

# From reads to results: differential expression analysis with RNA-seq

Alicia Oshlack

Head of Bioinformatics

Murdoch Childrens Research Institute

REVIEW

# From RNA-seq reads to differential expression results

Alicia Oshlack<sup>1\*</sup>, Mark D Robinson<sup>1,2</sup> and Matthew D Young<sup>1</sup>

## Abstract

Many methods and tools are available for preprocessing high-throughput RNA sequencing data and detecting differential expression.

High-throughput sequencing technologies are now in common use in biology. These technologies produce millions of short sequence reads and are routinely being applied to genomes, epigenomes and transcriptomes. Sequencing steady-state RNA in a sample, known as RNA-seq, is free from many of the limitations of previous

detection of alternative splicing [10-12], RNA editing [13] and novel transcripts [11,14]. However, the primary objective of many biological studies is gene expression profiling between samples. Thus, in this review we focus on the methodologies available to detect differences in gene level expression between samples. This sort of analysis is particularly relevant for controlled experiments comparing expression in wild-type and mutant strains of the same tissue, comparing treated versus untreated cells, cancer versus normal, and so on. For example, comparison of expression changes between the cultured pathogen *Acinetobacter baumannii* and the pathogen grown in the presence of ethanol - which is

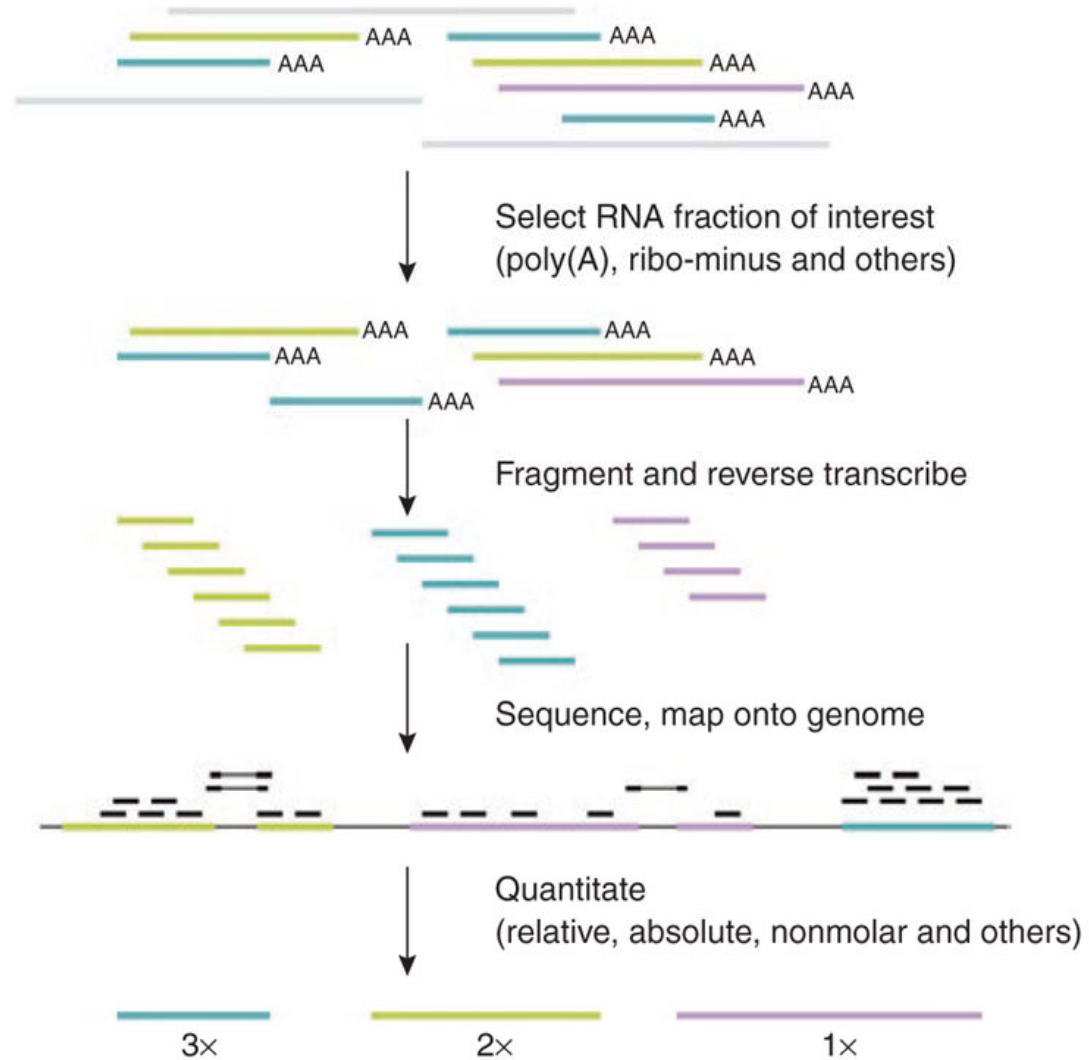
# Benefits and opportunities of RNA-seq

- All transcripts are sequenced not just ones for which probes are designed (cf microarrays)
- Annotation of new exons, transcribed regions, genes or non-coding RNAs
- Whole transcriptome sequencing
  - The ability to look at alternative splicing
  - Allele specific expression
  - RNA editing

# This talk

- Analysis of RNA-seq data for the purpose of determining differential expression
- How much are expression levels changing between samples?

