Lecture 6 - mRNAseq

This week

- Today: mRNAseq
- Tuesday: resequencing (Jeff Barrick)
- Wed: ChIP-seq (Mark Robinson)
- Thursday: building & using pipelines
- Friday: post-mortem

This week

- Tues, Th – bonfire
- Th – G&T party?
- Friday: ice cream at MSU dairy store

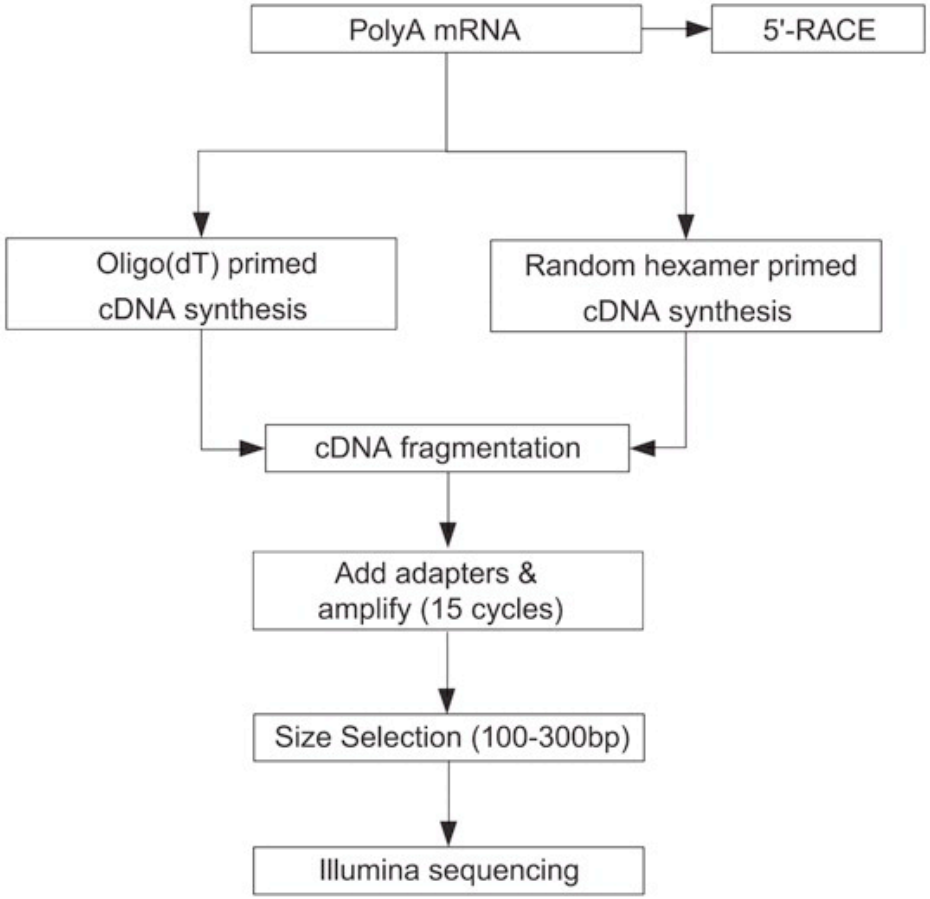
Every day: CTB dominates Adina at frisbee

Today

- Morning: mRNAseq mapping, counting, and normalization
- Afternoon: significance estimation (?)
- Afternoon: working with files & dirs in UNIX
- Evening: GO analysis

