RNA-seq quantification

Dusan Randjelovic
Seven Bridges Genomics, Inc.
Belgrade: 2016/2017

SevenBridges

Recap (1/3)

- DNA -> RNA -> Protein
- + DNA replication

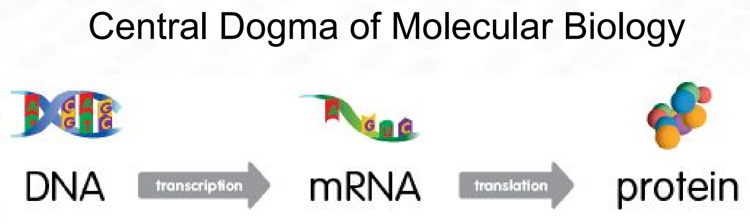
  (all cells in body have same DNA)
- NOT all cells express same genes

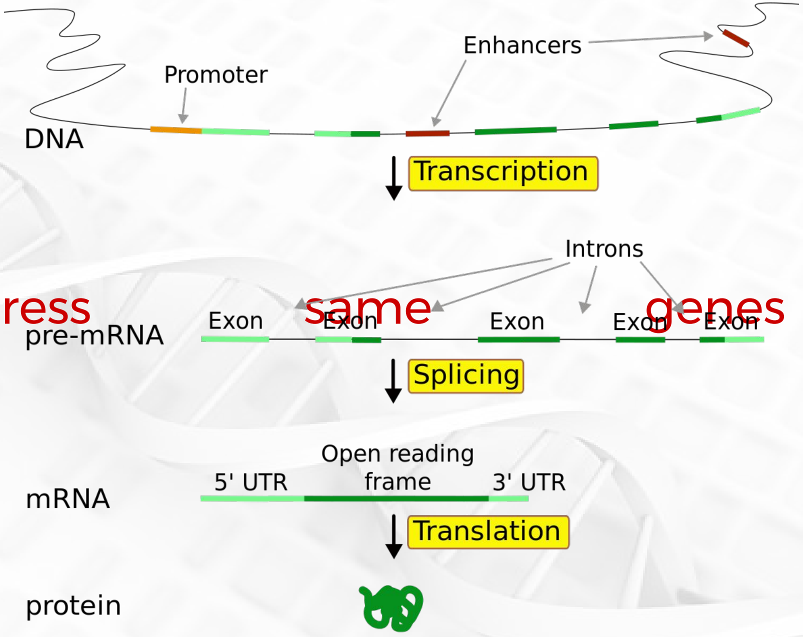
  (due to cell type or cell cycle)

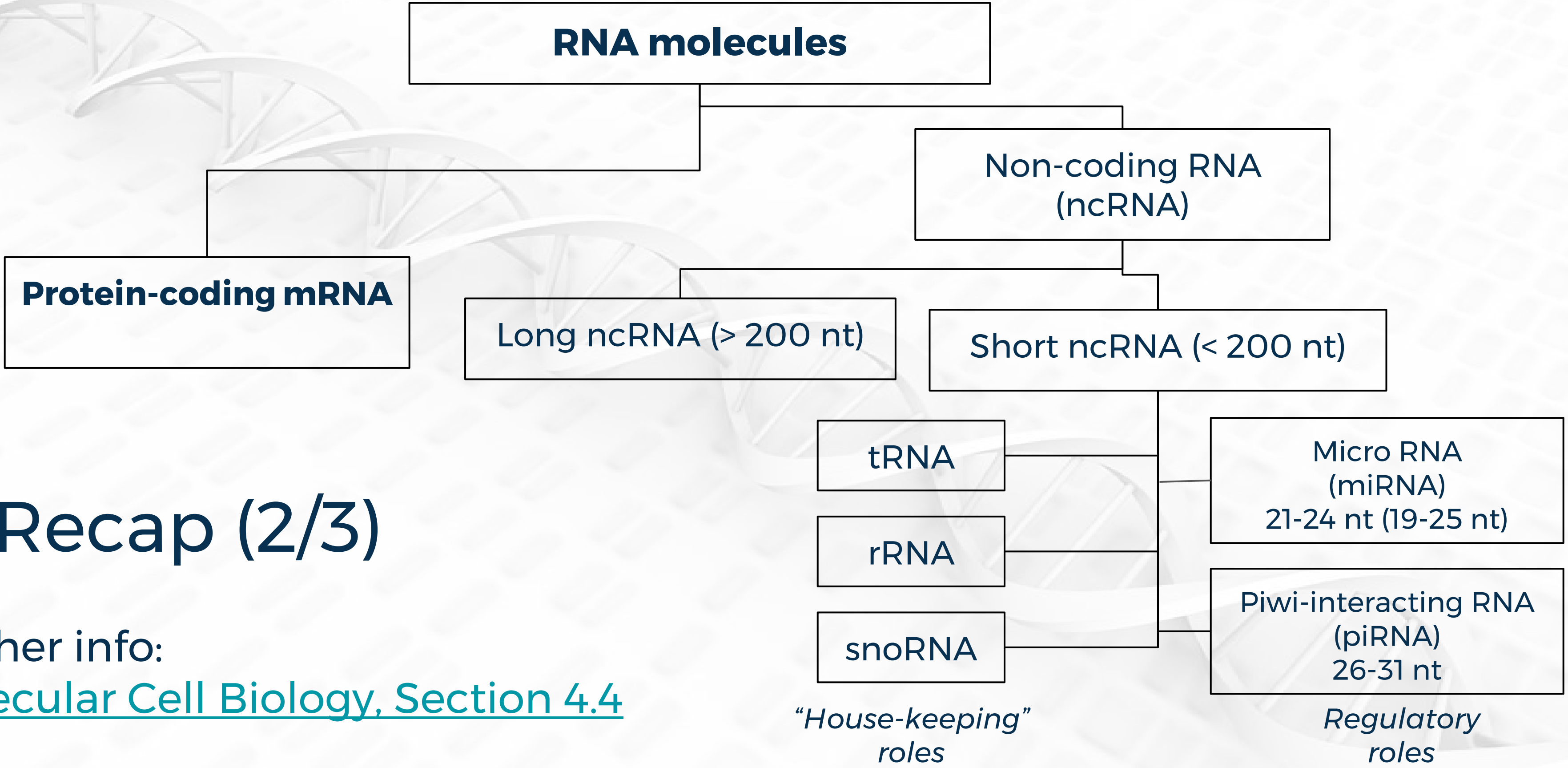
Further info:

