

# Measuring transcriptomes with RNA-Seq

---

BMI/CS 776

[www.biostat.wisc.edu/bmi776/](http://www.biostat.wisc.edu/bmi776/)

Spring 2015

Colin Dewey

[cdewey@biostat.wisc.edu](mailto:cdewey@biostat.wisc.edu)

# Overview

---

- Some motivation: axolotl
- RNA-Seq technology
- The RNA-Seq quantification problem
- Generative probabilistic models and Expectation-Maximization for the quantification task

# What I want you to get from this lecture

---

- What is RNA-Seq?
- How is RNA-Seq used to measure the abundances of RNAs within cells?
- What probabilistic models and algorithms are used for analyzing RNA-Seq?

# Some motivation

---



James Thomson



Ron Stewart

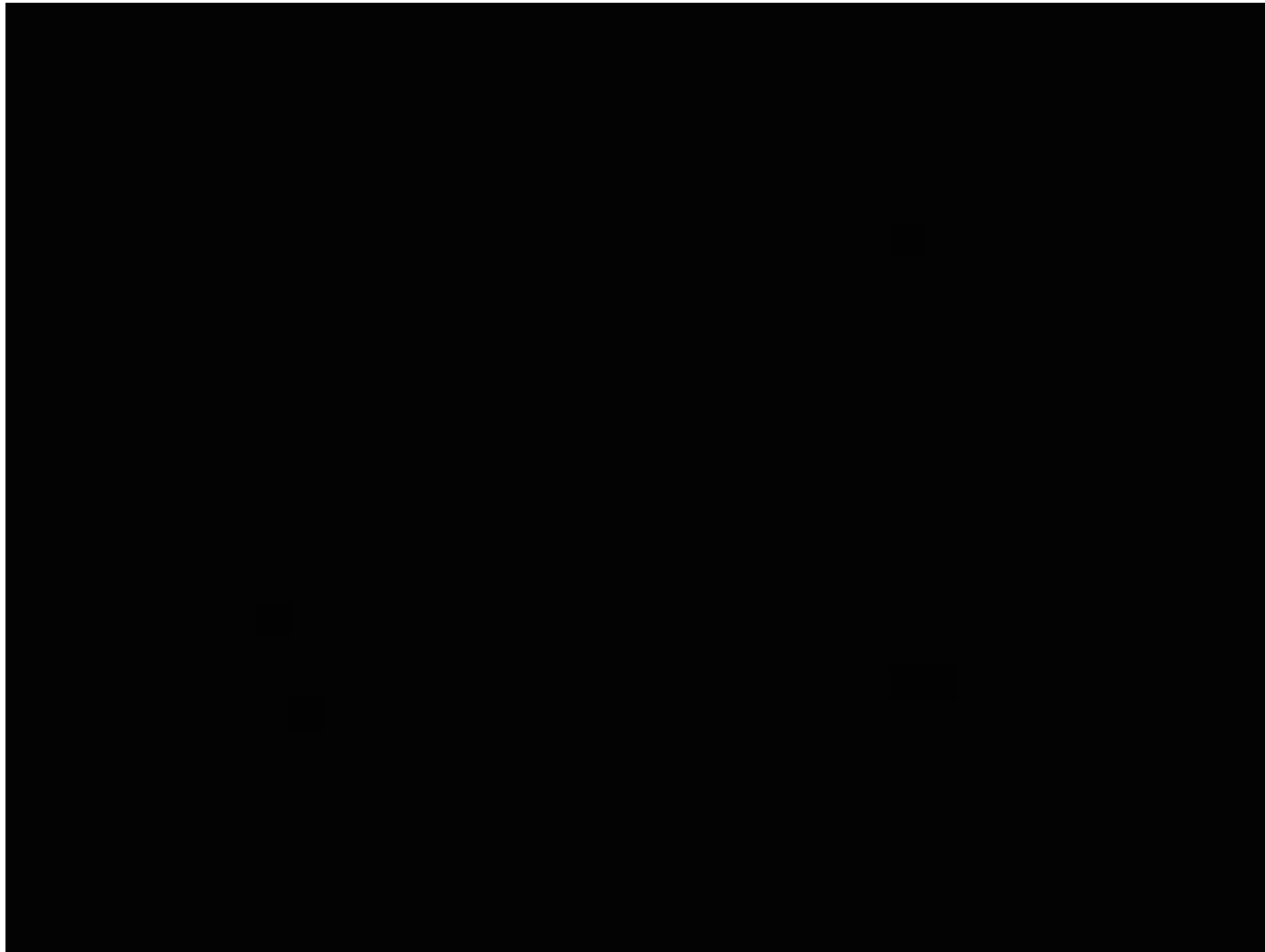


Axolotl

Regenerative Biology Laboratory, Morgridge Institute for Research, Madison, WI

# Axolotl limb regeneration

---



David Gardiner - HHMI-UCI

# Axolotl background

---



- *Ambystoma mexicanum*
- Neotenuous
- Natural habitats
  - Lake Xochimilco (canals)
  - Lake Chalco (drained)
  - Endangered
- Commonly sold as pets
- Regenerative abilities
  - Limbs
  - Portions of Heart
  - Portions of Brain
  - Tail and spinal cord

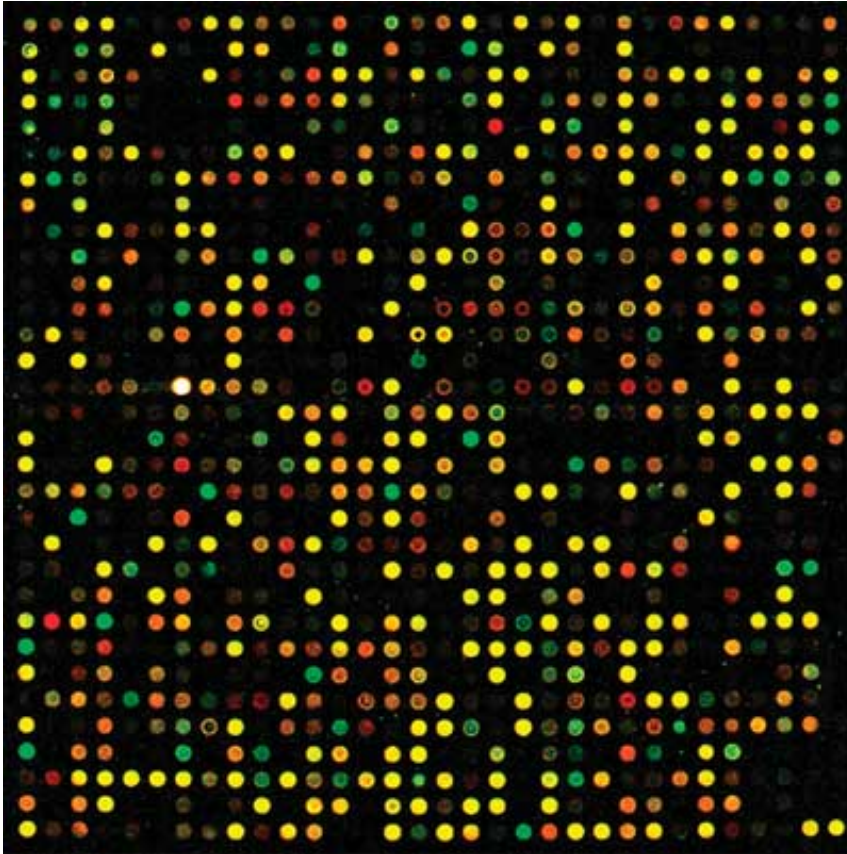
# Goals

---

- What are the axolotl **genes** that are responsible for this remarkable regenerative ability?
- Can this knowledge improve our medical treatments of severe wounds and tissue regeneration?

