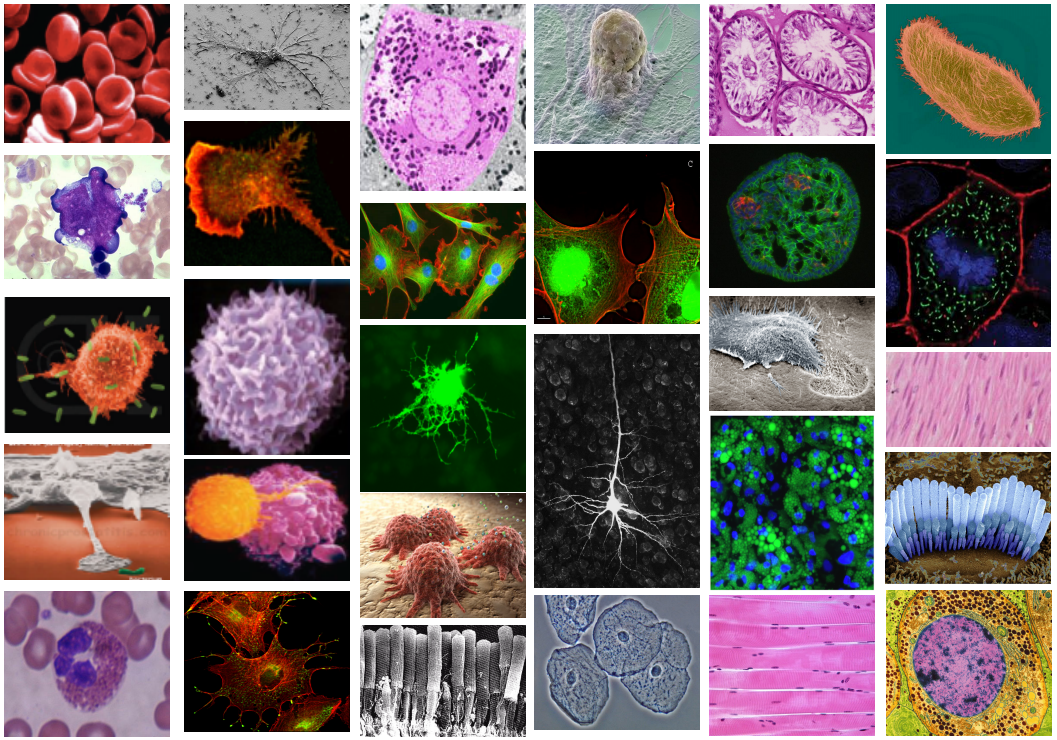
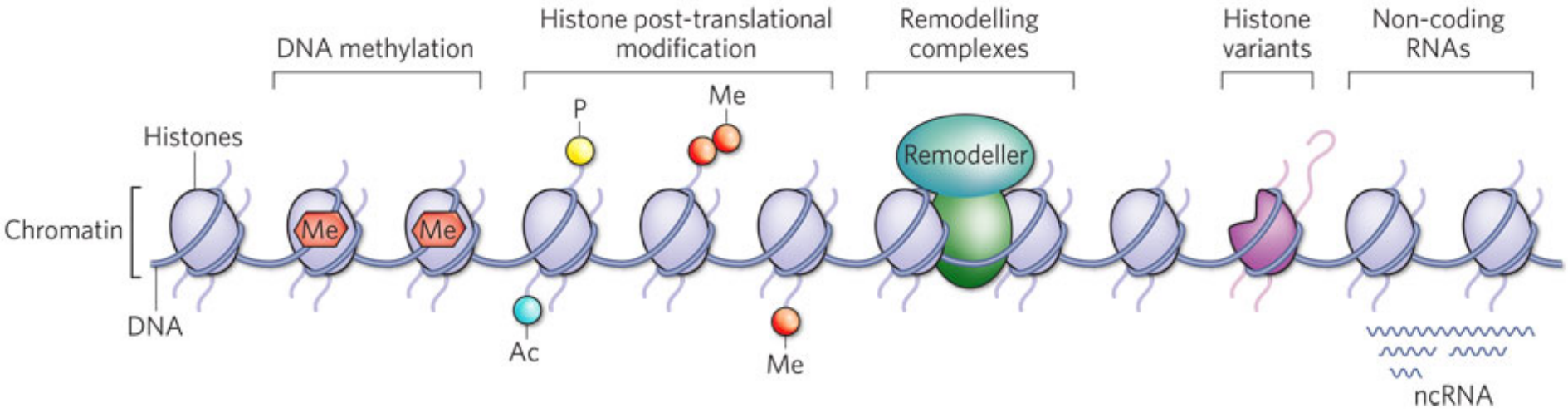
# Short read sequence analysis

Manuel Garber

# How does a single genome gives rise to more than 200 different cells?

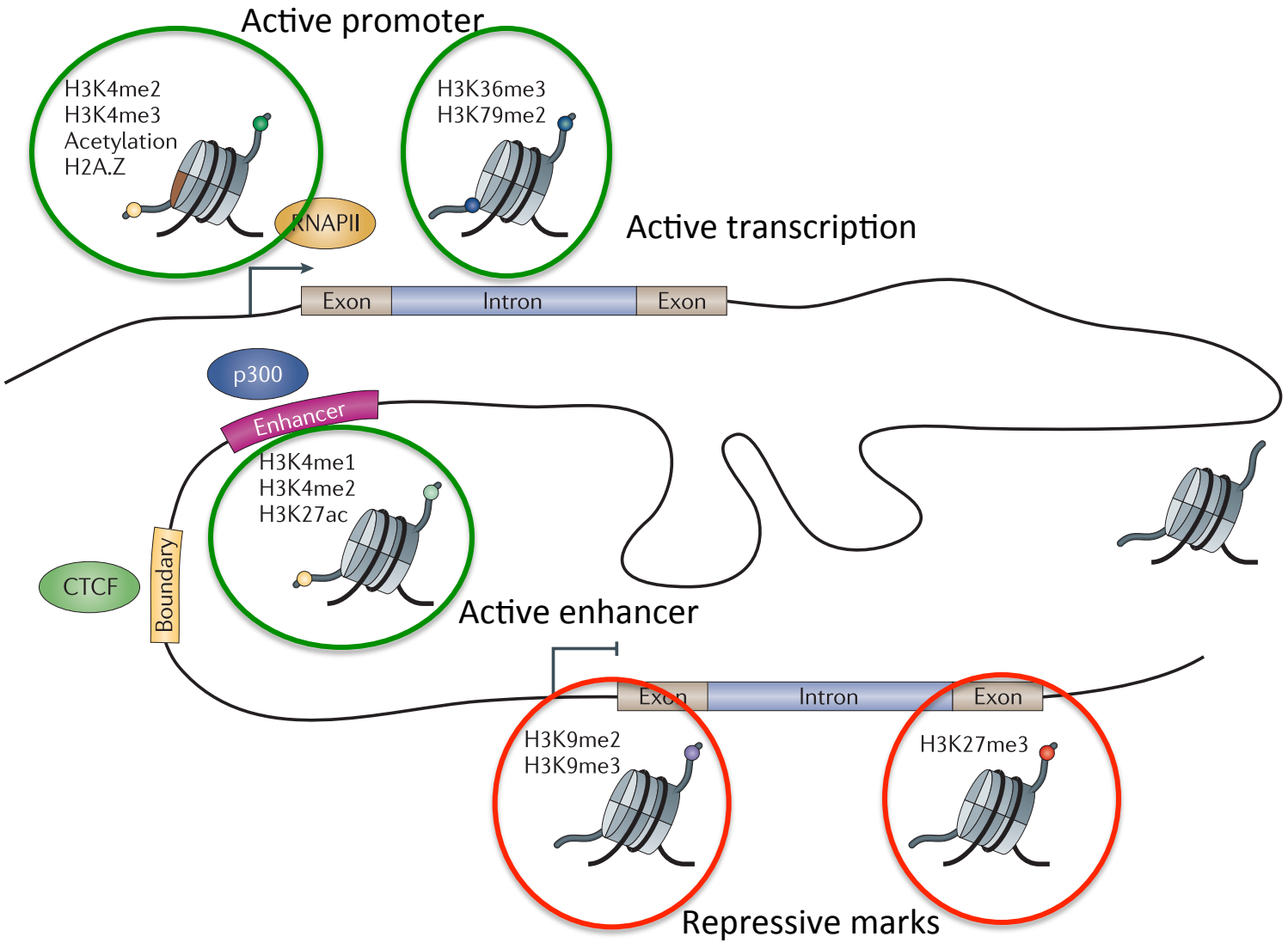


# Cell identity is determined by its epigenetic state




Catherine Dulac, Nature 2010

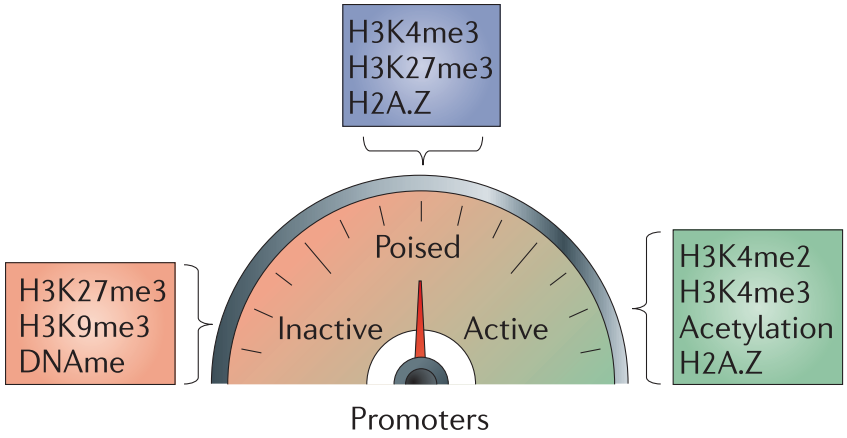
# Which controls the genome functional elements

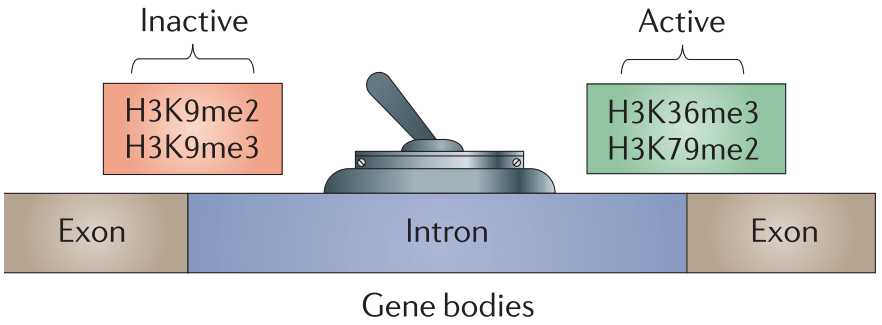



**Motivation: find the genome state using sequencing data**

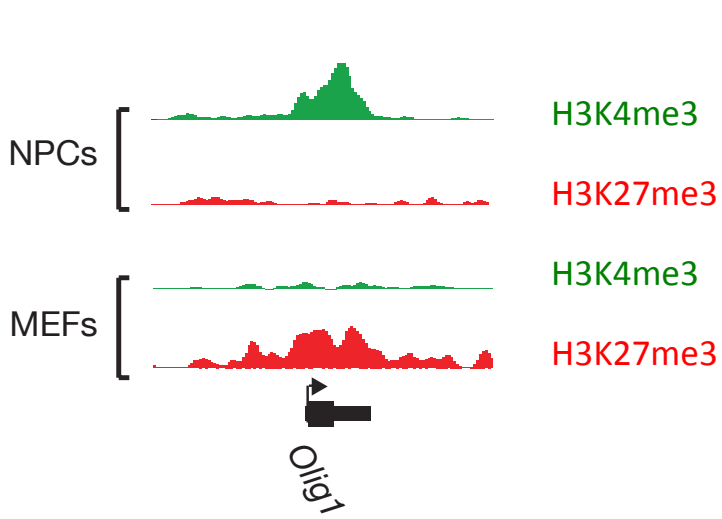
Zhou, Goren Berenstein, Nature Rev. Genetics 2011

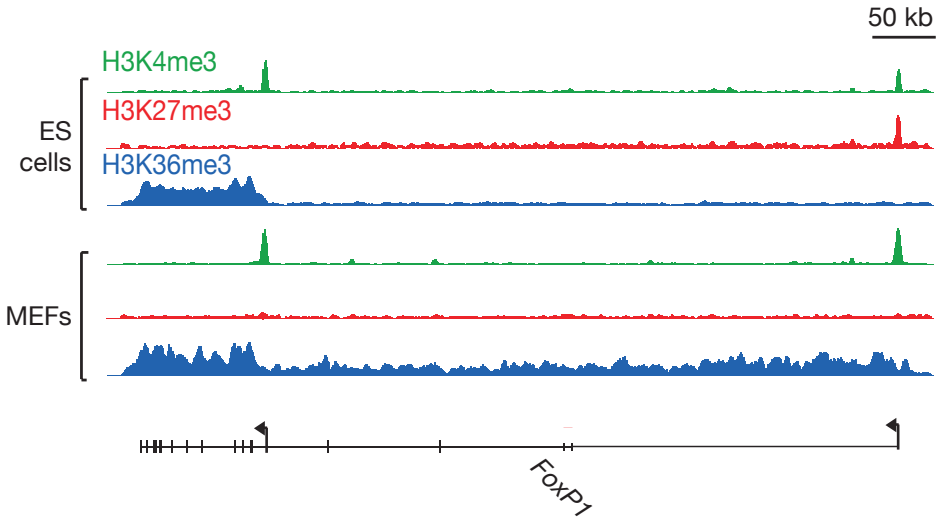
# How does it work?



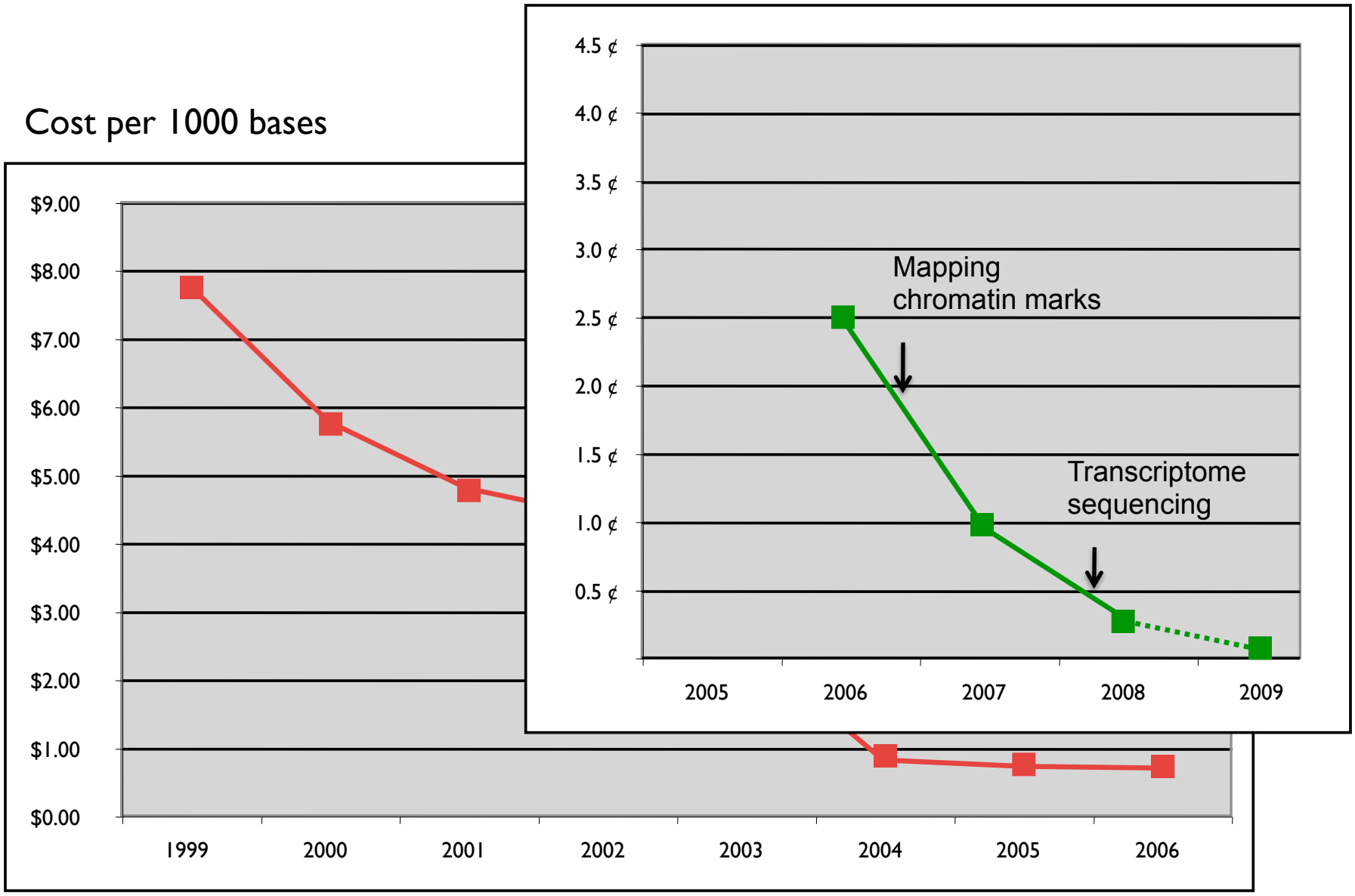

Zhou, Goren Berenstein, Nature Rev. Genetics 2011



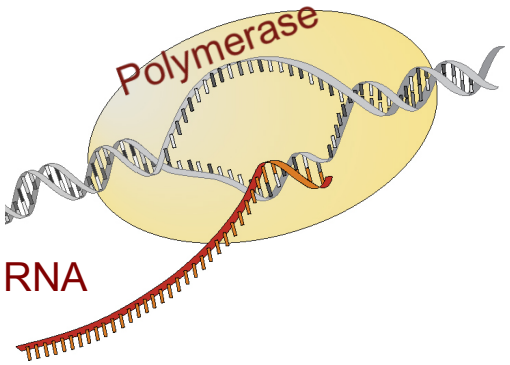

Mikkelsen et al, Nature 2007

# Enabler: Drop in cost of sequening

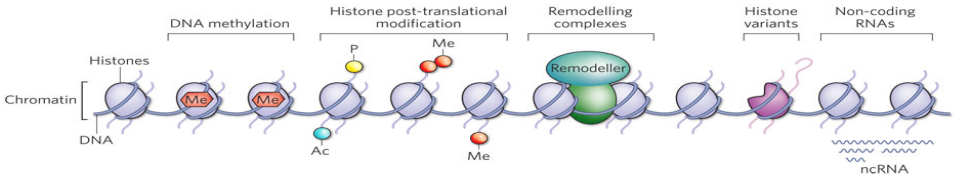


# Goal: Find the genome state and output

- Transcriptomics (output)




- Epigenomics (state)
  - Open promoters (H3K4me3)
  - Active enhancers (H3K4me1, H3K27Ac)
  - Transcribed regions (PolII, H3K36me3)
  - Repressed genes (H3K27me3)




Catherine Dulac, Nature 2010

**The goal of this session is to survey computational tools to analyze sequencing data to measure state and output**

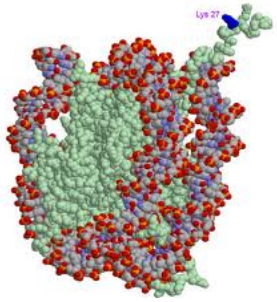
## Which require three main approaches

We'll cover the  3 main computational challenges of sequence analysis for *counting applications*:

- Read mapping (alignment): Placing short reads in the genome
- Reconstruction: Finding the regions that originated the reads
- Quantification:
    - Assigning scores to regions
    - Finding regions that are differentially represented between two or more samples.

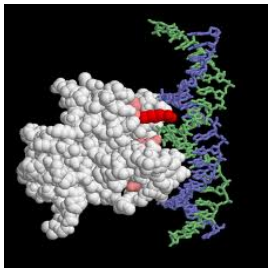
# 1. ChIP-Seq: Genome state

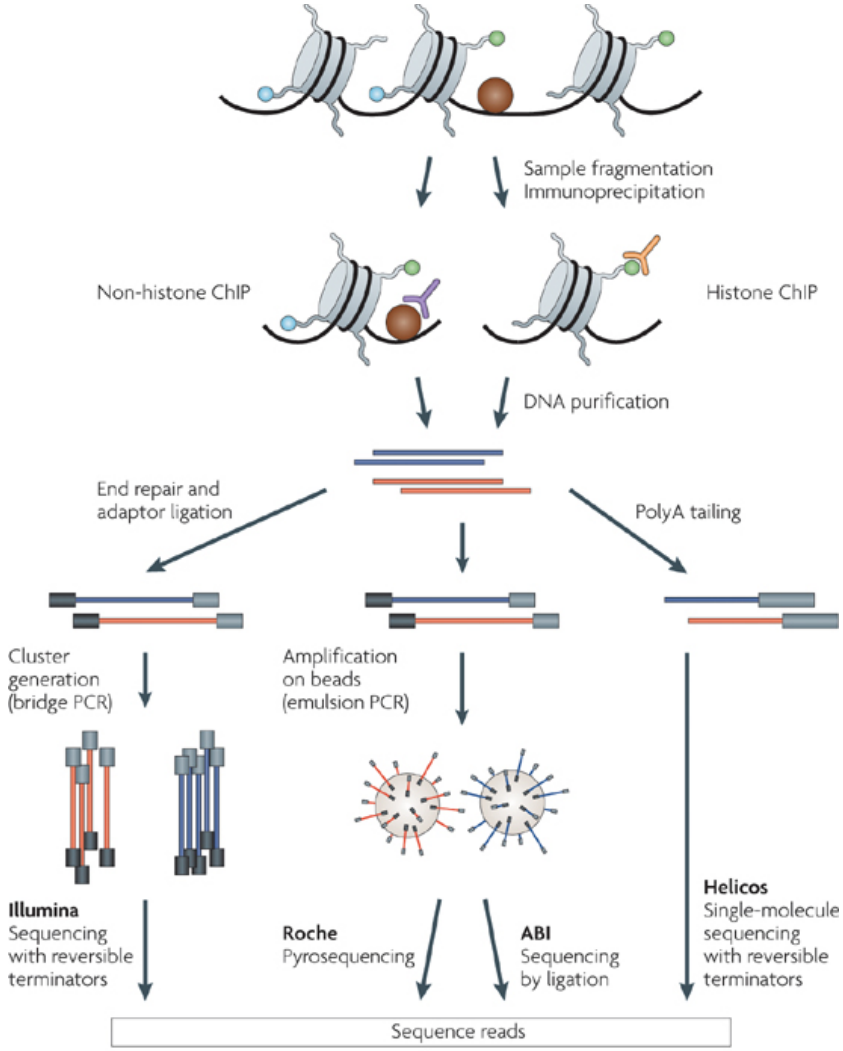
Histone Marks



**ChIP-Seq**

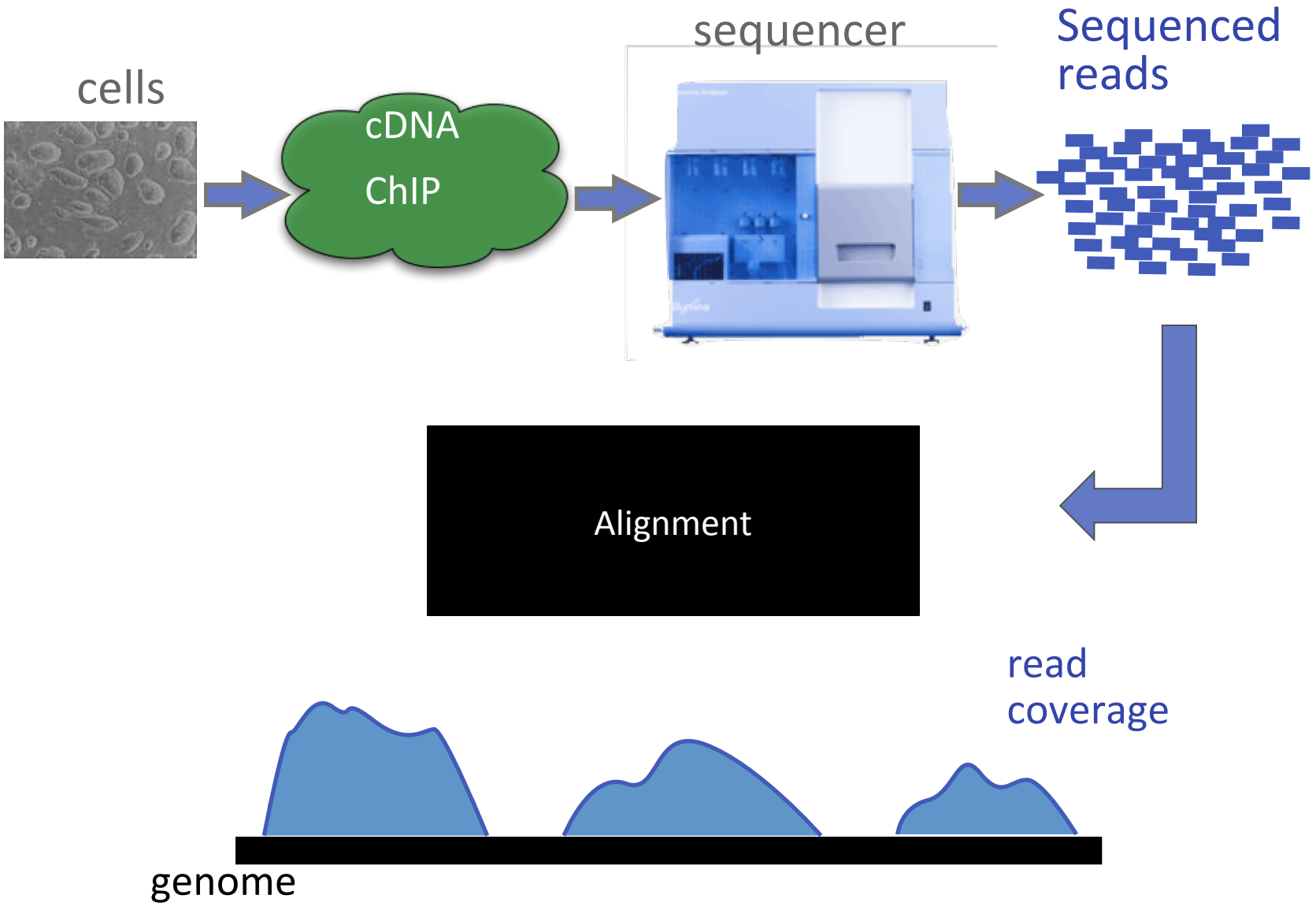
Transcription Factors






Park, P Nature Reviews | Genetics

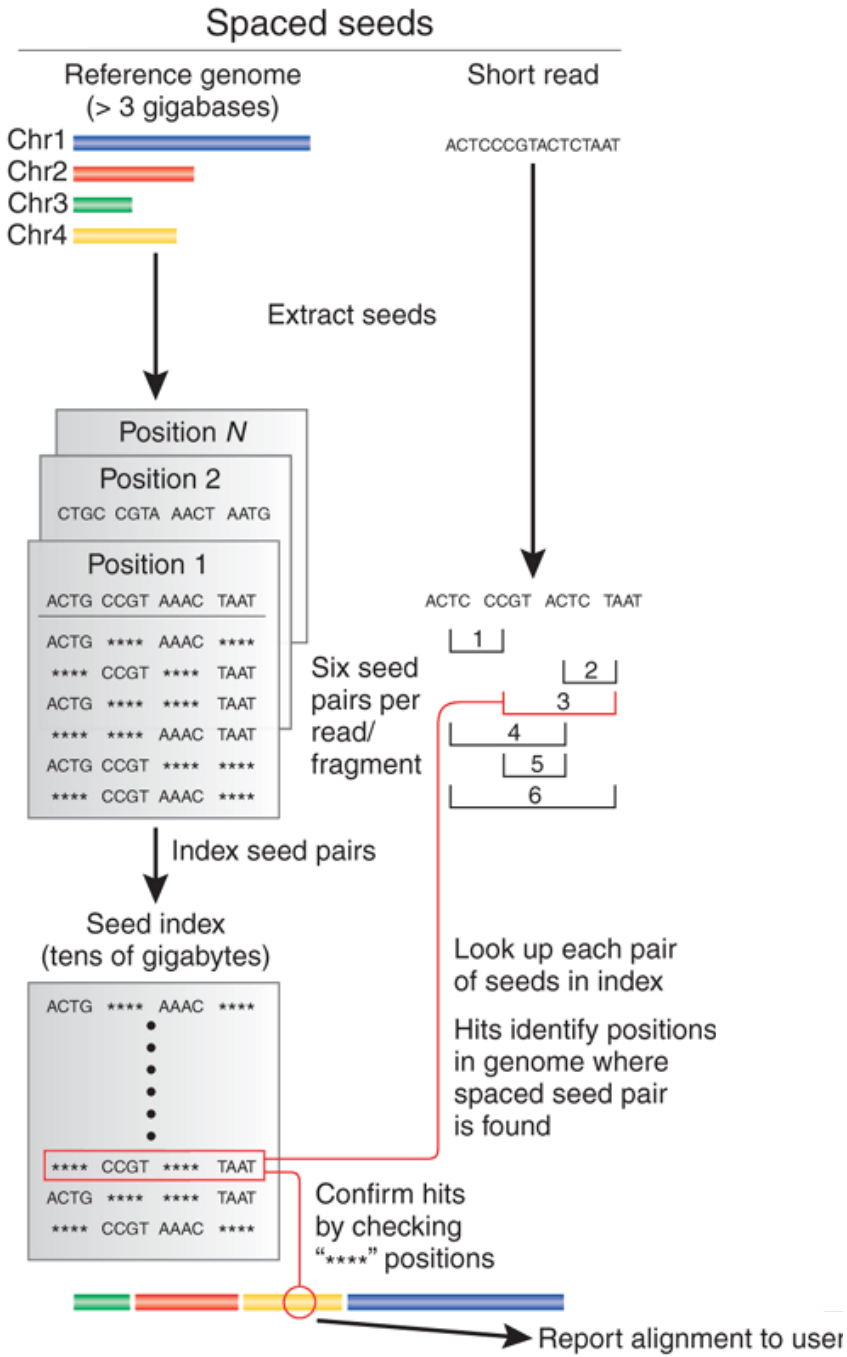
# Once sequenced the problem becomes computational



# Overview of the session

We'll cover the 3 main computational challenges of sequence analysis for *counting applications*:

- Read mapping (alignment): Placing short reads in the genome
- Reconstruction: Finding the regions that originated the reads
- Quantification:
  - Assigning scores to regions
  - Finding regions that are differentially represented between two or more samples.