# RNA-seq

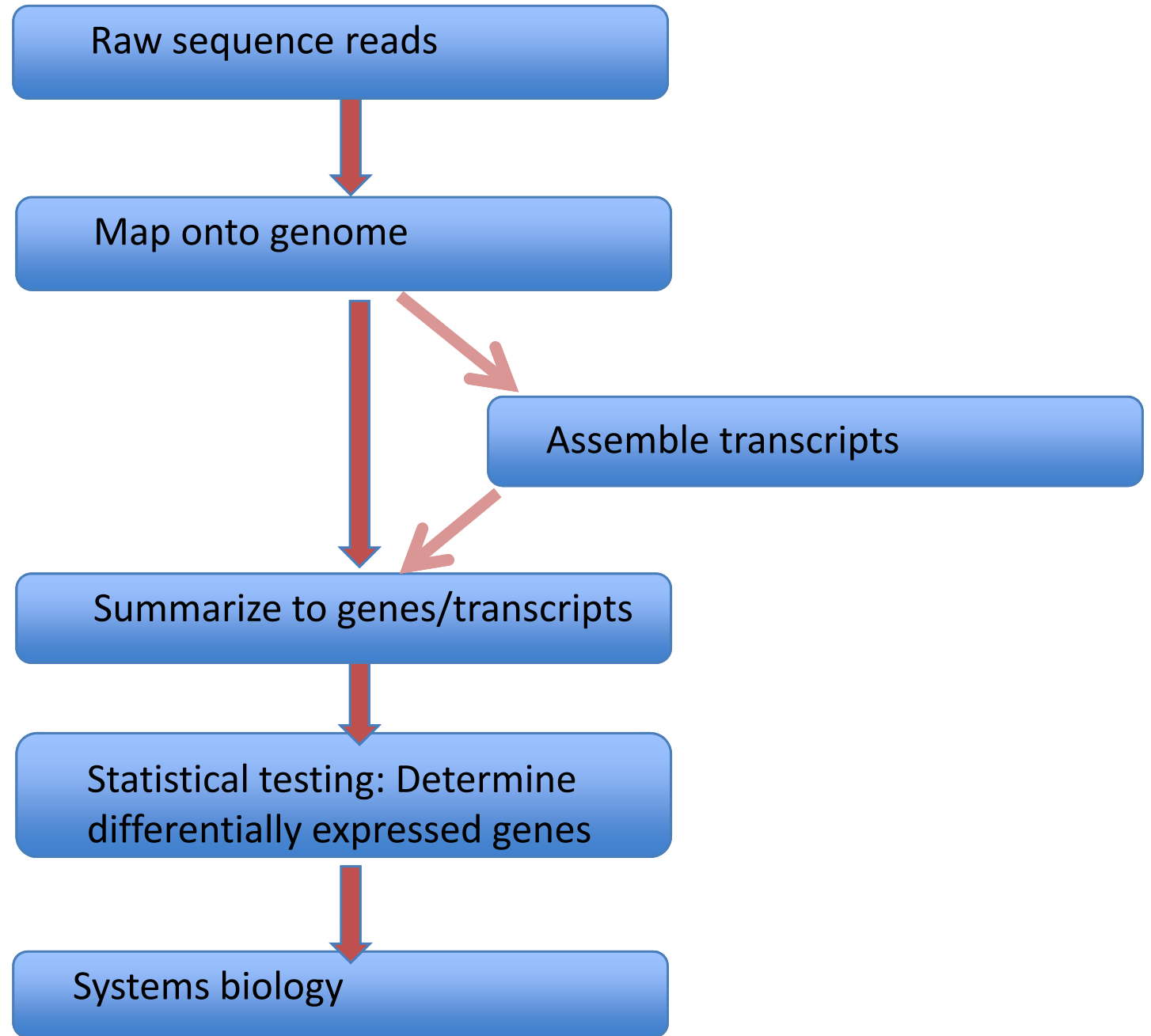


# Raw data

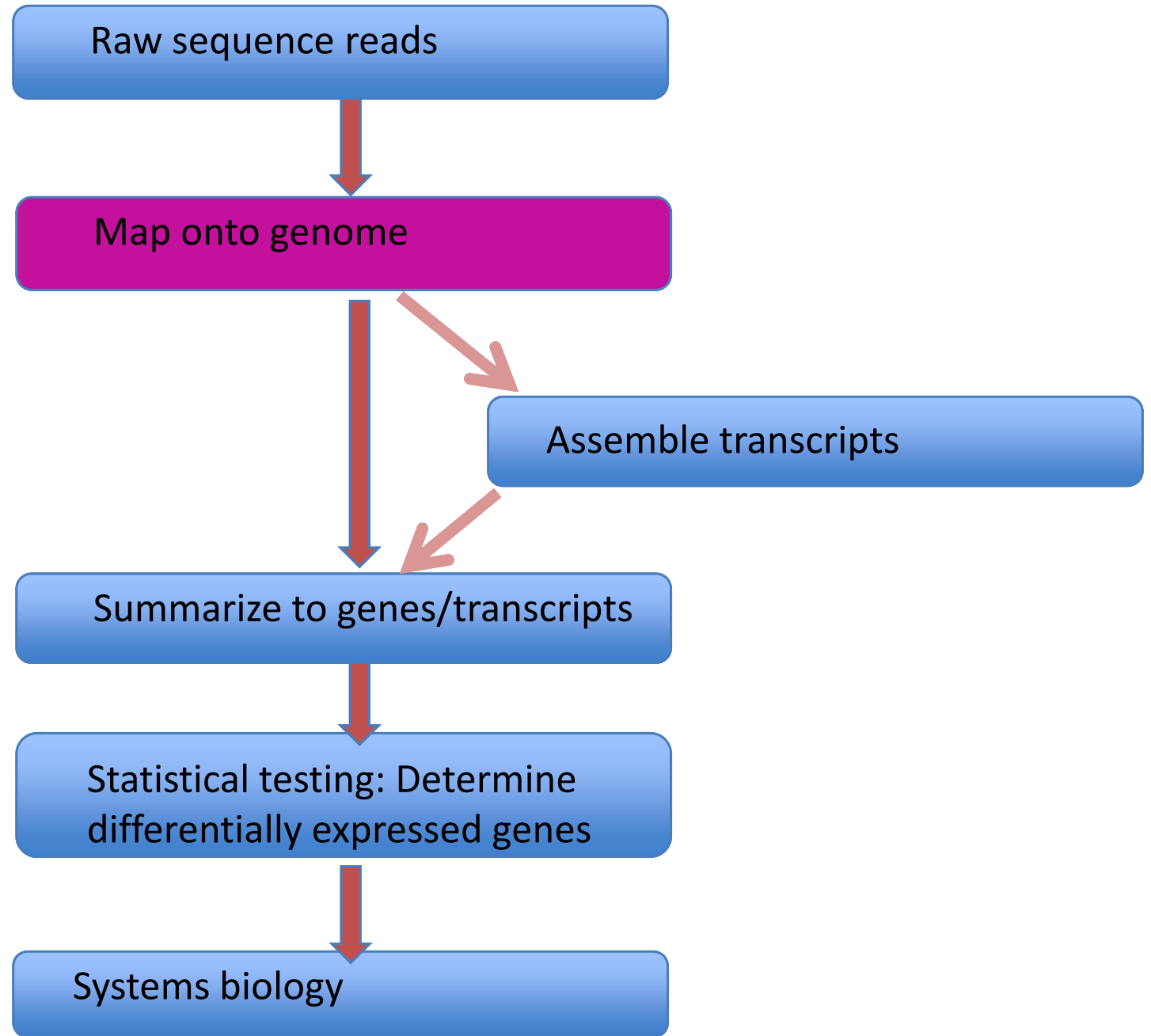
- Short sequence reads
- Quality scores

```
@SEQ ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!' '*((( (***) ) %%%++) (%%%) .1***-+*'') ) **55CCF>>>>>CCCCCCC65
```

# RNA-seq analysis steps

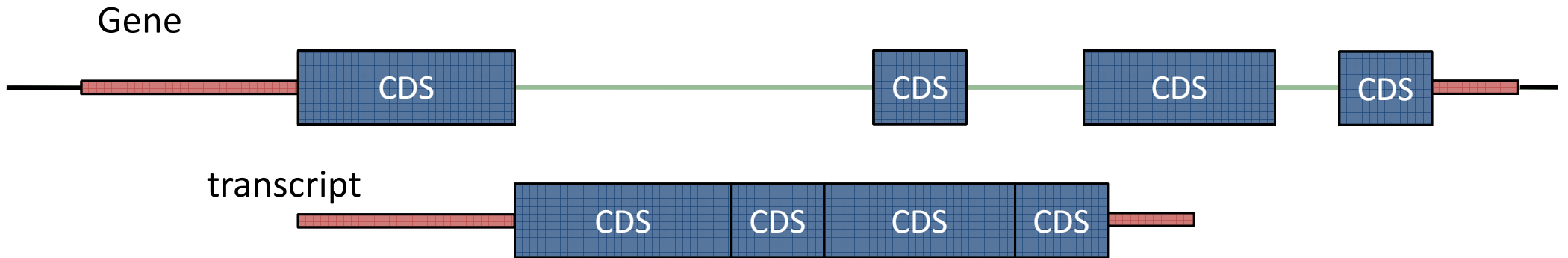


# RNA-seq analysis steps



Mapping: RNA-seq

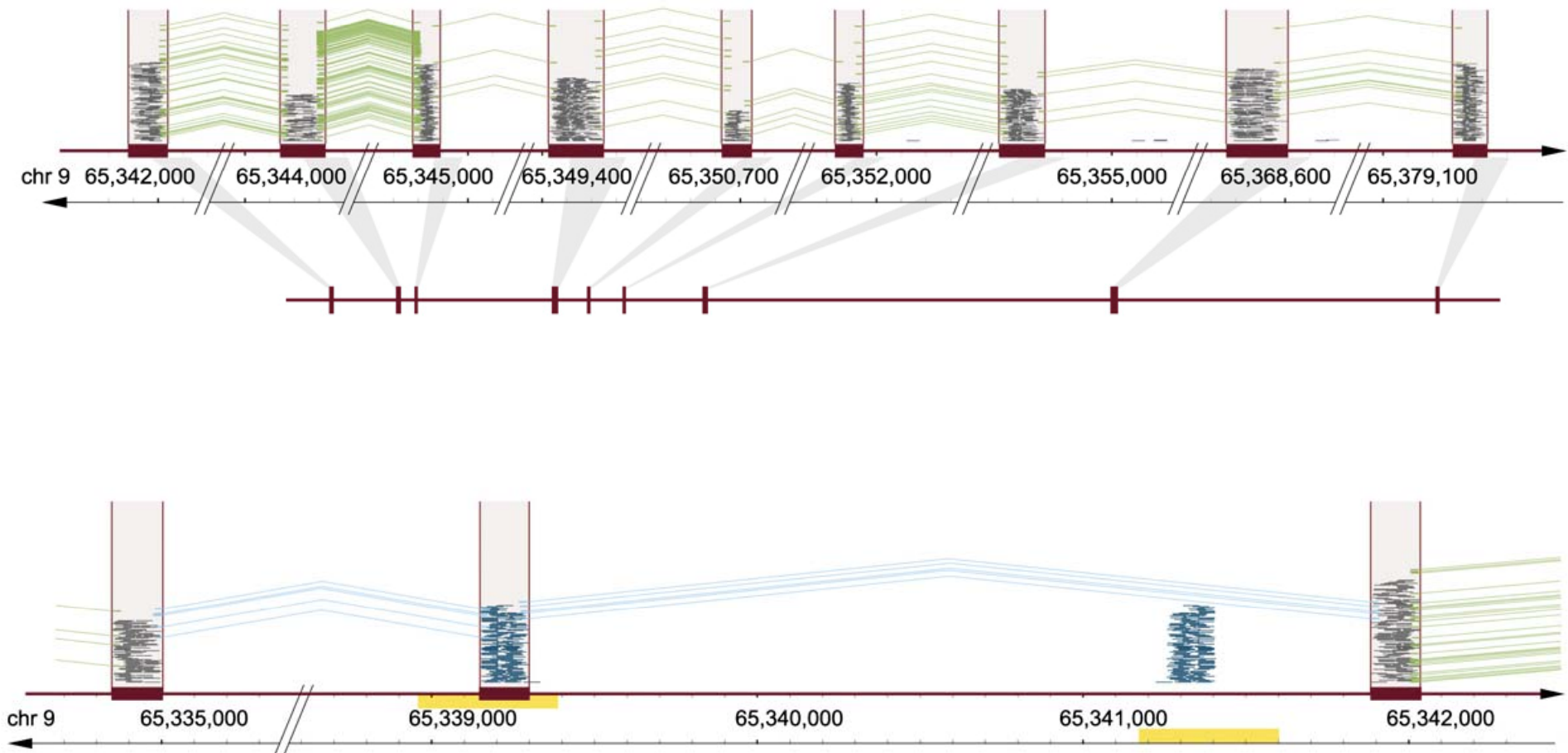
# Sequencing transcripts not the genome



# Specific issues for RNA-seq

- Unlike ChIP-seq or DNA -seq, with RNA-seq we really want to align to the transcriptome.
- The transcriptome is built from the genome, but exon boundary or junction reads will not align to the genome.
- The longer the reads, the more likely a read is to hit a junction.

