Application of NGS to Transcriptomics

Fernando García Alcalde

Department of Bioinformatics and Genomics
Centro de Investigación Príncipe Felipe, Valencia, Spain

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

International Course on MASSIVE Data Analysis
Valencia, March, 2011

Acknolowegdments

Sonia Tarazona

Pablo Escobar

José Carbonell

Ana Conesa

Outline

1. Introduction

2. RNA-Seq Data Mapping

3. RNA-Seq Data Analysis

Outline

1. Introduction
   - Basic Biology
   - From Microarrays to RNA-Seq
   - RNA-Seq
2. RNA-Seq Data Mapping
3. RNA-Seq Data Analysis

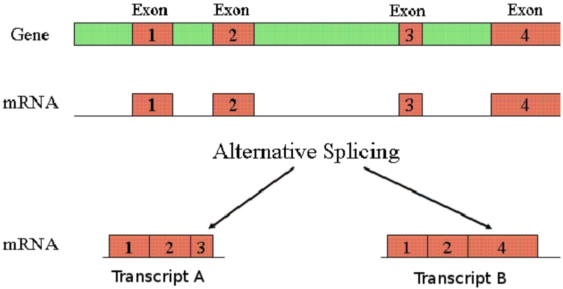
Inicial Concepts

Definitions

- Gene: a hereditary DNA sequence that determines a particular characteristic in an organism.
- Exon: a region of a gene that codes information for protein synthesis that is transcribed to mRNA.
- Intron: a region of a gene which is not translated into protein and is removed before translation of mRNA.
- Splicing: a process in which the introns are removed and exons are joined to be translated into a single transcript.

Alternative Splicing

Alternative splicing: process in which exons can be spliced out in different combinations named transcripts to generate the mature RNA molecule.



Microarrays

Features

- Allow measuring the abundance of thousands of DNA and RNA sequences simultaneously in different cell samples.
- Make use of the hibridatory properties of the nucleic acids to observe their abundance.
- Probes: Short (known) DNA sequences fixed in the array.
- Targets: DNA sample that one wants to monitorize.
- The abundance of each sequence is a function of the fluorescence level recovered after the hybridization process.

Hight-throughput sequencing

Brief Summary

- Improvements in the efficiency, quality and cost of genemo-wide sequencing have made biologist to abandon microarrays in favor of so-called next-generation sequencing (NGS)
- Plataforms: SOLiD, Illumina, Roche's 454, HeliScope
- Allow to obtain *digital* measures for the secuence abundances (read counts)

Pros / Cons

Microarrays

Pros

- Price
- Well-established protocols
- Wide computational analysis tools accessible.

Cons

- Limited to known genomes/transcriptomes.
- Limited sensitivity
- Problems in the hybridization (e.g. cross-hybridization, affinity effects, ...)
- Specific designs for each particular problem

Sequencing

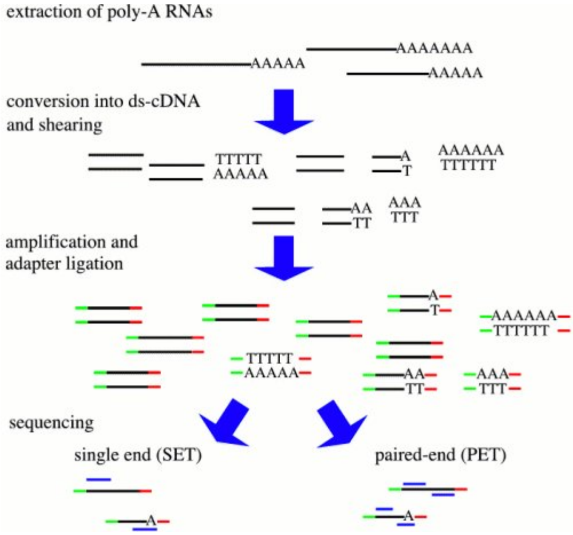
Pros

- Potential for the discovery of novel / not annotated regions
- Discrete measure of abundance (read counts)
- Improved quality and versatility of the data