Trapnell, Salzberg, Nature Biotechnology 2009

# Spaced seed alignment – Hashing the genome

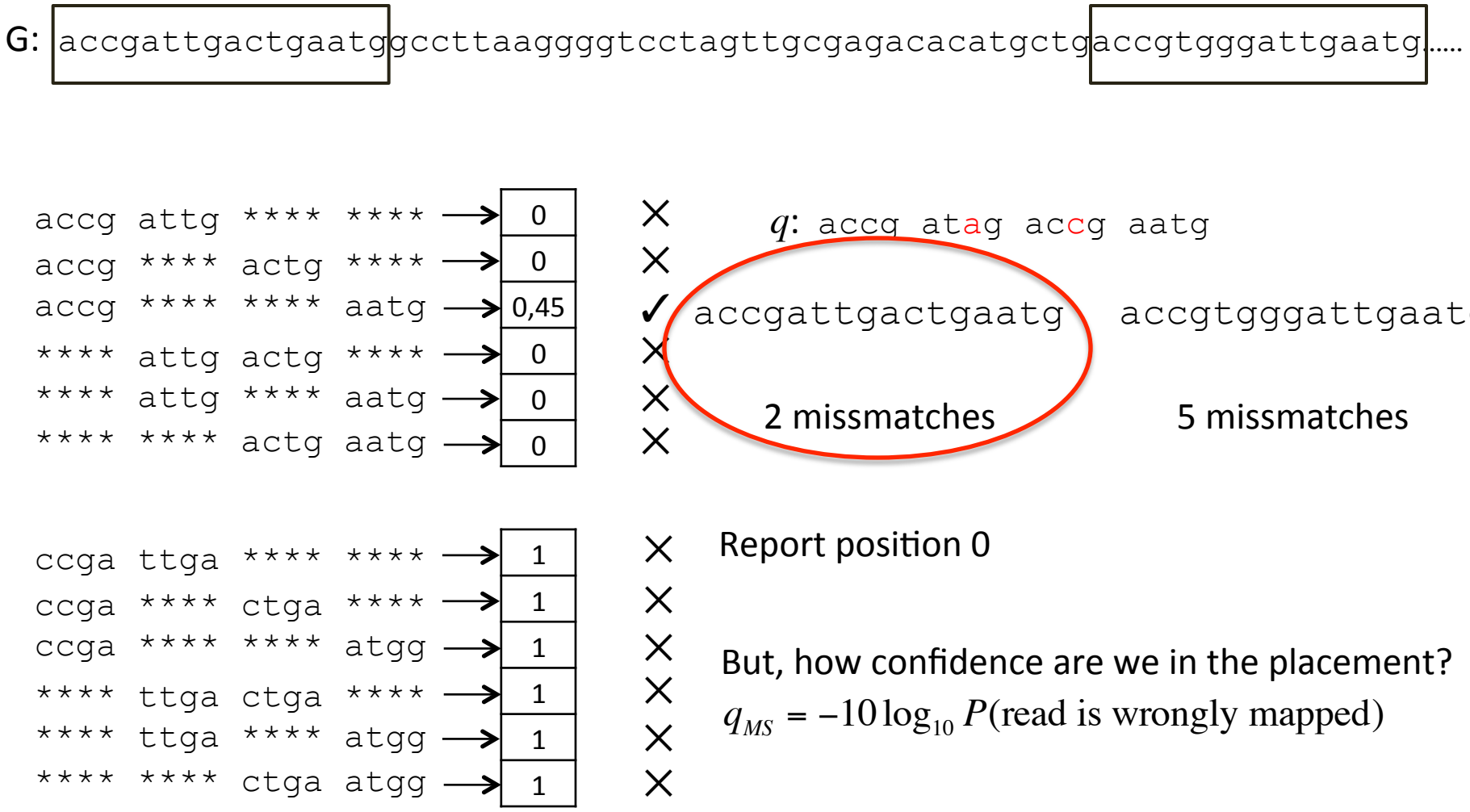
G: accgattgactgaatggccttaaggggtcctagttgcgagacacatgctgaccgtgggattgaatg……

Store spaced seed positions

| | |
|---|---|
| accg attg **** **** → | 0 |
| accg **** actg **** → | 0 |
| accg **** **** aatg → | 0,45 |
| **** attg actg **** → | 0 |
| **** attg **** aatg → | 0 |
| **** **** actg aatg → | 0 |

| | |
|---|---|
| ccga ttga **** **** → | 1 |
| ccga **** ctga **** → | 1 |
| ccga **** **** atgg → | 1 |
| **** ttga ctga **** → | 1 |
| **** ttga **** atgg → | 1 |
| **** **** ctga atgg → | 1 |

# Spaced seed alignment – Mapping reads



# Mapping quality

What does $q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$ mean?

Lets compute the probability the read originated at genome position i

$q$: accg atag accg aatg

$q_s$: 30 40 25 30 30 20 10 20 40 30 20 30 40 40 30 25

$q_s[k] = -10\log_{10} P(\text{sequencing error at base k})$, the PHRED score. Equivalently:

$$P(\text{sequencing error at base k}) = 10^{-\frac{q_s}{10}}$$

So the probability that a read originates from a given genome position i is:

$$P(q \mid G,i) = \prod_{j\ \text{match}} P(q_j \text{good call}) \prod_{j\ \text{missmatch}} P(q_j \text{bad call}) \approx \prod_{j\ \text{missmatch}} P(q_j \text{bad call})$$

In our example

$$P(q \mid G,0) = \left[(1-10^{-3})^6 (1-10^{-4})^4 (1-10^{-2.5})^2 (1-10^{-2})^2\right]\left[10^{-1}10^{-2}\right] = [0.97]*[0.001] \approx 0.001$$

# Mapping quality

What does $q_{MS} = -10\log_{10} P(\text{read is wrongly mapped})$ mean?

$$P(q \mid G,i) = \prod_{j\ \text{match}} P(q_j \text{good call}) \prod_{j\ \text{missmatch}} P(q_j \text{bad call}) \approx \prod_{j\ \text{missmatch}} P(q_j \text{bad call})$$

But what we need is the posterior probability, the probability that the region starting at i was sequenced *given* that we observed the read *q*:

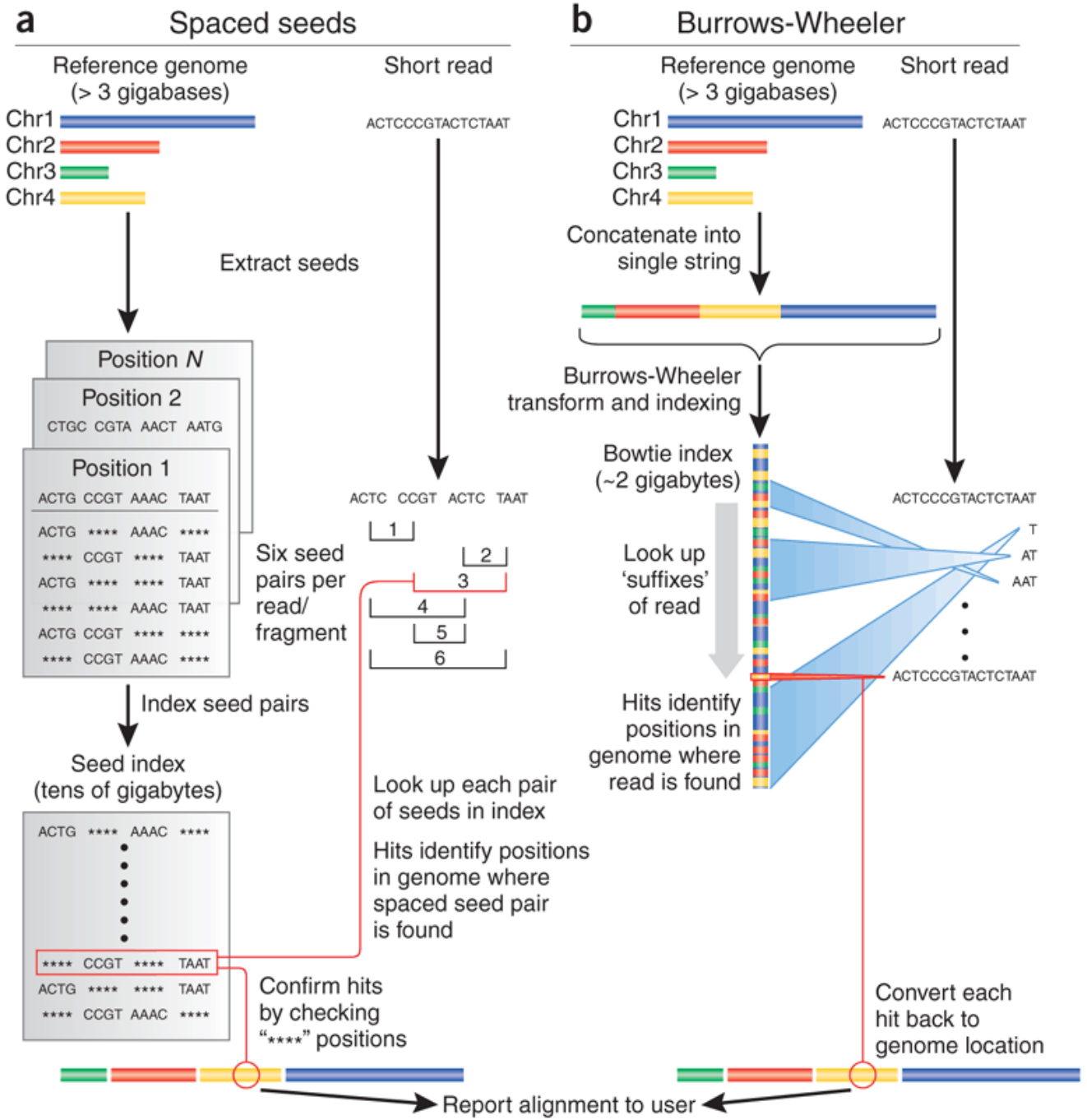
$$P(i \mid G,q) = \frac{P(q \mid G,i)P(i \mid G)}{P(q \mid G)} = \frac{P(q \mid G,i)P(i \mid G)}{\sum_j P(q \mid G,j)}$$

Fortunately, there are efficient ways to approximate this probability (see Li, H *genome Research* 2008, for example)

$$q_{MS} = -10\log_{10}(1 - P(i \mid G,q))$$

## Considerations

- Trade-off between sensitivity, speed and memory
  - Smaller seeds allow for greater mismatches at the cost of more tries
  - Smaller seeds result in a smaller tables (table size is at most $4^k$), larger seeds increase speed (less tries, but more seeds)




Trapnell, Salzberg, Nature Biotechnology 2009

# Considerations

- BWT-based algorithms rely on perfect matches for speed
- When dealing with mismatches, algorithms “backtrack” when the alignment extension fails.
- Backtracking is expensive
- As read length increases novel algorithms are required
- Smaller seeds result in a smaller tables (table size is at most $4^k$), so larger seeds increase speed (less *fishing* but more seeds

# Short read mapping software for ChIP-Seq

**Seed-extend**

| | Short indels | Use base qual |
|---|---|---|
| Maq | **No** | **YES** |
| BFAST | Yes | NO |
| GASSST | Yes | NO |
| RMAP | Yes | **YES** |
| SeqMap | Yes | NO |
| SHRiMP | Yes | NO |

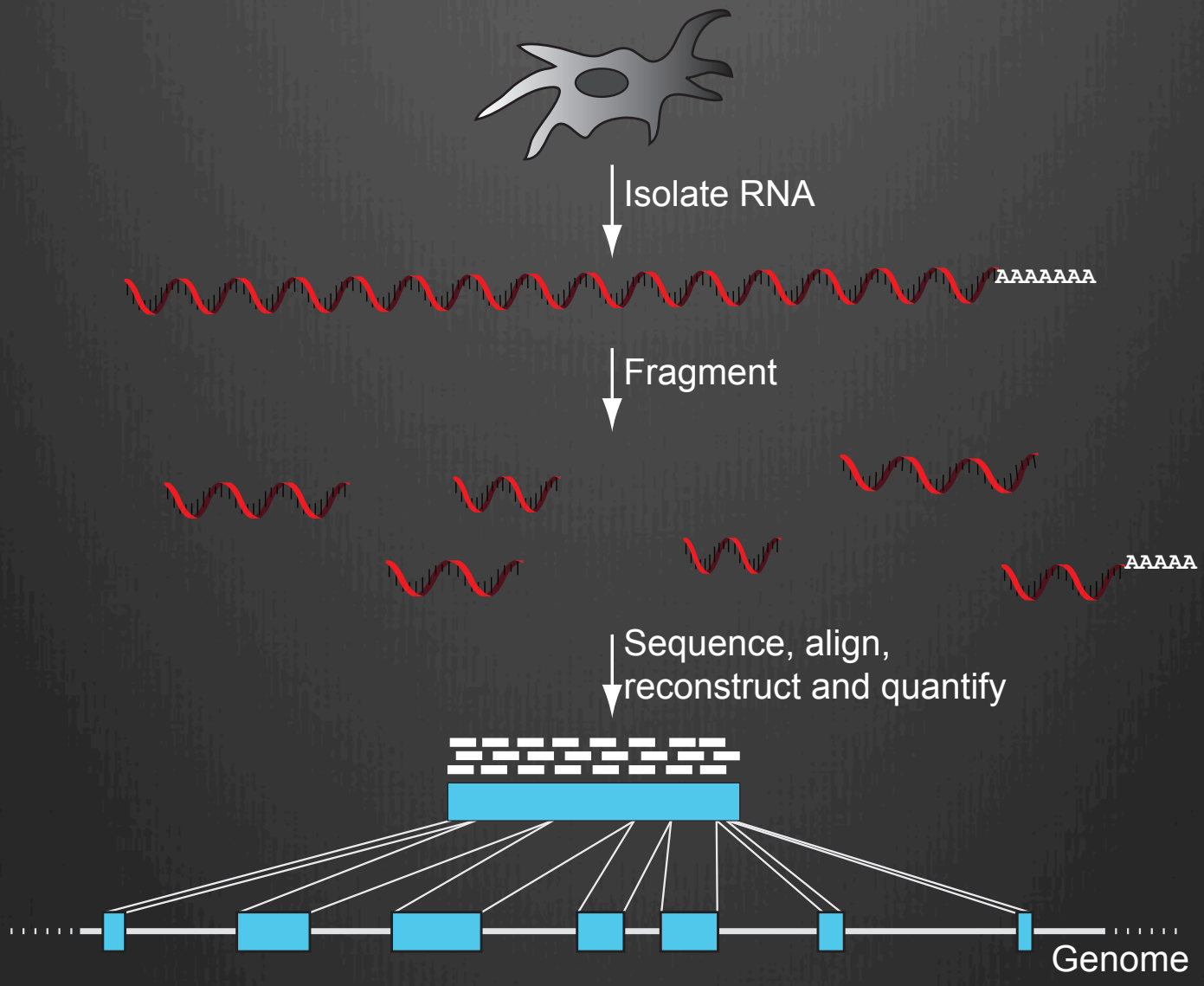
**BWT**

| | Use Base qual |
|---|---|
| BWA | **YES** |
| Bowtie | NO |
| Soap2 | NO |
| Stampy* | YES |
| Bowtie2* | (NO, I think) |
| | |

***Stampy is a hybrid approach which first uses BWA to map reads then uses seed-extend only to reads not mapped by BWA**

***Bowtie2 breaks reads into smaller pieces and maps these "seeds" using a BWT genome.**

# RNA-Seq


Isolate RNA
AAAAAAA
Fragment
AAAAA
Sequence, align,
reconstruct and quantify
Genome

# What's the fuss

| | Expression arrays | Exon Arrays | Tiling Arrays | **RNASeq** |
|---|---|---|---|---|
|  | ✓ | ✗ | ✗ | ✓ |
|  | ✗ | ✓ | ✗ | ✓ |
|  | ✗ | ✗ | ✓ | ✓ |

- ~~Until recently transcriptomics required:~~
  - ~~A "finished" grade genome~~
  - ~~A clone based cDNA and EST annotation~~

*RNASeq requires none*

# RNASeq as a revolutionary tool

Splicing

nature Vol 456|27 November 2008|doi:10.1038/nature07509

ARTICLES

**Alternative isoform regulation in human tissue transcriptomes**

Eric T. Wang[1,2]*, Rickard Sandberg[1,3]*, Shujun Luo[4], Irina Khrebtukova[4], Lu Zhang[4], Christine Mayr[5], Stephen F. Kingsmore[6], Gary P. Schroth[4] & Christopher B. Burge[1]

>90% genes Alternatively spliced and splicing is tissue specific

Annotation

*Ab initio* reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs

Mitchell Guttman[1,2,6], Manuel Garber[1,6], Joshua Z Levin[1], Julie Donaghey[1], James Robinson[1], Xian Adiconis[1], Lin Fan[1], Magdalena J Koziol[1,3], Andreas Gnirke[1], Chad Nusbaum[1], John L Rinn[1,3], Eric S Lander[1,2,4] & Aviv Regev[1,2,5]

1,500 mouse annotated lincRNAs

**Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses**

Moran N. Cabili,[1,2,3] Cole Trapnell,[1,3] Loyal Goff,[1,4] Magdalena Koziol,[1,3] Barbara Tazon-Vega,[1,3] Aviv Regev,[1,5,6] and John L. Rinn[1,3,6,7]

8,000 human annotated lincRNAs

Expression

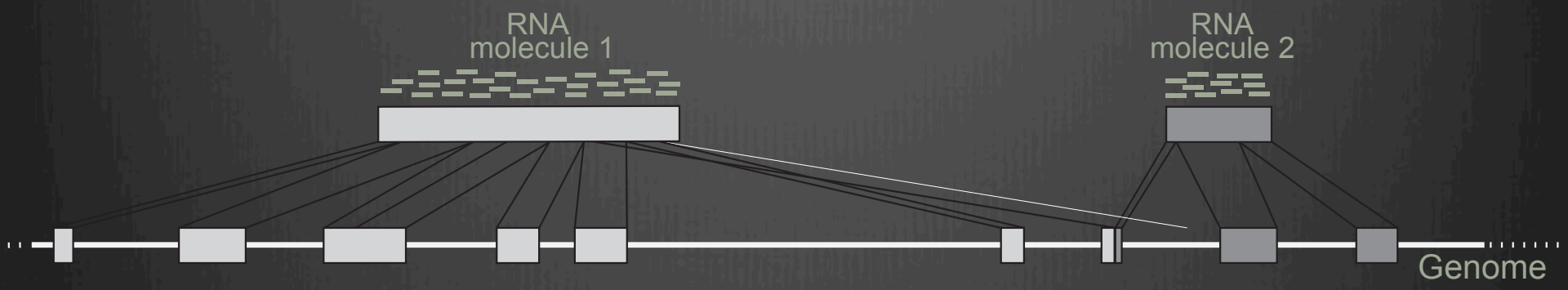
Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

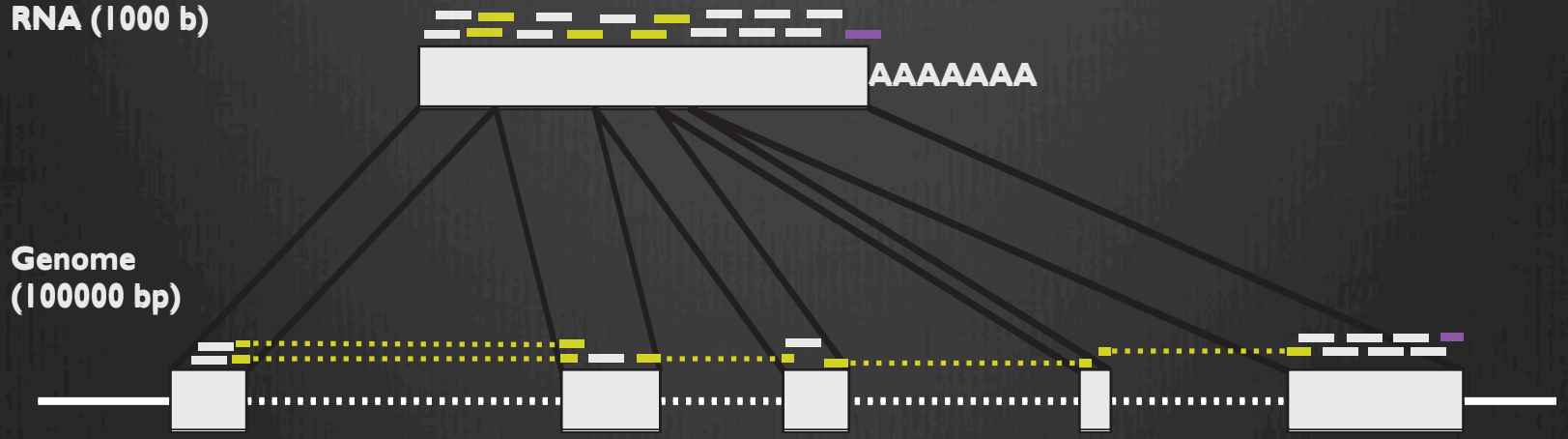
Cole Trapnell[1–3], Brian A Williams[4], Geo Pertea[2], Ali Mortazavi[4], Gordon Kwan[4], Marijke J van Baren[5], Steven L Salzberg[1,2], Barbara J Wold[4] & Lior Pachter[3,6,7]

Ability to measure isoform specific expression

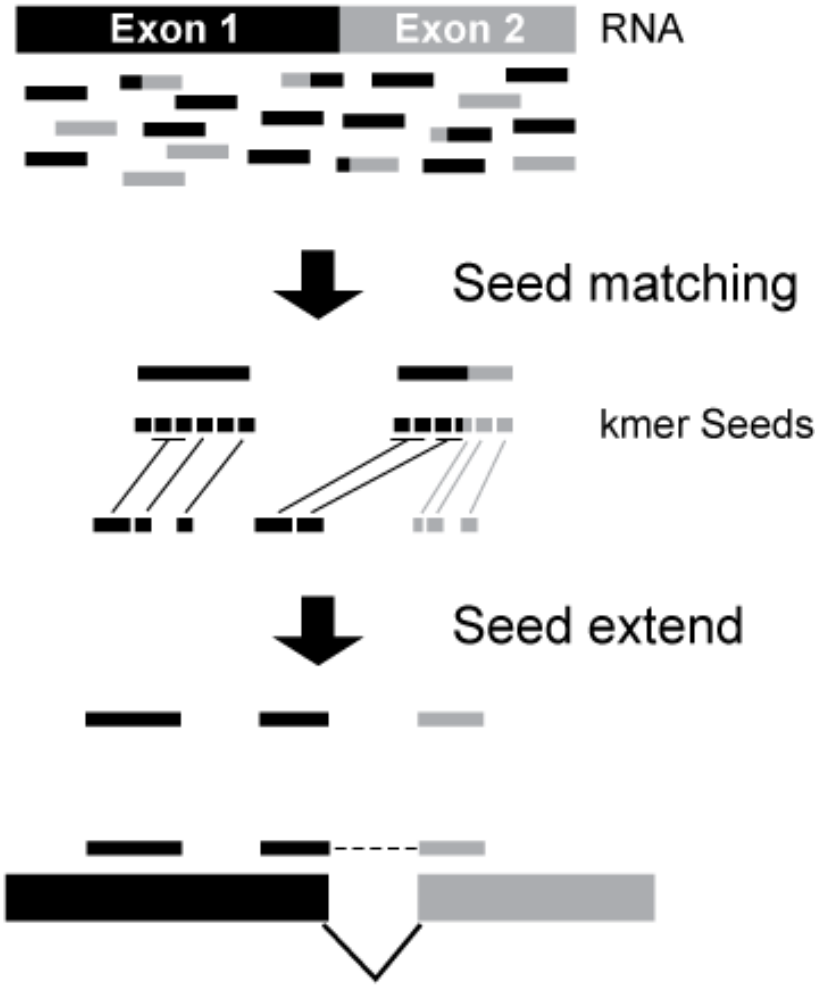
**RNASeq is a one stop offer for transcriptomics**

# RNA-Seq Read mapping

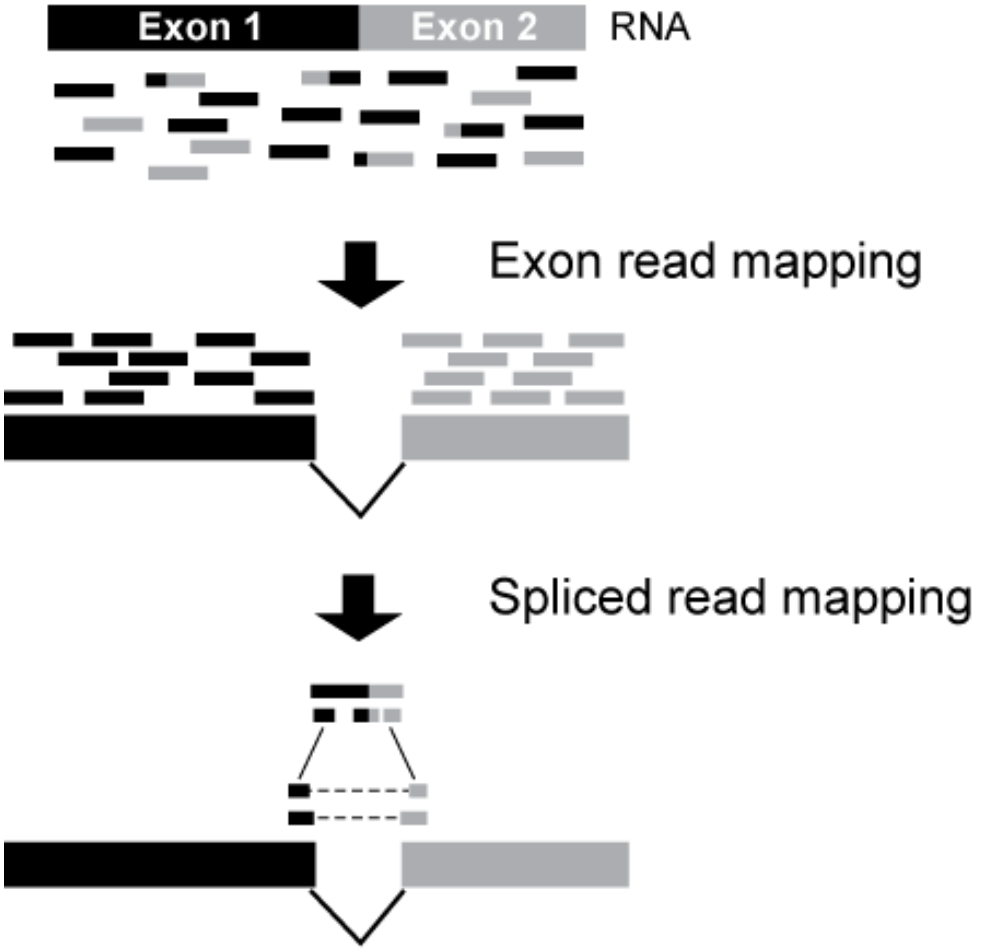


# Mapping RNA-Seq reads: Seed-extend spliced alignment



# Mapping RNA-Seq reads: Exon-first spliced alignment (e.g. TopHat)



# Short read mapping software for RNA-Seq

**Seed-extend**

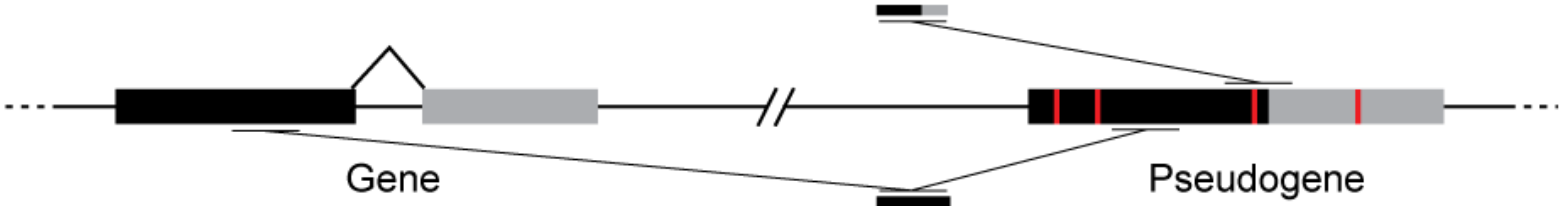
| | Short indels | Use base qual |
|---|---|---|
| GSNAP | **No** | NO |
| QPALMA | Yes | NO |
| BLAT | Yes | NO |