# The problem



# Option 1: Don't worry about it!

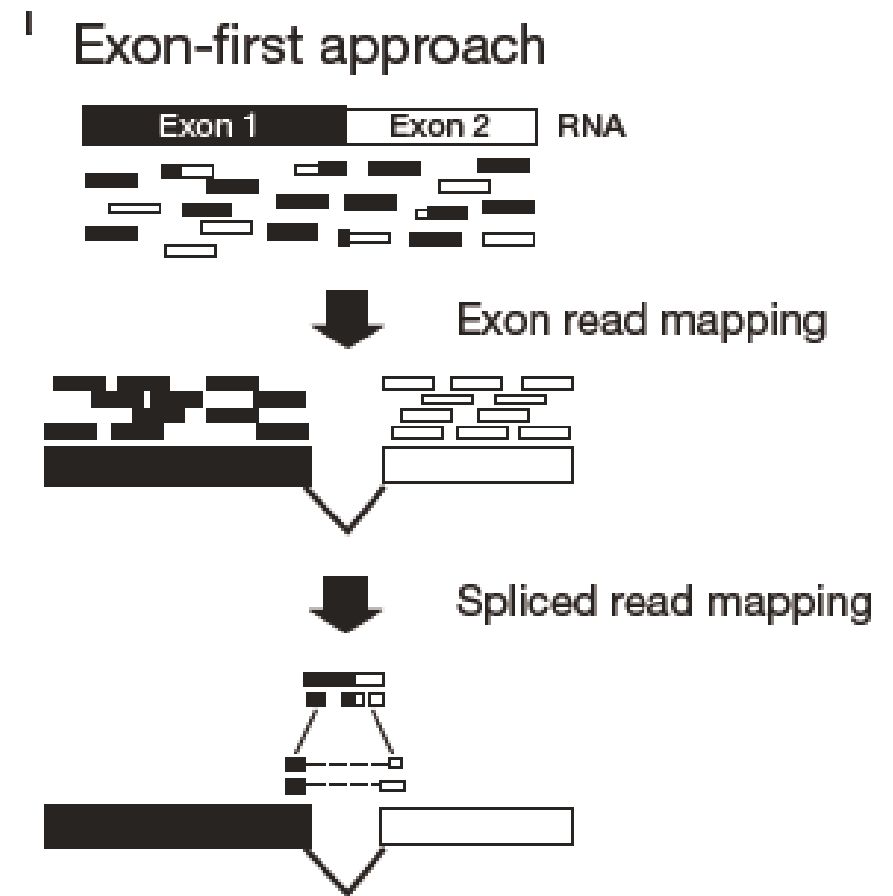
- Map to genome.
- Captures any reads in unannotated exons.
- Can't map any reads that cross exon boundaries.
- Not dependent on any annotation.

# Option 2: Build a junction library

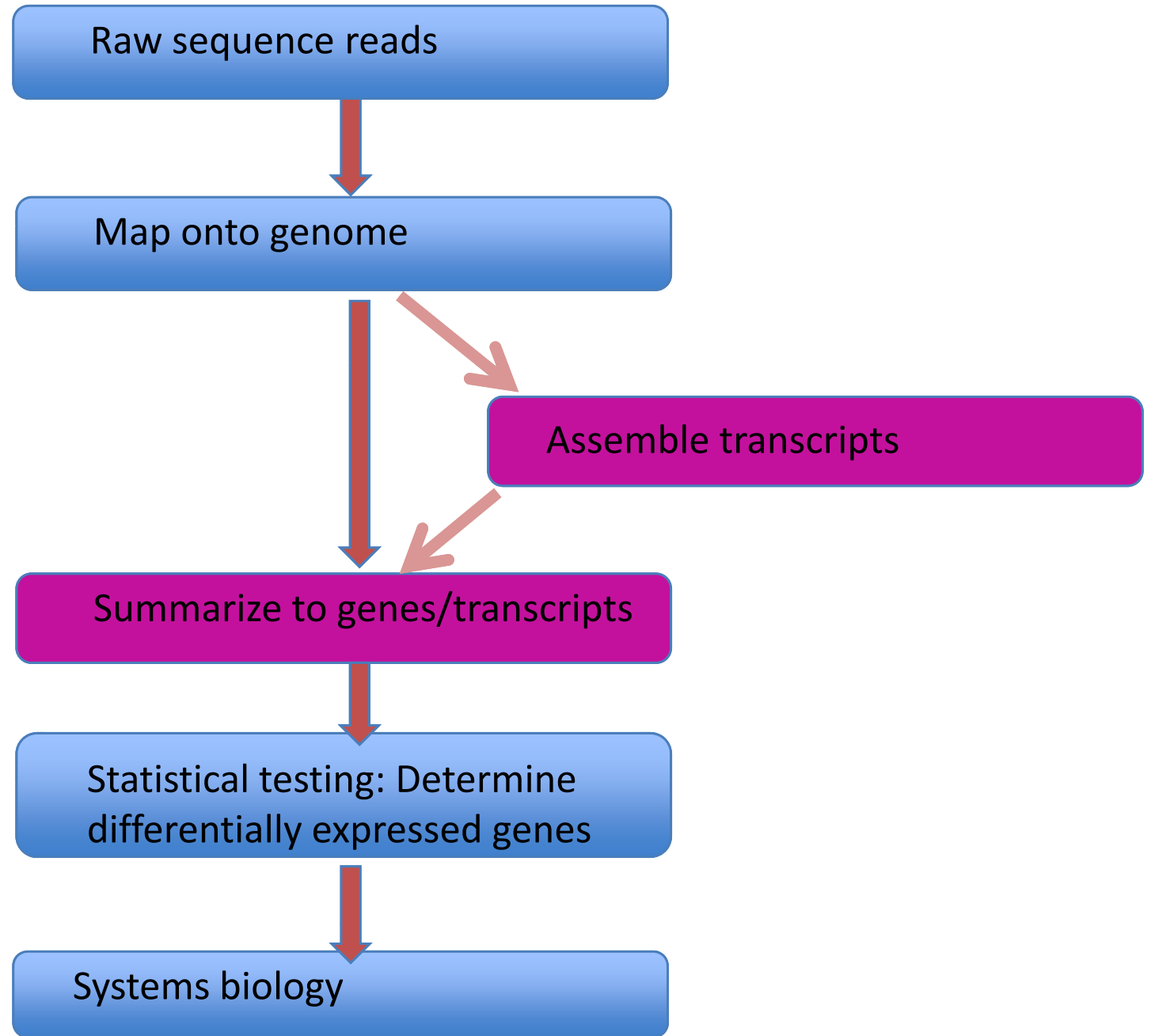
- Make a new reference by combining all known exons.
- Captures some junction reads.
- Biased towards well annotated genes.
- Mapping to the transcriptome is just a special junction library.

# Option 3: splice site discovery

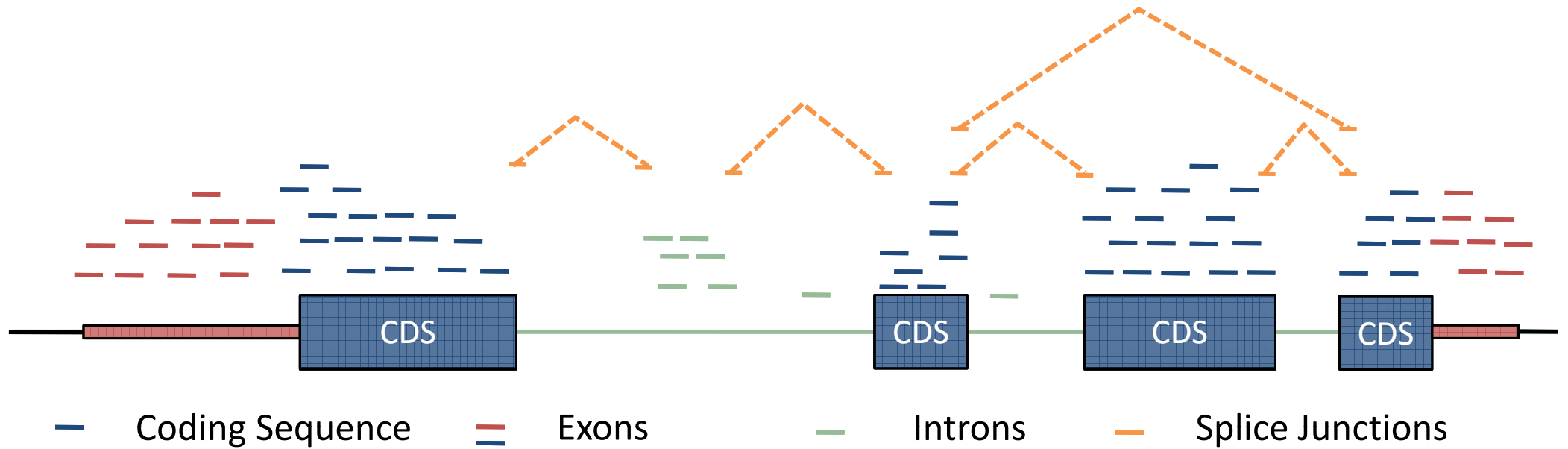
- Truly de novo transcript assembly uses no reference (oases, TransABySS).
- More often we try and determine where splice junctions occur using the data itself.
- Computationally intensive, but unbiased by annotation.
- Several software packages to do this such as TopHat, SplitSeek, SpliceMap...



