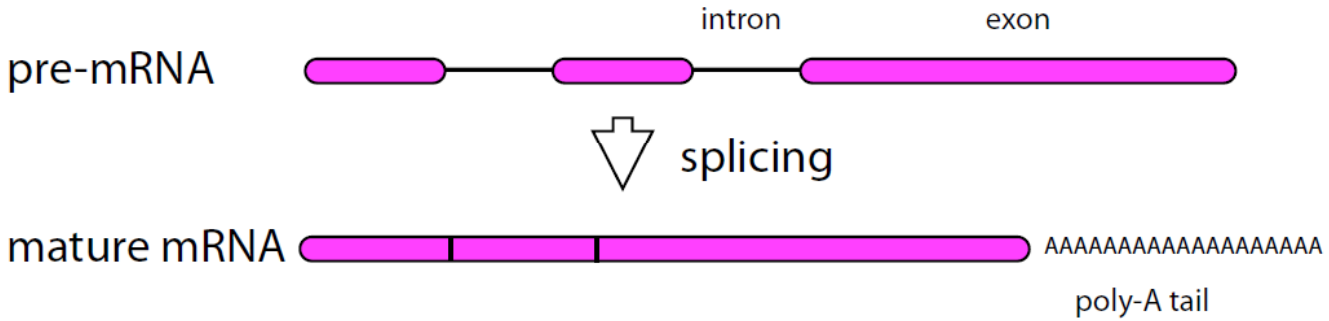
RNA sequencing

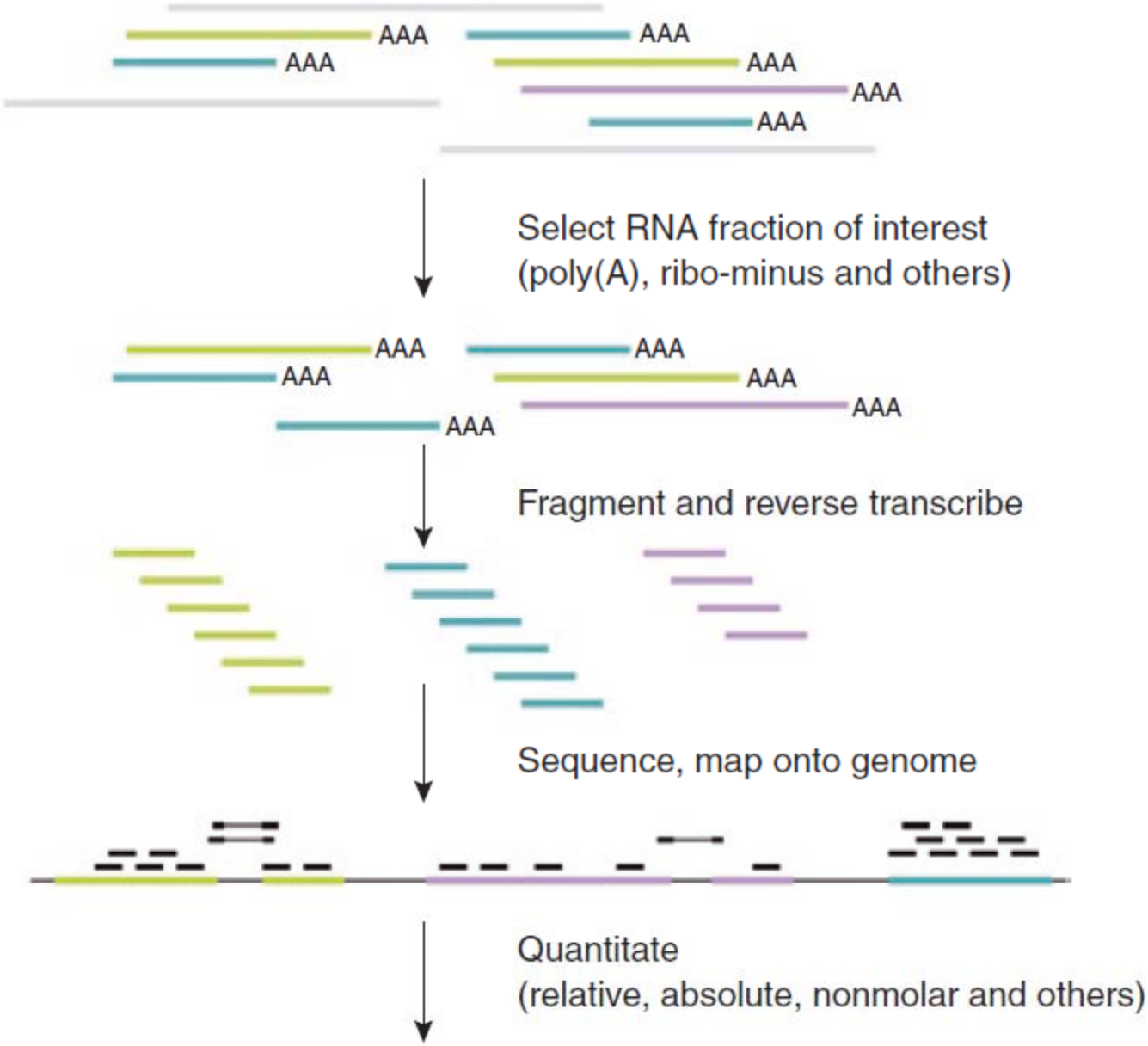
RNAs

- Coding RNA
  - The usual “genes” , with poly-A



- Non-coding RNA (ncRNA)
  - Short non-coding RNAs , microRNA
  - Long non-coding RNAs
- Ribosomal RNA (Most of the RNA in the cell, but most people are not interested in this part)