**Exon-first**

| | Use base qual |
|---|---|
| MapSplice | NO |
| SpliceMap | NO |
| **TopHat** | NO |

**Exon-first alignments will map contiguous first at the expense of spliced hits**

# Exon-first aligners are faster but at cost




**How do we visualize the results of these programs**

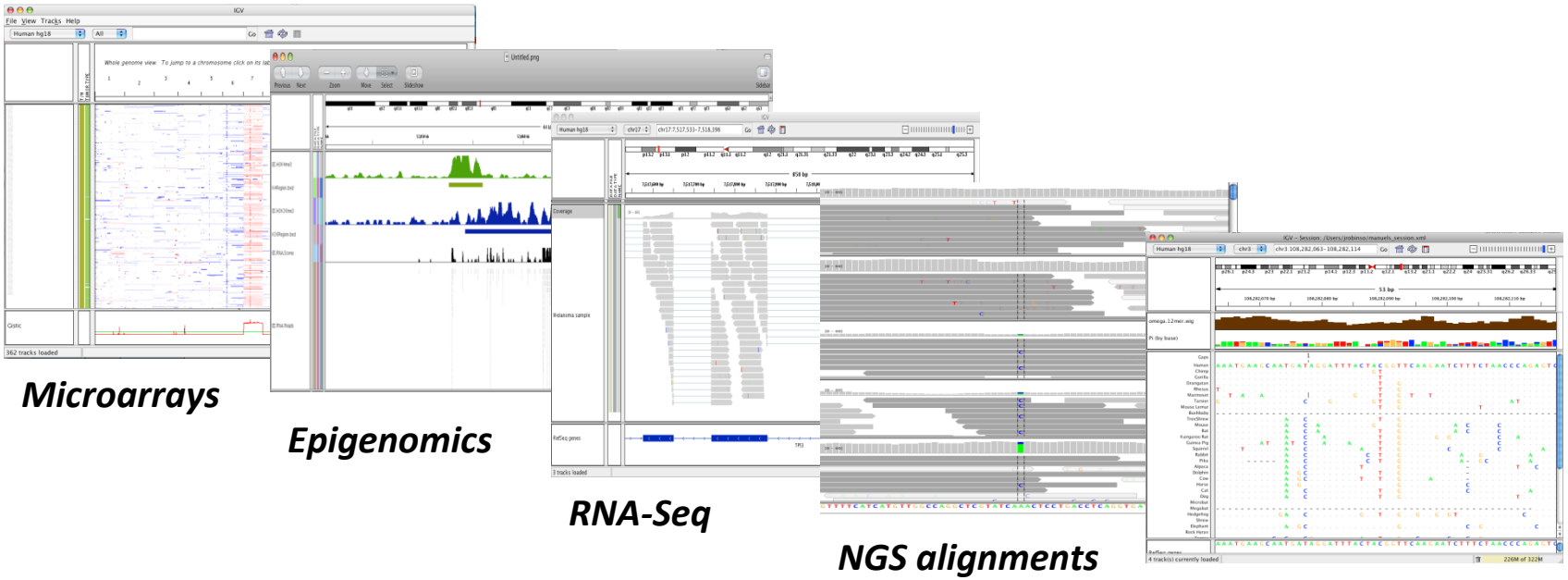
# IGV: Integrative Genomics Viewer

Integrative Genomics Viewer

A desktop application

for the visualization and interactive exploration

of genomic data



*Microarrays*

*Epigenomics*

*RNA-Seq*

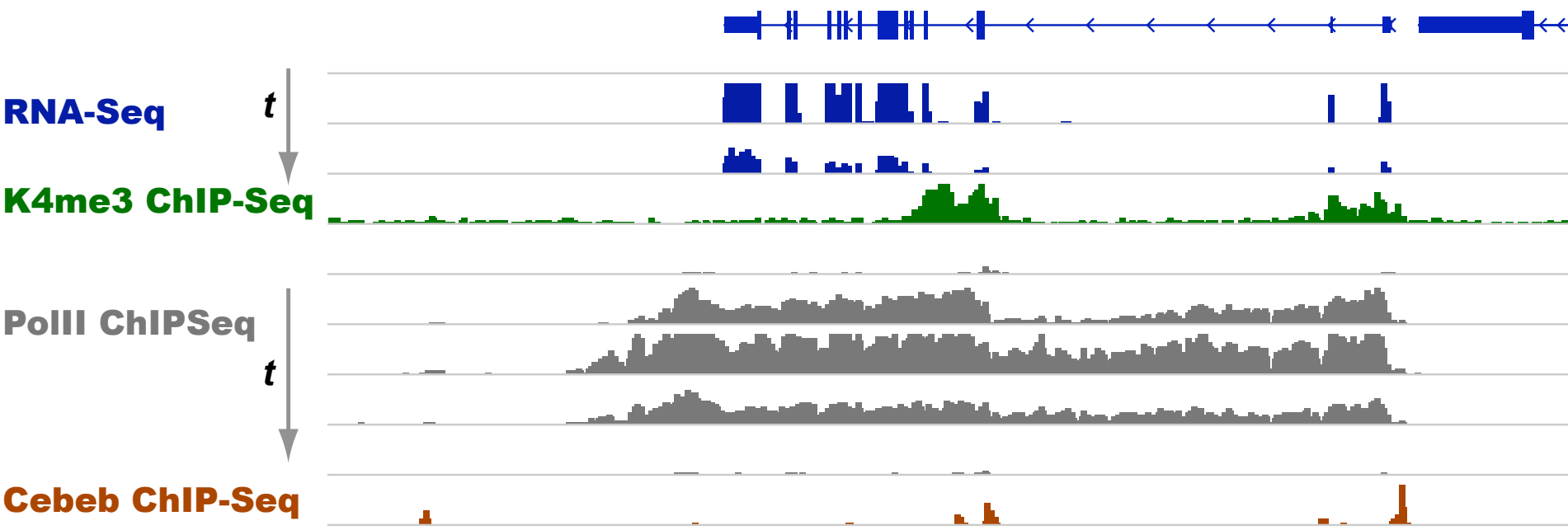
*NGS alignments*

*Comparative genomics*

BROAD INSTITUTE

# Visualizing read alignments with IGV — RNASeq

# Visualizing read alignments with IGV — zooming out



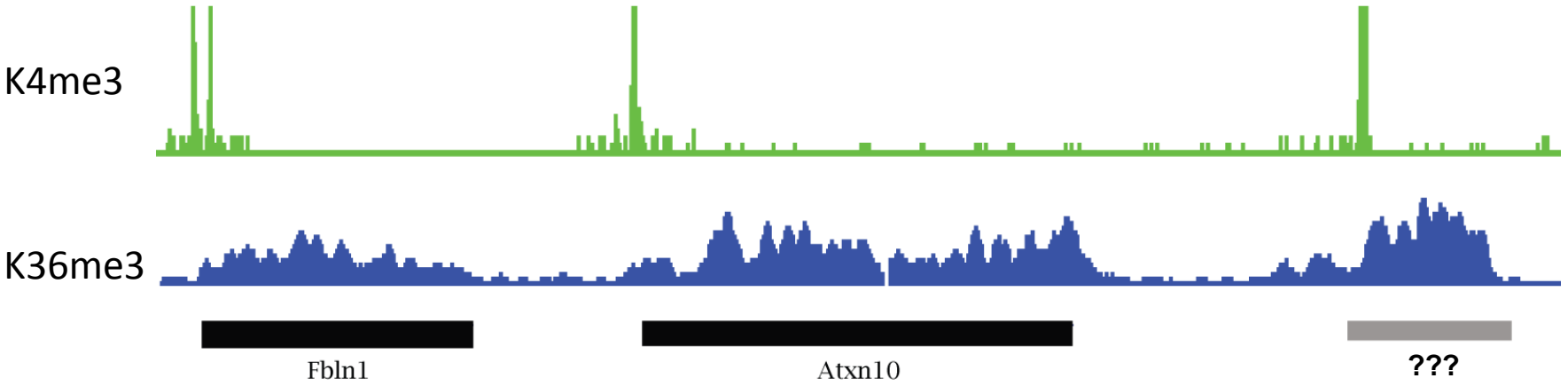

**How can we identify regions enriched in sequencing reads?**

# Overview of the session

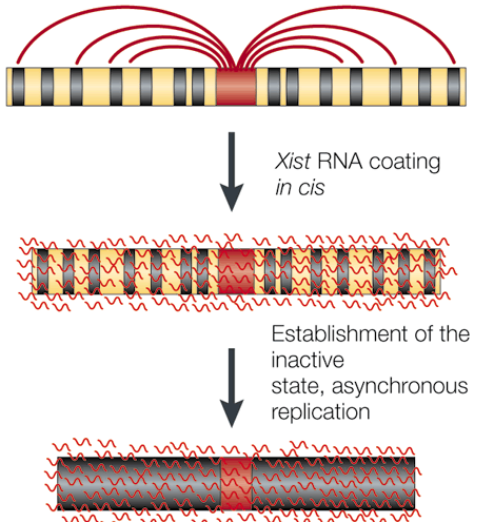
The 3 main computational challenges of sequence analysis for *counting applications*:

- Read mapping: Placing short reads in the genome
- Reconstruction: Finding the regions that originate the reads
- Quantification:
  - Assigning scores to regions
  - Finding regions that are differentially represented between two or more samples.

# Chromatin domains demarcate interesting surprises in the transcriptome




**XIST**

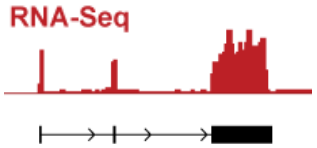

**These regions likely contain similar non-coding RNA genes**

Mitch Guttman

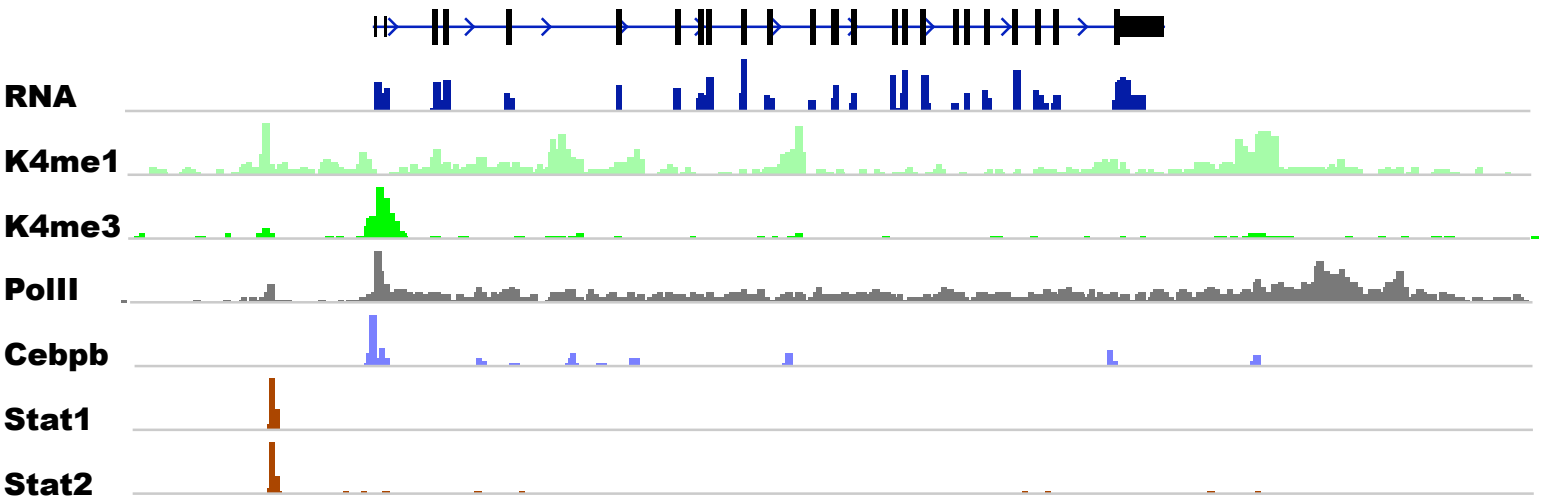
# How can we identify these chromatin marks and the genes within?


Short modification


Long modification

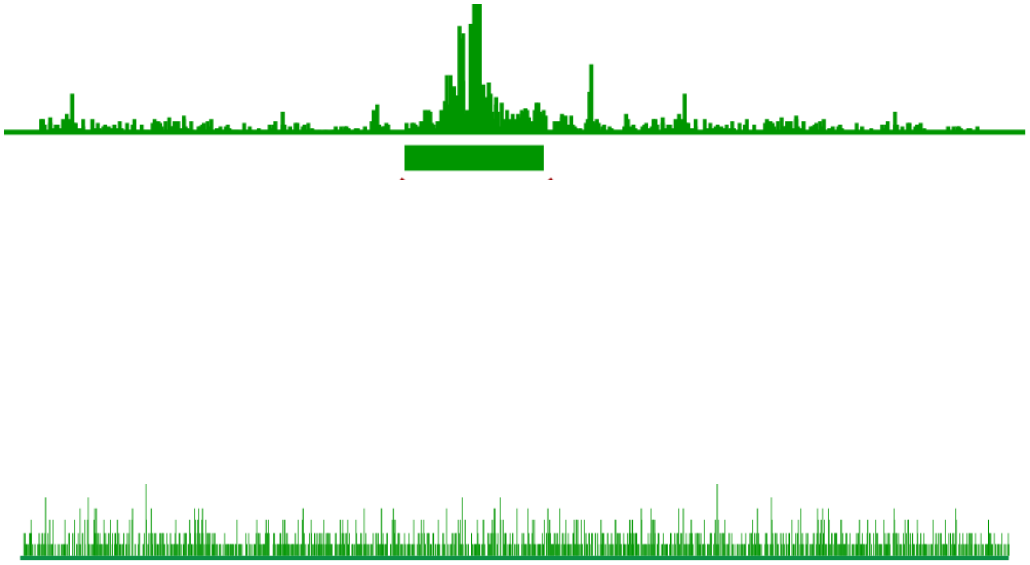


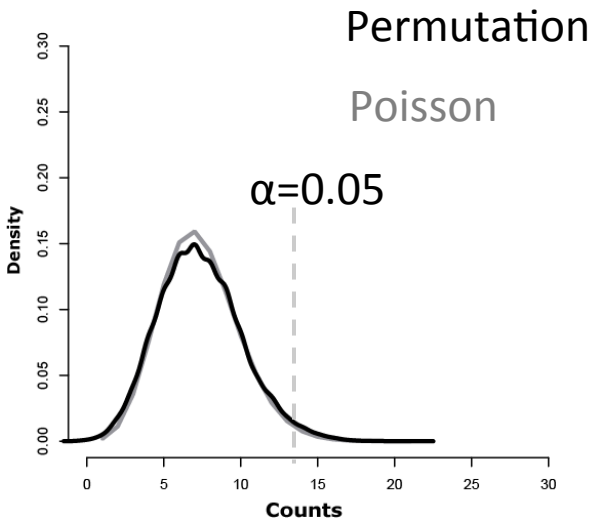

Discontinuous data




**Scripture is a method to solve this general question**

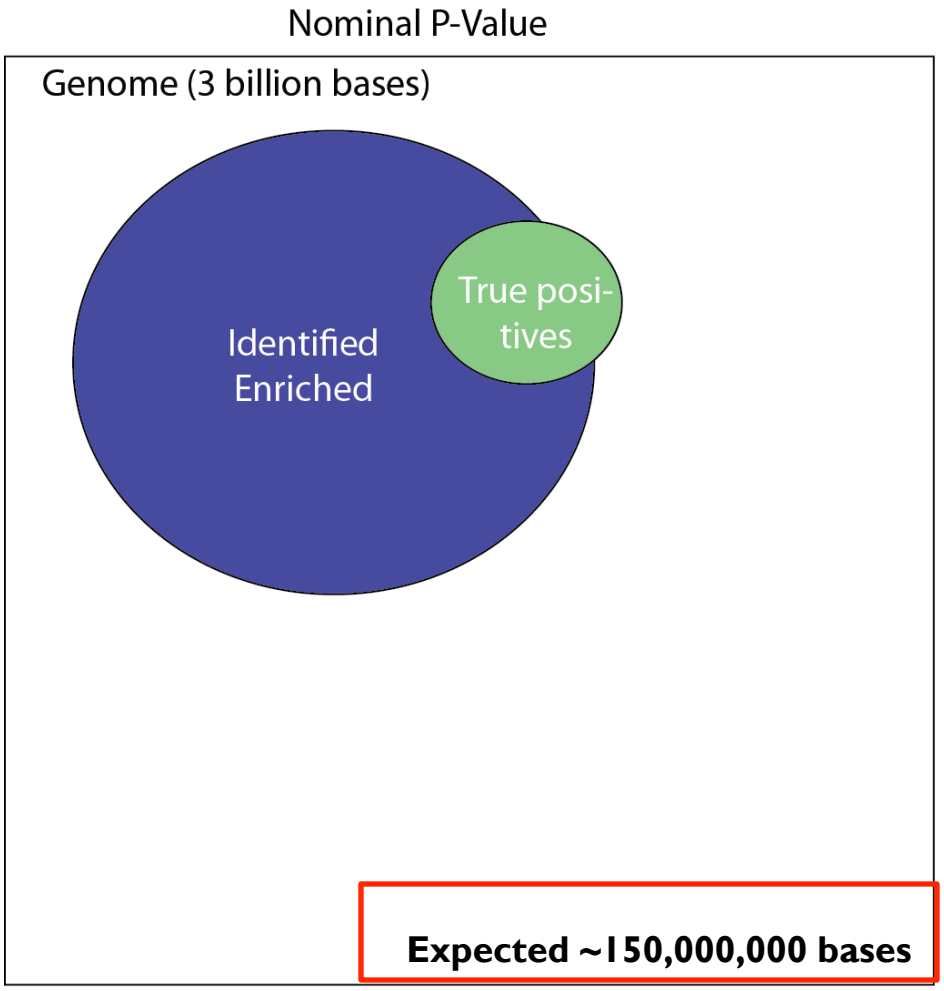
# Our approach





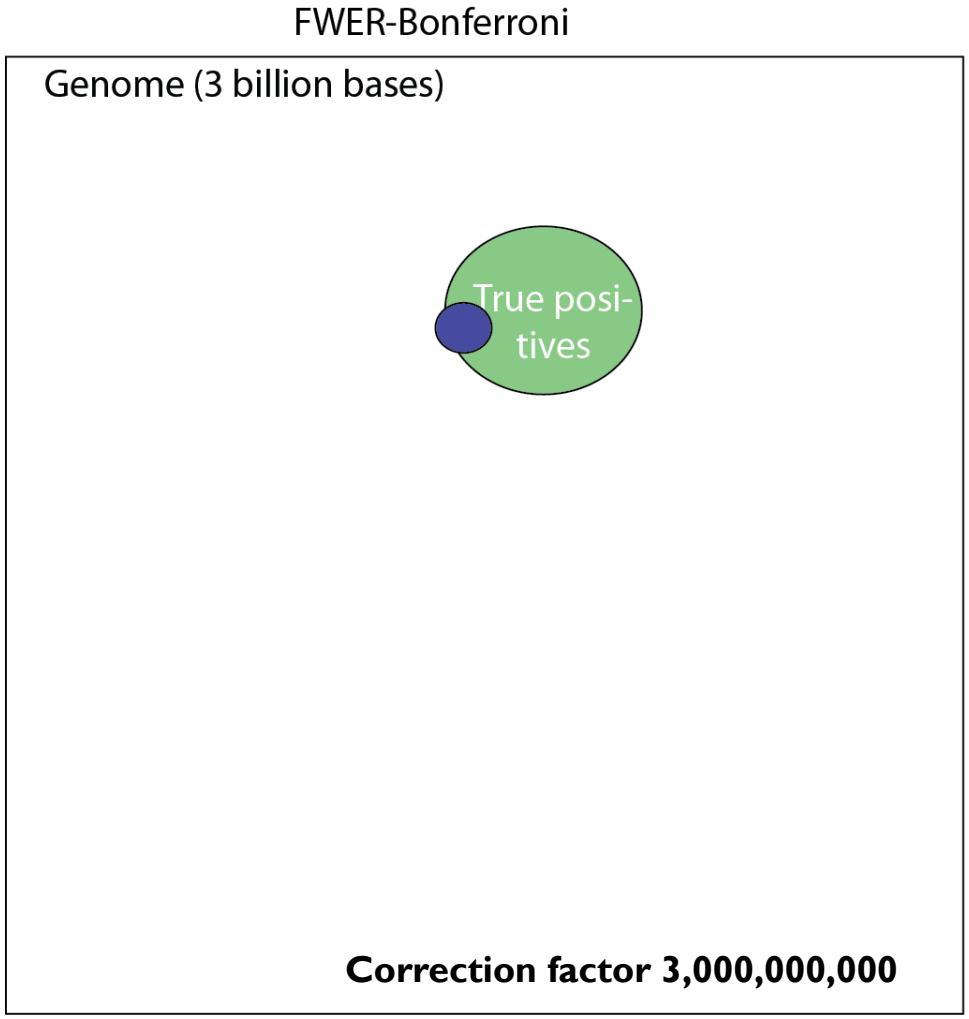

**We have an efficient way to compute read count p-values …**

# The genome is big, many things happen by chance




**We need to correct for multiple hypothesis testing**

# Bonferroni correction is way to conservative

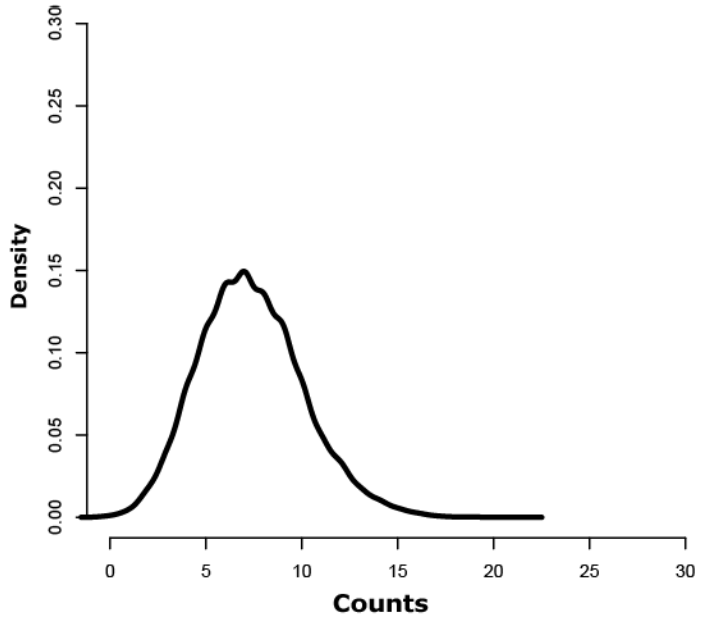



**Bonferroni corrects the number of hits but misses many true hits because its too conservative – How do we get more power?**

# Controlling FWER

Max Count distribution

$\alpha$=0.05 $\alpha_{FWER}$=0.05




Count distribution (Poisson)

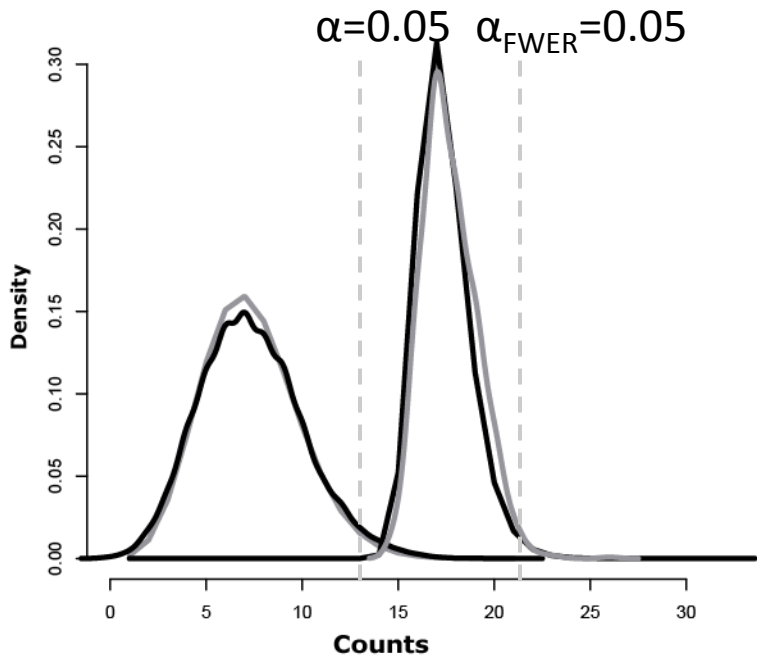
Given a region of size w and an observed read count n. What is the probability that one or more of the $3x10^9$ regions of size w has read count >= n under the null distribution?

We could go back to our permutations and compute an FWER: **max of the genome-wide distributions of same sized region**)→
but really really really slow!!!

# Scan distribution, an old problem

- Is the observed number of read counts over our region of interest high?
- Given a set of Geiger counts across a region find clusters of high radioactivity
- Are there time intervals where assembly line errors are high?

Scan distribution




Poisson distribution

Thankfully, there is a distribution called the Scan Distribution which computes a closed form for this distribution.

ACCOUNTS for dependency of overlapping windows thus more powerful!