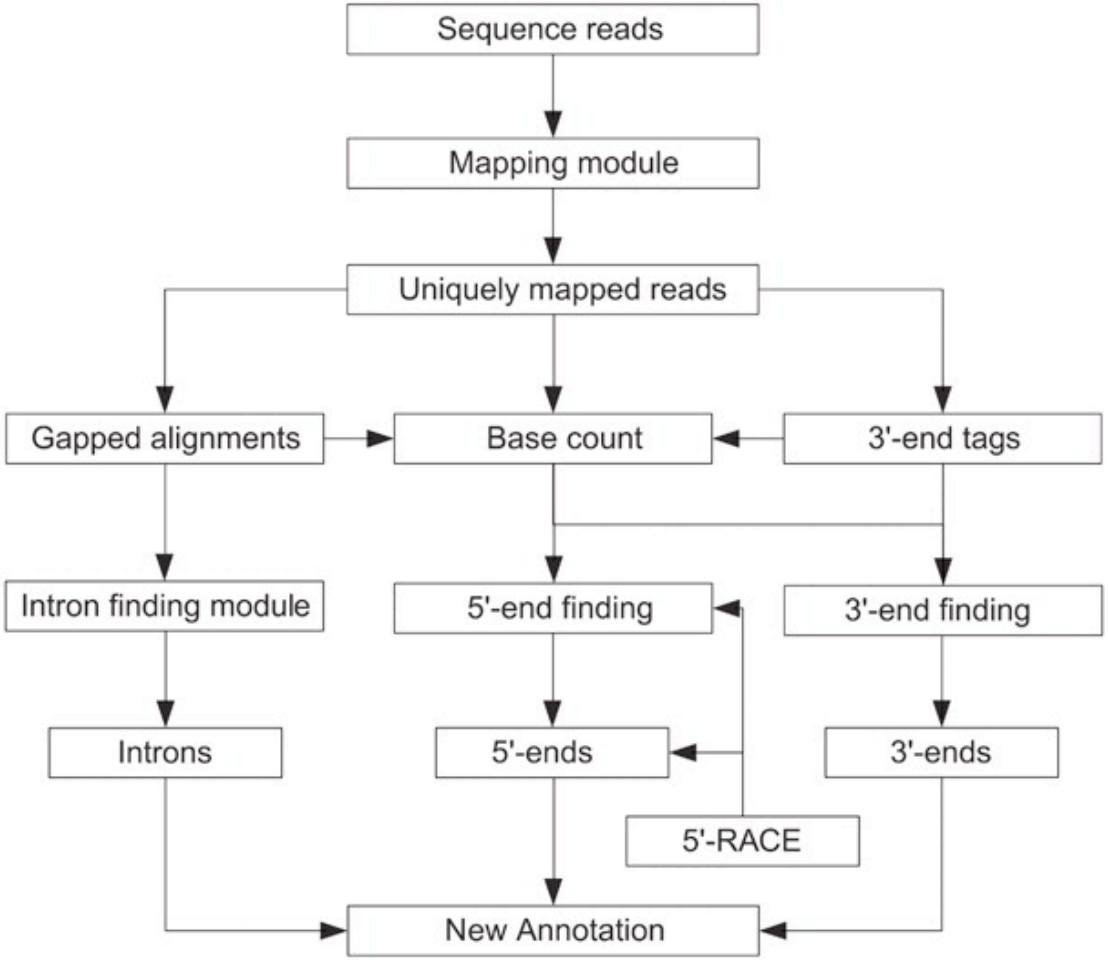
Lecture 6 - mRNAseq

Sequencing the transcriptome



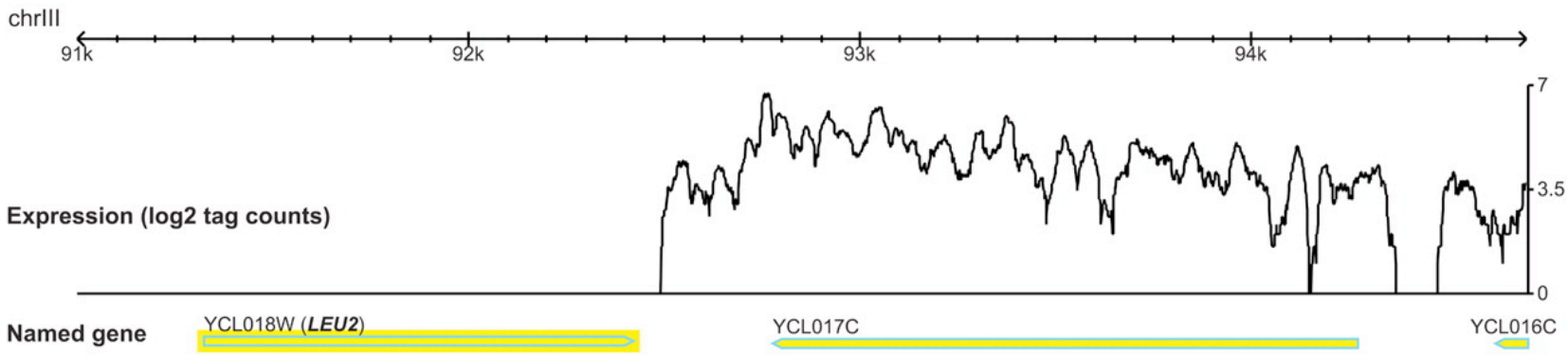
Nagalakshmi et al., Science, 2009