[animation: DNA from the beginning]

[Nature Scitable: Gene expression]



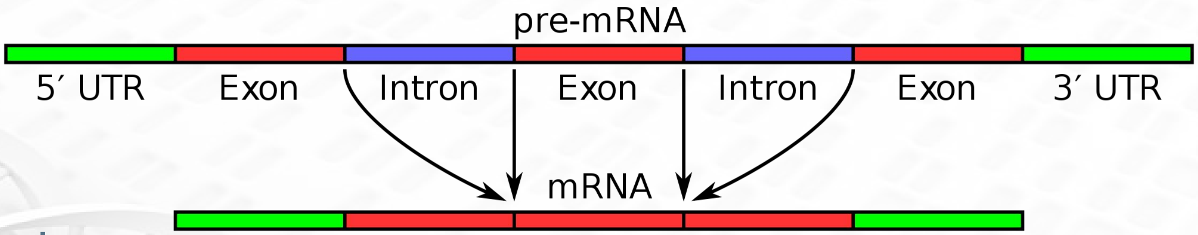




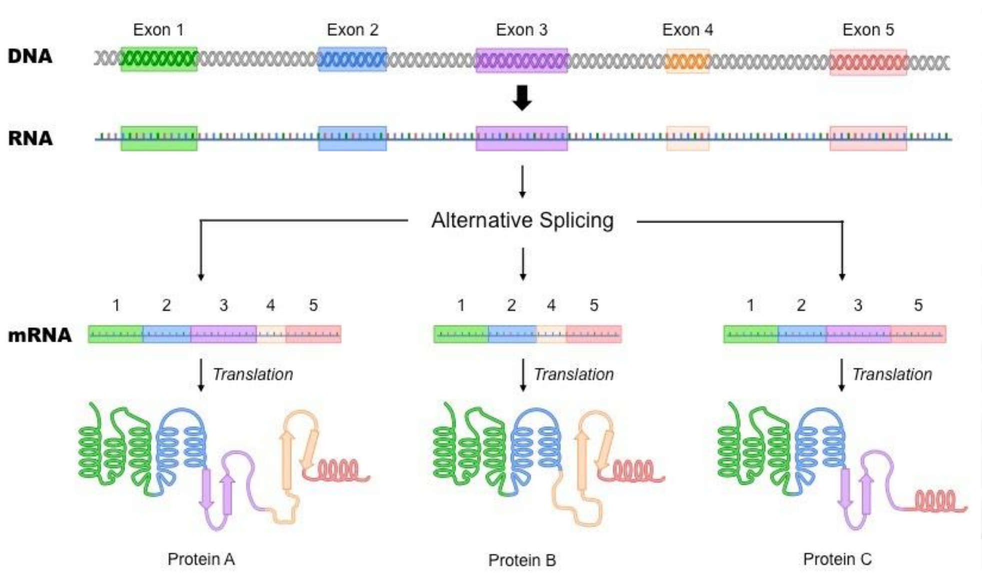
Recap (2/3)

Further info:
[Molecular Cell Biology, Section 4.4]

Recap (3/3)



- mRNA = spliced transcript + poly-A tail
- Alternative splicing (isoforms of genes)



