RNA-seq differential expression analysis

bioconnector.org/workshops

Agenda

- Our data: source, pre-processing, structure
- Importing & exploring data
- Processing and analysis with DESeq2
    - Structuring the count data and metadata
    - Running the analysis
    - Extracting results
- Data visualization
- Alternative approaches

What this class is *not*

- This is ***not*** an introductory R class. Pre-requisites:
  - Basic R skills: data frames, packages, importing data, saving results
  - Manipulating data with dplyr and %>%
  - Tidy data & advanced manipulation
  - Data Visualization with ggplot2
- This is ***not*** a statistics course.
- This is ***not*** a comprehensive RNA-seq theory/practice course. Refer to the Conesa 2016 and Soneseson 2015 references on the workshop website.
  - We only discuss a simple 2-group design (treated vs. control).
  - Complex designs, multifactorial experiments, interactions, batch effects, etc.
  - Transcriptome assembly & reference-free approaches
  - Upstream analysis...

What this class is *not*

- **This class does *not* cover upstream pre-processing.**
- Sequence read QA/QC
- Our quantitation path: (Kallisto/Salmon + txImport):
  - "Alignment-free" transcript abundance estimation
  - Gene-level abundance summarization
- Alternative path 1 (STAR + featureCounts):
  - Spliced alignment to genome
  - Counting reads overlapping exons
- Alternative path 2 (~~Tophat+Cufflinks~~; HISAT+StringTie):
  - Spliced alignment to genome
  - Transcriptome assembly
  - Transcript abundance estimation

Course website: bioconnector.org

- Data
- Setup instructions
- Lessons dropdown: *RNA-seq: airway*
- **?** dropdown: FAQs, resources, etc.

Our data: Background

- Himes et al. "RNA-Seq Transcriptome Profiling Identifies CRISPLD2 as a Glucocorticoid Responsive Gene that Modulates Cytokine Function in Airway Smooth Muscle Cells." *PLoS ONE*. 2014 Jun 13;9(6):e99625. PMID: 24926665.
- Glucocorticoids inhibit inflammatory processes, used to treat asthma because of anti-inflammatory effects on airway smooth muscle (ASM) cells.
- RNA-seq to profile gene expression changes in 4 ASM cell lines treated w/ dexamethasone (synthetic glucocorticoid).
- Results: many differentially expressed genes. Focus on CRISPLD2
  - Encodes a secreted protein involved in lung development
  - SNPs in CRISPLD2 in previous GWAS associated w/ inhaled corticosteroid resistance and bronchodilator response in asthma patients.
  - Confirmed the upregulated CRISPLD2 w/ qPCR and increased protein expression w/ Western blotting.
- They analyzed with Tophat and Cufflinks. We're taking a different approach with DESeq2. See recommended reading and resources page for more info.

Data pre-processing

- Analyzing RNA-seq data starts with sequencing reads.
- Many different approaches, see references on class website.
- Our workflow (previously done):
  - Reads downloaded from GEO (GSE:GSE52778)
  - Quantify transcript abundance (*kallisto*).
  - Summarize to gene-level abundance – length-scaled counts (*txImport*).
- Our starting point is a **count matrix**: each cell indicates the number of reads originating from a particular gene (in rows) for each sample (in columns).

Data structure: counts + metadata

countData

gene	ctrl_1	ctrl_2	exp_1	exp_1
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
…	…	…	…	…

countData is the count matrix
(number of reads coming from
each gene for each sample)

