Lecture 7 - mRNAseq

This week

- Today: mRNAseq
- Tuesday: ChIP-seq (Istvan Albert)
- Wed: Resequencing (Ian Dworkin)
- Thursday: Genomes & future (Erich Schwarz)
- Friday: post-mortem

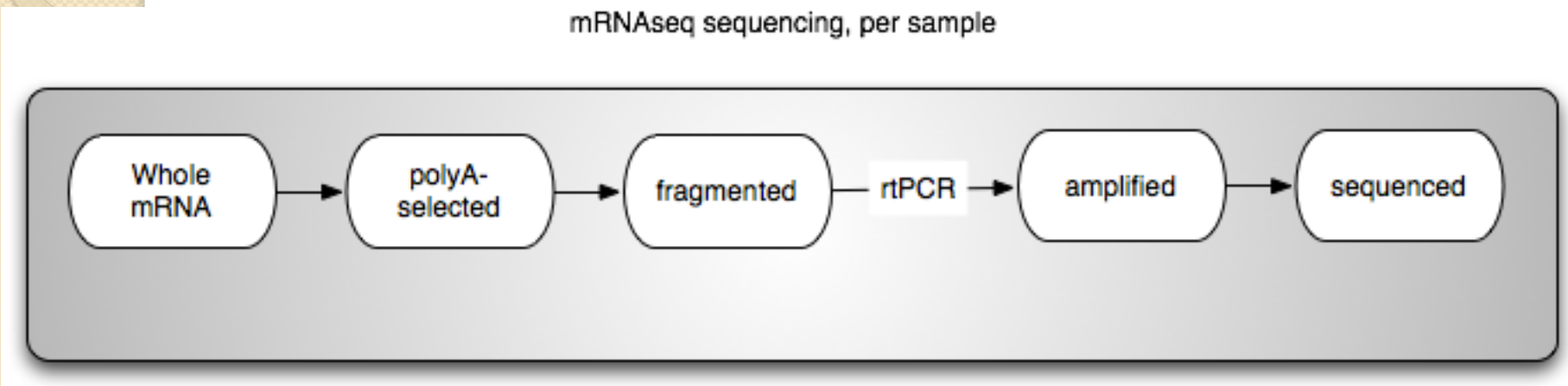
This week

- Monday 12:30pm – Ian, on data sharing
- Monday, ~7pm – Istvan, on metagenomics tools for population structure
- Monday, 9pm – "focus groups" discussion
- Tuesday 7pm – Titus, on computational challenges of soil metagenomics

- Tues, Th – bonfire?
- Th – G&T party?
- Friday: lunch in Kzoo?

