



RNA-Seq de novo assembly training

Training session aims

- Give you some keys elements to look at during read quality check.
- Transcriptome assembly is not completely a strait forward process :
 - Multiple strategies
 - Multiple software packages
 - Important to know how to check the results and select the best assembly
- Transcriptome assembly is hot :
 - Lot of new software packages and processing chains, small improvement in different parts of the process

Session organisation : Day 1

Afternoon :

- Transcriptome introduction
 - Transcriptome variability
 - RNA-Seq techniques
- RNA-Seq experiment set up
- Read quality assessment
- Read filtering

Session organisation : Day 2

Morning :

- Assembly quality assessment
 - Assemblathon stats
 - Read mapping stats
- Clustering
 - cd-hit
- Greedy assembly
 - vcake

Afternoon :

- Overlap Layout Consensus
 - cap3
 - tgi
- de Bruijn graph based
 - Velvet/Oases
 - Trinity
- Comparing results on test sets

Session organisation : Day 3

Morning :

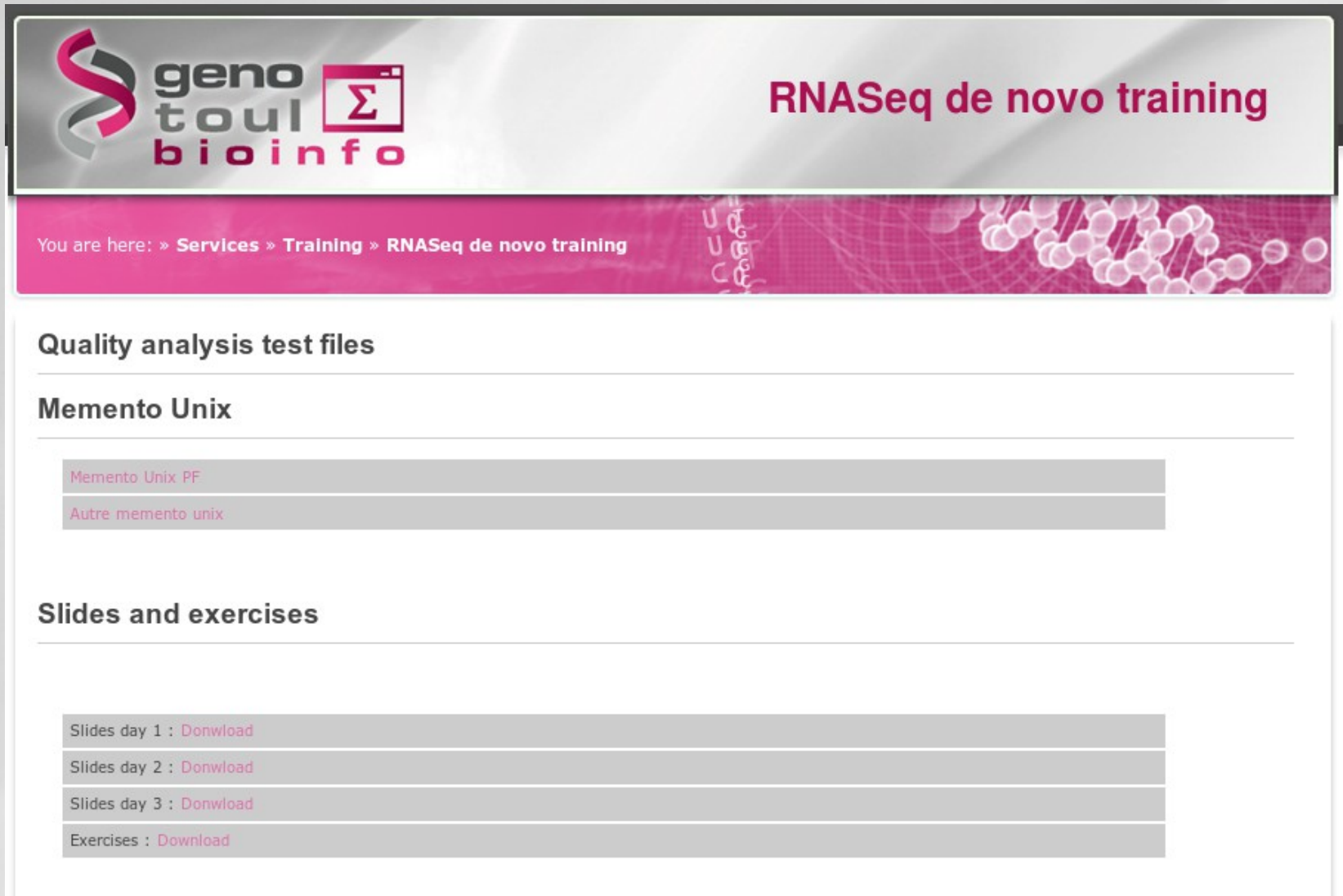
- Assembly quality common problems
 - Frame-shifts
 - Chimera
- Assembly quality assessment using biological knowledge
 - Cegma
 - Blat to reference

Afternoon :

- From transcript to unigene
- Publishing your transcriptome in TSA

Attendees presentation

1. Name and laboratory of origin
2. Species of interest
3. Scientific question
4. Experimental design
5. Data type
6. Current knowledge about de novo RNA-Seq data processing
7. Your expectations



genotoul bioinfo **RNASEq de novo training**

You are here: » [Services](#) » [Training](#) » [RNASEq de novo training](#)

Quality analysis test files

Memento Unix

- [Memento Unix PF](#)
- [Autre memento unix](#)

Slides and exercises

- Slides day 1 : [Download](#)
- Slides day 2 : [Download](#)
- Slides day 3 : [Download](#)
- Exercises : [Download](#)

<http://bioinfo.genotoul.fr/index.php?id=137>

The platform FAQ gives you information about :

- How to connect to the servers.
- How to set-up, run and monitor jobs.

<http://bioinfo.genotoul.fr/index.php?id=11>

Objectives for this first half day

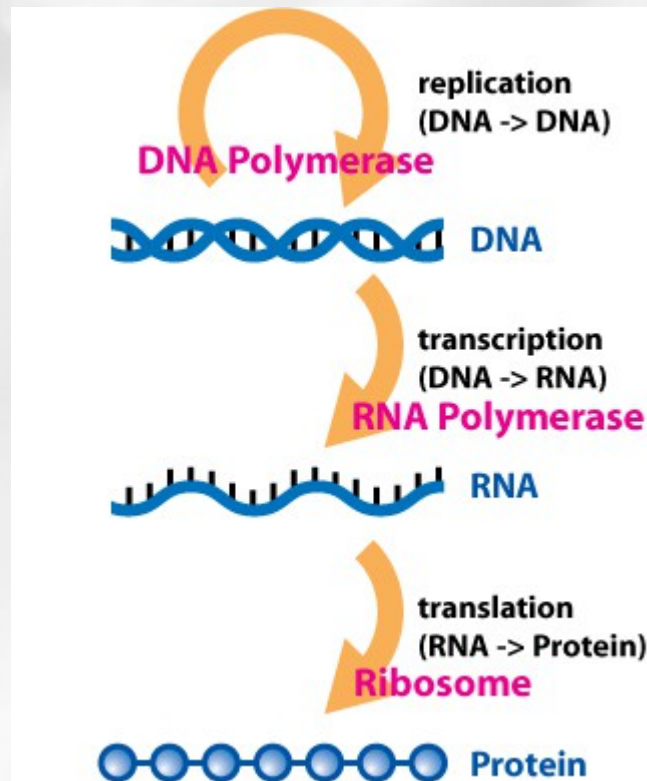
Answer the following questions :

- What is a transcriptome?
- What are the variability factors encountered?
- Why do we use RNA-Seq data?
- Which sequencing protocols are available?
- How do we check the quality of the data-sets?
- Do we keep all the reads, all the nucleotides for the assembly process?

Transcription

Molecular biology dogma

This dogma has been described as “DNA makes RNA makes protein”



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

First exercise

In two groups :

- Make a list of all the **transcription products** you know or heard of.
- Organize your list to present it to the other group members.

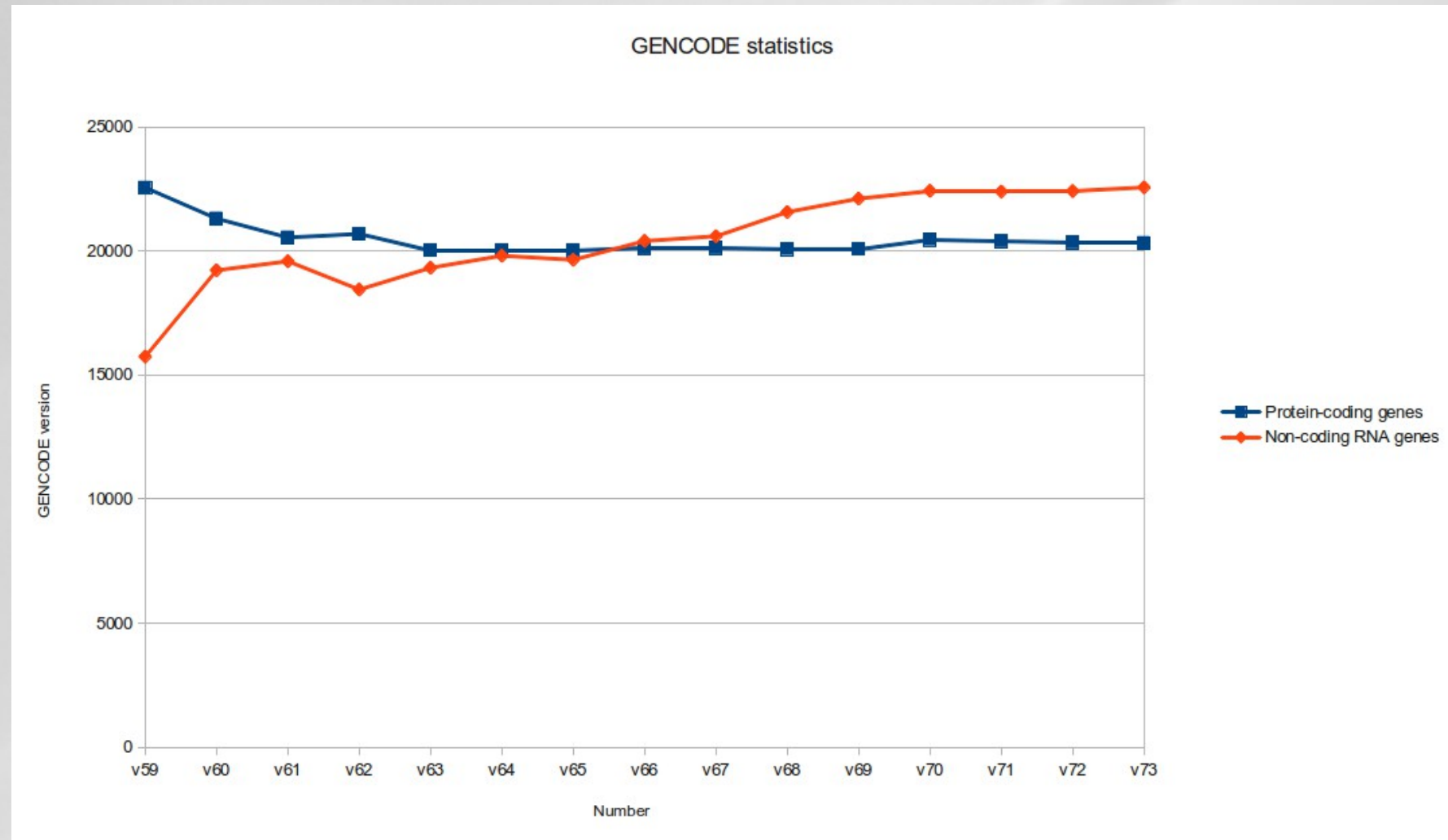
Version 18 (April 2013 freeze, GRCh37) - Ensembl 73

General stats

Total No of Genes	57445	Total No of Transcripts	195584
Protein-coding genes	20318	Protein-coding transcripts	81673
Long non-coding RNA genes	13562	- full length protein-coding:	56953
Small non-coding RNA genes	8998	- partial length protein-coding:	24720
Pseudogenes	14181	Nonsense mediated decay transcripts	12985
- processed pseudogenes:	10585	Long non-coding RNA loci transcripts	23105
- unprocessed pseudogenes:	2873		
- unitary pseudogenes:	165		
- polymorphic pseudogenes:	36		
- pseudogenes:	292		
Immunoglobulin/T-cell receptor gene segments		Total No of distinct translations	61482
- protein coding segments:	386	Genes that have more than one distinct translations	13602
- pseudogenes:	230		

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

GENCODE gene statistics

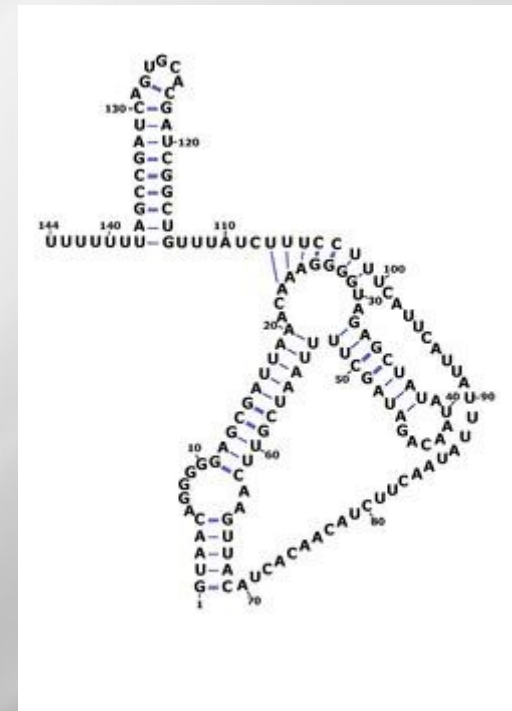


Transcription products

Protein coding gene: transcribed in mRNA

ncRNA : highly abundant and functionally important RNA

- tRNA,
- rRNA,
- Regulatory RNA
 - * snoRNAs (rRNA maturation)
 - * microRNAs (post-transcriptional regulators)
 - * siRNAs (mRNA degradation)
 - * piRNAs (block the activity of the mobile elements)
 - * LincRNA (regulators of diverse cellular processes)
 - * VlincRNA...



