RNA-Seq data analysis

17-18 octobre 2019

Céline Noirot et Matthias Zytnicki

Material

- **Slides:**
  - pdf : one per page
    http://genoweb.toulouse.inra.fr/~formation/19_Rnaseq_Cli/doc/Rnaseq_training_012019.pdf
  - pdf : three per page with comment lines
    http://genoweb.toulouse.inra.fr/~formation/19_Rnaseq_Cli/doc/Rnaseq_training_012019_3p.pdf
- **Hands on:**
  - Exercises:
    http://genoweb.toulouse.inra.fr/~formation/19_Rnaseq_Cli/doc/Tps/RNAseq_TP_ligne_cmd_ennonce-Octobre2019.pdf
  - Data files: http://genoweb.toulouse.inra.fr/~formation/19_Rnaseq_Cli/data
  - Results files: http://genoweb.toulouse.inra.fr/~formation/19_Rnaseq_Cli/doc/Tps/Correction.txt

Session organisation

Day 1

Morning (9h00 -12h30) :

- Biological reminds
- Sequence quality
  Theory & exercises
- Spliced read mapping
  Theory & Exercises & Visualisation

Afternoon (14h-17h) :

- Expression quantification
  Theory + exercises
- mRNA calling
  Theory & exercises & Visualisation

Day 2

Morning (9h00 -12h30) :

- Models comparison
  Theory & exercises
- Hovering differential gene expression analyse

Summary – Biological reminds

- ✓ Transcriptome specificity
- ✓ High throughput sequencers
- ✓ Illumina protocol, paired-end library, directional library
- ✓ Experimental protocol
- ✓ RNAseq specific bias
- ✓ How to retrieve public data

GenoToul Bioinfo MIA TOULOUSE

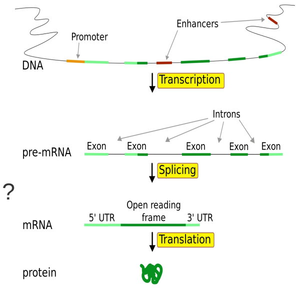
Context

Prerequis :

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

Try to answer to :

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

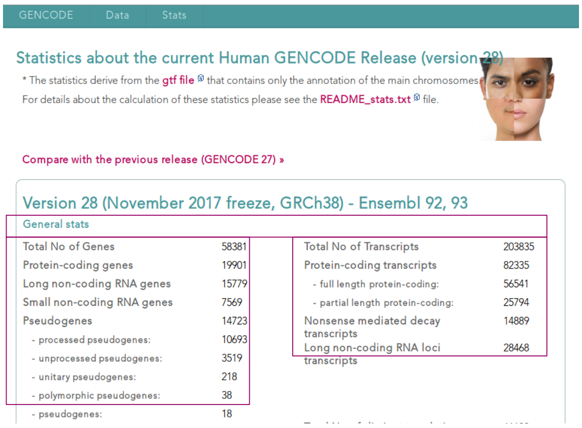


Source : en.wikipedia.org/wiki/User:Forluvoft/sandbox

Transcriptome variability

- Many types of transcripts (mRNA, ncRNA, cis-natural antisense, fusion gene ...)
- 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

Transcriptome variability *(ENCODE)*



GENCODE Data Stats

Statistics about the current Human GENCODE Release (version 28)

* The statistics derive from the **gtf file** that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the **README_stats.txt** file.

Compare with the previous release (GENCODE 27) »

Version 28 (November 2017 freeze, GRCh38) - Ensembl 92, 93

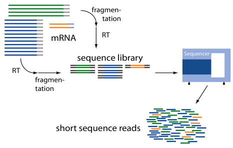
General stats

Total No of Genes	58381	Total No of Transcripts	203835
Protein-coding genes	19901	Protein-coding transcripts	82335
Long non-coding RNA genes	15779	- full length protein-coding:	56541
Small non-coding RNA genes	7569	- partial length protein-coding:	25794
Pseudogenes	14723	Nonsense mediated decay transcripts	14889
- processed pseudogenes:	10693	Long non-coding RNA loci transcripts	28468
- unprocessed pseudogenes:	3519		
- unitary pseudogenes:	218		
- polymorphic pseudogenes:	38		
- pseudogenes:	18		

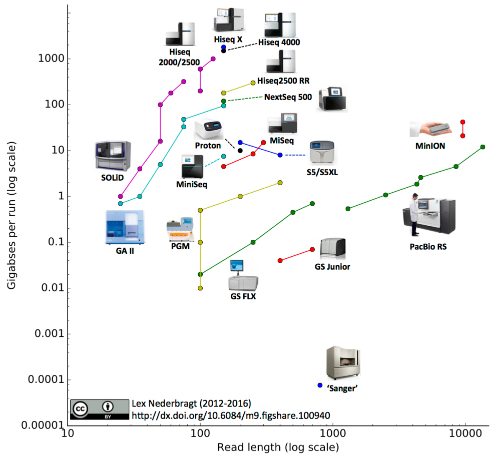
https://www.gencodegenes.org/stats/current.html

What is « new » 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, ...