colData

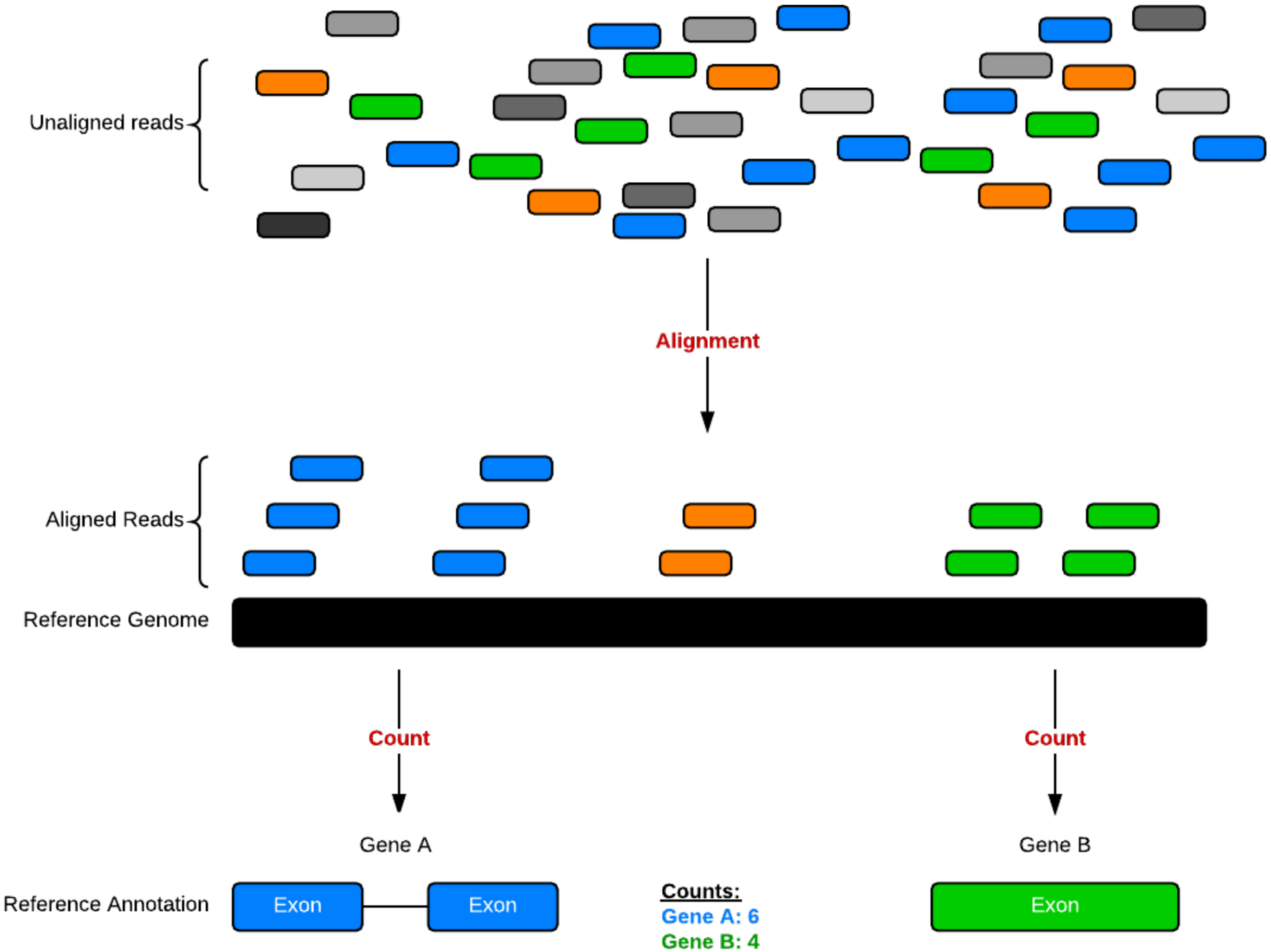
id	treatment	sex	...
ctrl_1	control	male	...
ctrl_2	control	female	...
exp_1	treatment	male	...
exp_2	treatment	female	...

