<sup>1</sup> [Felix Schlesinger](#),<sup>1</sup> [Jorg Drenkow](#),<sup>1</sup> [Chris Zaleski](#),<sup>1</sup> [Sonali Jha](#),<sup>1</sup> [Philippe Batut](#),<sup>1</sup> [Mark Chaisson](#),<sup>2</sup> and [Thomas R. Gingeras](#)<sup>1</sup>

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

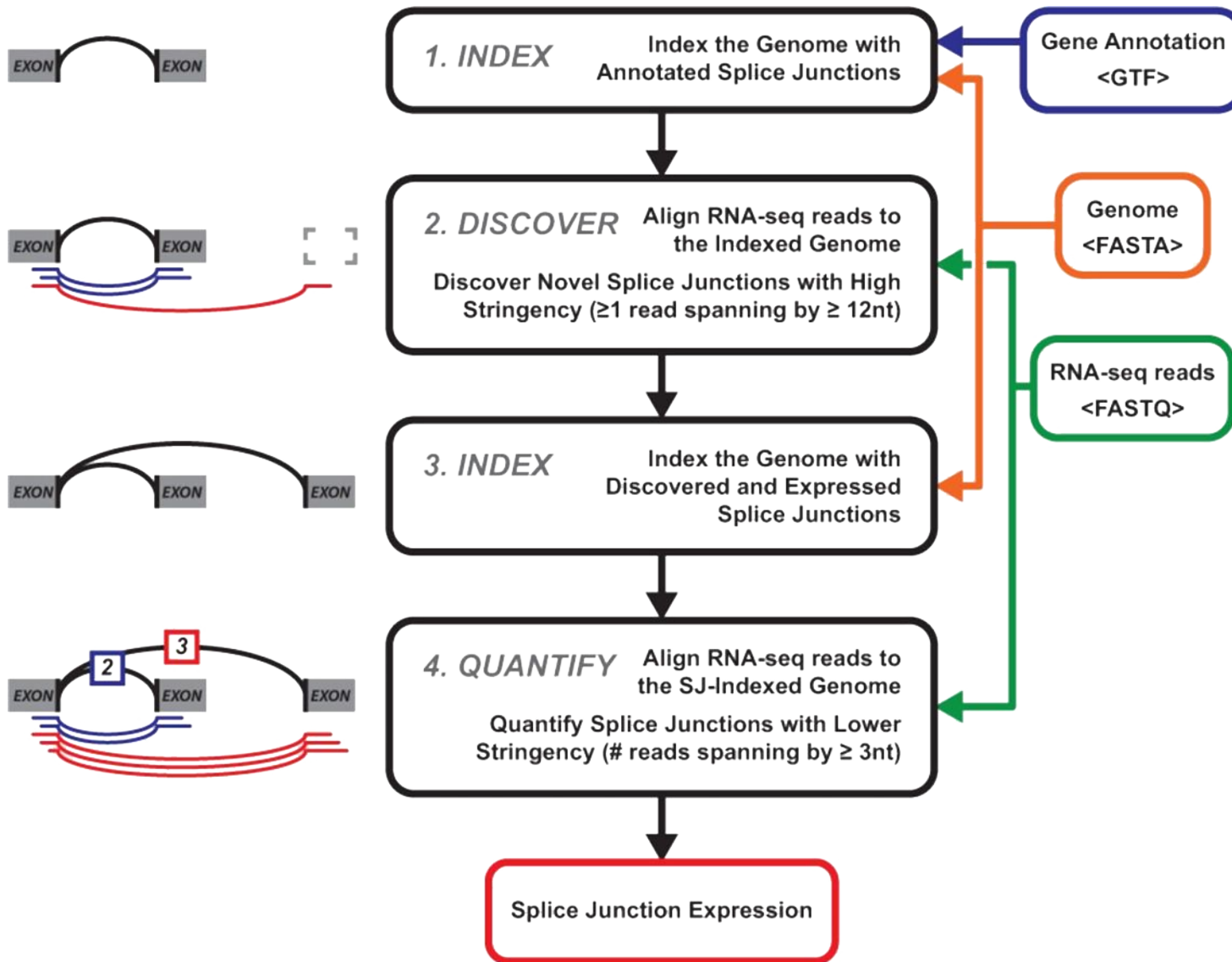
Another strategy:

- 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



# Two passes strategy



« 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	1000	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	1000	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";

ENST01.1 [1000-2000] [3000-4000]

ENST01.2 [1000-4000]

ENSG01, ABC

ENSG02, ENST02.1, 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

## Exercise n°5 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/qcsh 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        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	8M2I4M1D3M =		37	39	TTAGATAAAGGATACTG	*	
r002	0	ref	9	30	3S6M1P1I4M *		0	0	AAAAGATAAGGATA	*	
r003	0	ref	9	30	5S6M *		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,+,5S6M,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	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*	
r002	0	ref	9	30	3S6M1P1I4M	*	0	0	AAAAGATAAAGGATA	*	
r003	0	ref	9	30	5S6M	*	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,+,5S6M,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

```

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	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*	
r002	0	ref	9	30	3S6M1P1I4M	*	0	0	AAAAGATAAAGGATA	*	
r003	0	ref	9	30	5S6M	*	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,+,5S6M,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	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*	
r002	0	ref	9	30	3S6M1P1I4M	*	0	0	AAAAGATAAGGATA	*	
r003	0	ref	9	30	5S6M	*	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,+,5S6M,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

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

Exercise n°5 B/

Map the 2 FASTQ files.

Do not forget to provide a different output file name for each set.

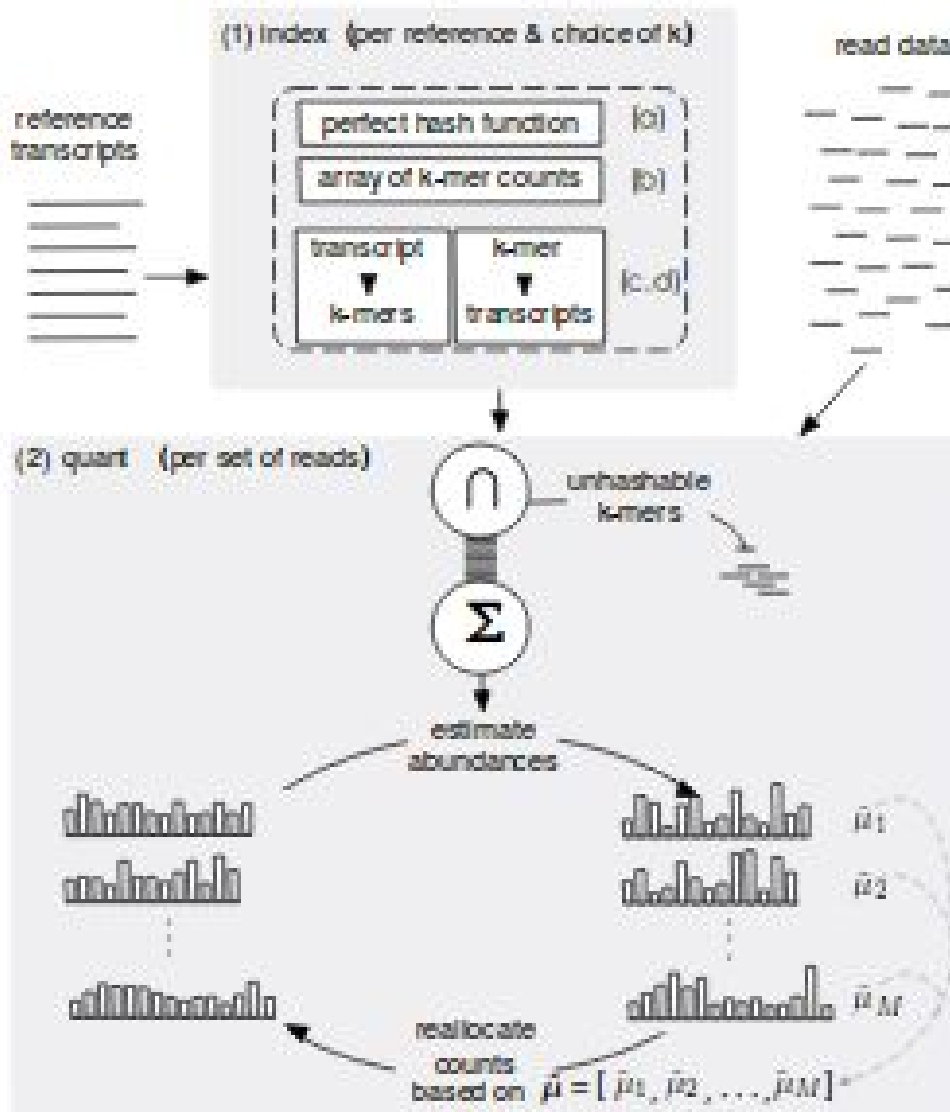
Index the output BAM files with:

```
samtools index file.bam
```

Get some stats with:

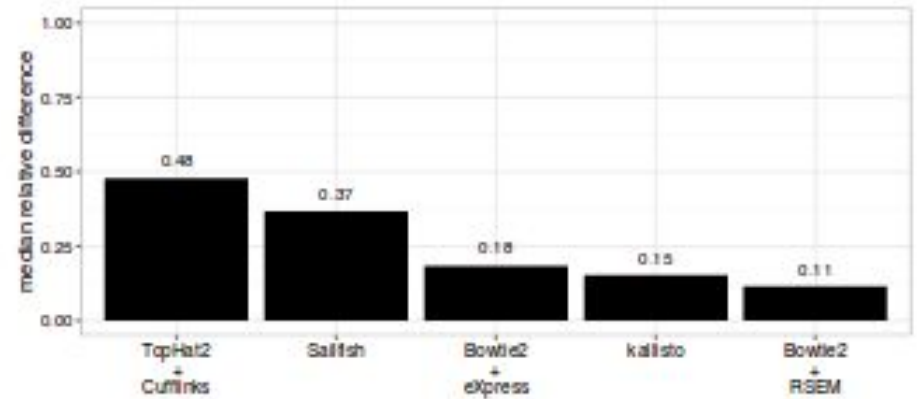
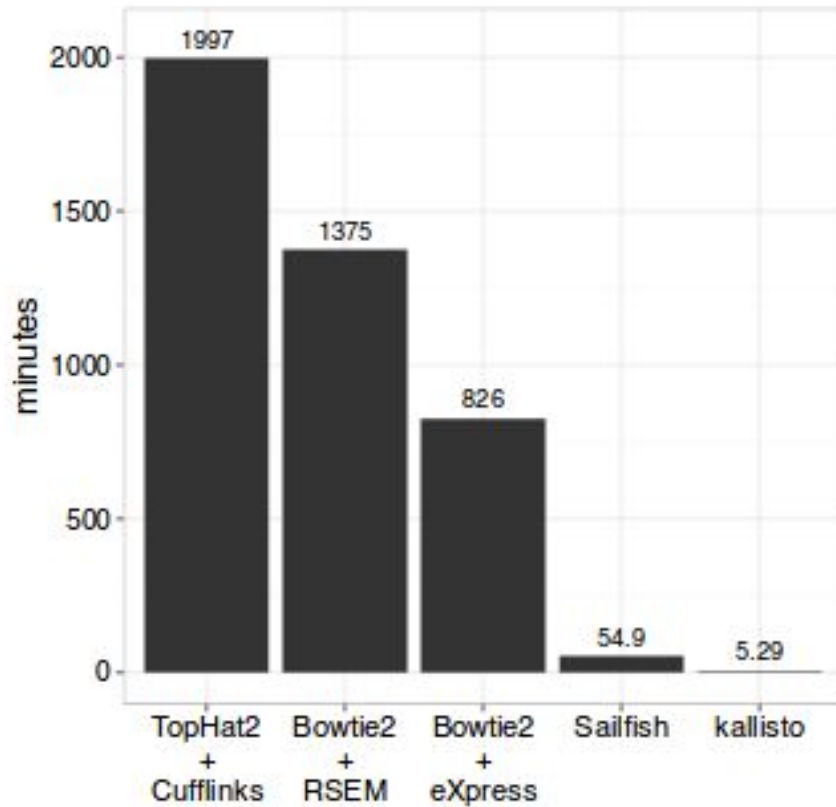
```
samtools flagstat file.bam
```

# Quasi-mapping: Sailfish



- 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

- Exercise n°5 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).

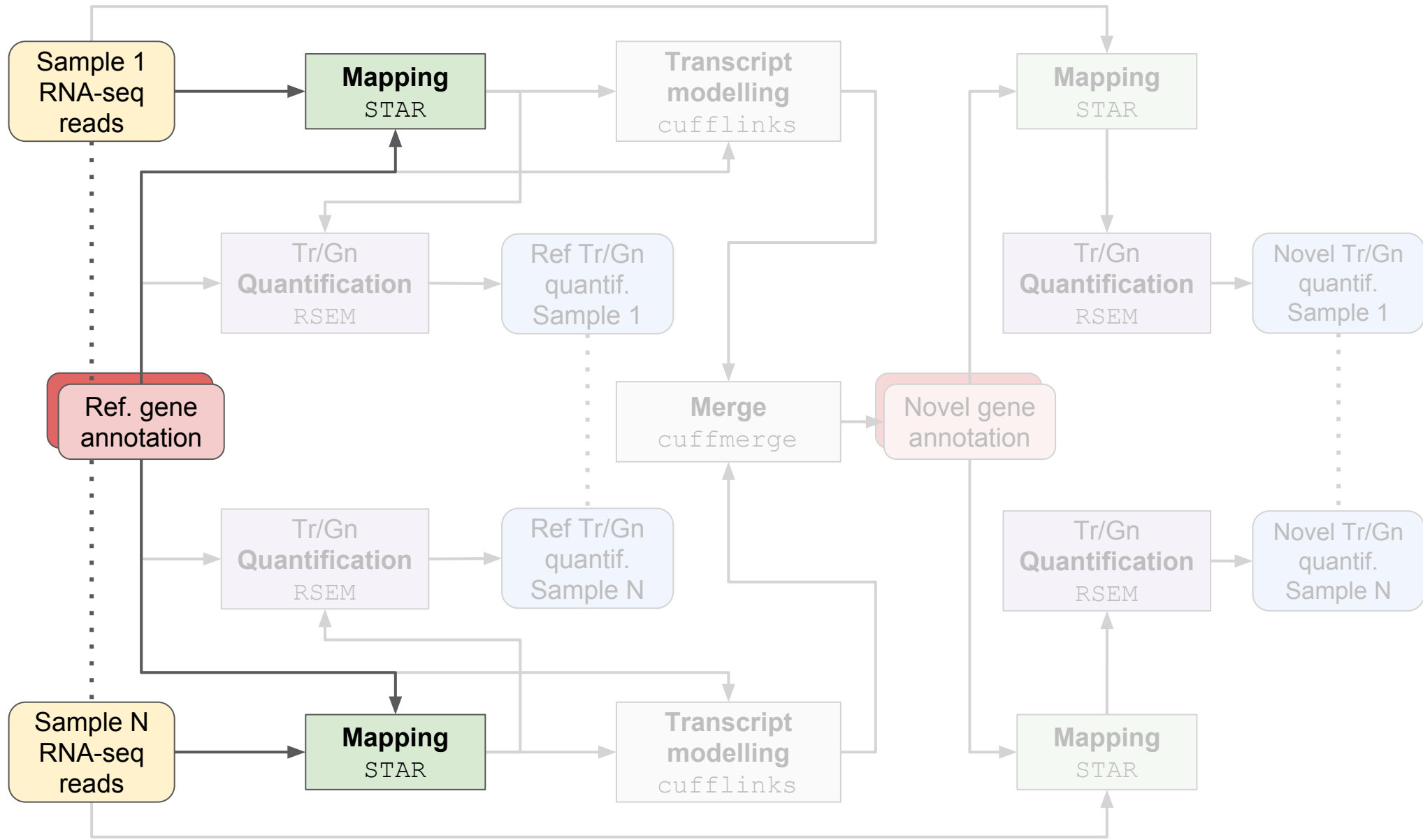
Some interesting loci:

- 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.

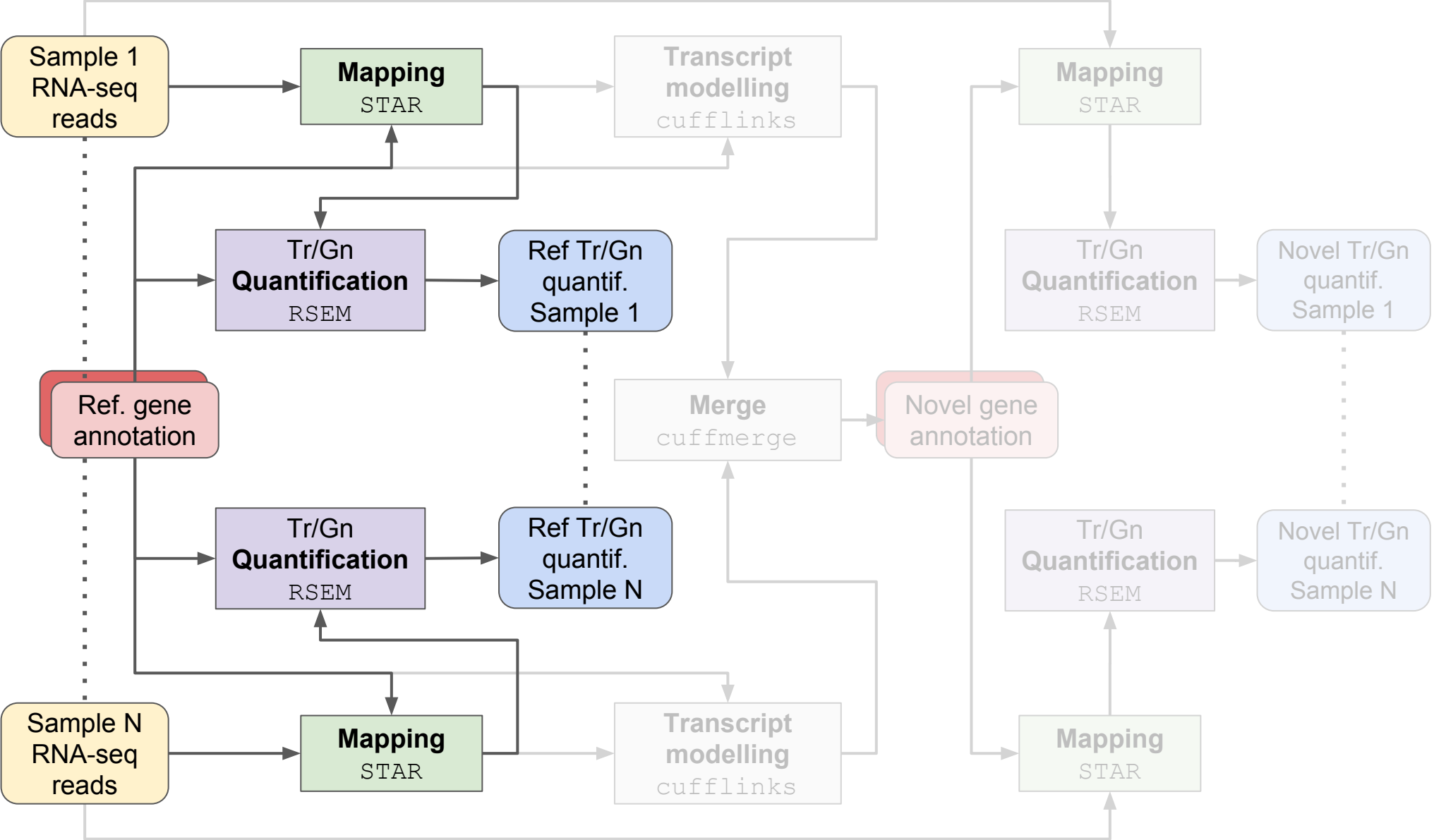
# Explore IGV

- 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

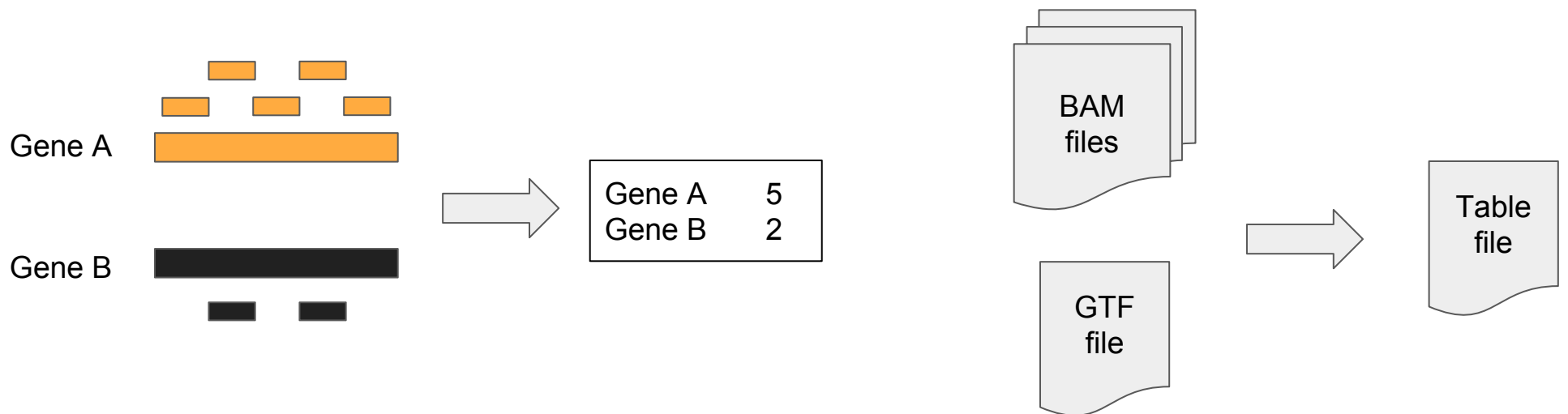


# Analysis workflow



# Quantification

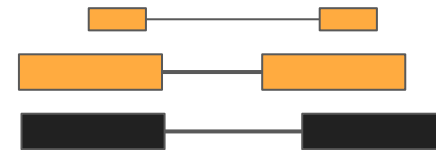
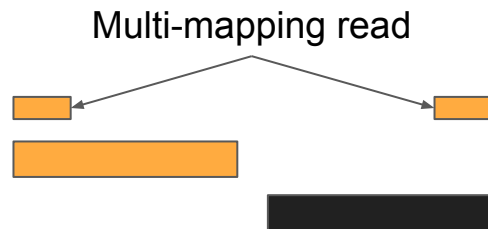
Quantification: estimation of expression based on a read count.



Estimation of:

- gene expression
- transcript expression
- exon expression

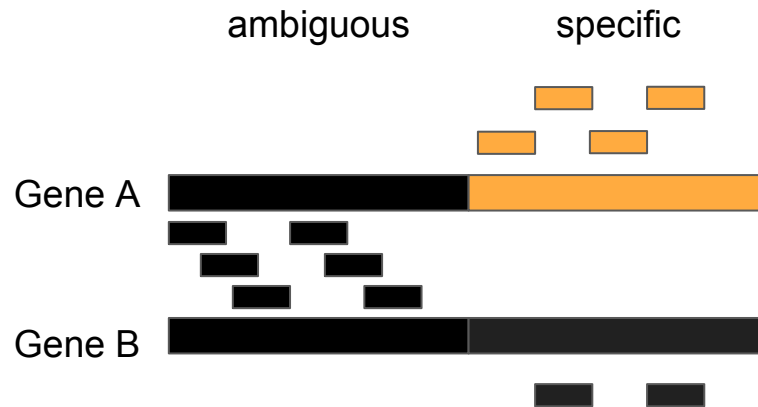
# Difficult cases



Every quantification tools uses its own rules!

# Raw counts vs estimation

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



Pros estimation:

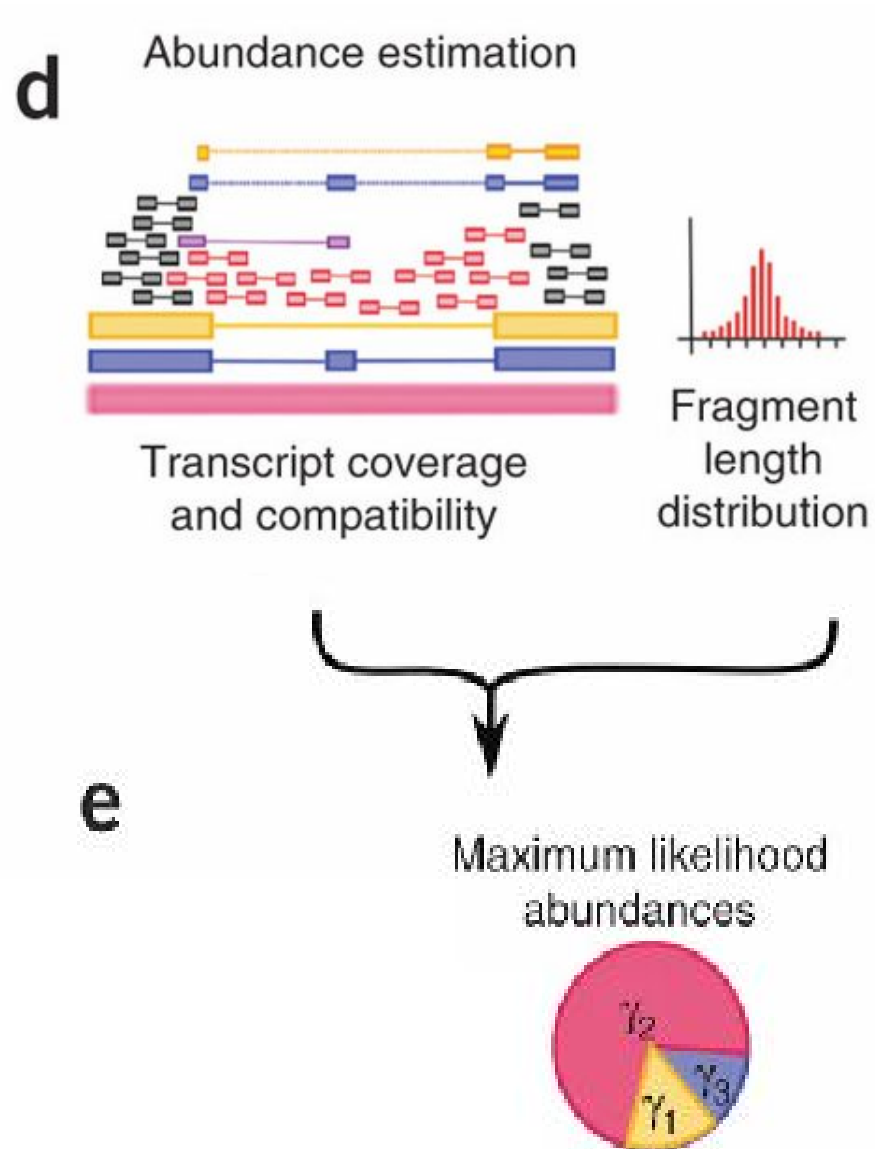
- Use more reads.
- More accurate?

Cons estimation:

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



# Transcript expression



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

# Raw counts tool: featureCounts

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

Yang Liao<sup>1,2</sup>, Gordon K. Smyth<sup>1,3</sup> and Wei Shi<sup>1,2,\*</sup>

<sup>1</sup>Bioinformatics Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052,

<sup>2</sup>Department of Computing and Information Systems and <sup>3</sup>Department 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
  - Assignment 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

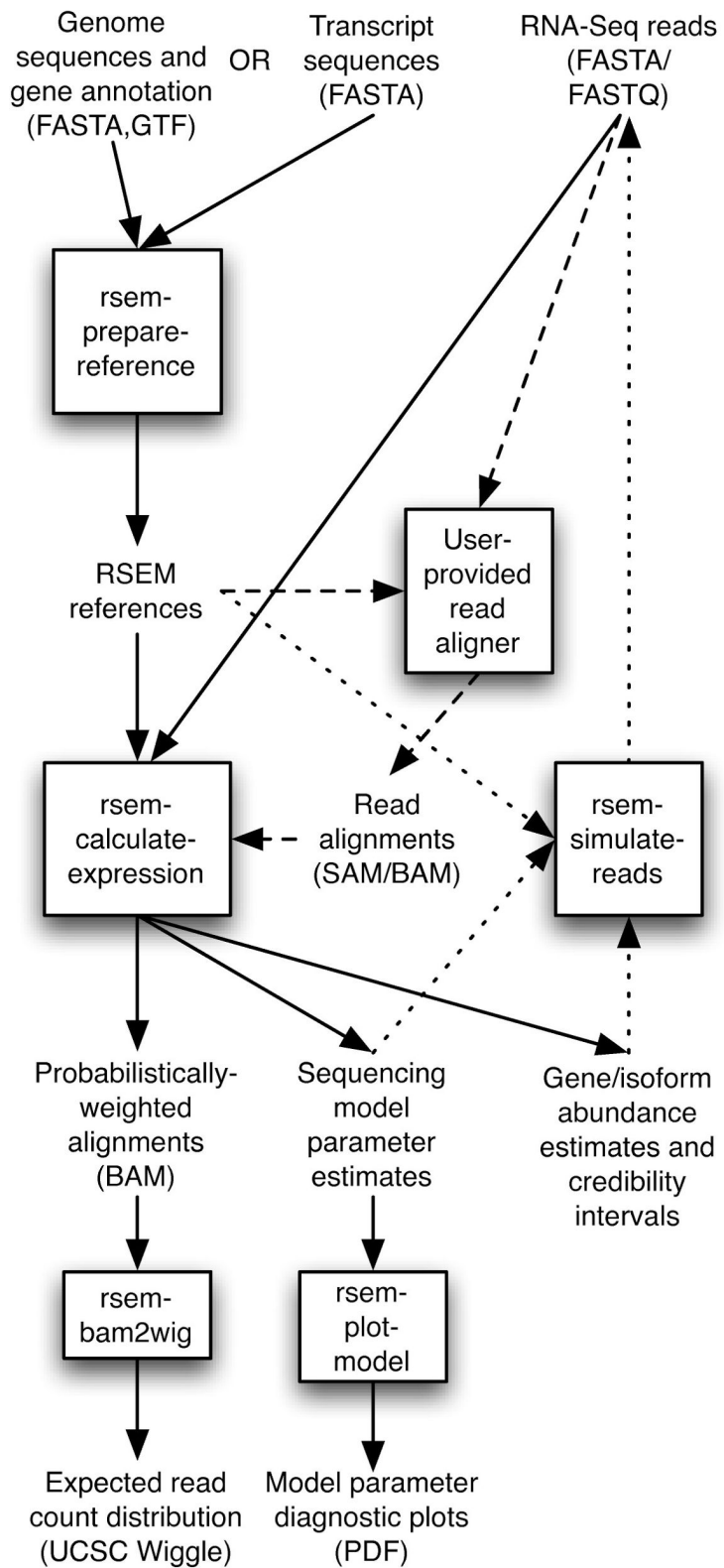
Bo Li and Colin N Dewey 

*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

# RSEM workflow



Two main steps:

- rsem-prepare-reference
- rsem-calculate-expression

In default use case, RSEM maps the reads (with Bowtie).

Using a different mapping tool (STAR) requires:

- Extra parameters for STAR
- Extra parameters for RSEM

# Hands-in: prepare reference

## Exercise n°6

Command line:

```
/usr/local/bioinfo/src/RSEM/RSEM-1.3.0/rsem-p  
repare-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  $i$ .

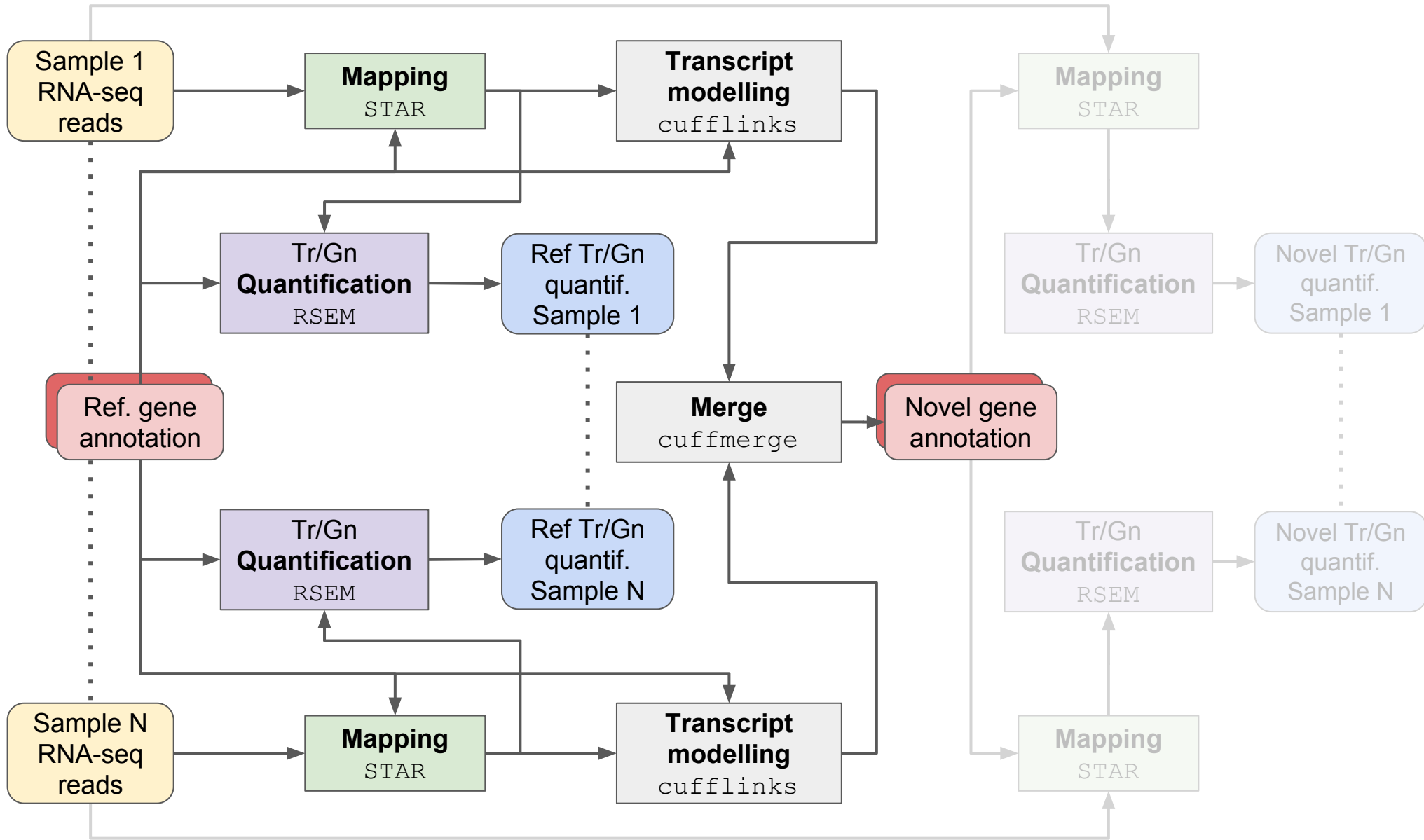
# RPKM vs FPKM vs TPM

- 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



# New transcriptome: why?

Ensembl Release 88 (March 2017)

## Homo sapiens

Coding genes	20,310 (incl 556 readthrough)
Non coding genes	22,529
Pseudogenes	14,589 (incl 6 readthrough)
Gene transcripts	<u>199,234</u>

## Mus musculus

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

## Rattus norvegicus

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

## Bos taurus

Coding genes	19,994
Non coding genes	3,825
Pseudogenes	797
Gene transcripts	<u>26,740</u>

## Oryctolagus cuniculus

Coding genes	19,293
Non coding genes	3,375
Pseudogenes	1,001
Gene transcripts	<u>24,964</u>

## Sus scrofa

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

## Gallus gallus

Coding genes	18,346
Non coding genes	6,492
Pseudogenes	43
Gene transcripts	<u>38,118</u>

# Transcript reconstruction



Gene location



Exon location



Junctions :