Sequencing the transcriptome



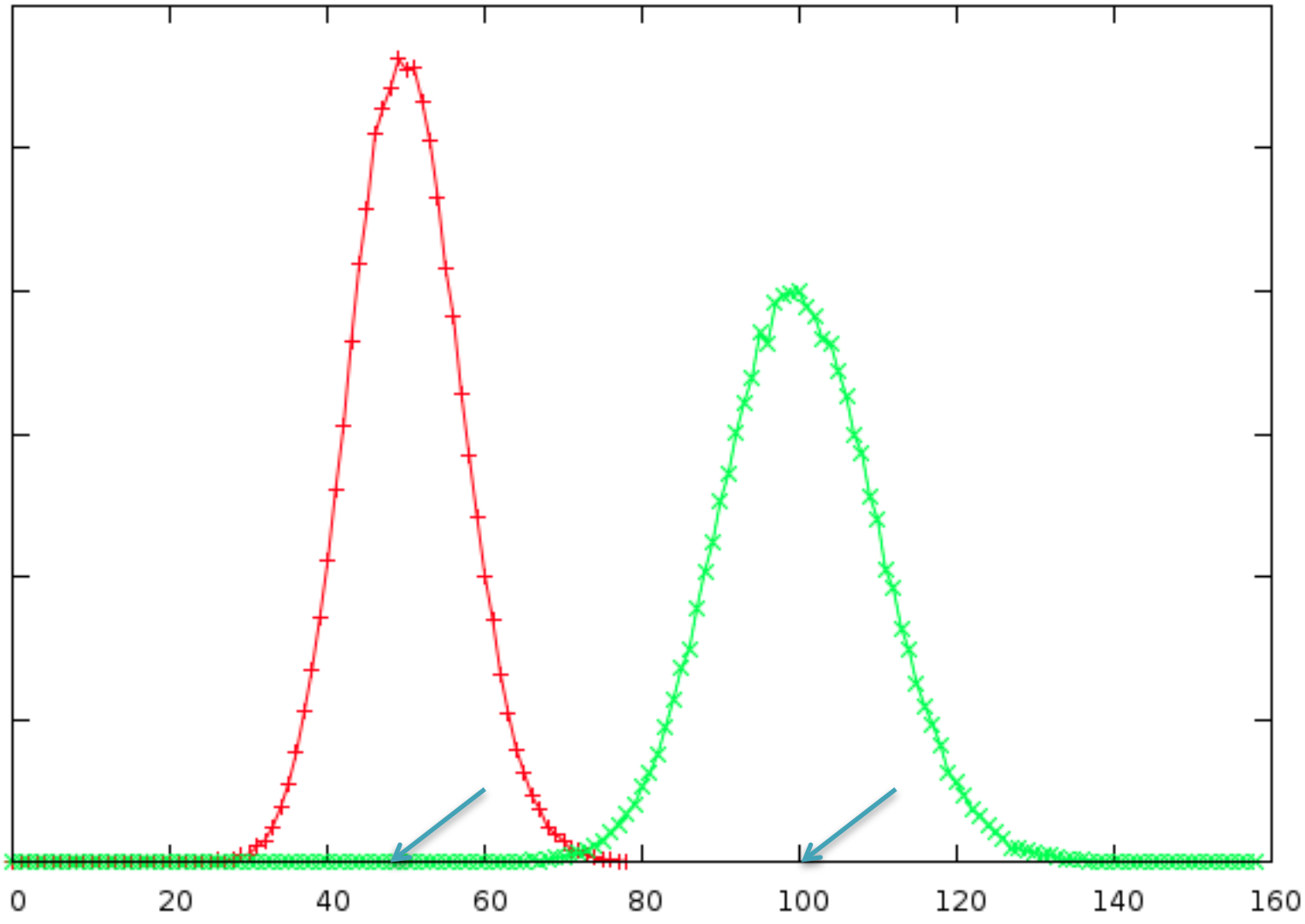
Nagalakshmi et al., Science, 2009

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.