Sample names:
ctrl_1, ctrl_2, exp_1, exp_2

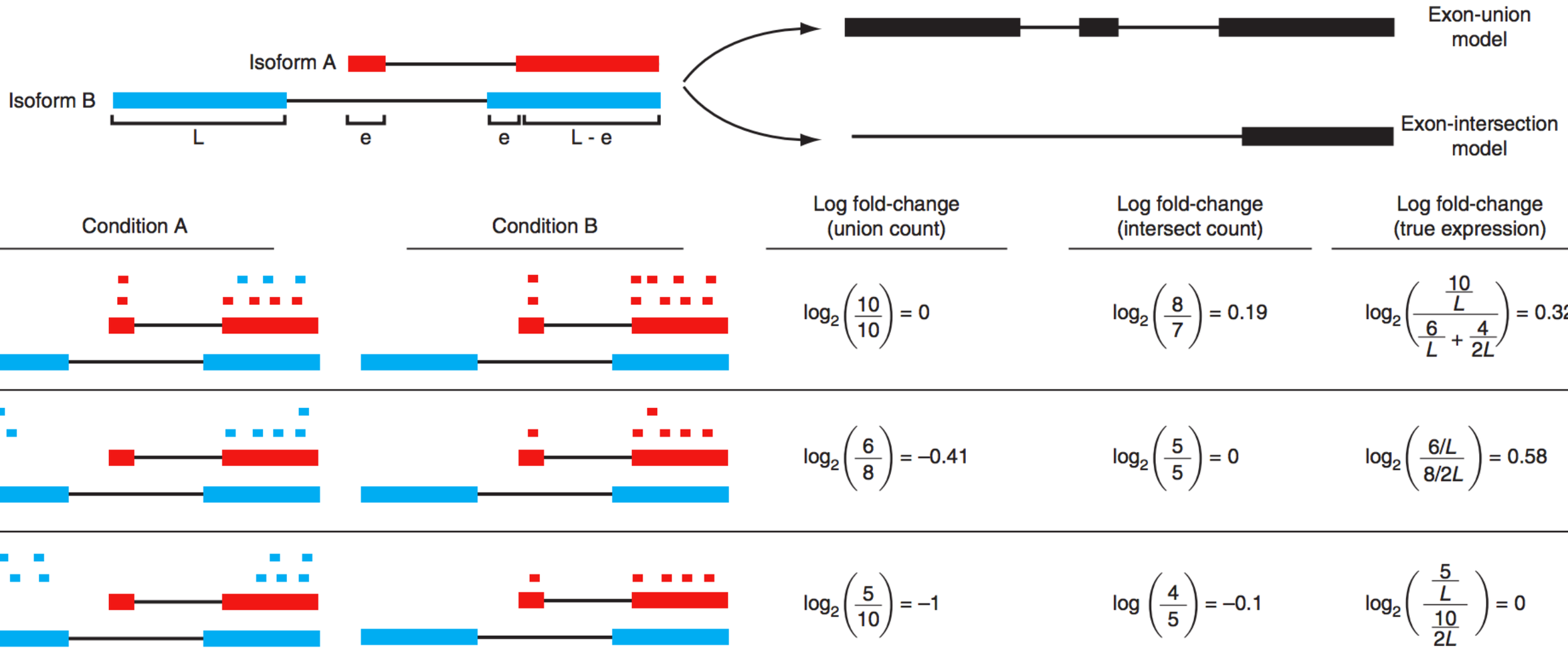
colData describes metadata
about the *columns* of countData

First column of **colData** must match column names of **countData** (-1st)

Counting is (relatively) easy:



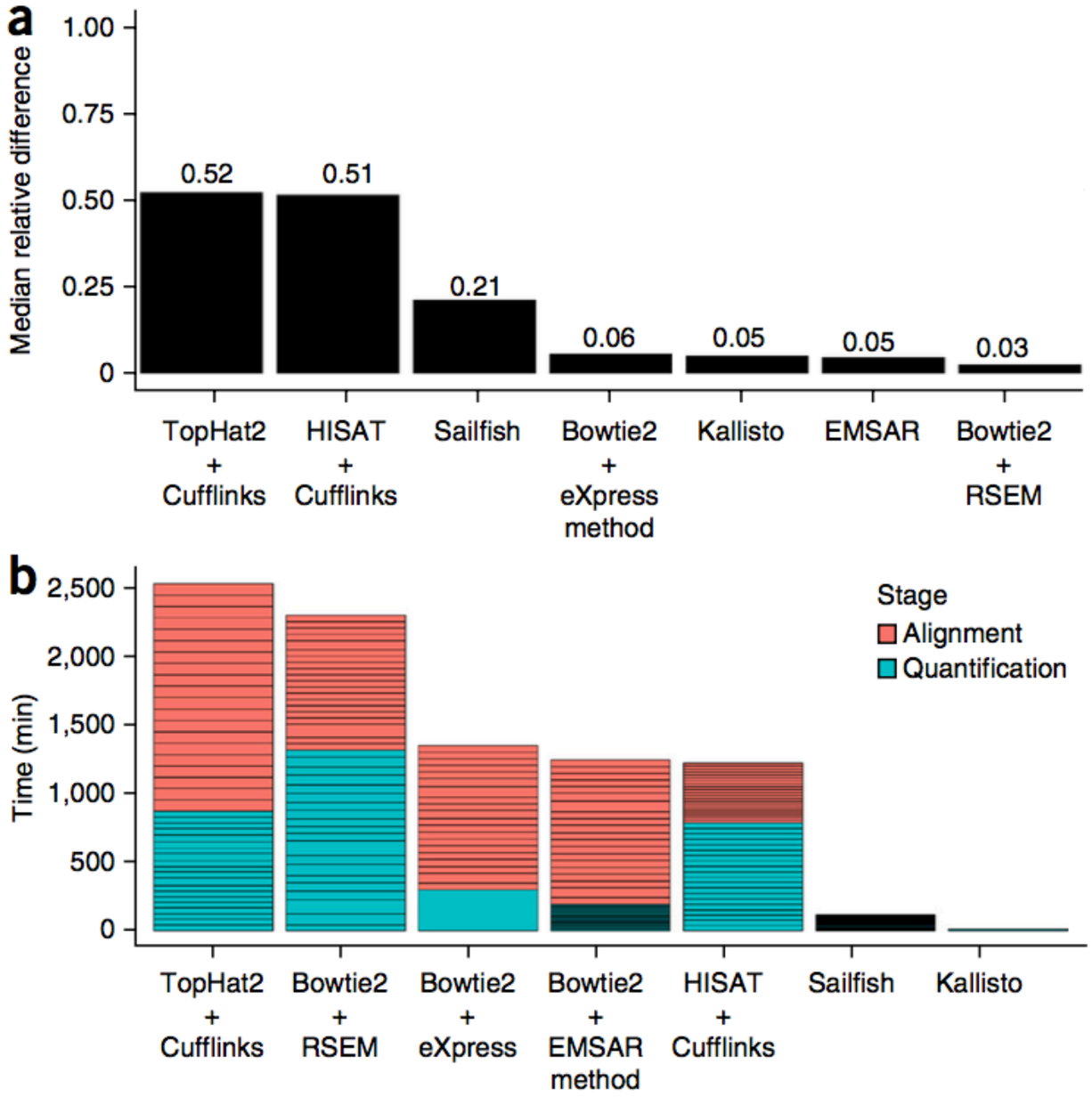
Problem: transcript length bias



Trapnell, Cole, et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq." Nature biotechnology 31.1 (2013): 46-53.