# Measuring transcription the old way: Microarrays

---

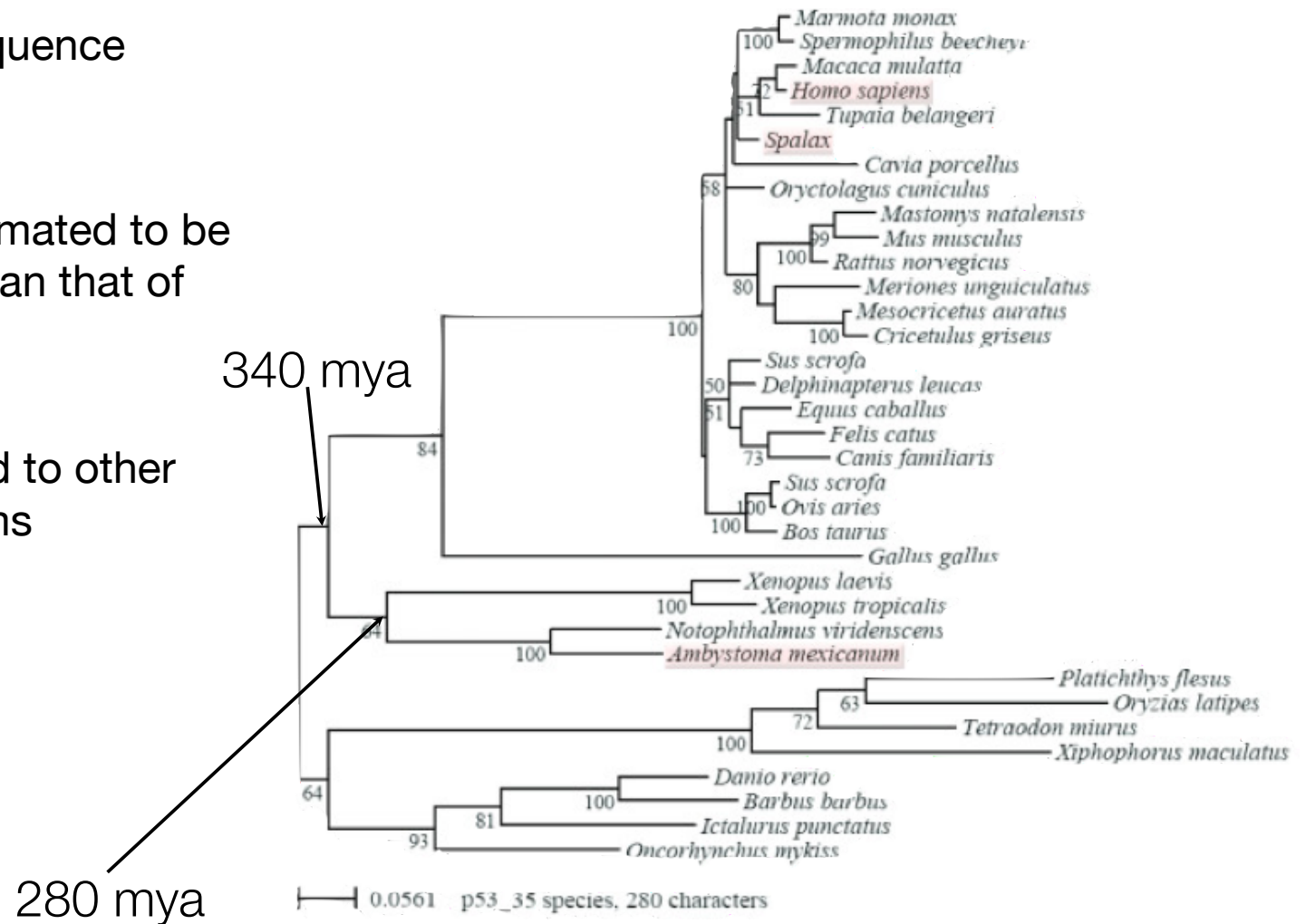


- Each spot has “probes” for a certain gene
- Probe: a DNA sequence complementary to a certain gene
- Relies on complementary hybridization
- Intensity/color of light from each spot is measurement of the number of transcripts for a certain gene in a sample
- Requires knowledge of gene sequences



# Challenges with genomic studies of Axolotl

- No genome sequence available
  - genome estimated to be 10x larger than that of human
- Distantly related to other model organisms



Villiard et al. 2007

# Prior gene expression studies in Axolotl

---

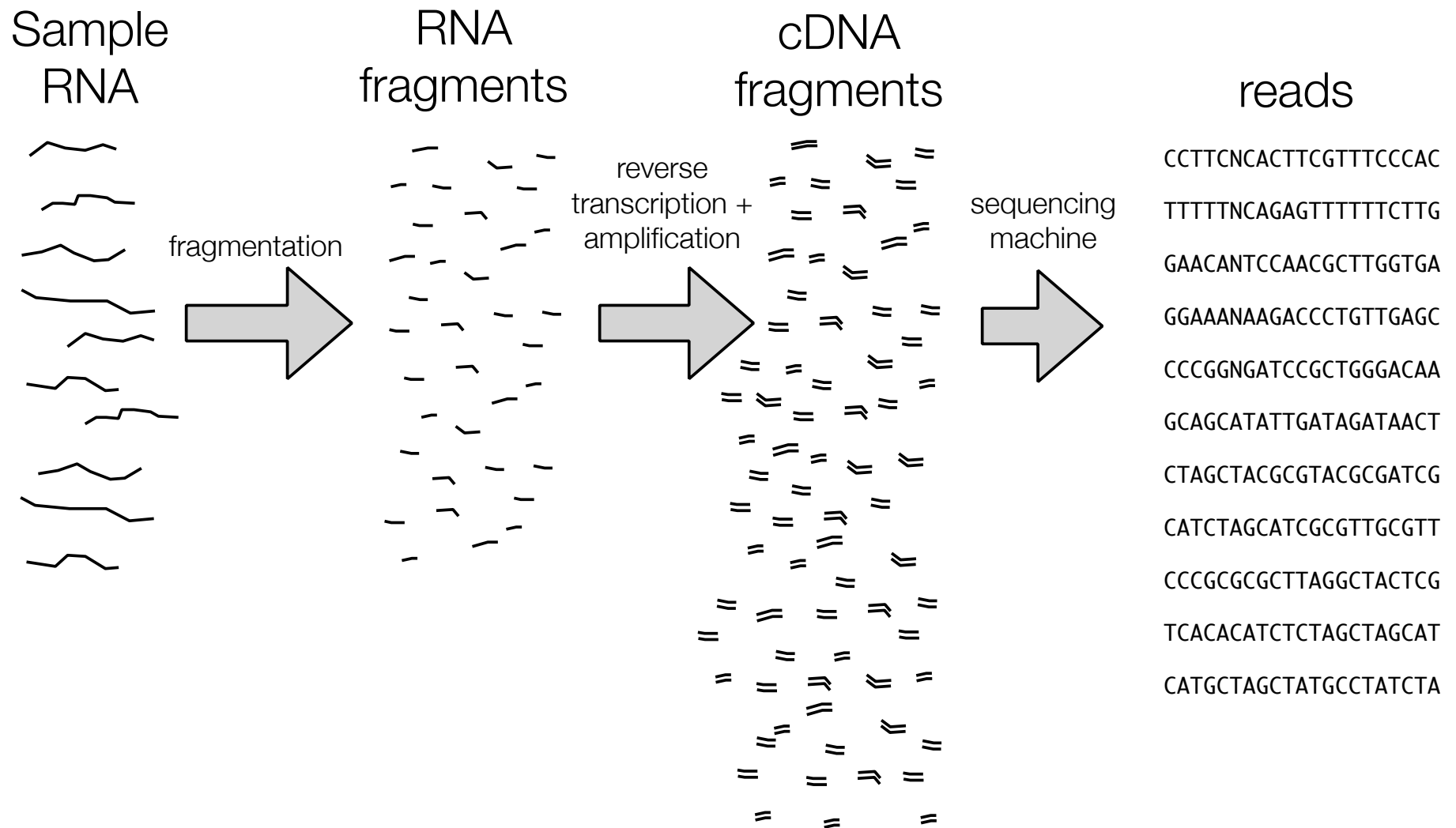
- Microarrays
  - Exist, but not very complete
  - Limited amount of mRNA sequence data from Axolotl
  - No genome, so can't use predicted gene sequences

# RNA-Seq technology

---

- Leverages rapidly advancing sequencing technology (e.g., Illumina, SOLiD)
- Transcriptome analog to whole genome shotgun sequencing
- Two key differences from genome sequencing:
  1. Transcripts sequenced at different levels of coverage - expression levels
  2. Sequences already known (in many cases) - coverage is measurement

# A generic RNA-Seq protocol



# RNA-Seq data

```
@HWUSI-EAS1789_0001:3:2:1708:1305#0/1
CCTTCNCACTTCGTTTCCCACTTAGCGATAATTTG
+HWUSI-EAS1789_0001:3:2:1708:1305#0/1
VVULVBVYVYZXZZ\ee[a^b`[a\ a[\a^^^
@HWUSI-EAS1789_0001:3:2:2062:1304#0/1
TTTTTNCAGAGTTTTTTCTTGAAGTGGAAATTTTT
+HWUSI-EAS1789_0001:3:2:2062:1304#0/1
a__[\Bbbb`edeeefd`cc`b]bffff`ffffff
@HWUSI-EAS1789_0001:3:2:3194:1303#0/1
GAACANTCCAACGCTTGGTGAATTCTGCTTCACAA
+HWUSI-EAS1789_0001:3:2:3194:1303#0/1
ZZ[[VBZZY][TWQQZ\ZS\[ZZXV__\OX`a[ZZ
@HWUSI-EAS1789_0001:3:2:3716:1304#0/1
GGAAANAAGACCCTGTTGAGCTTGACTCTAGTCTG
+HWUSI-EAS1789_0001:3:2:3716:1304#0/1
aaXWYBZVTXZX_]Xdccdfbb_\`a\ aY_^]LZ^
@HWUSI-EAS1789_0001:3:2:5000:1304#0/1
CCCGGNGATCCGCTGGGACAAGCAGCATATTGATA
+HWUSI-EAS1789_0001:3:2:5000:1304#0/1
aaaaaBeeeeffffehhhhhhggdhhhhahhhhadh
```

