Analysis of RNA-seq data

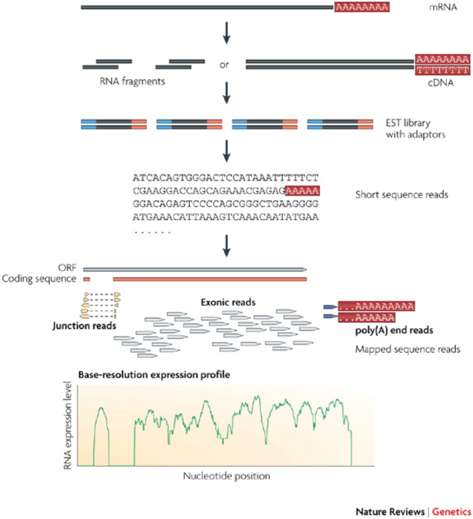
FOS 2017

Maarten van Iterson

Leiden University Medical Center
Department of Molecular Epidemiology

June 1, 2017

A typical RNA-seq experiment



Wang *et al.* Nature Reviews Genetics 10, 57–63(2009)

Experimental Design

- clear simple research question:
    - Good: Which genes are differentially expressed between disease and control, treated and untreated samples
    - Bad: Which genes are differentially expressed between treated and untreated samples with different dosages and different time points after treatment
- randomization
    - not all the controls on one day with one batch of chemicals and the cases on the other day with another batch of chemicals
- think ahead
    - sequencer (mostly Illumina nowadays)
    - aligner
    - annotation
    - statistical analysis

Analysis of RNA-seq data

- Read mapping
  - Input: FASTQ file generated by the sequencer
  - Output: BAM file

- Summarization
  - Input: BAM file
  - Output: count table

- Quality control

- Normalization
  - Input: count table
  - Output: scale factors

- Differential Expression Analysis
  - Input: count table plus scale factors
  - Output: list of differentially expressed genes

FASTQ file

file format for sequences plus quality scores
quality score indicates the probability that a given base is called incorrectly by the sequencer

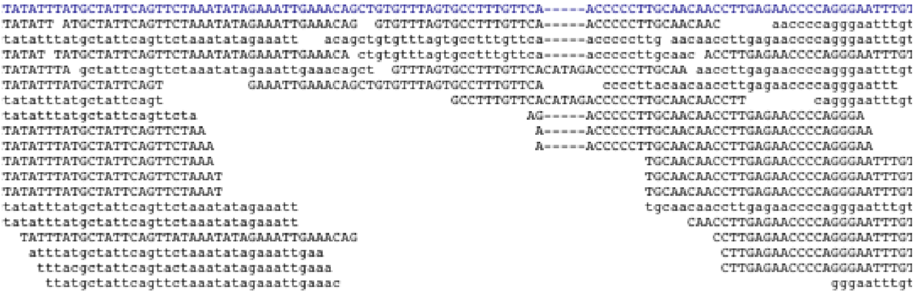
```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#"""""""""""7F@71,";C?,B;?6B;:EA1EA
1EA59B:?:#9EA0D@2EA5:>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

@title and optional description
sequence line(s)
+optional repeat of title line

quality line(s)

Cock *et al.* Nucleic Acids Research,38(6), 1767–1771(2010)

Read mapping