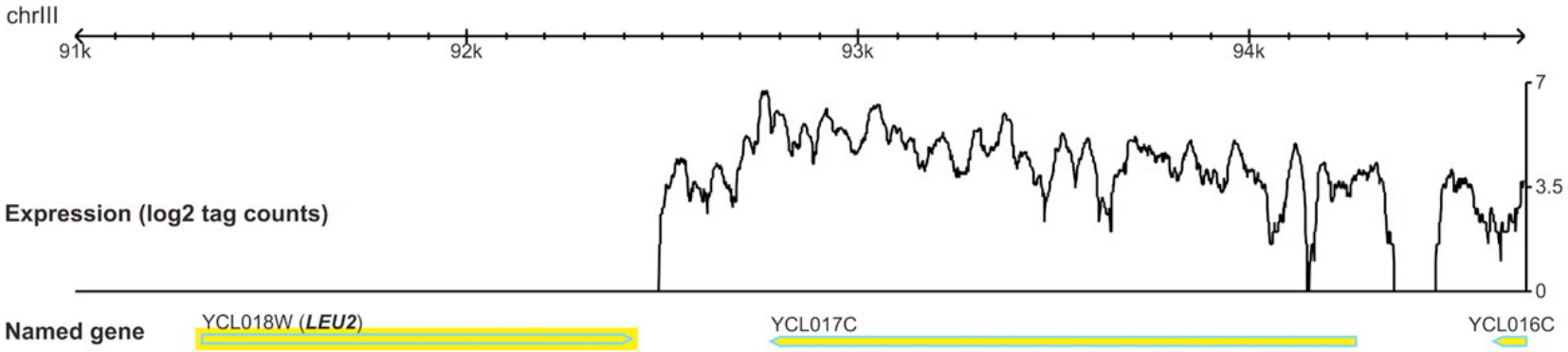
Lecture 6 - mRNAseq

Illumina mRNAseq protocol



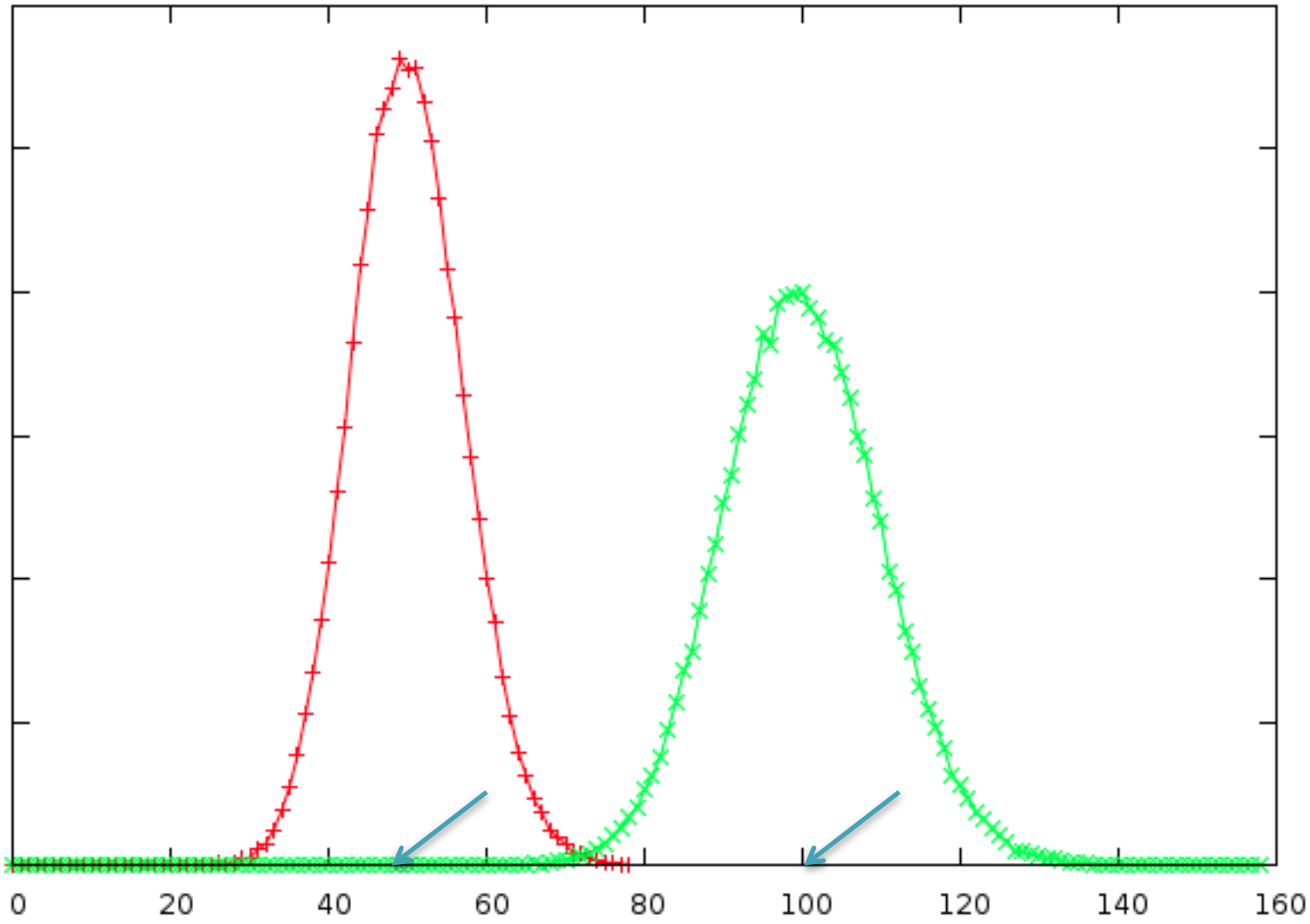
- Avoids 3' bias in sequencing

Sequencing the transcriptome



Nagalakshmi et al., Science, 2009

mRNAseq quantitation



mRNAseq vs microarrays

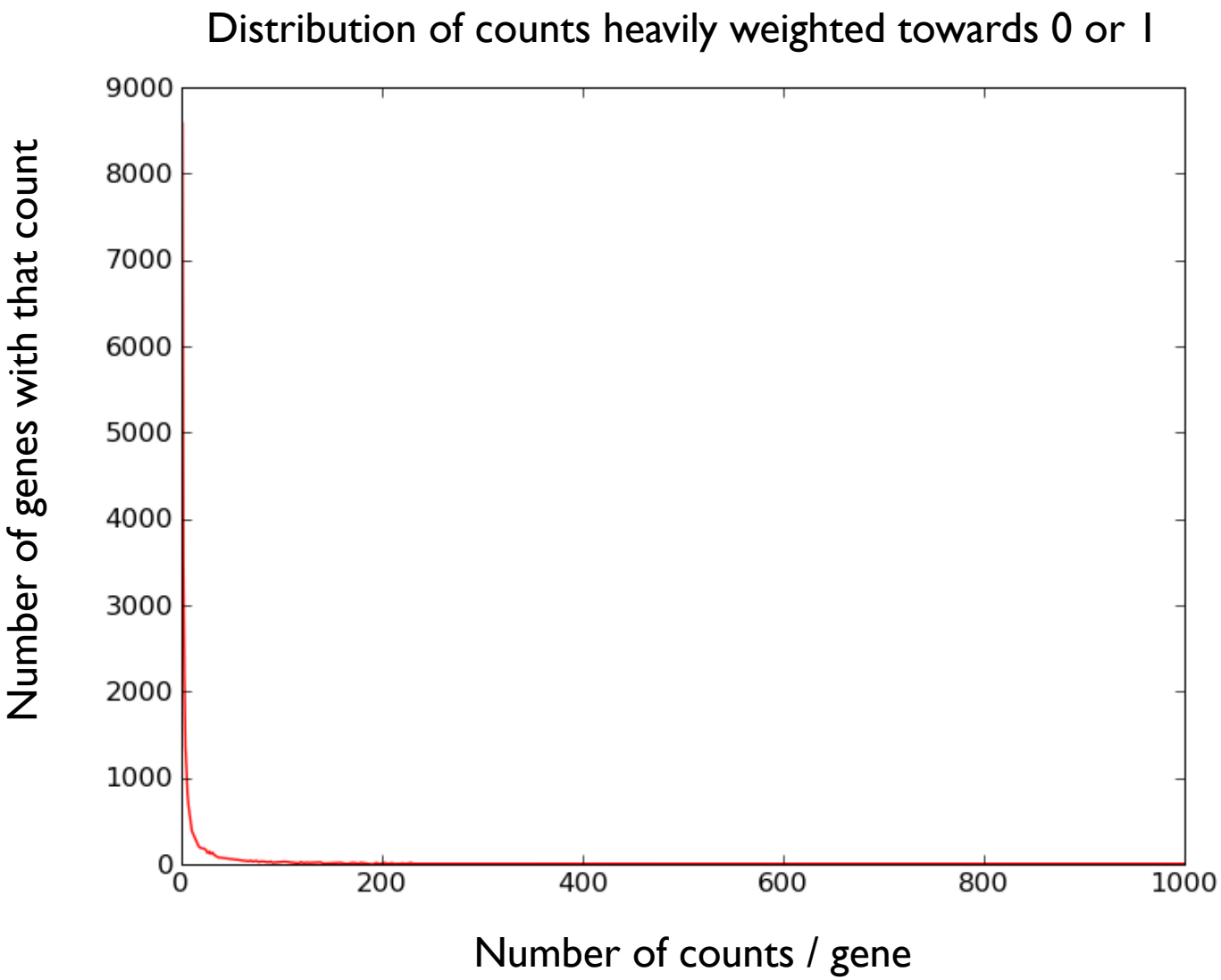
- No genome needed for mRNAseq
- Microarrays typically (always?) require internal comparison; mRNAseq does not.
- mRNAseq seems to be more reproducible & sensitive.