Transcript quantification: kallisto

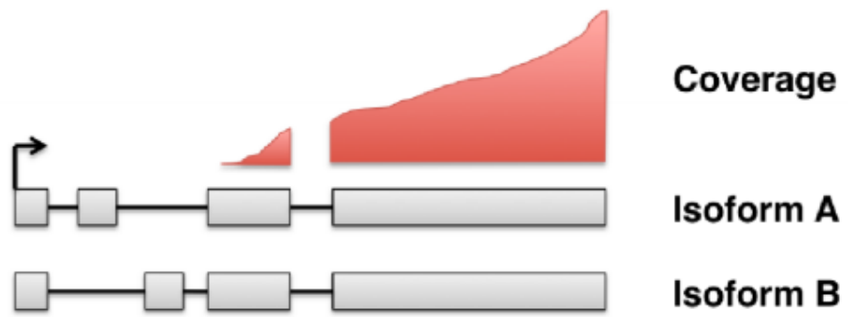
- Don't need a basepair-to-basepair alignment. Only need to know abundance.
- Kallisto determines which transcripts are compatible with the reads (and their abundance).



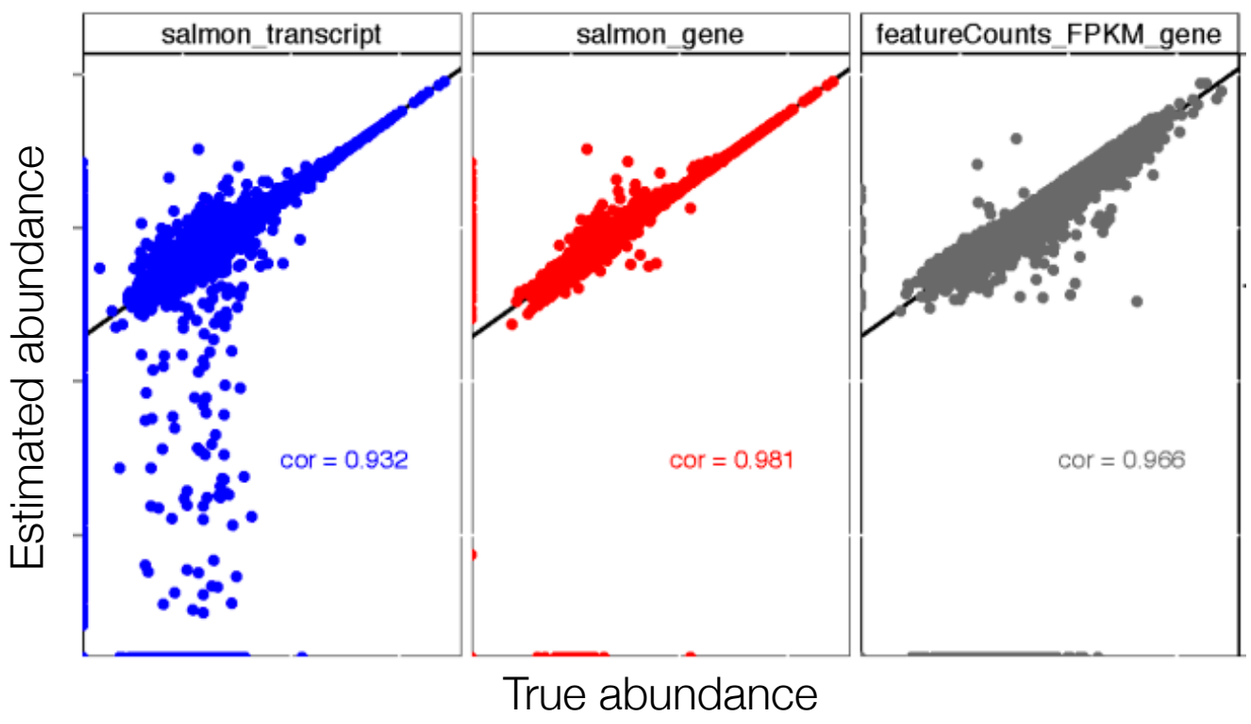
Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature biotechnology, 34(5), 525-527.