Further info:
DNA Learning Center

Motivation for RNA quantification

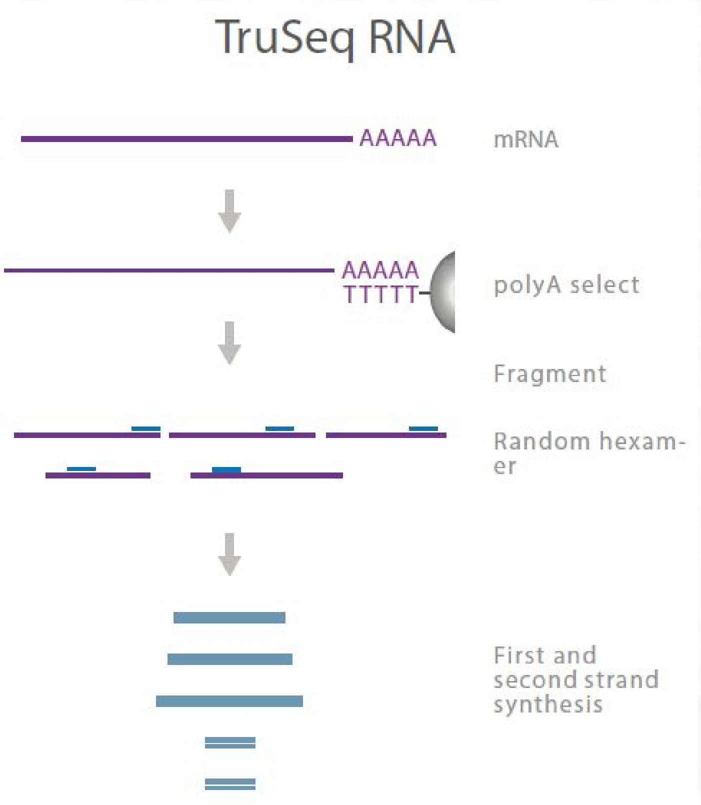
- We (usually) want to check if there is **change in transcription** between conditions (healthy/sick, treated/untreated, different tissues, etc..)
- Typical studies:
  - DNA-seq -> alignment -> variant calling
  - RNA-seq -> (alignment) -> quantification -> differential expression

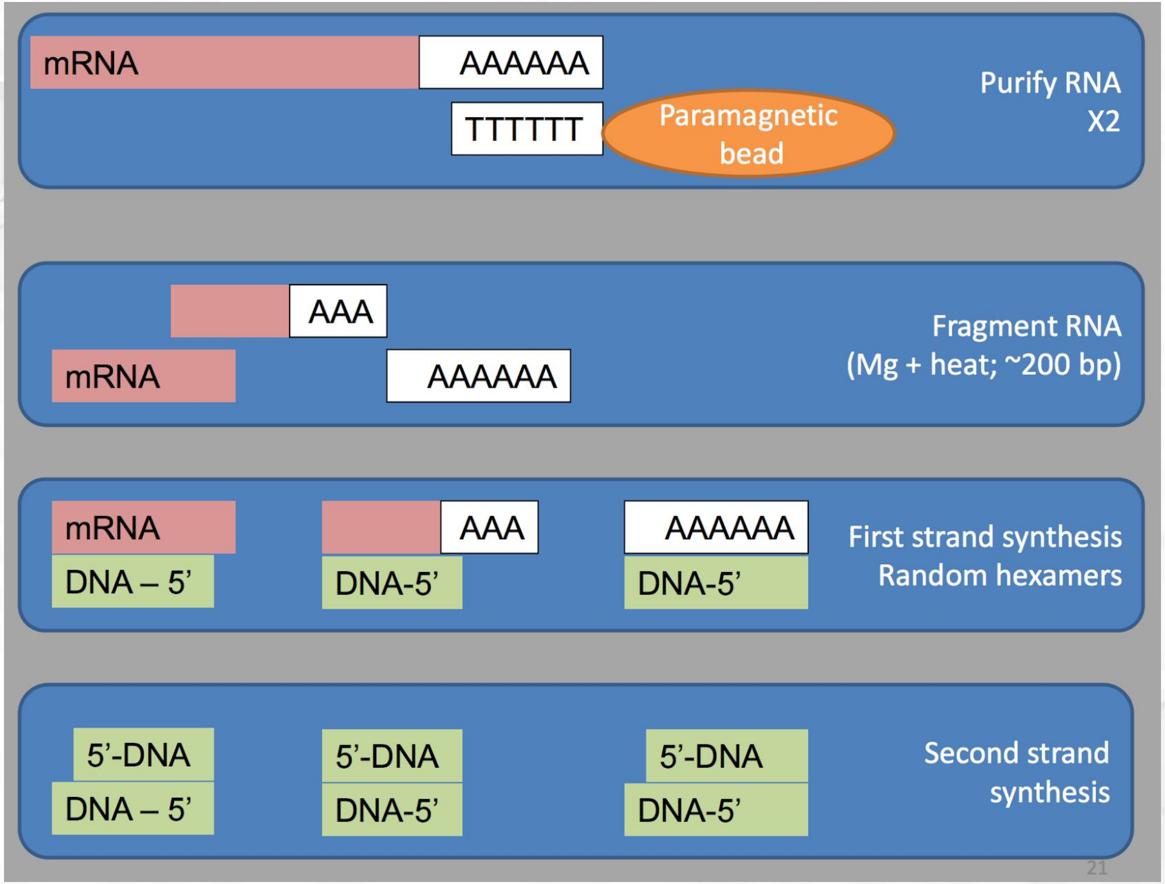
Further info:
Number of mRNAs/cell

RNA-seq library prep (1/2)

Protocols differ in :

- what types of RNA they target (total RNA, mRNA)
- on fragment sizes
- strand specificity (more info)
- bulk or single-cell
- ligation, priming...