# Scan distribution for a Poisson process

The probability of observing k reads on a window of size w in a genome of size L given a total of N reads can be approximated by (Alm 1983):

$$P(k|\lambda w, N, L) \approx 1 - F_p(k-1|\lambda w)e^{-\frac{k-w\lambda}{k}\lambda(T-w)P(k-1|\lambda w)}$$

where

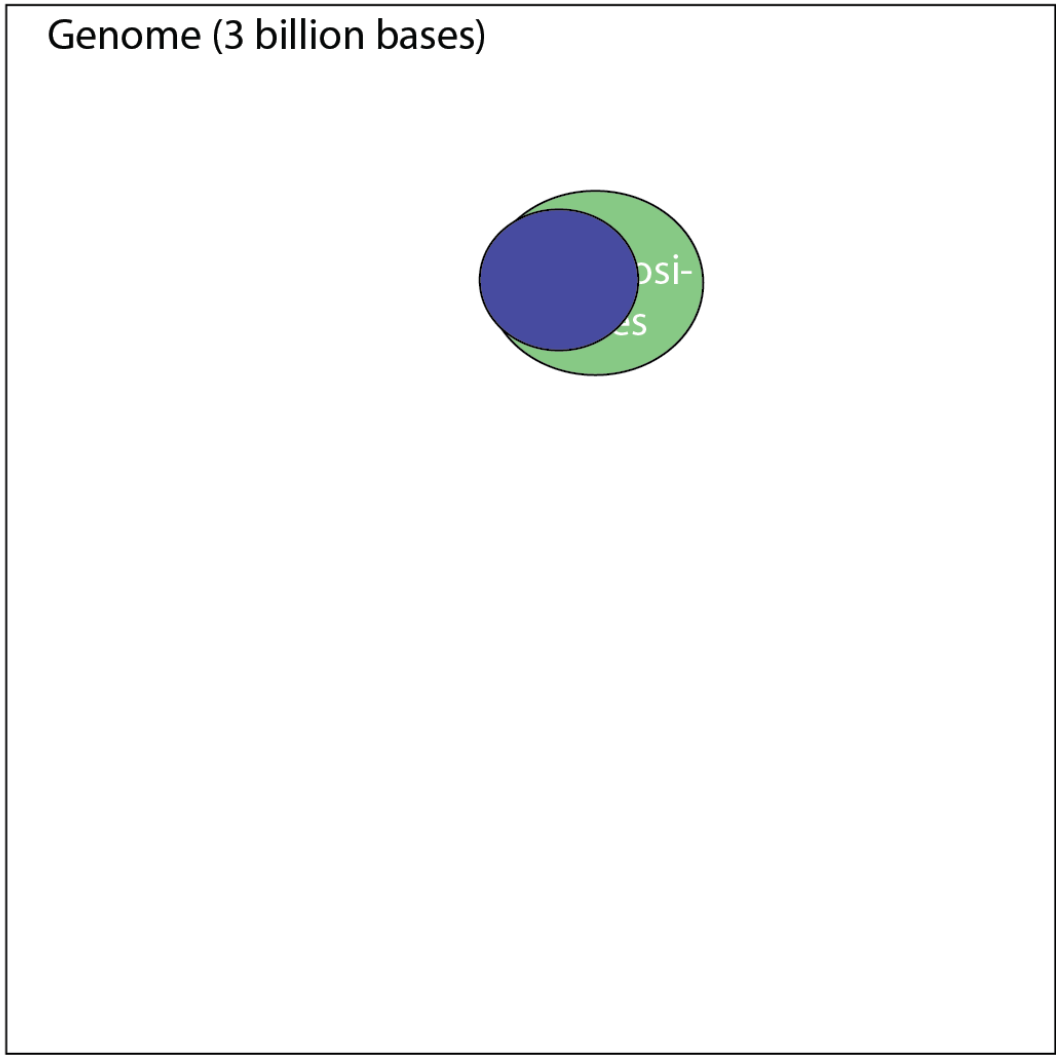
$P(k-1|\lambda w)$ is the Poisson probability of observing $k-1$ counts given an expected count of $\lambda w$

and

$F_p(k-1|\lambda w)$ is the Poisson probability of observing $k-1$ or fewer counts given an expectation of $\lambda w$ reads

**The scan distribution gives a computationally very efficient way to estimate the FWER**

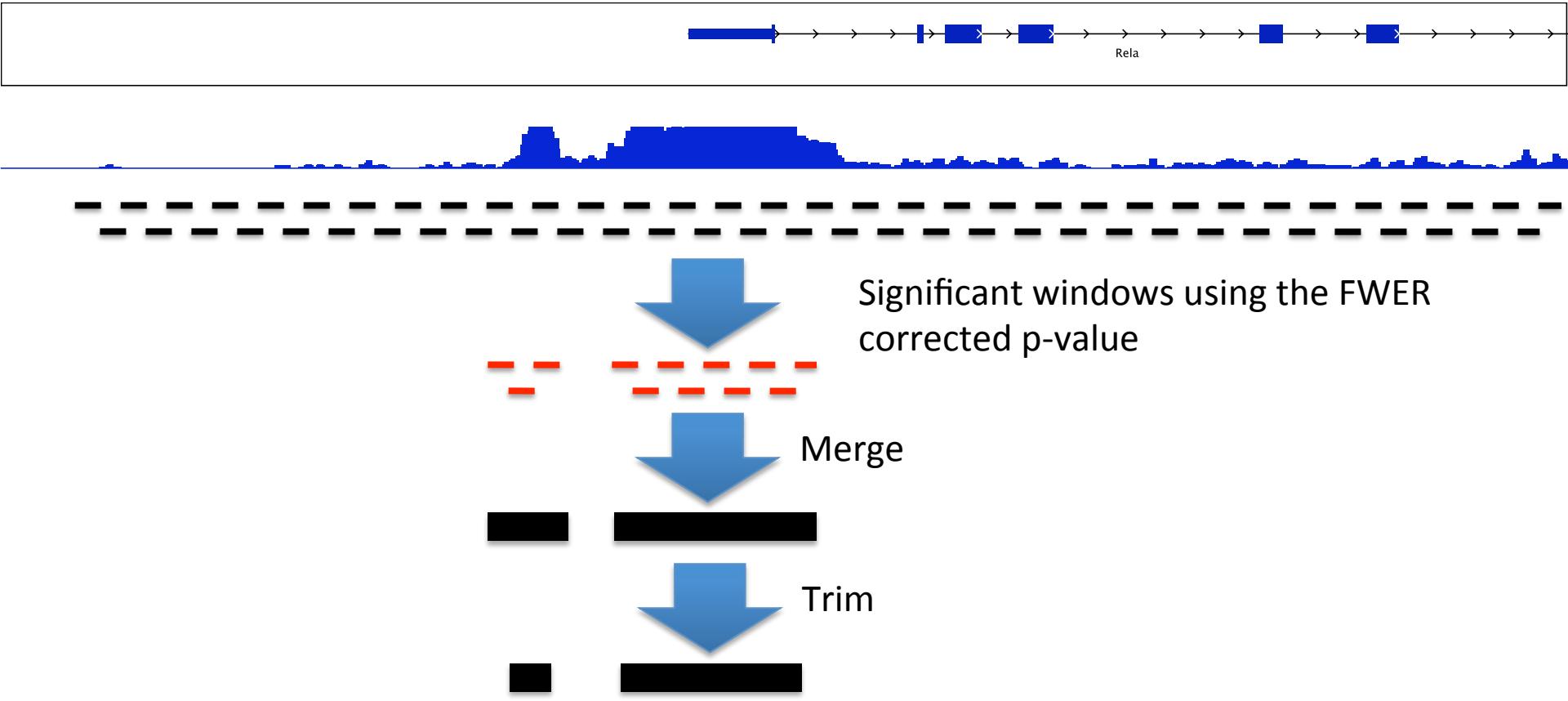
FWER-Scan Statistics




**By utilizing the dependency of overlapping windows we have greater power, while still controlling the same genome-wide false positive rate.**

# Segmentation method for contiguous regions

Example : PolII ChIP

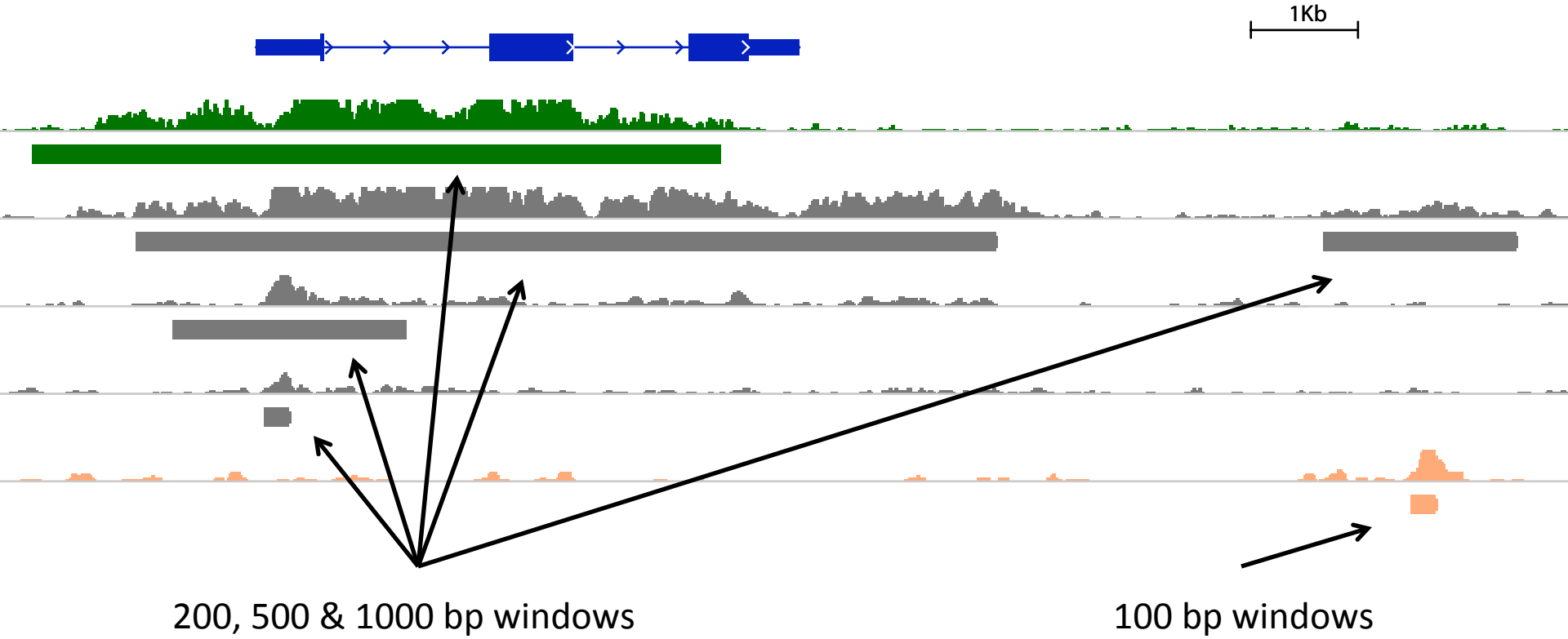



**But, which window?**

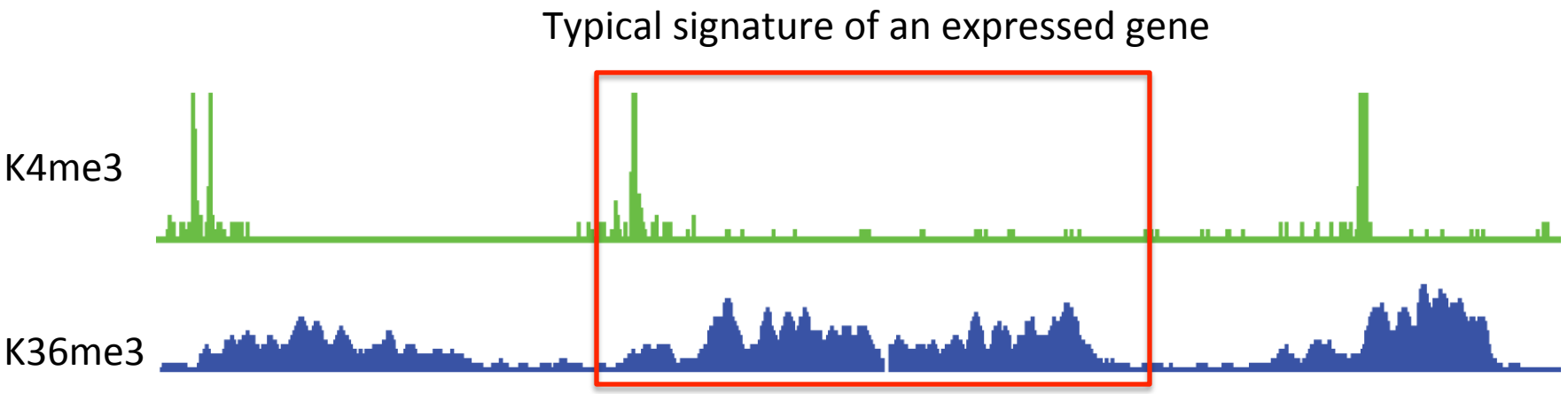
# We use multiple windows

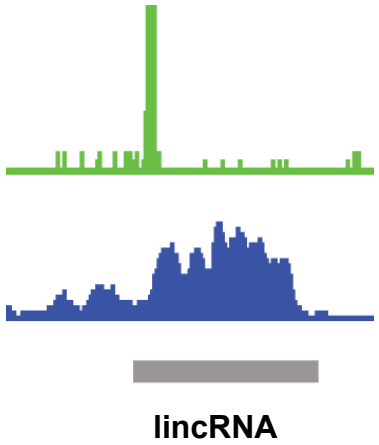
- Small windows detect small punctuate regions.
- Longer windows can detect regions of moderate enrichment over long spans.
- In practice we scan different windows, finding significant ones in each scan.
- In practice, it helps to use some prior information in picking the windows although globally it might be ok.

# Applying Scripture to a variety of ChIP-Seq data



# Application of scripture to mouse chromatin state maps




Identifed
- ~1500 lincRNAs
  - Conserved
  - Noncoding
  - Robustly expressed

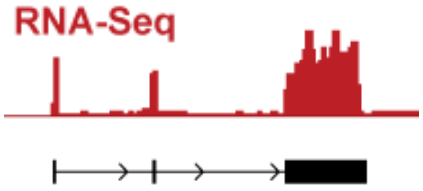
Mitch Guttman

# Can we identify enriched regions across different data types?


Short modification ✓


Long modification ✓

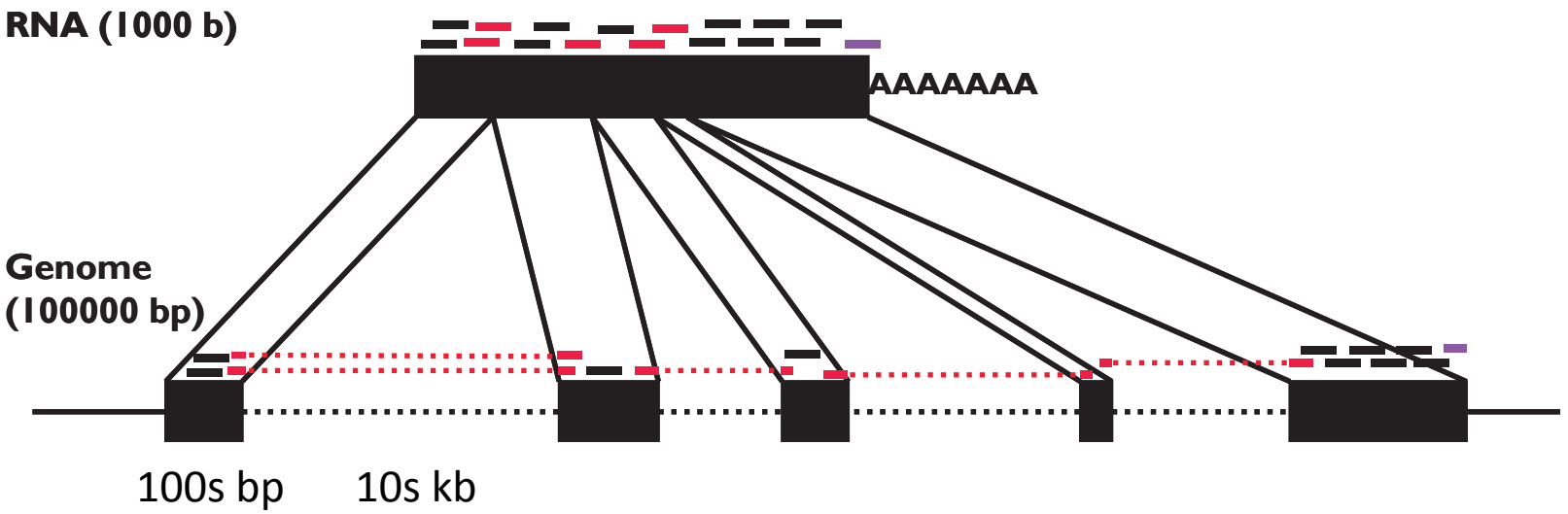
Using chromatin signatures we discovered hundreds of putative genes.
**What is their structure**?




Discontinuous data: RNA-Seq to find gene structures for this gene-like regions

# Scripture for RNA-Seq:
# Extending segmentation to discontiguous regions

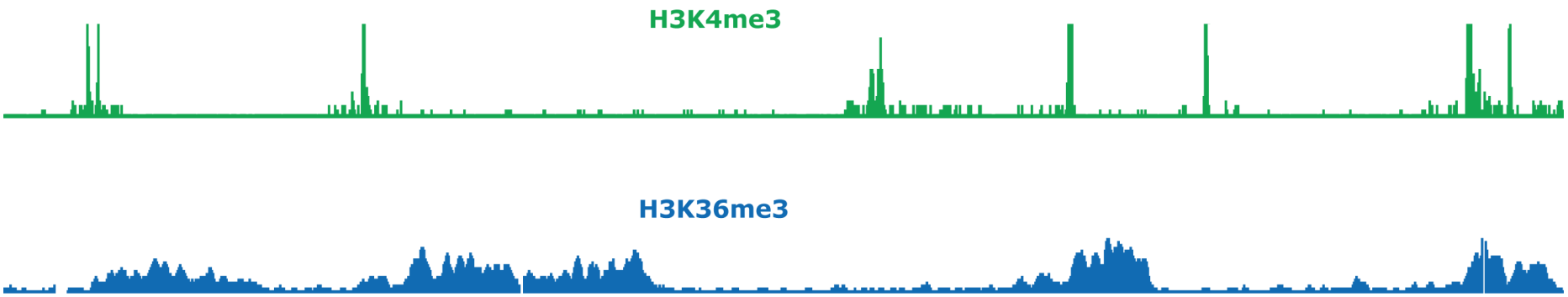
# The transcript reconstruction problem as a segmentation problem



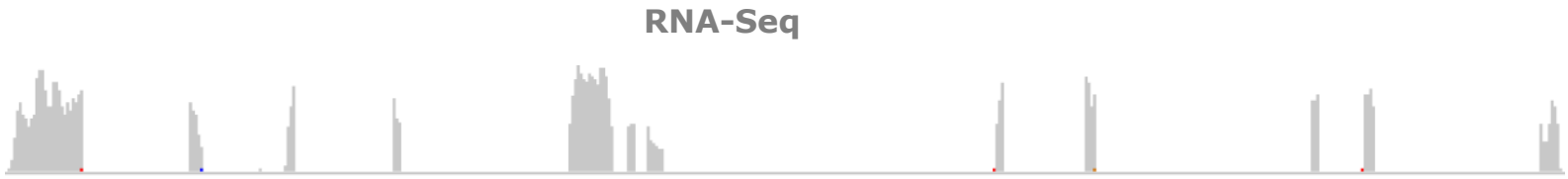

**Challenges:**

- Genes exist at many different expression levels, spanning several orders of magnitude.
- Reads originate from both mature mRNA (exons) and immature mRNA (introns) and it can be problematic to distinguish between them.
- Reads are short and genes can have many isoforms making it challenging to determine which isoform produced each read.

# Scripture: A statistical genome-guided transcriptome reconstruction




**Statistical segmentation of chromatin modifications uses continuity of segments to increase power for interval detection**

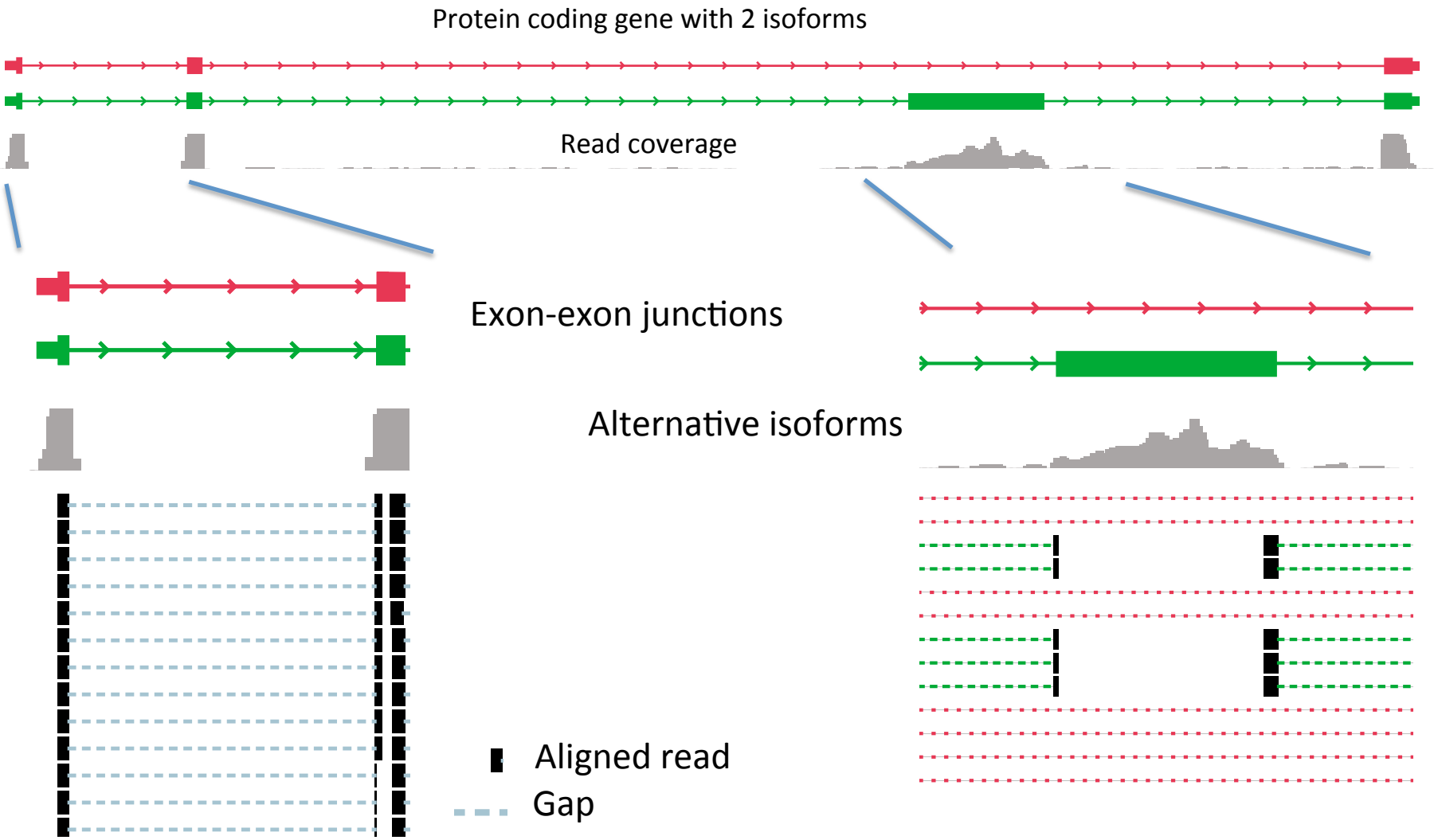



**If we know the connectivity of fragments, we can increase our power to detect transcripts**

# Longer (76) reads provide increased number of junction reads

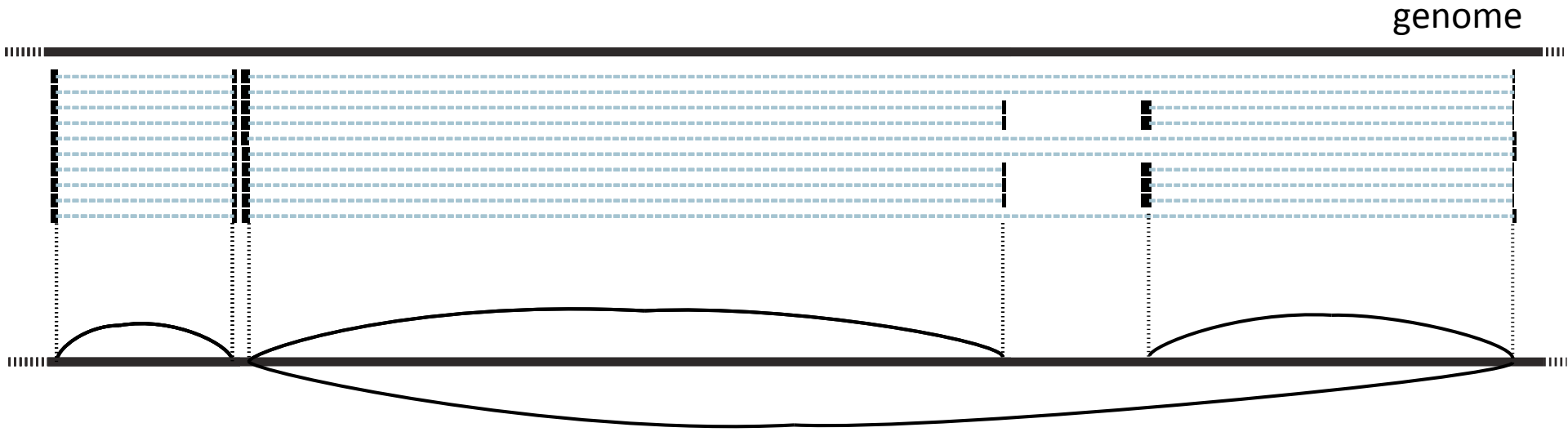

**Exon junction spanning reads provide the connectivity information.**

# The power of spliced alignments



# Statistical reconstruction of the transcriptome

Step 1: Align Reads to the genome allowing gaps flanked by splice sites

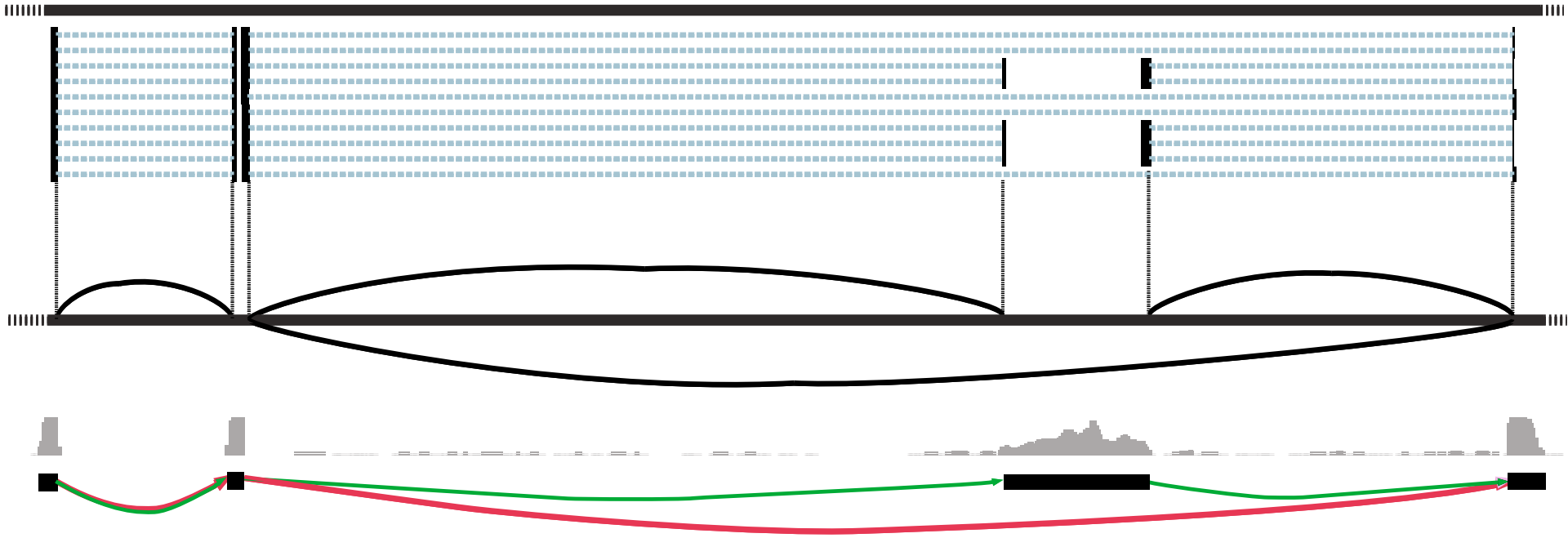



Step 2: Build an oriented connectivity graph using every spliced alignment and orienting edges using the flanking splicing motifs

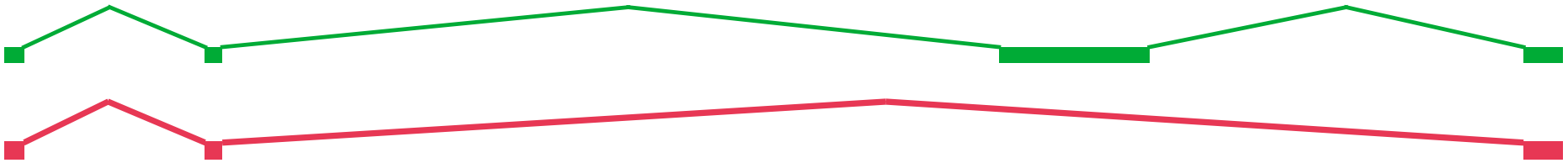
**The “connectivity graph” connects all bases that are directly connected within the transcriptome**