name  
sequence  
qualities

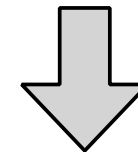
read

paired-end reads

read1

read2

1 Illumina (GAII) lane

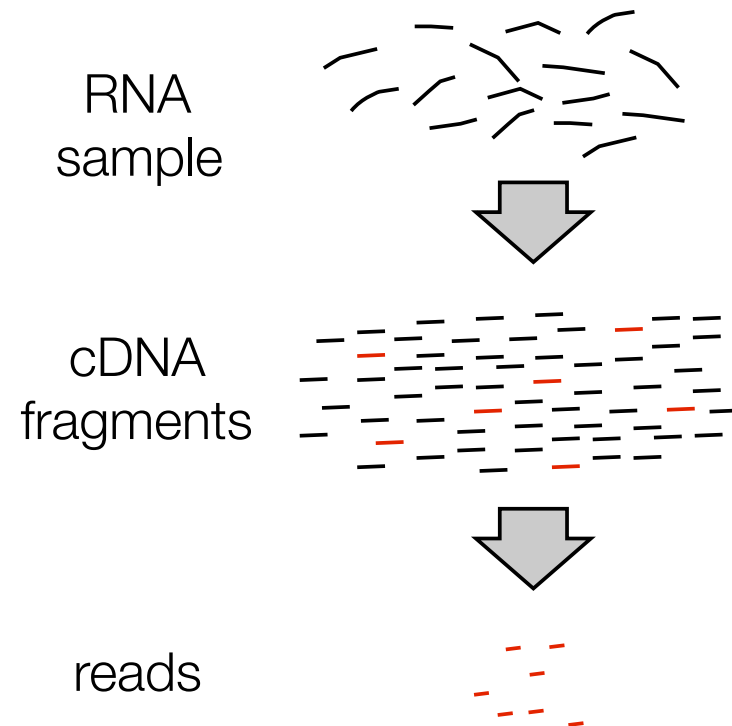


~20 million reads

# RNA-Seq is a *relative* abundance measurement technology

---

- RNA-Seq gives you reads from the ends of a random **sample** of fragments in your library
- Without additional data this only gives information about **relative** abundances
- Additional information, such as levels of “spike-in” transcripts, are needed for absolute measurements



# Issues with relative abundance measures

---

Gene	Sample 1 absolute abundance	Sample 1 relative abundance	Sample 2 absolute abundance	Sample 2 relative abundance
1	20	10%	20	5%
2	20	10%	20	5%
3	20	10%	20	5%
4	20	10%	20	5%
5	20	10%	20	5%
6	100	50%	300	75%

- Changes in absolute expression of high expressors is a major factor
- Normalization is required for comparing samples in these situations

# Advantages of RNA-Seq over microarrays

---

- No reference sequence needed
  - With microarrays, limited to the probes on the chip
- Low background noise
- Large dynamic range
  - $10^5$  compared to  $10^2$  for microarrays
- High technical reproducibility



# Tasks with RNA-Seq data

---

- Assembly:
  - Given: RNA-Seq reads (and possibly a genome sequence)
  - Do: reconstruct full-length transcript sequences from the reads
- Quantification:
  - Given: RNA-Seq reads and transcript sequences
  - Do: Estimate the relative abundances of transcripts (“gene expression”)
- Differential expression:
  - Given: RNA-Seq reads from two different samples and transcript sequences
  - Do: Predict which transcripts have different abundances between the two samples

# Public sources of RNA-Seq data

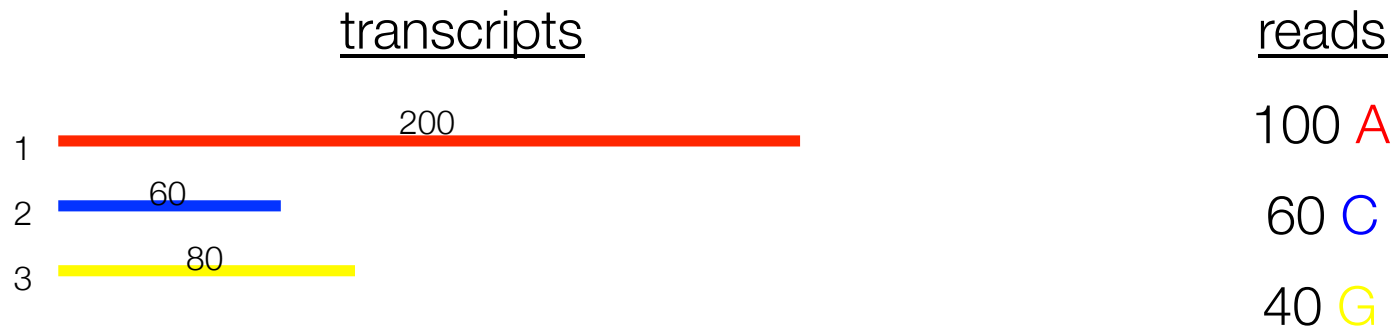
---

- Gene Expression Omnibus (GEO): <http://www.ncbi.nlm.nih.gov/geo/>
  - Both microarray and sequencing data
- Sequence Read Archive (SRA): <http://www.ncbi.nlm.nih.gov/sra>
  - All sequencing data (not necessarily RNA-Seq)
- ArrayExpress: <https://www.ebi.ac.uk/arrayexpress/>
  - European version of GEO
- All of these have links between them

# The basics of quantification with RNA-Seq data

---

- For simplicity, suppose reads are of length **one** (typically they are > 35 bases)






- What relative abundances would you estimate for these genes?

# Length dependence

---

- probability of a read coming from a transcript  $\propto$  relative abundance  $\times$  length

	<u>transcripts</u>	<u>reads</u>
1	 200	100 <span style="color: red;">A</span>
2	 60	60 <span style="color: blue;">C</span>
3	 80	40 <span style="color: yellow;">G</span>

$$\hat{f}_1 \propto \frac{\frac{100}{200}}{200} = \frac{1}{400}$$

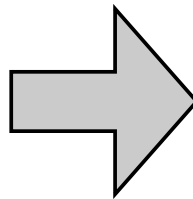
$$\hat{f}_1 = 0.25$$

$$\hat{f}_2 \propto \frac{\frac{60}{200}}{60} = \frac{1}{200}$$

$$\hat{f}_2 = 0.5$$

$$\hat{f}_3 \propto \frac{\frac{40}{200}}{80} = \frac{1}{400}$$

$$\hat{f}_3 = 0.25$$



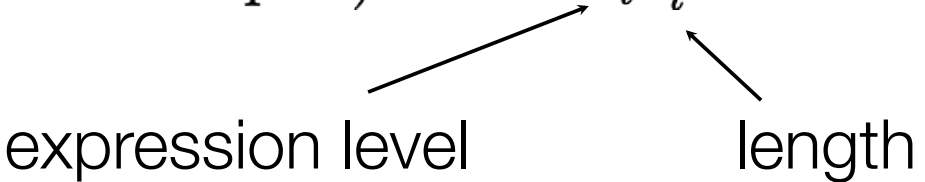
# The basics of quantification from RNA-Seq data

---

- Basic assumption:

$$\theta_i = P(\text{read from transcript } i) = Z^{-1} \tau_i \ell'_i$$

expression level                      length



- Normalization factor is the mean length of expressed transcripts

$$Z = \sum_i \tau_i \ell'_i$$

# The basics of quantification from RNA-Seq data

---

- Estimate the probability of reads being generated from a given transcript by counting the number of reads that align to that transcript

$$\hat{\theta}_i = \frac{c_i}{N}$$

← # reads mapping to transcript i

← total # of mappable reads

- Convert to expression levels by normalizing by transcript length

$$\hat{\tau}_i \propto \frac{\hat{\theta}_i}{\ell'_i}$$

# The basics of quantification from RNA-Seq data

---

- Basic quantification algorithm
  - Align reads against a set of reference transcript sequences
  - Count the number of reads aligning to each transcript
  - Convert read counts into relative expression levels

# Counts to expression levels

---

- RPKM - **R**eads **P**er **K**ilobase per **M**illion mapped reads

$$\text{RPKM for gene } i = 10^9 \times \frac{c_i}{\ell'_i N}$$

- TPM - **T**ranscripts **P**er **M**illion

$$(\text{estimate of}) \text{ TPM for isoform } i = 10^6 \times Z \times \frac{c_i}{\ell'_i N}$$

- Prefer TPM to RPKM/FPKM because of normalization factor
- TPM is a technology-independent measure (simply a fraction)



# What if reads do not uniquely map to transcripts?

---

- The approach described assumes that every read can be uniquely aligned to a single transcript
- This is generally not the case
  - Some genes have similar sequences - gene families, repetitive sequences
  - Alternative splice forms of a gene share a significant fraction of sequence