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

Second exercise

In two groups :

- List the **transcription variability factors** you know.
- Figure out the impact of these factors on the view of the transcriptome given by the assembly.
- Are there other phenomena which could hinder the assembly?

Which transcriptome variability factors can impact the assembly process?

Assembly take place on the mRNA sequence level :

- Biological elements which tend to blur the signal
 - * Repeats
 - * Gene families
 - * Pseudogenes
 - * (Cis-)natural anti-sens transcript
 - * Fusion genes
 - * Alternative splicing
 - * Intron retention

Elements removing or masking the signal :

- Expression level
- Transcript decay
- Sequencing protocol biases
- Sequencing depth

Other elements :

- PolyA tails
- Adapters
- Contamination

Gene expression law (SAGE data)

VOLUME 90, NUMBER 8

PHYSICAL REVIEW LETTERS

week ending
28 FEBRUARY 2003

Zipf's Law in Gene Expression

Chikara Furusawa

Center for Developmental Biology, The Institute of Physical and Chemical Research (RIKEN), Kobe 650-0047, Japan

Kunihiko Kaneko

Department of Pure and Applied Sciences, University of Tokyo, Komaba, Meguro-ku, Tokyo 153-8902, Japan

(Received 27 September 2002; published 26 February 2003)

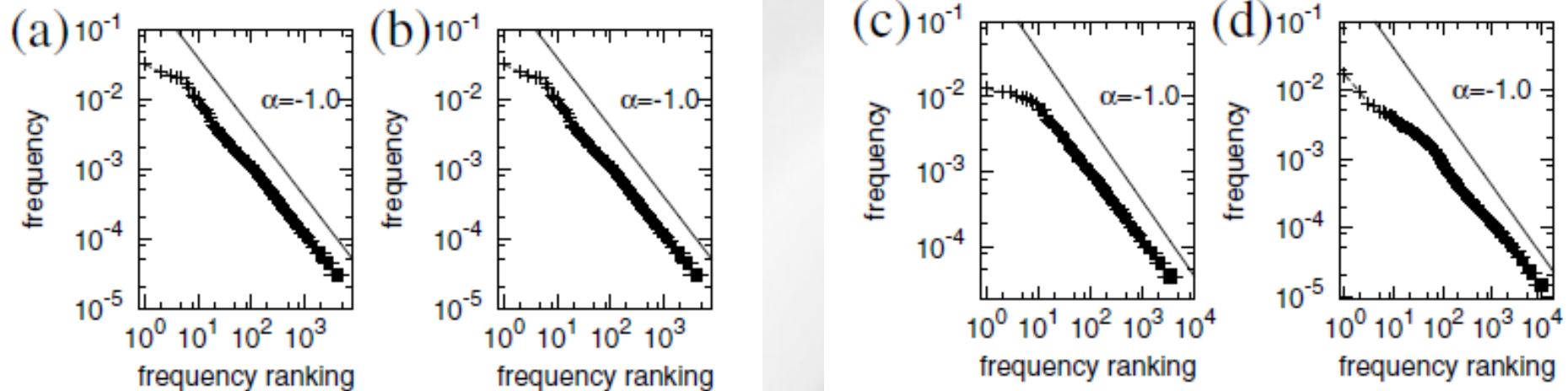
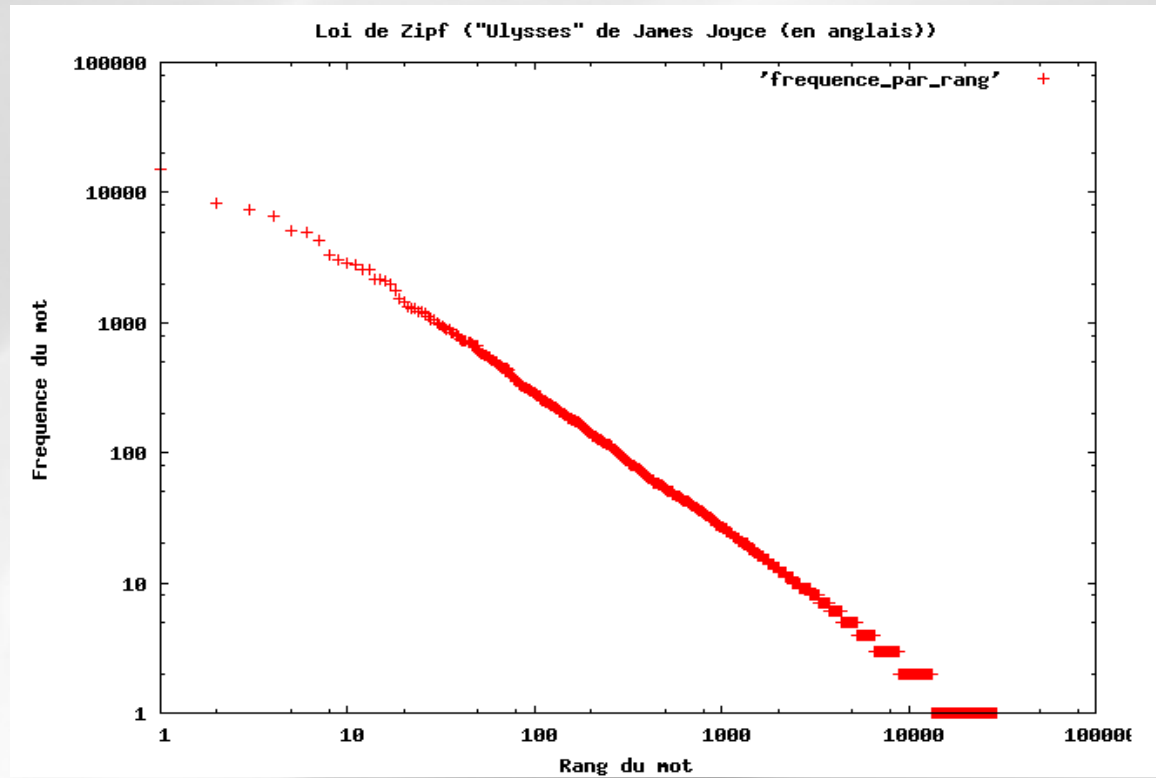


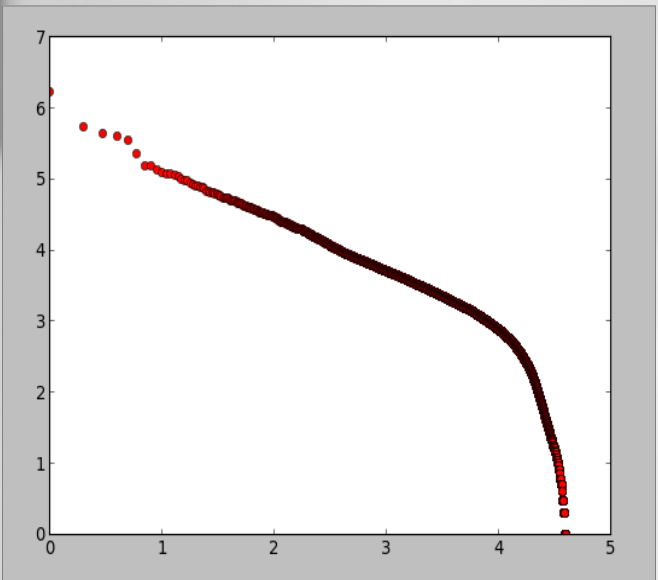
FIG. 1. Rank-ordered frequency distributions of expressed genes. (a) Human liver, (b) kidney, (c) human colorectal cancer, (d) mouse embryonic stem cells, (e) *C. elegans*, and (f) yeast (*S. cerevisiae*). The exponent of the power law is in the range from -1 to -0.86 for all the samples inspected, except for two plant data (seedlings of *Arabidopsis* and the trunk of *Pinus taeda*), whose exponents are approximately -0.63 .

Zipf's law

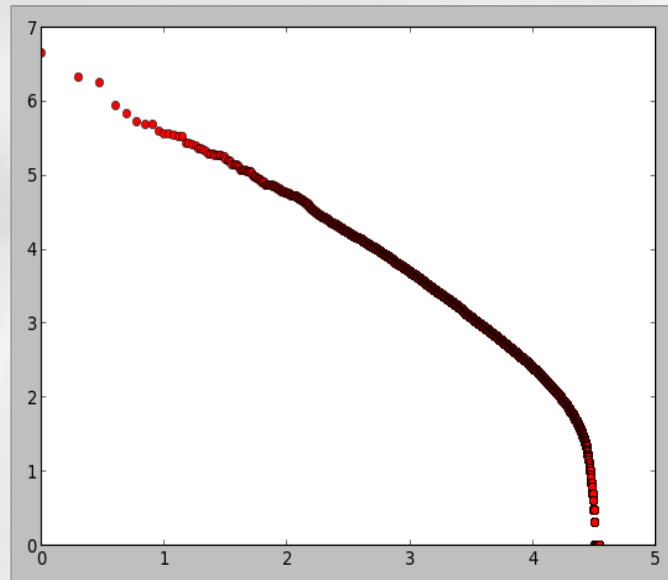
- Zipf's law is an empirical observation on the frequency of words in a text.
- Highlights the relationship between the occurrence of a word in a text and its rank in the order of occurrences.
- Highlights the difference in magnitude of occurrences.



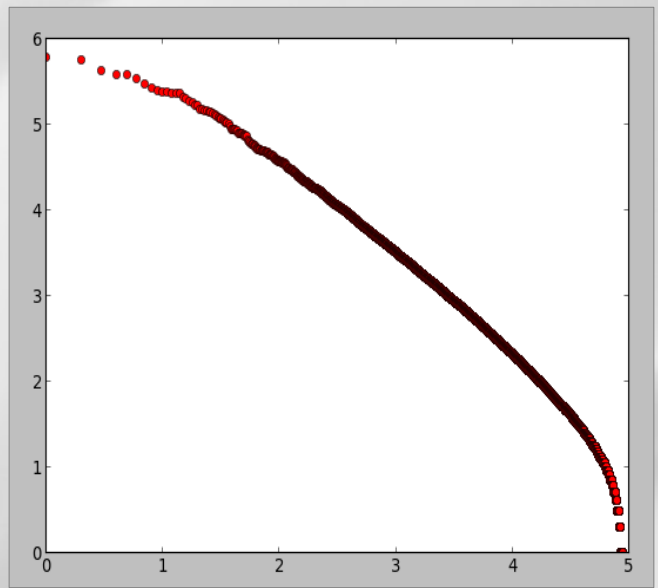
Some examples on local data



Centella asiatica



Dicentrarchus labrax



Meta transcriptome

Zipf like, but the end!

How can we study the transcriptome?

How can we study the transcriptome?