# Statistical reconstruction of the transcriptome

Step 3: Identify “segments” across the graph



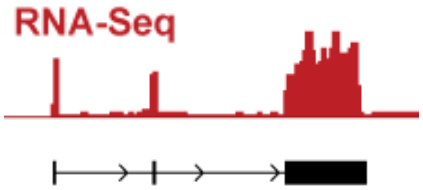
Step 4: Find significant segments



# Can we identify enriched regions across different data types?


Short modification ✓

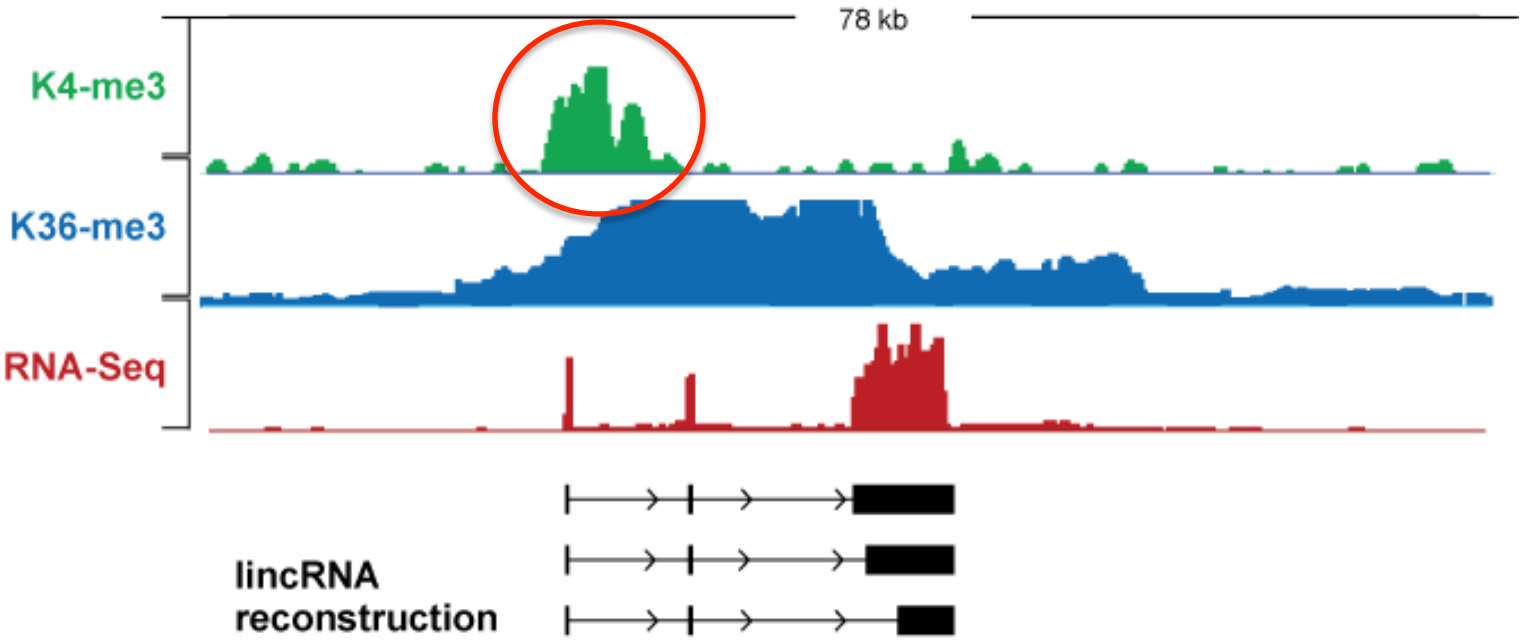

Long modification ✓




Discontinuous data ✓

**Are we really sure reconstructions are complete?**

# RNA-Seq data is incomplete for comprehensive annotation

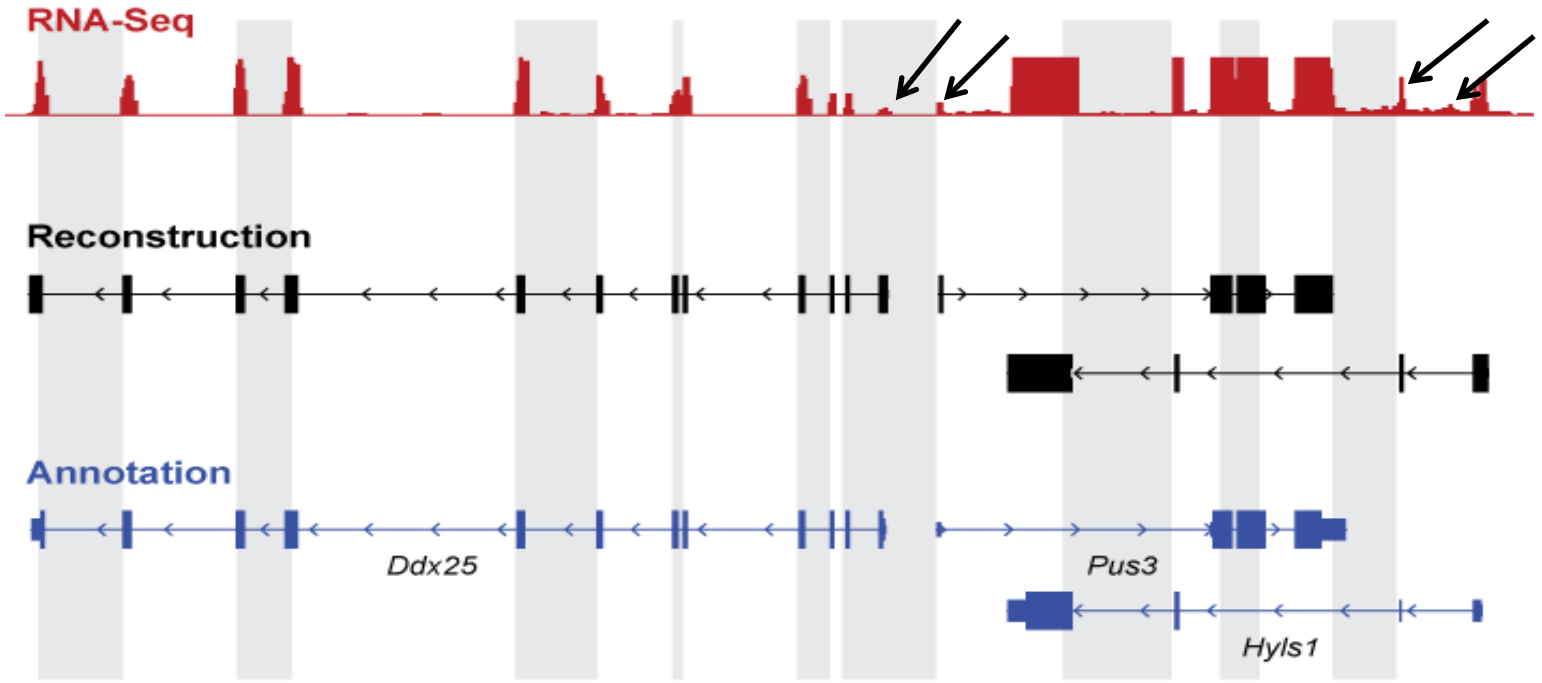



**Library construction can help provide more information. More on this later**

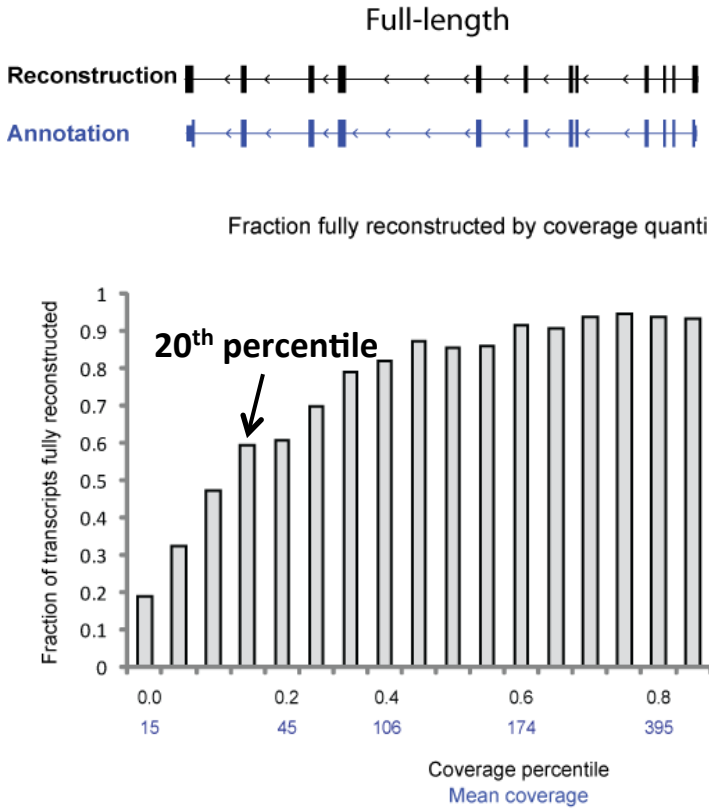
# Applying scripture: Annotating the mouse transcriptome

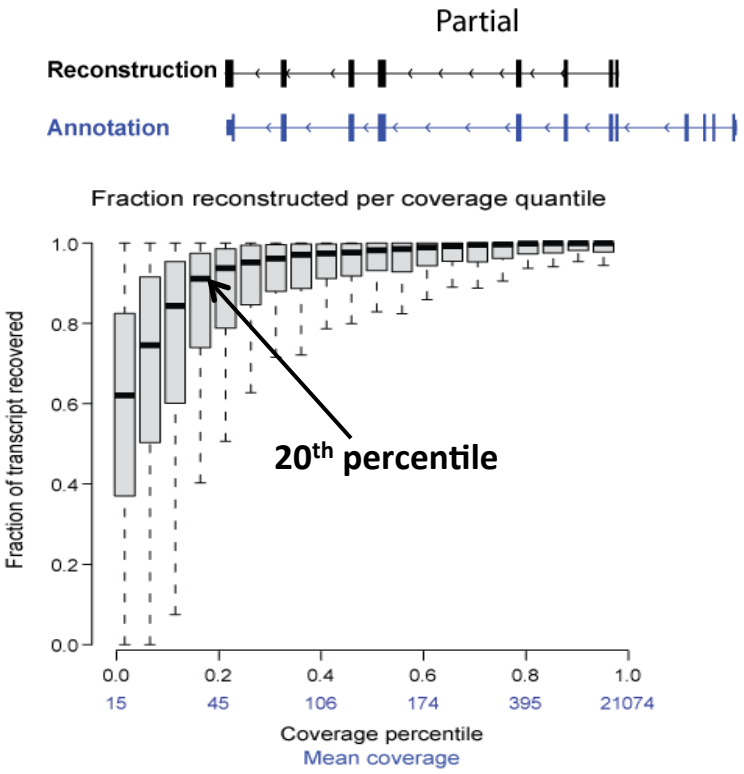
# Reconstructing the transcriptome of mouse cell types


**Mouse Cell Types** → **Sequence** → **Reconstruct**



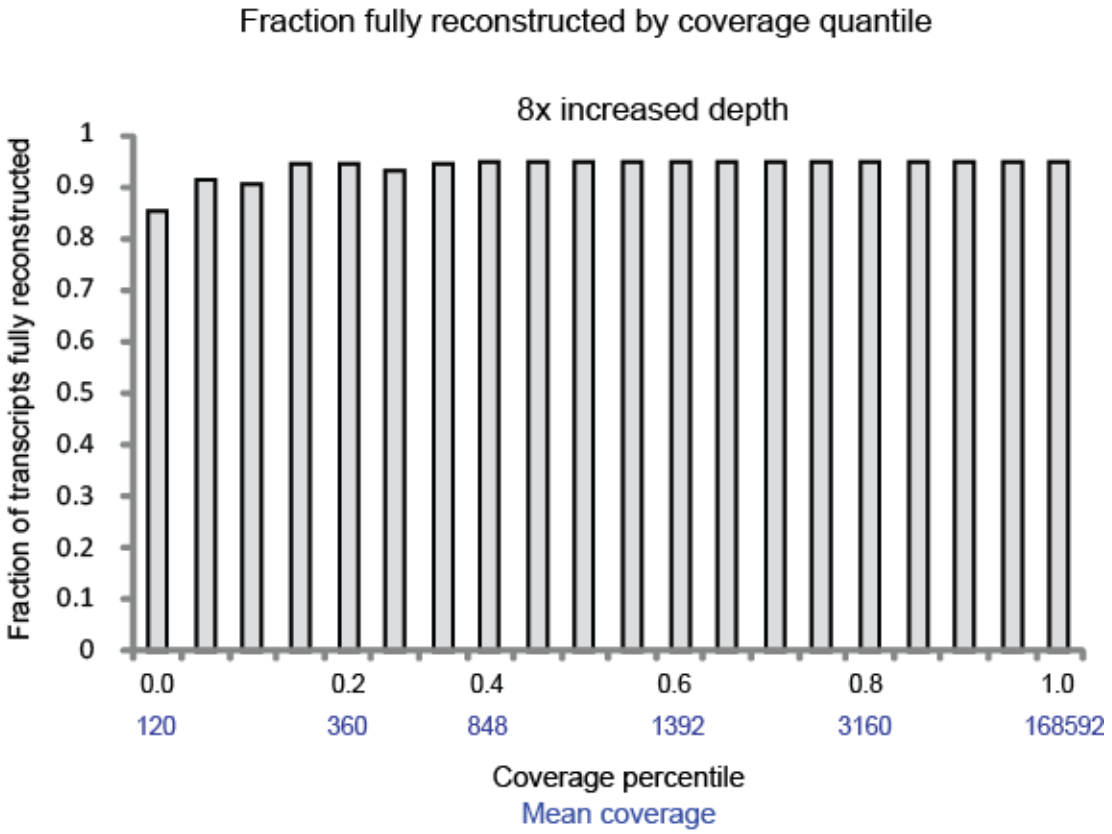
# Sensitivity across expression levels




**Even at low expression (20th percentile), we have:**
**average coverage of transcript is ~95% and 60% have full coverage**

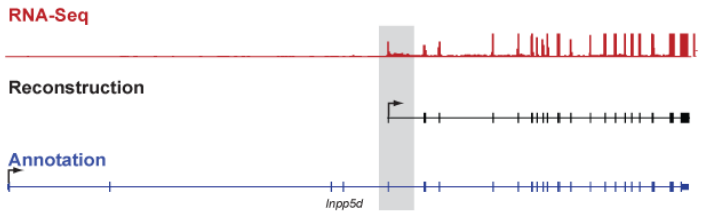
# Sensitivity at low expression levels improves with depth




**As coverage increases we are able to fully reconstruct a larger percentage of known protein-coding genes**

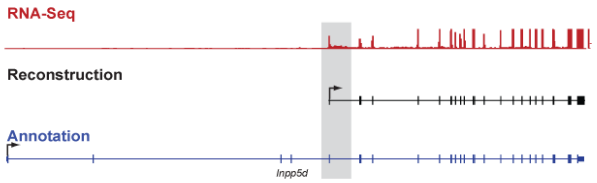
# Novel variation in protein-coding genes

| | ES cells | 3 cell types |
|---|---|---|
| Novel 5’ Start Sites | **1,804** | **3,137** |
| | **1,310** | **2,477** |
| | **588** | **903** |

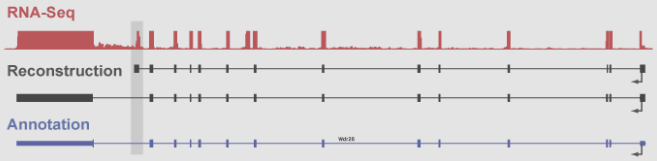


# Novel variation in protein-coding genes

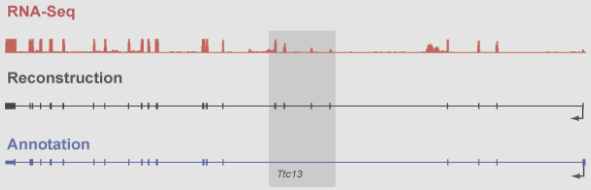
Novel 5’ Start Sites

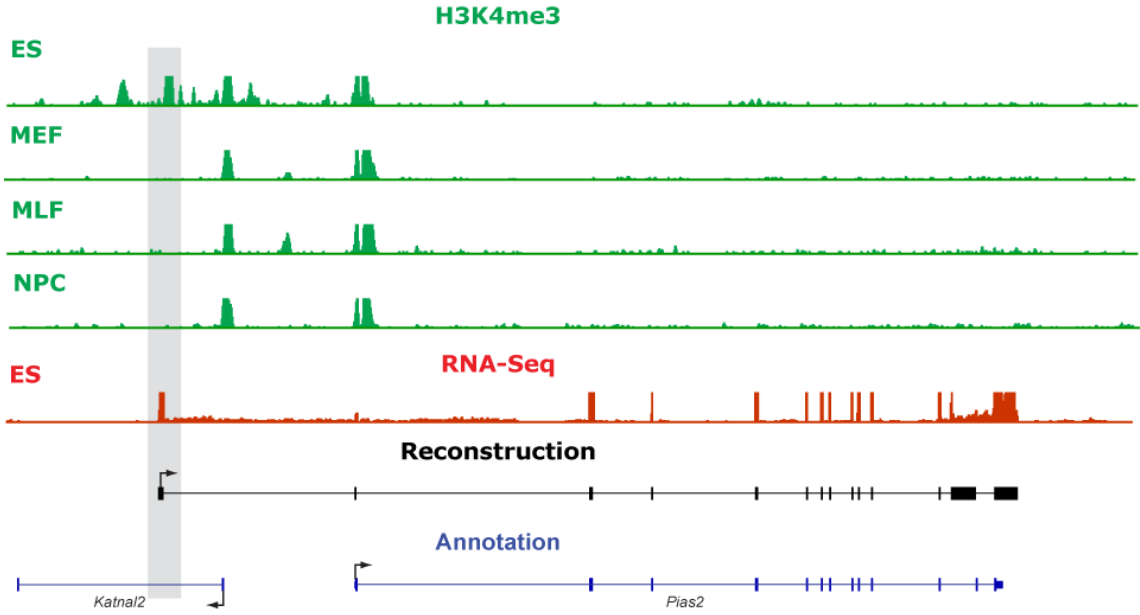



Novel 3’ End




Novel Coding Exons

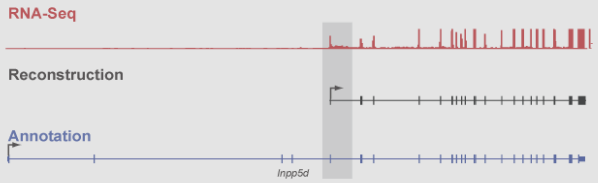


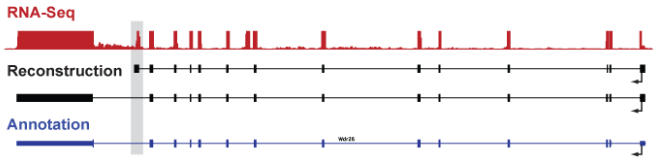

**~85% overlap K4me3**

# Novel variation in protein-coding genes

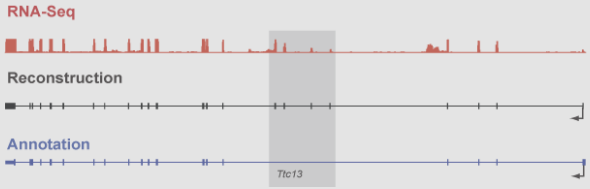
Novel 5’ Start Sites




Novel 3’ End

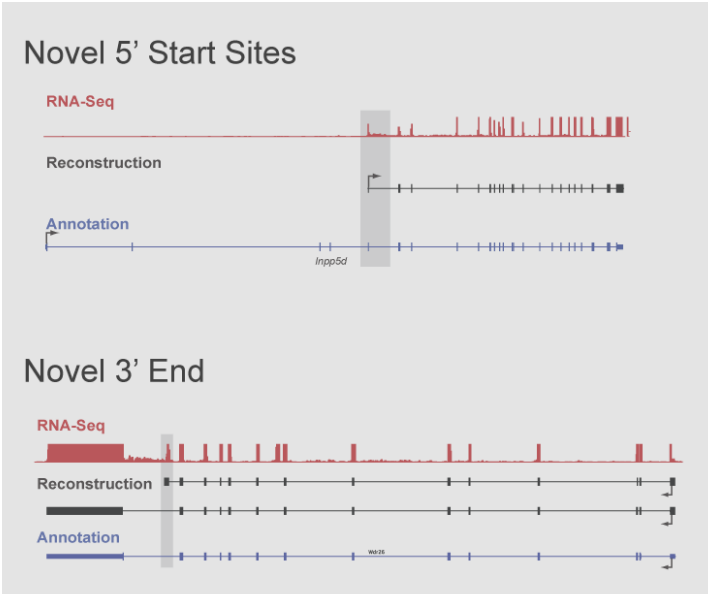


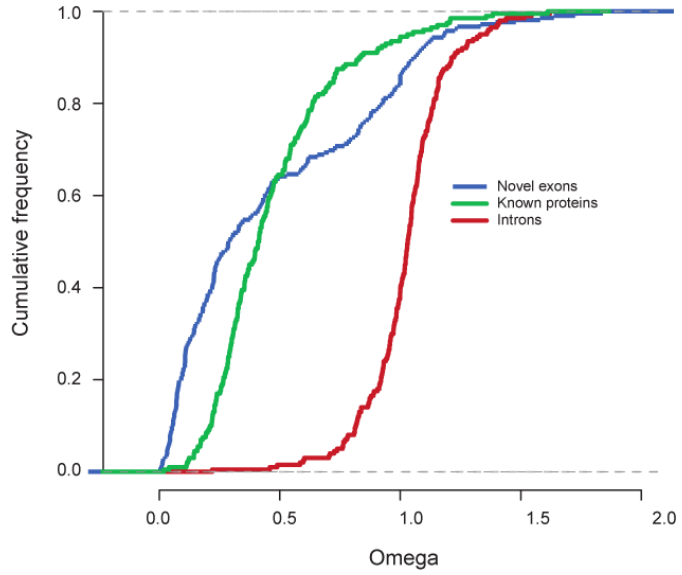

Novel Coding Exons

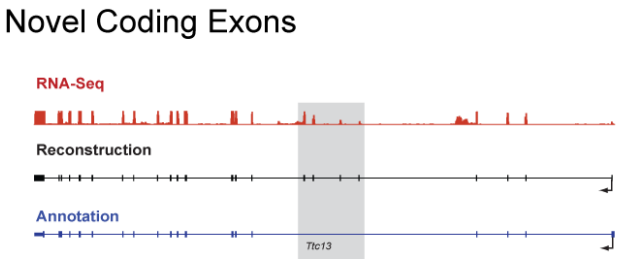



**~50% contain polyA motif**
**Compared to ~6% for random**

# Novel variation in protein-coding genes




**~80% retain ORF**

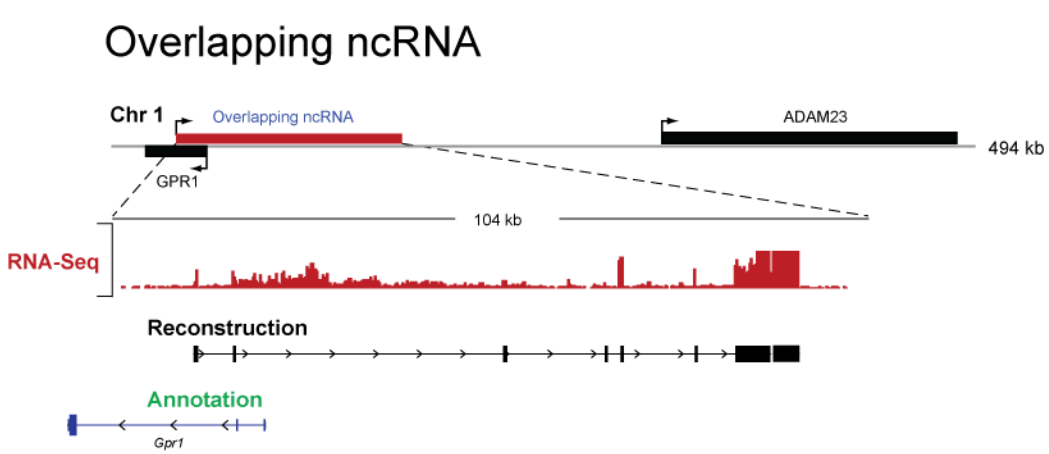
# What about novel genes?

Class 1: Overlapping ncRNA


Class 2: Large Intergenic ncRNA (lincRNA)

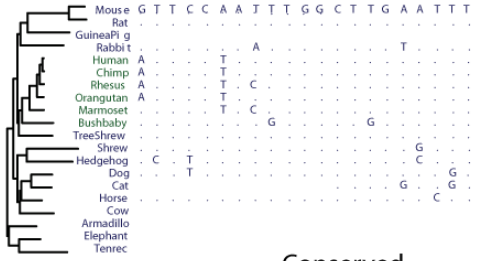

Class 3: Novel protein-coding genes

# Class 1: Overlapping ncRNA

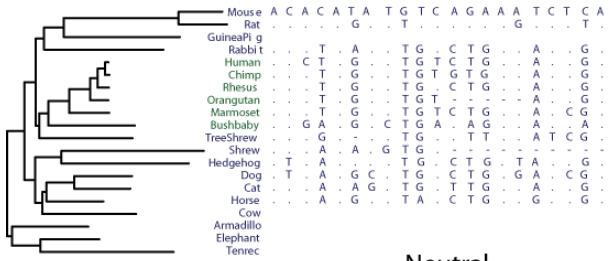



| ES cells | 3 cell types |
| --- | --- |
| **201** | **446** |

# Overlapping ncRNAs: Assessing their evolutionary conservation

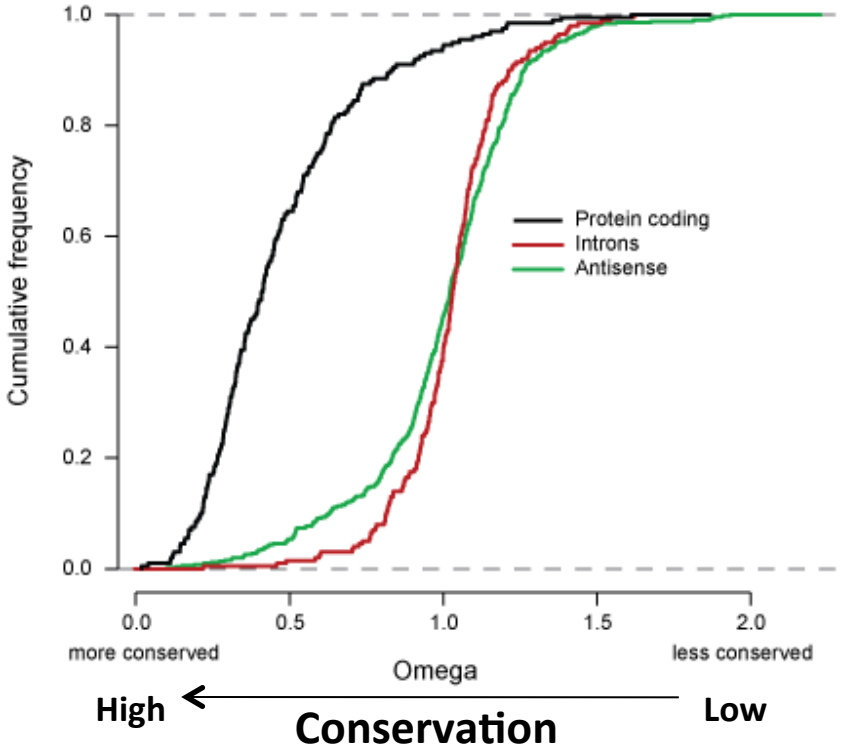



Conserved




Neutral

SiPhy – (Garber et al. *Bioinformatics*, 2009




**Overlapping ncRNAs show little evolutionary conservation**

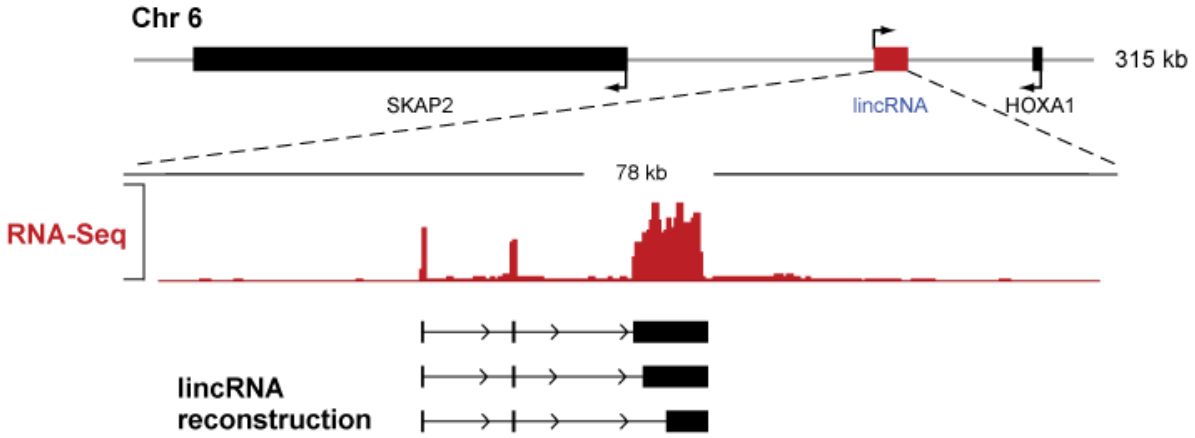
# What about novel genes?