Sequencing platforms



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

https://flxlexblog.wordpress.com/2016/07/08/developments-in-high-throughput-sequencing-july-2016-edition/#more-790

Illumina Sequencing platforms

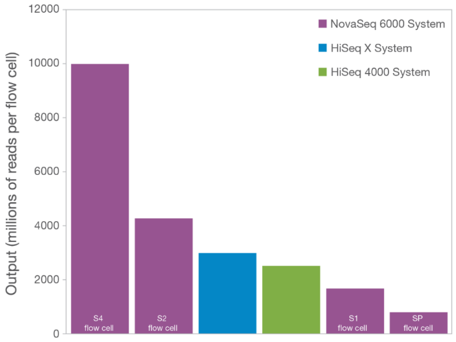
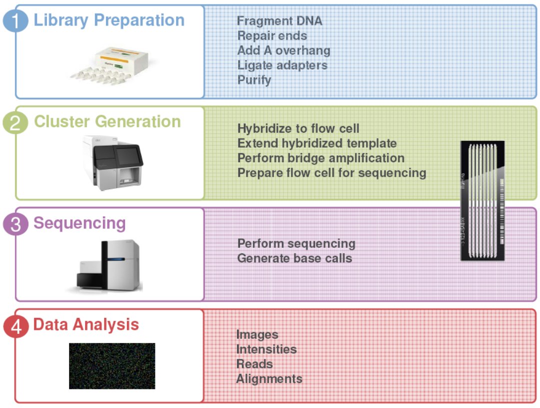


Figure 2: The NovaSeq 6000 System offers the broadest output range—The NovaSeq 6000 System generates from 80 Gb and 800 M reads to 3 Tb and 10 B reads of data in single flow cell mode. In dual flow cell mode, output can be up to 6 Tb and 20 B reads. The tunable output makes the NovaSeq 6000 System accessible for a wide range of applications.

https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/novaseq-6000-system-specification-sheet-770-2016-025.pdf

Illumina RNA-Seq protocol



1 Library Preparation

Fragment DNA
Repair ends
Add A overhang
Ligate adapters
Purify

2 Cluster Generation

Hybridize to flow cell
Extend hybridized template
Perform bridge amplification
Prepare flow cell for sequencing

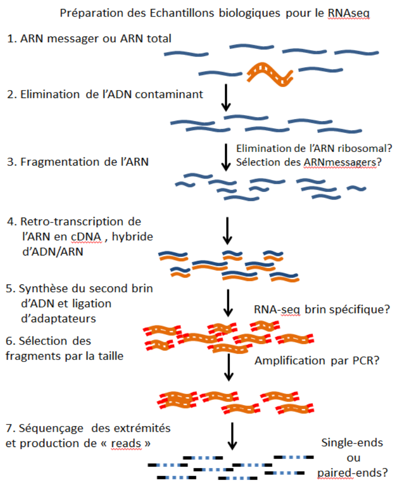
3 Sequencing

Perform sequencing
Generate base calls

4 Data Analysis

Images
Intensities
Reads
Alignments

RNA-Seq library preparation



Préparation des Echantillons biologiques pour le RNAseq

1. ARN messager ou ARN total

2. Elimination de l'ADN contaminant

Elimination de l'ARN ribosomal?
Sélection des ARNmessagers?

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

RNA-seq brin spécifique?

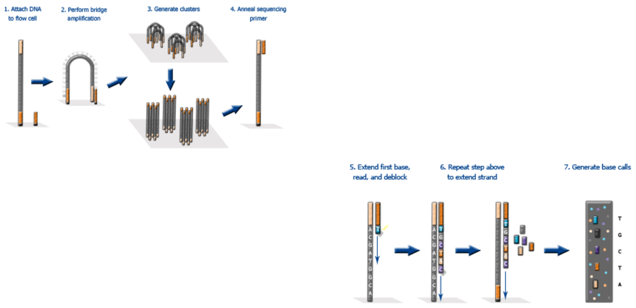
6. Sélection des fragments par la taille

Amplification par PCR?

7. Séquençage des extrémités et production de « reads »

Single-ends
ou
paired-ends?

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
A
C
G
A
T
G
G
C
A
T
G
C
T
A

https://www.eurofinsgenomics.co.in/en/eurofins-genomics/product-faqs/next-generation-sequencing/general-technical-questions/what-is-the-principal-of-the-illumina-sequencing-technology.aspx

How to define experimental protocol ?

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

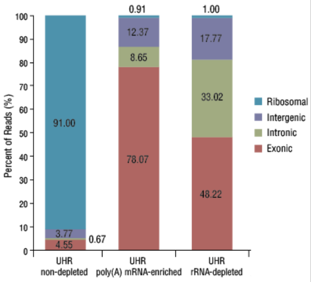
Déplétion / Enrichissement ?