mRNAseq *and* microarrays

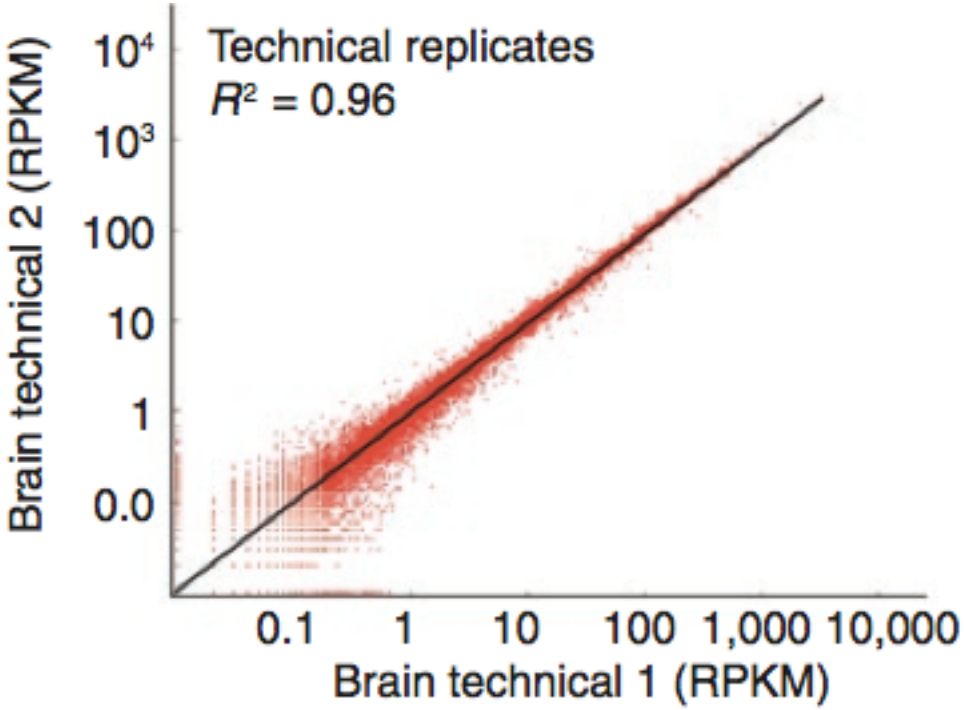
- Annotation still critical (and lacking)
- Good for hypothesis generation
- Not so good for hypothesis validation

- Statistical analysis still needed.
- Multiple samples still required ☺

Counting


Distribution of counts heavily weighted towards 0 or 1
Number of genes with that count
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
200
400
600
800
1000
Number of counts / gene