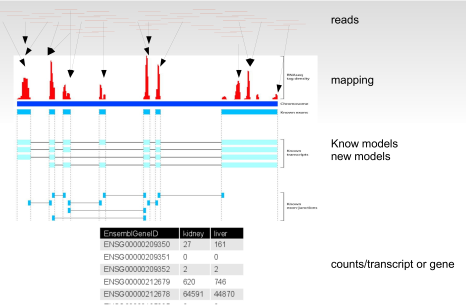
Cons

- Dependence in the sequencing depth
- Price
- Complex data processing and analysis
- Lack of a well-defined benchmark

RNA-Seq. General Protocol



RNA-Seq. Schema



reads

mapping

Know models
new models

EnsemblGeneID	kidney	liver
ENSG00000209350	27	161
ENSG00000209351	0	0
ENSG00000209352	2	2
ENSG00000212679	620	746
ENSG00000212678	64591	44870

counts/transcript or gene

General Objectives

- Quantify transcript abundances
- Identify gene transcriptional structure: splicing, 5’ and 3’ sites, etc
- Quantify expression level changes in each transcript

RNA-Seq. Data

Raw Data

Reads from the sequencer (sequences + qualities)

Formats

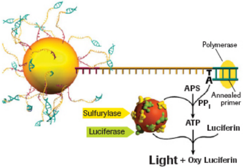
- FASTQ $\Longrightarrow$ nucleotides
- Colorspace $\Longrightarrow$ colors for each change

Basic Features

- Single-end / Parired-end
- Length: 35bp, 50bp, 75bp, 400bp,....
- Strand specificity
- Quality
- Depth $\Longrightarrow$ Tipically 10 millions per *lane* (growing)

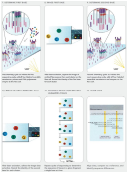
Platforms

Roche



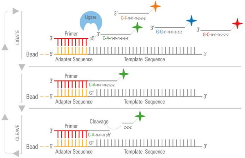
- “Long reads” (400nts)
- Good for *de novo*
- Errors: Poly-n’s

Illumina



- Reads 35-150nts
- Paired-end
- Errores: hexámeros

SOLiD



- Reads 50-100nts
- Strand specific
- Colorspace

Índice

1 Introduction

2 RNA-Seq Data Mapping

- Before we start
- Background
- TopHat

3 RNA-Seq Data Analysis

What do we have?

Sequencer Output

- Obtained sequence (read) → Different techniques and protocols
- Estimated quality → Sequencer calibration

Main Problem

VERY big files → How can we have an idea of what is in them?

Related problems

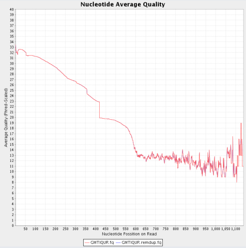
- Detect wrong reads
- What to do with the wrong ones (trimming, removing, ...)
- Take into account specific problems of each platform

Read Quality

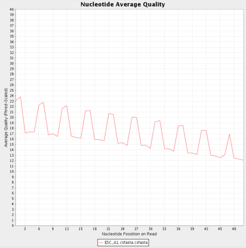
Theory: Same scale ⟹ Comparable results

Reality: Different platforms ⟹ Different behviours

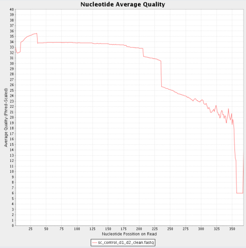
Roche



SOLiD



Illumina



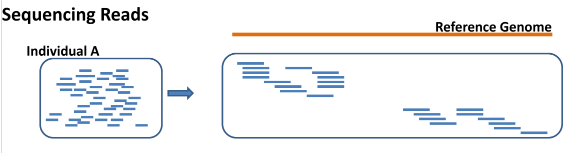
FastQC

- Covered in the previous class
- Software for the sequencing quality control
- Very useful to get an quick idea of the quality of the data and where problems can be expected

Ejemplos

- Datos OK:
  http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/good_sequence_short_fastqc/fastqc_report.html
- Datos with problems:
  http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/bad_sequence_fastqc/fastqc_report.html

RNA-Seq. Mapping



Main Issues

- Number of allowed mismatches
- Number of multi-hits
- Distance between pairs
- Consider exon junctions