- Similar results

Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling, BMC Genomics , 2014

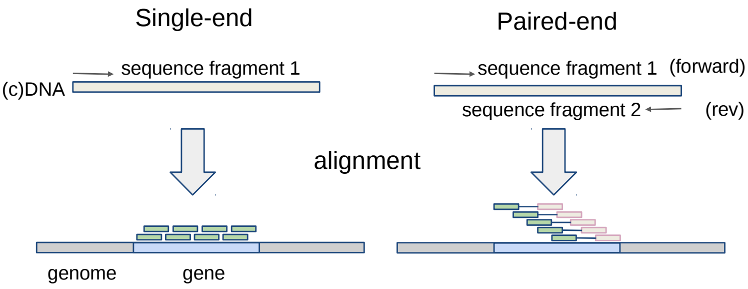
- RNA depletion:
  - For bacterial
  - ARN more varied
  - CircRNA
  - Some ncRNA

- polyA enrichment:
  - More reads into exons
  - Less biological material
  - No transcript without PolyA or partially degraded
  - No circRNA biais



https://content.neb.com/products/e6310-nebnext-rrna-depletion-kit-human-mouse-rat

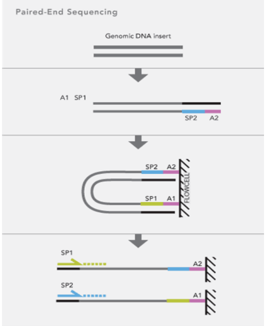
Paired-end VS single-end



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

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



Adapter (A1 and A2) with sequencing primer sites (SP1 and SP2) are ligated onto DNA fragments. Template clusters are formed on the flow cell by bridge amplification and then sequenced by synthesis from the paired primers sequentially.

Strand specific RNA-Seq protocol

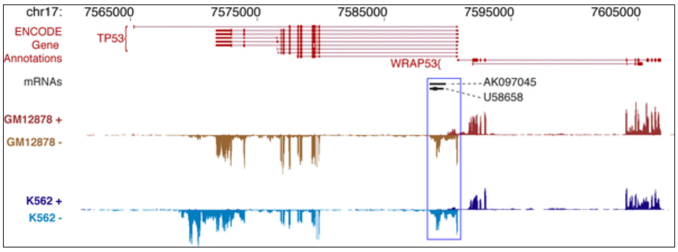
Nat Methods. 2010 Sep;7(9):709-15. Epub 2010 Aug 15.

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

Abstract

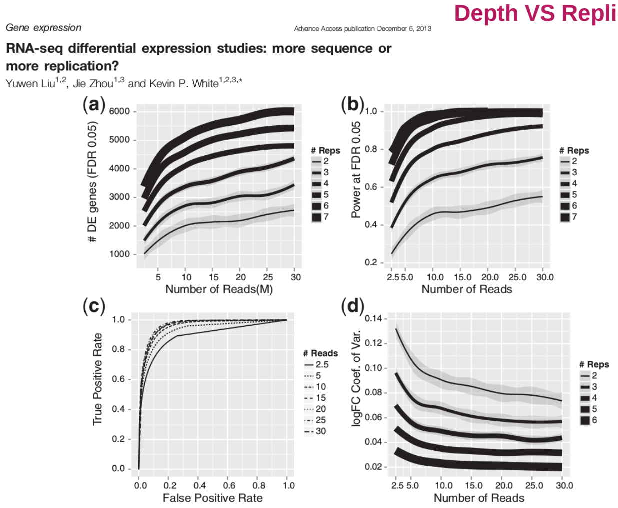


Experimental protocol:
Depth VS Replicates

- Encode (2016):
  - 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.

https://www.encodeproject.org/documents/cede0cbe-d324-4ce7-ace4-f0c3eddf5972/@@download/attachment/ENCODE%20Best%20Practices%20for%20RNA_v2.pdf

Experimental protocol:
Depth VS Replicates



Gene expression

Advance Access publication December 6, 2013

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

Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3,*]

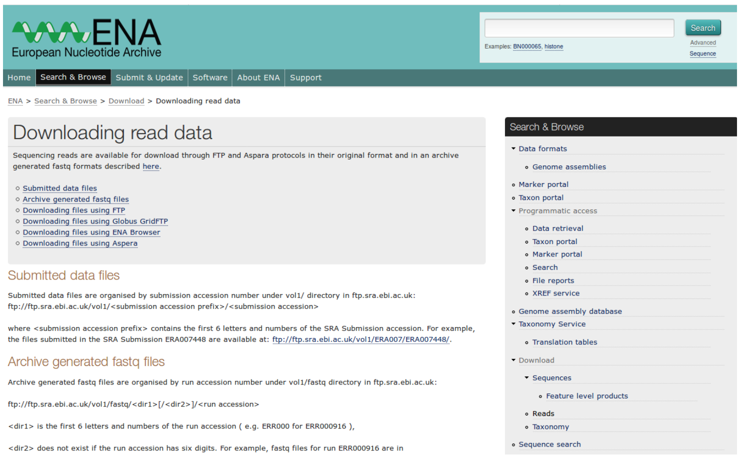
Retrieve public data

Why ?