Good r^2 for tech replicates



Mortazavi et al., Nature Methods 2008

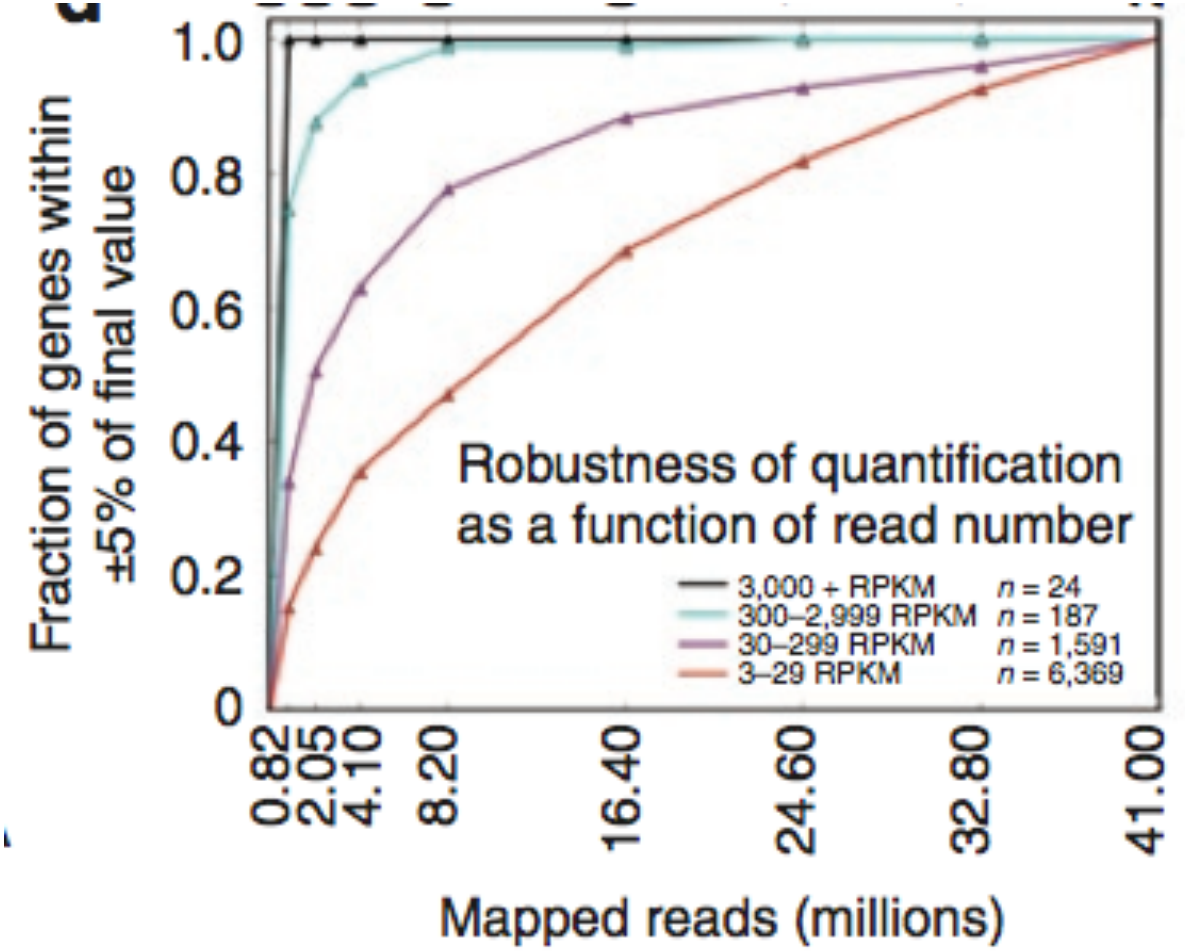
High sensitivity

Mortazavi et al. (Nat Met, 2008) estimate that a 2kb mRNA transcript can be robustly detected (~30 reads/gene) at ~.3 transcripts/cell with 50m reads.

RPKM: "reads per kilobase of mRNA", a measure that normalizes to length.

…does not normalize for sequencing bias!