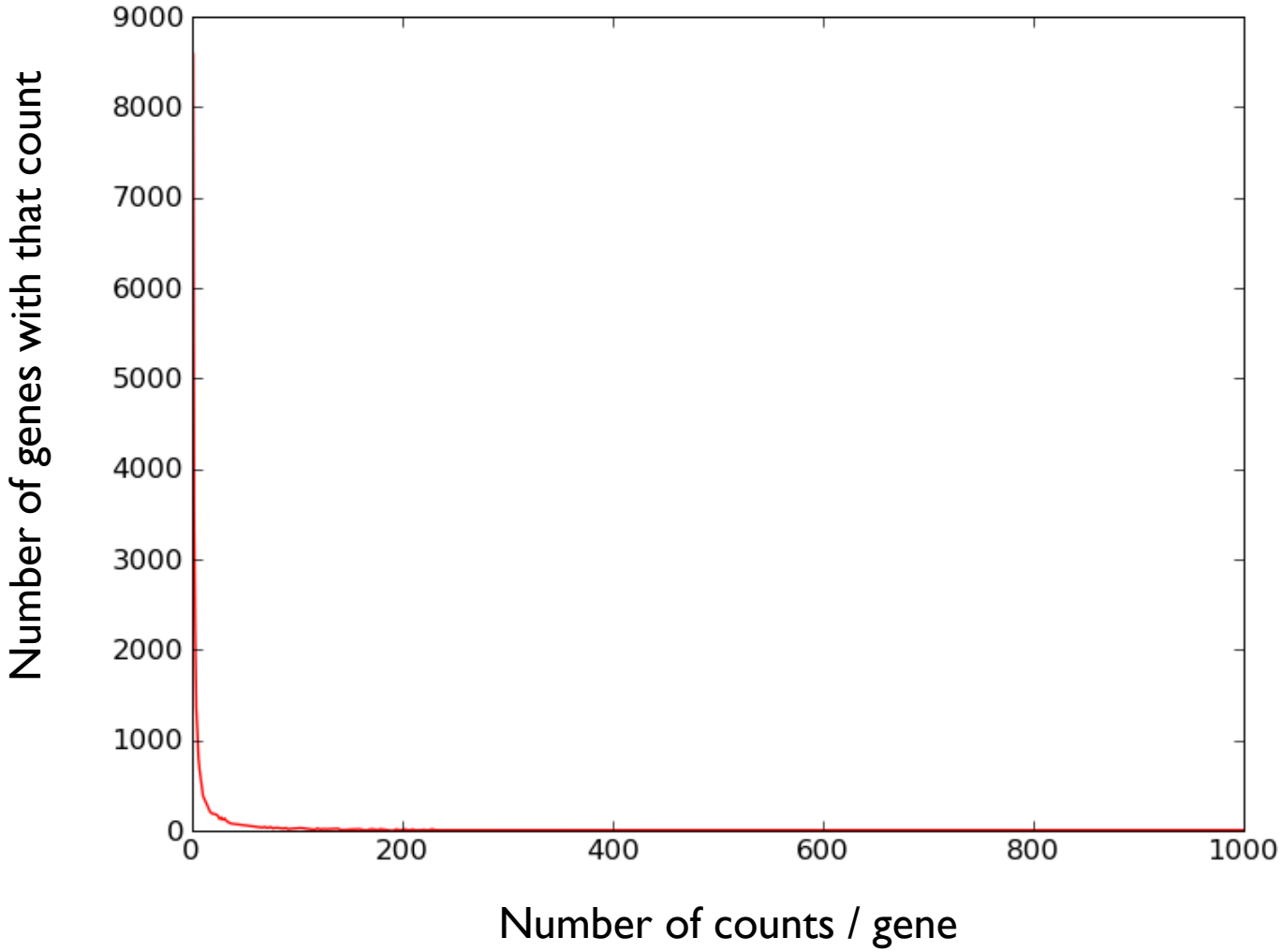
mRNAseq *and* microarrays

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

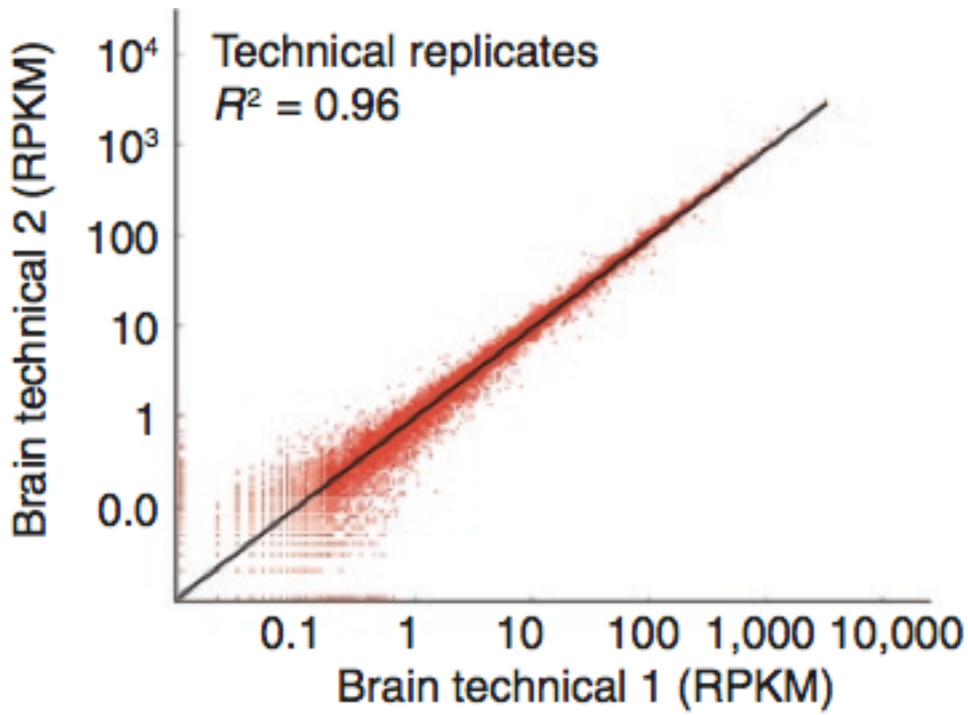
- Statistical analysis still needed.
- Multiple samples still basically required.