- Because there is 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 to sequence your own data

Retrieve public data

ENA https://www.ebi.ac.uk/ena



ENA > Search & Browse > Download > Downloading read data

Downloading read data

Sequencing reads are available for download through FTP and Aspara 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),

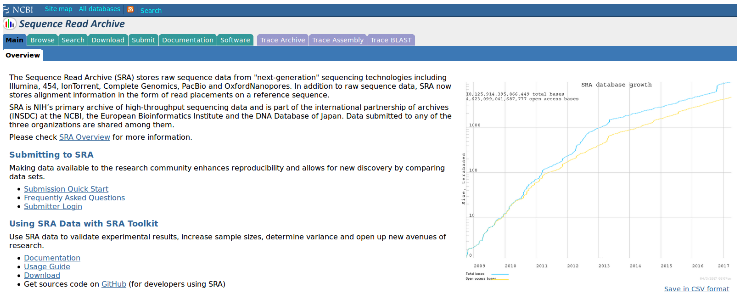
<dir2> does not exist if the run accession has six digits. For example, fastq files for run ERR000916 are in

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

SRA https://www.ncbi.nlm.nih.gov/sra



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.

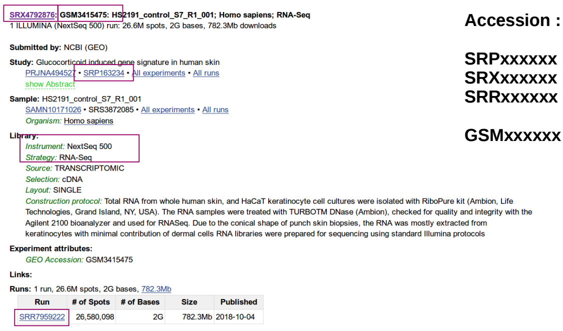
- Submission Quick Start
- Frequently Asked Questions
- Submitter Login

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



SRX4792876: GSM3415475: HS2191_control_S7_R1_001; Homo sapiens; RNA-Seq
1 ILLUMINA (NextSeq 500) run: 26.6M spots, 2G bases, 782.3Mb downloads

Submitted by: NCBI (GEO)

Study: Glucocorticoid induced gene signature in human skin
PRJNA494527 • SRP163234 • All experiments • All runs
show Abstract

Sample: HS2191_control_S7_R1_001
SAMN10171026 • SRS3872085 • All experiments • All runs
Organism: Homo sapiens

Library:
Instrument: NextSeq 500
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: SINGLE
Construction protocol: Total RNA from whole human skin, and HaCaT keratinocyte cell cultures were isolated with RiboPure kit (Ambion, Life Technologies, Grand Island, NY, USA). The RNA samples were treated with TURBOTM DNase (Ambion), checked for quality and integrity with the Agilent 2100 bioanalyzer and used for RNASeq. Due to the conical shape of punch skin biopsies, the RNA was mostly extracted from keratinocytes with minimal contribution of dermal cells RNA libraries were prepared for sequencing using standard Illumina protocols

Experiment attributes:
GEO Accession: GSM3415475

Links:

Runs: 1 run, 26.6M spots, 2G bases, 782.3Mb

Run	# of Spots	# of Bases	Size	Published
SRR7959222	26,580,098	2G	782.3Mb	2018-10-04