Different techniques :

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

Techniques classification

	EST	PCR/ RT-QPCR	SAGE	Micro- Array
Quantification	No	Yes	Yes	Indirect
Throughput	Low	Low (hundreds)	High (thousands)	High (millions)
Discovery ?	yes	no	no	No (except tiling)

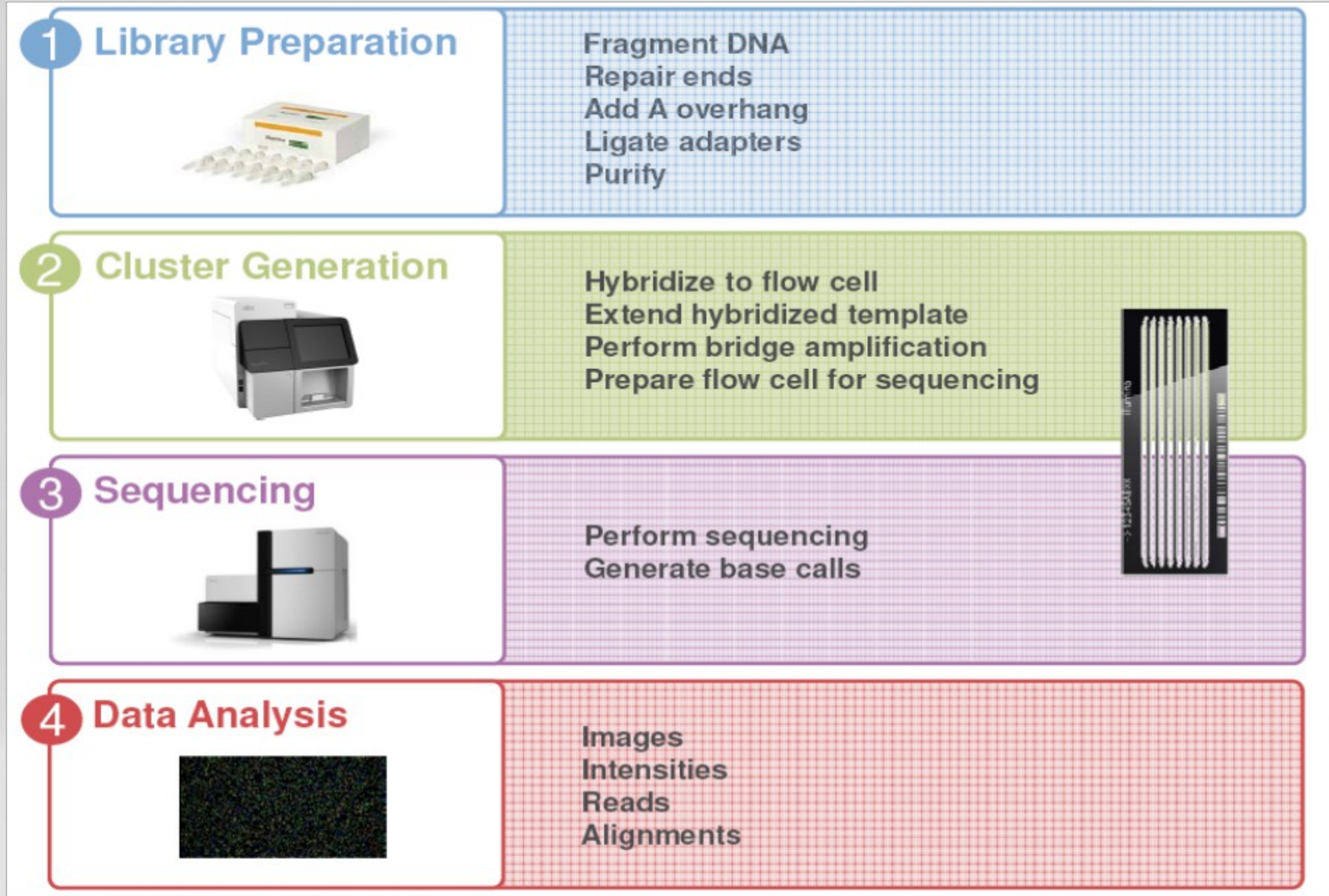
RNA-Seq platforms comparison

Séquenceurs 2 ^{ème} génération													
Société	Roche			Illumina				Life Technologies					
Plateforme													
Technologie	Titanium	FLX Titanium	FLX +					Chip 314	Chip 316	Chip 318			
Acides nucléiques (matrice)													
Ligation adaptateurs													
Méthode d'amplification	 PCR en émulsion			 « Bridge PCR »				 PCR en émulsion					
Méthode de séquençage	Synthèse (Pyroséquençage)			Synthèse				Ligation					
Durée de séquençage/run	10h	10h	20h	26h	8jrs	8jrs	14jrs	2h	12jrs	8jrs	8jrs		
Capacité (Mb) séquençage/run	50	500	900	1500	100000	200000	95000	>10	>100	>1000	70000	80000	150000
Taille moyenne des reads	400	400	700	150+150	100+100	100+100	150+150	100	>100	>100	50+35	75+35	75+35
Coût (\$) /run	1100	6200		750	10000	20000	11500	500	750	950	8150	6100	10500
Coût machine + annexes ((K\$))	110+25	500+30		125	560	690	250	50+20			480+55	350+55	600+55
Exactitude de séquençage (%)	99	99		99,9	99,9	99,9	99,9	99			99,95	99,95	99,99

Interests of the RNA-Seq approach?

- No prior knowledge of the sequenced genome 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, ...

Illumina RNA-Seq protocol



Library preparation

It is a very important step because it defines the transcripts which will be monitored.

How do we get rid of the ribosomal RNA?

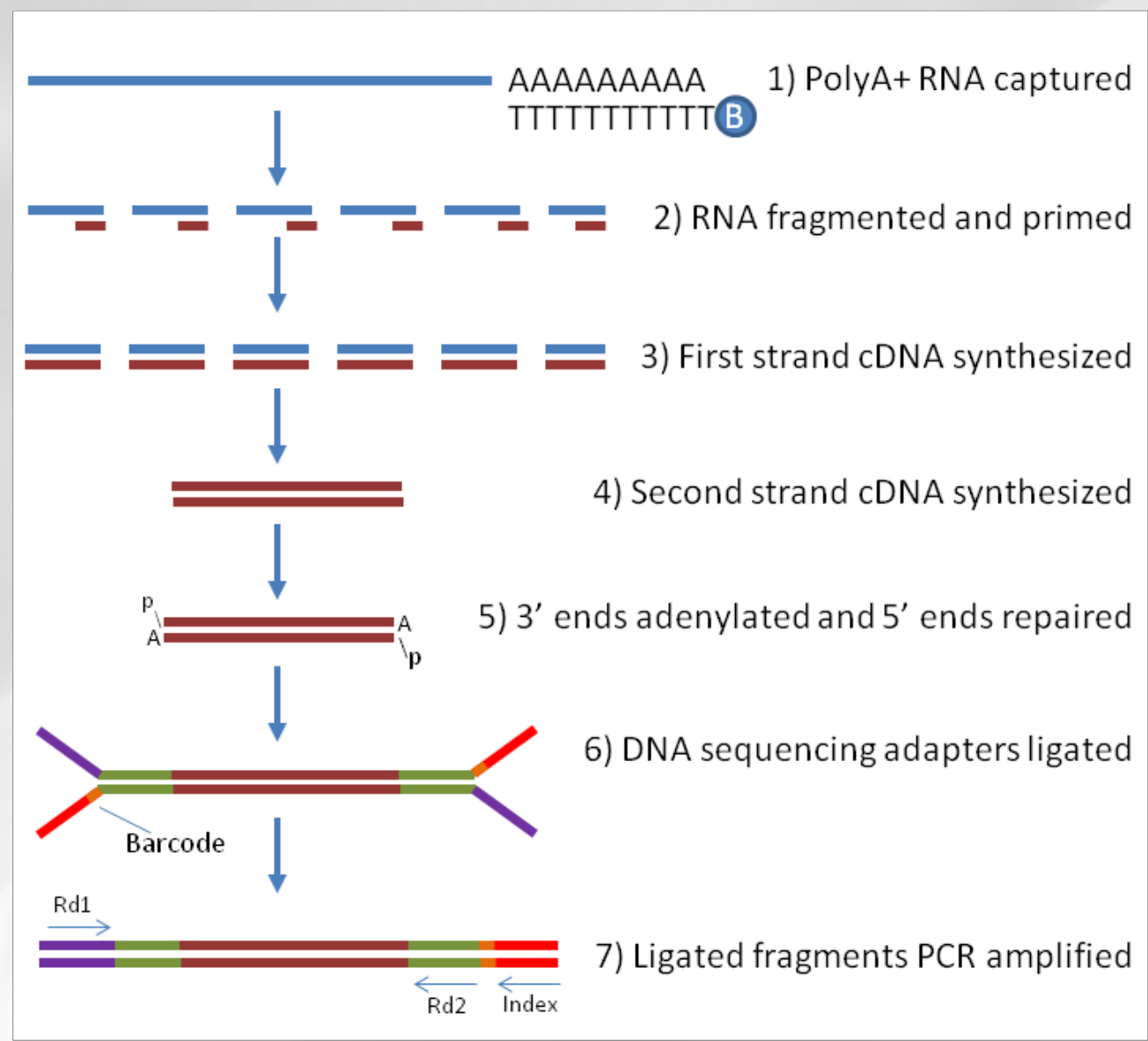
- PolyA tails picking
- Ribosomal RNA depletion

How do we get a complete view of the transcript?

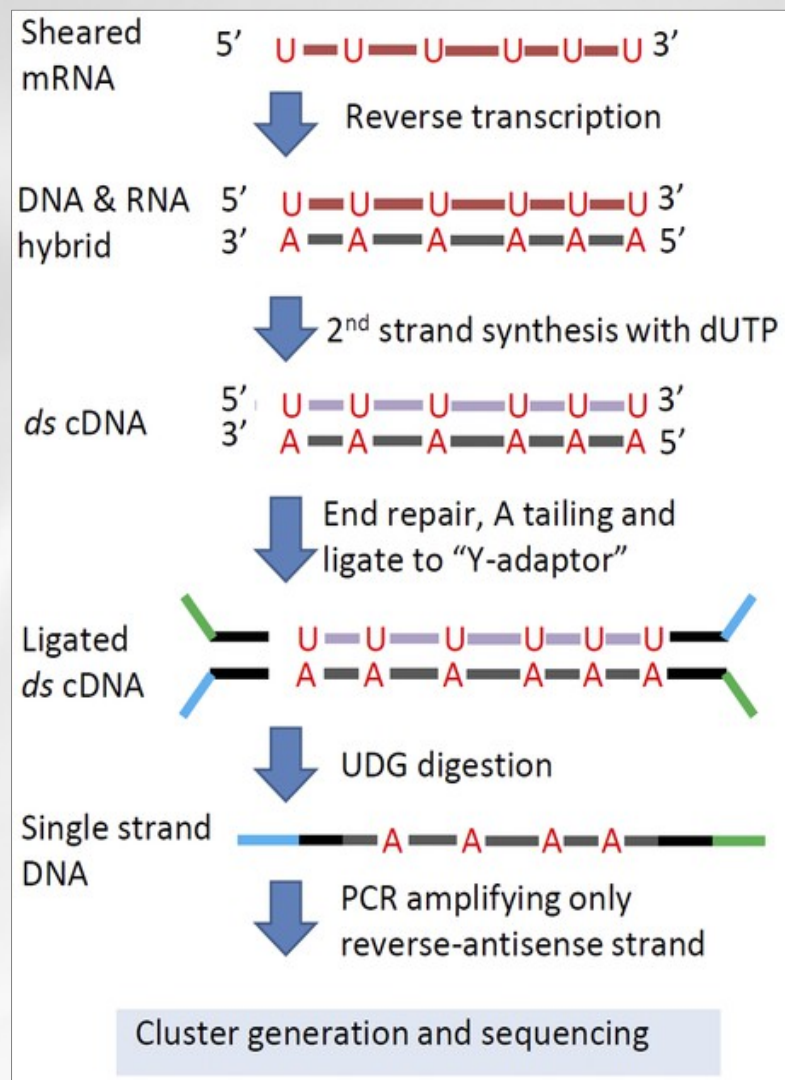
- 3' end priming
- Random priming
- Adapter priming (SMART)

How do we get a strand specific signal?

