

Analysis of RNA-seq data

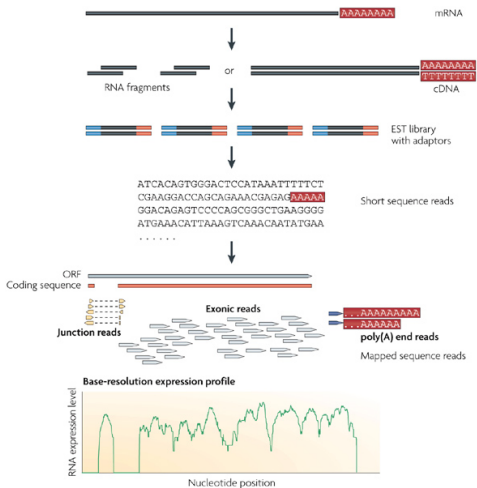
FOS 2017

Maarten van Iterson

Leiden University Medical Center
Department of Molecular Epidemiology

June 14, 2017

A typical RNA-seq experiment



Nature Reviews | Genetics

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

```
@SRRO14849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRRO14849.1 EIXKN4201CFU84 length=93
3+&$#"7F071,";C?,B;?6B;:EA1EA
1EA59B:?:#9EAOD@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

```
TATATTTATGCTATTCAGTTCTAAATATAGAAATTGAAACAGCTGTGTTTAGTGCCCTTTGTTCA-----ACCCCTTGCAACAACCTTGAGAACCCAGGGAATTGT
TATATT ATGCTATTCAGTTCTAAATATAGAAATTGAAACAG GTGTTTAGTGCCCTTTGTTCA-----ACCCCTTGCAACAAC aacccaggggaatttgt
tatatttatgetattcagttctaaatatagaaatt acagctgtgttttagtgccctttgttca-----accccttg aacaaccttgagaacccaggggaatttgt
TATAT TATGCTATTCAGTTCTAAATATAGAAATTGAAACA ctgtgttttagtgccctttgttca-----accccttgcaac ACCTTGAGAACCCAGGGAATTGT
TATATTTA getattcagttctaaatatagaaattgaaacagct GTTAGTGCCCTTTGTTTACATAGACCCCTTGCAA aaccttgagaacccaggggaatttgt
TATATTTATGCTATTCAGT GAAATTGAAACAGCTGTGTTTAGTGCCCTTTGTTCA ccccttacaacaaccttgagaacccaggggaattt
tatatttatgetattcagt GCCTTTGTTTACATAGACCCCTTGCAACAACCTT caggggaatttgt
tatatttatgetattcagttcta AG-----ACCCCTTGCAACAACCTTGAGAACCCAGGGAA
TATATTTATGCTATTCAGTTCTAA A-----ACCCCTTGCAACAACCTTGAGAACCCAGGGAA
TATATTTATGCTATTCAGTTCTAAA A-----ACCCCTTGCAACAACCTTGAGAACCCAGGGAA