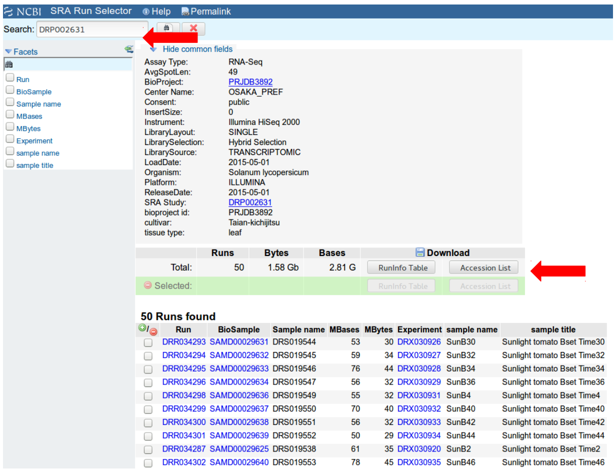
Accession : SRX/ERX/DRX

SRPxxxxxx : Project
SRXxxxxxx : Experiement
SRRxxxxxx : Run

GSMxxxxxx : GEO id

http://bioinfo.genotoul.fr/index.php/faq/bioinfo_tips_faq/

Retrieve public data



Search: DRP002631

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

Hide common fields

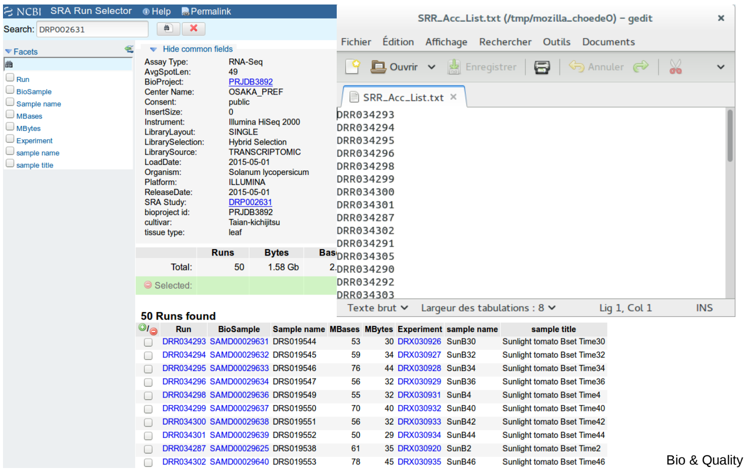
Assay Type:	RNA-Seq
AvgSpotLen:	49
BioProject:	PRJDB3892
Center Name:	OSAKA_PREF
Consent:	public
InsertSize:	0
Instrument:	Illumina HiSeq 2000
LibraryLayout:	SINGLE
LibrarySelection:	Hybrid Selection
LibrarySource:	TRANSCRIPTOMIC
LoadDate:	2015-05-01
Organism:	Solanum lycopersicum
Platform:	ILLUMINA
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	RunInfo Table	Accession List
Selected:				RunInfo Table	Accession List

50 Runs found

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

Retrieve public data



Search: DRP002631

□ 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:			

SRR_Acc_List.txt (/tmp/mozilla_choede0) – gedit

SRR_Acc_List.txt

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

50 Runs found

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

Retrieve public data

- On genologin, use sratoolkit to :
  - download raw file
  - and convert format.

```
mkdir ~/work/ncbi
ln -s ~/work/ncbi ~/ncbi
module load bioinfo/sratoolkit.2.8.2-1
prefetch <sra_accession> --max-size
(20G by default)
Files are created into:
~/work/ncbi/public/sra/
Convertion
fastq-dump --gzip  sra_file.sra
```

Summary - Sequence quality

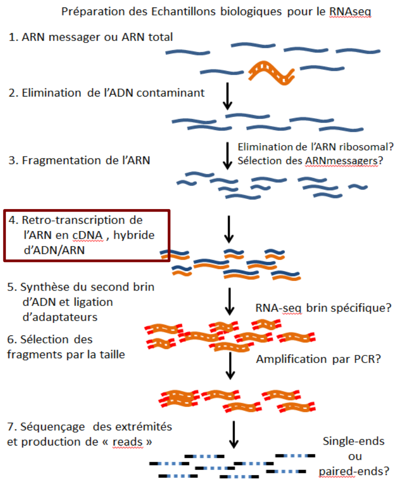
- Known RNAseq biais
- How to check the quality ?
- How to clean the data ?

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
- « Mappability »

Hexamer random priming bias

Random priming
→ not so random



Préparation des Echantillons biologiques pour le RNAseq

1. ARN messager ou ARN total

2. Elimination de l'ADN contaminant

3. Fragmentation de l'ARN

Elimination de l'ARN ribosomal?
Sélection des ARNmessagers?

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

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

RNA-seq brin spécifique?

6. Sélection des fragments par la taille

Amplification par PCR?

7. Séquençage des extrémités et production de « reads »

Single-ends
ou
paired-ends?

Hexamer random priming bias

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

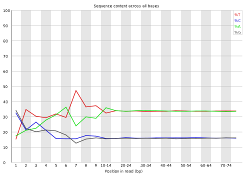
Biases in Illumina transcriptome sequencing caused by random hexamer priming

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

ABSTRACT

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

- 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

Transcript length bias

Biol Direct. 2009 Apr 16;4:14.

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

Oshlack A, Wakefield MJ.

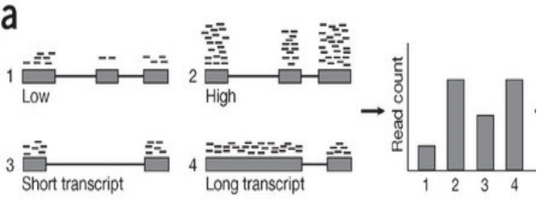
Abstract

Background: Several recent
transcriptome analysis (RNA
genome transcriptional profil
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 662–669
doi:10.1093/bioinformatics/btr005

Gene expression Advance Access publication January 19, 2011

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

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

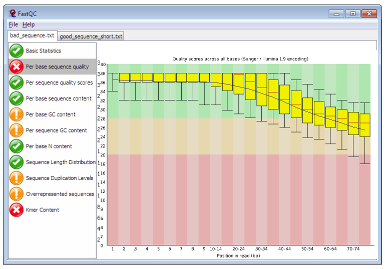
Bias “mappability”