Reasonably robust to # of reads



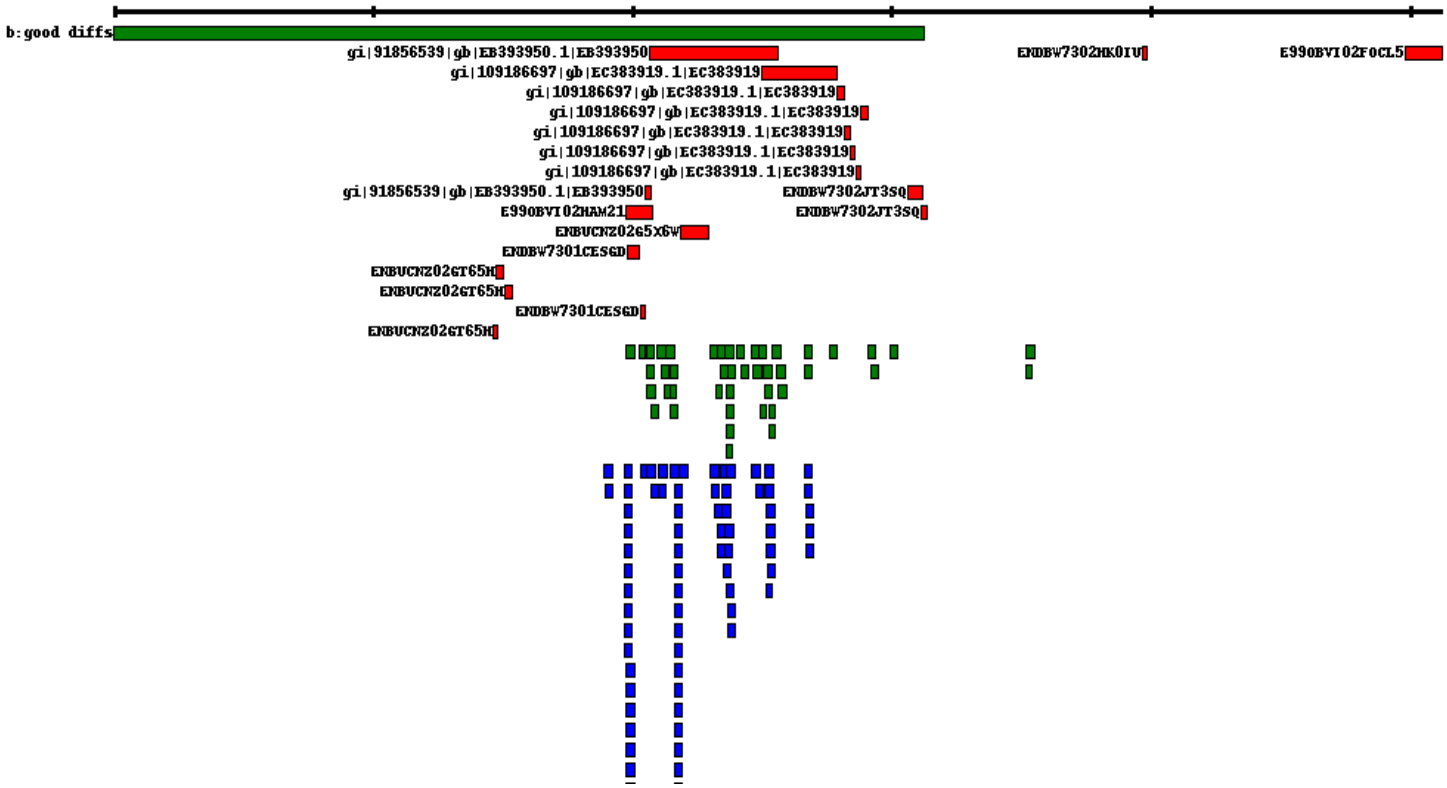
Mortazavi et al., Nature Methods 2008

mRNAseq analysis

1. Find or build transcriptome
2. Map reads to transcriptome
3. Count and normalize across samples
4. Compare between samples
5. Pathway/gene analysis