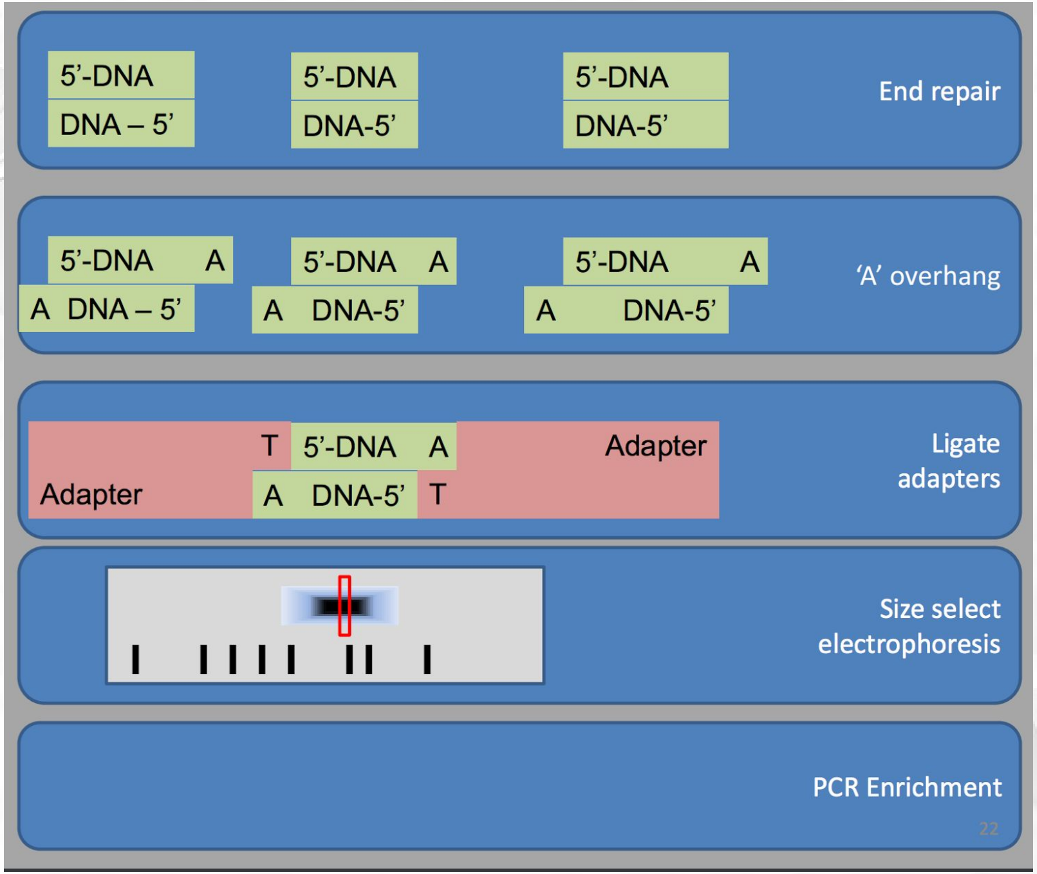


Purify RNA
X2

Fragment RNA
(Mg + heat; ~200 bp)

First strand synthesis
Random hexamers

Second strand
synthesis



End repair

'A' overhang

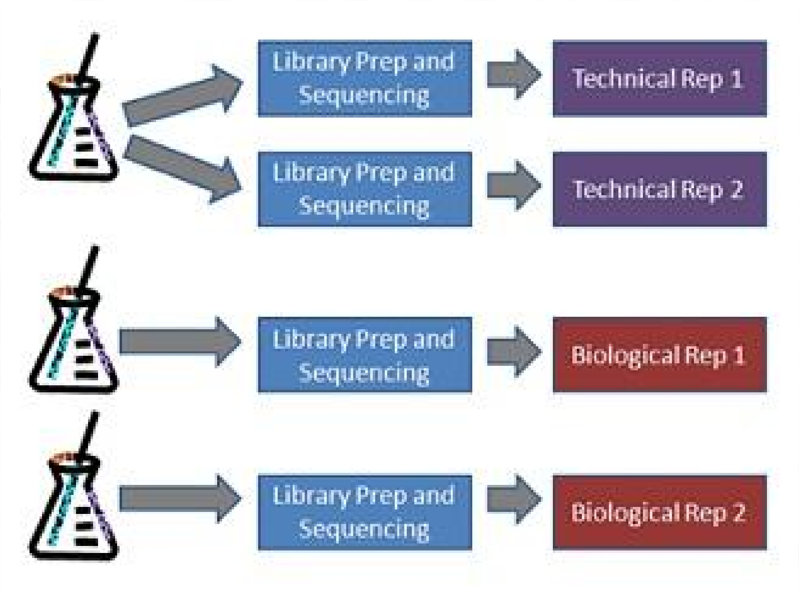
Ligate adapters

Size select electrophoresis

PCR Enrichment

RNA-seq library prep (2/2)

- technical or biological replicates
- Important to estimate # replicates (power analysis)


Library Prep and Sequencing
Technical Rep 1
Library Prep and Sequencing
Technical Rep 2
Library Prep and Sequencing
Biological Rep 1
Library Prep and Sequencing
Biological Rep 2

Source: http://hdl.handle.net/2345/3145

RNA-seq data analysis

- alignment, assembly, **relative abundance**, differential expression, functional enrichment analysis
- RNA-seq quantification

  vs.