Overview

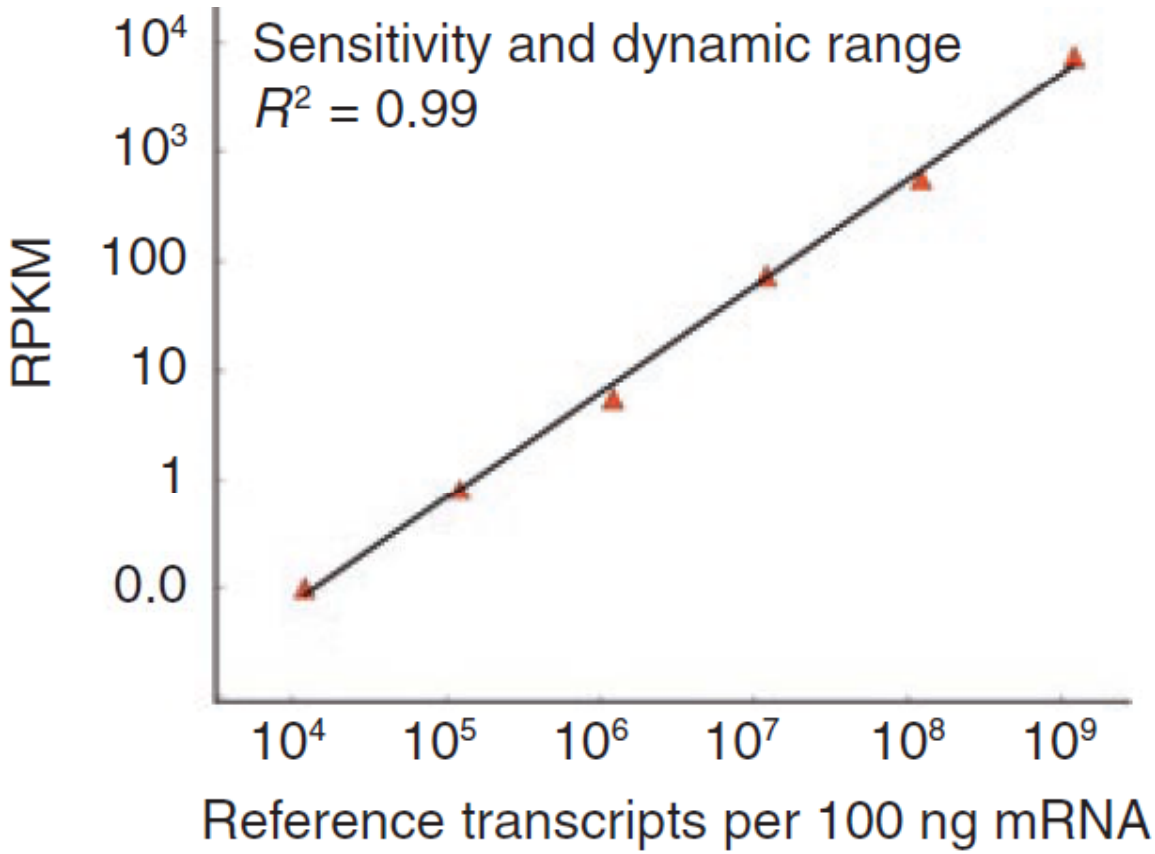

AAA
AAA
AAA
AAA
AAA
AAA
Select RNA fraction of interest
(poly(A), ribo-minus and others)
AAA
AAA
AAA
AAA
AAA
Fragment and reverse transcribe
Sequence, map onto genome
Quantitate
(relative, absolute, nonmolar and others)

What affects the counts on each gene?

- Gene expression level
- Sequencing depth
  - For the same biological sample, the more we count in total, the more counts are expected on each gene expressed
- Gene length
  - For the same expression level, longer genes tend to yield more fragments
- Mortazavi(2008) Nat. Method introduced RPKM

  Reads-per-kilobase-per-million

  $RPKM_{gi} = Y_{gi}/L_g N_i$



Mortazavi (2008) Nat. Methods