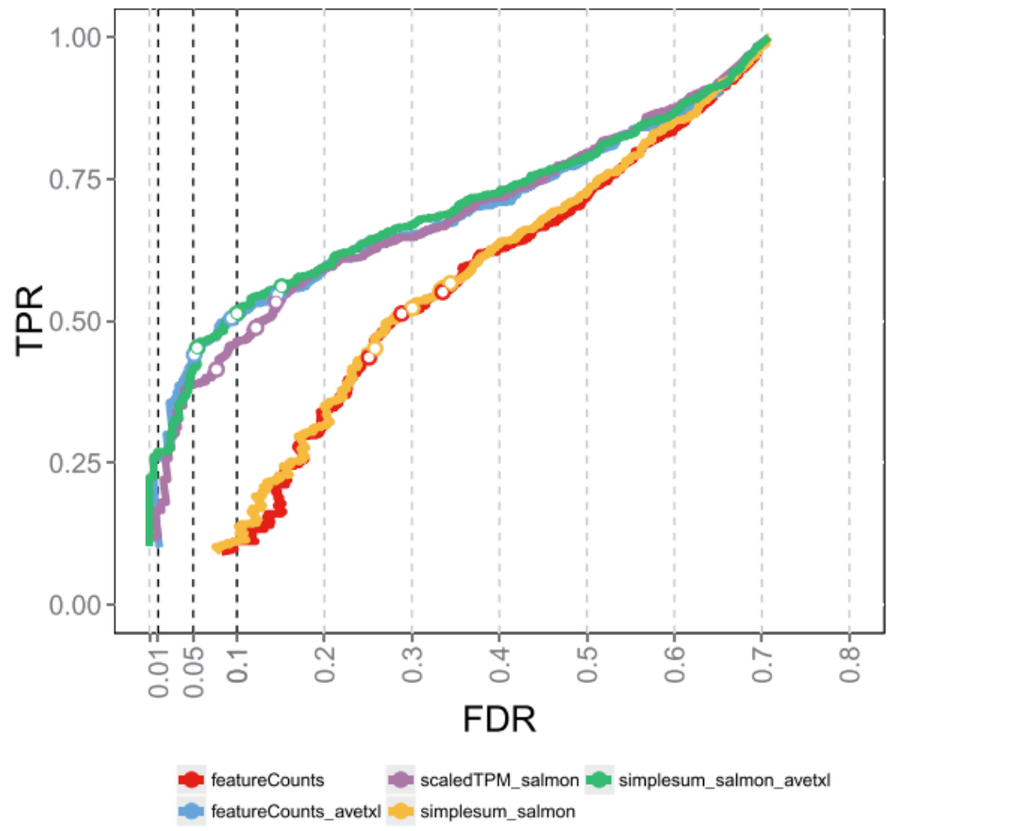
Gene-level summarization: tximport

- Differential gene expression (c/f transcript):
  - More powerful
  - More accurate
  - More interpretable



- Gene-level summaries from transcript abundance estimates are more accurate than simple counts.





Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*.

Getting Started

- Go to **bioconnector.org**. Hit the **data** link on the top navbar. Download the following files, save them somewhere on your computer you can easily find. E.g., create a new folder on your desktop called airway and save it there, or move them to your project directory.
    - **airway_scaledcounts.csv**
    - **airway_metadata.csv**
    - **annotables_grch38.csv**
- Using project management: Open your **.Rproj** file to start R running in the same folder as the data. File ➡️ New file ➡️ R script. Save this file as **airway_analysis.R**.
- Not using project management: Open RStudio. File ➡️ New file ➡️ R script. Save this file as **airway_analysis.R** in the same folder as the data. Quit RStudio, then double-click the R script to open R running in that folder.)
- Load the data:

```
library(tidyverse)
rawcounts <- read_csv("airway_rawcounts.csv")
metadata <-  read_csv("airway_metadata.csv")
```