# RNA-seq analysis steps



# Summarization



- Reads in exons
- Exons + junctions
- All reads start to end of transcript
- De novo methods

Number of  
reads

=

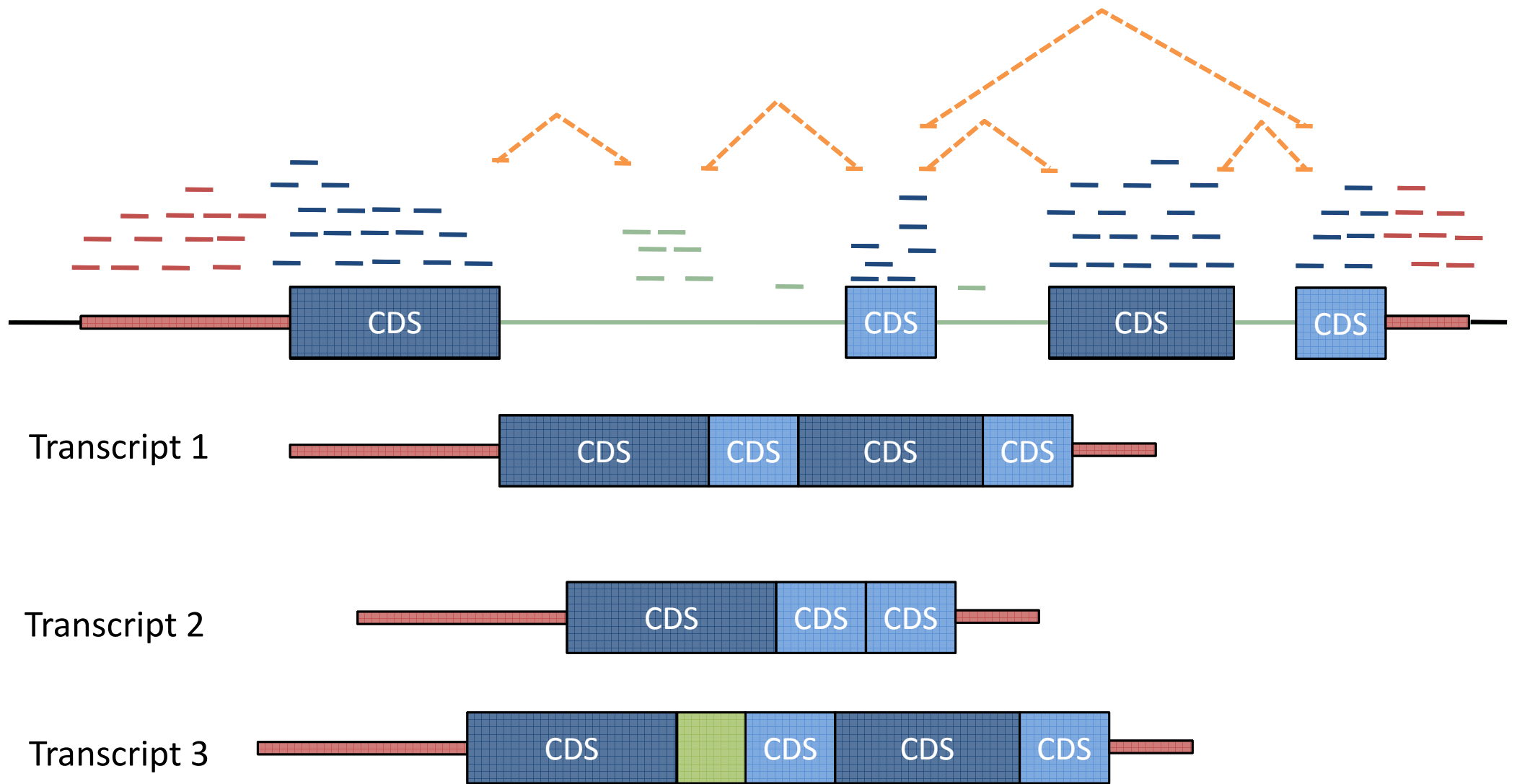
Number of  
transcripts

X

Length of  
transcript

# Transcript assembly

- Take mapped reads, exons and junctions and infer the possible transcripts

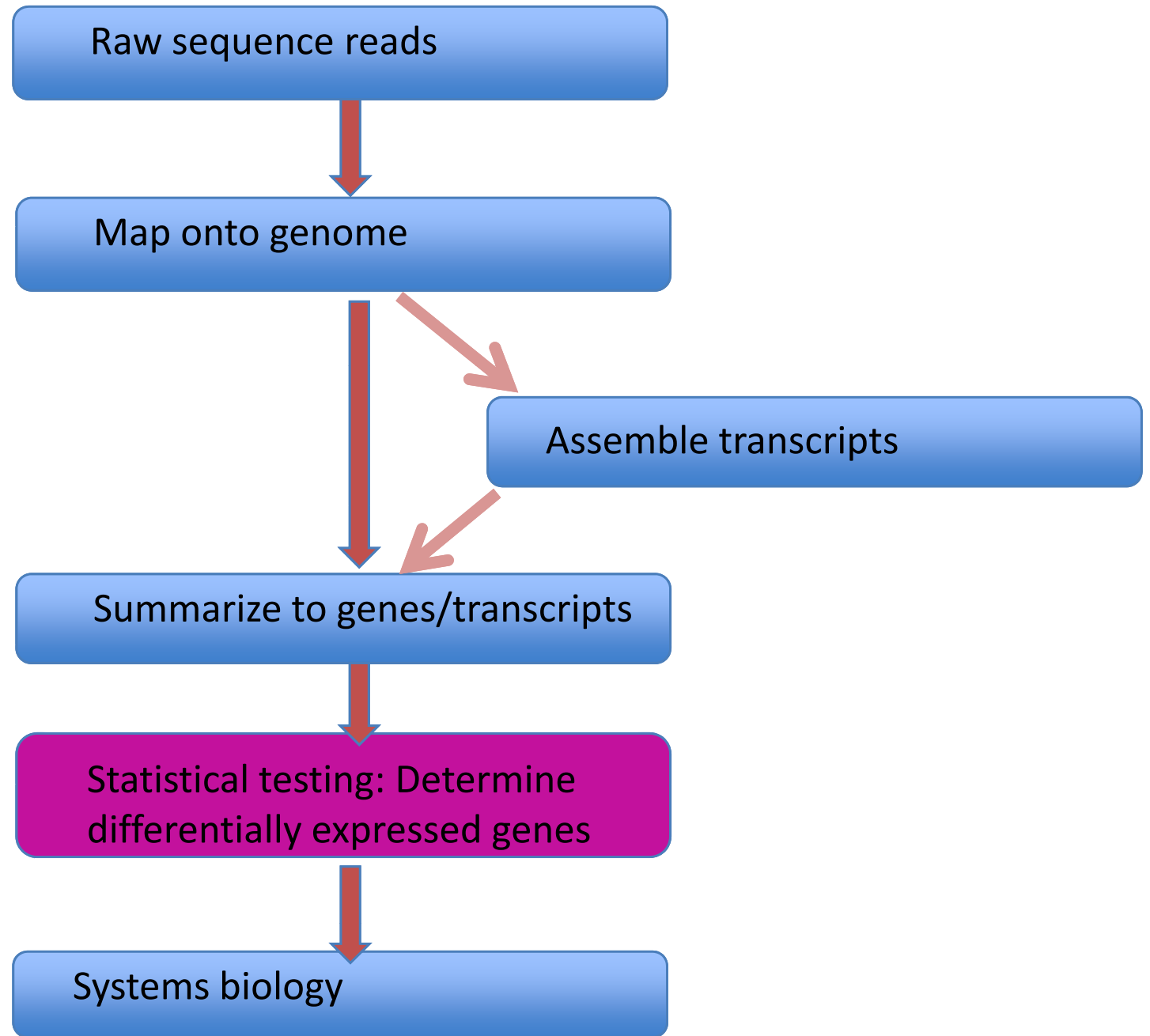


Even when all transcripts are “known” summarization or expression quantification is difficult. How do you assign reads to transcripts?

# Table of counts

gene	Sample A	Sample A	Sample B	Sample B
A	13	14	16	13
B	167	134	1095	1276
C	0	0	1	0
D	22	25	1	1
E	45	36	22	21
F	223	199	56	67
G	1	1	0	0
H	3	4	7	8
I	12	10	25	24
J	0	0	12	32
K	98	123	0	2
L	10023	9845	2365	2144
...				