Mapping Algorithms

BWA

- Short reads up to 200bp with error $< 5\,\%$
- Do not account for read quality
- Gapped alignment

Bowtie

- Very fast for short reads
- Does not align with gaps
- Use the read quality to evaluate the alignment

Tophat

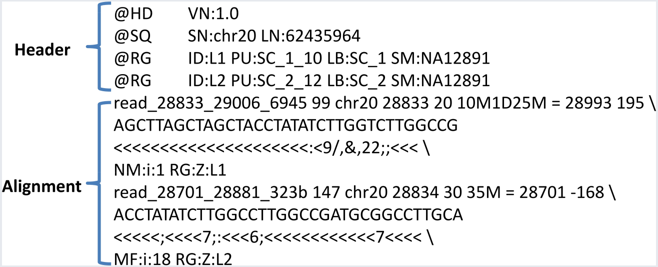
- Improved Bowtie with gap alignment

Other

- ELAND (Illumina software), SOAP, MAQ, etc.

SAM format

SAM file example



Header

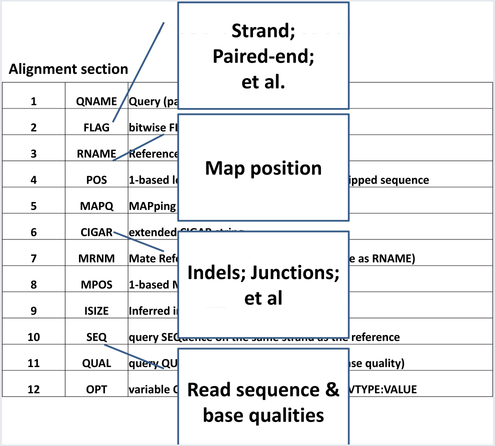
```
@HD     VN:1.0
@SQ     SN:chr20 LN:62435964
@RG     ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG     ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
```

Alignment

```
read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195 \
AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG
<<<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< \
NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M = 28701 -168 \
ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA
<<<<<;<<<<7;:<<<6;<<<<<<<<<<<<<7<<<< \
MF:i:18 RG:Z:L2
```

SAM Format

Information about the alignment



Alignment section

1	QNAME	Query (pa
2	FLAG	bitwise Fl
3	RNAME	Reference
4	POS	1-based l ipped sequence
5	MAPQ	MAPping
6	CIGAR	extended CIGAR string
7	MRNM	Mate Ref e as RNAME)
8	MPOS	1-based M
9	ISIZE	Inferred i
10	SEQ	query SEQuence on the same strand as the reference
11	QUAL	query QU se quality)
12	OPT	variable O VTYPE:VALUE

Strand; Paired-end; et al.

Map position

Indels; Junctions; et al

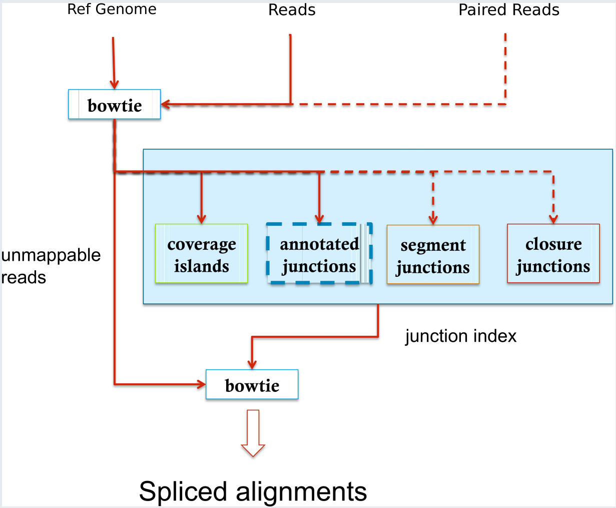
Read sequence & base qualities

Tophat (http://tophat.cbcb.umd.edu/)

Features

- Align the sequences against the genome AND the exon unions (with/without reference)
- Uses Bowtie, an ultrafasr aligner with low memory consumption
- Align segments (25bp by default) of each read, allowing up to 2 mismatches (by default)
- Recent support for colorspace
- It does not consider indels
- Highly configurable
- Continous improvement of the software → good but... Caution! New bugs sometimes