  microarrays or qRT-PCR
- Latest approach -> single-cell RNA-seq (preserve information of cell origin, usually by priming all transcripts from same cell identically)

RNA-seq data (1)

(variance and ambiguity)

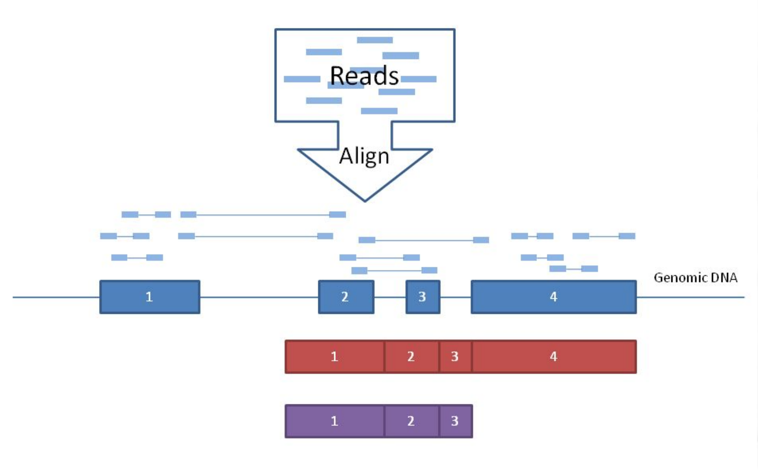
- sampling variance (biological replicates)
- technical or *biological* variance

  (both technical and biological rep.)

alternative splicing:

- mapping ambiguity

  (multiple mapping)



Source: http://dx.doi.org/10.13070/mm.en.3.203

RNA-seq data (2)

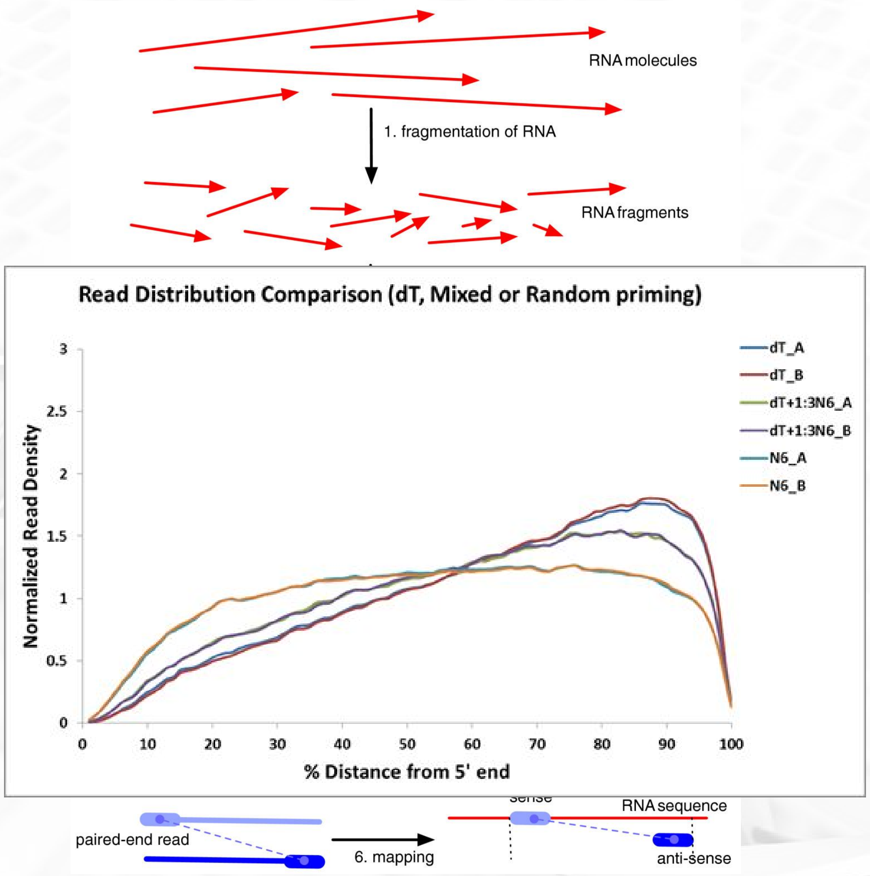
(biases)

- fragment length distribution

  (small fragments -> more ambiguity)
- positional and sequence-specific

  (due to priming or fragmentation)
- sequencing errors

  (error model - mismatches & indels)



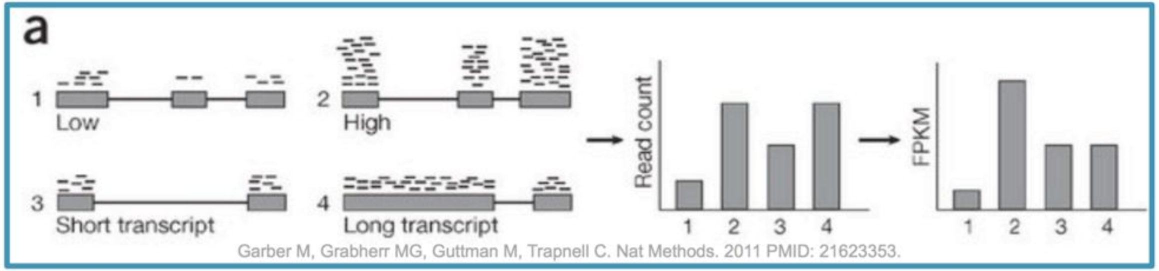
Source: doi:10.1186/gb-2011-12-3-r22

RNA-seq data (3)

(normalization)

- **within sample** normalization (transcript length + sequencing depth):

Let X_i be number of reads aligned to ith transcript
and l_i is length of ith transcript
$\sum_i X_i \neq$ expression of a gene



Garber M, Grabherr MG, Guttman M, Trapnell C. Nat Methods. 2011 PMID: 21623353.