How do we test for differential expression?

- Counts

$Y_1, Y_2, \ldots, Y_{n1}$ vs $Y_{n1+1}, Y_{n1+2}, \ldots, Y_{n1+n2}$

- Early approaches assumed that under null (non-differential expression)

$Y_{gi} \sim$ Poisson$(N_i L_g \mu_g)$

(Or $Y_{gi} \sim$ Binomial $(N_i, L_g \mu_g)$ which is equivalent to the Poisson model since N is large ($\sim 10^{7 \text{or} 8}$) and proportion for each gene is very small)

Sources of variation

- Technical replicates
  - Take tissue sample from individual
  - Extract RNA from a tissue sample
  - Convert RNA to DNA, library preparation
  - Sequencer

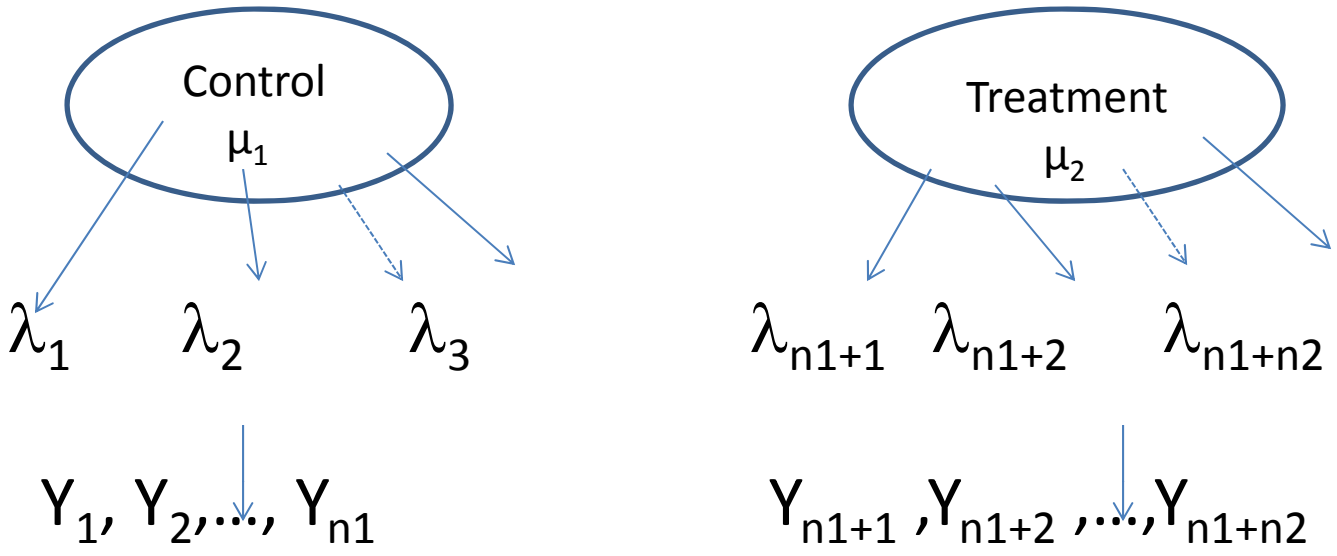
  All of the above induces technical variation, and several steps are independent of the sequencing technology.