# RNA-seq analysis steps



# Normalization

Removing technical sources of  
variation

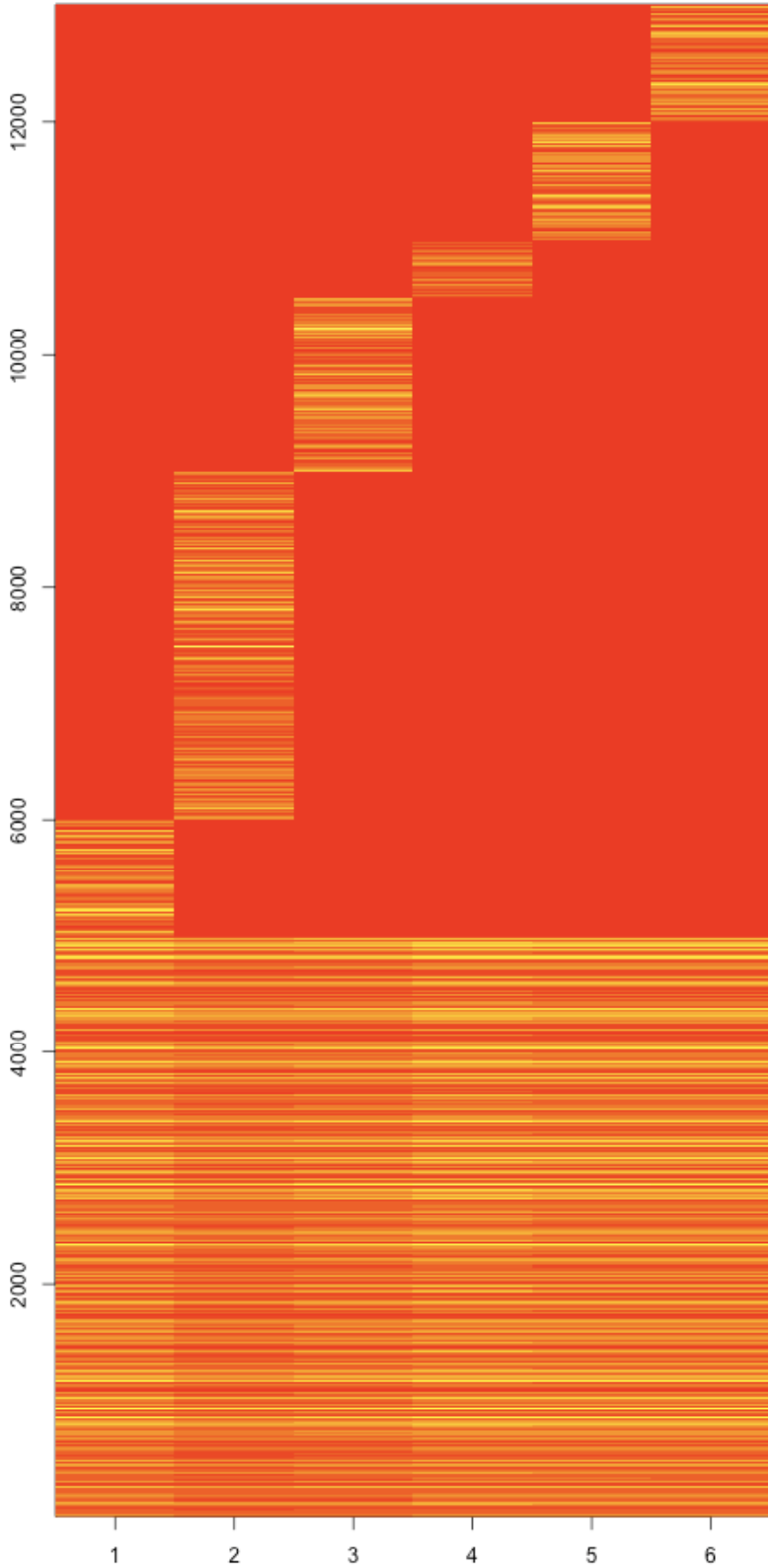
# Simple thought experiment

- Two samples A and B, sequenced to the same depth (same library size)
- Every gene that is expressed in A, is expressed in B at the same level
- Say there are a small number of genes that are expressed uniquely to sample B, but they are quite highly expressed (lots of reads)
- Many genes within the common set will appear differentially expressed ( $B < A$ )

# Another way to view it

- Hypothetical example: Sequence 6 libraries to the same depth, with varying levels of *unique-to-sample* expression
- Differences in observed counts among the common genes

Red=low, goldenyellow=high



# Marioni et al. 2008 RNA-seq data

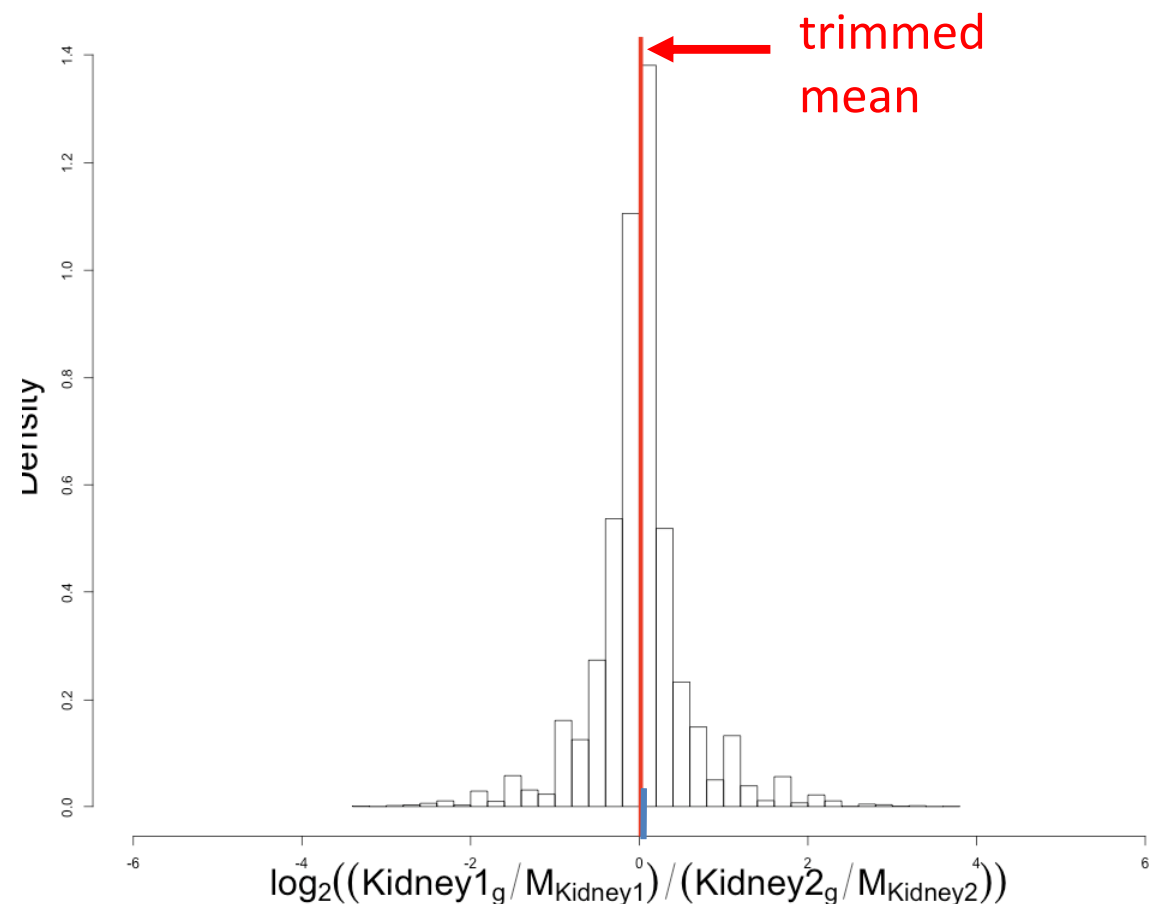
- 5 lanes of liver RNA, 5 lanes of kidney RNA
- Compare two single kidney libraries (technical replicates), after adjusting for library size

Distribution of log-ratios of counts

$M$  = library size

0 counts are omitted

Red line = trimmed mean



# Marioni et al. 2008 RNA-seq data

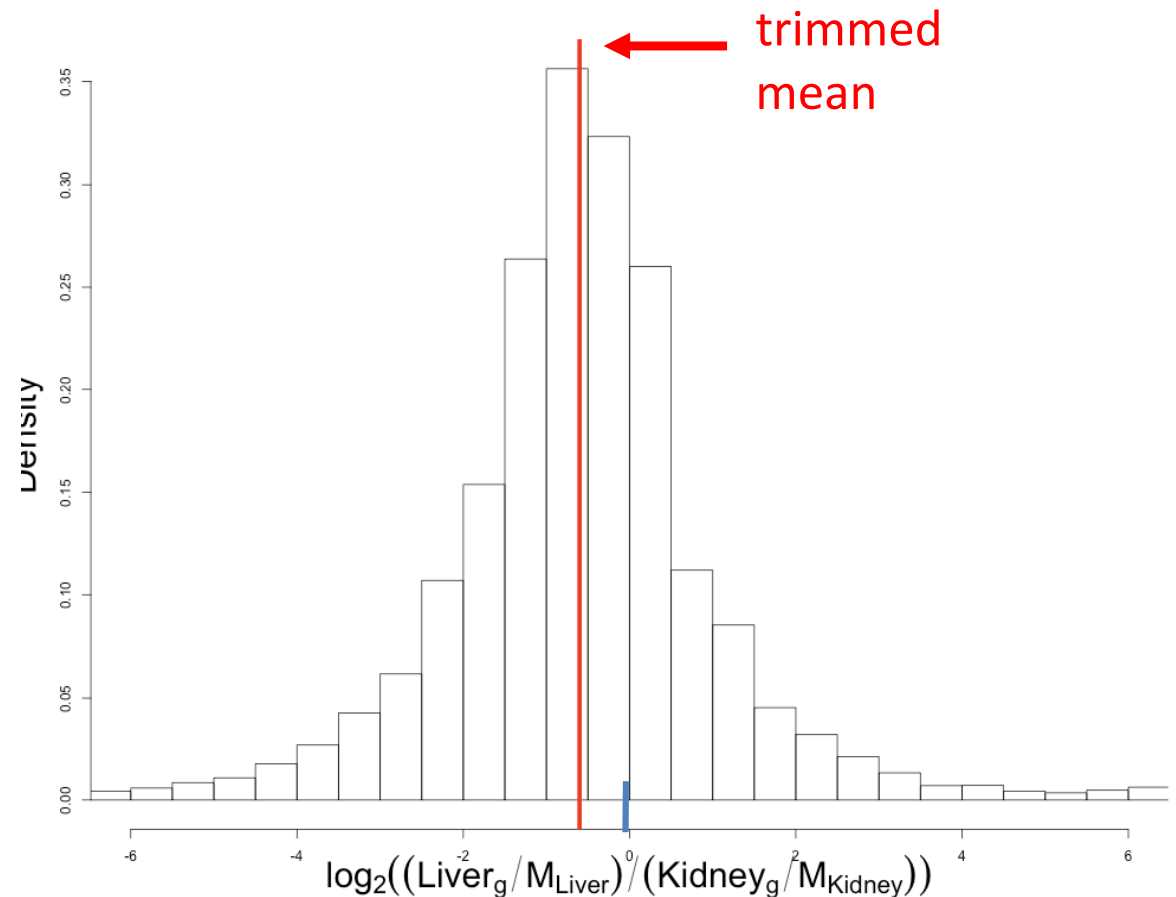
- 5 lanes of liver RNA, 5 lanes of kidney RNA
- Compare one liver to one kidney library, after adjusting for library size