Adjust for (effective) length

Let N = total number of mapped reads
$\widetilde{l}_i = l_i - \mu_{FLD} + 1$, μ is mean of fragment length dist.

$$\text{effCounts}_i = X_i \cdot \frac{l_i}{\widetilde{l}_i}$$

RNA-seq data (4)

(normalization)

$$\mathbf{TPM}_i = \frac{X_i}{\widetilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\widetilde{l}_j}} \right) \cdot 10^6, \ \mathbf{FPKM}_i = \frac{X_i}{\left(\frac{\widetilde{l}_i}{10^3}\right) \cdot \left(\frac{N}{10^6}\right)}$$

Some approaches:

- Raw counts: Quantile normalization
- Raw counts: Thinning to the minimum library size
- Transcripts per million
- Fragments per kilobase of exon per million reads

Further info: What the FPKM

RNA-seq data (5)

(features for quantification)

- Exons, **transcripts** or genes
- Raw counting (aligned reads)

vs.

probabilistic of

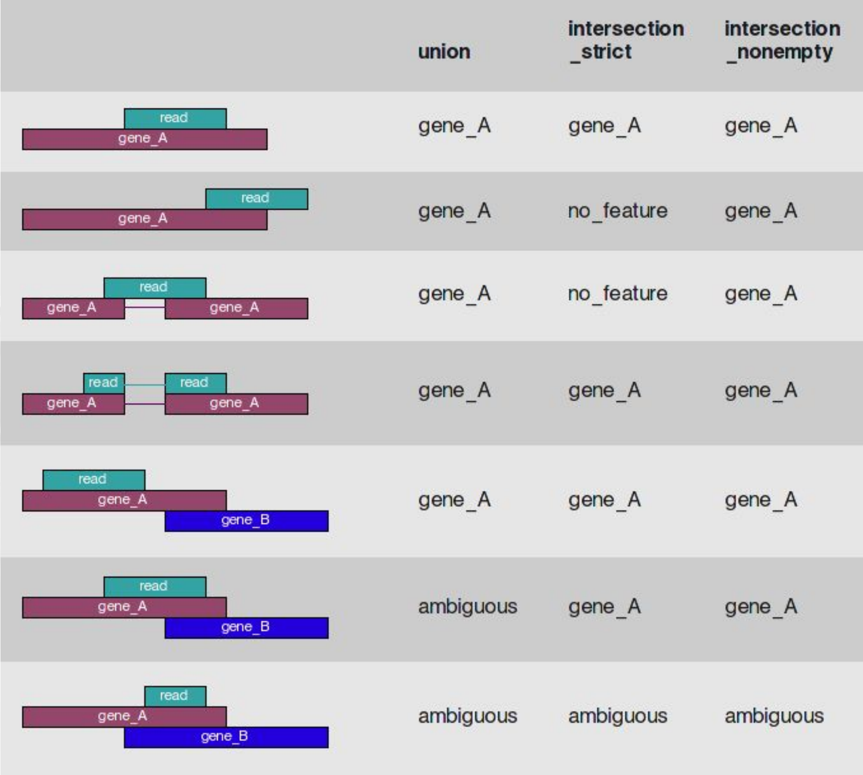
aligned reads to features

vs.

probabilistic of

unaligned/mapped kmers of reads

HTseq counting model



	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

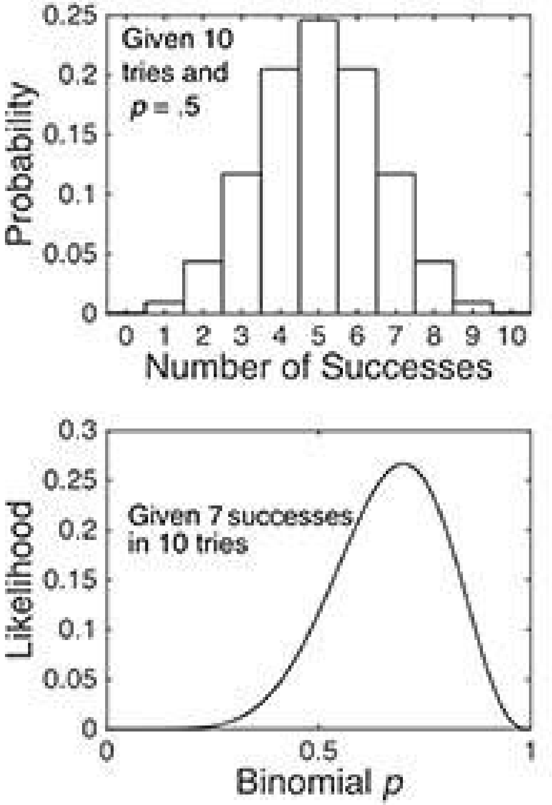
Statistical background (1)