Class 1: Overlapping ncRNA


Class 2: Large Intergenic ncRNA (lincRNA)


Class 3: Novel protein-coding genes

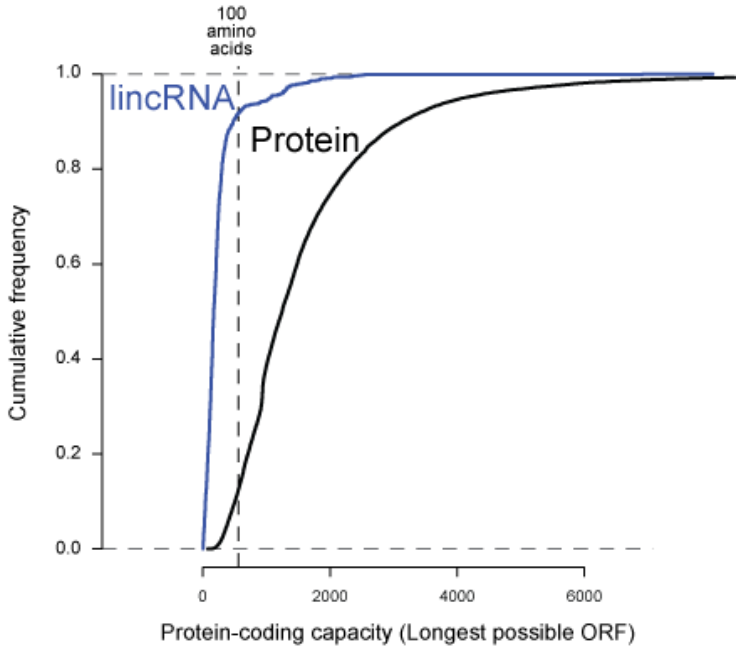
# Class 2: Intergenic ncRNA (lincRNA)



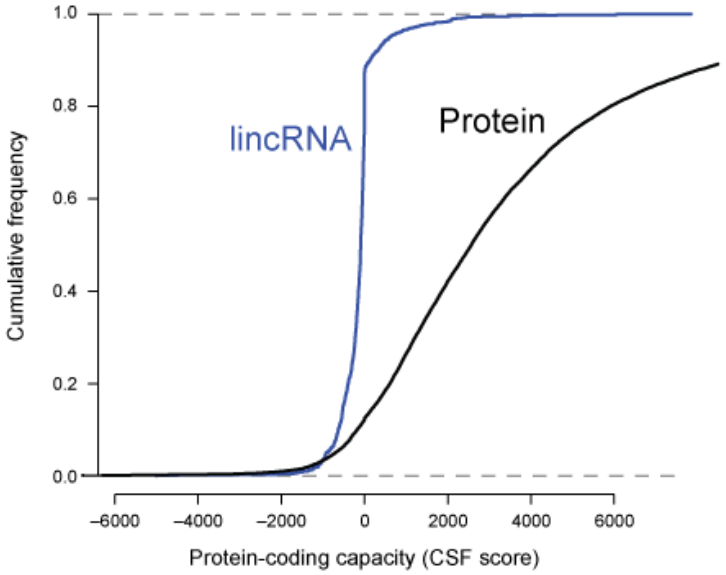

| ES cells | 3 cell types |
| --- | --- |
| **~500** | **~1500** |

# lincRNAs: How do we know they are non-coding?

ORF Length

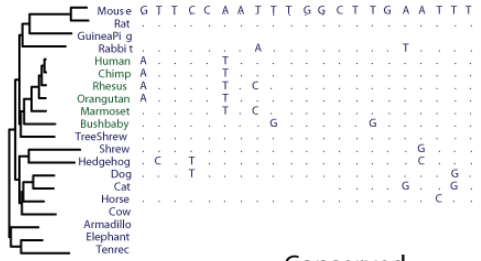



CSF (ORF Conservation)

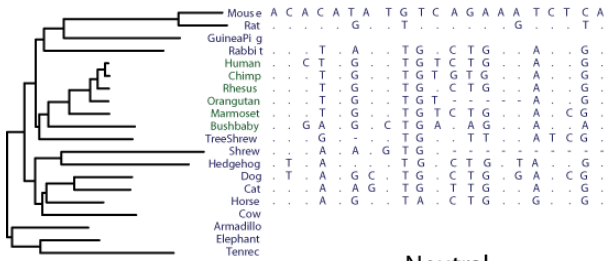



**>95% do not encode proteins**

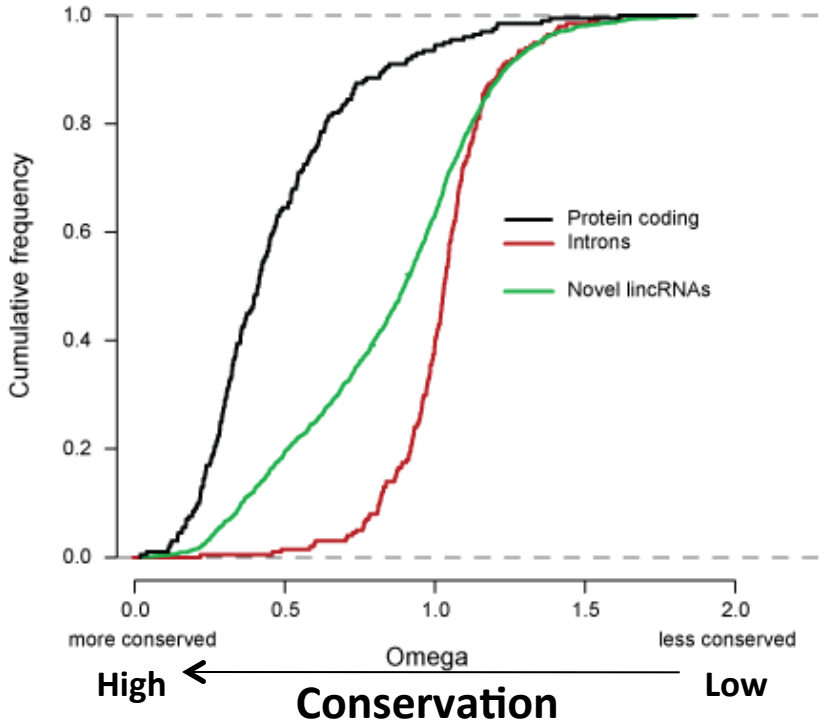
# lincRNAs: Assessing their evolutionary conservation




Conserved




Neutral



# What about novel coding genes?

Class 1: Overlapping ncRNA

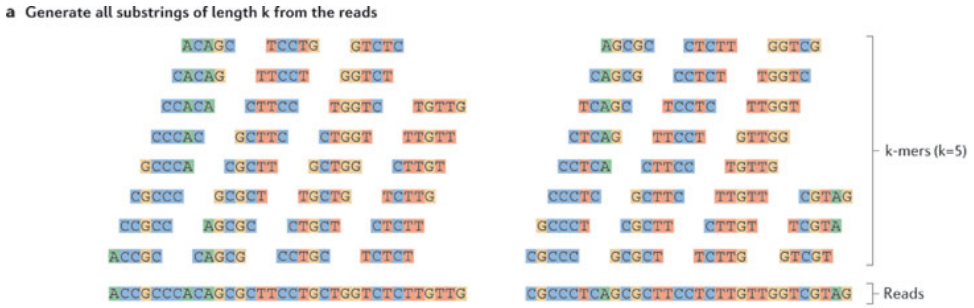

Class 2: Large Intergenic ncRNA (lincRNA)


Class 3: Novel protein-coding genes

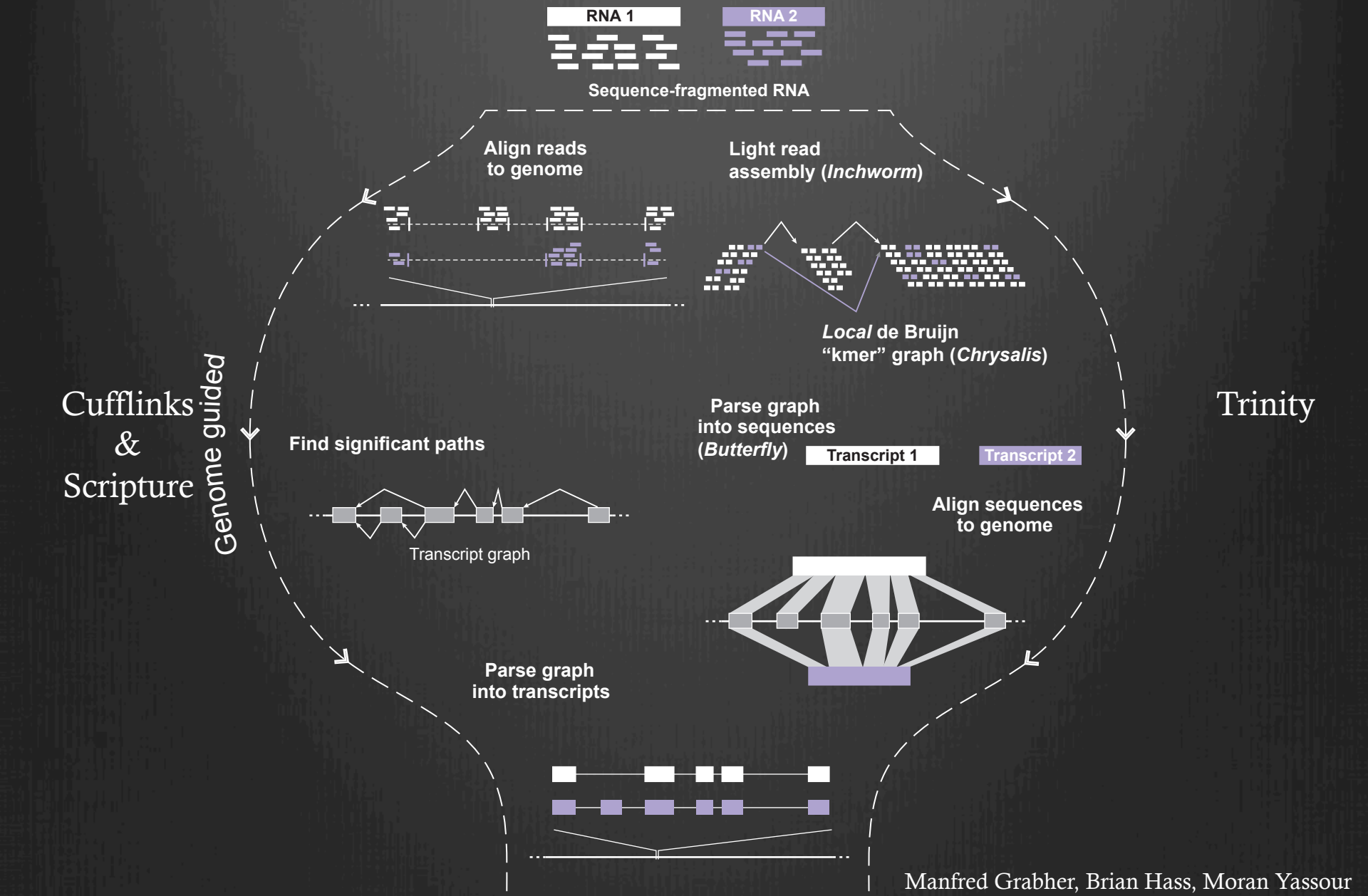

**~40 novel protein-coding genes**

# Other transcript reconstruction methods

# Direct assembly

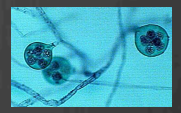


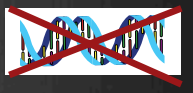
# Genome Independent

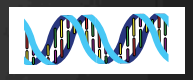


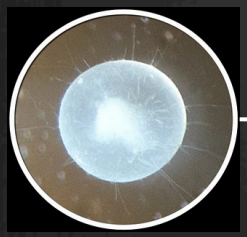

Manfred Grabher, Brian Hass, Moran Yassour

# Trinity: Reference free transcriptomics



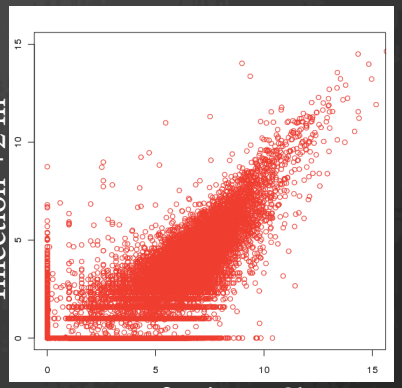




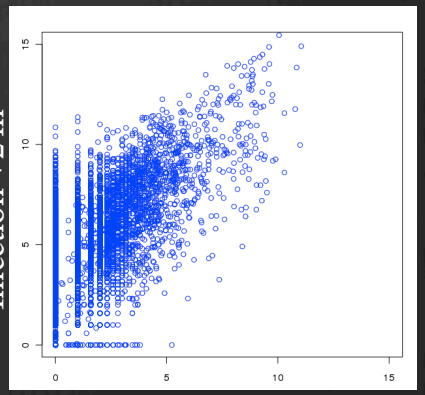


Infected fish cells




Trout




Saprolegnia

Brian Haas, Rays Jiang, Carsten Russ

## Pros and cons of each approach

- Transcript assembly methods are the obvious choice for organisms without a reference sequence.
- Genome-guided approaches are ideal for annotating high-quality genomes and expanding the catalog of expressed transcripts and comparing transcriptomes of different cell types or conditions.
- Hybrid approaches for lesser quality or transcriptomes that underwent major rearrangements, such as in cancer cell.
- More than 1000 fold variability in expression leves makes assembly a harder problem for transcriptome assembly compared with regular genome assembly.
- Genome guided methods are very sensitive to alignment artifacts.

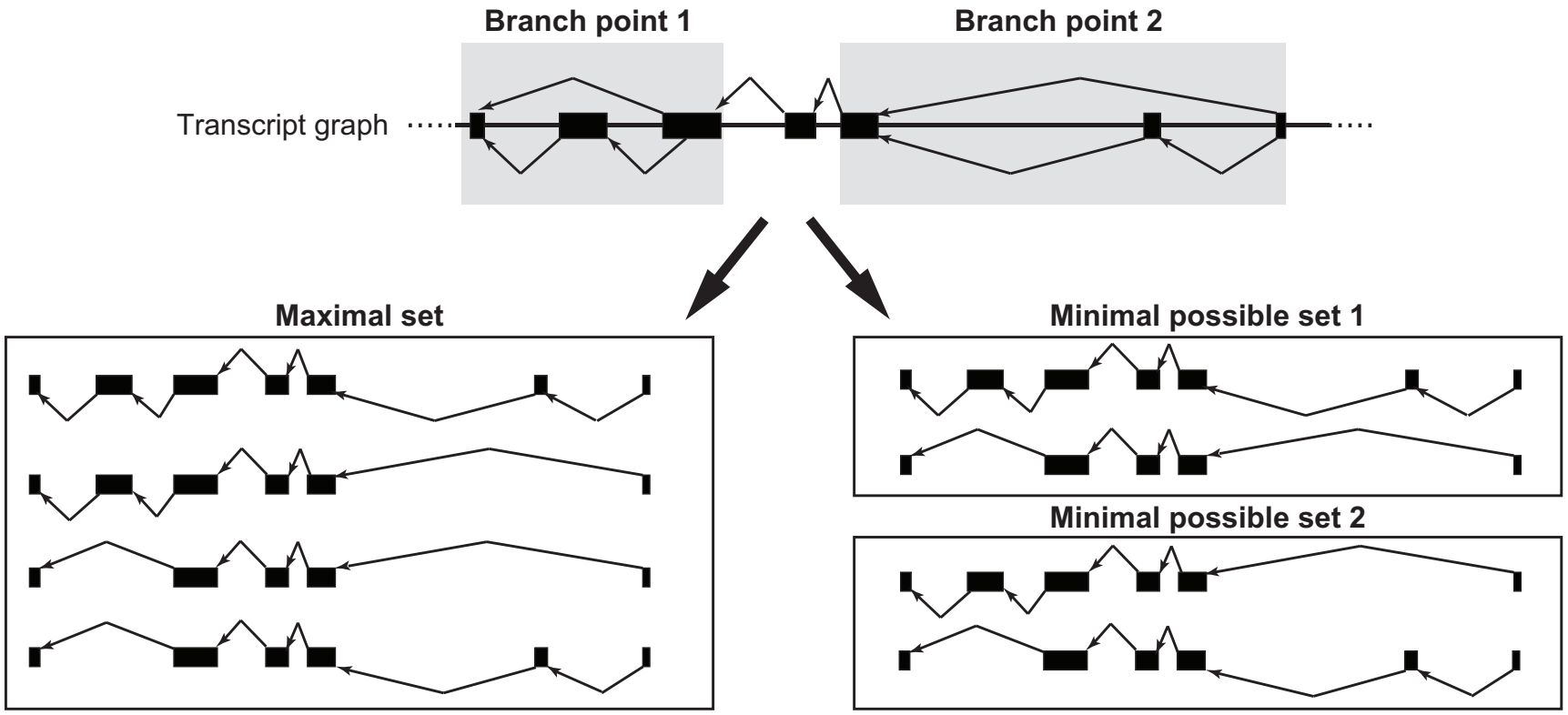
# RNA-Seq transcript reconstruction software

| Assembly | Genome Guided |
| --- | --- |
| Oasis (velvet) | Cufflinks |
| Trans-ABySS | Scripture |
| Trinity | |

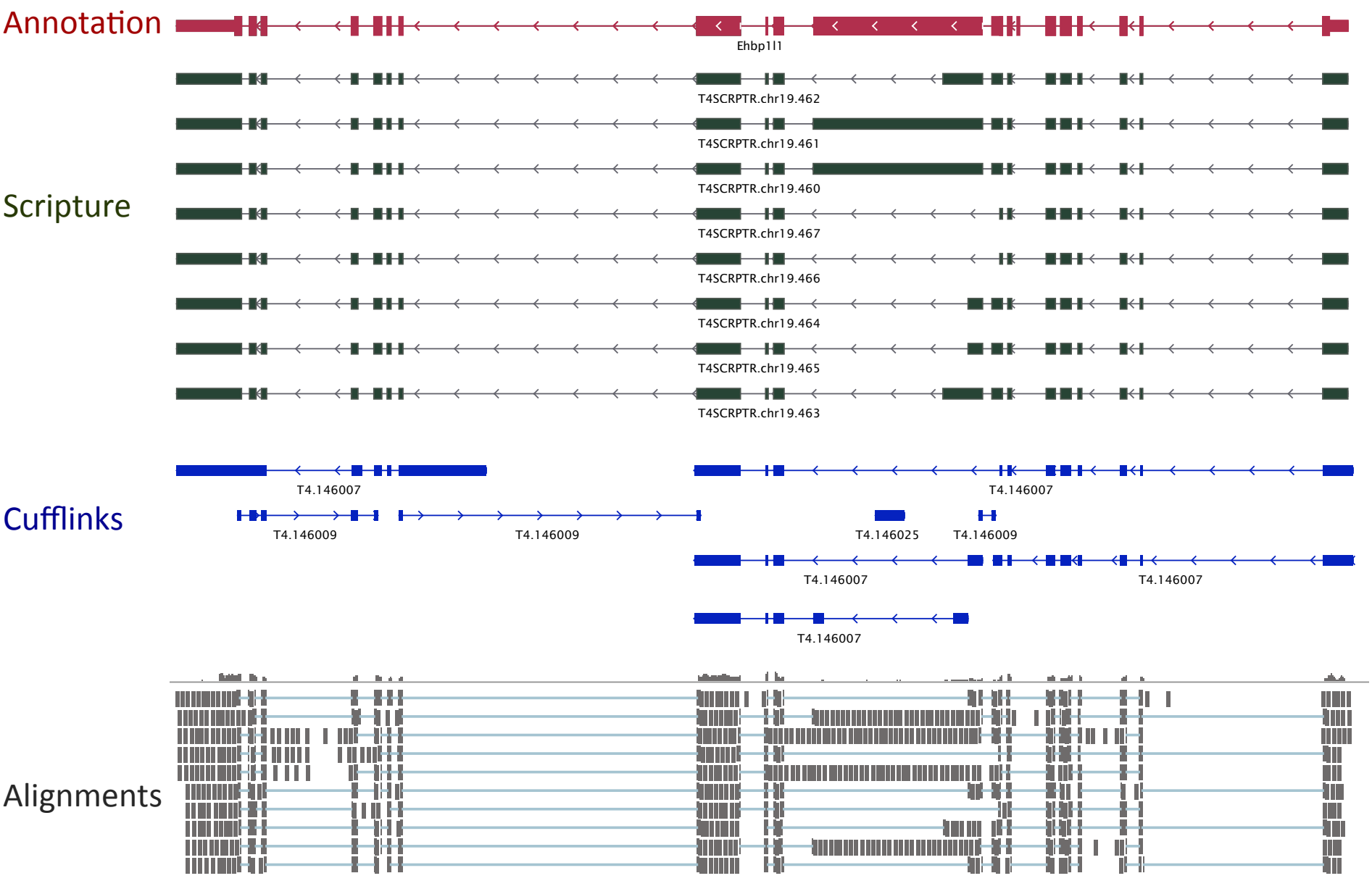
# Differences between Cufflinks and Scripture

- Scripture was designed with annotation in mind. It reports all possible transcripts that are *significantly expressed* given the aligned data (*Maximum sensitivity*).
- Cuffllinks was designed with quantification in mind. It limits reported isoforms to the minimal number that explains the data (*Maximum precision*).

# Maximum sensitivity vs. maximal precision



# Differences between Cufflinks and Scripture - Example

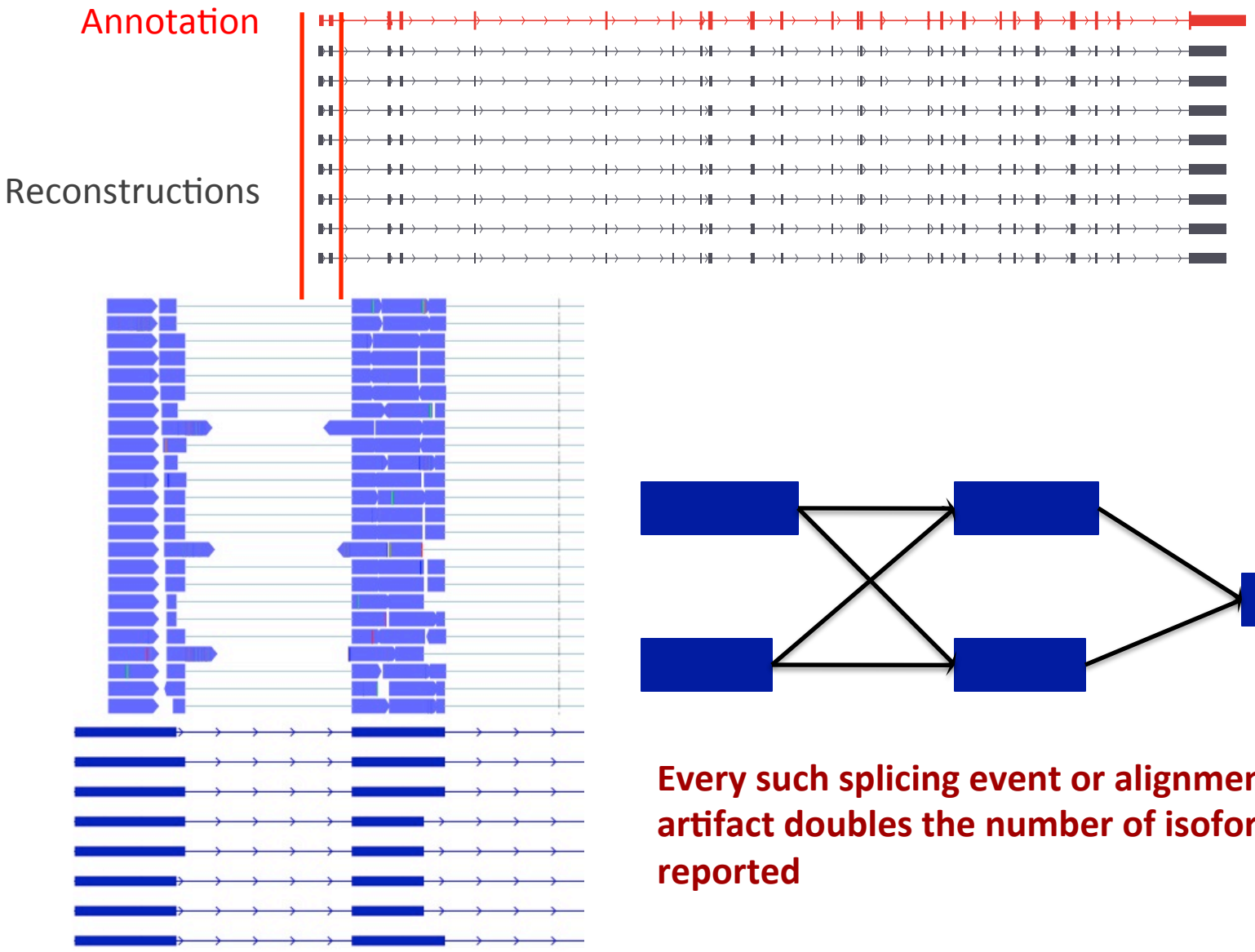


# Comparing reconstructions

| | CPU Hours | Total Memory | Genes fully reconstructed | Mean isoforms per reconstruction | Mean fragments per known annotation | Number of fragments predicted |
|---|---|---|---|---|---|---|
| **Cufflinks** | 10 | 1.4 G | 5,994 | 1.2 | 1.4 | **159,856** |
| **Scripture** | 16 | 3.5 G | 6,221 | **1.6** | 1.3 | 61,922 |
| **Trans- Abyss** | 650 | 120 G[4] | 3,330 | 4.7 | 2.6 | 3,117,238 |