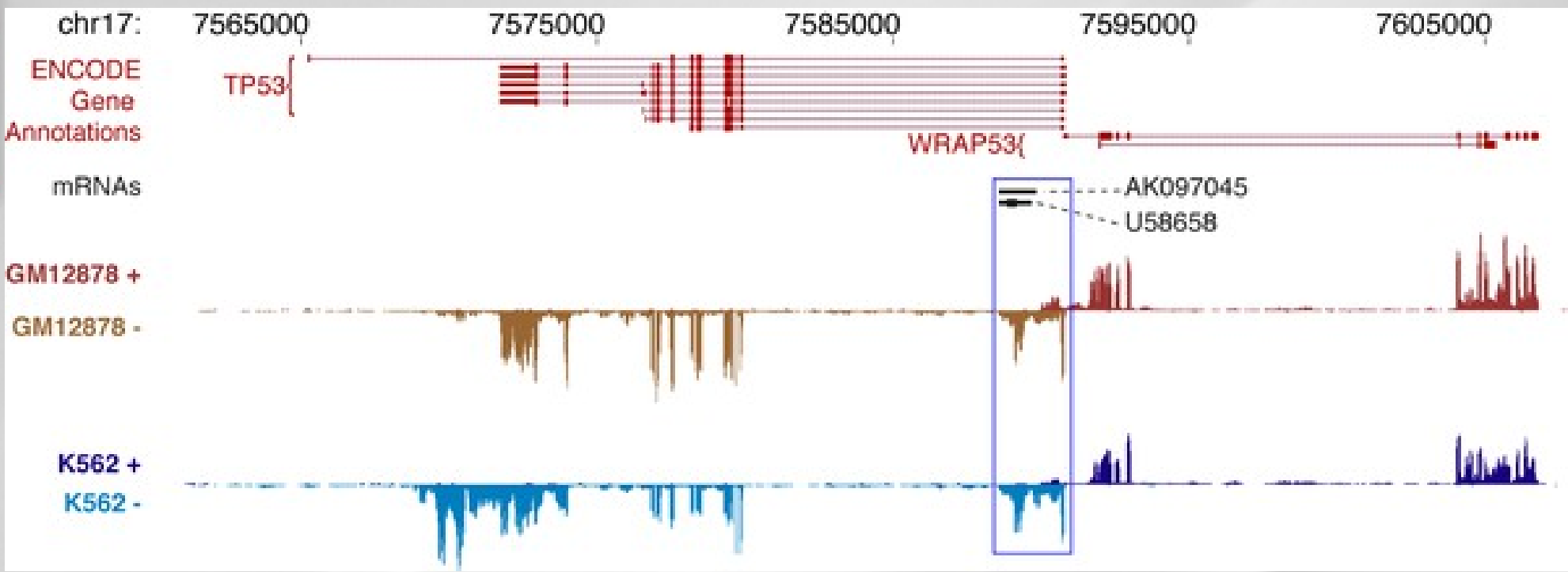
Illumina library preparation



Strand specific libraries



Strand specific alignment



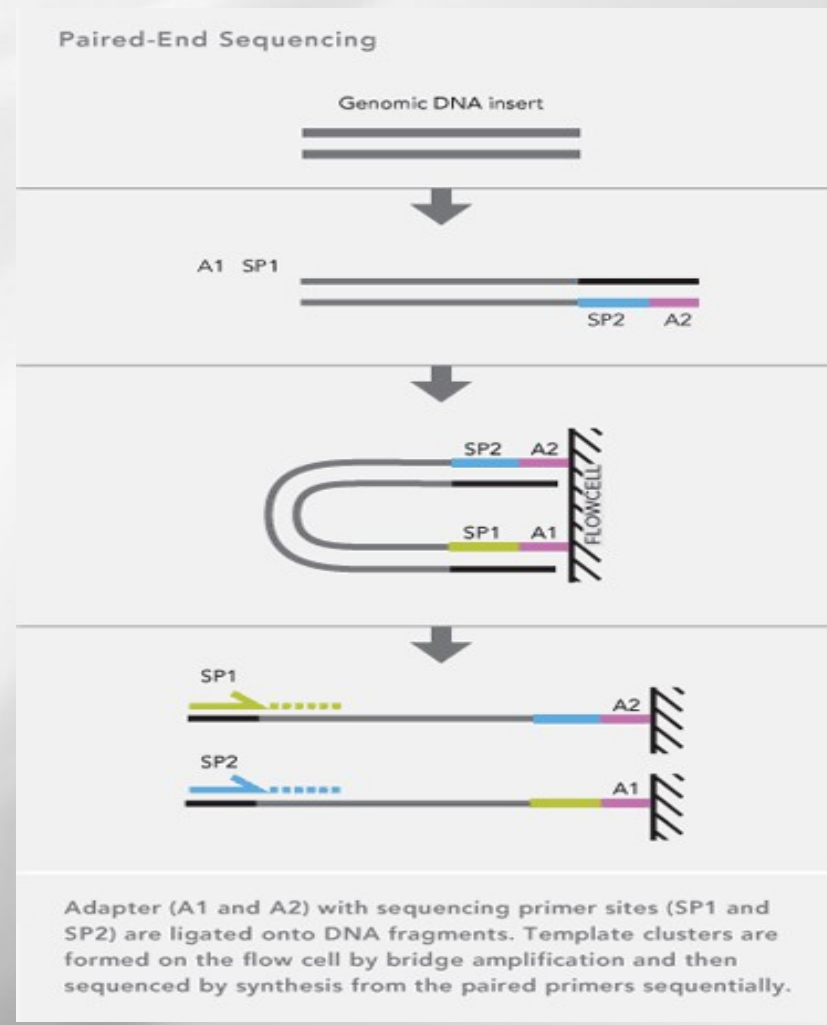
<http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.1001046>

We will discuss strand specific assembly further in this course.

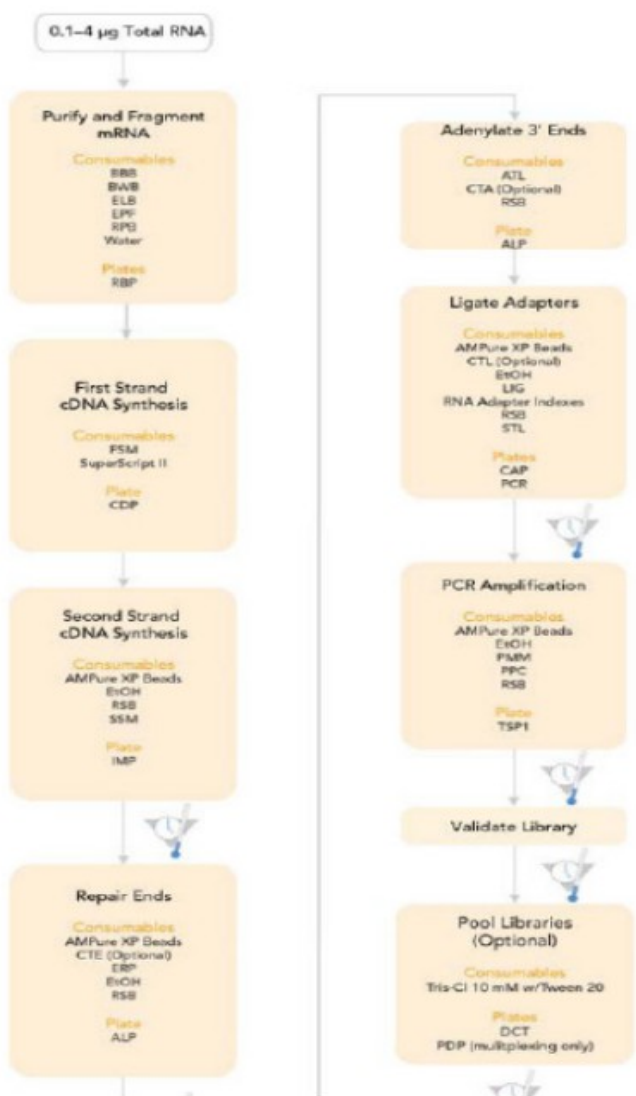
Paired-end sequencing

Modification of the standard single-read DNA library preparation facilitates reading both ends of each fragment.

Mapping improvement.



TruSeq library preparation



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

Library prep takes <2 days

What does an RNA-Seq experiment look like?

Different usages

- Differential expression study
 - Gene and transcript levels
- Gene/transcript annotation
- Phylogenomic analysis (gene evolution between species)
 - Gene level : comparing the longest proteins to produce a phylogenomic tree

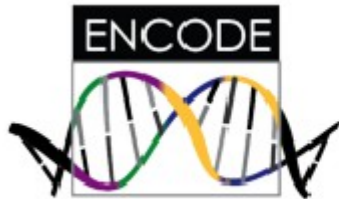
Usual questions on RNA-Seq !

- How many samples for my experimental design?
- 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 ?
- Should I remove duplicated reads from my results?

ENCODE answers (2009)

- RNA-Seq is not a mature technology.
- Experiments should be performed with **two or more biological replicates**, unless there is a compelling reason why this is impractical or wasteful
- A typical R^2 (Pearson) correlation of gene expression (RPKM) between two biological replicates, for RNAs that are detected in both samples using RPKM or read counts, should be between **0.92 to 0.98**. Experiments with biological correlations that fall below 0.9 should be either be repeated or explained.
- Between **30M and 100M reads** per sample depending on the study.

NB. Guidelines for the information to publish with the data.



Encyclopedia of DNA Elements

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

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

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

Conclusions

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

Produced data and Quality control

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. Fields², Naohisa Goto³, Michael L. Heuer⁴ and Peter M. Rice⁵

Table 1. The three described FASTQ variants, with columns giving the description, format name used in OBF projects, range of ASCII characters permitted in the quality string (in decimal notation), ASCII encoding offset, type of quality score encoded and the possible range of scores

Description, OBF name	ASCII characters		Quality score	
	Range	Offset	Type	Range
Sanger standard fastq-sanger	33–126	33	PHRED	0 to 93
Solexa/early Illumina fastq-solexa	59–126	64	Solexa	–5 to 62
Illumina 1.3+ fastq-illumina	64–126	64	PHRED	0 to 62

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

$$Q_{\text{Solexa}} = -10 \times \log_{10}\left(\frac{P_e}{1 - P_e}\right)$$

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

How to check reads?

Is the run OK ?

- Expected quantity :
 - number of sequences (expected sequencers throughput)
 - number of nucleotides (read length and total amount)
- fragments sizes,
- Expected quality (content) : presence of Ns ? If present, are they randomly distributed ?
- Every read should be picked up randomly among transcripts
- It implies no over-representation sequences (could be rRNA or adapter)
- Random selection of the nucleotides and the GC%

How do I clean my reads?

Different type of elements to clean :

- Contamination
- Unknown nucleotides
- Adapters
- Low quality
- PolyA tails

Cleaning can correspond to read removal or read clipping.

If you use paired-ends keep in mind that the assemblers usually check pairing.

FastQC

- FastQC provides a simple way to do some quality control checks on raw sequence data.
- Keep in mind that FastQC quality thresholds are adapted for DNA sequencing.

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later .
Initial Contact	Simon Andrews












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

Basic statistics with fastqc



Basic Statistics

Measure	Value
Filename	SRR334221.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	27006399
Filtered Sequences	0
Sequence length	180
%GC	45

-  [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](#)

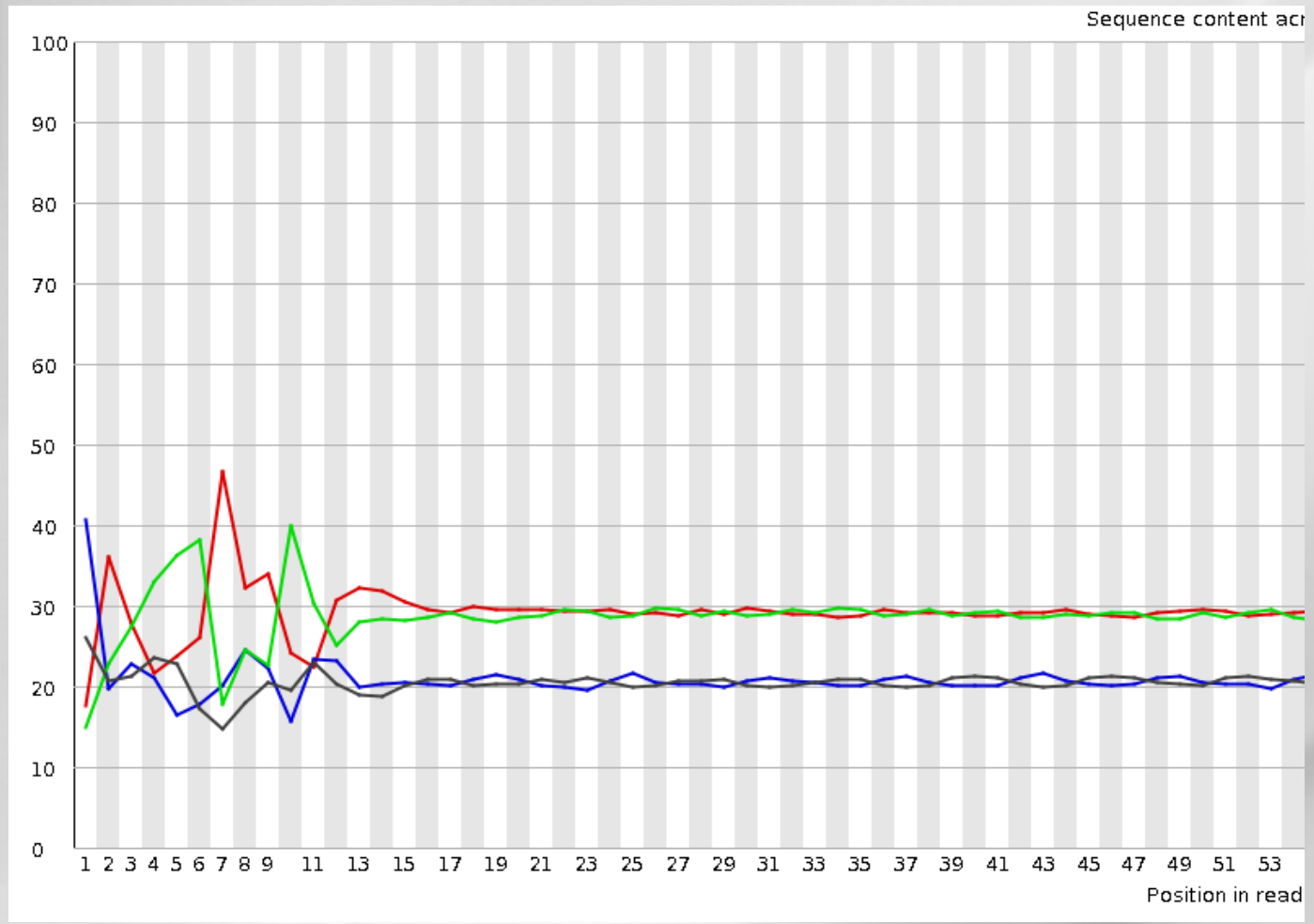
<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Third exercise