- type of problem -> counting transcripts (Poisson, binomial/multinomial)
- Likelihood vs. probability

If *probability* is a function of data given some params, then *likelihood* is function of those params given the data

$$L(p;x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

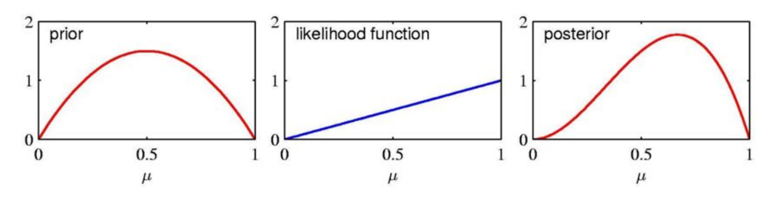
- Maximum Likelihood Estimation (MLE)



Statistical background (2)

- MLE: analytically or numerically
- Frequentist: data -> model fit-> validate model params (MLE)
- Bayesian: prior model + likelihood (given data) -> posterior model
- Bayesian interpretation closer to algorithmic approach (treats additional data, hidden params)

Further info: cruncher notebook



MLE example (RNA)

$i = 5$ single-end, equal-length reads (a,b,c,d,e)

$k = 3$ transcripts (blue, green, red)

$\rho = (\rho_{blue}, \rho_{green}, \rho_{red})$ relative abundances of transcripts

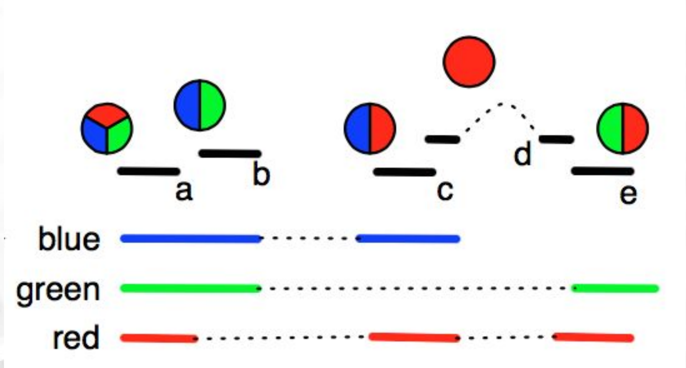
$\sum_k \rho_k = 1$, multinomial distribution

$P_i = \sum_k y_{i,k} \cdot \rho_k$, probability of detecting i-th read

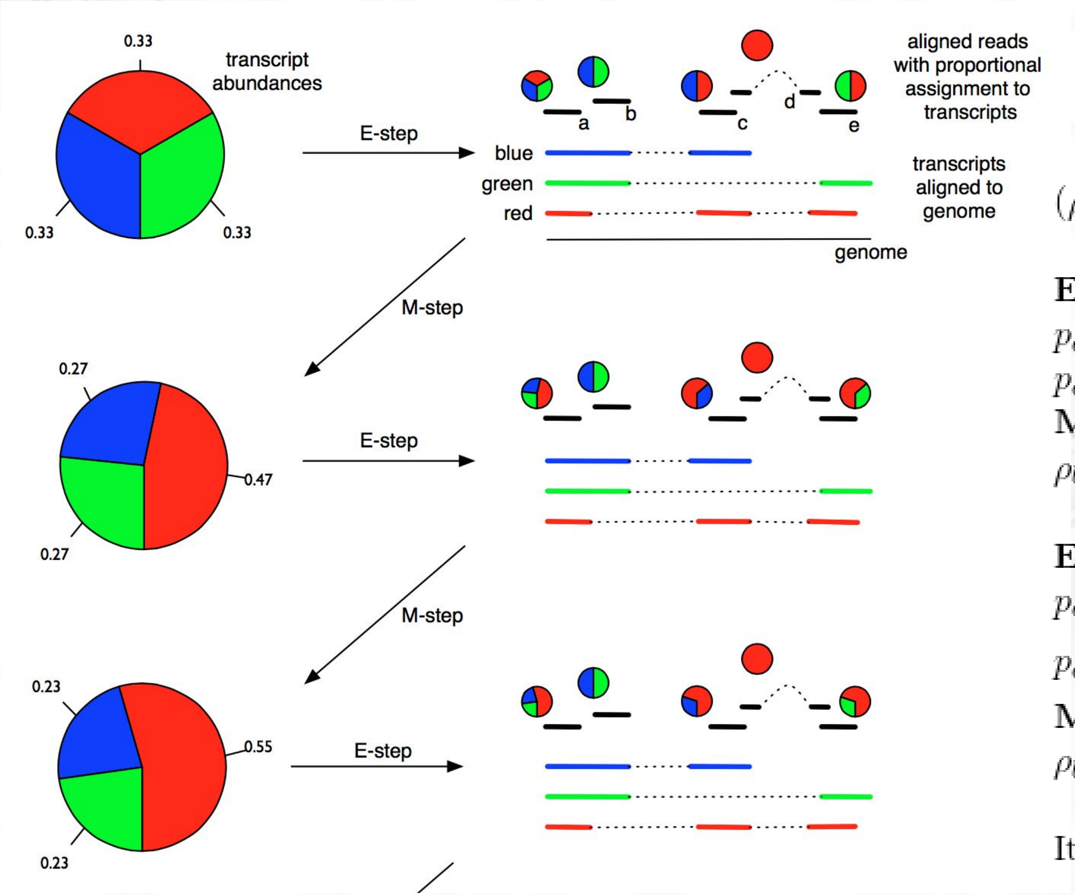
where $y_{i,k} = 1$ if i-th read aligns to k-th transcript, otherwise 0