```
TATATTTATGCTATTCAGTTCTAAATATAGAAATTGAAACAGCTGTGTTTAGTGCCTTTGTTCA-----ACCCCCTTGCAACAACCTTGAGAACCCCAGGGAATTTGT
TATATT ATGCTATTCAGTTCTAAATATAGAAATTGAAACAG  GTGTTTAGTGCCTTTGTTCA-----ACCCCCTTGCAACAAC      aaccccagggaatttgt
tatatttatgctattcagttctaaatatagaaatt   acagctgtgtttagtgcctttgttca-----acccccttg aacaaccttgagaaccccagggaatttgt
TATAT TATGCTATTCAGTTCTAAATATAGAAATTGAAACA ctgtgtttagtgcctttgttca-----acccccttgcaac ACCTTGAGAACCCCAGGGAATTTGT
TATATTTA gctattcagttctaaatatagaaattgaaacagct  GTTTAGTGCCTTTGTTCACATAGACCCCCTTGCAA aaccttgagaaccccagggaatttgt
TATATTTATGCTATTCAGT          GAAATTGAAACAGCTGTGTTTAGTGCCTTTGTTCA       ccccttacaacaaccttgagaaccccagggaattt
tatatttatgctattcagt                                  GCCTTTGTTCACATAGACCCCCTTGCAACAACCTT        cagggaatttgt
tatatttatgctattcagttcta                                       AG-----ACCCCCTTGCAACAACCTTGAGAACCCCAGGGA
TATATTTATGCTATTCAGTTCTAA                                       A-----ACCCCCTTGCAACAACCTTGAGAACCCCAGGGAA
TATATTTATGCTATTCAGTTCTAAA                                      A-----ACCCCCTTGCAACAACCTTGAGAACCCCAGGGAA
TATATTTATGCTATTCAGTTCTAAA                                                   TGCAACAACCTTGAGAACCCCAGGGAATTTGT
TATATTTATGCTATTCAGTTCTAAAT                                                  TGCAACAACCTTGAGAACCCCAGGGAATTTGT
TATATTTATGCTATTCAGTTCTAAAT                                                  TGCAACAACCTTGAGAACCCCAGGGAATTTGT
tatatttatgctattcagttctaaatatagaaatt                                         tgcaacaaccttgagaaccccagggaatttgt
tatatttatgctattcagttctaaatatagaaatt                                              CAACCTTGAGAACCCCAGGGAATTTGT
  TATTTATGCTATTCAGTTATAAATATAGAAATTGAAACAG                                          CCTTGAGAACCCCAGGGAATTTGT
   atttatgctattcagttctaaatatagaaattgaa                                               CTTGAGAACCCCAGGGAATTTGT
    tttacgctattcagtactaaatatagaaattgaaa                                              CTTGAGAACCCCAGGGAATTTGT
     ttatgctattcagttctaaatatagaaattgaaac                                                          gggaatttgt
```

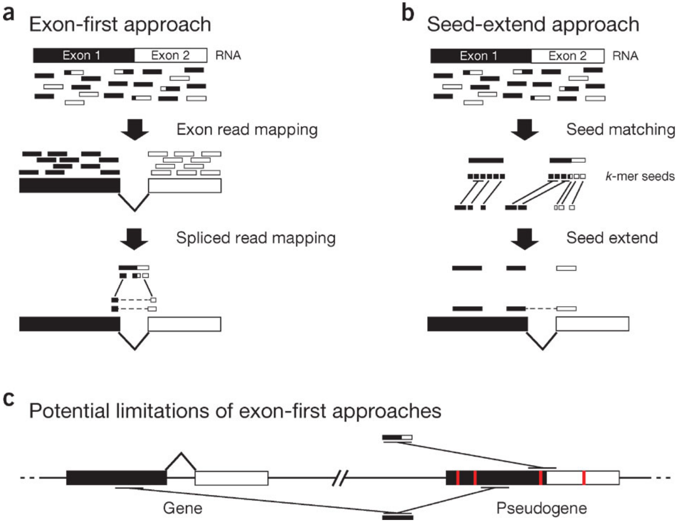
Align against genome or transcriptome