Tophat. Schema



Tophat. Example

Input data → `rawReads1.fastq`, `rawReads2.fastq`

`rawReads1.fastq`:

1000 reads

50 % gene `ARHGAP5` (two exons) and 50 % gene `CMA1` (two exons)

Tophat. Example

Input data → `rawReads1.fastq`, `rawReads2.fastq`

`rawReads1.fastq`:

1000 reads

50 % gene `ARHGAP5` (two exons) and 50 % gene `CMA1` (two exons)

`rawReads2.fastq`:

500 reads

20 % gene `ARHGAP5` (two exons) and 80 % gene `APEX1` (three exons)

Tophat. Example

Input data → `rawReads1.fastq`, `rawReads2.fastq`

`rawReads1.fastq`:

1000 reads

50 % gene `ARHGAP5` (two exons) and 50 % gene `CMA1` (two exons)

`rawReads2.fastq`:

500 reads

20 % gene `ARHGAP5` (two exons) and 80 % gene `APEX1` (three exons)

Reference → `HS_chr14.*`

Homo sapiens, chromosome 14 (pre-indexed)

Tophat. Example

Input data → rawReads1.fastq, rawReads2.fastq

rawReads1.fastq:
1000 reads
50 % gene ARHGAP5 (two exons) and 50 % gene CMA1 (two exons)

rawReads2.fastq:
500 reads
20 % gene ARHGAP5 (two exons) and 80 % gene APEX1 (three exons)

Reference → HS_chr14.*

Homo sapiens, chromosome 14 (pre-indexed)

Command

1. `cd /home/biouser/rnaseq`
2. `mkdir results`
3.
```
tophat -o /home/biouser/rnaseq/results/exp1/ -p 1 /home/biouser/rnaseq/data/HS.chr14
/home/biouser/rnaseq/data/rawReads1.fastq
```
4.
```
tophat -o /home/biouser/rnaseq/results/exp2/ -p 1 /home/biouser/rnaseq/data/HS.chr14
/home/biouser/rnaseq/data/rawReads2.fastq
```

Tophat. Exercices

Alignments

- Examinate and understand the generated SAM files (`accepted_hits.sam`)
- Load the SAM files with IGV and observe the alignment

  Hint: Look for the regions of interest

Junctions

- Observe and understand the generated BED files (`junctions.bed`)

  Hint: `http://genome.ucsc.edu/FAQ/FAQformat.html#format1`

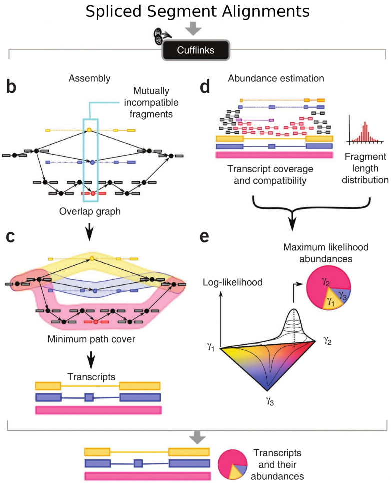
BONUS

- Which are the genomic coordinates of the junctions?
- Can you explain the situation for the CMA1 gene?

Outline

1. Introduction
2. RNA-Seq Data Mapping
3. RNA-Seq Data Analysis
   - Transcript Reconstruction
   - Counting Regions of Interest
   - Final Exercise

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


Spliced Segment Alignments
Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts
d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
γ_1
γ_2
γ_3
Transcripts and their abundances

- Not restricted to a previous annotation
- Accounts for alternative splicing
- Receives a set of mapped reads (SAM/BAM)
- Detects compatible fragments and search for a parsimonous explanation

Cufflinks. Example

Exercise

Assembly the transcripts found in the previous alignment

Cufflinks. Example

Exercise

Assembly the transcripts found in the previous alignment

Commands

From /home/biouser/rnaseq/results/exp1/ and /home/biouser/rnaseq/results/exp2/:

```
cufflinks -p 1 accepted_hits.sam
```

Cufflinks. Example

Exercise

Assembly the transcripts found in the previous alignment

Commands

From /home/biouser/rnaseq/results/exp1/ and /home/biouser/rnaseq/results/exp2/:

```
cufflinks -p 1 accepted_hits.sam
```

Question

Were the transcripts reconstructed as you expected?