# Multi-mapping reads in RNA-Seq

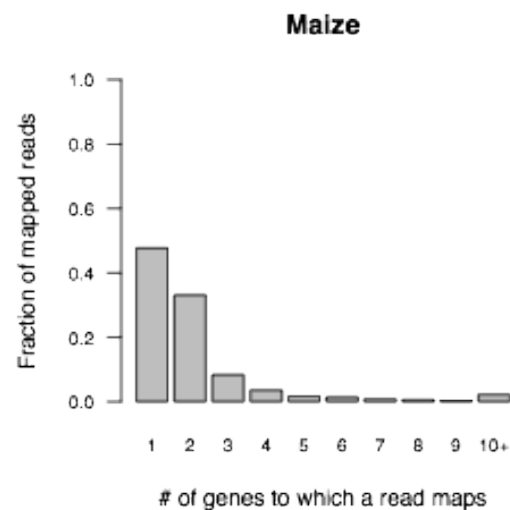
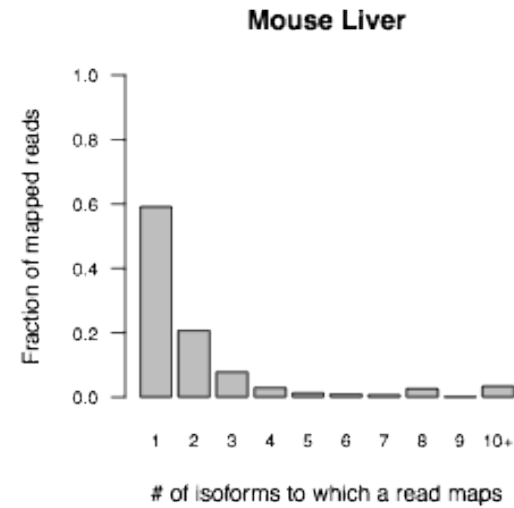
---

Species	Read length	% multi-mapping reads
Mouse	25	17%
Mouse	75	10%
Maize	25	52%
Axolotl	76	23%

- Throwing away multi-mapping reads leads to
  1. Loss of information
  2. Potentially biased estimates of abundance

# Distributions of alignment counts

---



# What if reads do not uniquely map to transcripts?

---

- “multiread”: a read that could have been derived from multiple transcripts



- How would you estimate the relative abundances for these transcripts?

# Some options for handling multireads

---

- Discard all multireads, estimate based on uniquely mapping reads only
- Discard multireads, but use “unique length” of each transcript in calculations
- “Rescue” multireads by allocating (fractions of) them to the transcripts
  - Three step algorithm
    1. Estimate abundances based on uniquely mapping reads only
    2. For each multiread, divide it between the transcripts to which it maps, proportionally to their abundances estimated in the first step
    3. Recompute abundances based on updated counts for each transcript

# Rescue method example - Step 1

---

	<u>transcripts</u>	<u>reads</u>
1	 200	90 <span style="color: red;">A</span>
2	 60	40 <span style="color: blue;">C</span>
3	 80	40 <span style="color: yellow;">G</span>
		30 <span style="color: green;">T</span>

Step 1

$$\hat{f}_1^{unique} = \frac{\frac{90}{200}}{\frac{90}{200} + \frac{40}{60} + \frac{40}{80}} = 0.278$$

$$\hat{f}_2^{unique} = 0.412$$

$$\hat{f}_3^{unique} = 0.309$$

## Rescue method example - Step 2

---



Step 2

$$c_1^{rescue} = 90 + 30 \times \frac{0.278}{0.278 + 0.412} = 102.1$$

$$c_2^{rescue} = 40 + 30 \times \frac{0.412}{0.278 + 0.412} = 57.9$$

$$c_3^{rescue} = 40 + 0 = 40$$

## Rescue method example - Step 3

---

	<u>transcripts</u>	<u>reads</u>
1	 200	90 <span style="color: red;">A</span>
2	 60	40 <span style="color: blue;">C</span>
3	 80	40 <span style="color: yellow;">G</span>
		30 <span style="color: green;">T</span>

$$\begin{aligned}
 \hat{f}_1^{\text{rescue}} &= \frac{\frac{102.1}{200}}{\frac{102.1}{200} + \frac{57.9}{60} + \frac{40}{80}} = 0.258 \\
 \hat{f}_2^{\text{rescue}} &= \frac{\frac{57.9}{60}}{\frac{102.1}{200} + \frac{57.9}{60} + \frac{40}{80}} = 0.488 \\
 \hat{f}_3^{\text{rescue}} &= \frac{\frac{40}{80}}{\frac{102.1}{200} + \frac{57.9}{60} + \frac{40}{80}} = 0.253
 \end{aligned}$$

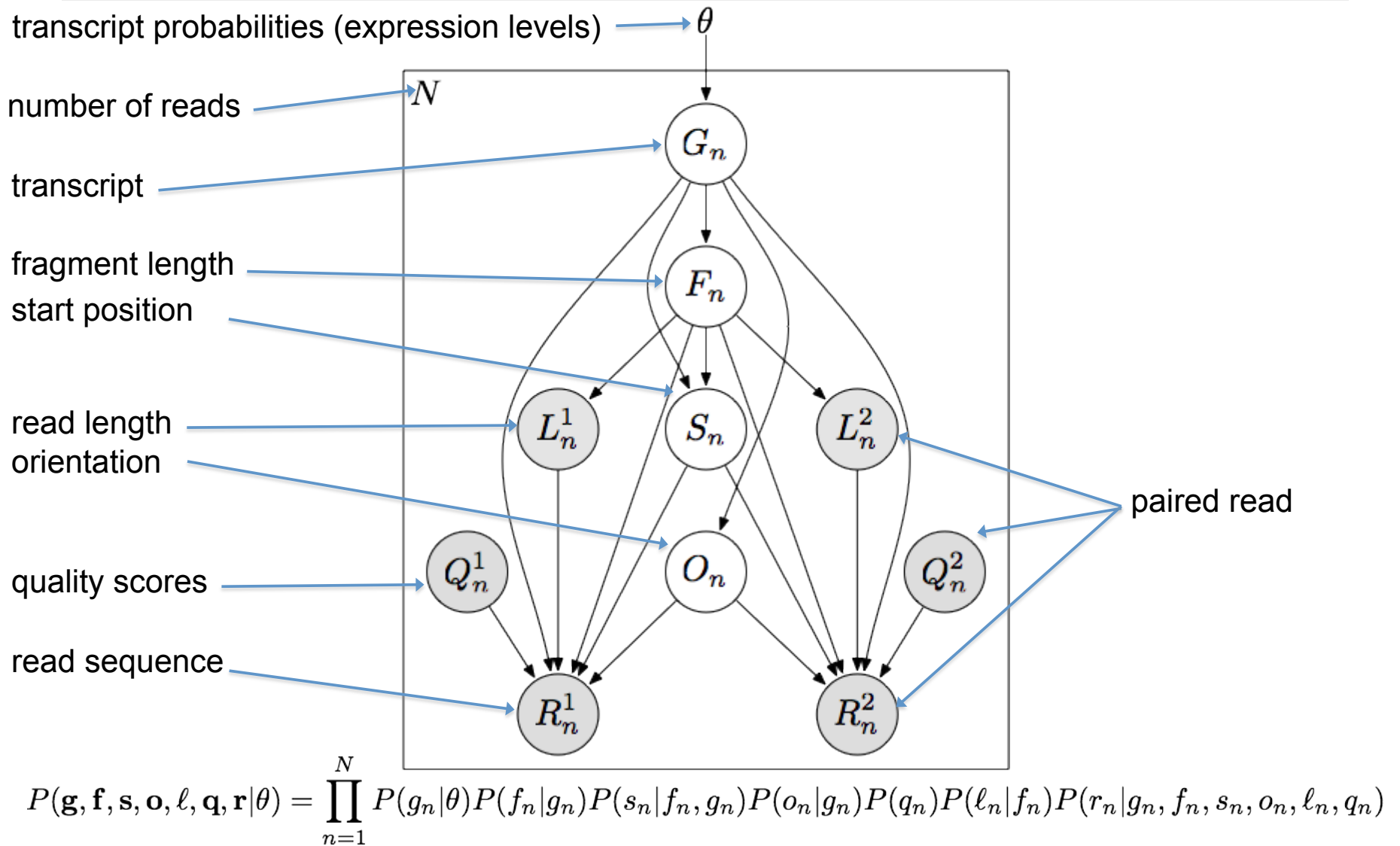


# An observation about the rescue method

---

- Note that at the end of the rescue algorithm, we have an updated set of abundance estimates
- These new estimates could be used to reallocate the multireads
- And then we could update our abundance estimates once again
- And repeat!
- This is the intuition behind the statistical approach to this problem

# Our solution - a generative probabilistic model



# Quantification as maximum likelihood inference

---

- Observed data likelihood

$$P(\mathbf{r}, \ell, \mathbf{q} | \theta) = \prod_{n=1}^N \sum_{i=0}^M \theta_i \sum_{j=0}^{L_i} \sum_{k=0}^{L_i} \sum_{o=0}^1 P(R_n = r_n, L_n = \ell_n, Q_n = q_n, S_n = j, F_n = k, O_n = o | G_n = i)$$

- Likelihood function is concave w.r.t.  $\theta$ 
  - Has a global maximum (or global maxima)
- Expectation-Maximization for optimization

*“RNA-Seq gene expression estimation with read mapping uncertainty”*

Li, B., Ruotti, V., Stewart, R., Thomson, J., Dewey, C.