6. …challenges for the future

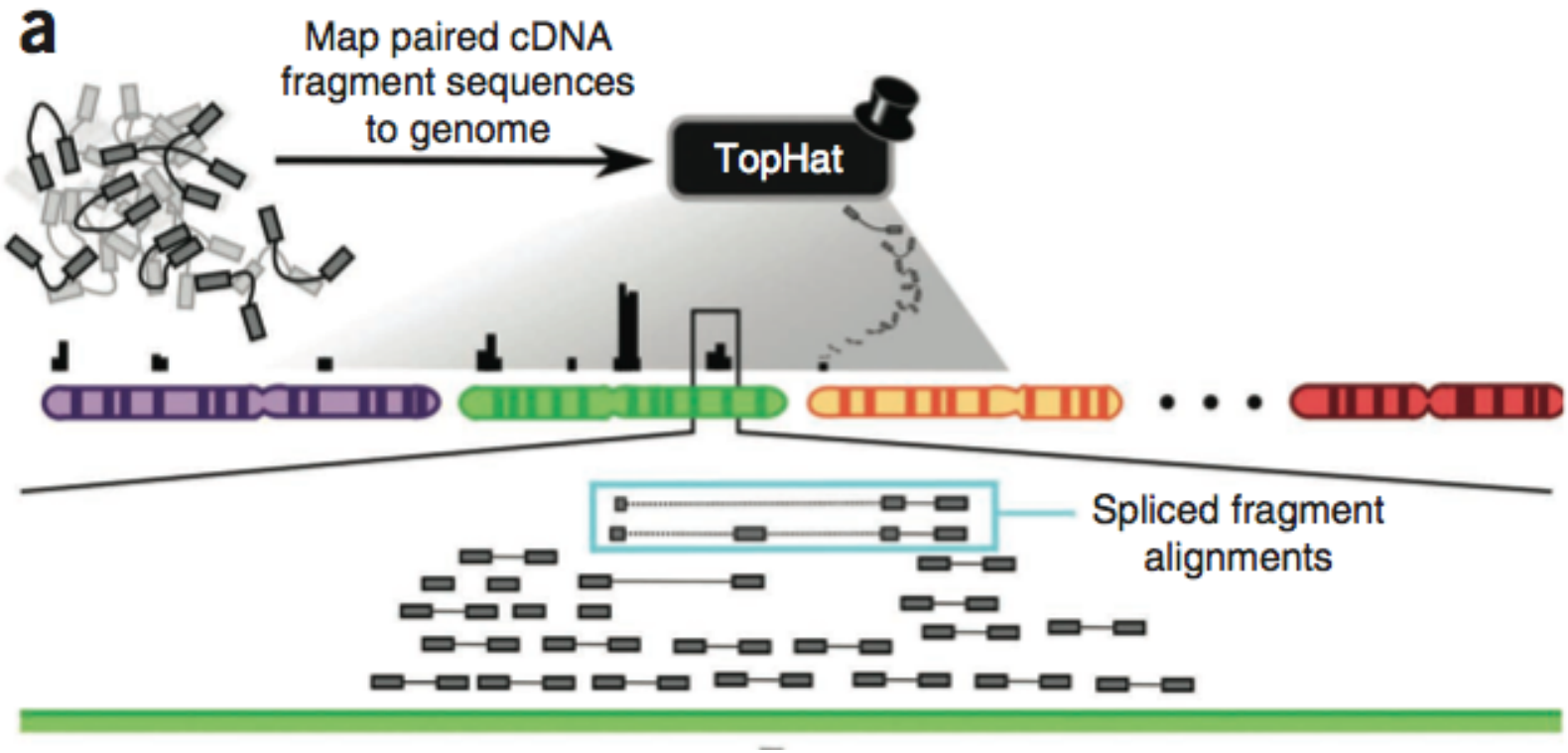
mRNAseq example: lamprey



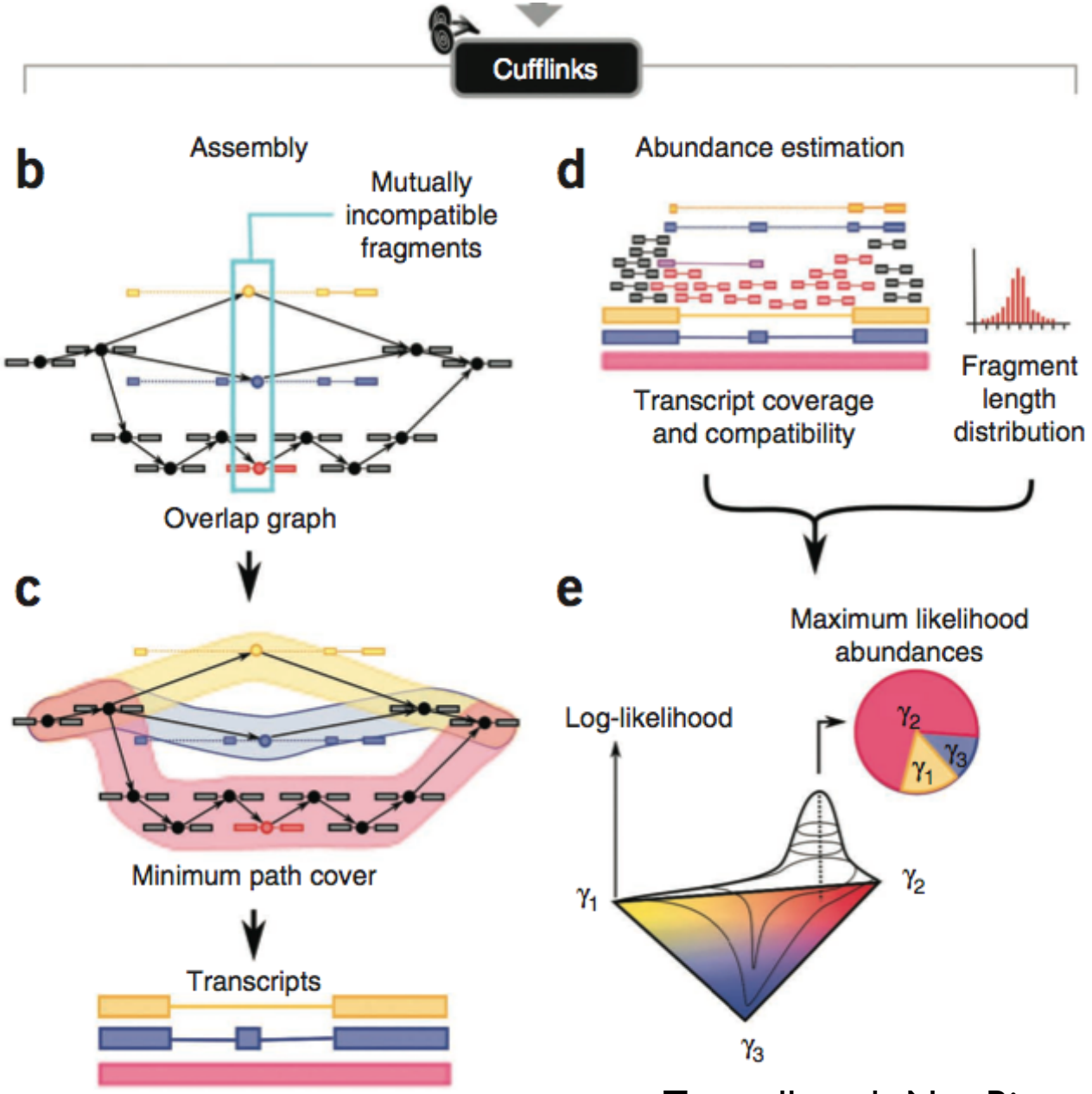
Mapping-based approaches

- Require reference!
- Cufflinks, TopHat, MAKER.