Hints:

Observe the transcripts.gtf files
GTF definition: http://genome.ucsc.edu/FAQ/FAQformat.html#format3
Finding the actual transcripts in *ensembl* might help

Comparing Experiments → cuffcompare

Exercise

cuffcompare cufflinks application for comparing results from different experiments

Compare the outputs obtained in the previous experiment

Comparing Experiments → cuffcompare

Exercise

cuffcompare cufflinks application for comparing results from different experiments

Compare the outputs obtained in the previous experiment

Commands

Create /home/biouser/rnaseq/results/compare/ and from there:

```
cuffcompare -r /home/biouser/rnaseq/data/HS.chr14.gtf
   ../exp1/transcripts.gtf ../exp2/transcripts.gtf
```

Comparing Experiments → cuffcompare

Exercise

cuffcompare cufflinks application for comparing results from different experiments

Compare the outputs obtained in the previous experiment

Commands

Create /home/biouser/rnaseq/results/compare/ and from there:

```
cuffcompare -r /home/biouser/rnaseq/data/HS.chr14.gtf
    ../exp1/transcripts.gtf ../exp2/transcripts.gtf
```

Questions

- How many transcripts in total? Why?
  Hint: Observe the file `stdout.combined.gtf`
- Can you identify the specific transcripts of each experiment?
  Hint: Observe the file `stdout.tracking`

Counting Regions of Interest → htseq-count

htseq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html)

htseq-count: Receives an alignment file (`SAM/BAM`) and a list of genomics features (p.ej. `GTF`)
Returns the number of reads that fall within the selected feature

Counting Regions of Interest → htseq-count

htseq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html)

htseq-count: Receives an alignment file (`SAM/BAM`) and a list of genomics features (p.ej. `GTF`)
Returns the number of reads that fall within the selected feature

Exercise

Compute the counts at the gene level for the previous experiments

Counting Regions of Interest → htseq-count

htseq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html)

htseq-count: Receives an alignment file (`SAM/BAM`) and a list of genomics features (p.ej. `GTF`)
Returns the number of reads that fall within the selected feature

Exercise

Compute the counts at the gene level for the previous experiments

Commands

From `/home/biouser/rnaseq/results/exp1/` and `/home/biouser/rnaseq/results/exp2/`:

```
htseq-count -s no -i gene_name accepted_hits.sam
/home/biouser/rnaseq/data/HS.chr14.gtf > counts.txt
```

Counting Regions of Interest → htseq-count

htseq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html)

htseq-count: Receives an alignment file (`SAM/BAM`) and a list of genomics features (p.ej. `GTF`)
Returns the number of reads that fall within the selected feature

Exercise

Compute the counts at the gene level for the previous experiments

Commands

From `/home/biouser/rnaseq/results/exp1/` and `/home/biouser/rnaseq/results/exp2/`:

```
htseq-count -s no -i gene_name accepted_hits.sam
/home/biouser/rnaseq/data/HS.chr14.gtf > counts.txt
```

Questions

- How many reads has each gene in each experiment
- Does it matches with what you expected? Can you explain it? Hint: Think of the effects of the transcript length and the sequencing depth

BONUS

RPKM

Compute the RPKM (Reads per kilobase per million reads) value for a gene in a given experiment and compare your results with those obtained by `cufflinks`

$$RPKM = \frac{\text{number of reads of the region}}{\frac{\text{total reads}}{1000000} \times \frac{\text{region length}}{1000}}$$

Real problem

Objective

Use what you have learnt in a real dataset

Data

Two real sequencing experiments (reduced due to memory issues):

- `brain.fastq`
- `uhr.fastq`

Hints

- Reference genome: hg19.*
- Annotation: hg19.gtf
- Examine the problem and follow the previously used pipeline
- Ask if you find yourself lost