In two groups :

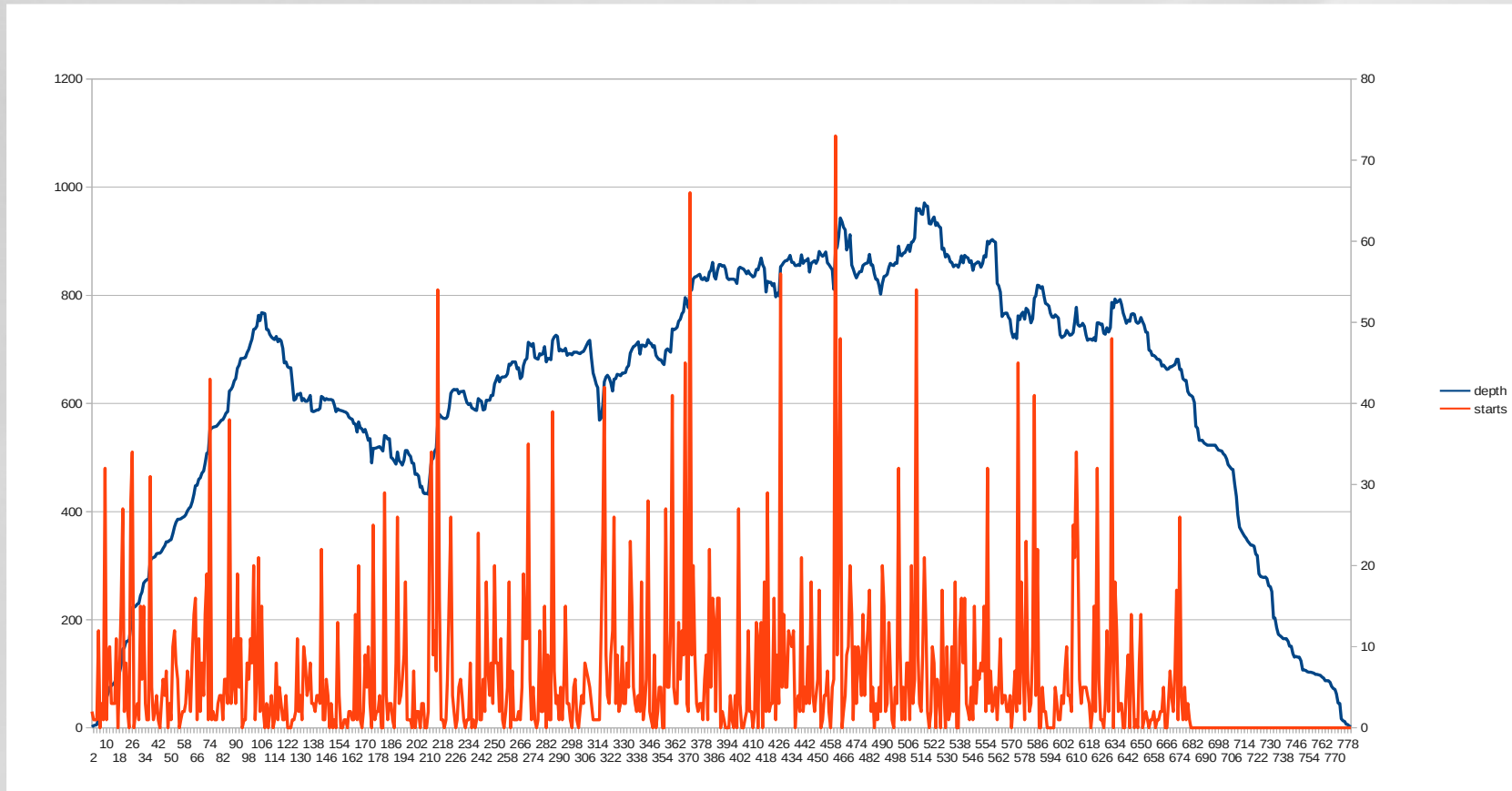
- Explain the graphics which have been given to you to the other group.
- Find the remarkable elements and explain where they come from.

Hexamer random priming bias



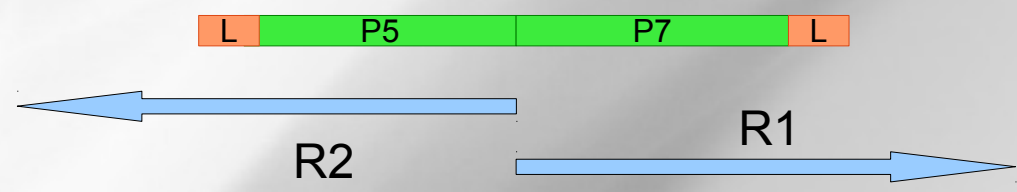
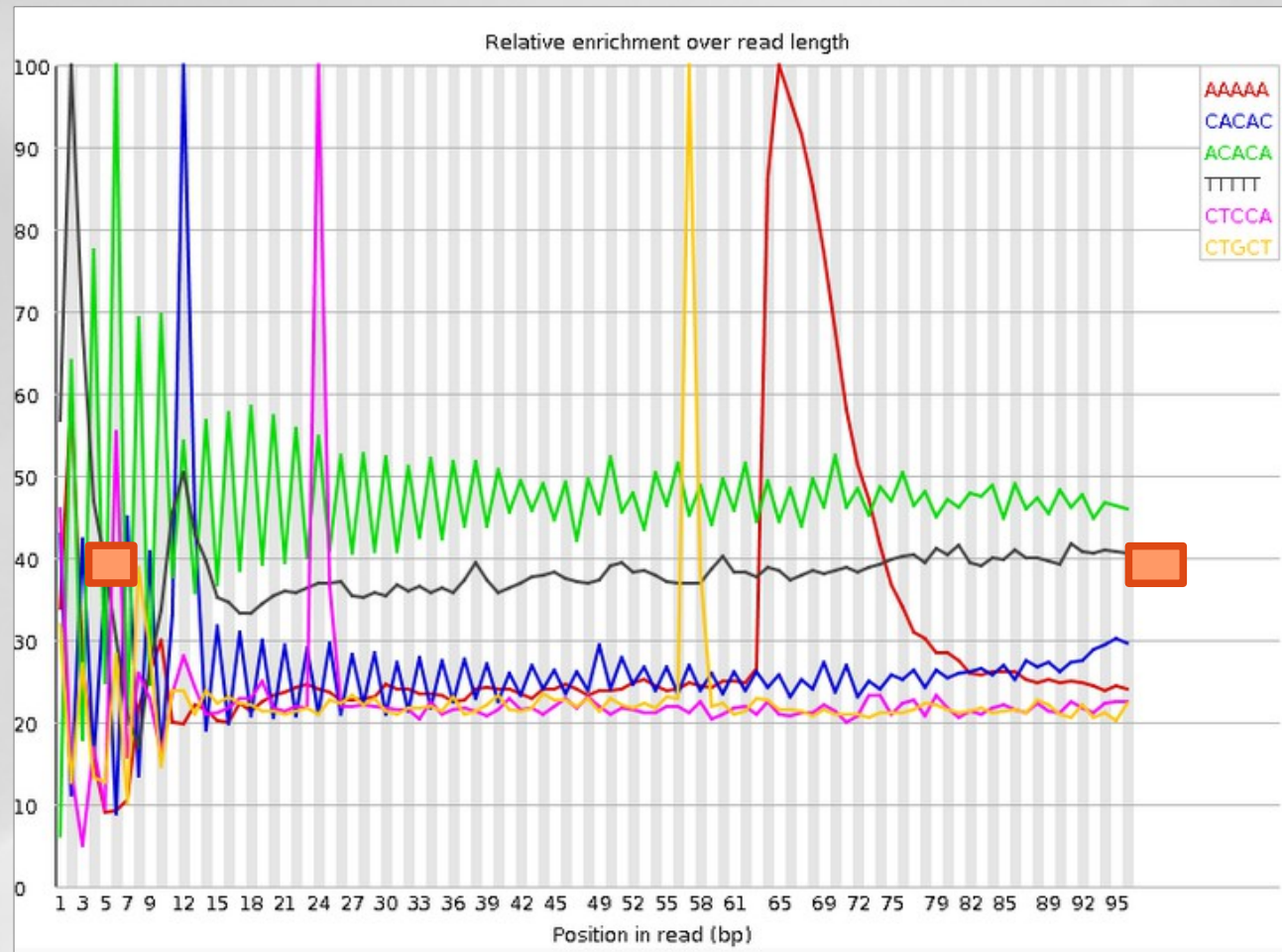
- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✗ [Per base sequence content](#)
- ✗ [Per base GC content](#)
- ✓ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✗ [Sequence Duplication Levels](#)
- !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

Hexamer random effect

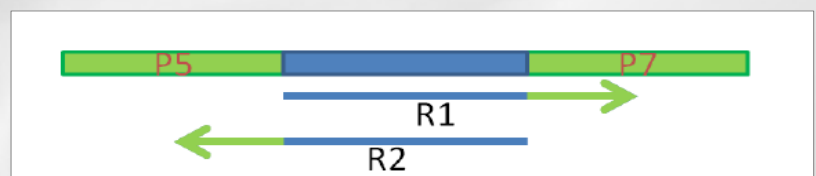
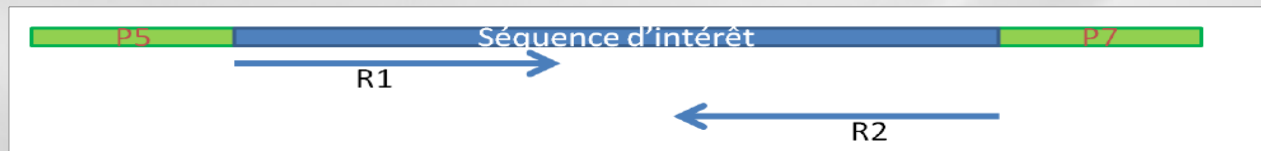


- Orange = reads start sites
- Blue = coverage

Reads with no inserts



How to check read pairs?




- Depending on the fragments size the reads will overlap or not
- If the reads are overlapping then the fragments size histogram can be checked
- Fragment sizes (from library protocol) and reads length may lead to sequence adapter


FLASH

- FLASH (Fast Length Adjustment of SHort reads) is a very fast and accurate software tool to merge paired-end reads.
- FLASH is designed to merge pairs of reads when the original DNA fragments are shorter than twice the length of reads.
- The resulting longer reads can significantly improve genome assemblies. They can also improve transcriptome assembly when FLASH is used to merge RNA-seq data.

FLASH: fast length adjustment of short reads to improve genome assemblies

Tanja Magoč^{*} and **Steven L. Salzberg**

 Author Affiliations

 * To whom correspondence should be addressed.

Received June 14, 2011.
Revision received August 25, 2011.
Accepted August 31, 2011.