- 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

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
NCTAAGTGTTAGGGGGTTTCCGCCCCTTAGTGCTGCAGCTAACGCATTAAGCACTCCGCCTGGGGAGTACGGTCGCAAGACTGAAAA
+
#<3?BFGGGGGGEGGGGGGGEGGGGGG@F1FGGGGGGDDGG1FB</9FE=EGGGGGGGG>GGGGBGGGGG<<C/BDGGGGGGC=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 th 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

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

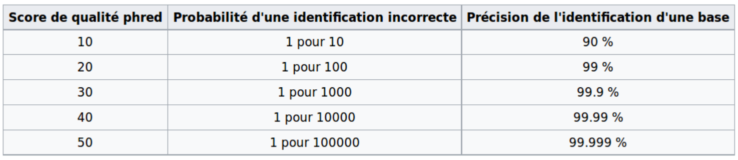
fastq format

- Base quality (Sanger standard)

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

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

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 %

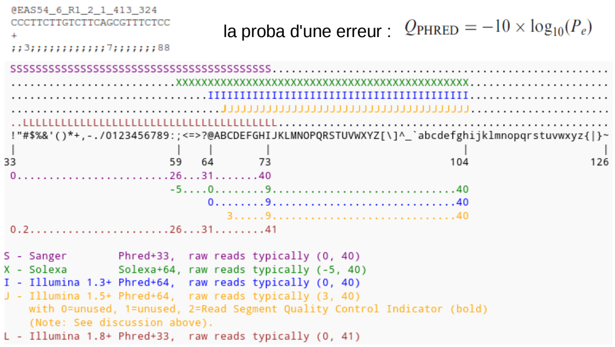
fastq format

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

SURVEY AND SUMMARY

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

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



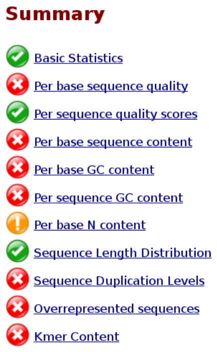
```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
```

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

```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                       59   64       73                            104                   126
0........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                 3......9.............................40
0.2......................26...31........41

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

fastQC Report



Summary

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

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

Quick evaluation of whether the results of the module seem :

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

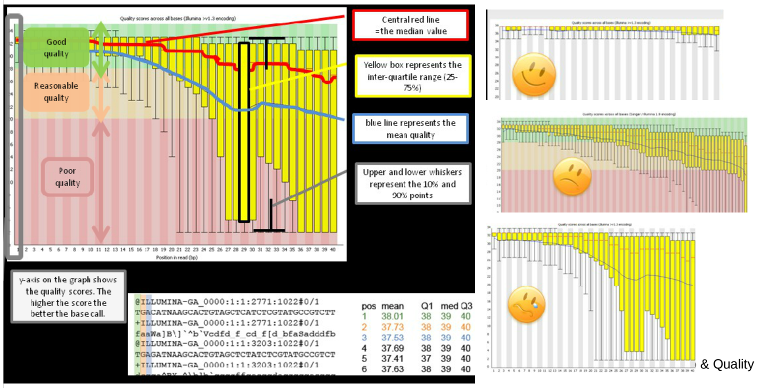
These evaluations must be taken in the context of what you expect from your library. A 'normal' sample as far as FastQC is concerned is random and diverse.

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.



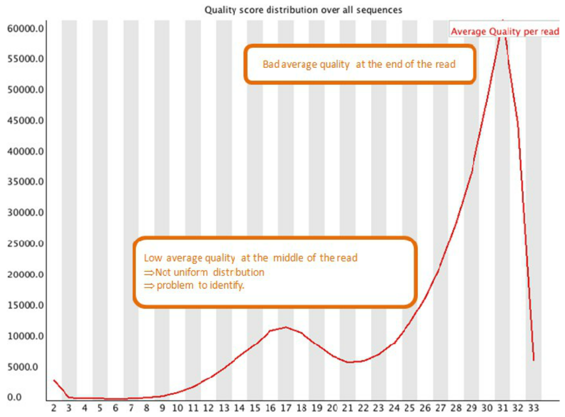
```
@ILLUMINA-GA_0000:1:1:2771:1022#0/1
TGACATNAAGCACTGTAGCTCATCTCGTATGCCGTCTT
+ILLUMINA-GA_0000:1:1:2771:1022#0/1
faaWa]B\]`^b`Vcdfd_f_cd_f[d_bfaSadddfb
@ILLUMINA-GA_0000:1:1:3203:1022#0/1
TGAGATNAAGCACTGTAGCTCTATCTCGTATGCCGTCT
+ILLUMINA-GA_0000:1:1:3203:1022#0/1
```

```
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.