  Other factors that will continue to affect data:

  Day-to-day, lab effect, technician effect…

How does Poisson model work?

- OK for technical replicates
  - Several publications showed that the counts in technical replicates are Poisson
- Too many positives in biological replicates
  - Gene Expression is a stochastic process
  - The expression levels in two biological replicates under the same treatment are not identical
  - Poisson model does not account for this variability



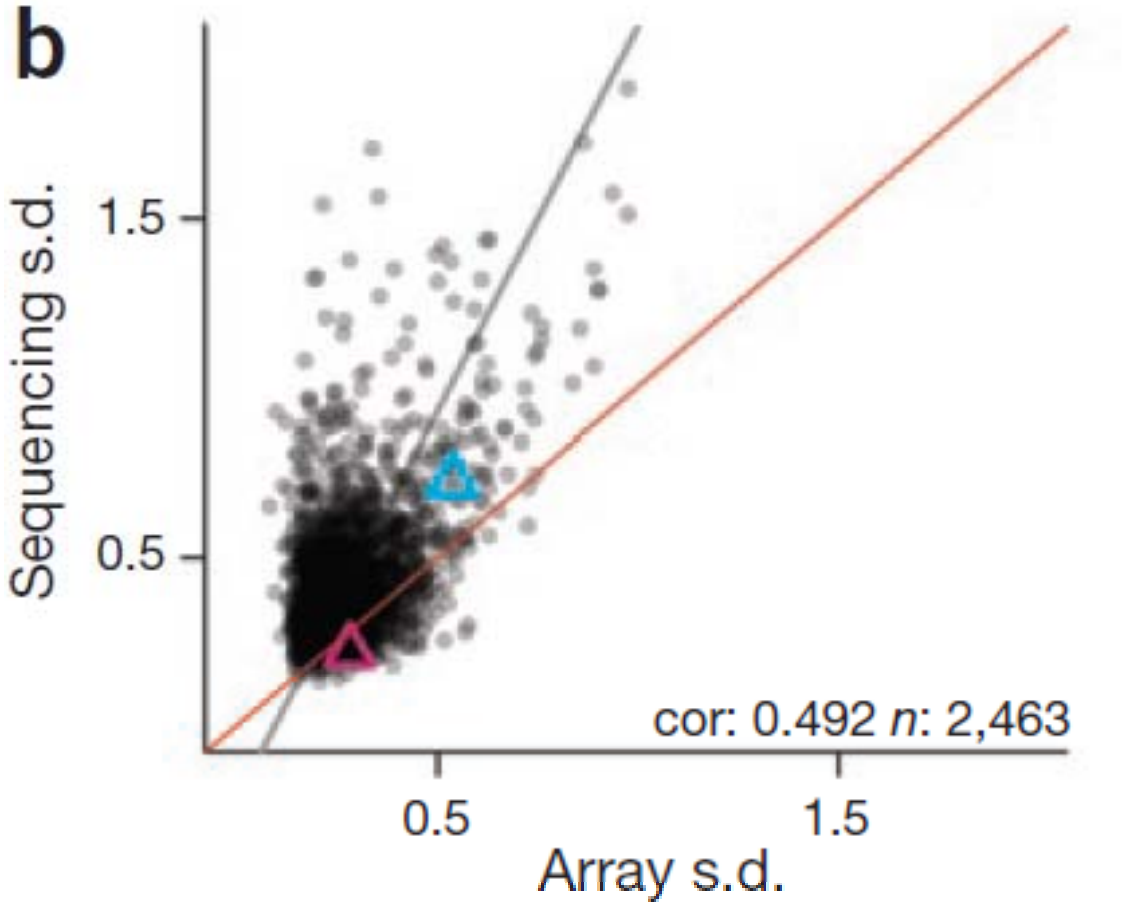
$E(Y_1)=E\{\ E[Y|\ \lambda_1]\ \}=E\{\lambda_1\}=\mu_1$