against transcriptome: easier, because no gapped alignment necessary

but: risk to miss possible alignments!

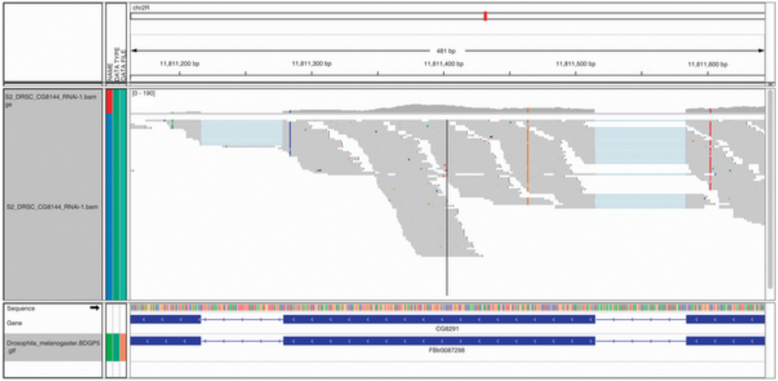
many tools available see e.g.,
http://massgenomics.org/short-read-aligners

Read mapping



Garber *et al.* Nature Methods 8, 469–477(2011)

Summarizing mapped reads

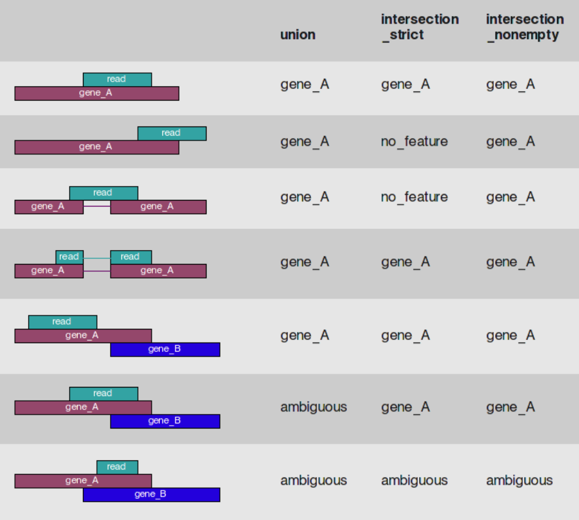


Count each read at most once

Discard a read if

- it cannot be uniquely mapped
- its alignment overlaps with several genes
- the alignment quality score is bad

Summarizing mapped reads: Counting rules



	union	intersection_strict	intersection_nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous	gene_A	gene_A
read gene_A gene_B	ambiguous	ambiguous	ambiguous

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

Quality Control: on 'raw' reads

- Basic information (total reads, sequence length, etc.)
- Per base sequence quality
- Overrepresented sequences (e.g., ribosomal RNAs)
- GC content
- Duplication level
- Etc

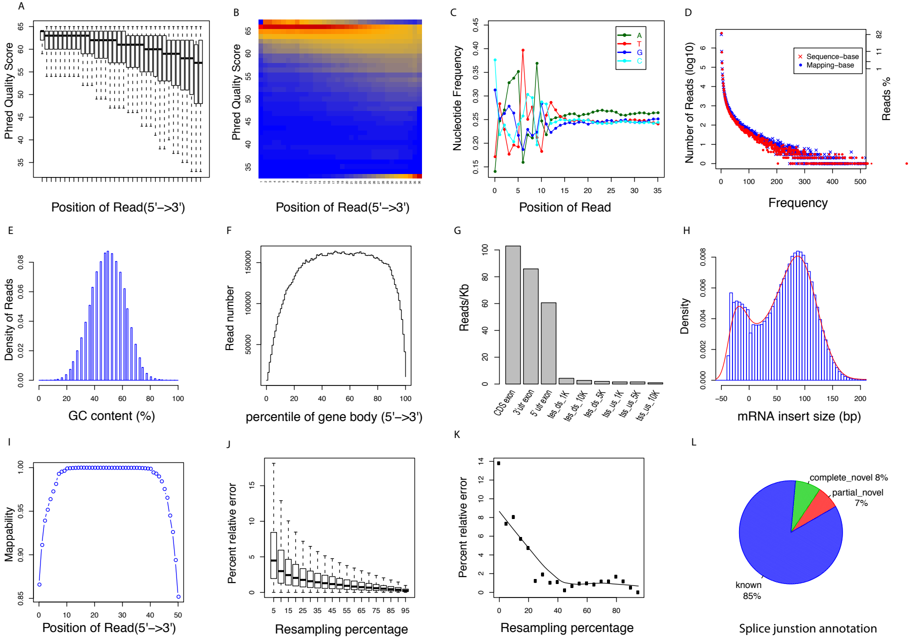
Tools: samtools, Fastqc, Fastx, Galaxy fastq tools, $\cdots$

Quality Control: on aligned reads

- Percentage of reads properly mapped or uniquely mapped
- Among the mapped reads, the percentage of reads in exon, intron, and intergenic regions.
- 5' or 3' bias
- The percentage of expressed genes

Tools: RSeQC[1], RNA-SeqQC, $\cdots$

[1] Wang *et al.* Bioinformatics, 28(16), 2184–2185(2012)


A
Phred Quality Score
65
60
55
50
45
40
35
Position of Read(5'−>3')
B
Phred Quality Score
65
60
55
50
45
40
35
Position of Read(5'−>3')
C
Nucleotide Frequency
0.45
0.40
0.35
0.30
0.25
0.20
0.15
A
T
G
C
0
5
10
15
20
25
30
35
Position of Read
D
Number of Reads (log10)
6
5
4
3
2
1
0
Sequence−base
Mapping−base
Reads %
82
11
1
0
100
200
300
400
500
Frequency
E
Density of Reads
0.08
0.06
0.04
0.02
0.00
0
20
40
60
80
100
GC content (%)
F
Read number
150000
100000
50000
0
20
40
60
80
100
percentile of gene body (5'−>3')
G
Reads/Kb
100
80
60
40
20
0
CDS exon
3'utr exon
5'utr exon
tes_ds_1K
tes_ds_10K
tes_ds_5K
tss_us_1K
tss_us_5K
tss_us_10K
H
Density
0.008
0.006
0.004
0.002
0.000
−50
0
50
100
150
200