Counting

Distribution of counts heavily weighted towards 0 or 1



Good r^2 for tech replicates

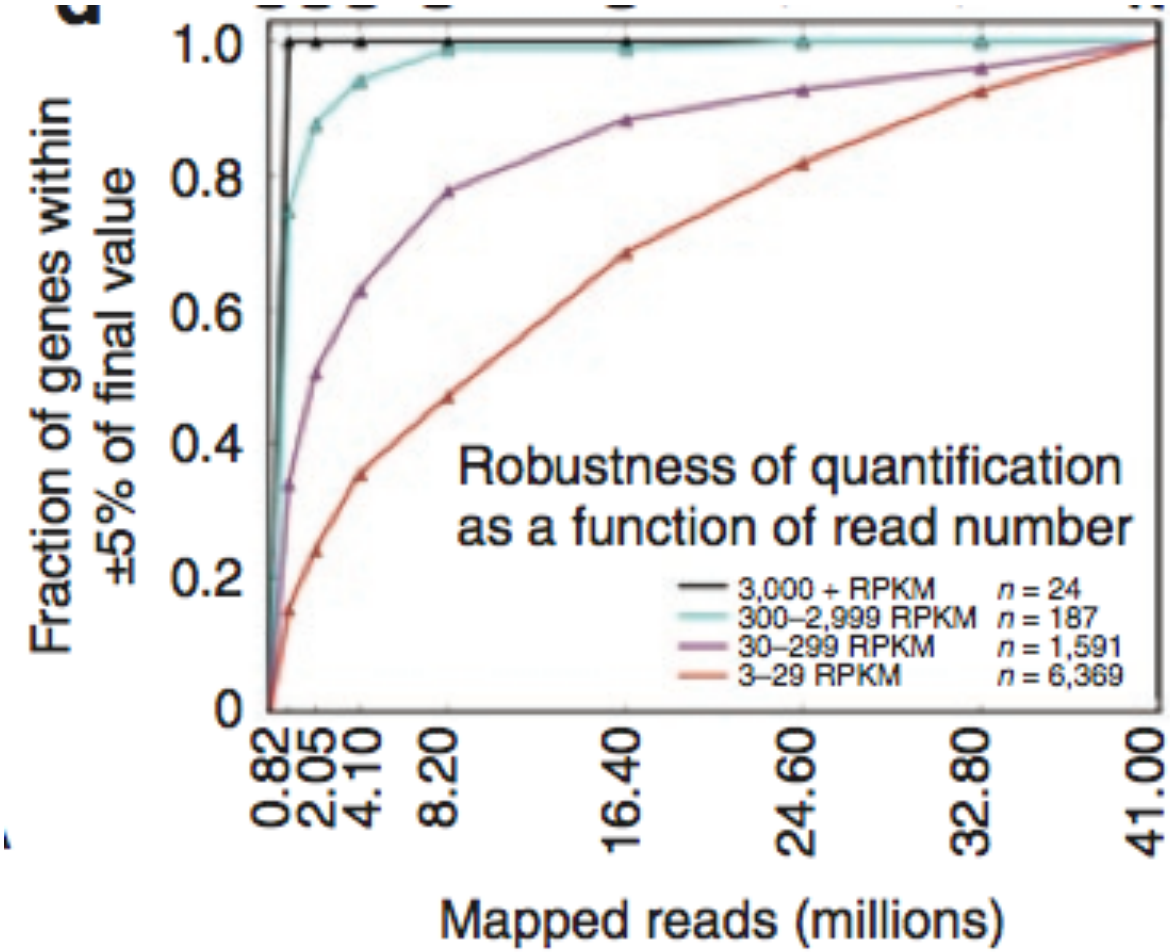


Mortazavi et al., Nature Methods 2008

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.