$$\boxed{L(\rho) = \prod_i \sum_k y_{i,k} \cdot \rho_k}$$

Analytical solution $\rho = (0.18, 0.18, 0.64)$



Adapted from: Lior Pachter 2011, arxiv: 1104.3889v2



RNA example EM

$(\rho_{blue}, \rho_{green}, \rho_{red}) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, uniform prior

E1 step: Proportional assignment

$p_a = (1/3, 1/3, 1/3),\ p_b = (1/2, 1/2, 0),$

$p_c = (1/2, 0, 1/2),\ p_d = (0, 0, 1),\ p_e = (0, 1/2, 1/2)$

M1 step: recalculate abundances

$\rho_{blue} = (1/3 + 1/2 + 1/2 + 0 + 0)/5 = 0.27$

E2 step: prior $= (0.27, 0.27, 0.46)$

$p_a = (0.27, 0.27, 0.46),\ p_b = (1/2, 1/2, 0),$

$p_c = (\frac{0.27}{0.46 + 0.27}, 0, \frac{0.46}{0.46 + 0.27}),\ p_d = (0, 0, 1), \ldots$

M2 step:

$\rho_{blue} = (0.27 + 1/2 + 0.37 + 0 + 0)/5 = 0.23$

Iterative convergance $\rho_{blue} = 0.33, 0.27, 0.23, ..., 0.18$

Raw counts implementations

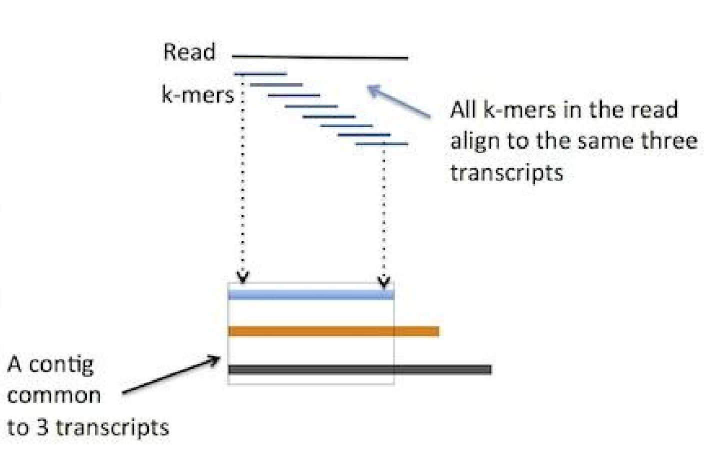
- HTseq count
- featureCounts

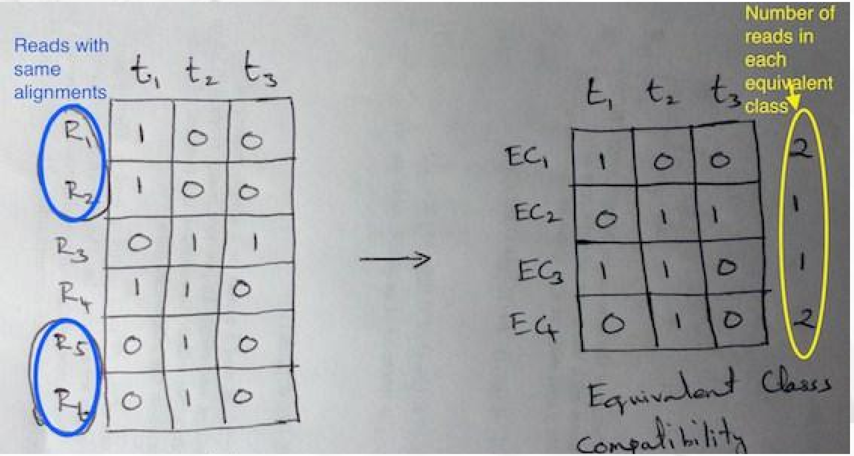
EM implementations

- RSEM
- eXpress
- ... EM in other areas of genomics

Pseudo-alignment methods

- Find set of transcripts that a read belongs to, BUT not exact position
- Counting kmers instead of reads
- RESULT: no local alignment, much faster EM


Read
k-mers
All k-mers in the read align to the same three transcripts
A contig common to 3 transcripts


Reads with same alignments
Number of reads in each equivalent class
Equivalent Class Compatibility

	t_1	t_2	t_3
R_1	1	0	0
R_2	1	0	0
R_3	0	1	1
R_4	1	1	0
R_5	0	1	0
R_6	0	1	0

	t_1	t_2	t_3	
EC_1	1	0	0	2
EC_2	0	1	1	1
EC_3	1	1	0	1
EC_4	0	1	0	2