$Var(Y_1)=Var\{\ E[Y|\ \lambda_1]\ \}+E\{\ var[Y|\ \lambda_1]\ \}$

$=Var\{\lambda_1\}+E\{\lambda\}\ >\mu_1$ (unless $var(\lambda_1)=0$)

Over-dispersion (compared to Poisson)

Biological variation



Negative binomial distribution

Consider one gene of interest,

- Suppose under a given biological condition, the expression rate λ follows a gamma distribution with shape α and scale β (then mean $\mu = \alpha\beta$ is variance is $\alpha\beta^2$)
- Suppose the counts from the gene follows Poisson expression

  $Y|\ \lambda \sim \text{Poisson}(\lambda N)$
- Then marginally, Y follows negative binomial distribution (NB)

  $E(Y)=E\{\ E[Y|\ \lambda]\ \}=E\{\lambda N\}= N\mu$

  $Var(Y)=E\{\ var[Y|\ \lambda]\ \} + Var\{\ E[Y|\ \lambda]\ \}$

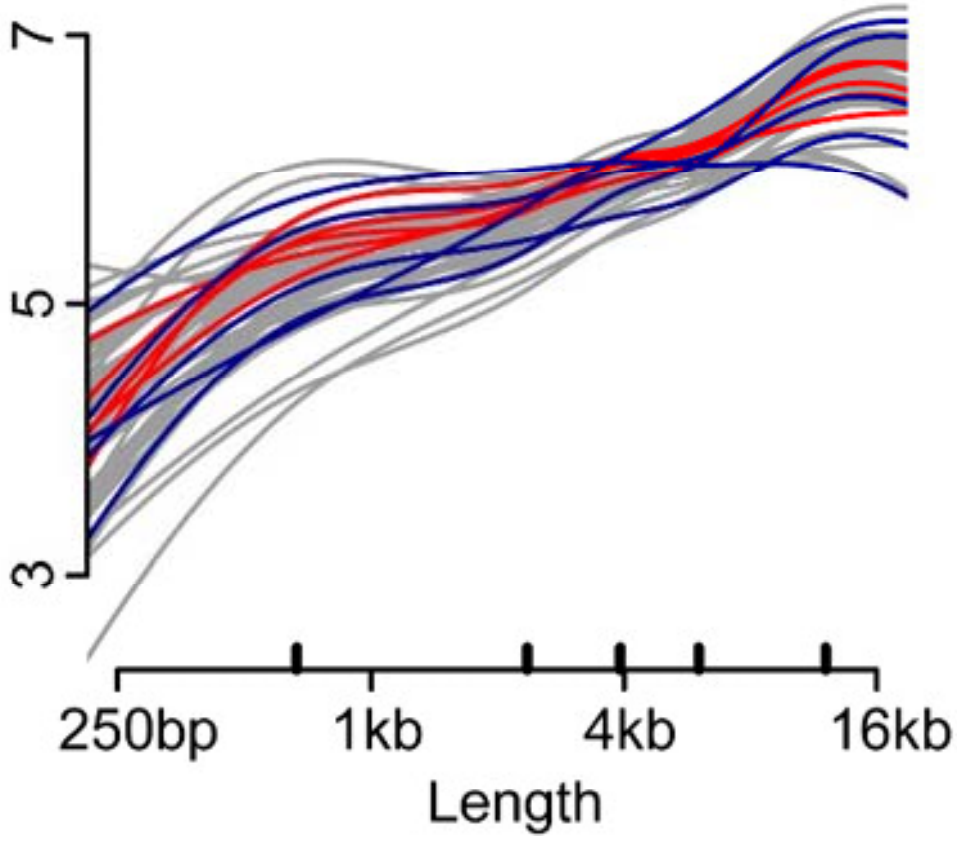
  $=E\{N\lambda\} +Var\{N\lambda\} = N\mu + (N\mu)^2/\alpha = N\mu + (N\mu)^2\ \phi$

$\Phi$is often referred to as the “dispersion parameter”