- between read pair junction

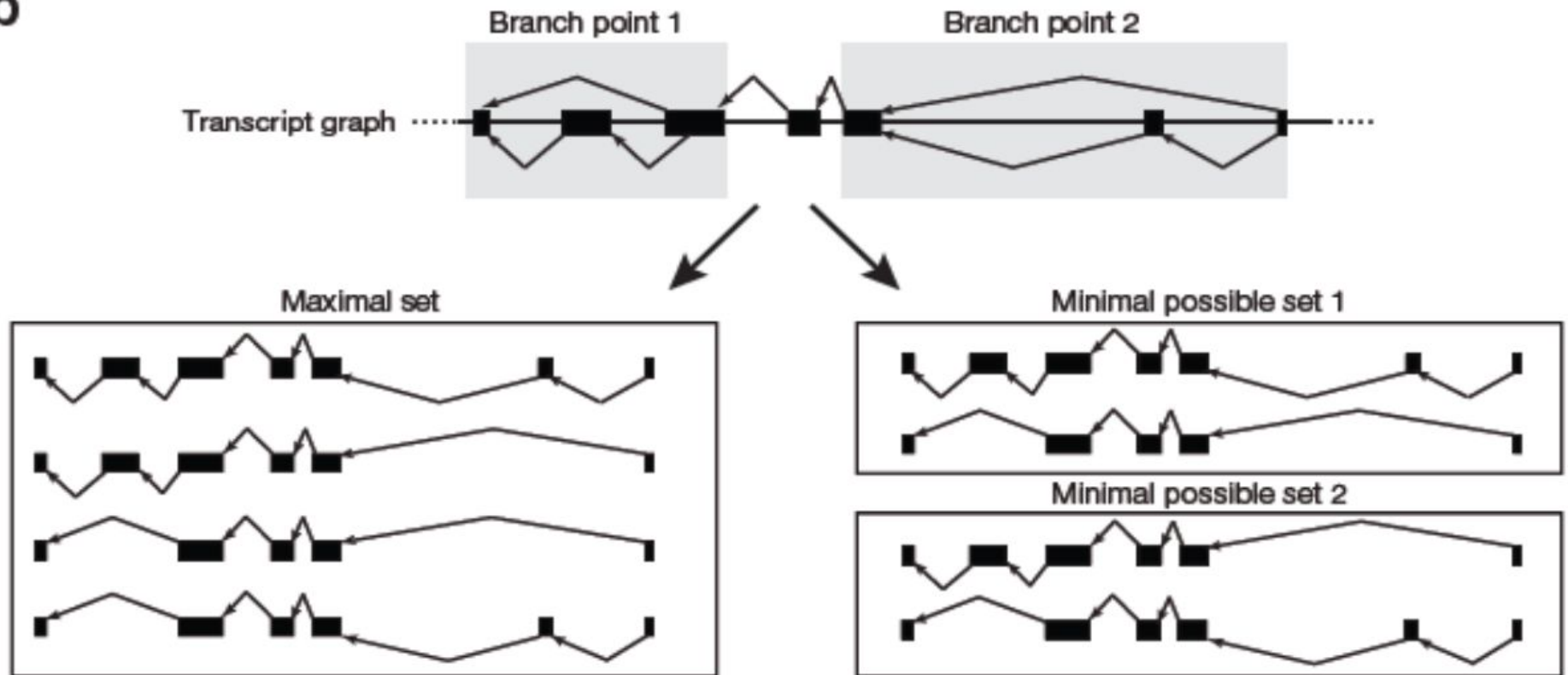


- within read junction



# Model building strategies

b



REVIEW

## Computational methods for transcriptome annotation and quantification using RNA-seq

Manuel Garber<sup>1</sup>, Manfred G Grabherr<sup>1</sup>, Mitchell Guttman<sup>1,2</sup> & Cole Trapnell<sup>1,3</sup>



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

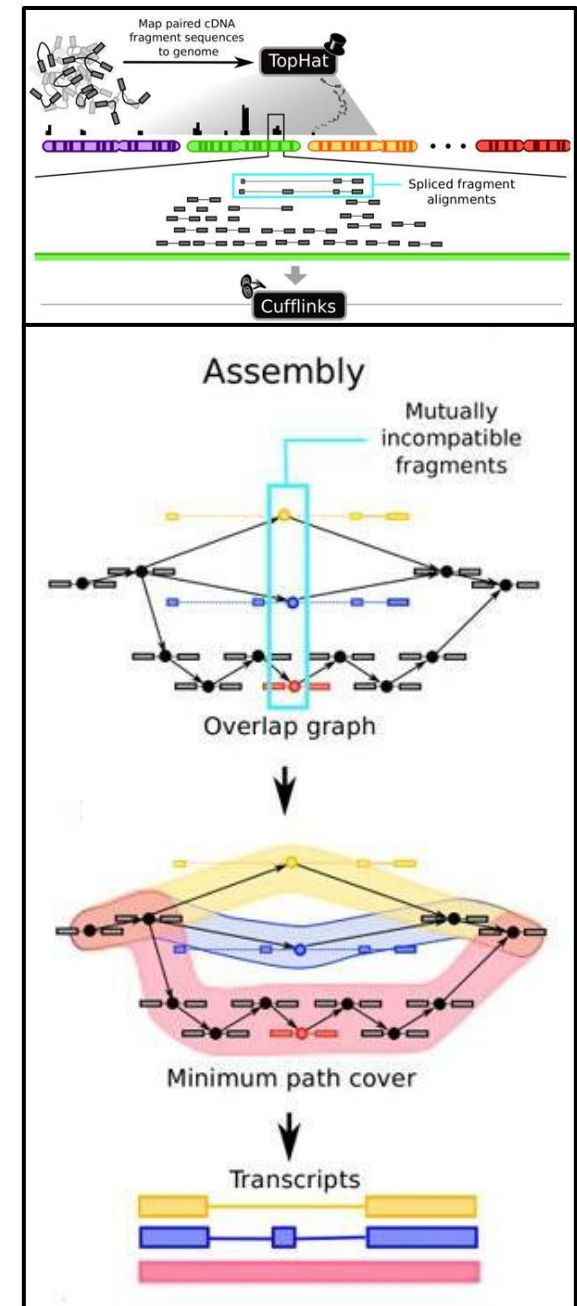
# Cufflinks

<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

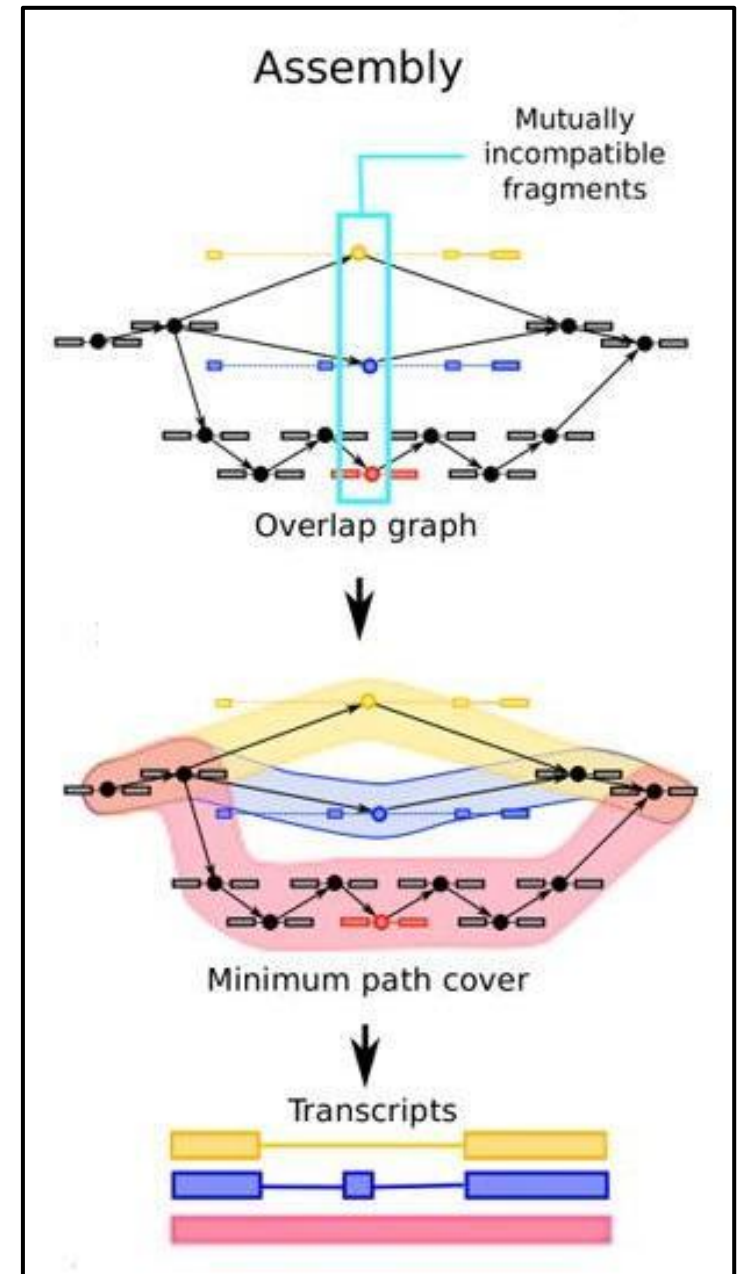
- Transcripts assembly:
  - fragments are divided into non-overlapping loci
  - each locus is assembled independently
- Cufflinks assembler
  - find the mini nb of transcripts that explain the reads
  - find a minimum path cover (Dilworth's theorem):
    - nb incompatible read = mini nb of transcripts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



# Cufflinks transcript assembly

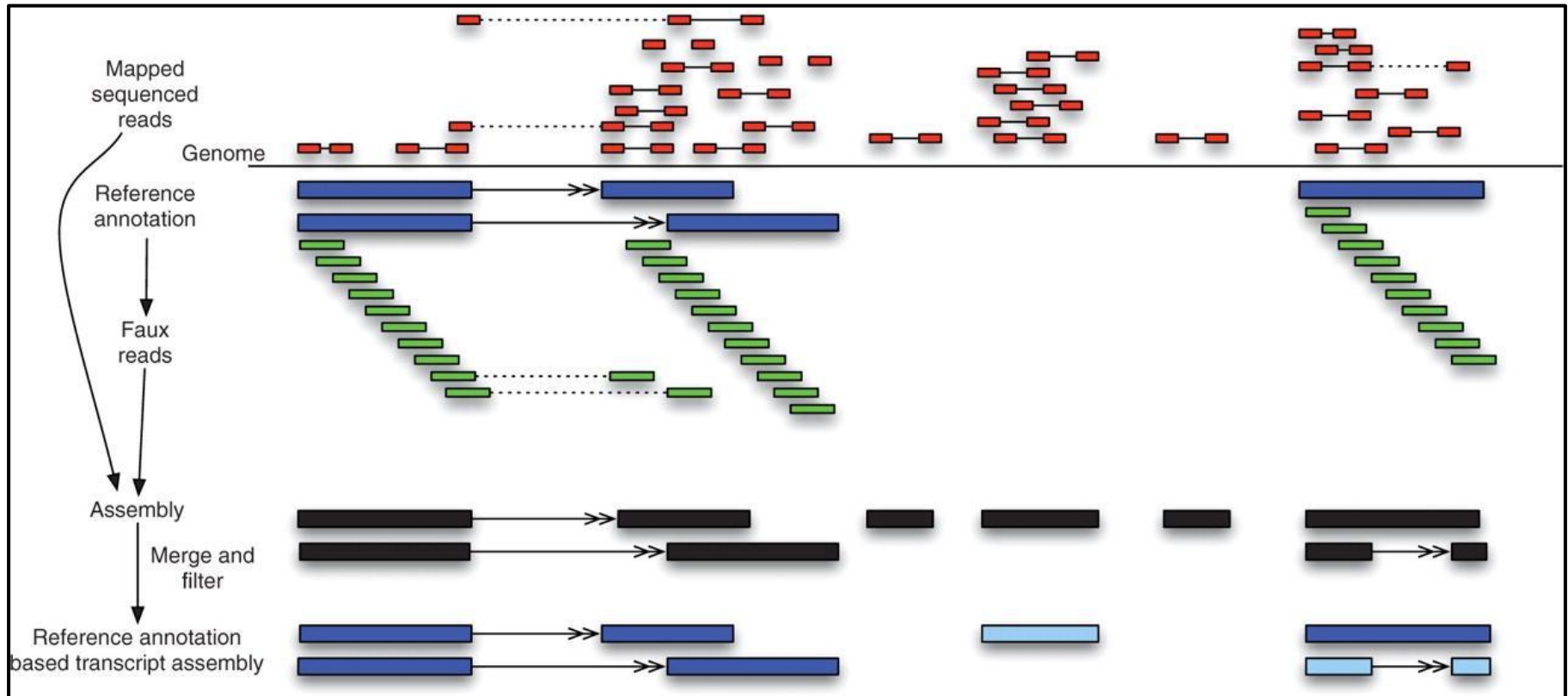
Transcripts assembly:

- 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



# Cufflinks transcript assembly

## Reference Annotation Based Transcripts Assembly



Assembling novel transcripts in the context of an existing annotation

# Cufflinks inputs and options

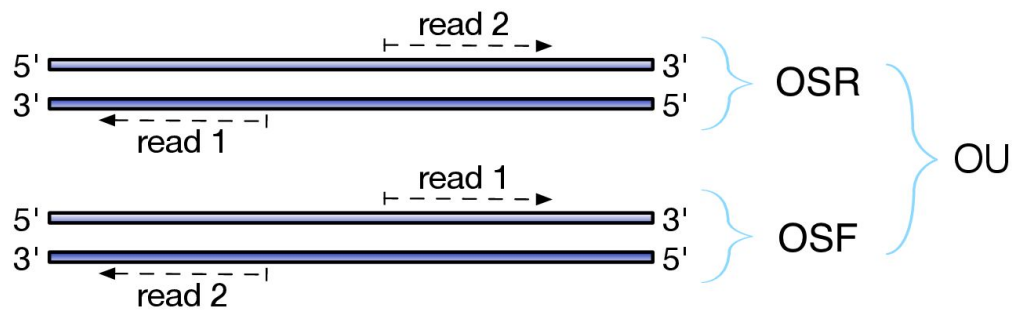
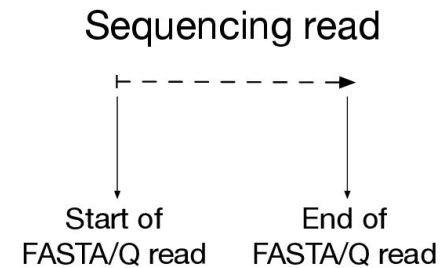
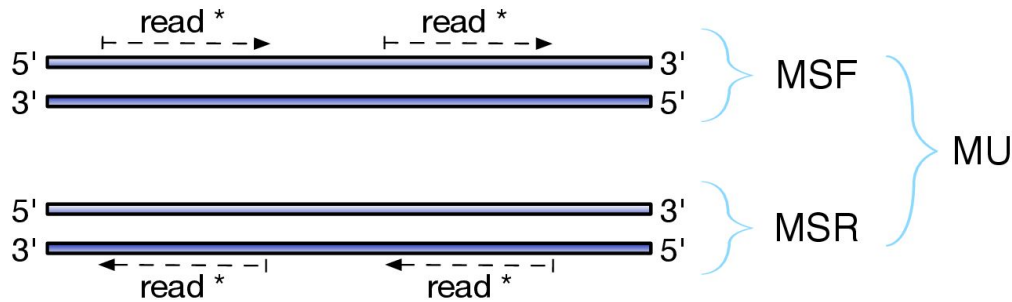
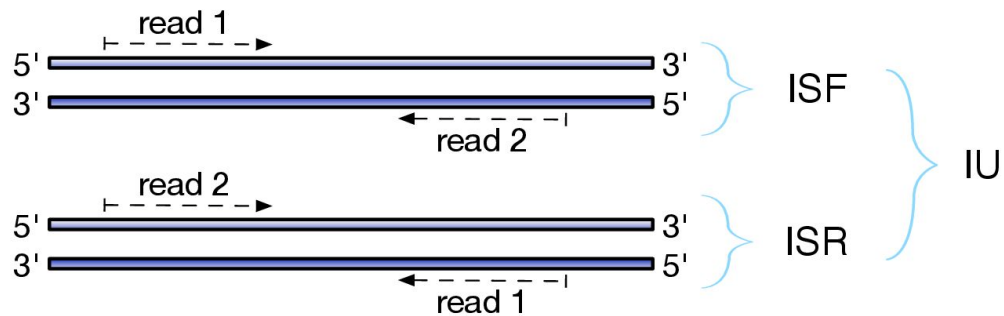
- Command line:

```
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



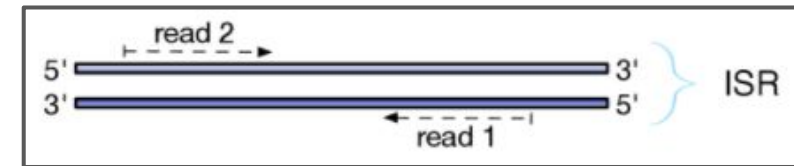
I = inward  
O = outward  
M = matching

S = stranded  
U = unstranded

F = read 1 (or single-end read) comes from the forward strand  
R = read 1 (or single-end read) comes from the reverse strand

# 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 ISR	-1 SR
-fr-secondstrand	-1 ISF	-1 SF

[http://salmon.readthedocs.io/en/latest/library\\_type.html](http://salmon.readthedocs.io/en/latest/library_type.html)

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

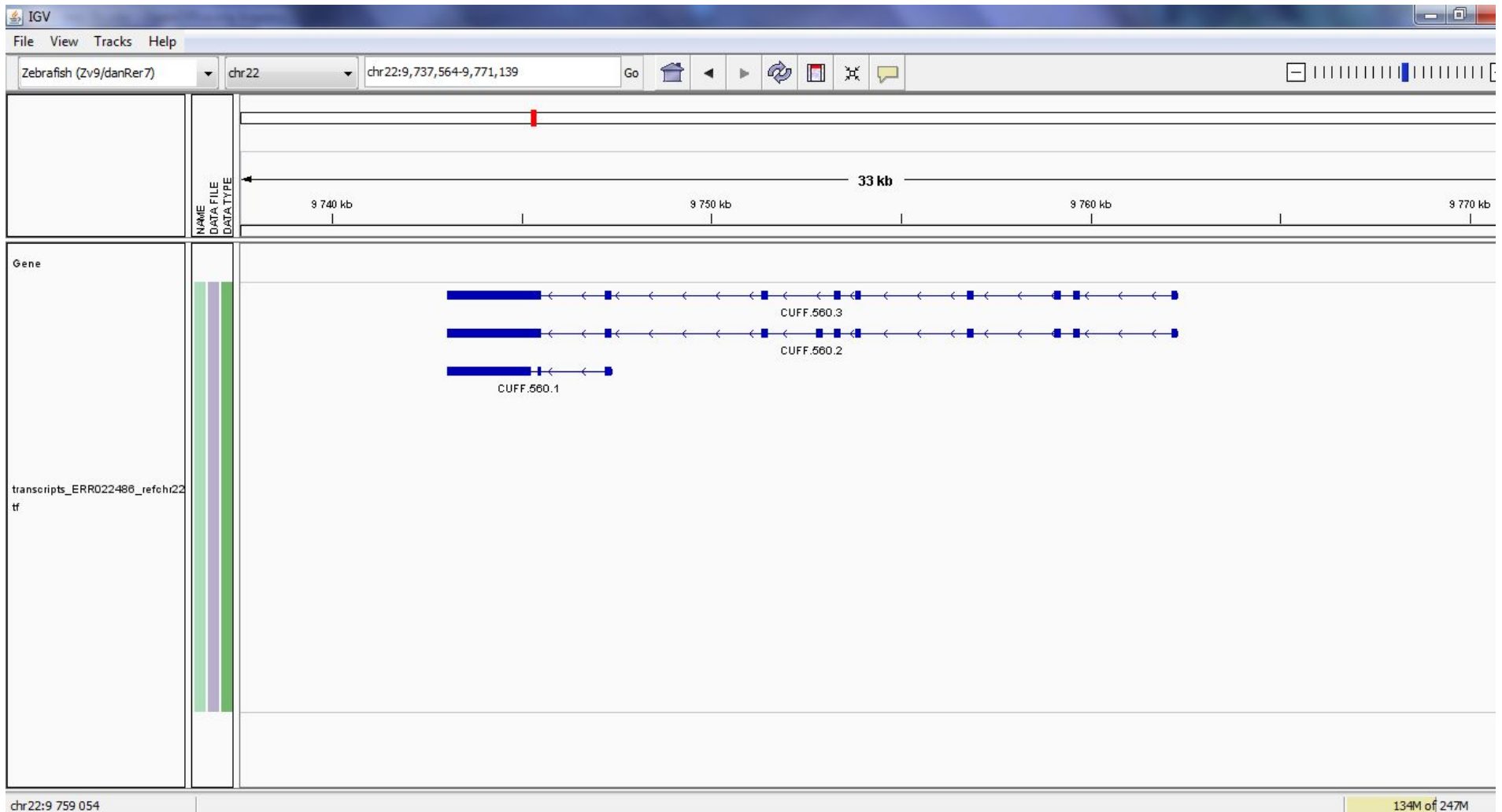
```
1 Cufflinks transcript 459812 460830 1 - .
1 Cufflinks exon 459812 460830 1 - .
1 Cufflinks transcript 463572 478996 1000 - .
1 Cufflinks exon 463572 463746 1000 - .
1 Cufflinks exon 466228 466405 1000 - .
```

```
gene_id "ENSBTAG00000013841"; transcript_id "ENSBTAT00000018387"; FPKM "0.0000000000"; frac "0.000000";
gene_id "ENSBTAG00000013841"; transcript_id "ENSBTAT00000018387"; exon_number "1"; FPKM "0.0000000000"; frac "0.000000";
gene_id "CUFF.2"; transcript_id "ENSBTAT00000015319"; FPKM "25.4745974237"; frac "1.000000";
gene_id "CUFF.2"; transcript_id "ENSBTAT00000015319"; exon_number "1"; FPKM "25.4745974237"; frac "1.000000";
gene_id "CUFF.2"; transcript_id "ENSBTAT00000015319"; 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