Bad average quality at the end of the read

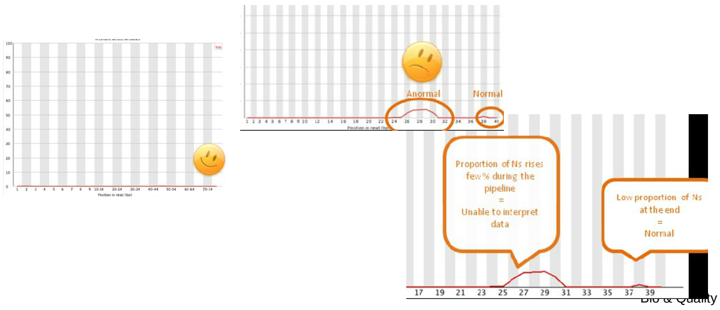
Low average quality at the middle of the read
⇒ Not uniform distribution
⇒ problem to identify.

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

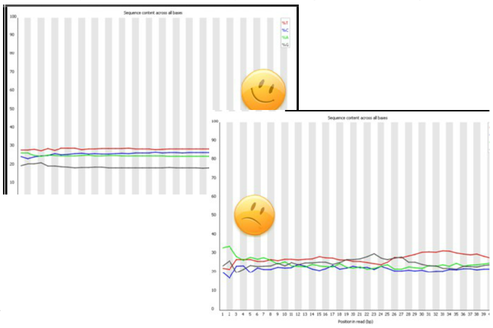
fastqQC Report

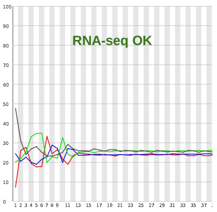
Statistics Per Base Sequence Content

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.





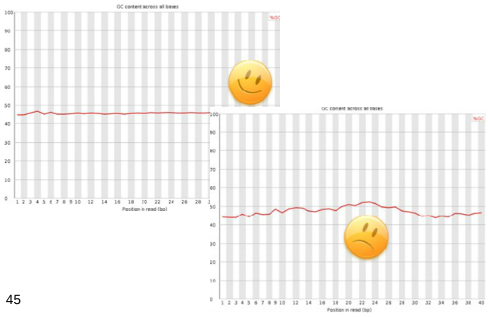
fastqQC Report

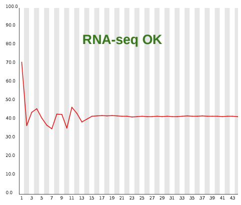
Statistics per Base GC Distribution

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.

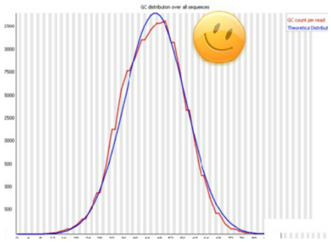


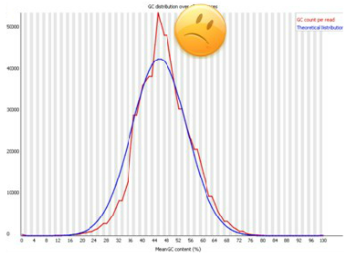


fastqQC Report

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.



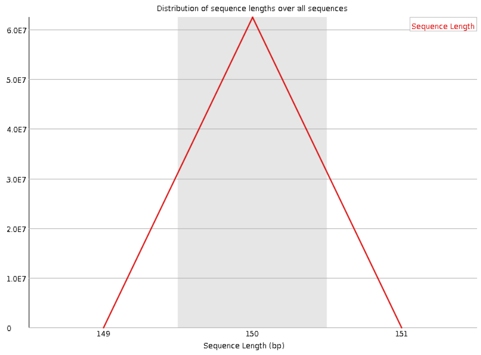


fastqQC Report

Statistics per Sequence Length Distribution

Some sequence fragments contain reads of wildly varying lengths.

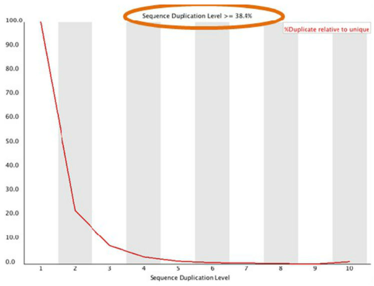
Even within uniform length libraries some pipelines will trim sequences to remove poor quality base calls from the end.

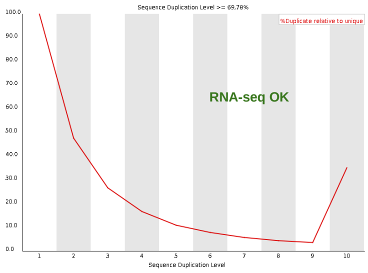


fastqQC Report

Statistics per Duplicate Sequences

High level of duplication indicate an enrichment biais.