Size factor

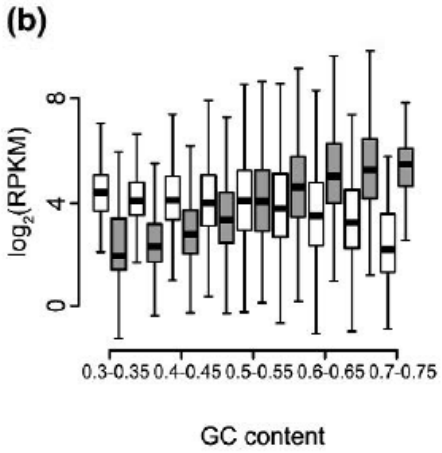
- N is a normalization number that reflects the factors that affect the gene counts that is not due to the variation of gene expression
  - One simple option is the library size (number of reads)
  - Number of mapped reads
  - Other scaling factors
    - Bullard(2010) BMC Bioinformatics (“upper quartile”)
    - Robinson (2010) Genome Biology (“TMM”)

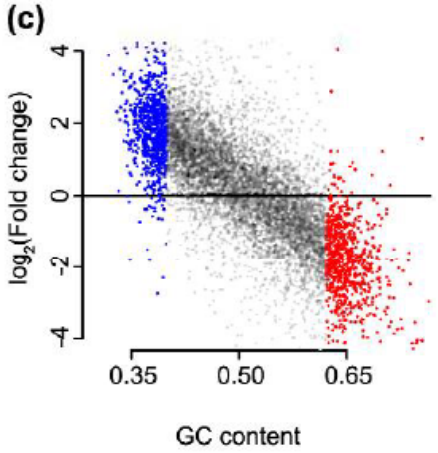
Length Bias

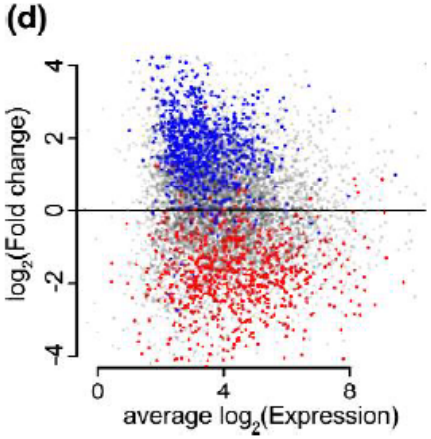


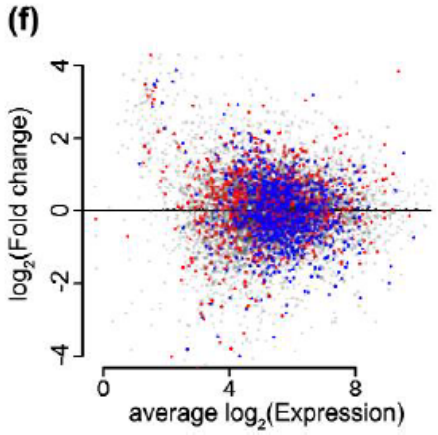
GC bias

Sample-specific
GC content effect









Back to NB distribution

For two class comparison, i=1 vs i=2, for gene j, dispersion ϕ

$$Y_{ij} \sim \mathrm{NB}(\mu_{ij}, \phi)$$

$$E(Y_{ij}) = \mu_{ij} \text{ and } Var(Y_{ij}) = \mu_{ij}(1 + \mu_{ij}\phi)$$

$$\mu_{ij} = m_{ij}\lambda_i$$

$$H_0: \lambda_1 = \lambda_2$$

ϕ is the parameter representing biological variation, and it affects the test result.

When we have a large number of replicates, estimating ϕ from each gene is not difficult. But we face the same problem in estimating gene specific variance as we did in microarray data: limited N.