Bioinformatics, 2010

# Approximate inference with read alignments

---

$$P(\mathbf{r}, \ell, \mathbf{q} | \theta) = \prod_{n=1}^N \sum_{i=0}^M \theta_i \sum_{j=0}^{L_i} \sum_{k=0}^{L_i} \sum_{o=0}^1 P(R_n = r_n, L_n = \ell_n, Q_n = q_n, S_n = j, F_n = k, O_n = o | G_n = i)$$

- Full likelihood computation requires  $O(NML^2)$  time

- N (number of reads)  $\sim 10^7$

- M (number of transcripts)  $\sim 10^4$

- L (average transcript length)  $\sim 10^3$

- Approximate by alignment

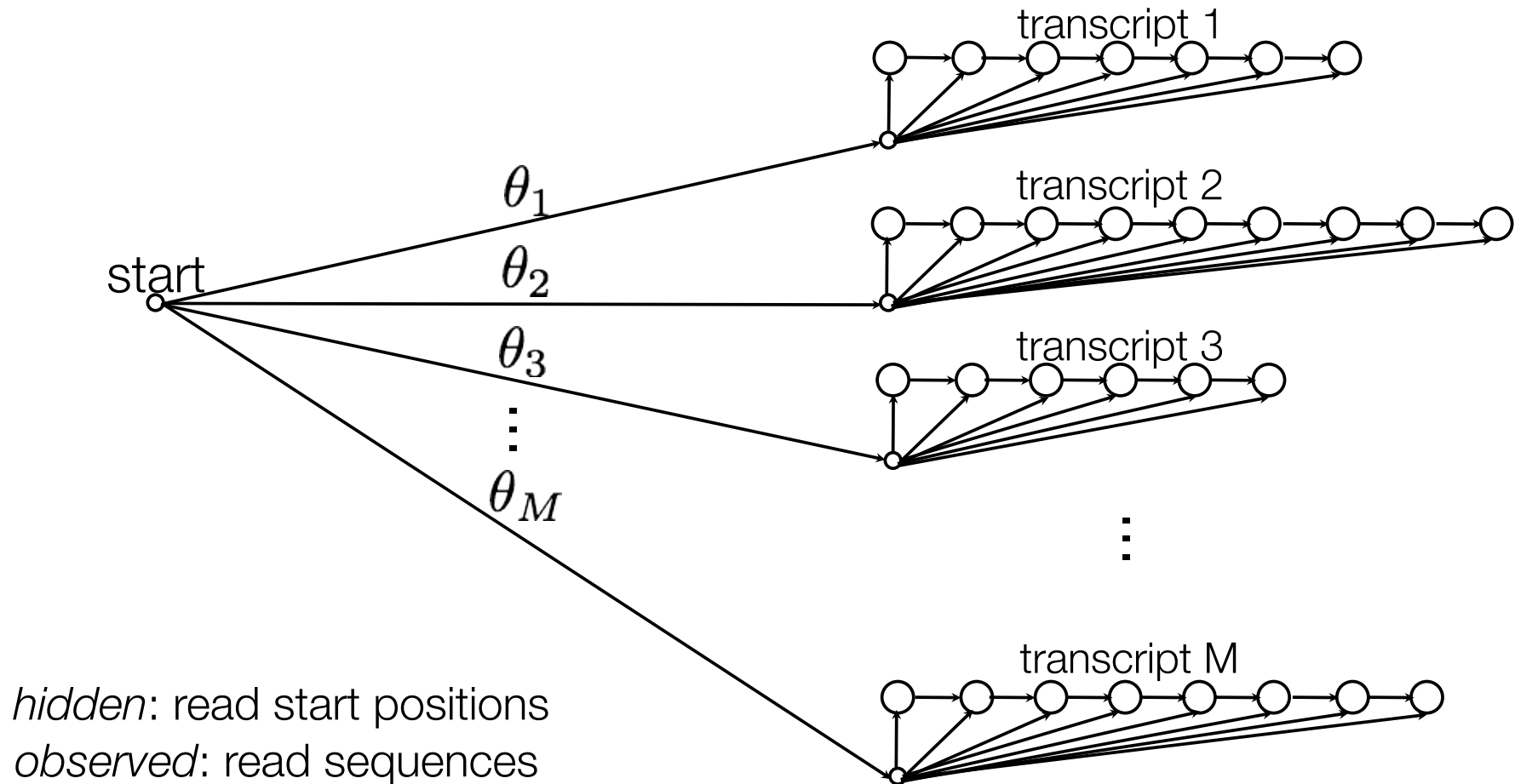
$$P(\mathbf{r}, \ell, \mathbf{q} | \theta) = \prod_{n=1}^N \sum_{(i,j,k,o) \in \pi_n^x} \theta_i P(R_n = r_n, L_n = \ell_n, Q_n = q_n, Z_{nijko} = 1 | G_n = i)$$



all local alignments of read n with at most x mismatches

# HMM Interpretation

---



Learning parameters: Baum-Welch Algorithm (EM for HMMs)  
Approximation: Only consider a subset of paths for each read

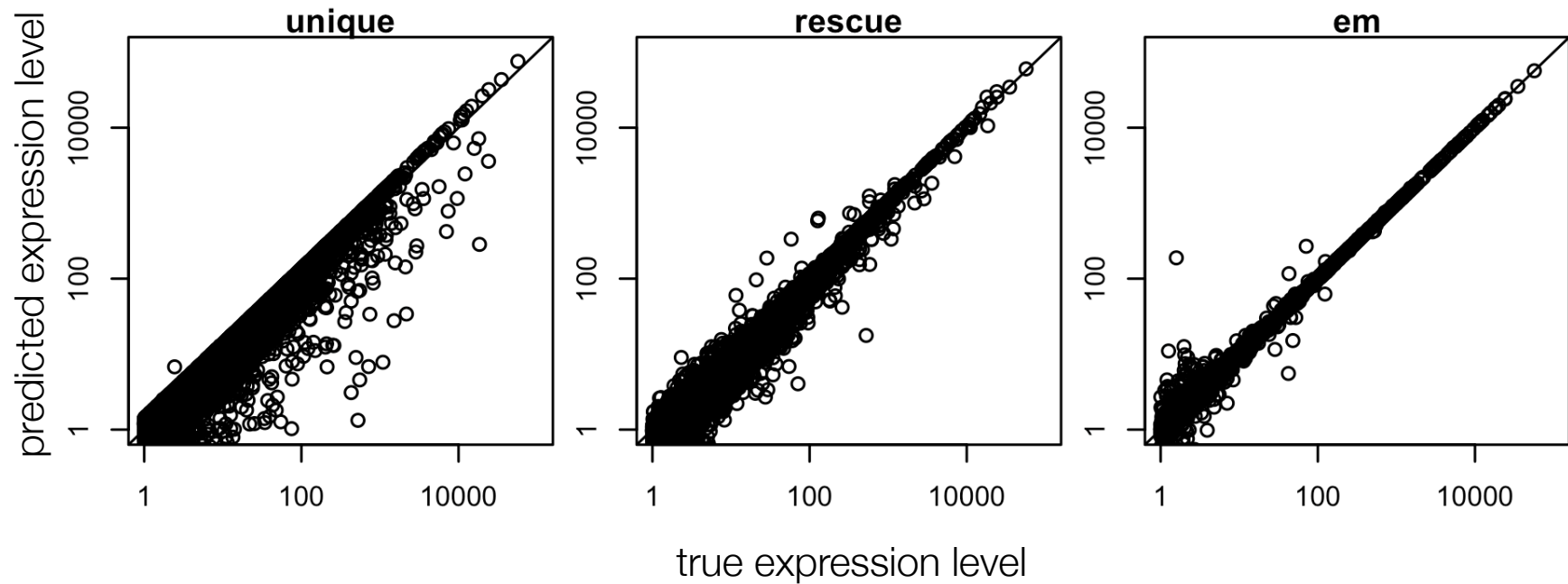
# EM Algorithm

---

- Expectation-Maximization for RNA-Seq
  - E-step: Compute expected read counts given current expression levels
  - M-step: Compute expression values maximizing likelihood given expected read counts
- Rescue algorithm  $\approx$  1 iteration of EM