**Many of the bogus locus and isoforms are due to alignment artifacts**

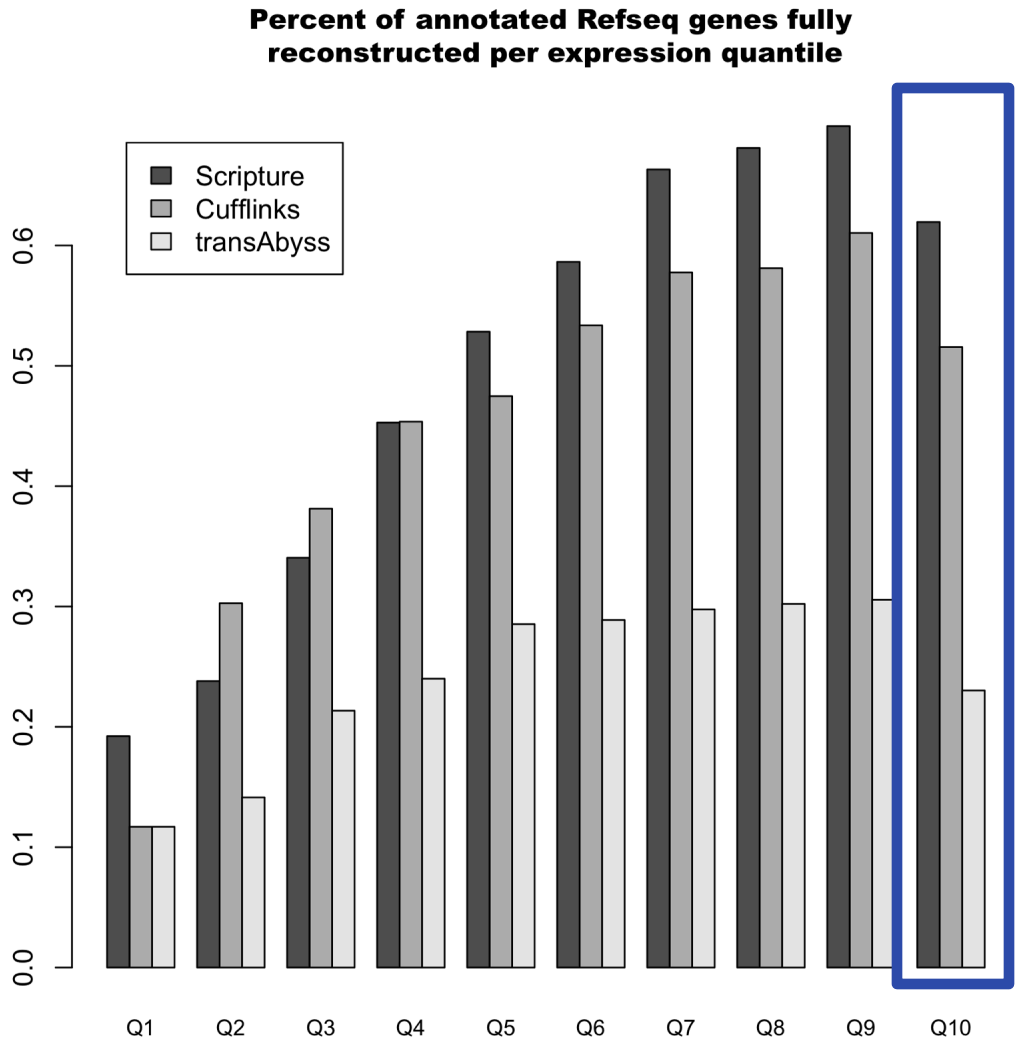
# Why so many isoforms





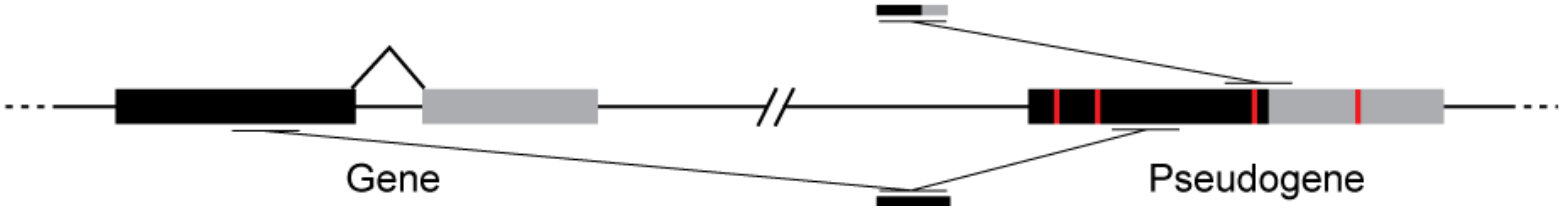

**Every such splicing event or alignment artifact doubles the number of isoforms reported**

# Reconstruction comparison



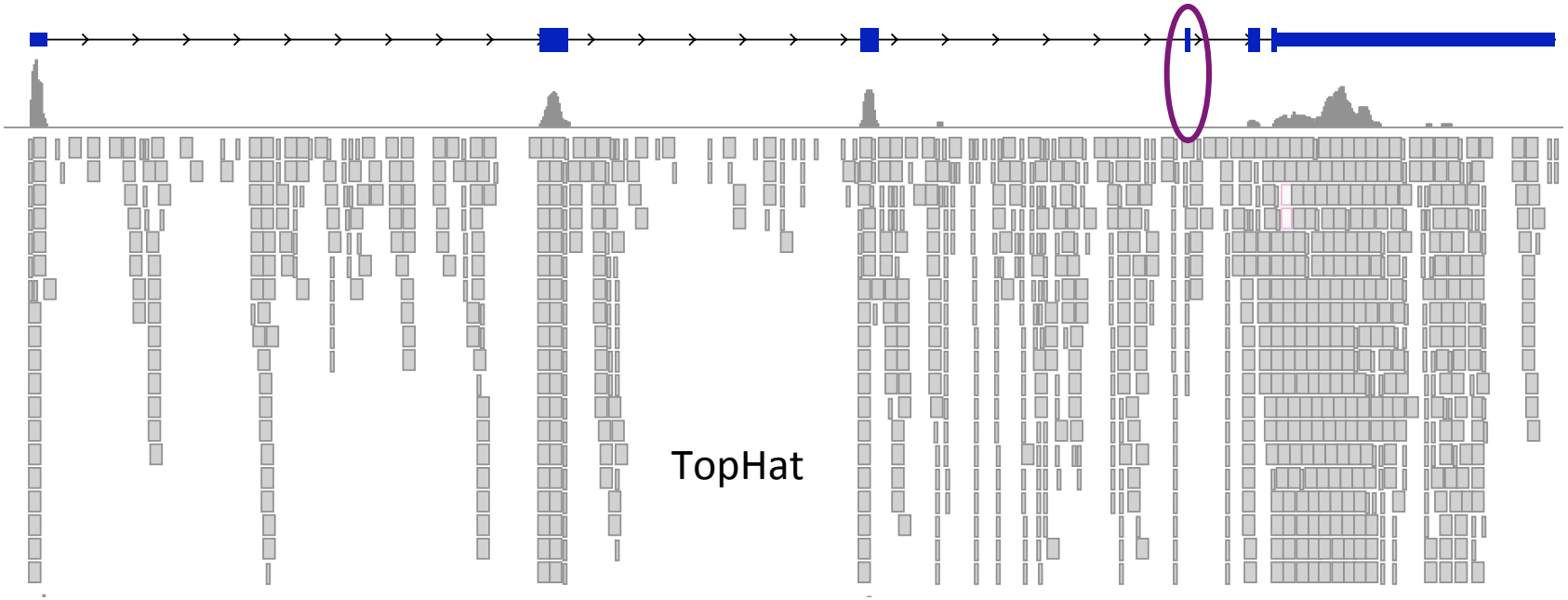
Alignment revisited — spliced alignment is still work in progress

# Exon-first aligners are faster but at cost




**Alignment artifacts can also decrease sensitivity**

# Missing spliced reads for highly expressed genes



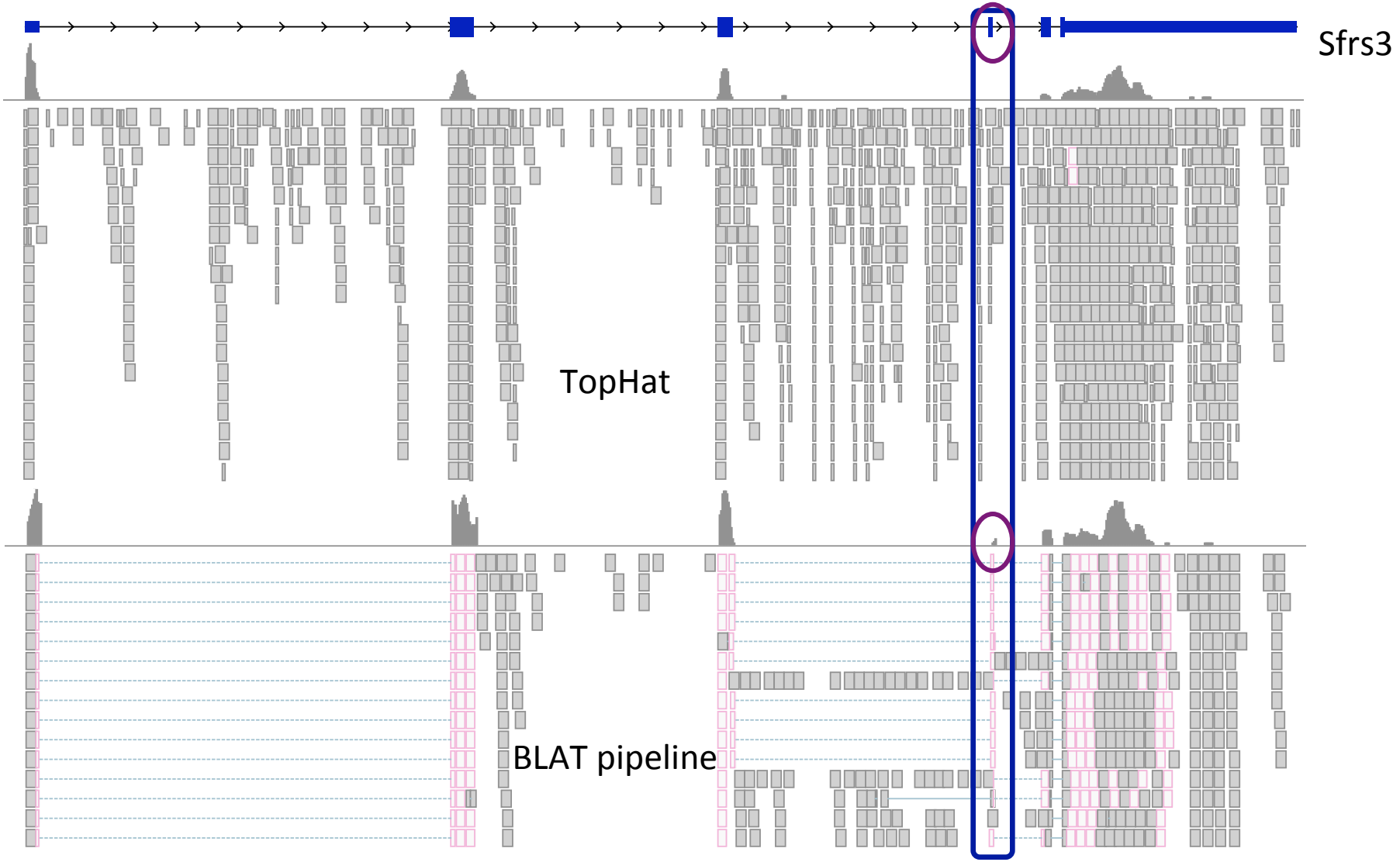

Read mapped uniquely

Read ambiguously mapped

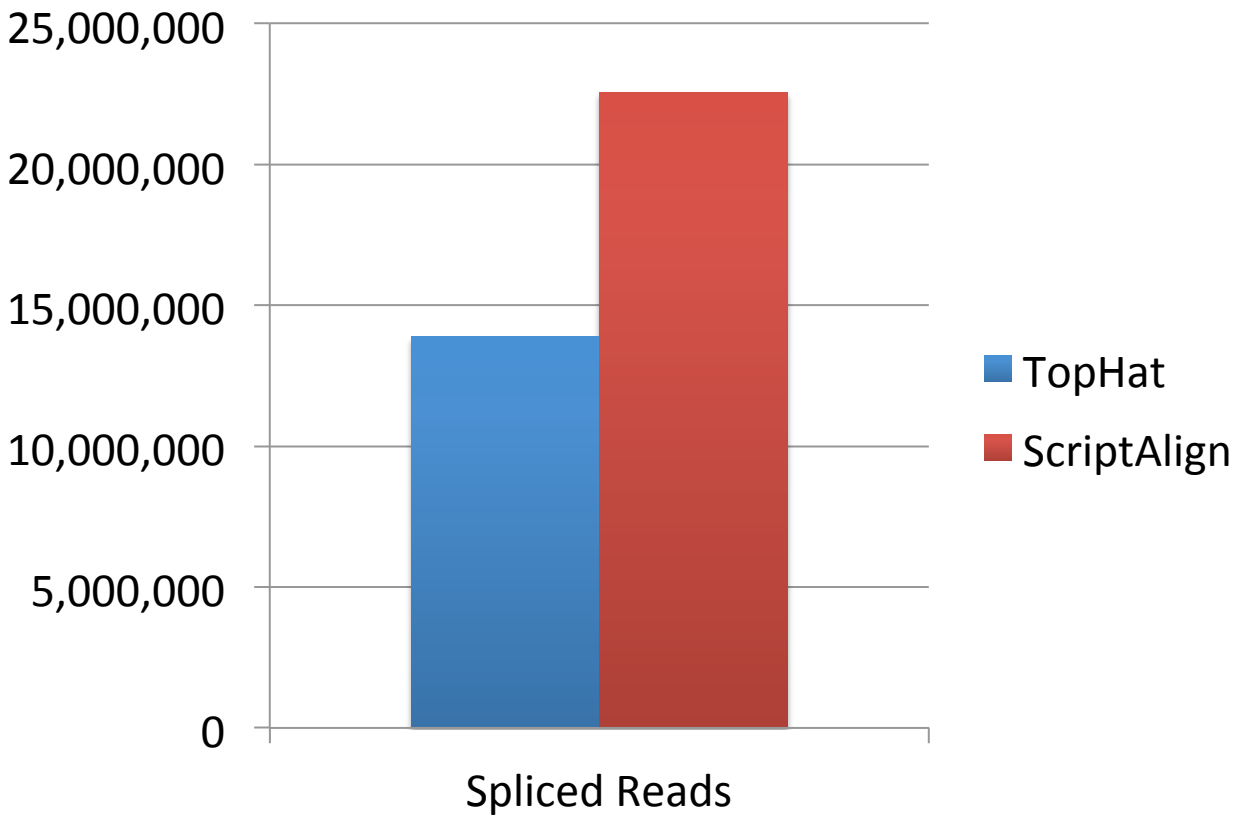
# Can more sensitive alignments overcome this problem?

- Use gapped aligners (e.g. BLAT) to map reads
  - Align all reads with BLAT
  - Filter hits and build candidate junction "database" from BLAT hits (Scripture light).
  - Use a short read aligner (Bowtie) to map reads against the connectivity graph inferred transcriptome
  - Map transcriptome alignments to the genome

# Many junctions can be rescued



# ScriptAlign: Can increase alignment across junctions




**"Map first" reconstruction approaches directly benefit with mapping improvements**
**We even get more uniquely aligned reads (not just spliced reads)**

## Practical advise on spliced alignment

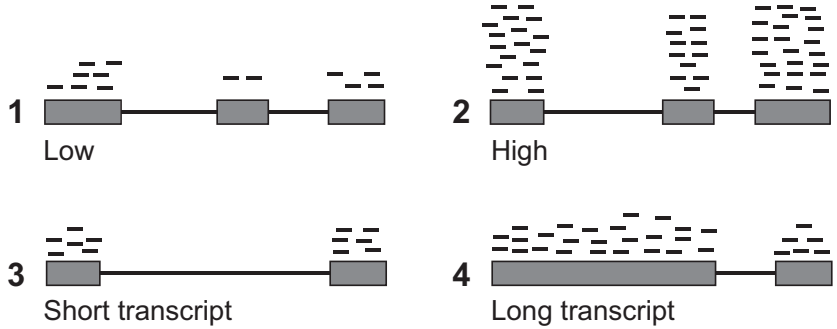
- Use tophat, specify a transcript set if one is availble
- Align twice:
  - Align once, keep splice junctions
  - Realign using both transcript and junction set

# Overview of the session

The 3 main computational challenges of sequence analysis for *counting applications*:

- Read mapping: Placing short reads in the genome
- Reconstruction: Finding the regions that originate the reads
- Quantification:
  - Assigning scores to regions
  - Finding regions that are differentially represented between two or more samples.

# Quantification: only one isoform



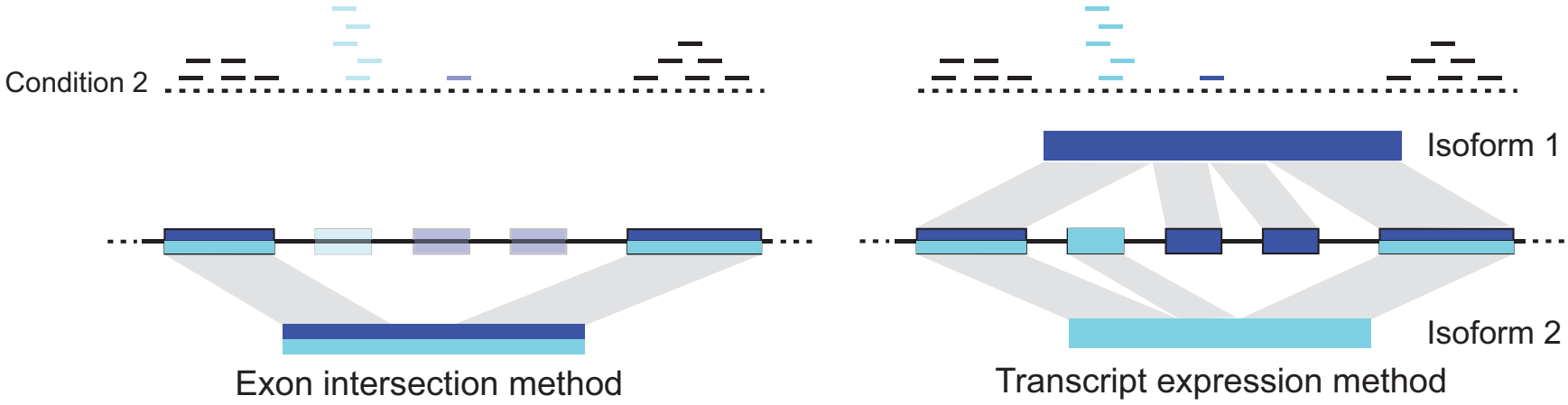

$$RPKM = 10^9 \frac{\#reads}{length \times TotalReads}$$

Reads per kilobase of exonic sequence per million mapped reads
(*Mortazavi* et al Nature methods 2008)

- Fragmentation of transcripts results in length bias: longer transcripts have higher counts
- Different experiments have different yields. Normalization may be required for cross lane comparisons

**This is all good when genes have one isoform.**

# Quantification: gene expression with multiple isoforms

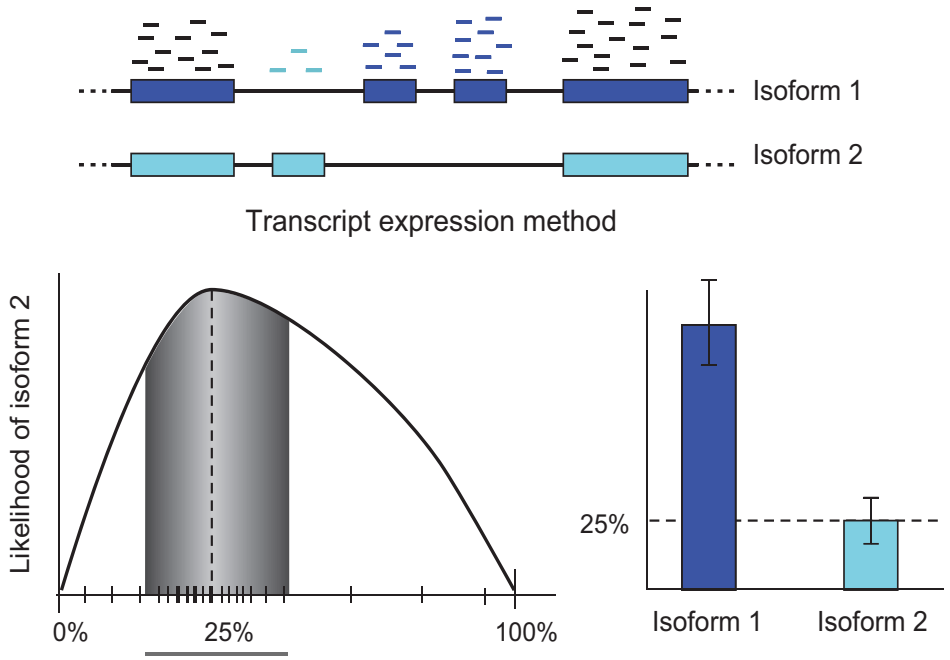



Exon intersection model: Score constituent exons

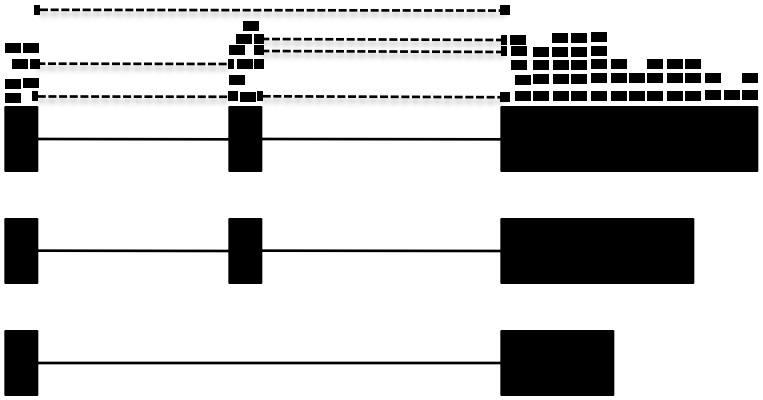
Exon union model: Score the the “merged” transcript

Transcript expression model: Assign reads uniquely to different isoforms. *Not a trivial problem!*

# Quantification: read assignment method

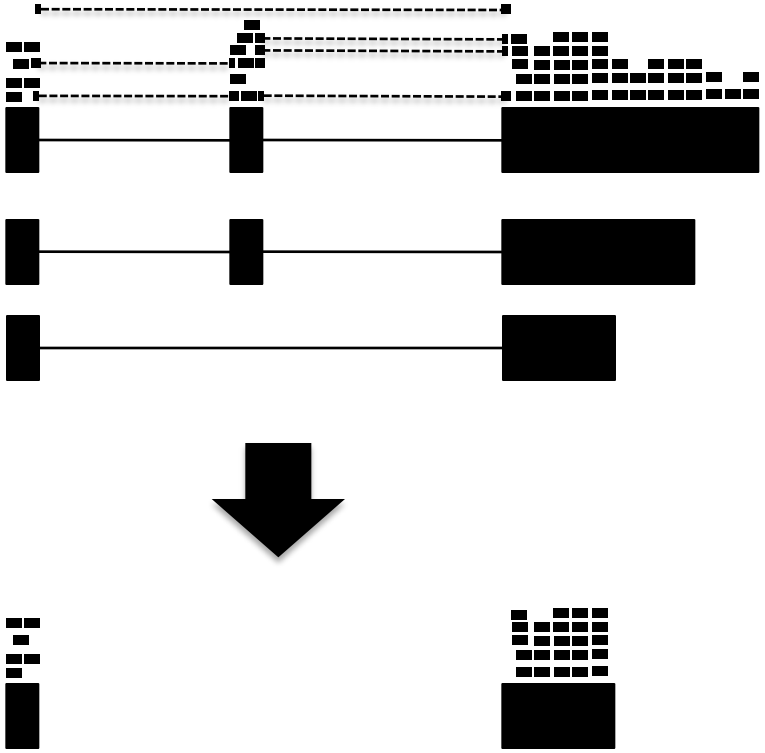


# Quantification with multiple isoforms



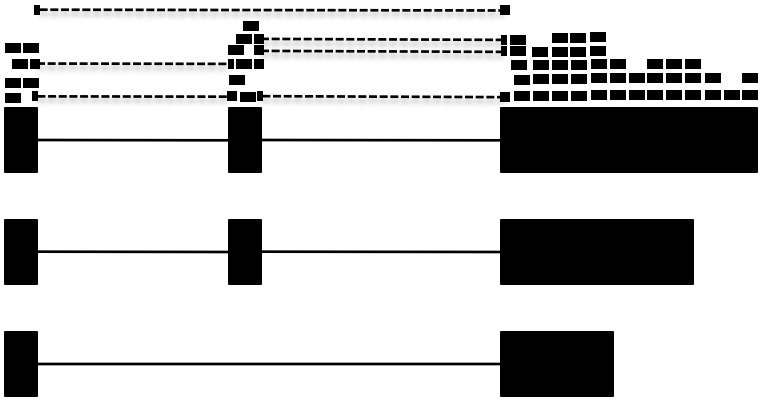
How do we define the gene expression?
How do we compute the expression of each isoform?

# Computing gene expression

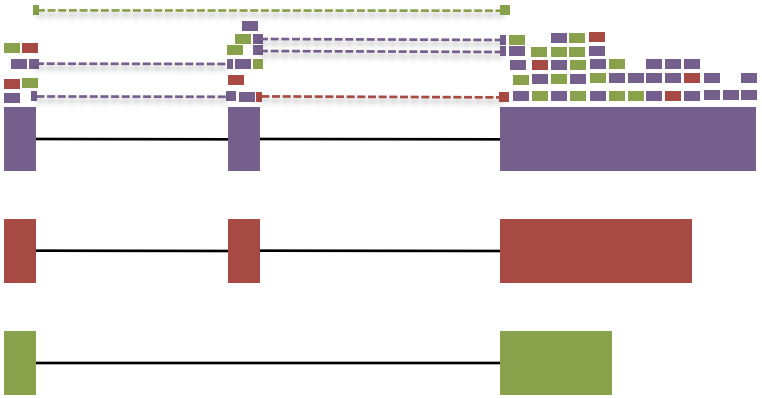


Idea1: RPKM of the constitutive reads (Neuma, Alexa-Seq, Scripture)

# Computing gene expression — isoform deconvolution



# Computing gene expression — isoform deconvolution



If we knew the origin of the reads we could compute each isoform's expression. The gene's expression would be the sum of the expression of all its isoforms.

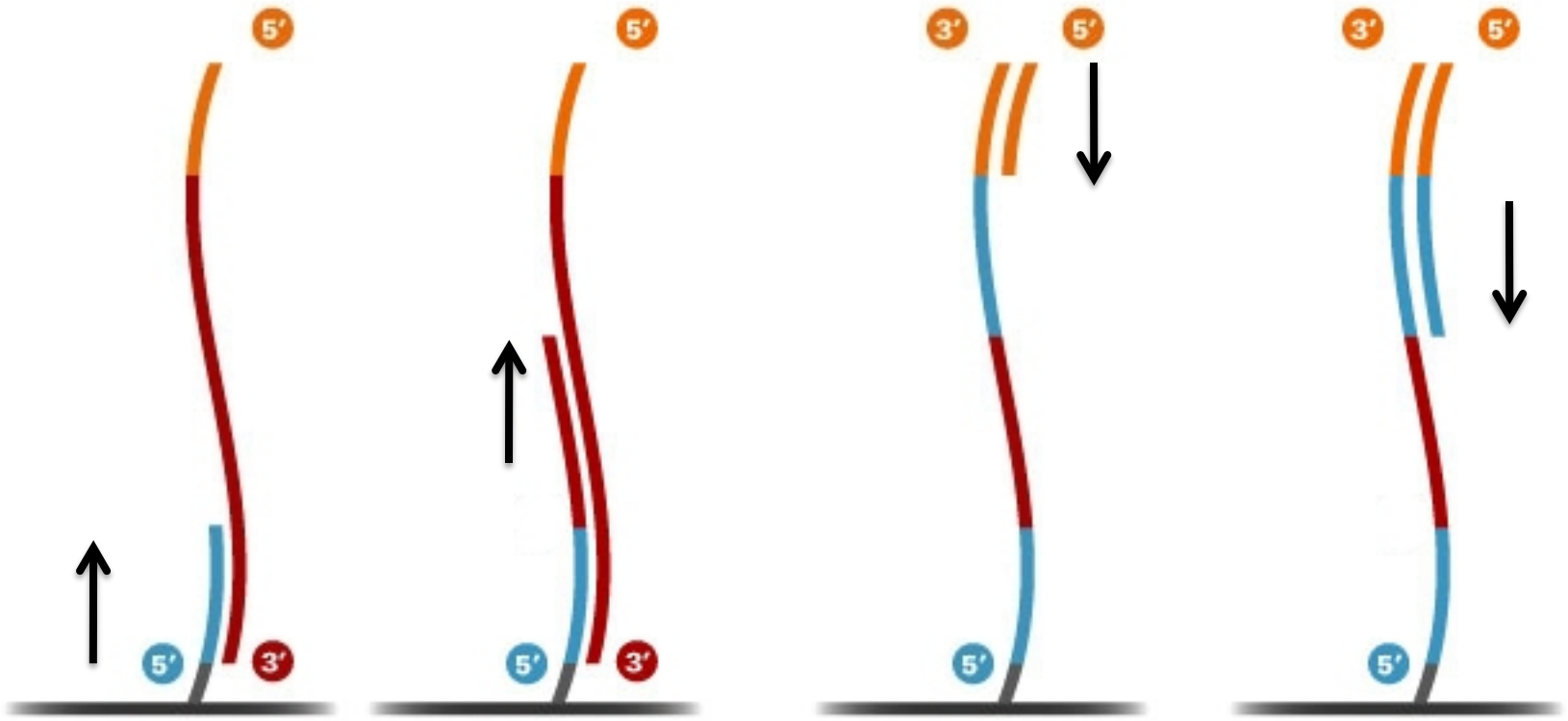
$$E = RPKM_1 + RPKM_2 + RPKM_3$$

# Programs to measure transcript expression

| | Implemented method |
|---|---|
| Alexa-seq | Gene expression by constitutive exons |
| ERANGE | Gene expression by using all Exons |
| Scripture | Gene expression by constitutive exons |
| **Cufflinks** | **Transcript deconvolution by solving the maximum likelihood problem** |
| **MISO** | **Transcript deconvolution by solving the maximum likelihood problem** |
| **RSEM** | **Transcript deconvolution by solving the maximum likelihood problem** |