Distribution of log-ratios of counts

$M$  = library size

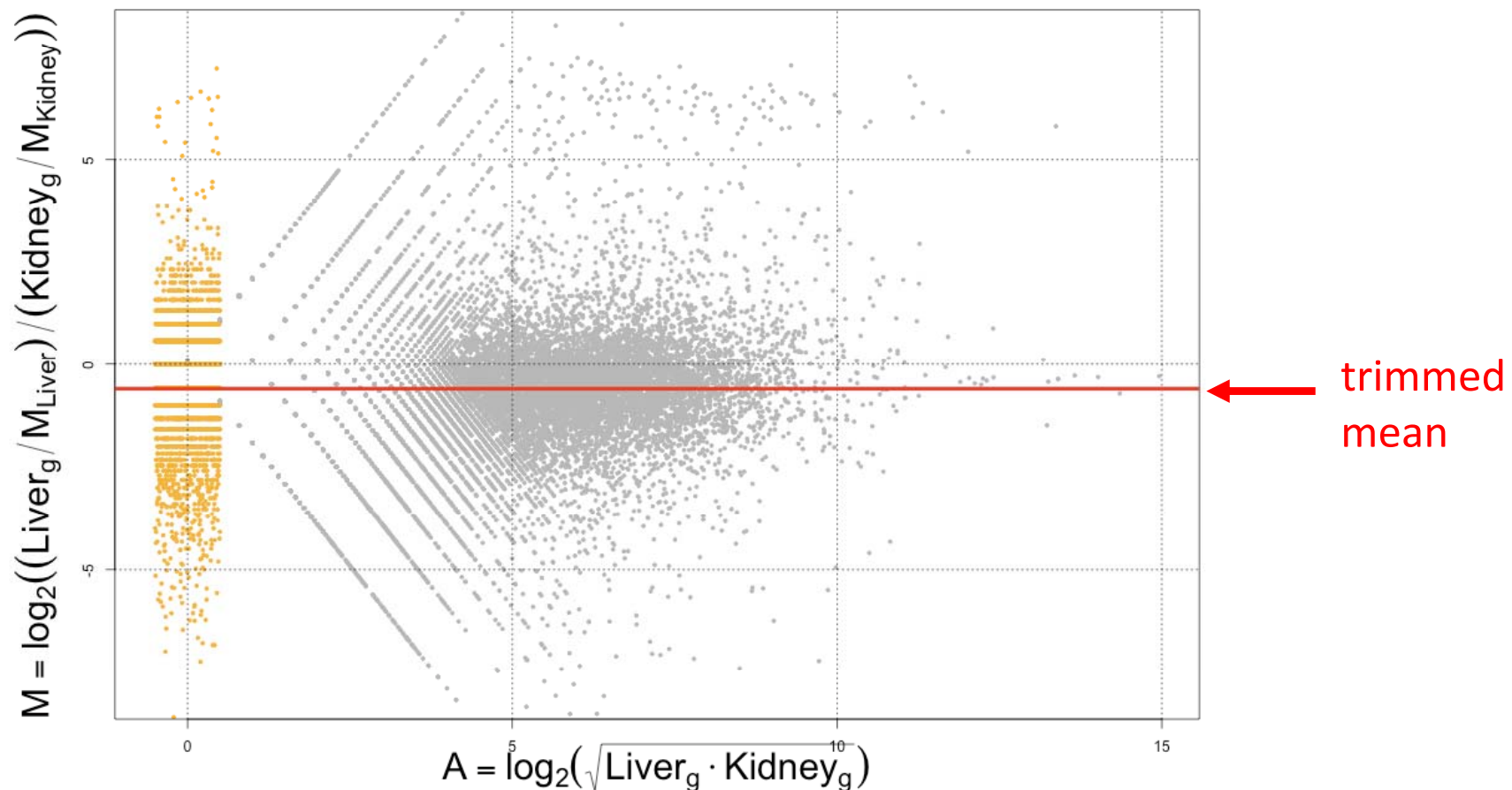
0 counts are omitted

Red line = trimmed mean



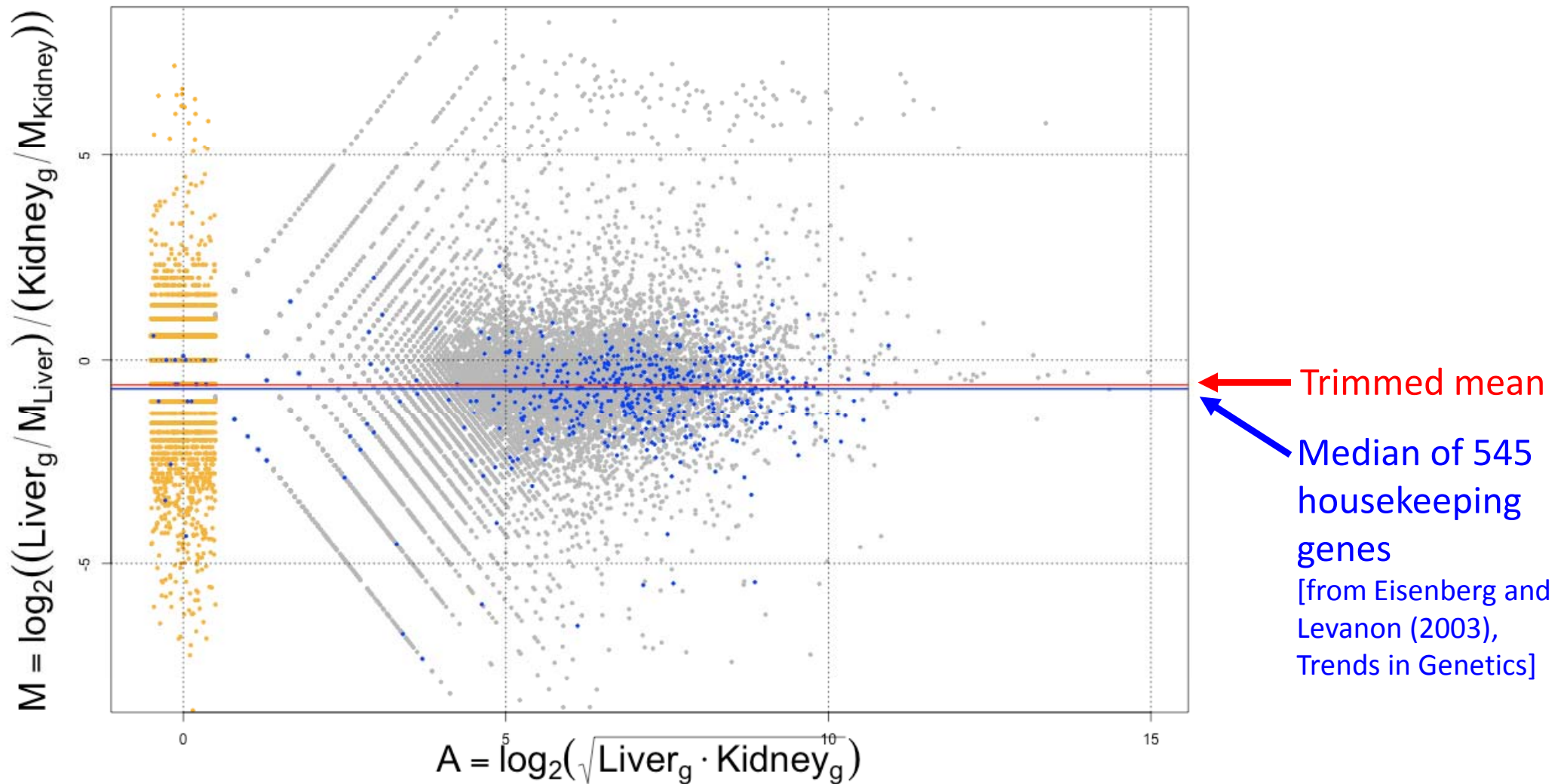
# Should we expect the log-ratio to be 0 after “normalization”?