Algorithm

Flash processes each read pair separately and searches for the correct overlap between the paired-end reads. When the correct overlap is found, the two reads are merged, producing an extended read that matches the length of the original DNA fragment from which the paired-end reads were generated.

It uses ungapped alignments only.

The overlap is tested one position after the other while the overlap is longer than *min-olap*.

- *Calculate length and score for each position*
 - * *If the score is smaller than the best one keep it*
 - * *If the score is equal : calculate the average quality of the mismatches*
 - *If it is lower keep it*
- *If the score of the best overlap is over the mismatch threshold then no good overlap is reported.*

Parameters

```
-m, --min-overlap=NUM    The minimum required overlap length between two
-M, --max-overlap=NUM    Maximum overlap length expected in approximately
-x, --max-mismatch-density=NUM
-p, --phred-offset=OFFSET
-r, --read-len=LEN
-f, --fragment-len=LEN
-s, --fragment-len-stddev=LEN
--interleaved-input      Instead of requiring files MATES_1.FASTQ and
--interleaved-output    Write the uncombined pairs in interleaved format.
-I, --interleaved        Equivalent to specifying both --interleaved-input
-o, --output-prefix=PREFIX
-d, --output-directory=DIR
-c, --to-stdout          Write the combined reads to standard output; do not
-z, --compress           Compress the FASTQ output files directly with zlib.
--compress-prog=PROG    Pipe the output through the compression program
--compress-prog-args=ARGS
--suffix=SUFFIX, --output-suffix=SUFFIX
-t, --threads=NTHREADS  Set the number of worker threads. This is in
-q, --quiet              Do not print informational messages. (Implied with
-h, --help               Display this help and exit.
-v, --version            Display version.
```

Command line

```
[klopp@genotoul RNASeq]$ flash --min-overlap=20 --output-prefix=Prefix ERR029942_1_500000.fastq.gz ERR029942_2_500000.fastq.gz
[FLASH] Starting FLASH v1.2.6
[FLASH] Fast Length Adjustment of SHort reads
[FLASH]
[FLASH] Input files:
[FLASH]   ERR029942_1_500000.fastq.gz
[FLASH]   ERR029942_2_500000.fastq.gz
[FLASH]
[FLASH] Output files:
[FLASH]   ./Prefix.extendedFrag.s.fastq
[FLASH]   ./Prefix.notCombined_1.fastq
[FLASH]   ./Prefix.notCombined_2.fastq
[FLASH]   ./Prefix.hist
[FLASH]   ./Prefix.histogram
[FLASH]
[FLASH] Parameters:
[FLASH]   Min overlap:      20
[FLASH]   Max overlap:     65
[FLASH]   Phred offset:    33
[FLASH]   Combiner threads: 24
[FLASH]   Max mismatch density: 0.250000
[FLASH]   Output format:   text
[FLASH]   Interleaved input: false
[FLASH]   Interleaved output: false
[FLASH]
```

Outputs

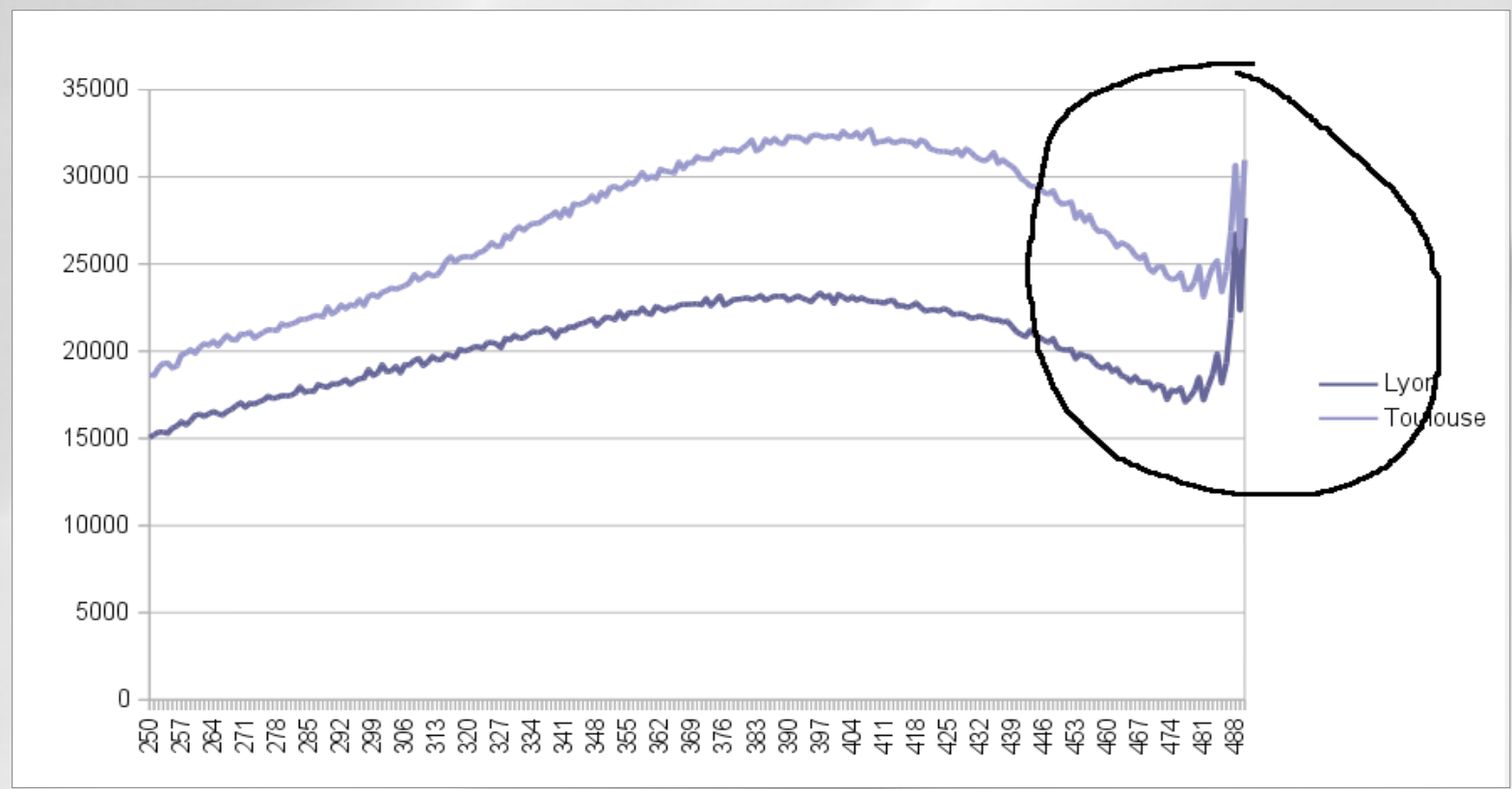
- 5 output files :
 - * Prefix.extendedFrag.fastq
 - * Prefix.notCombined_2.fastq
 - * Prefix.notCombined_1.fastq
 - * Prefix.hist
 - * Prefix.histogram

```
[klopp@genotoul Project_GAN0SEQ.273]$ head Shotgun.hist
250 11001
251 11491
252 11574
253 12097
254 12530
255 13028
256 13175
257 13657
258 13974
259 14658
```

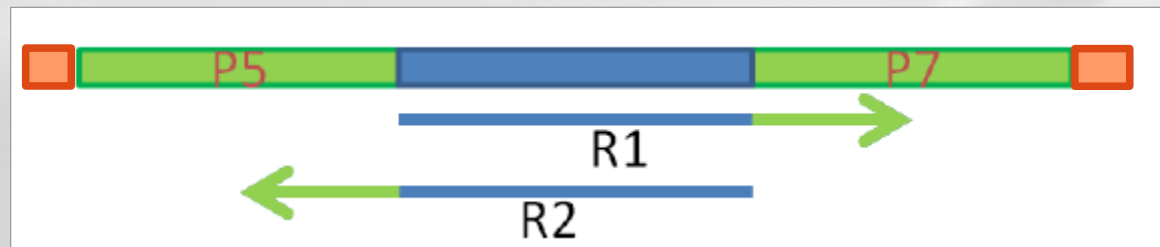
```
[klopp@genotoul Project_GAN0SEQ.273]$ head -20 Shotgun.histogram
250 *****
251 *****
252 *****
253 *****
254 *****
255 *****
256 *****
257 *****
258 *****
259 *****
260 *****
261 *****
262 *****
263 *****
264 *****
265 *****
266 *****
267 *****
268 *****
269 *****
```

Classical problem

The number of overlapping sequences increases when reaching read length.



Explanation and correction



Once you have sequences the adapter you start sequencing the anchor (link to the plate).

The anchor is a polyA of close to 10 nucleotides.

The anchors can be bridged by FLASH.

Change parameters :

- `--min-overlap=20`
- `--max-mismatch-density=0.1`

FastQC and FLASH : exercises

Data location for the exercises :

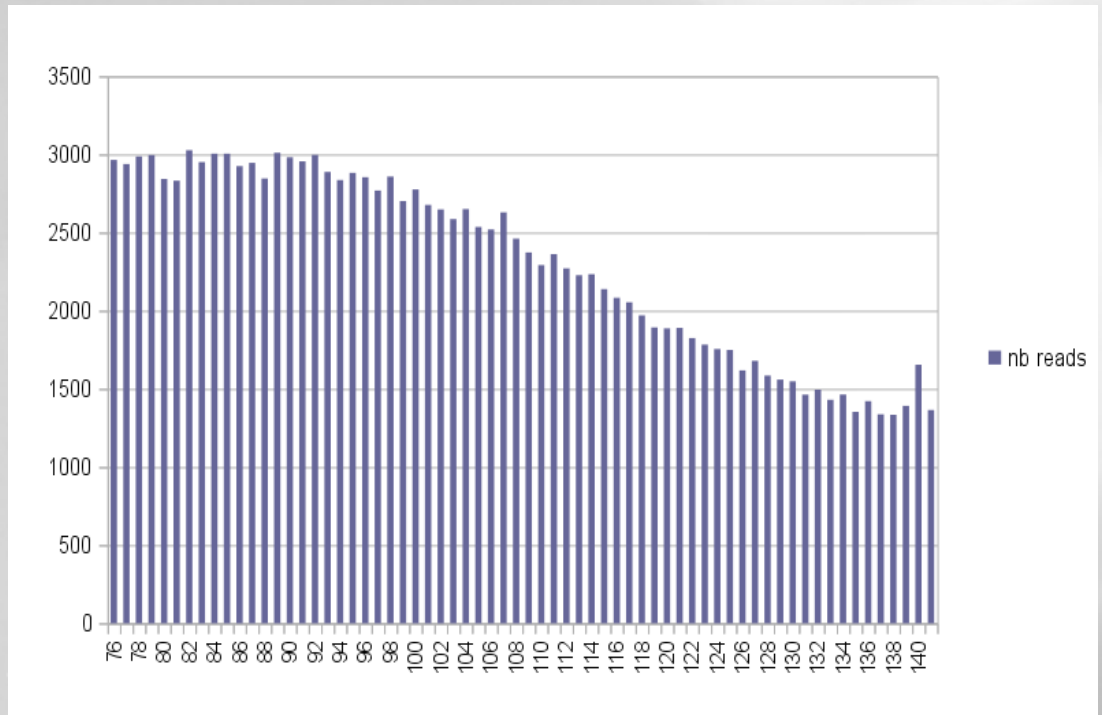
<http://bioinfo.genotoul.fr/index.php?id=137>

Use the fastq files and process them with fastqc.
Note all the remarkable elements found.

Find the average insert size with FLASH for sample :
ERR029942

Flash

flash ERR029942_1_500000.fastq.gz ERR029942_2_500000.fastq.gz



```
[FLASH] Read combination statistics:  
[FLASH] Total reads: 500000  
[FLASH] Combined reads: 153553  
[FLASH] Uncombined reads: 346447  
[FLASH] Percent combined: 30.71%  
[FLASH] Writing histogram files.  
[FLASH] FLASH v1.2.6 complete!  
[FLASH] 10.947 seconds elapsed
```

Read cleaning

- Cutadapt : overview
 - Originally design for remove adapter sequences from reads
 - Features coming :
 - Remove initial or trailing N characters
 - Bam format support
 - Add multi-threading
 - ...