TATATTTATGCTATTCAGTTCTAAA TGCAACAACCTTGAGAACCCAGGGAATTGT
TATATTTATGCTATTCAGTTCTAAAT TGCAACAACCTTGAGAACCCAGGGAATTGT
TATATTTATGCTATTCAGTTCTAAAT TGCAACAACCTTGAGAACCCAGGGAATTGT
tatatttatgetattcagttctaaatatagaaatt tgaacaaccttgagaacccaggggaatttgt
tatatttatgetattcagttctaaatatagaaatt CAACCTTGAGAACCCAGGGAATTGT
TATTTATGCTATTCAGTTATAAATATAGAAATTGAAACAG CCTTGAGAACCCAGGGAATTGT
atttatgetattcagttctaaatatagaaattgaa CTTGAGAACCCAGGGAATTGT
tttaacgetattcagtaactaaatatagaaattgaaa CTTGAGAACCCAGGGAATTGT
ttatgetattcagttctaaatatagaaattgaaac ggggaatttgt
```

Align against genome or transcriptome

against transcriptome: easier, because no gapped alignment necessary

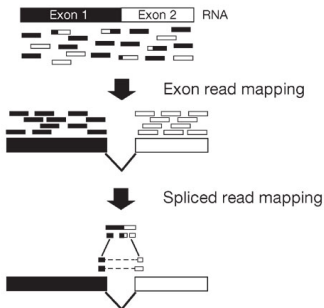
but: risk to miss possible alignments!

many tools available see e.g.,

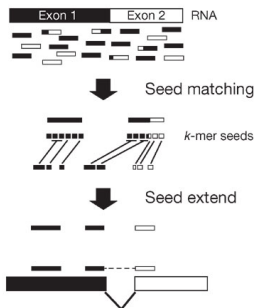
<http://massgenomics.org/short-read-aligners>

Read mapping

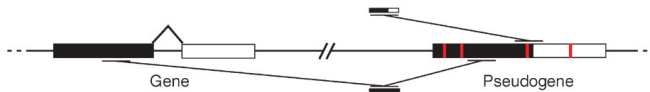
a Exon-first approach



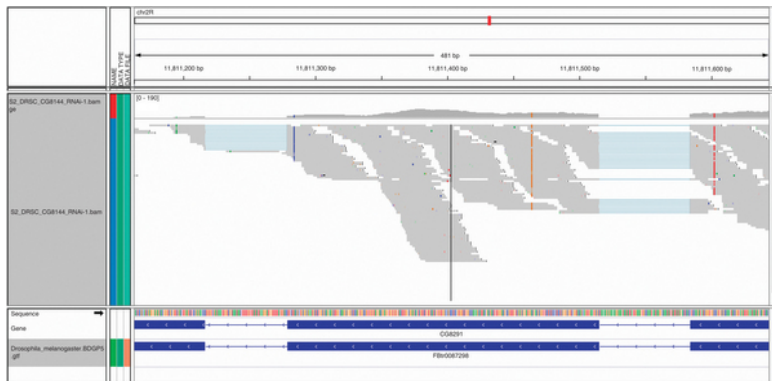
b Seed-extend approach



c Potential limitations of exon-first approaches



Summarizing mapped reads

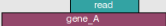
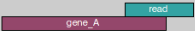



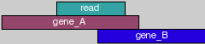
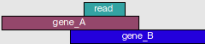


Count each read at most once

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
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	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