Reasonably robust to # of reads



Mortazavi et al., Nature Methods 2008

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

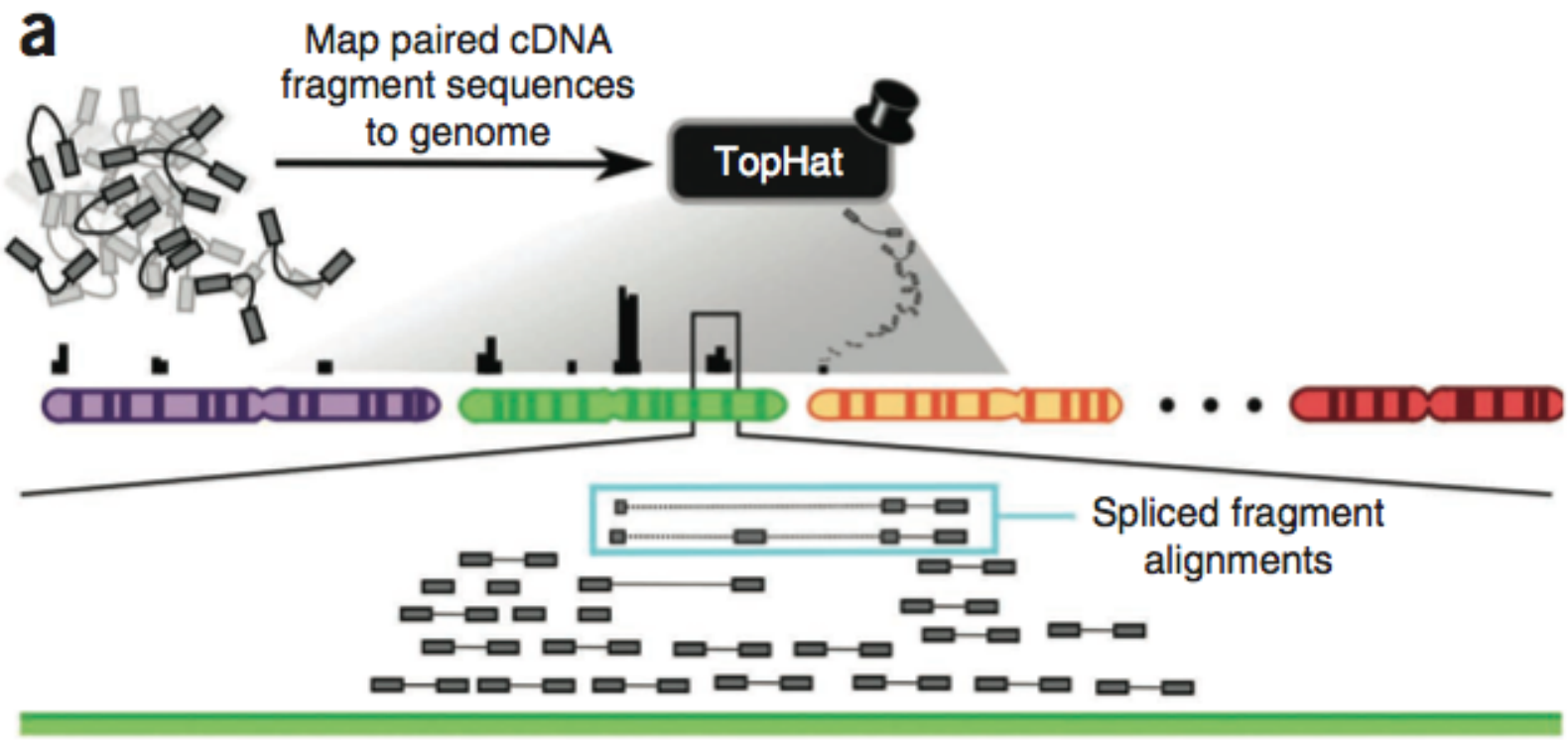
- Normalize to a consistently expressed gene (e.g. "housekeeping")
  - Finding housekeeping genes is challenging!
- Normalize to maximum expressed gene, or sum, or average.
- *Quantile normalization* normalizes for shape of distribution.
- Will show you the method in tutorial.