Cutadapt removes adapter sequences from high-throughput sequencing reads

Marcel Martin

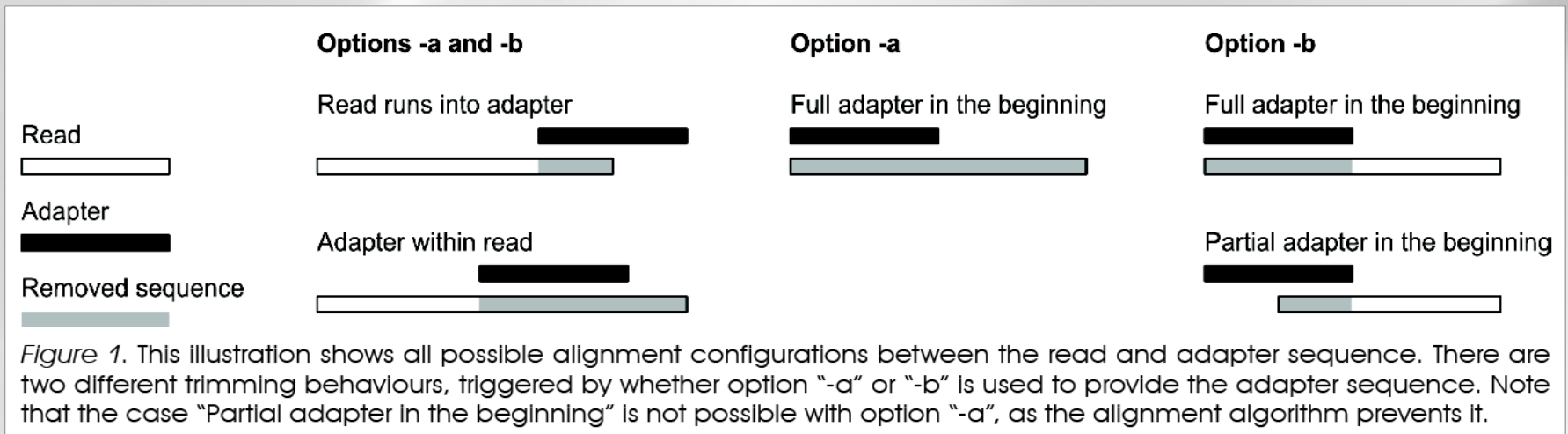
Abstract

When small RNA is sequenced on current sequencing machines, the resulting reads are usually longer than the RNA and therefore contain parts of the 3' adapter. That adapter must be found and removed error-tolerantly from each read before read mapping. Previous solutions are either hard to use or do not offer required features, in particular support for color space data. As an easy to use alternative, we developed the command-line tool cutadapt, which supports 454, Illumina and SOLiD (color space) data, offers two adapter trimming algorithms, and has other useful features.

Cutadapt, including its MIT-licensed source code, is available for download at <http://code.google.com/p/cutadapt/>

M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, North America, 17, May 2011. Available at: <http://journal.embnet.org/index.php/embnetjournal/article/view/>

1. Compute optimal alignment between the read and the adapter sequences. The type of alignment produced is called end-space (or regular semi-global) alignment. It does not penalize initial or trailing gaps.
2. Depending on the parameter used (-a -b -g) cutadapt considers that you know where the adapter is located or not.



Parameters

```

--version          show program's version number and exit
-h, --help        show this help message and exit
-f FORMAT, --format=FORMAT
  trimmed (but see the --times option).
-a ADAPTER, --adapter=ADAPTER
-b ADAPTER, --anywhere=ADAPTER
-g ADAPTER, --front=ADAPTER
-e ERROR_RATE, --error-rate=ERROR_RATE
-n COUNT, --times=COUNT
-O LENGTH, --overlap=LENGTH
--match-read-wildcards
-N, --no-match-adapter-wildcards
--discard-trimmed, --discard
--discard-untrimmed, --trimmed-only
-m LENGTH, --minimum-length=LENGTH
-M LENGTH, --maximum-length=LENGTH
-o FILE, --output=FILE
--info-file=FILE  Write information about each read and its adapter
-r FILE, --rest-file=FILE
--wildcard-file=FILE
--too-short-output=FILE
--untrimmed-output=FILE
-q CUTOFF, --quality-cutoff=CUTOFF
--quality-base=QUALITY_BASE
-x PREFIX, --prefix=PREFIX
-y SUFFIX, --suffix=SUFFIX
--strip-suffix=STRIP_SUFFIX
-c, --colorspace  Colorspace mode: Also trim the color that is adjacent
-d, --double-encode
-t, --trim-primer  When in color space, trim primer base and the first
--strip-f3        For color space: Strip the _F3 suffix of read names
--maq, --bwa      MAQ- and BWA-compatible color space output. This
--length-tag=TAG  Search for TAG followed by a decimal number in the
-z, --zero-cap    Change negative quality values to zero (workaround to

```

Adapter parameters

```
-a ADAPTER, --adapter=ADAPTER
    Sequence of an adapter that was ligated to the 3' end.
    The adapter itself and anything that follows is
    trimmed.
-b ADAPTER, --anywhere=ADAPTER
    Sequence of an adapter that was ligated to the 5' or
    3' end. If the adapter is found within the read or
    overlapping the 3' end of the read, the behavior is
    the same as for the -a option. If the adapter overlaps
    the 5' end (beginning of the read), the initial
    portion of the read matching the adapter is trimmed,
    but anything that follows is kept.
-g ADAPTER, --front=ADAPTER
    Sequence of an adapter that was ligated to the 5' end.
    If the adapter sequence starts with the character '^',
    the adapter is 'anchored'. An anchored adapter must
    appear in its entirety at the 5' end of the read (it
    is a prefix of the read). A non-anchored adapter may
    appear partially at the 5' end, or it may occur within
    the read. If it is found within a read, the sequence
    preceding the adapter is also trimmed. In all cases,
    the adapter itself is trimmed.
```

Why does the "-g" option delete adapters even if they occur at the end or within the read?

The only difference between the "-a" and "-g" options is that "-g" finds the adapter anywhere within the read and removes everything **before** it. If you expect the read to begin with the adapter, then add the character "^" before the adapter sequence on the command line. For example:

```
cutadapt -g ^ADAPTER input.fasta > output.fasta
```

- Cutadapt : command line

```
cutadapt -a ACACTCTTTCCCTACACGACGCTCTTCCGATCT \  
-a ACACTCTTTCCCTACACGACGCTCTTCCGATCT \  
--info-file=FDm1_ATCACG_L008_R1.cutadapt.info \  
-o FDm1_ATCACG_L008_R1.cutadapt.fastq \  
FDm1_ATCACG_L008_R1.fastq.gz
```

Cutadapt reports

```

=== Adapter 1 ===

Adapter 'GCTAGCTAGCATCG', length 14, was trimmed 391411 times.

No. of allowed errors:
0-9 bp: 0; 10-14 bp: 1

Lengths of removed sequences
length  count  expected  max. errors
3       329495  421975.0  0
4       49868  105493.7  0
5       8902   26373.4  0
6       1516   6593.4   0
7       725    1648.3   0
8       153    412.1    0
9       250    103.0    0
10      326    25.8     1
11      117    6.4      1
12      30     1.6      1
13      8      0.4      1
14      1      0.1      1
16      1      0.1      1
21      1      0.1      1
32      1      0.1      1
34      1      0.1      1
36      1      0.1      1
40      1      0.1      1
43      2      0.1      1
47      1      0.1      1
62      1      0.1      1
66      1      0.1      1
70      1      0.1      1
75      2      0.1      1
81      2      0.1      1
82      2      0.1      1
83      1      0.1      1
84      1      0.1      1
    
```

Global

```

Command line parameters: -a GCTAGCTAGCATCG SRR334221_1.fq
Maximum error rate: 10.00%
No. of adapters: 1
Processed reads:      27006399
Processed bases:     2430575910 bp (2430.6 Mbp)
Trimmed reads:       391411 (1.4%)
Trimmed bases:       1256290 bp (1.3 Mbp) (0.05% of total)
Too short reads:     0 (0.0% of processed reads)
Too long reads:      0 (0.0% of processed reads)
Total time:          789.93 s
Time per read:       0.03 ms
    
```


Cutadapt : cleaning pairs

Paired-end adapter trimming

Cutadapt supports paired-end trimming, but currently two passes over the data are required.

Assume the input is in `reads.1.fastq` and `reads.2.fastq` and that `ADAPTER_FWD` should be trimmed from the forward reads (first file) and `ADAPTER_REV` from the second reverse reads (second file). There are two cases.

If you do not use any of the options that discard reads, such as `--discard`, `--minimum-length` or `--maximum-length`, then run cutadapt on each file separately:

```
cutadapt -a ADAPTER_FWD -o trimmed.1.fastq reads1.fastq
cutadapt -a ADAPTER_REV -o trimmed.2.fastq reads2.fastq
```

If you use one of the read-discarding options, then the `--paired-output` option is needed to keep the two files synchronized. First trim the forward read, writing output to temporary files:

```
cutadapt -a ADAPTER_FWD --minimum-length 20 --paired-output tmp.2.fastq -o tmp.1.fastq reads.1.fastq
```


Then trim the reverse read, using the temporary files as input:

```
cutadapt -a ADAPTER_REV --minimum-length 20 --paired-output trimmed.1.fastq -o trimmed.2.fastq tmp.2.f
```

Finally, remove the temporary files:

```
rm tmp.1.fastq tmp.2.fastq
```

trim_galore



Babraham Bioinformatics

[About](#) | [People](#) | [Services](#) | [Projects](#) | [Training](#) | [Publications](#)

Trim Galore!

Function	A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisulfite-Seq) libraries.
Language	Perl
Requirements	A functional version of Cutadapt and optionally FastQC are required.
Code Maturity	Stable.
Code Released	Yes, under GNU GPL v3 or later .
Initial Contact	Felix Krueger

[Download Now](#)

http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Running trim_galore

Command line :

```
mkdir ERR145651_trim_galore

trim_galore -a
AATGATACGGCGACCACCGAGATCTACTCTTTCCCTACACGACGCTCTTCCGATC
T -o ERR145651_trim_galore --paired
ERR145651_chr3_star_R1.fastq.gz
ERR145651_chr3_star_R2.fastq.gz
```

```
=== Adapter 1 ===

Adapter 'AATGATACGGCGACCACCGAGATCTACTCTTTCCCTACACGACGCTCTTCCGATCT', length 58, was trimmed 513451 times.

No. of allowed errors:
0-9 bp: 0; 10-19 bp: 1; 20-29 bp: 2; 30-39 bp: 3; 40-49 bp: 4; 50-58 bp: 5

Overview of removed sequences
length count expect max.err error counts
1 311757 557769.5 0 311757
2 171443 139442.4 0 171443
3 18617 34860.6 0 18617
4 7197 8715.1 0 7197
5 3558 2178.8 0 3558
6 814 544.7 0 814
7 28 136.2 0 28
8 12 34.0 0 12
9 6 8.5 0 3 3
10 16 2.1 1 0 16
11 3 0.5 1 0 3
```

```
1879 Nov 24 10:39 ERR145651_chr3_star_R1.fastq.gz_trimming_report.txt
2078 Nov 24 10:41 ERR145651_chr3_star_R2.fastq.gz_trimming_report.txt
173245101 Nov 24 10:41 ERR145651_chr3_star_R2_val_2.fq.gz
170582556 Nov 24 10:41 ERR145651_chr3_star_R1_val_1.fq.gz
```

Other software pieces

Trimmomatic: A flexible read trimming tool for Illumina NGS data

Citation

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B. RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W622-7.

<http://www.usadellab.org/cms/?page=trimmomatic>

sickle - A windowed adaptive trimming tool for FASTQ files using quality

About

Most modern sequencing technologies produce reads that have deteriorating quality towards the 3'-end and some towards the 5'-end as well. Incorrectly called bases in both regions negatively impact assemblies, mapping, and downstream bioinformatics analyses.

<https://github.com/najoshi/sickle>

Removing Ns in reads

- Assemblers (de Bruijn) discard reads containing N (even 1 N)
- Different options :
 - Removing reads with Ns (in case of BMS this can remove a lot of reads)
 - Removing the part of the reads with the Ns

NNNGTCAGC>NNNNGCTAGCTAGCTGCATCGATCGATNNN
= **GCTAGCTAGCTGCATCGATCGAT**

In house script : `fastq_longest_subseq_without_Ns.py`
Able to keep corresponding pairs

Read clipping : Fastx toolkit

FASTX-Toolkit

FASTQ/A short-reads pre-processing tools

Available Tools

- FASTQ-to-FASTA converter
Convert FASTQ files to FASTA files.
- FASTQ Information
Chart Quality Statistics and Nucleotide Distribution
- FASTQ/A Collapser
Collapsing identical sequences in a FASTQ/A file into a single sequence (while maintaining reads counts)
- FASTQ/A Trimmer
Shortening reads in a FASTQ or FASTQ files (removing barcodes or noise).
- FASTQ/A Renamer
Renames the sequence identifiers in FASTQ/A file.
- FASTQ/A Clipper
Removing sequencing adapters / linkers
- FASTQ/A Reverse-Complement
Producing the Reverse-complement of each sequence in a FASTQ/FASTA file.
- FASTQ/A Barcode splitter
Splitting a FASTQ/FASTA files containning multiple samples
- FASTA Formatter
changes the width of sequences line in a FASTA file
- FASTA Nucleotide Changer
Convets FASTA sequences from/to RNA/DNA
- FASTQ Quality Filter
Filters sequences based on quality
- FASTQ Quality Trimmer
Trims (cuts) sequences based on quality
- FASTQ Masker
Masks nucleotides with 'N' (or other character) based on quality

Contamination search

- BWA (BLAST on a subset)
 - Comparing rates for different samples
 - Homology search in common contaminant organism databases and large scale database
 - nr (actually a subset of nr database : blast would be too slow)
 - E. coli
 - Fungi
 - Yeast
 - Phage
 - ...