Empirical Bayes methods

- Common dispersion model
  - Robinson and Smyth (2007) Biostatistics
  - Shrink all genes towards a common dispersion
- Dispersion depends on mean
  - Anders and Huber(2010) *DESeq*
- Weighted conditional likelihood

Borrow information across genes

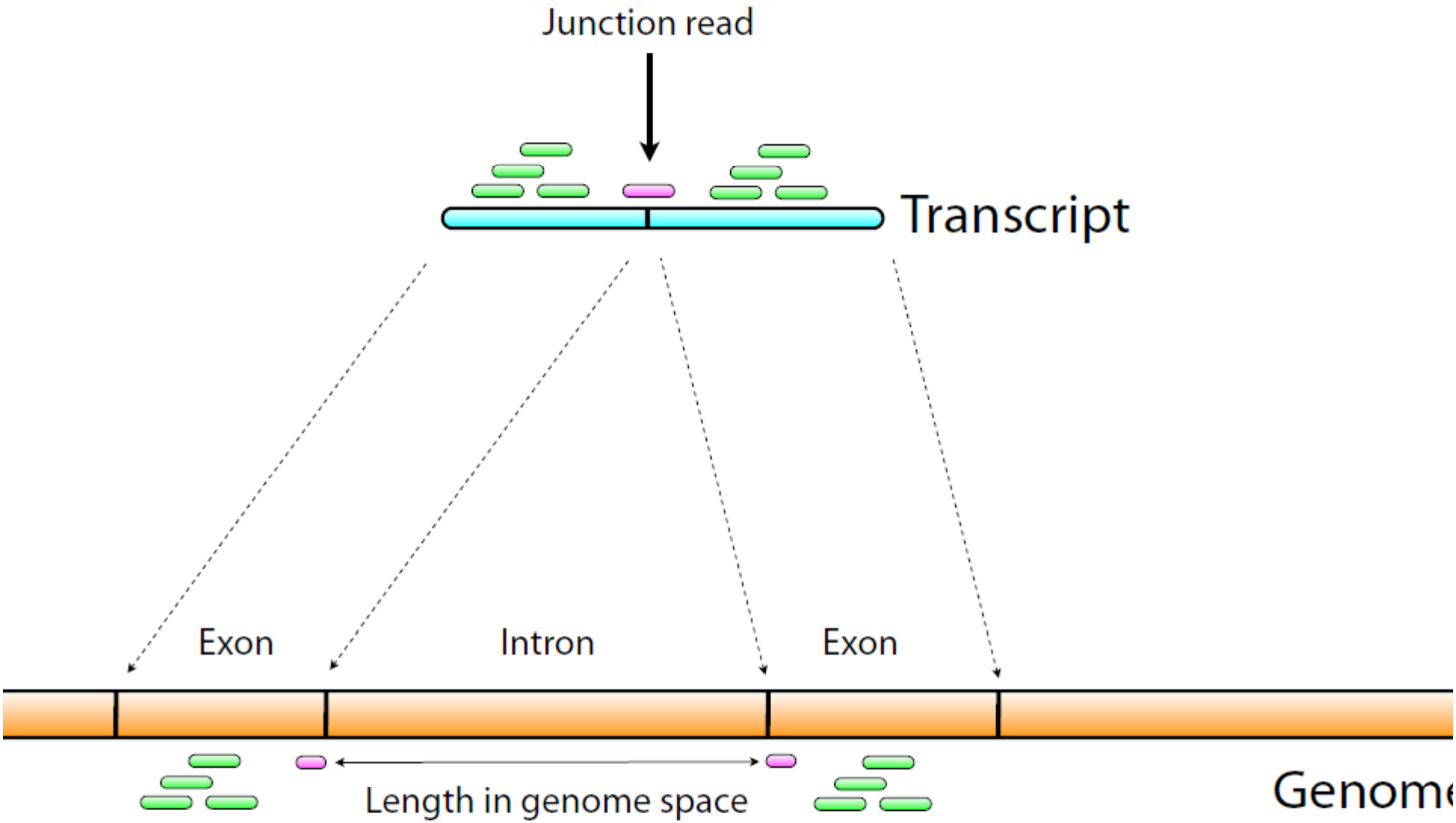
$$\hat{\phi}_g|\phi_g \sim N(\phi_g, \tau_g^2), \quad \phi_g \sim N(\phi_0, \tau_0^2), \quad g = 1, \ldots, G.$$

$$\hat{\phi}_g^B = E(\phi_g|\hat{\phi}_g) = \frac{\hat{\phi}_g/\tau_g^2 + \phi_0/\tau_0^2}{1/\tau_g^2 + 1/\tau_0^2}$$