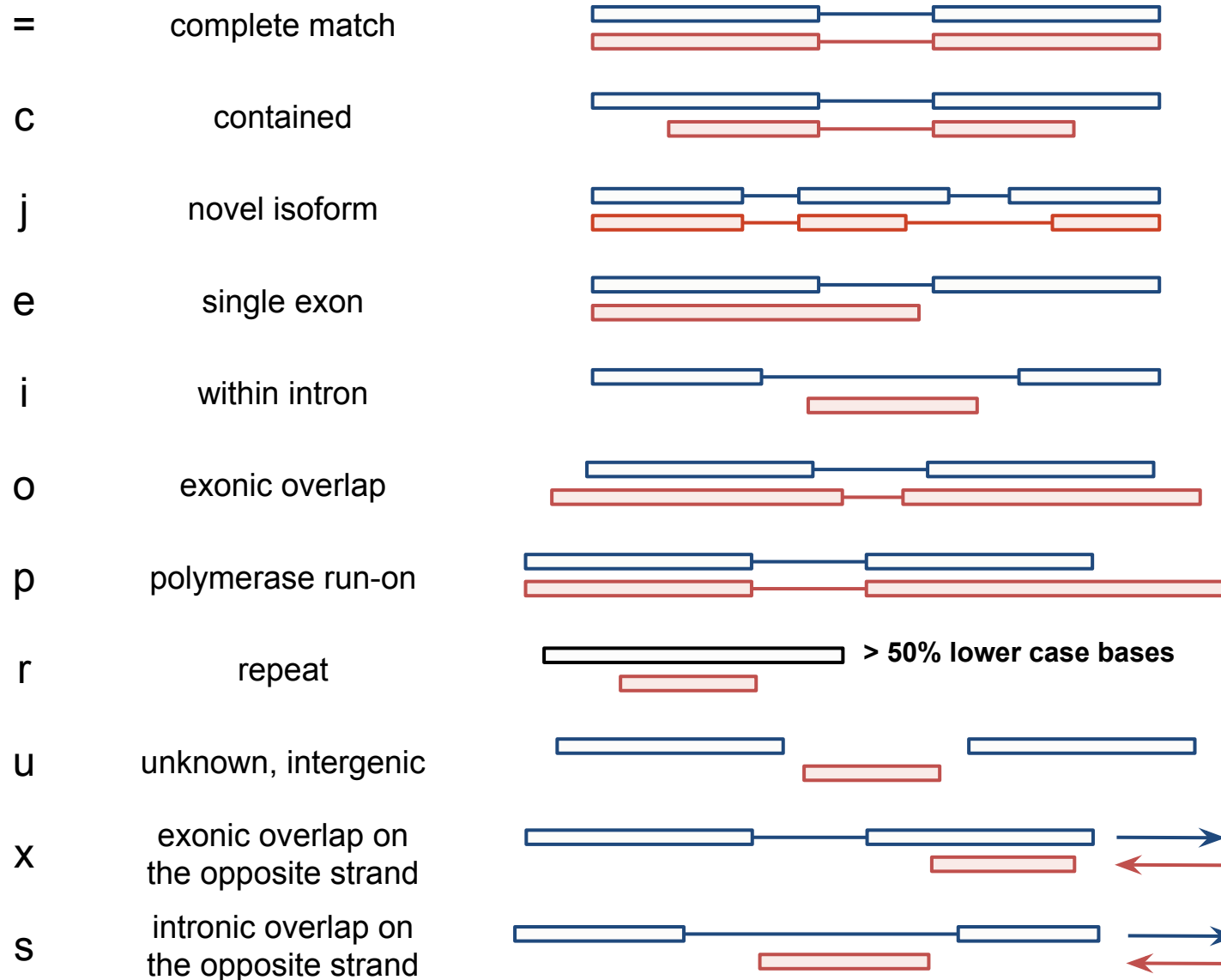
Command:

```
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

## Class code de cuffcompare



# 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 242646 . + .
1 Cufflinks exon 275623 275681 . + .
1 Cufflinks exon 242402 242646 . + .
1 Cufflinks exon 254559 254693 . + .
1 Cufflinks exon 247340 249673 . + .
1 Cufflinks exon 351546 351874 . + .
1 Cufflinks exon 355064 355237 . + .
1 Cufflinks exon 357793 357952 . + .
1 Cufflinks exon 361144 362915 . + .
```

```
gene_id "XLOC_000001"; transcript_id "TCONS_00000001"; exon_number "1"; gene_name "ENSBTAG00000006858";
gene_id "XLOC_000002"; transcript_id "TCONS_00000002"; exon_number "1"; gene_name "CBX3";
gene_id "XLOC_000002"; transcript_id "TCONS_00000002"; exon_number "2"; gene_name "CBX3";
gene_id "XLOC_000002"; transcript_id "TCONS_00000003"; exon_number "1";
gene_id "XLOC_000002"; transcript_id "TCONS_00000003"; exon_number "2";
gene_id "XLOC_000003"; transcript_id "TCONS_00000004"; exon_number "1";
gene_id "XLOC_000004"; transcript_id "TCONS_00000005"; exon_number "1"; gene_name "RCAN1";
gene_id "XLOC_000004"; transcript_id "TCONS_00000005"; exon_number "2"; gene_name "RCAN1";
gene_id "XLOC_000004"; transcript_id "TCONS_00000005"; exon_number "3"; gene_name "RCAN1";
gene_id "XLOC_000004"; transcript_id "TCONS_00000005"; exon_number "4"; gene_name "RCAN1";
```

```
oId "ENSBTAT00000009004"; nearest_ref "ENSBTAT00000009004"; class_code "="; tss_id "TSS1";
oId "CUFF.1.1"; nearest_ref "ENSBTAT00000007283"; class_code "x"; tss_id "TSS2";
oId "CUFF.1.1"; nearest_ref "ENSBTAT00000007283"; class_code "x"; tss_id "TSS2";
oId "CUFF.1.2"; class_code "u"; tss_id "TSS2";
oId "CUFF.1.2"; class_code "u"; tss_id "TSS2";
oId "CUFF.2.1"; class_code "u"; tss_id "TSS3";
oId "CUFF.3.1"; nearest_ref "ENSBTAT00000037243"; class_code "j"; tss_id "TSS4";
oId "CUFF.3.1"; nearest_ref "ENSBTAT00000037243"; class_code "j"; tss_id "TSS4";
oId "CUFF.3.1"; nearest_ref "ENSBTAT00000037243"; class_code "j"; tss_id "TSS4";
oId "CUFF.3.1"; nearest_ref "ENSBTAT00000037243"; 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

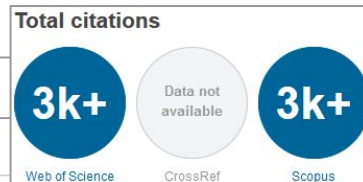
Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn & Lior Pachter

[Affiliations](#) | [Contributions](#) | [Corresponding author](#)

*Nature Protocols* **7**, 562–578 (2012) | doi:10.1038/nprot.2012.016

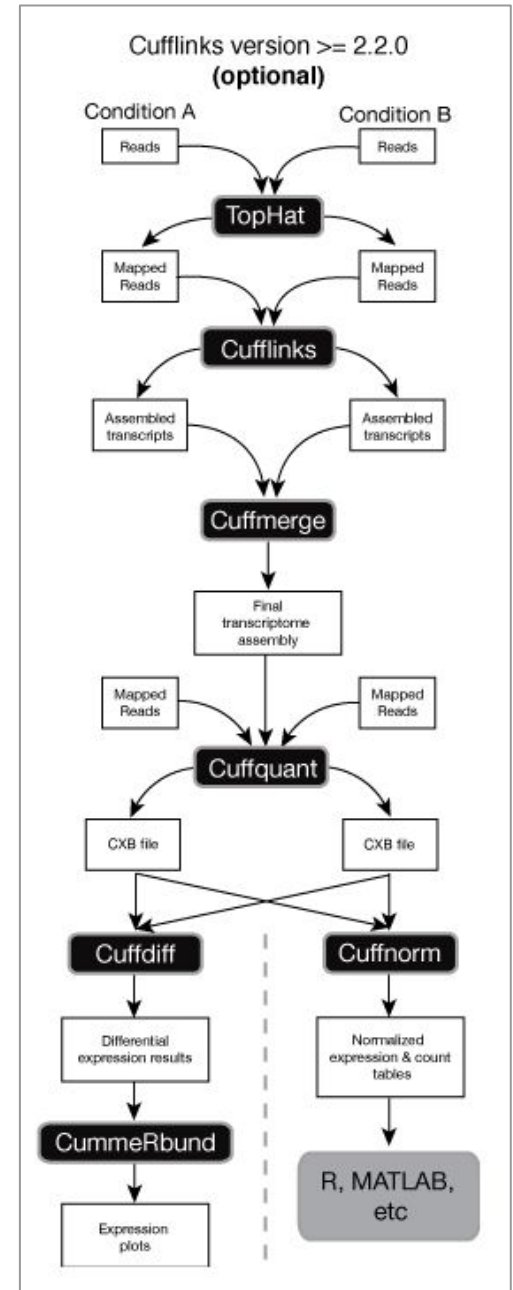
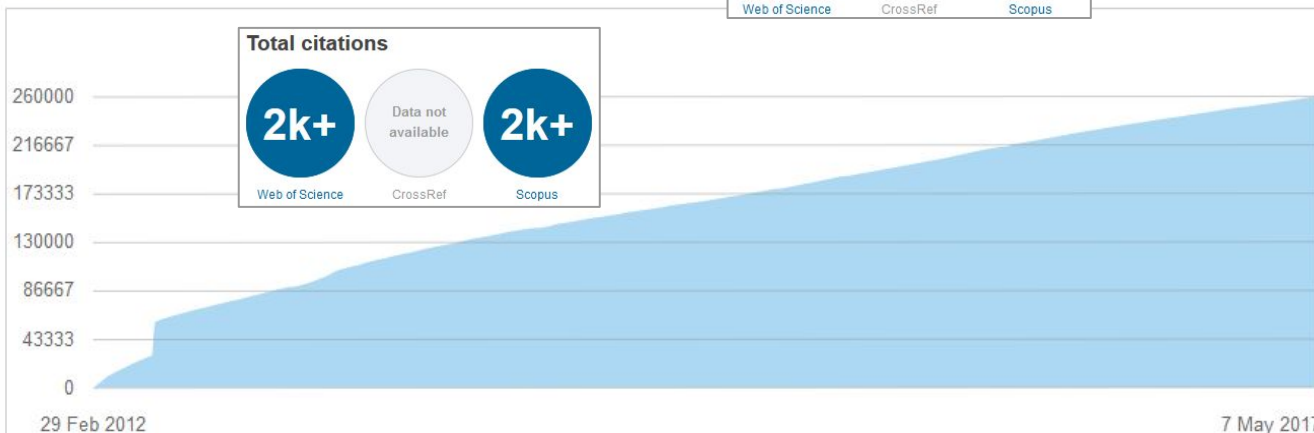
Published online 01 March 2012 | Corrected online **07 August 2014**

[Corrigendum \(October, 2014\)](#)



### Page views

256,089







日本語要約

## 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

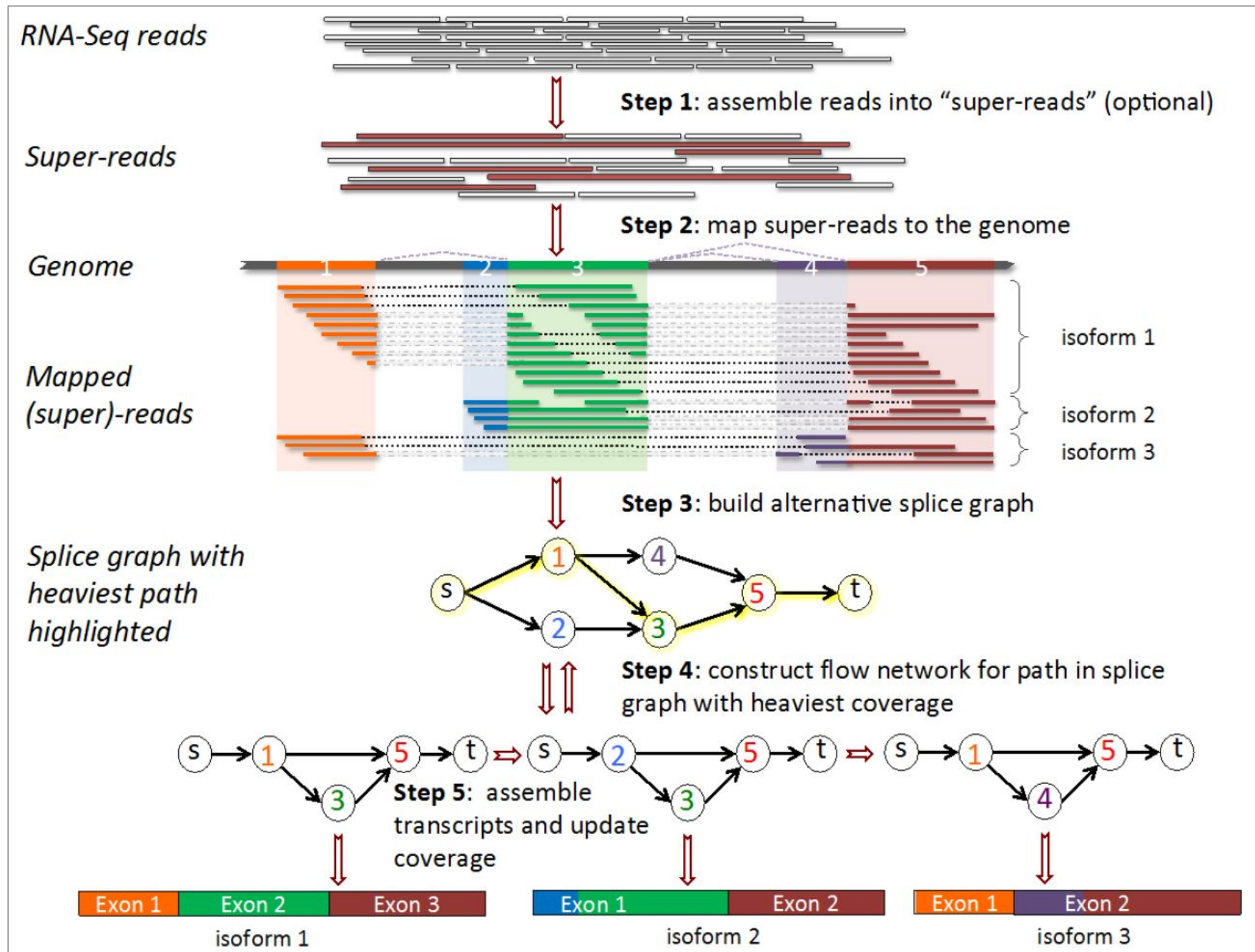
Received 15 April 2014 | Accepted 09 December 2014 | Published online 18 February 2015

# StringTie

<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



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

NATURE PROTOCOLS | PROTOCOL



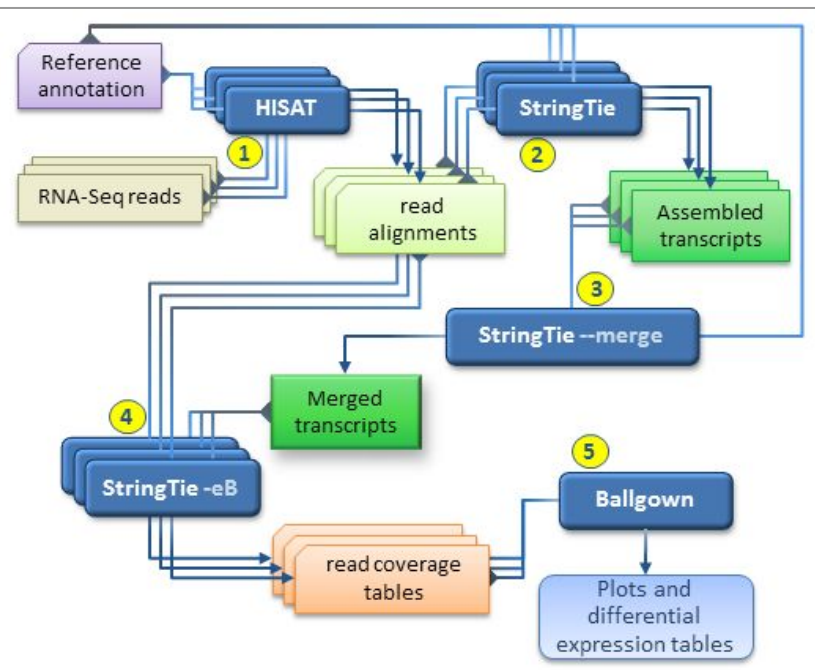
## Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek & Steven L Salzberg

[Affiliations](#) | [Contributions](#) | [Corresponding author](#)

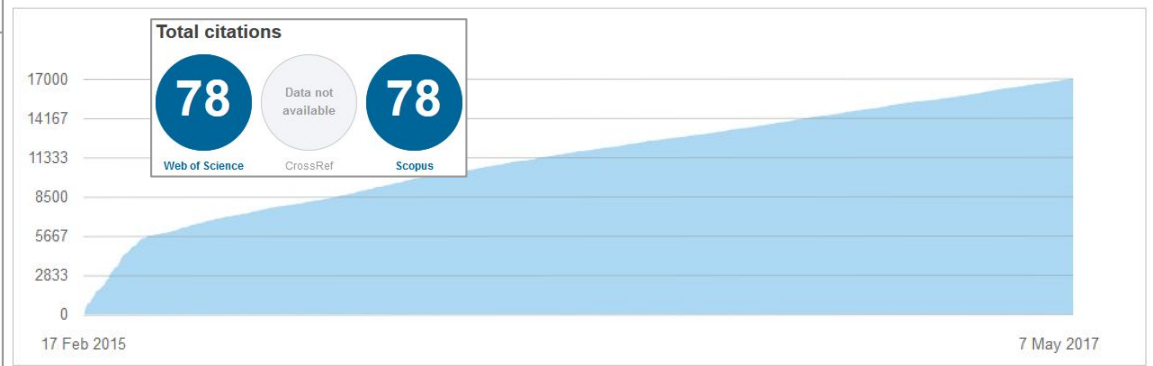
*Nature Protocols* **11**, 1650–1667 (2016) | doi:10.1038/nprot.2016.095

Published online 11 August 2016



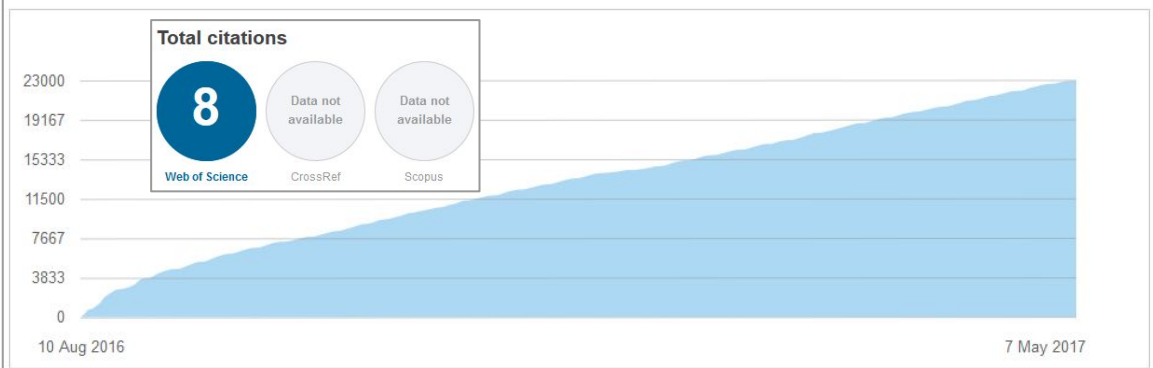
Page views

17,258



Page views

22,856



# 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

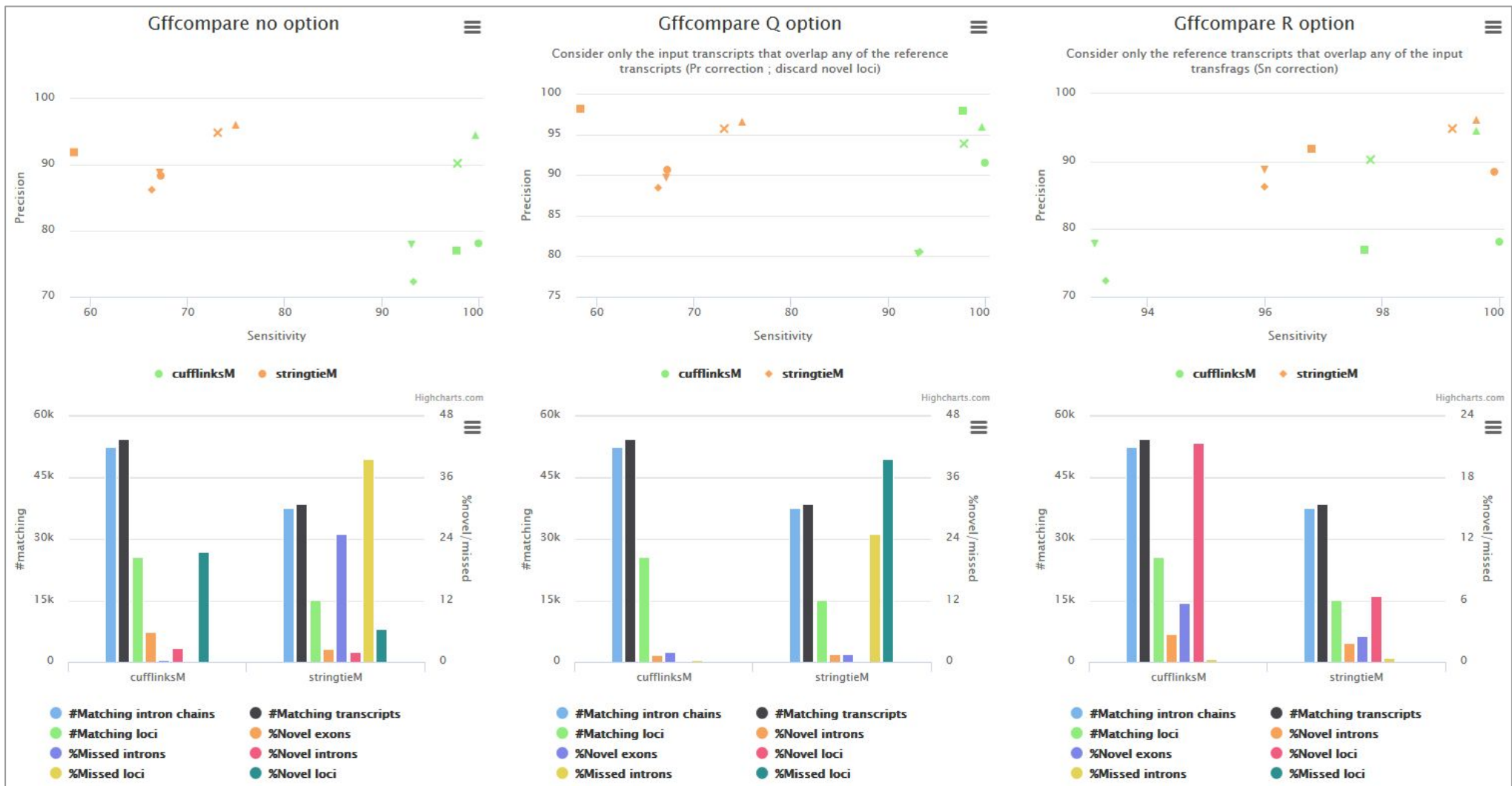
**strtcmp.stats** (transcript assembly accuracy comparison)