- I would argue: **yes**.
- Another way to look at it: M vs A plots



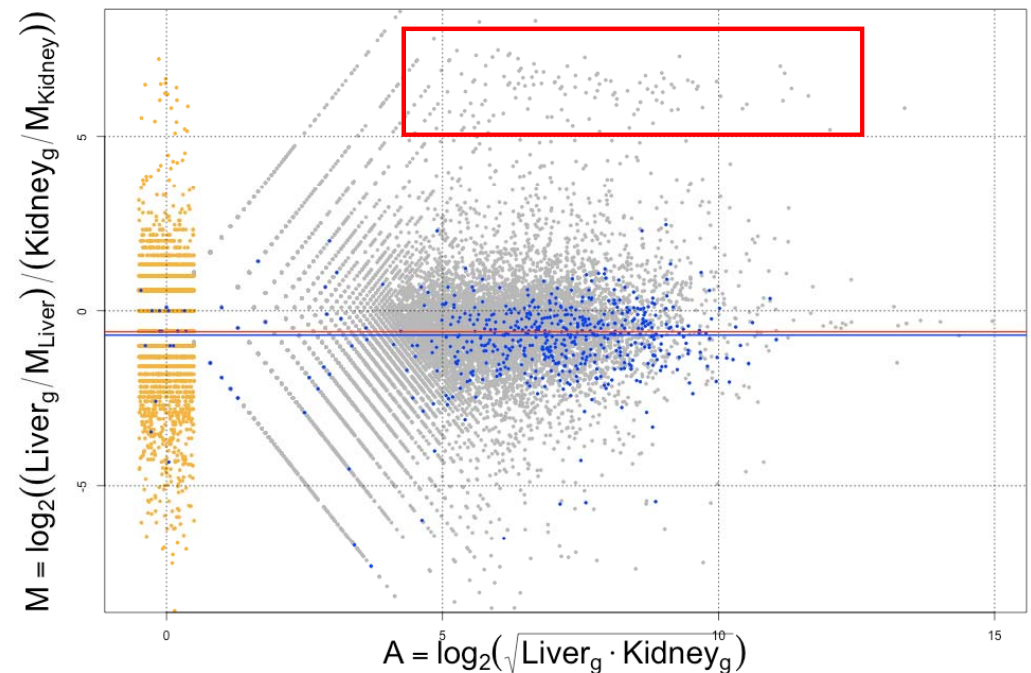
# Should we expect the log-ratio to be 0 after normalization?

- Housekeeping genes shown in blue.



# Shift in log-ratios is caused by RNA composition

- Sequencing “real estate” is fixed.
- Underlying RNA composition can be very different
  - e.g. **several liver-specific genes**
- An adjustment at the analysis stage should be made



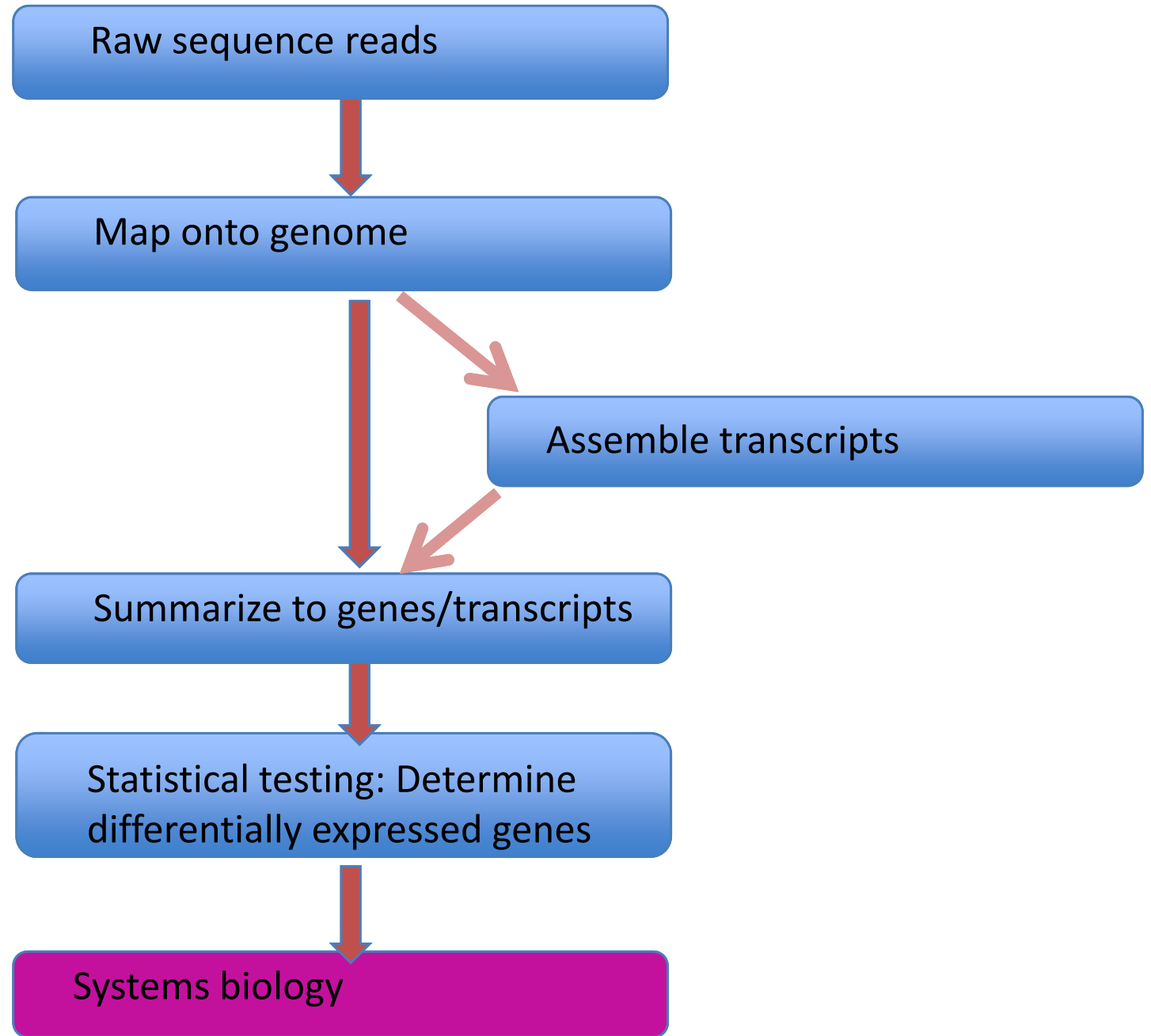
# The adjustment to data analysis is straightforward

- Assumption: core set of genes that do not change in expression.
- Pick a reference sample, compute TMM relative to reference
- TMM (Trimmed Mean of M values)  $M = \log$  ratio
- Adjustment to statistical analysis:
  - Use additional offset (GLM)
  - Use “effective” library size (Fisher’s exact test)

# Differential expression

- Statistical testing for differences in expression level
- Several methods available
- edgeR – next talk
- List of DE genes

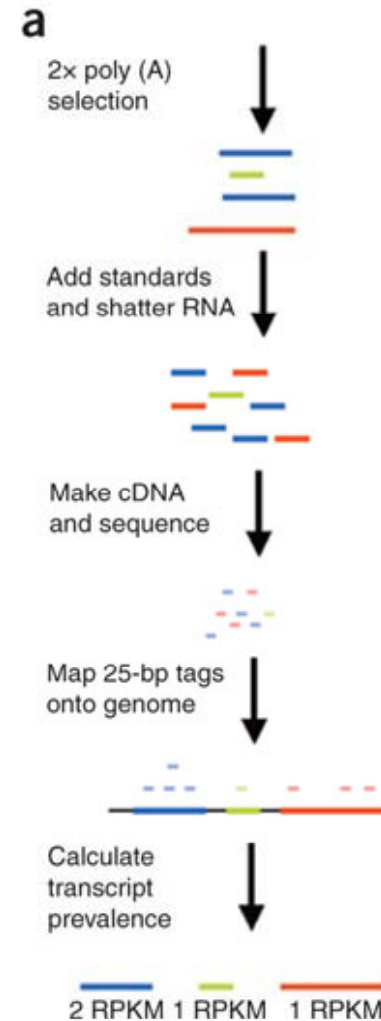
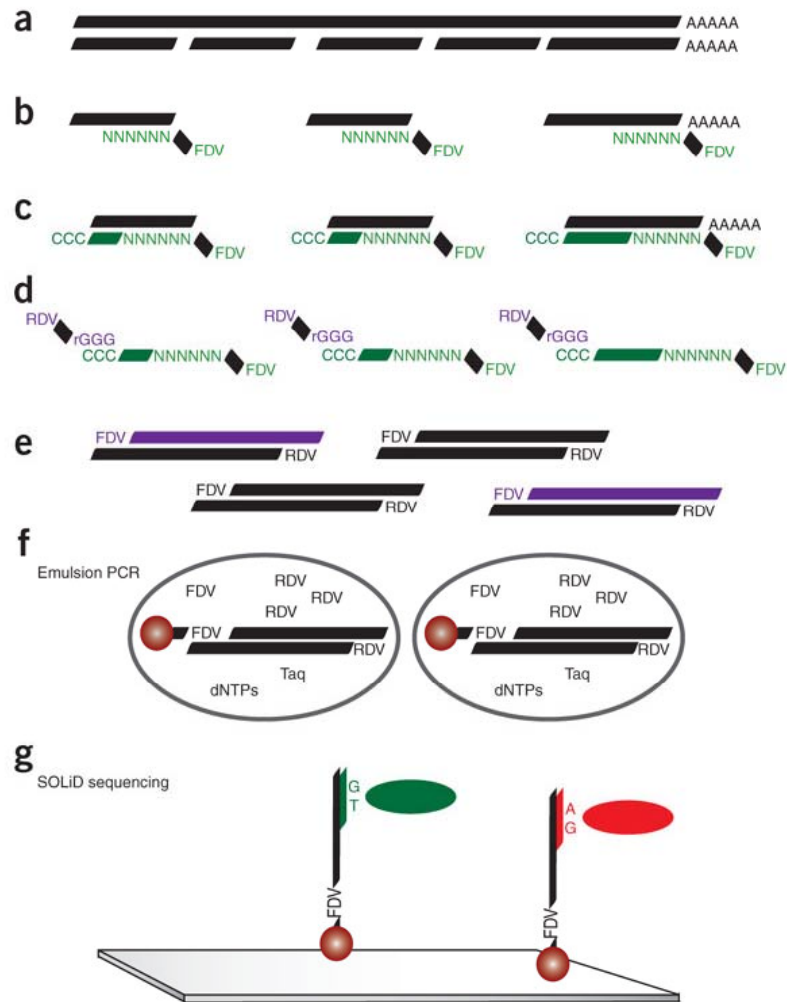
# RNA-seq analysis steps