# Improved accuracy over unique and rescue

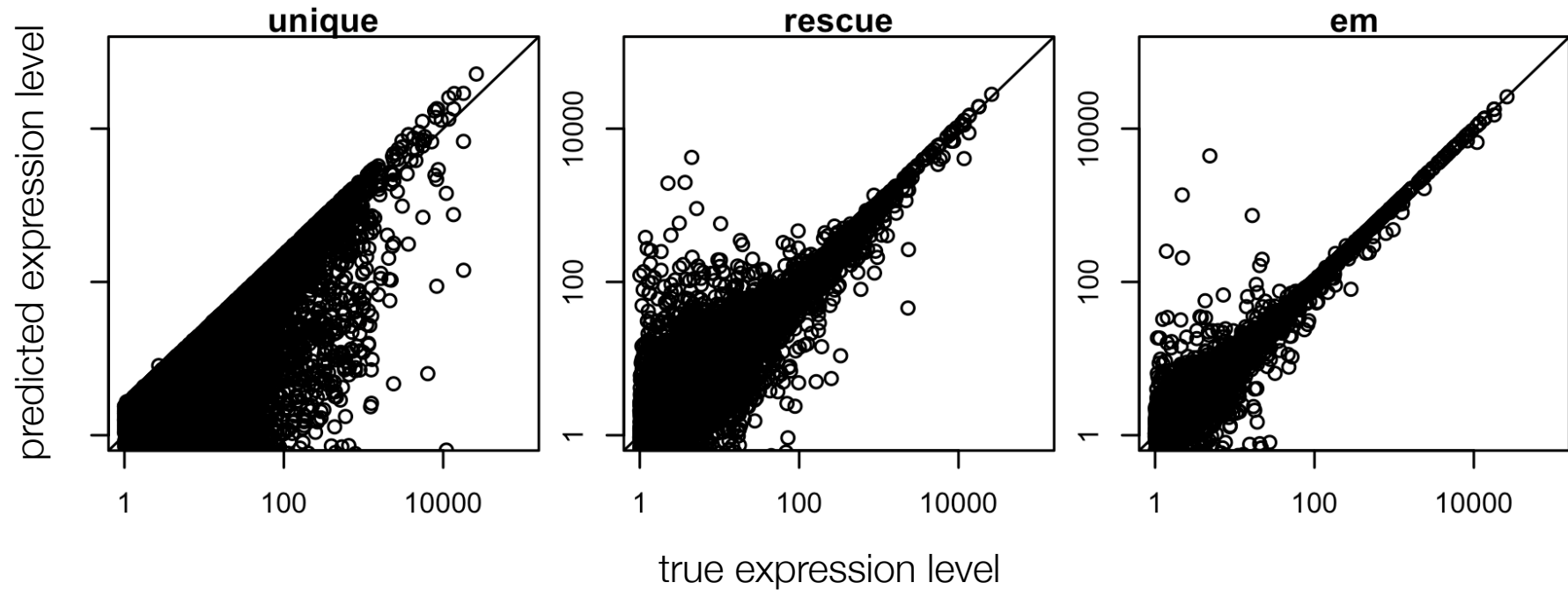
---



Gene-level expression estimation

# Improving accuracy on repetitive genomes: maize

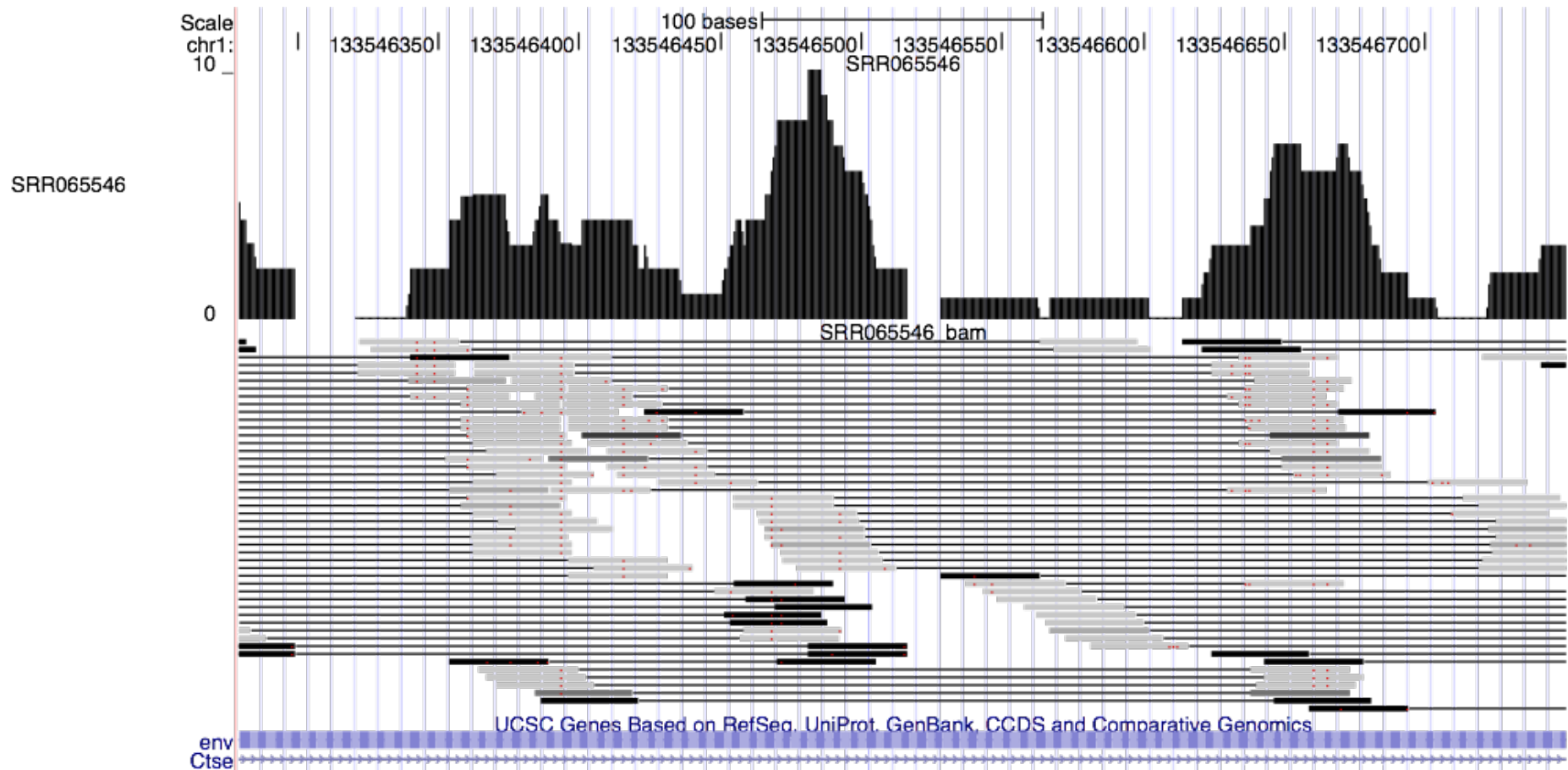
---



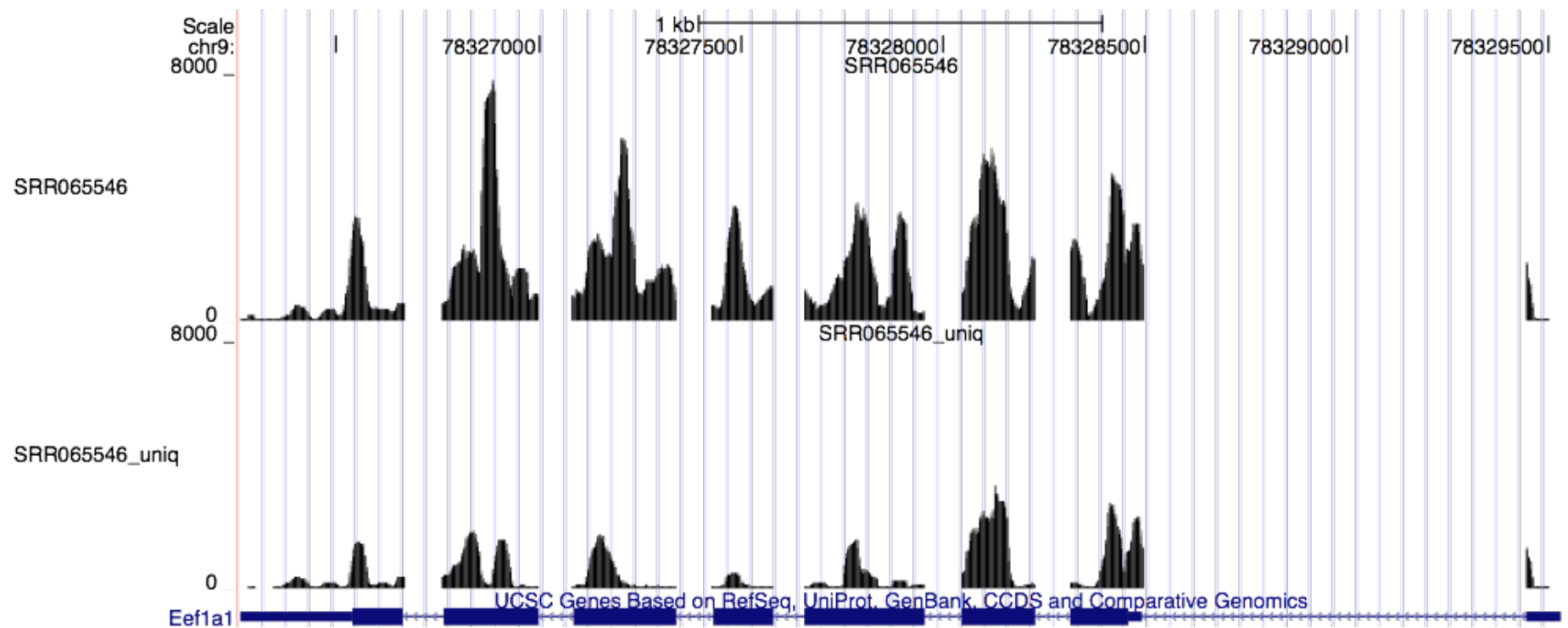
Gene-level expression estimation