- and more ...

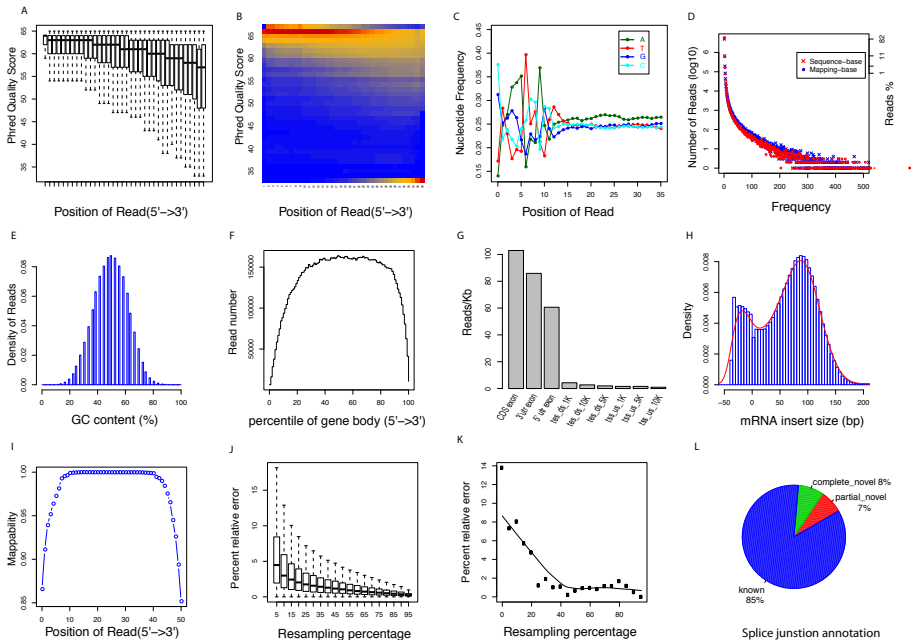
Tools: samtools, Fastqc, Fastx, Galaxy fastq tools, ...

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: RSeQC¹, RNA-SeqQC, ...

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



Normalization

Each sample different library size/sequencing depth

For example:

	sample A	sample B
gene 1	100	10
gene 2	20	2
...
gene 100	1000	100
Total Counts:	10.000	1.000

library size sample A $10\times$ sample B

But normalization cont'd

- intuitive normalization methods: total count, cpm, rpkm
- not appropriate for differential expression

For 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

sequencing depth the same but a single gene highly differentially expressed

Normalization cont'd

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

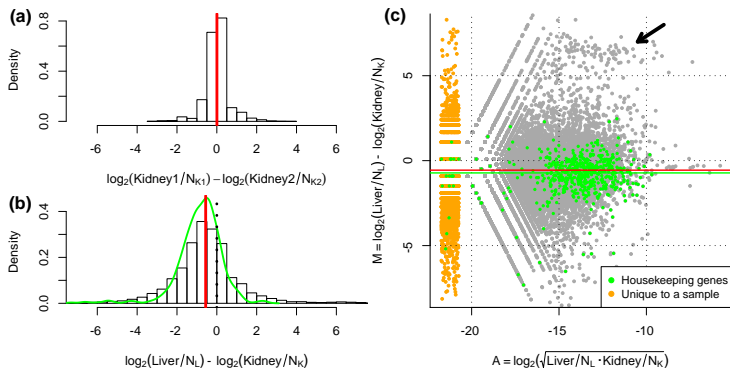


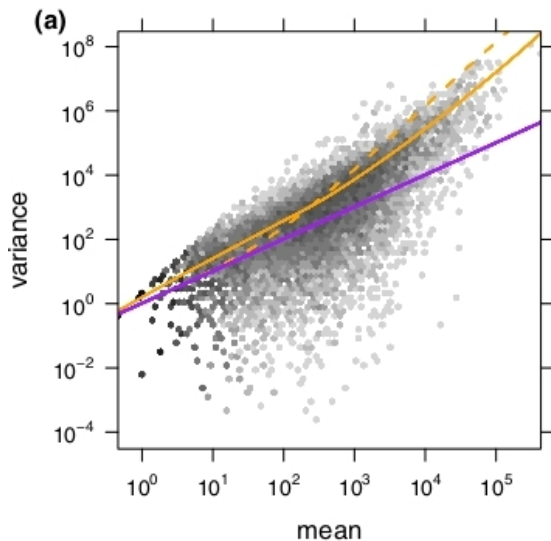
Figure : (a) technical replicates and (b) liver versus kidney (c) An M versus A plot comparing liver and kidney.

Testing for differential expression

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

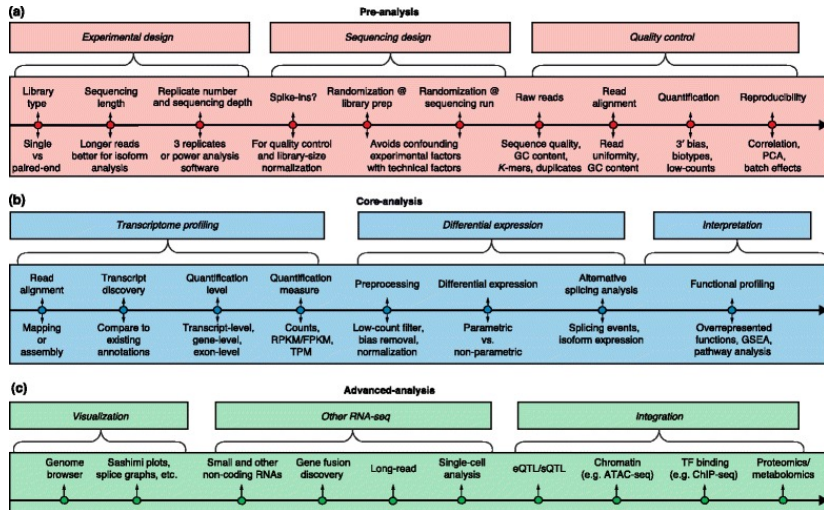
Tools: edgeR, DESeq2, Cuffdiff, Myrna, \dots

Mean variance relationship



Anders *et al.* Genome Biology, 11(10), (2010)

A generic roadmap for RNA-seq computational analyses



Conesa *et al.* Genome Biology, 17(13), (2016)