```
#= 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*







# Hands-on : star, RSEM with new gtf

## Exercise n°8 (Optional)

Commands :

Star and RSEM: see exercise n°5 and 6

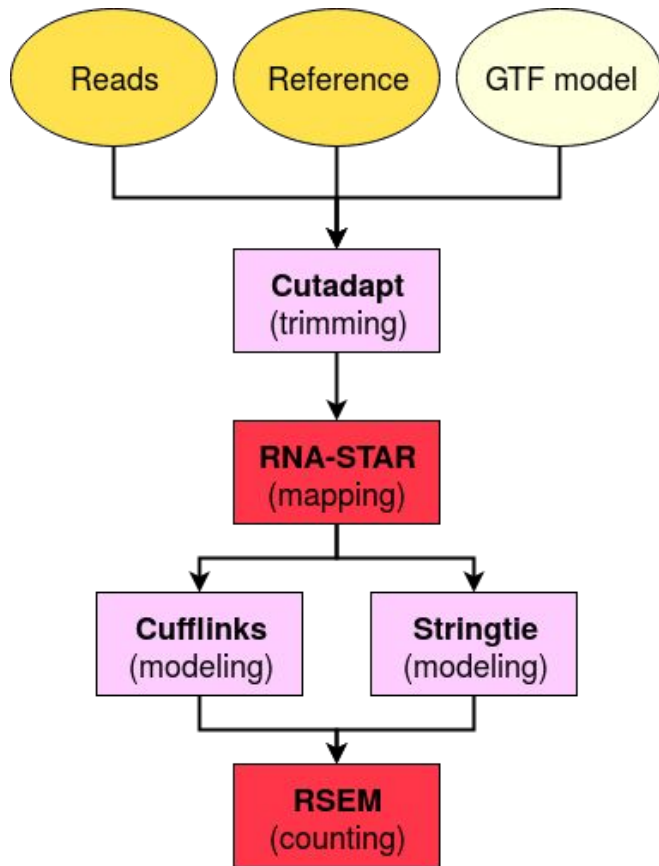
# How to choose count matrix ?