Simultaneous annotation + abundance calc

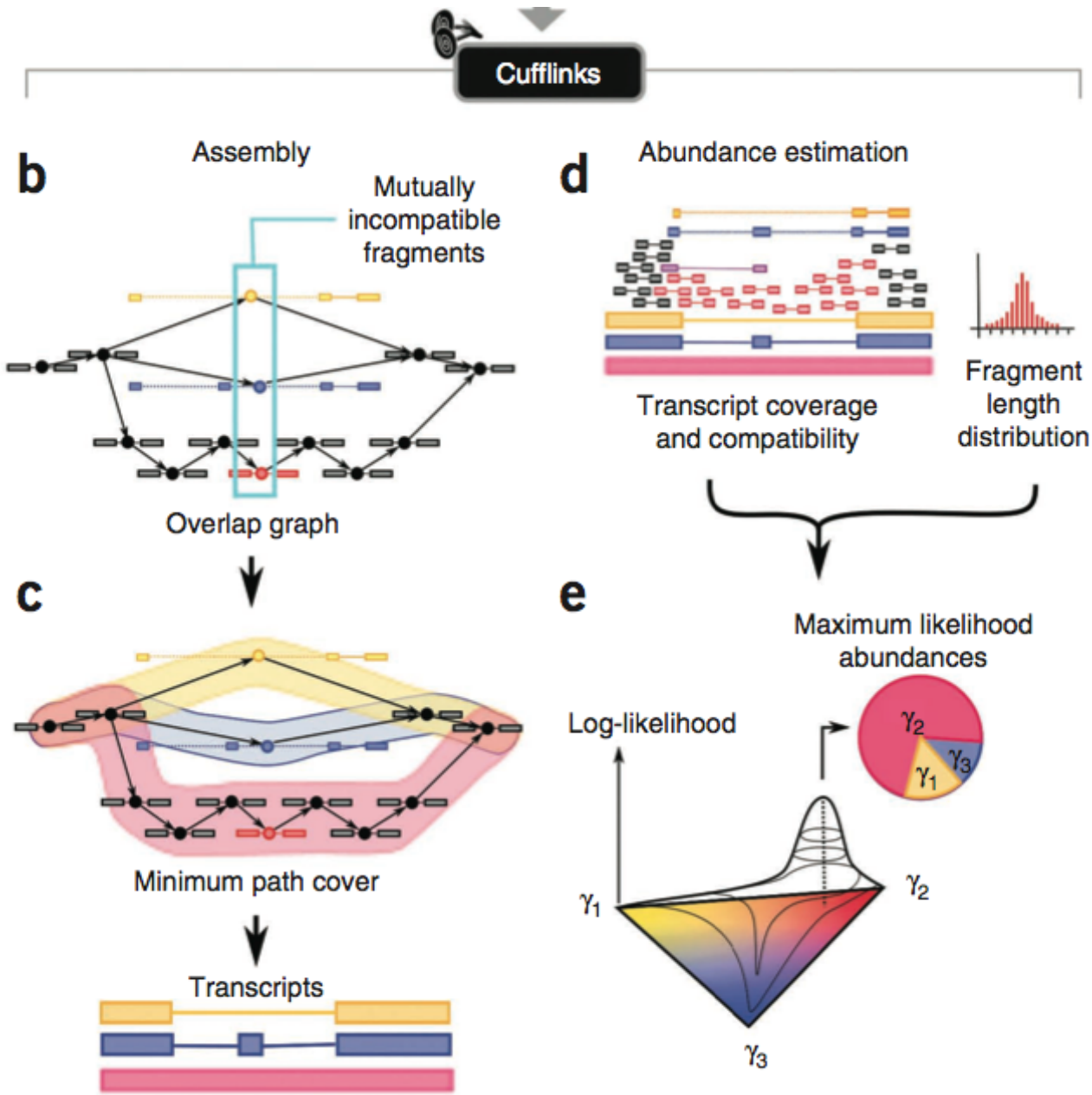
Challenging:

- Want to characterize exons & join isoforms
- …across multiple time points.
- Computationally difficult
- Requires paired-end sequencing

Cufflinks: pipeline for mRNAseq analysis

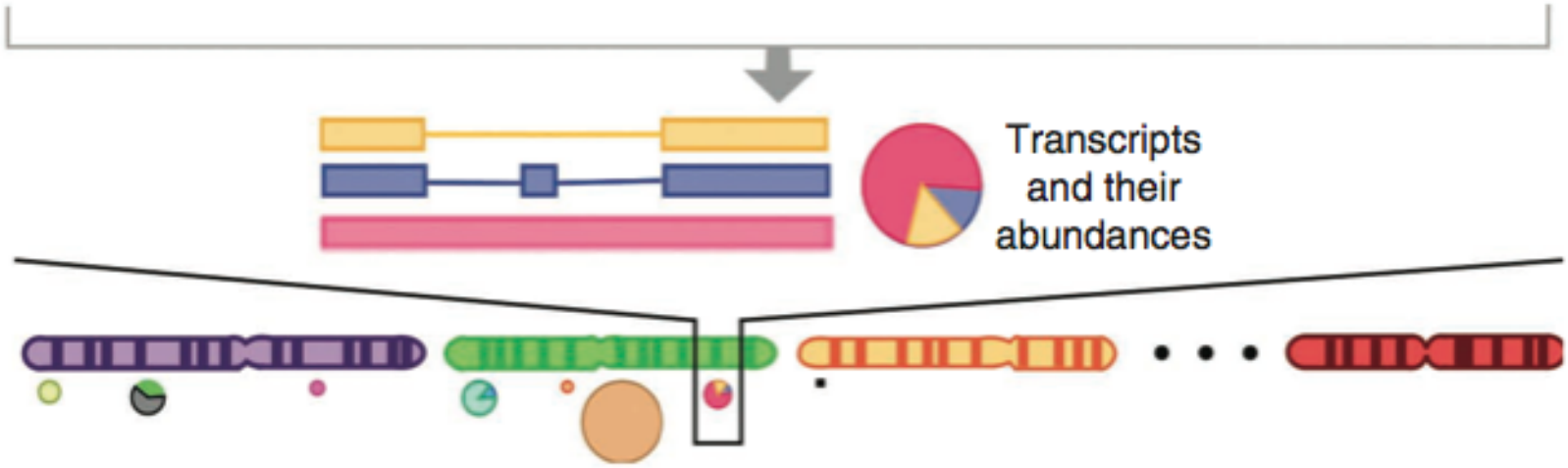


Trapnell et al., Nat. Biotech 2010



Trapnell et al., Nat. Biotech 2010

Cufflinks cont.