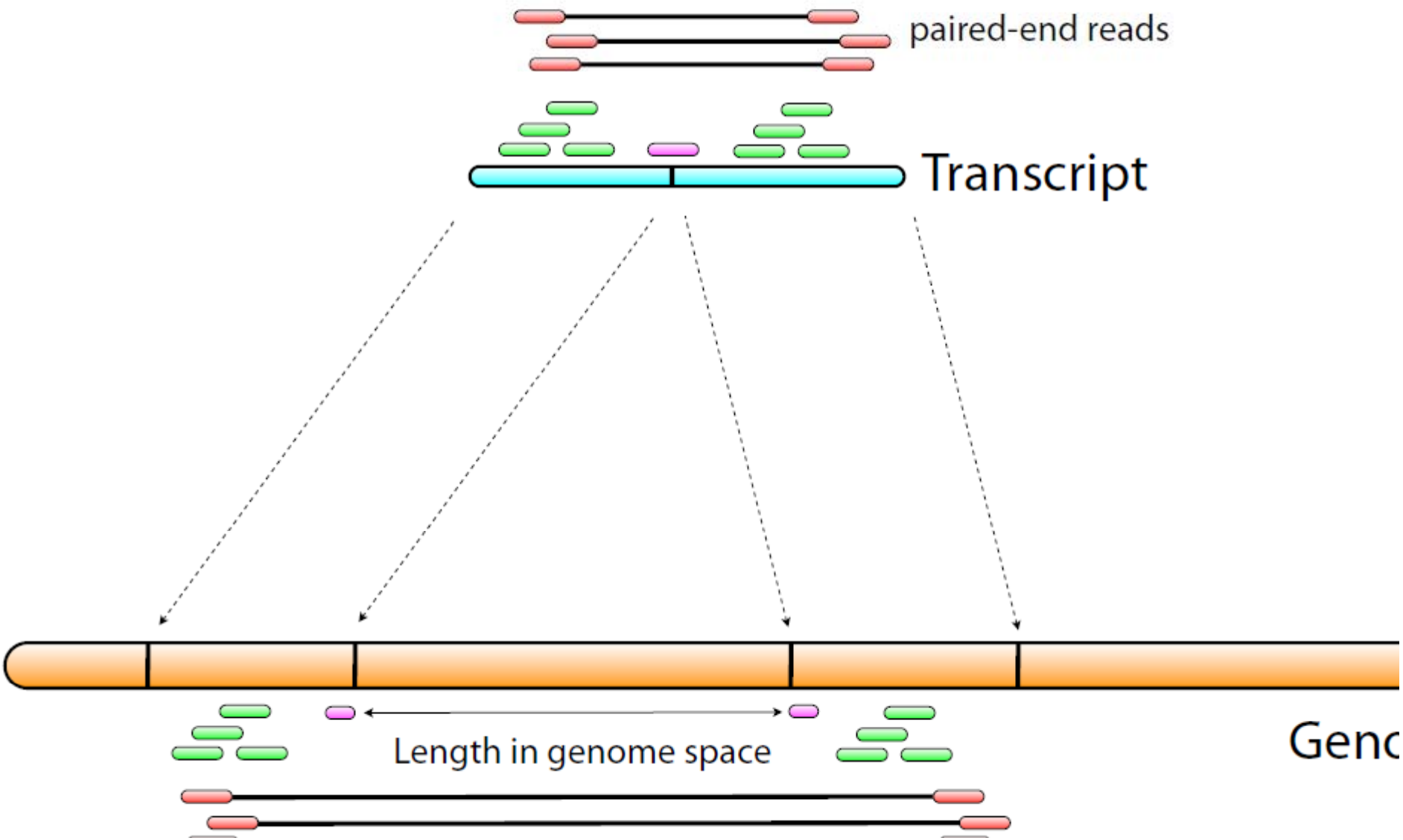
Where RNAseq is more powerful and analysis more challenging

- The definition of a gene/transcript
- Splicing and Isoforms

Mapping transcripts



Mapping transcripts



R packages for RNAseq

- For normalization
  - CQN
- For differential expression
  - EdgeR, DESeq, DSS, Bayseq
- For manipulating sequences, visualizing
  - Biostrings, Bsgenome, GenomicFeatures