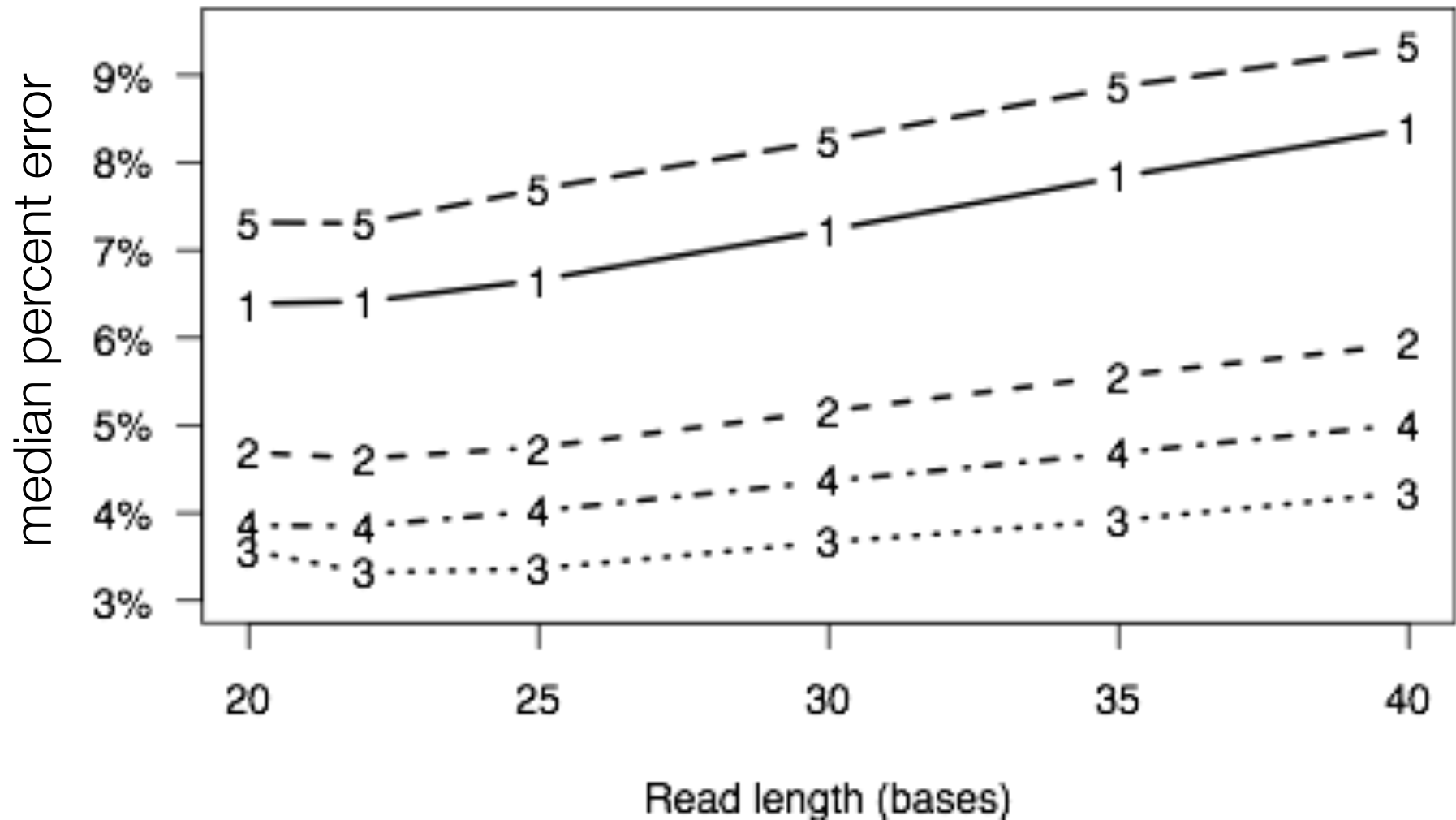
# Probabilistically-weighted alignments



# Expected read count visualization



## Finding the optimal read length



# Axolotl experimental setup

---

## Samples

Stylopod (upper arm) (3)

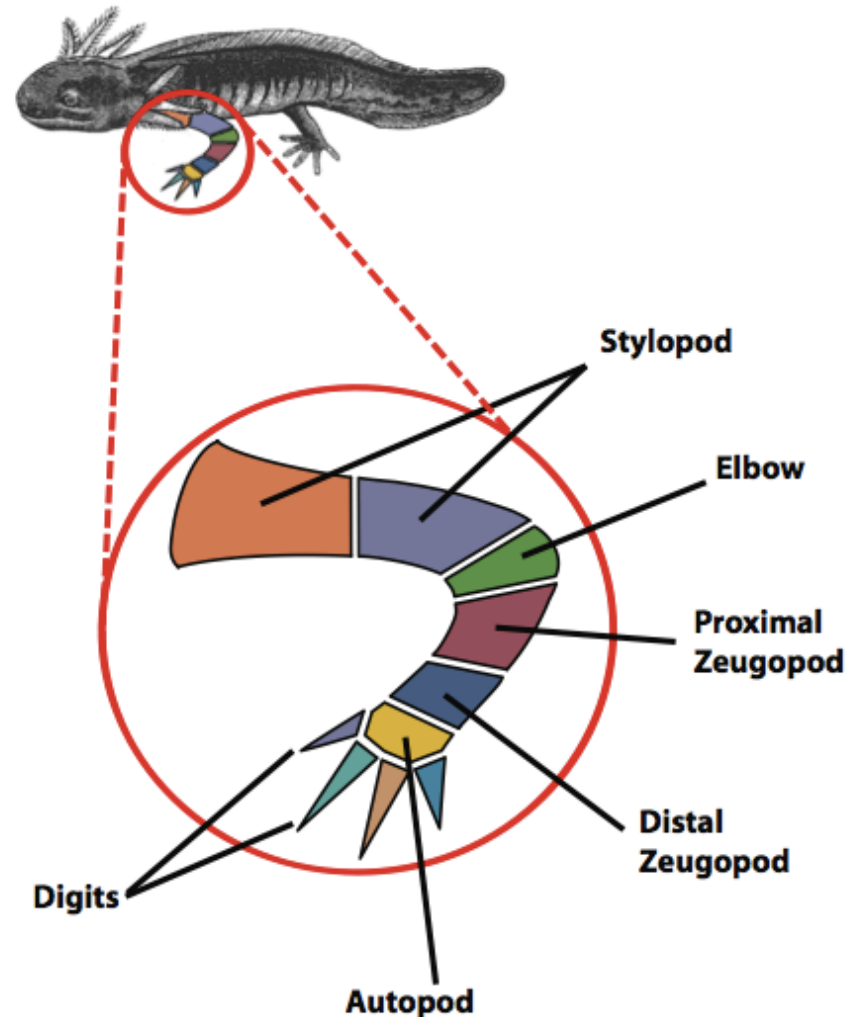
Zeugopod (lower arm) (3)

Autopod (hand) (3)

Digits (3)