In mouse muscle cell line,

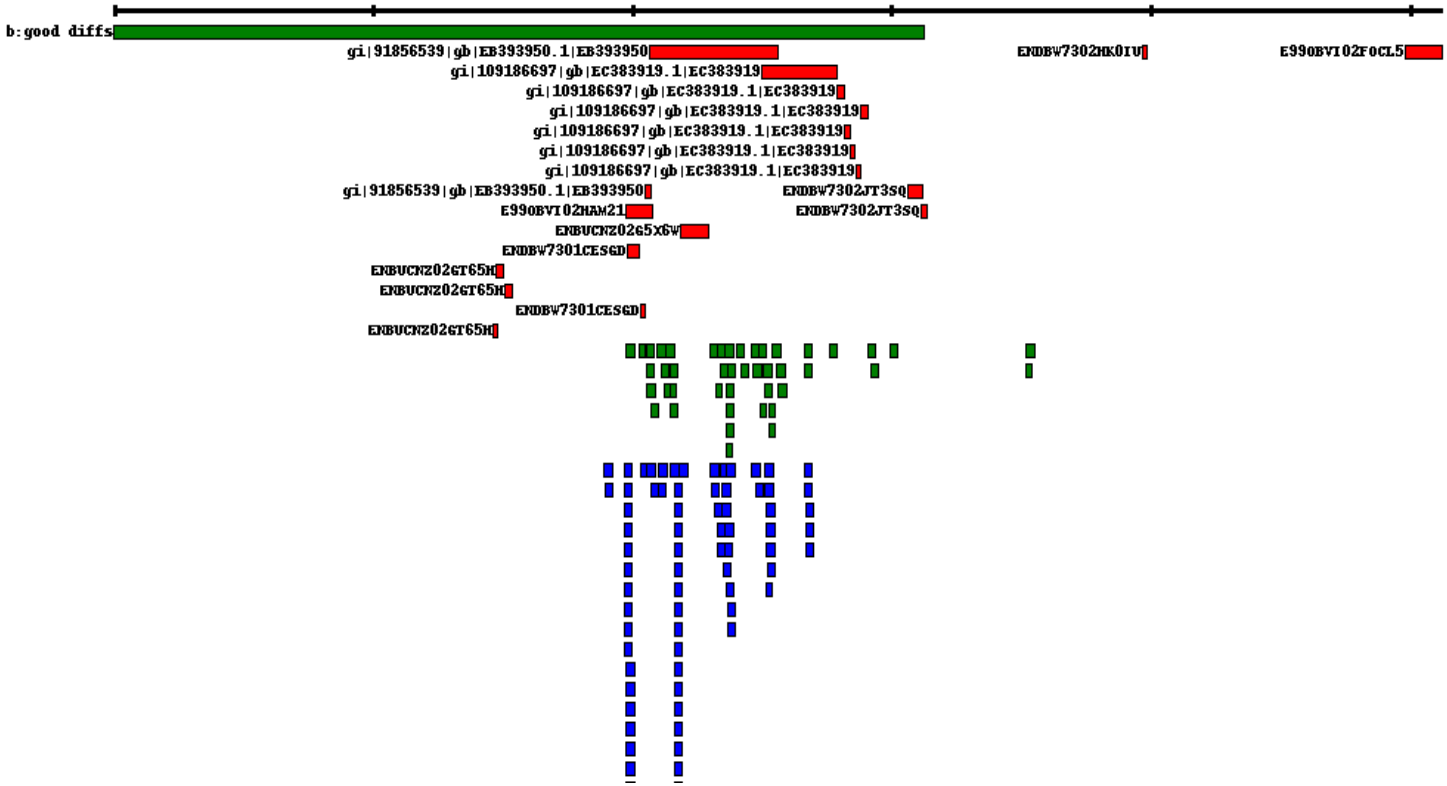
- 13.6k known splice isoforms
- 12.7k novel
- Differential regulation of isoforms from 1600 genes.

Trapnell et al., Nat. Biotech 2010

Primary mRNAseq issues for "emerging model organisms"

- Do you have a good genome + gene prediction set?
- What kind of transcriptome prior knowledge do you have?
  - If none, you must assemble rnatigs.
  - How do you evaluate assembly?
- Are you "close" to a good reference organism (vertebrates, worm, Drosophila, Arabidopsis, yeast, E. coli, …?)

mRNAseq example: lamprey



Concluding thoughts

- mRNAseq provides great power and resolution for annotation & quantitation.
- mRNAseq for emerging model organisms is challenging for reasons of
  - No good assembly (rna or genome)
  - No good annotation
- As usual, paired-end sequencing is important.
- Unlike de novo genomic sequencing, *mate-pair* sequencing is not so important for mRNAseq.