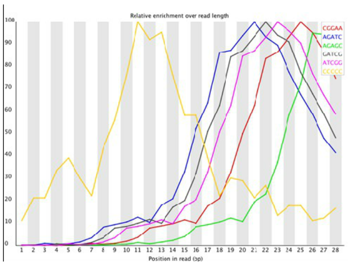


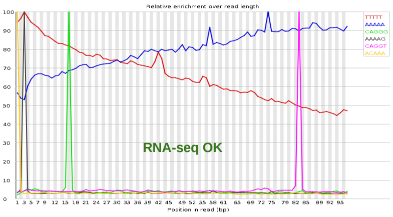
RNA-seq OK

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





RNA-seq OK

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,
- 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

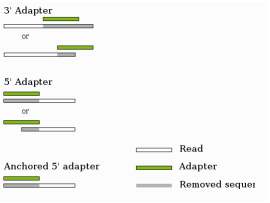
- Cleaning :
  - Low quality bases
  - Adaptors
- Software :
  - Trim_galore
  - Cutadapt
  - Trimmomatic
  - Sickle
  - PRINSEQ
  - ...

Cutadapt

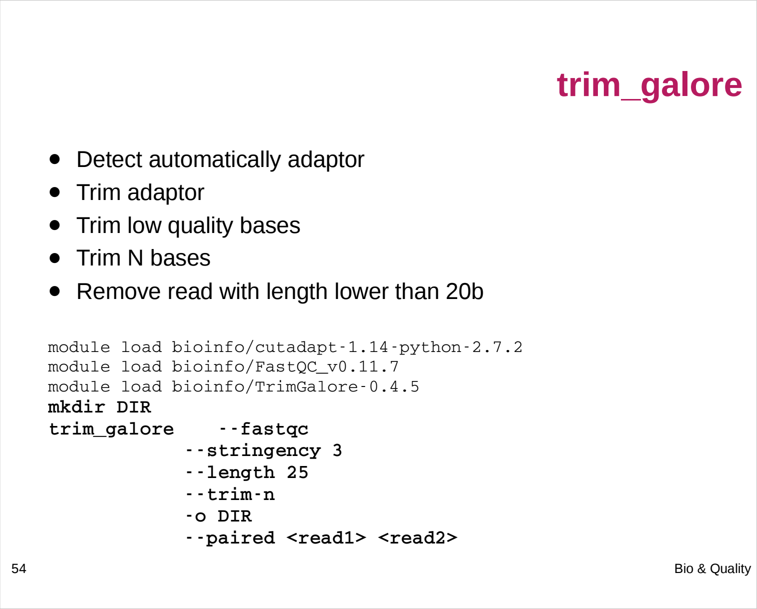
- Searches & removes adapter & tag in all reads.
- Trim quality
- Filter too short or untrimmed reads (in a separate output file).

```
module load bioinfo/cutadapt-1.8.3-python-2.7.2
cutadapt -a ADAPTER_FWD -A ADAPTER_REV -o out1.fatsq -p
out2.fastq  reads1.fastq reads2.fastq
```

Ex.: cutadapt -a AACCGGTT -o output.fastq input.fastq
(3' adapter, single read)
Input file : fasta, fastq or compressed (gz, bz2, xz).



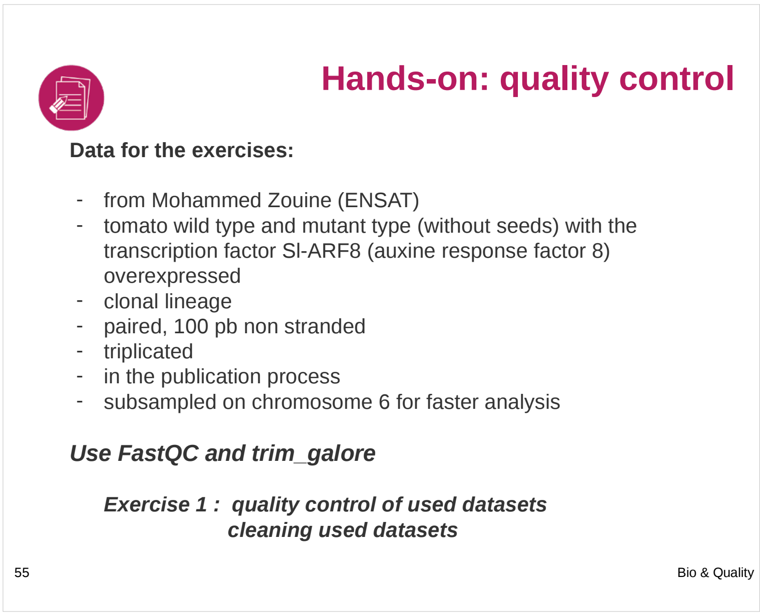
Source : http://cutadapt.readthedocs.io/en/stable/guide.html



trim_galore

- Detect automatically adaptor
- Trim adaptor
- Trim low quality bases
- Trim N bases
- Remove read with length lower than 20b

```
module load bioinfo/cutadapt-1.14-python-2.7.2
module load bioinfo/FastQC_v0.11.7
module load bioinfo/TrimGalore-0.4.5
mkdir DIR
trim_galore    --fastqc
            --stringency 3
            --length 25
            --trim-n
            -o DIR
            --paired <read1> <read2>
```



Hands-on: quality control

Data for the exercises:

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

Use FastQC and trim_galore

Exercise 1 : quality control of used datasets
cleaning used datasets