# Going beyond gene lists

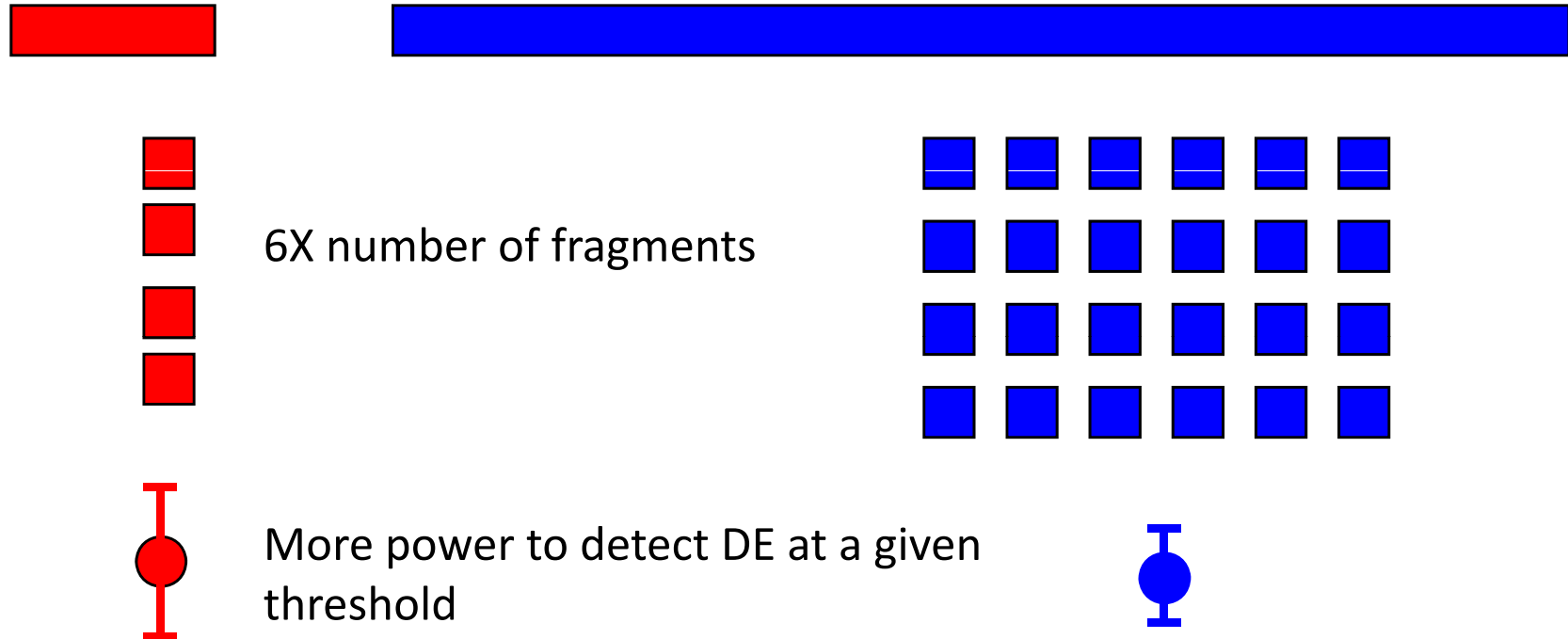
Gene length bias and  
GO analysis

# Sequencing the whole transcript - RNA fragmentation



# Length bias in RNA-seq

Equal number of transcripts 6X length



- For genes of the same expression level longer transcripts will have more reads
- Therefore there is more information for longer transcripts than shorter ones
- Longer genes have higher power to detect DE
- This length bias should not be present for microarrays

# Proportion of DE genes

# Gene Ontology analysis

- Gene Ontology categorises genes into functional groups
- We wish to know if certain gene ontologies have more DE genes than expected
- If a category has lots of long genes we expect it to have more DE genes
- Can't use simple statistical tests

# GO analysis for RNA-seq

Develop a computational method for gene category testing that can account for gene level biases in differential expression detection

Category testing refers to testing if a set of genes has an over representation of DE genes

# Three step procedure

1. Determine which genes are differentially expressed
2. Define a probability weighting function
3. Generate many random samples to produce a null distribution in order to calculate significance of a category

# Probability weighting function

Fit a function to the binary data  
series 1=DE, 0=!DE

We chose a spline with a  
monotonicity requirement

# Random sampling

- Select a random set of genes the same size as the set of DE genes
- However the probability of selecting a gene is weighted by the value of the probability weighting function (based on the length or read count of the gene)
- Then count how many genes in the DE set have the GO category of interest
- Repeat this many times
- Calculate a p-value for the category

# Results

Categories with short genes get a higher rank in GOseq

Categories with long genes get a lower rank

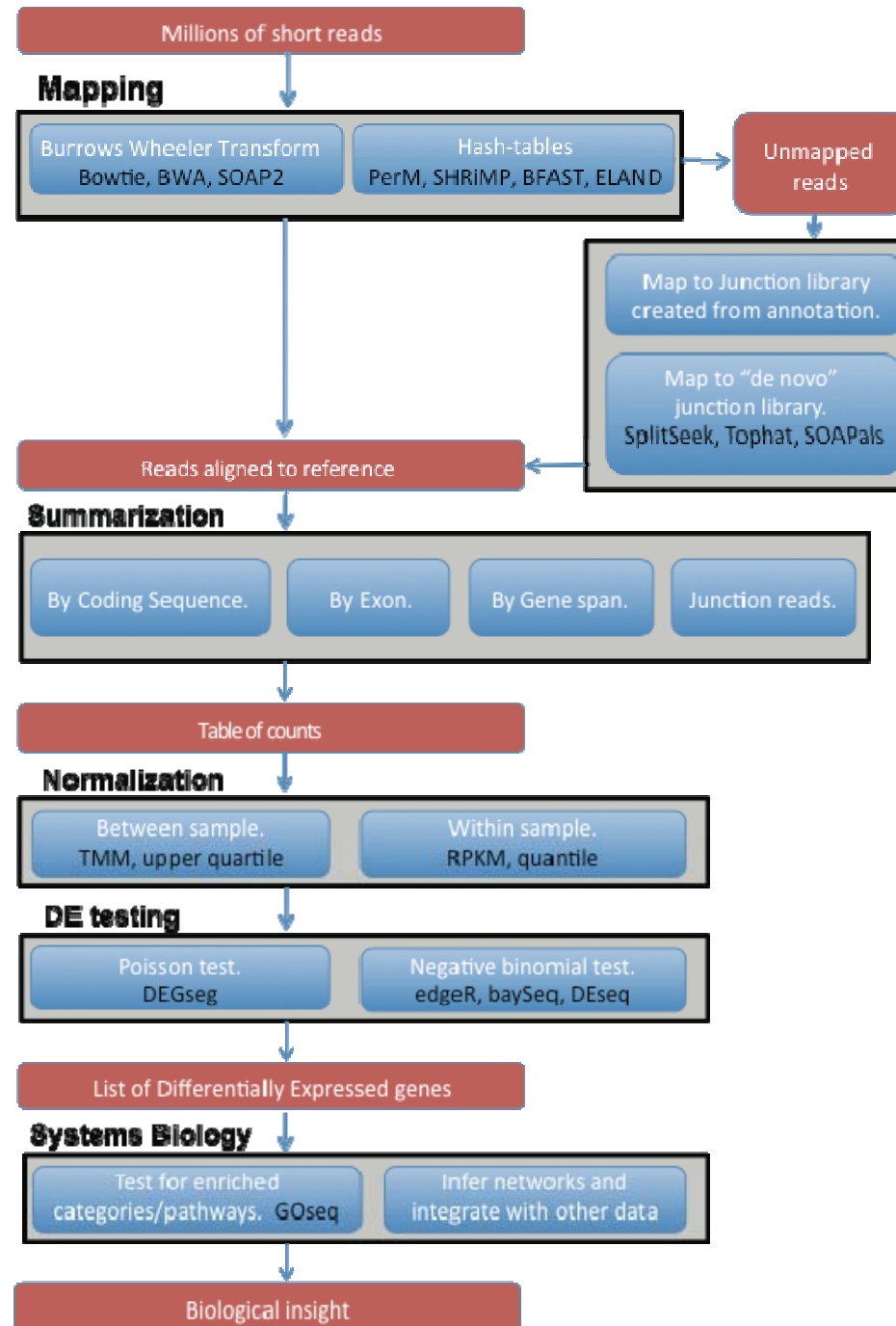
Even for a small number of categories there are a significant number of discrepancies between the methods.

Over all 20% different

# Systems Biology: Integrating data

- Gene expression
- ChIP-seq
  - Transcription factors
  - Histone modifications
- Epigenetic data
- Genome sequencing
  - Copy number
  - SNPs

# Overview of analysis



# The future

- Capacity is increasing
- Analysis methodology is critical and still developing
- Integration of different types of data
- Opportunities to use this data in new and imaginative ways

# Acknowledgements

- Mark Robinson
- Matthew Young
- Matthew Wakefield
- Gordon Smyth
- Terry Speed
- Matthew Ritchie
- Natalie Thorne
- Davis McCarthy

Research Assistant position available in my lab later this year contact:  
[alicia.oshlack@mcri.edu.au](mailto:alicia.oshlack@mcri.edu.au)