Cufflinks: pipeline for mRNAseq analysis

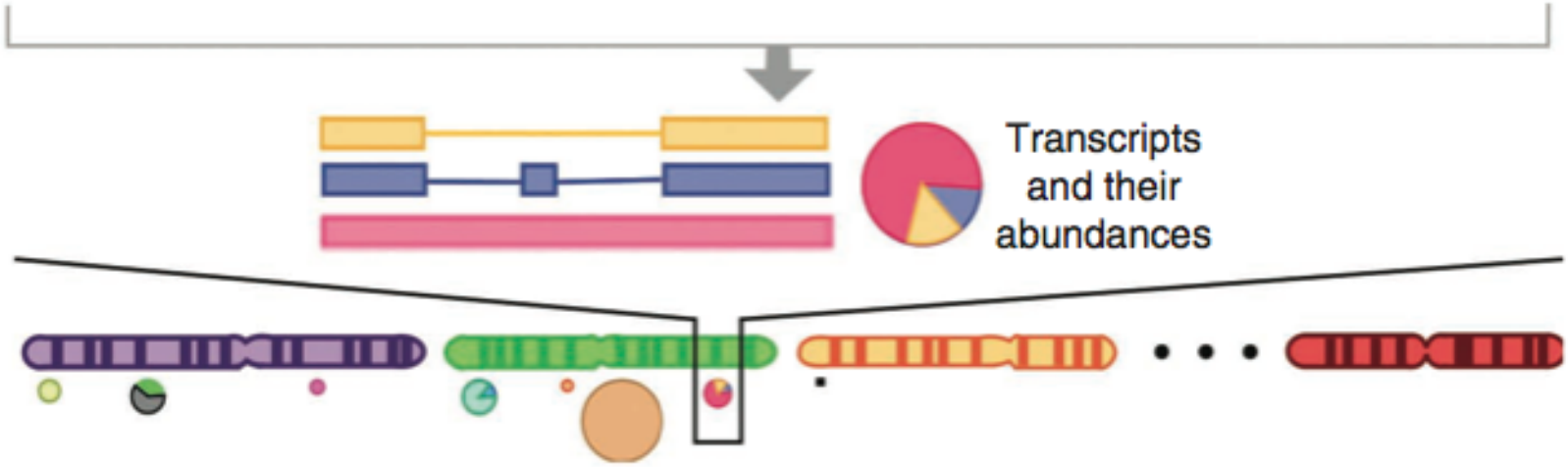


Trapnell et al., Nat. Biotech 2010



Trapnell et al., Nat. Biotech 2010

Cufflinks cont.



In mouse muscle cell line,

- 13.6k known splice isoforms
- 12.7k novel
- Differential regulation of isoforms from 1600 genes.

Trapnell et al., Nat. Biotech 2010

mRNAseq assembly

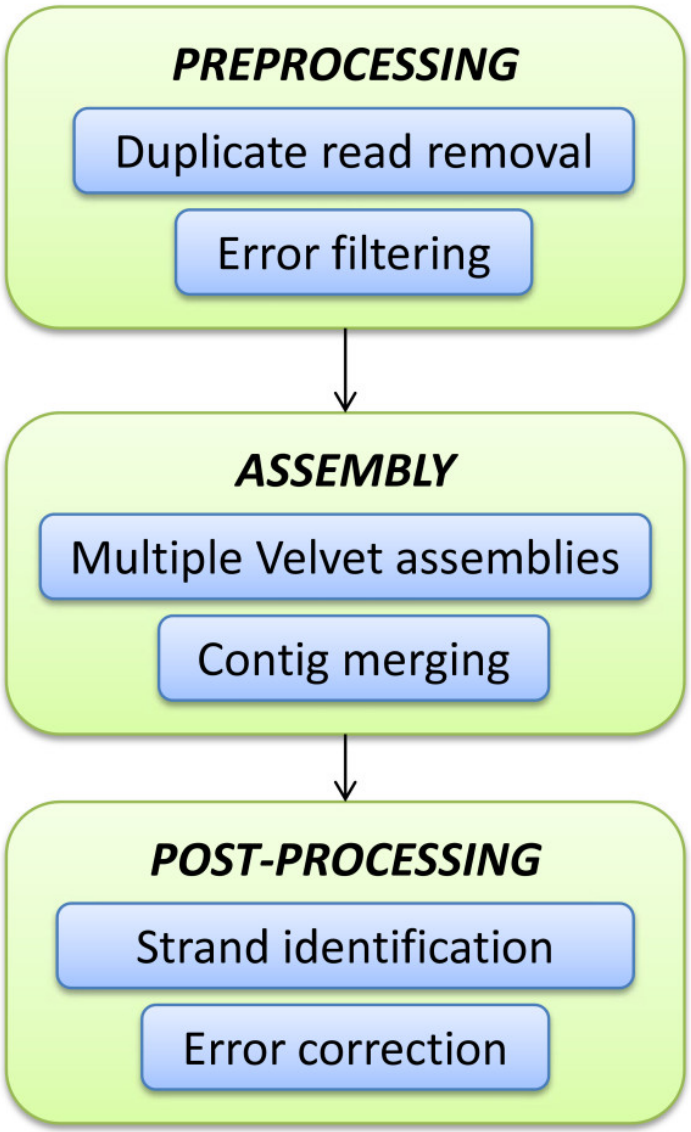
- No reference required!
- Resolution of splice variants will depend critically on depth of sequencing.

mRNAseq assembly challenges

- Several programs (Trinity, Oases, …) but all have poor scaling.
- Multiple k values yield different isoforms, and this is poorly handled by existing software.
- Integrating multiple sources of evidence is challenging (454, Sanger, Illumina)

Rnnotator ex.

- Cleans up reads
- Handles multiple K
- Post-processes to id strand, fix errors



Transcriptome annotation

- If no RefSeq…?
- Usually done via BLAST.