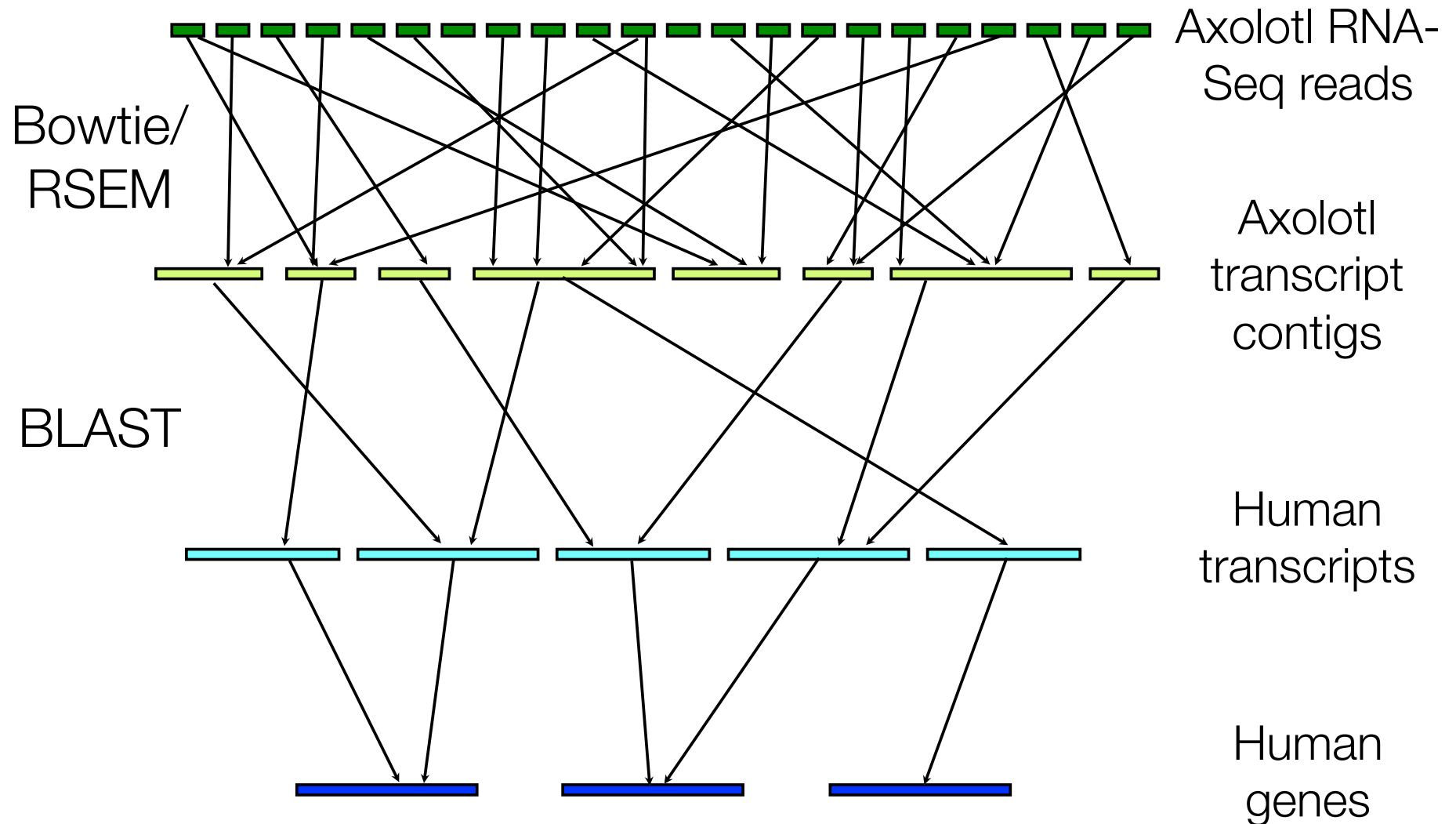
30 day blastema (5)

*Comparative RNA-seq analysis in the unsequenced axolotl: The oncogene burst highlights early gene expression in the blastema*  
R. Stewart, C. Rascón, S. Tian, J. Nie, C. Barry, L. Chu, R. Wagner, M. Probasco, J. Bolin, N. Leng, S. Sengupta, M. Volkmer, B. Habermann, E. Tanaka, J. Thomson, and C. Dewey  
*PLoS Computational Biology. In press.*



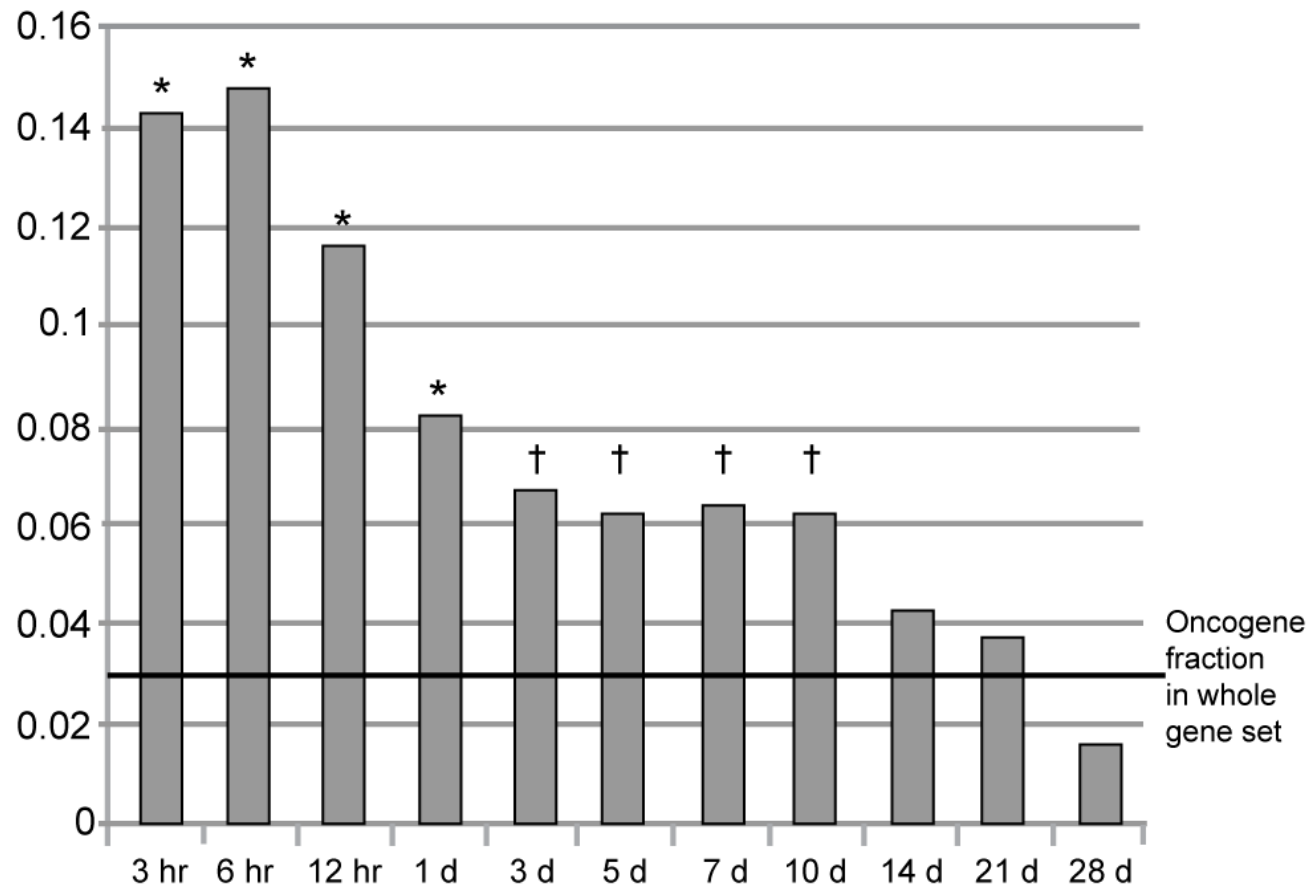
# Human-based analysis of axolotl transcription

---



# The oncogene burst

Fraction of Upregulated Genes That Are Oncogenes

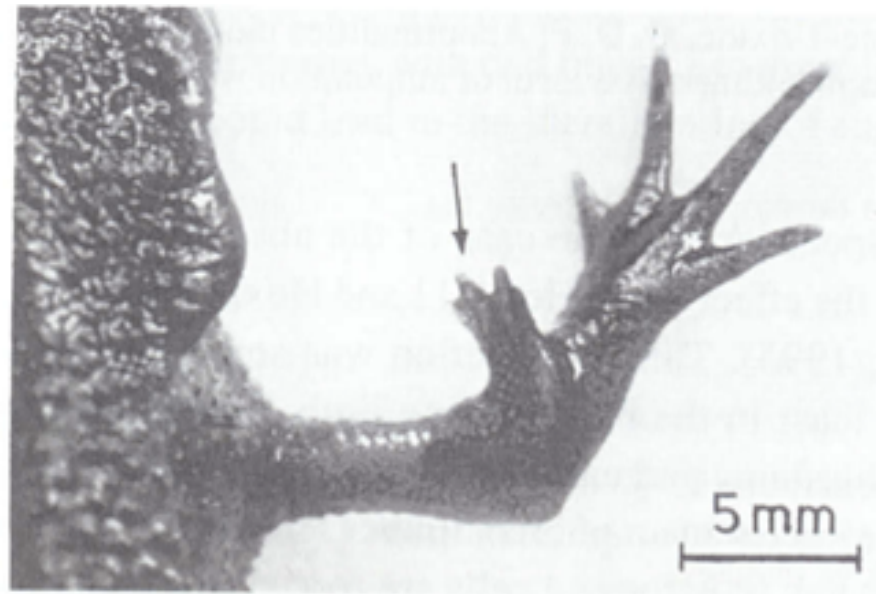


\* P value < 1e-5 by Fisher's exact test

† P value < 0.05 by Fisher's exact test

# Regeneration as controlled cancer

---



**Figure 11.1** Induction of supernumerary limb formation in the Japanese newt *Cynops pyrrhogaster* by carcinogen treatment. The carcinogen used was N-methyl-N'-nitro-N-nitrosoguanidine.

P Tsonis, Limb Regeneration, 1996, Cambridge University Press

Limb Regeneration -- Oncogenes and tumor suppressors  
“Controlled Cancer” --> development and differentiation  
Salamanders very resistant to tumorigenesis by carcinogens

# Summary

---

- **RNA-Seq** is likely the future of transcriptome analysis
- The major challenge in analyzing RNA-Seq data: the reads are much **shorter** than the transcripts from which they are derived
- Tasks with RNA-Seq data thus require handling **hidden** information: which gene/isoform gave rise to a given read
- The **Expectation-Maximization** algorithm is extremely powerful in these situations