mRNA insert size (bp)
I
Mappability
1.00
0.95
0.90
0.85
0
10
20
30
40
50
Position of Read(5'−>3')
J
Percent relative error
15
10
5
0
5
15
25
35
45
55
65
75
85
95
Resampling percentage
K
Percent relative error
14
12
10
8
6
4
2
0
0
20
40
60
80
Resampling percentage
L
complete_novel 8%
partial_novel
7%
known
85%
Splice junstion annotation

Normalization

Each sample (library) will have different number of total reads
Total count, Counts per million and Reads Per Kilobase per Million mapped reads (RPKM)
For differential expression these are not appropriate!!!

Toy example:

	sample A	sample B
gene 1	100	80
gene 2	100	80
...	...	...
gene 100	100	80
gene 101	0	2.000
Total Counts:	10.000	10.000

Normalization

Using Counts per million $\frac{X_{ij}}{X_{.j}} 10^6$ with $X_{.j} = \sum_{i=1}^{101} X_{ij}$

	sample A	sample B
gene 1	10.000	8.000
gene 2	10.000	8.000
...	...	...
gene 100	10.000	8.000
gene 101	0	20.000

All genes are differentially expressed!
Is this really true?

Normalization

Using TMM (trimmed mean of M-values)

	sample A	sample B
gene 1	10.000	10.000
gene 2	10.000	10.000
...	...	...
gene 100	10.000	10.000
gene 101	0	250.000

One gene differentially expressed!
Seems more realistic

TMM

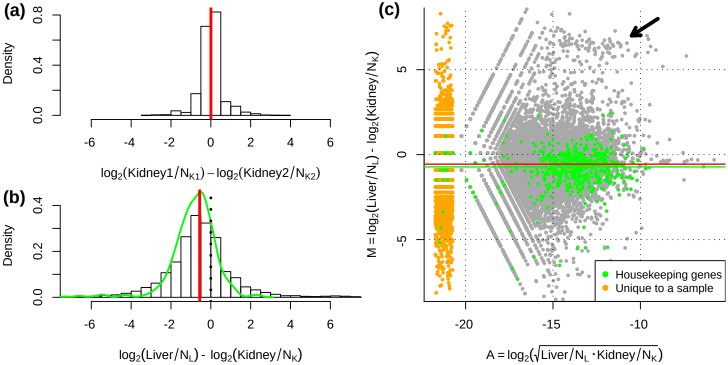


Figure : (a) technical replicates and (b) liver versus kidney (c) An M versus A plot comparing liver and kidney.

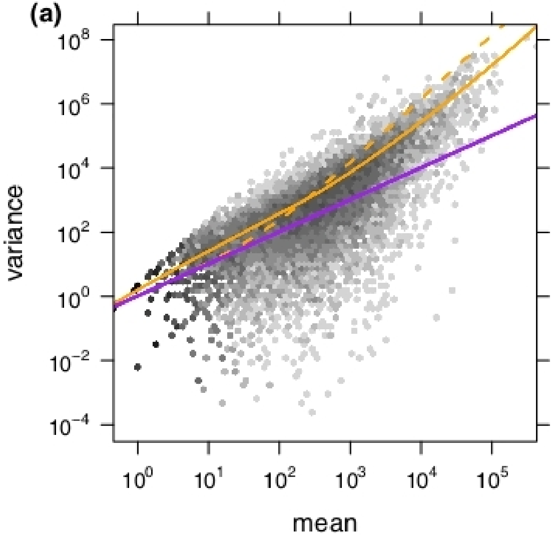
Robinson *et al.* Genome Biology, 11(3), (2010)

Testing for differential expression

- counts are discrete: $0, 1, 2, \cdots$
- large dynamic range $[0, > 100000]$
- Poisson or Negative Binomial distributed
- mean is approximately equal to the variance
- generalized linear model
- likelihood ratio test

Tools: edgeR, DESeq2, Cuffdiff, Myrna, $\cdots$

Mean variance relationship



Anders *et al.* Genome Biology, 11(10), (2010) From differential expression to Biology

- gene set enrichment (GO and KEGG)
- network construction (co-regulated genes)
- data integration (eQTL, meQTL, $\cdots$)
- $\ldots$

Other things you can do with RNAseq

- allele specific expression
- isoform (transcript) expression
- variant detection
- $\ldots$