NG6 contamination results

Contamination Results [Parameters](#) [Downloads](#)

10 records per page Search:

Samples (20)	ecoli536	phi	yeast
Heart	21	703	11 190
Heart	54	959	18 713
Intestine	54	267	758
Intestine	34	221	679
Liver	3 768	177 376	495
Liver	6 310	324 717	824
Muscle	15	484	1 739
Muscle	27	741	3 011
Testis	48	1 296	52 907
Testis	30	935	46 761

Showing 11 to 20 of 20 entries
[← Previous](#)
[1](#)
[2](#)
[Next →](#)

Sequencing error correction

Error occur during the sequencing process.

These errors impact the assembly process (less identity, larger graphs,...)

Removing these errors before assembly :

- Limits the errors in the contigs
- Speeds the assembly

Many different software packages. One adapted to RNA-Seq reads = Seecer.

The challenge is to separate errors from rare polymorphisms in an efficient manner.

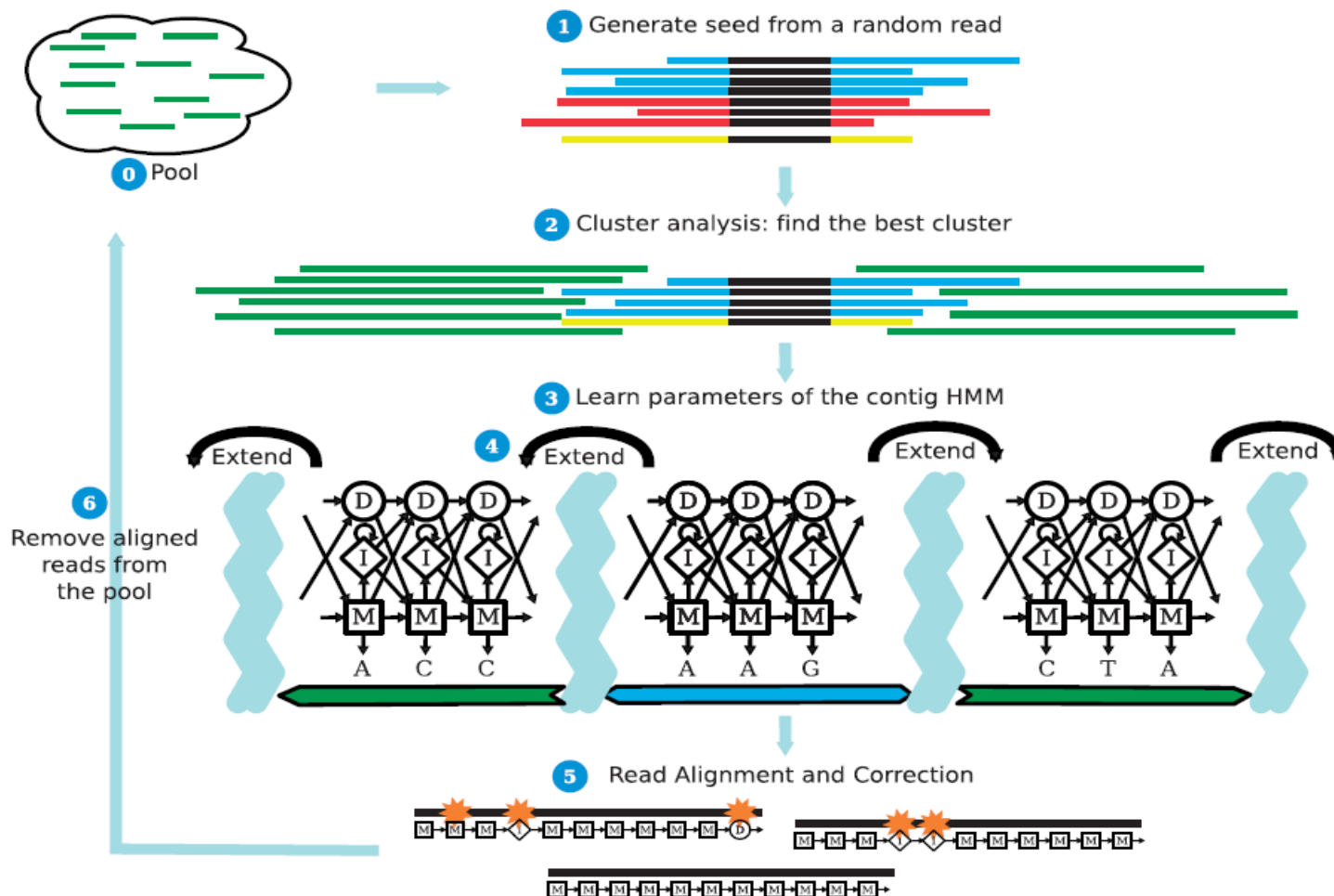
Probabilistic error correction for RNA sequencing

Hai-Son Le¹, Marcel H. Schulz², Brenna M. McCauley³, Veronica F. Hinman³ and Ziv Bar-Joseph^{1,2,*}

¹Machine Learning Department, Carnegie Mellon University, 5000 Forbes Avenue Pittsburgh, PA 15217, USA, ²Lane Center for Computational Biology, Carnegie Mellon University, 5000 Forbes Avenue Pittsburgh, PA 15217, USA and ³Department of Biological Sciences, Carnegie Mellon University, 5000 Forbes Avenue Pittsburgh, PA 15217, USA

Received December 3, 2012; Revised March 4, 2013; Accepted March 7, 2013

Nucleic Acids Research, 2013 3



Impact of error correction

<https://peerj.com/articles/113/>

PeerJ

Improving transcriptome assembly through error correction of high-throughput sequence reads

Matthew D. MacManes¹ and Michael B. Eisen^{1,2,3}

¹ California Institute for Quantitative Biosciences, University of California, Berkeley, CA, USA

² Howard Hughes Medical Institute, USA

³ Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA

Table 1 Number of raw sequencing reads, sequencing reads corrected, nucleotides (nt) corrected, and approximate runtime for each of the datasets. Note that neither ALLPATHS nor SGA provides information regarding the number of reads affected by the correction process.

Simulated dataset	Total reads	Num reads corr	Num nt corr	Runtime
Raw reads	30M PE	n/a	n/a	n/a
ALLPATHSLG Corr.	30M PE	?	139,592,317	~8 h
SGA Corr.	30M PE	?	19,826,919	~38 min
REPTILE Corr.	30M PE	2,047,088	7,782,594	~3 h
SEECER Corr.	30M PE	8,782,350	14,033,709	~5 h

Seecer results

SEECER, is the only dedicated error-correction software package dedicated to RNAseq reads. Though SEECER is expected to handle RNAseq datasets better than the other correction programs, its results were disappointing. More than 14 million nucleotides were changed, affecting approximately 8.8M sequencing reads. Upon assembly 54,574 nucleotide errors remained which is equivalent to the number of errors contained in the assembly of uncorrected reads.

Interesting, SEECER, the only error correction method designed for RNAseq reads, performed relatively poorly. In simulated reads, SEECER slightly increased the number of errors in the assembly, though with applied to empirically derived reads, results were more favorable, decreasing error by $\sim 3\%$. Though the effects of coverage on correction efficiency were not explored in the manuscript describing SEECER (*Le et al., 2013*), their empirical dataset contained nearly 90 million sequencing reads, a size $3\times$ larger than the dataset we analyze here. Future work investigating the effects of coverage on error correction is necessary.

Impact on the assembled contigs

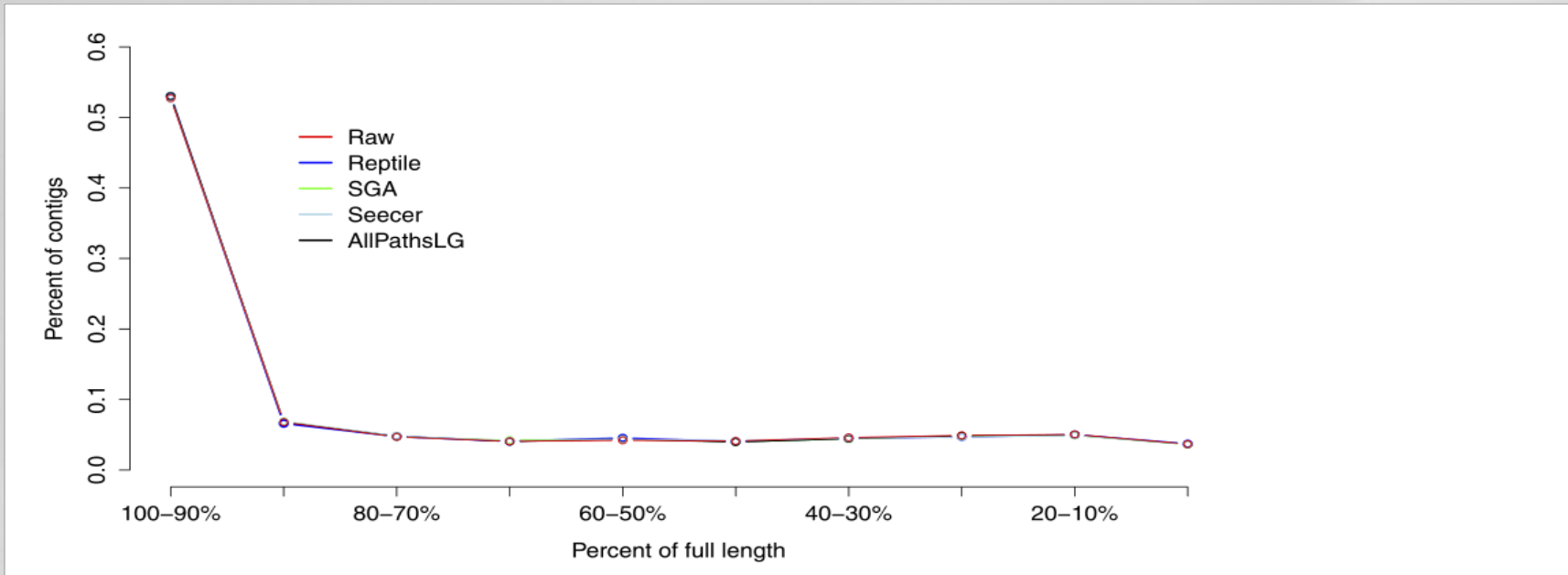


Figure 3 Assembly contiguity did not vary significantly between assemblies of reads using the different error correction methods. Each error correction methods, as well as assembly of raw reads, produced an assembly that is dominated by full length (both start and stop codon present) or nearly full length assembled transcripts.

Though sequence read error correction failed to have a large effect on global assembly metrics, there was substantial improvement at the nucleotide level.

Error rates

- Empirical reads (simulated data) :
 - 21,406 contigs
 - 14.7k nucleotide mismatches
 - 0.68 mismatches per contig (SD = 3.60 max = 197)
- Reptile :
 - 21580 contigs
 - 13k nucleotide mismatches
 - 10% error decrease
- SGA :
 - 9% error correction

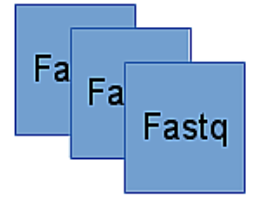
Hands-on

1/ Clean all five read sets with and the provided adapter files using cutadapt

NB. The adapter file is in the same directory.

Do you find any adapter?

Sum-up



Quality control : fastqc

Insert size control : FLASH

Adapter removal : cutadapt

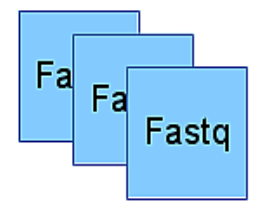
PolyA tail removal : cutadapt

Contamination search and removal : bwa + samtools

N blocks removal : script

Low quality removal : FastX toolkit

Sequencing error corrections : reptile



See you tomorrow!

Questions?