- 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

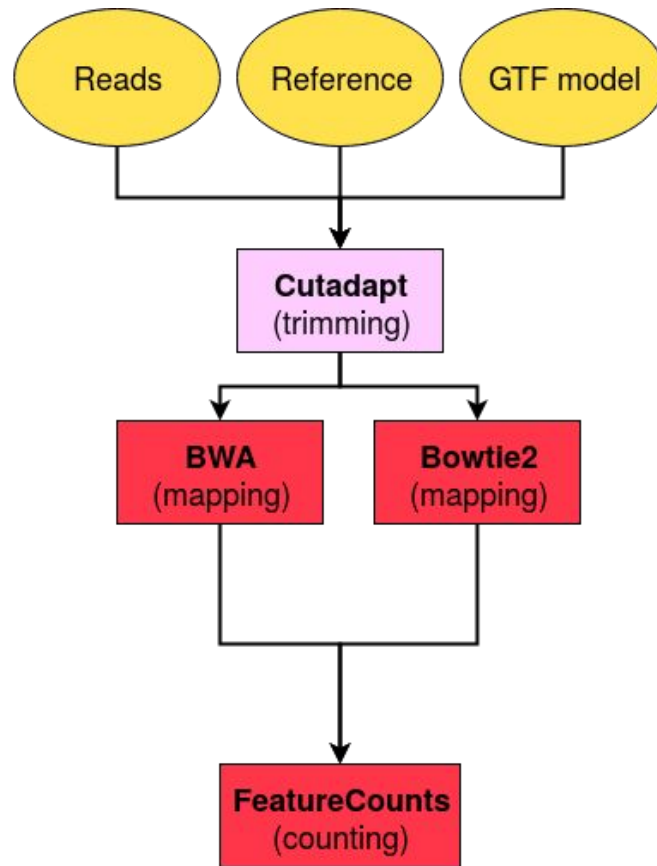
- Workflow management system
- ⇒ configuration, parallelization and monitoring
- Launch a workflow with one command line
  
  - Available on the Genotoul platform
    - `/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py`  
`<workflow_name> <workflow_parameters...>`

# Rna-Seq Workflows on Jflow

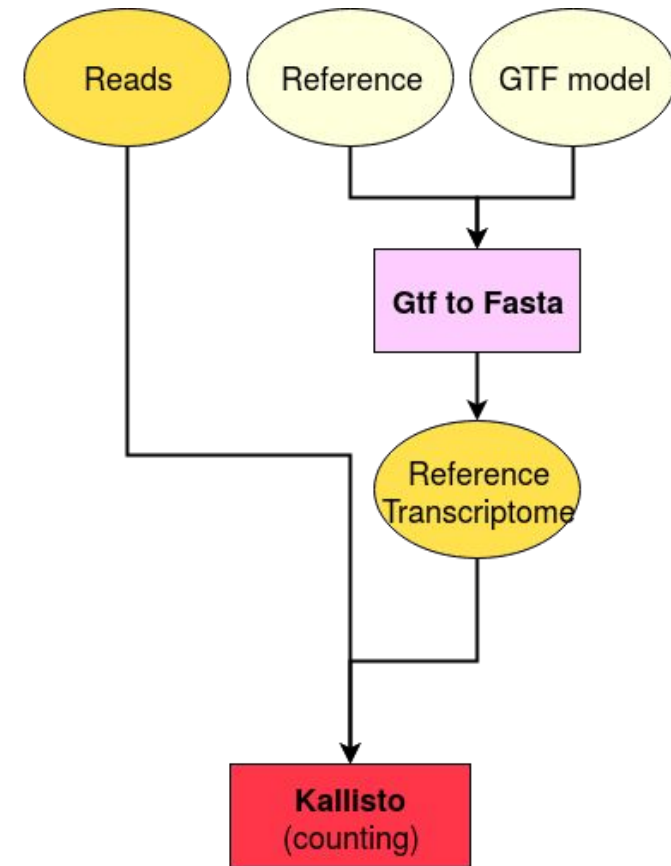
RNA-Seq for Eucaryotes  
(workflow: rnaseq)



RNA-Seq for Procaryotes  
(workflow: rnaseqproc)

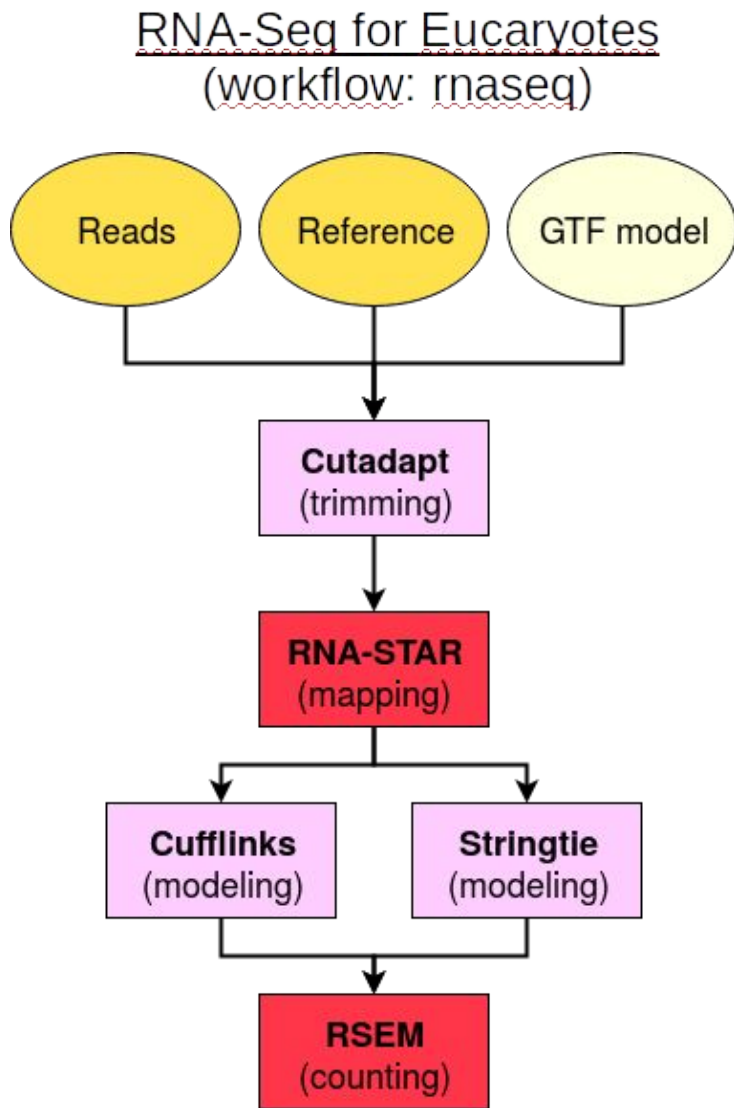


RNA-Seq without alignment  
(workflow: rnaseqnoalign)



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

# Rna-Seq Workflows on Jflow



## Launch workflow:

```
/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py rnaseq  
--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:

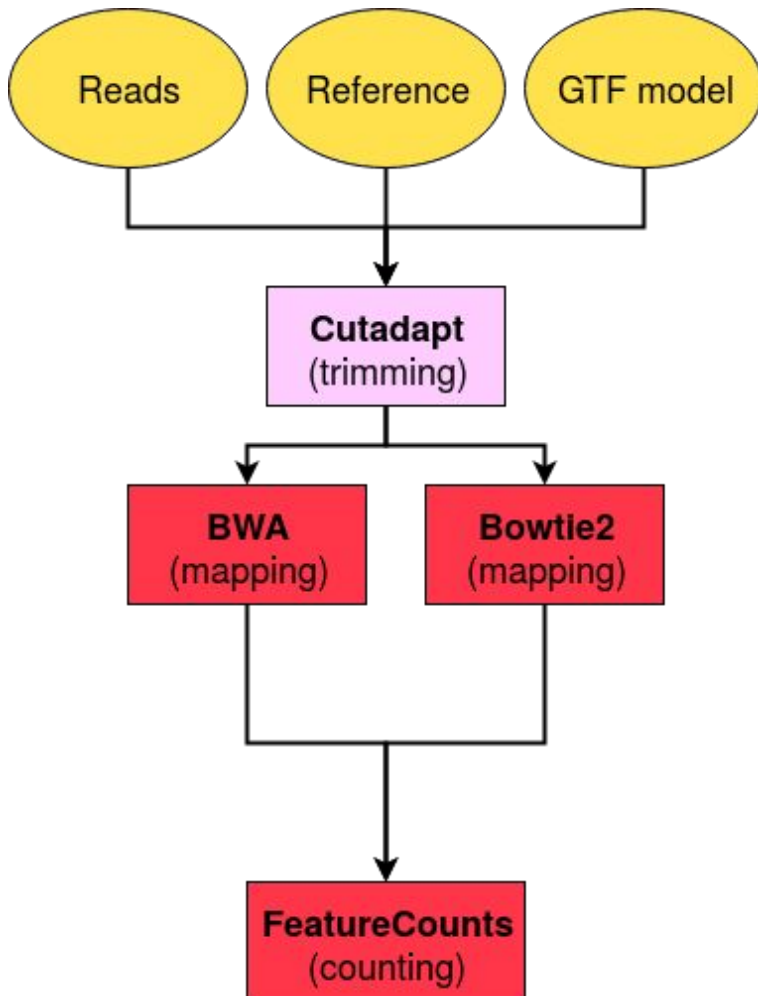
```
--trim-reads : to trim reads before proceeding  
default: TruSeq Adapter  
--compute-gtf-model : to compute a new gtf  
model (gtf-file parameter is optional in this  
case)  
--modeling-software [cufflinks|stringtie]  
(default: cufflinks)
```

To list all parameters available for this workflow:

```
/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py rnaseq --help
```

# Rna-Seq Workflows on Jflow

RNA-Seq for Procarvotes  
(workflow: rnaseqproc)



## Launch workflow:

```
/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py  
rnaseqproc
```

```
--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:

```
--trim-reads: to trim reads default: TruSeq  
Adapter
```

```
--use-bowtie2: use bowtie2 instead of default  
bwa
```

```
--multi-map: If specified, multi-mapping  
reads/fragments will be counted
```

```
--multi-assign: If specified, reads will be allowed  
to be assigned to multiple meta-feature
```

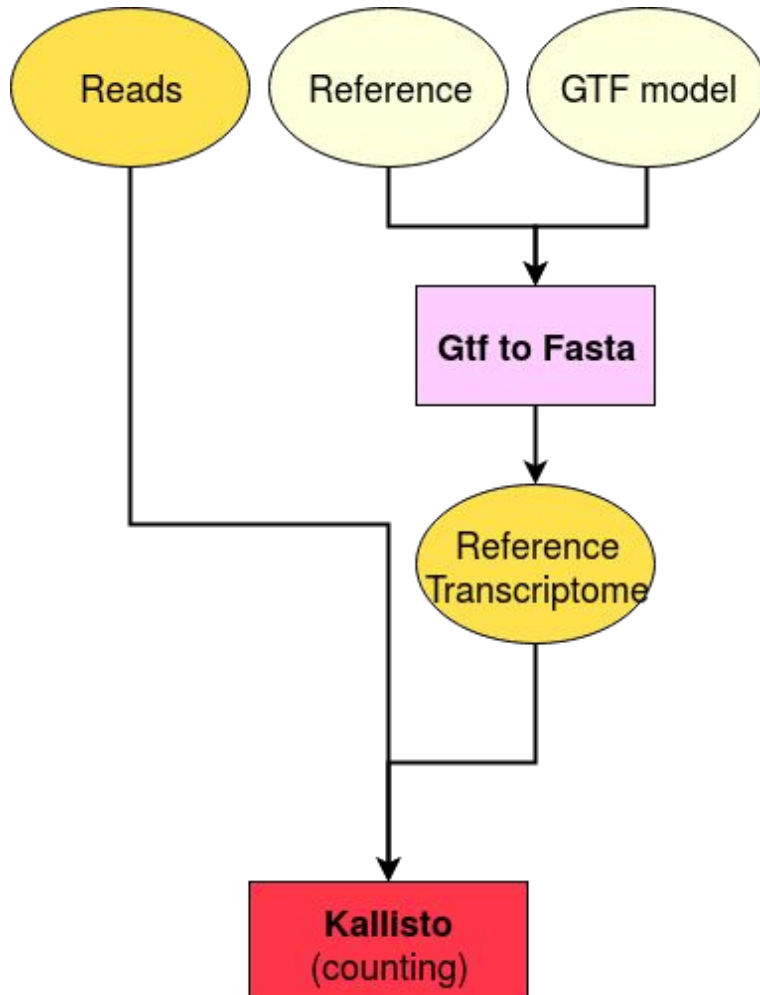
To list all parameters available for this workflow:

```
/usr/local/bioinfo/src/Jflow/jflow/bin/jflow_cli.py rnaseqproc --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.
- The results are in:  
`/work/login/jflow_results/workflowName/wf*/*`. They are specified in the stdout when the pipeline ended.
- In development:
  - A log file containing: the list of commands launched (to have the parameters) and versions of 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

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

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

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

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