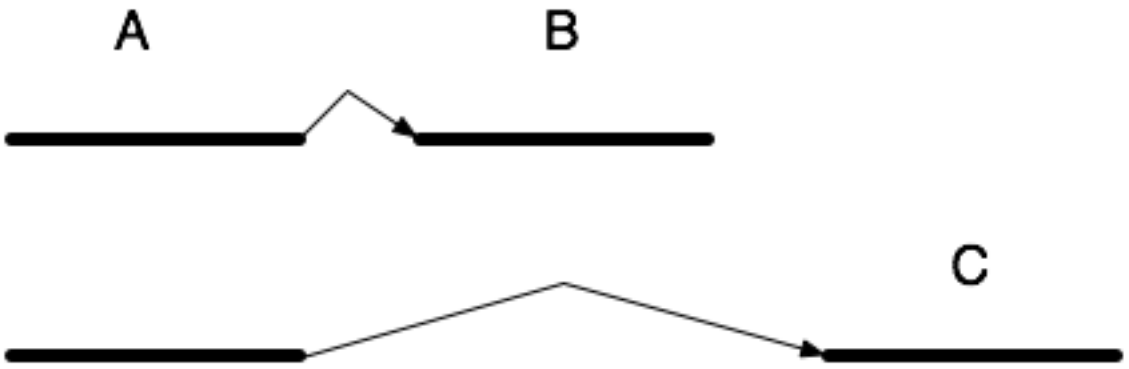
- Consideration: do you require reciprocal best hits, or not?

Quantifying transcripts

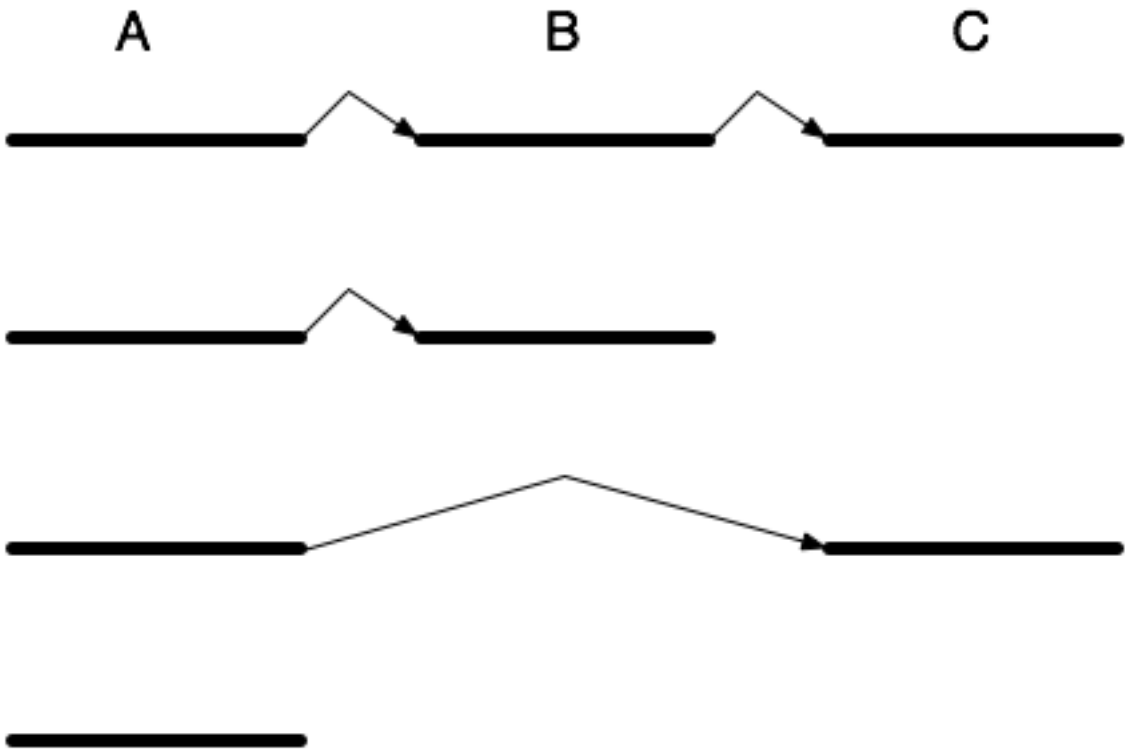
- Predominant method: map with bowtie/ BWA/etc.

- Should you eliminate reads that map to more than one transcript?
  - Yes…
  - Think repeats, in particular.
  - …but this will cause trouble for certain kinds of splice variants!

Isoform analysis – some easy…



Isoform analysis – some hard



Counting methods currently rely on presence of *unique sequence* to which to map.

Normalization

- In order to compare between mRNAseq samples, you must normalize.
- Think qPCR, "standard" genes.
- This controls for:
  - Different mRNA amounts
  - Different RT efficiency
  - Different sequencing depth/error rates/etc.
- No good way to control for *differential* RT or sequencing efficiency.
- So, can only compare changes in transcript levels between treatments/time points.

Normalization techniques

- Normalize to a consistently expressed gene (e.g. "housekeeping")
  - Finding housekeeping genes is challenging!
- Normalize to maximum expressed gene, or sum, or average.
- *Quantile normalization* normalizes counts assuming source distribution is the same shape.
- Will show you the method in tutorial.

Comparison across samples

- Why not log2 ratio?
- May be ok for naïve hypothesis discovery.
- Significance is diluted/difficult because of
  - Multiple hypothesis testing
  - Gene expression is not independent!
- Likit will demonstrate DEGseq.

"Gene list" …now what?

- (Same dilemma as microarrays :)
- Look for interesting genes.
- Pathway analysis?
  - GoMiner: classifies genes into "biology coherent" categories and looks for over/underrepresentation
  - DAVID: same.
- Major challenge (FFS…) turns out to be robust interconversion between gene ID formats. Sigh.
- ???

Research challenges

- Better artifact/graph filtering.
- Improvements in assembly methodology
  - Scaling
  - Combining multiple K
- High-quality isoform analysis