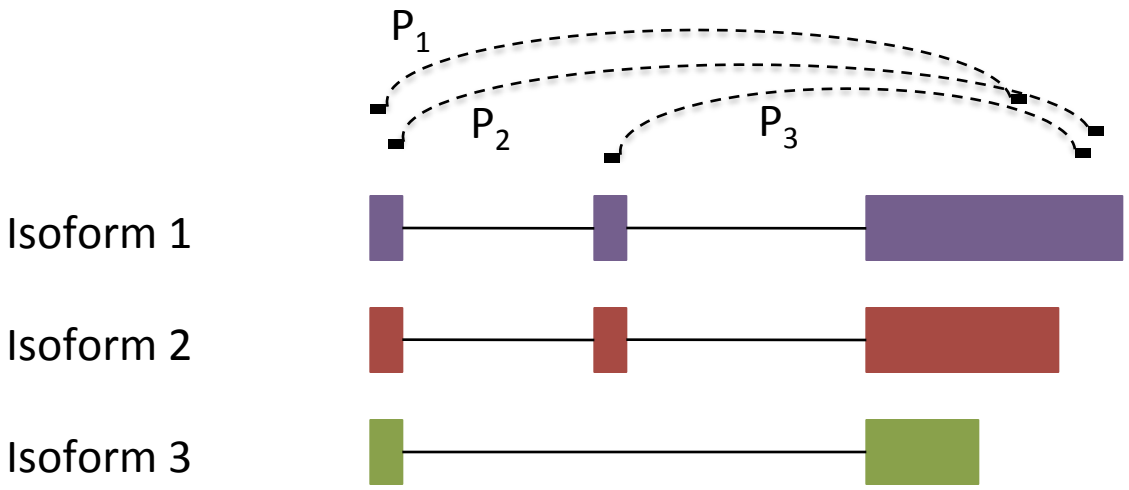
# Impact of library construction methods

# Paired-end sequencing impact in analysis




Adapted from the Helicos website

# Paired-end reads are easier to associate to isoforms

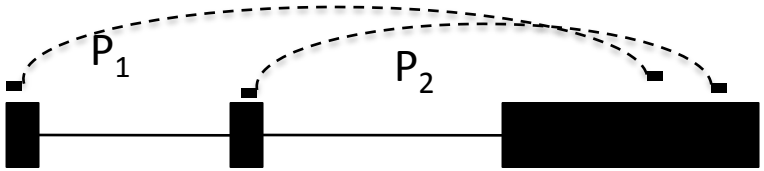



Paired ends increase isoform deconvolution confidence

- $P_1$ originates from isoform 1 or 2 but not 3.
- $P_2$ and $P_3$ originate from isoform 1

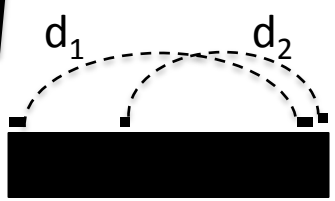
**Do paired-end reads also help identifying reads originating in isoform 3?**

# We can estimate the insert size distribution

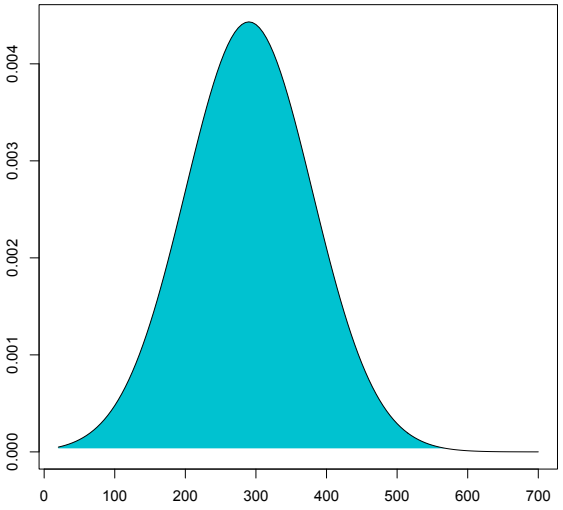



Get all single isoform reconstructions

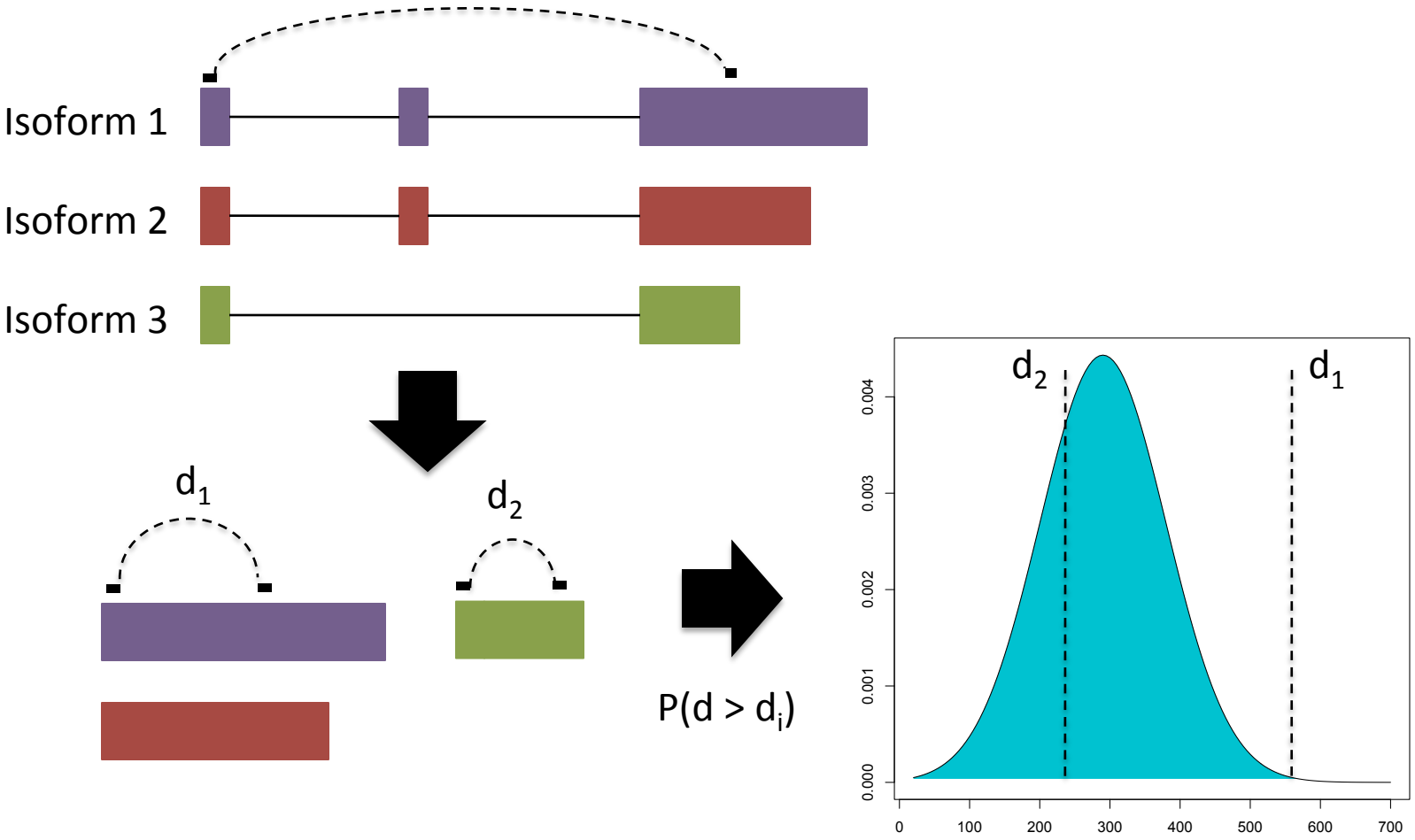
Splice and compute insert distance



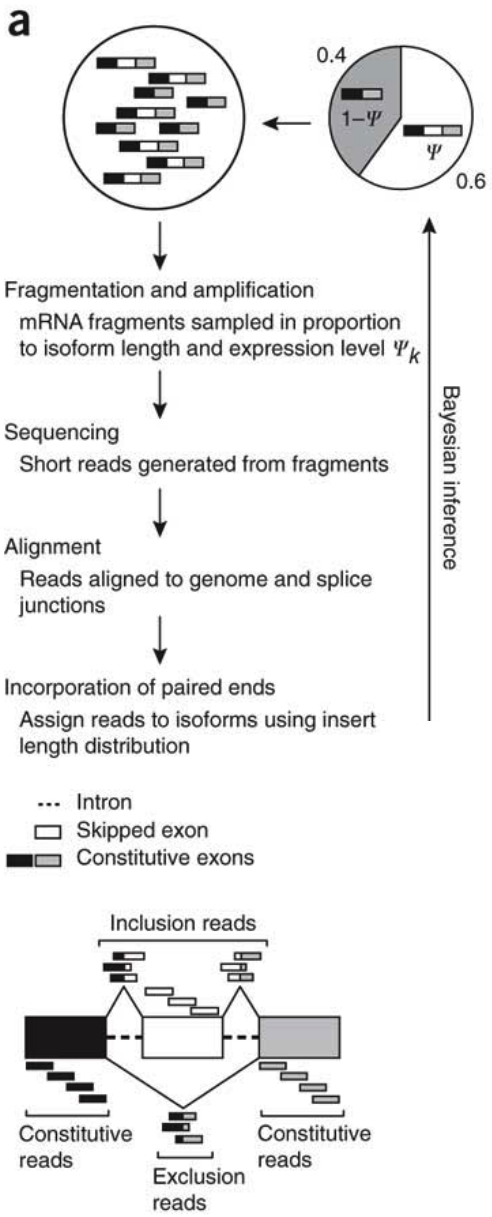

Estimate insert size empirical distribution



# … and use it for probabilistic read assignment



# And improve quantification

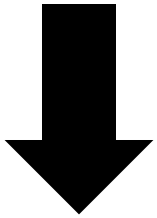



Katz et al Nature Methods 2008

# Quantification with paired ends (FPKM)

Cufflinks leverages paired ends to quantify fragments rather than raw reads. The extension of RPKM.

$$RPKM = 10^9 \frac{\#\,reads}{length \times Total\,\mathrm{Re}\,ads}$$

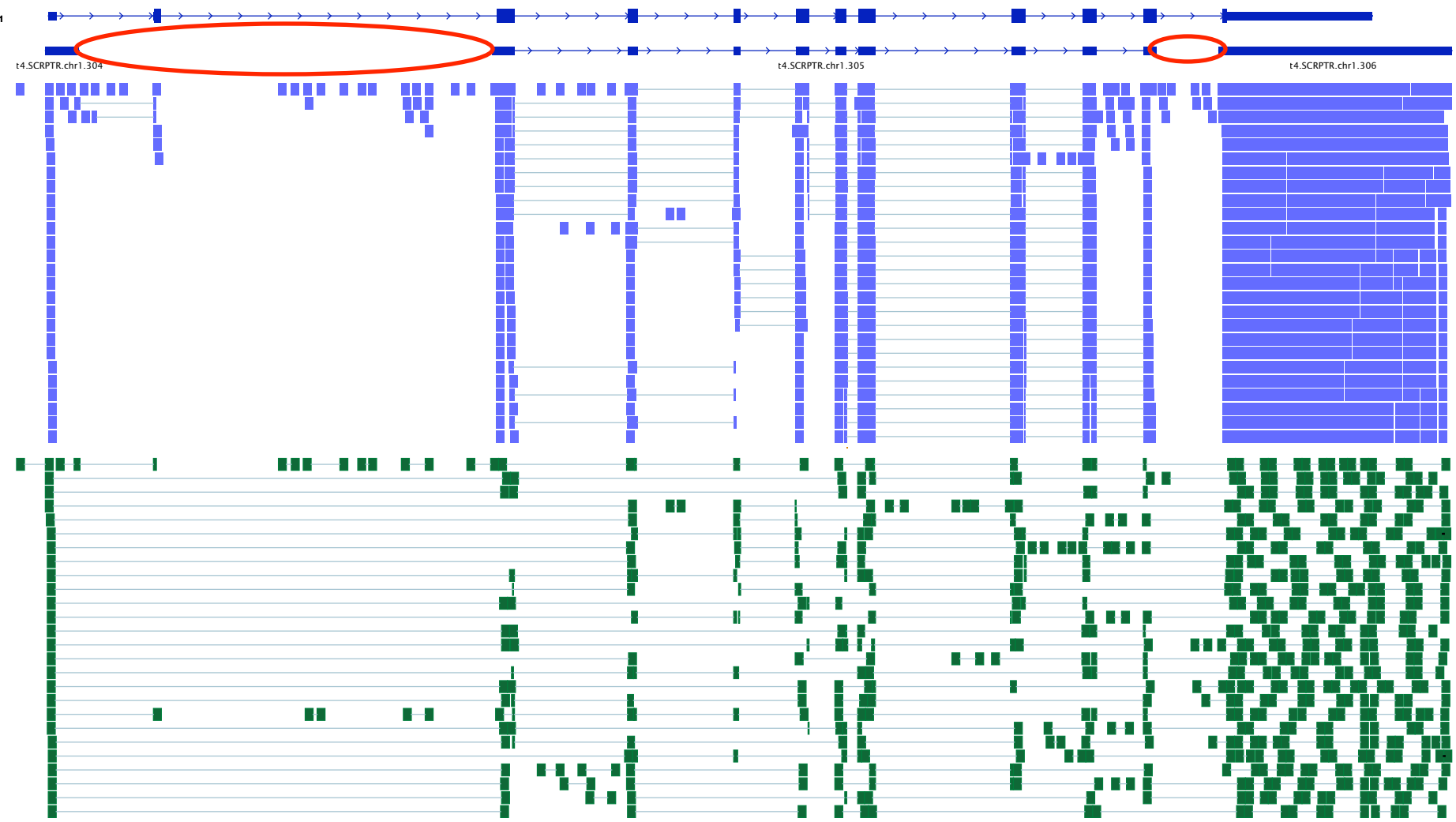


$$FPKM = 10^9 \frac{\#\,fragments}{length \times TotalFragments}$$

Fragments per kilobase of exonic sequence per million mapped fragments
(*Trapnel* et al Nature Biotechnology 2010)

**paired-end reads improve quantification accuracy**

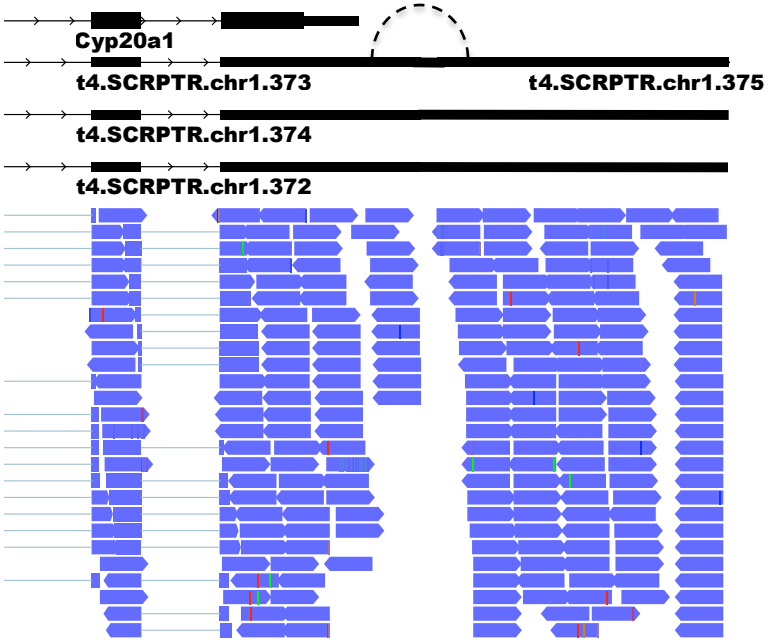
# Paired-end improve reconstructions




**Paired-end data complements the connectivity graph**

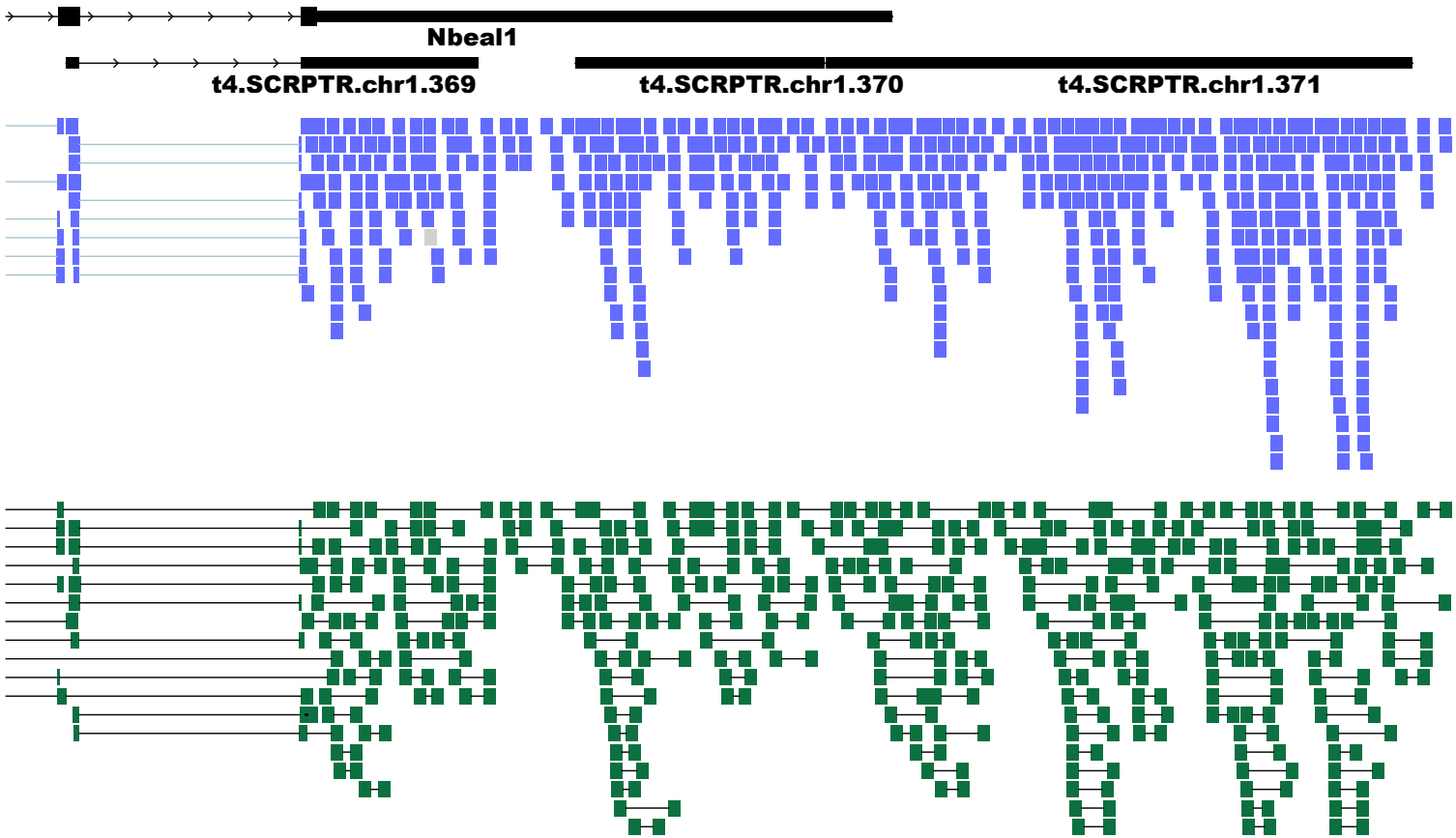
# And merge regions

Single reads




Paired reads

# Or split regions

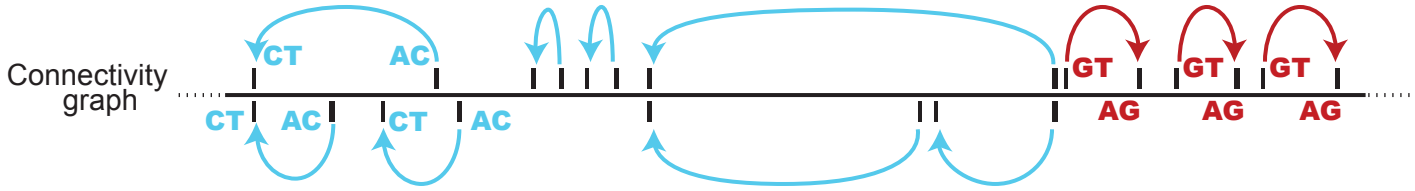


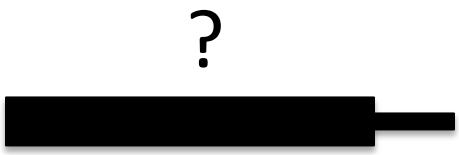
# Summary

- Paired-end reads are now routine in Illumina and SOLiD sequencers.
- Paired end alignment is supported by most short read aligners
- Transcript quantification depends heavily in paired-end data
- Transcript reconstruction is greatly improved when using paired-ends (work in progress)

# Giving orientation to transcripts — Strand specific libraries

Scripture relies on splice motifs to orient transcripts. It orients every edge in the connectivity graph.



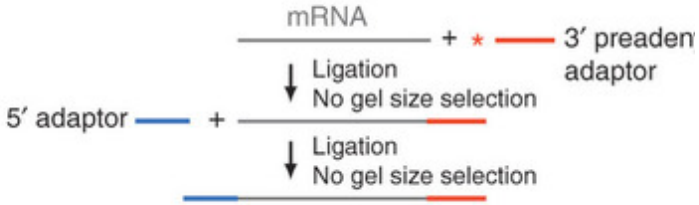

Single exon genes are left unoriented

# Strand specific library construction results in oriented reads.

**Illumina RNA ligation**

3′ preadenylated adaptors and 5′ adaptors ligated sequentially to RNA without cleanup

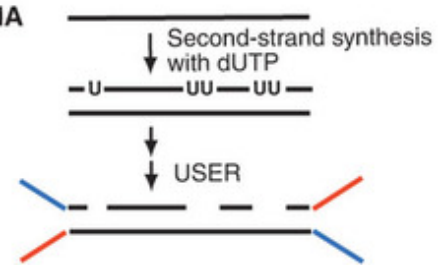
(S. Luo and G. Schroth, personal communication)




Sequence depends on the adapters ligated

**dUTP second strand**[13]

Second-strand synthesis with dUTP; remove 'U's after adaptor ligation and size selection




The second strand is destroyed, thus the cDNA read is always in reverse orientation to the RNA

Adapted from Levine et al Nature Methods

Scripture & Cufflinks allow the user to specify the orientation of the reads.

# The libraries we will work with are strand sepcific

# Summary

- Several methods now exist to build strand sepecific RNA-Seq libraries.
- Quantification methods support strand specific libraries. For example Scripture will compute expression on both strand if desired.

# Overview of the session

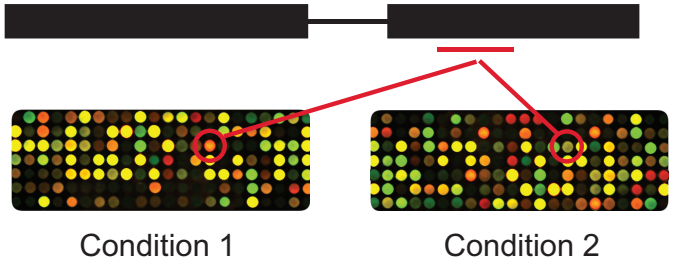
The 3 main computational challenges of sequence analysis for *counting applications:*

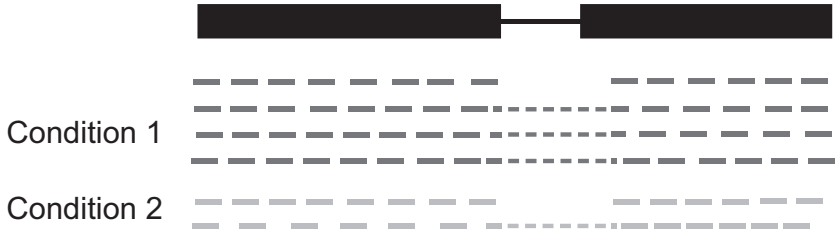
- Read mapping: Placing short reads in the genome
- Reconstruction: Finding the regions that originate the reads
- Quantification:
  - Assigning scores to regions
  - Finding regions that are differentially represented between two or more samples.

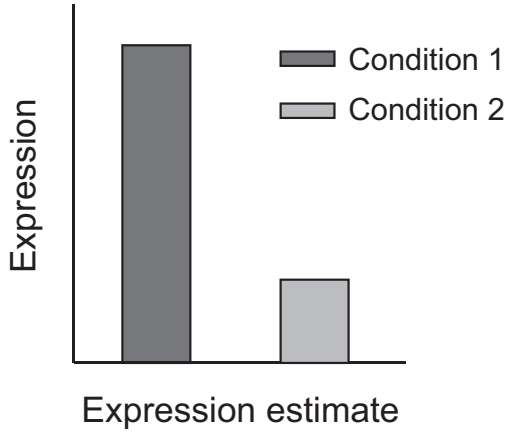
# The problem.

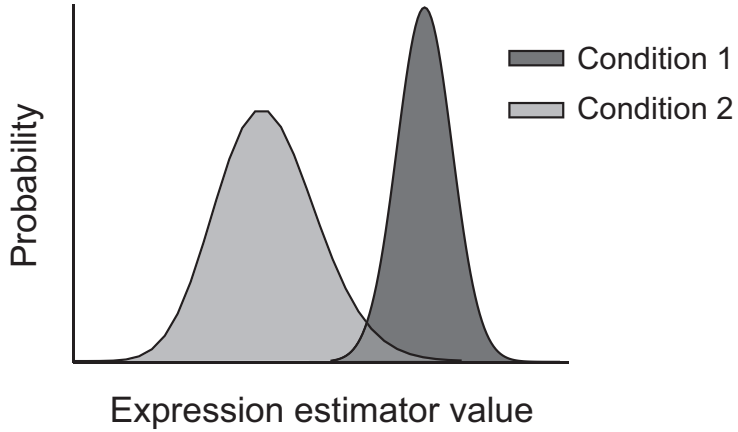
- Finding genes that have different expression between two or more conditions.
- Find gene with isoforms expressed at different levels between two or more conditions.
  - Find differentially used slicing events
  - Find alternatively used transcription start sites
  - Find alternatively used 3’ UTRs

# Differential gene expression using RNA-Seq




•(Normalized) read counts ←→ Hybridization intensity

## Differential analysis strategies

- Use read counts
  - Standard Fisher exact (no preplicates) or $\chi^2$ test (replicates)

| | Condition A | Condition B |
|---|---|---|
| Gene A reads | $n_a$ | $n_b$ |
| Rest of reads | $N_a$ | $N_b$ |

  - Model read counts (Poisson, negative binomial) and test whether models are distinct

# Cufflinks differential issoform ussage

Let a gene G have *n* isoforms and let $p_1, \ldots, p_n$ the estimated fraction of expression of each isoform.

Call this a the isoform expression distribution *P* for G

Given two samples the differential isoform usage amounts to determine whether $H_0$: $P_1 = P_2$ or $H_1$: $P_1$ != $P_2$ are true.

To compare distributions Cufflinks utilizes an information content based metric of how different two distributions are called the Jensen-Shannon divergence:

$$JS(p^1, \ldots, p^m) = H\left(\frac{p^1 + \cdots + p^m}{m}\right) - \frac{\sum_{j=1}^{m} H(p^j)}{m}.$$

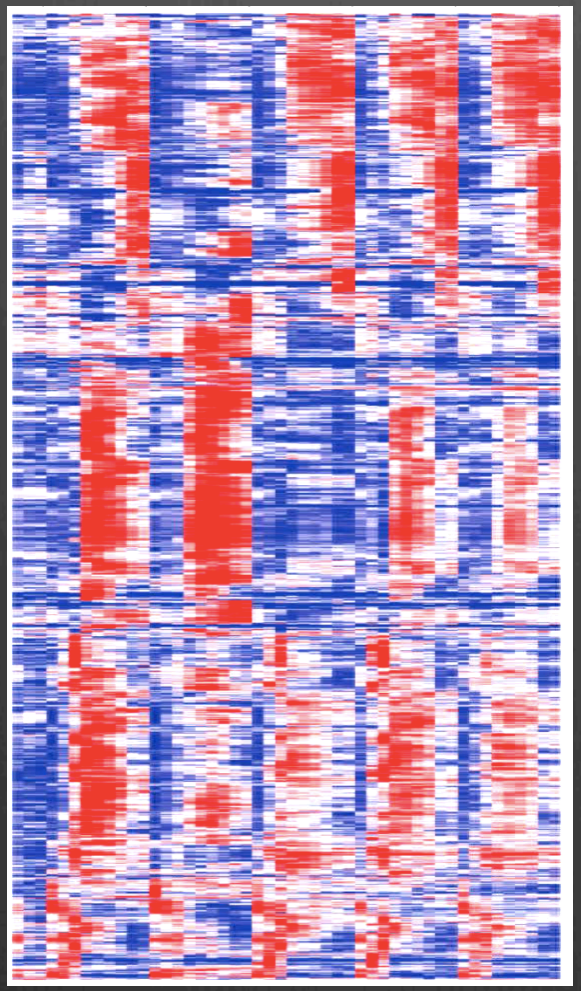
$$H(p) = -\sum_{i=1}^{n} p_i log p_i.$$

The square root of the JS distributes normal.

# RNA-Seq differential expression software

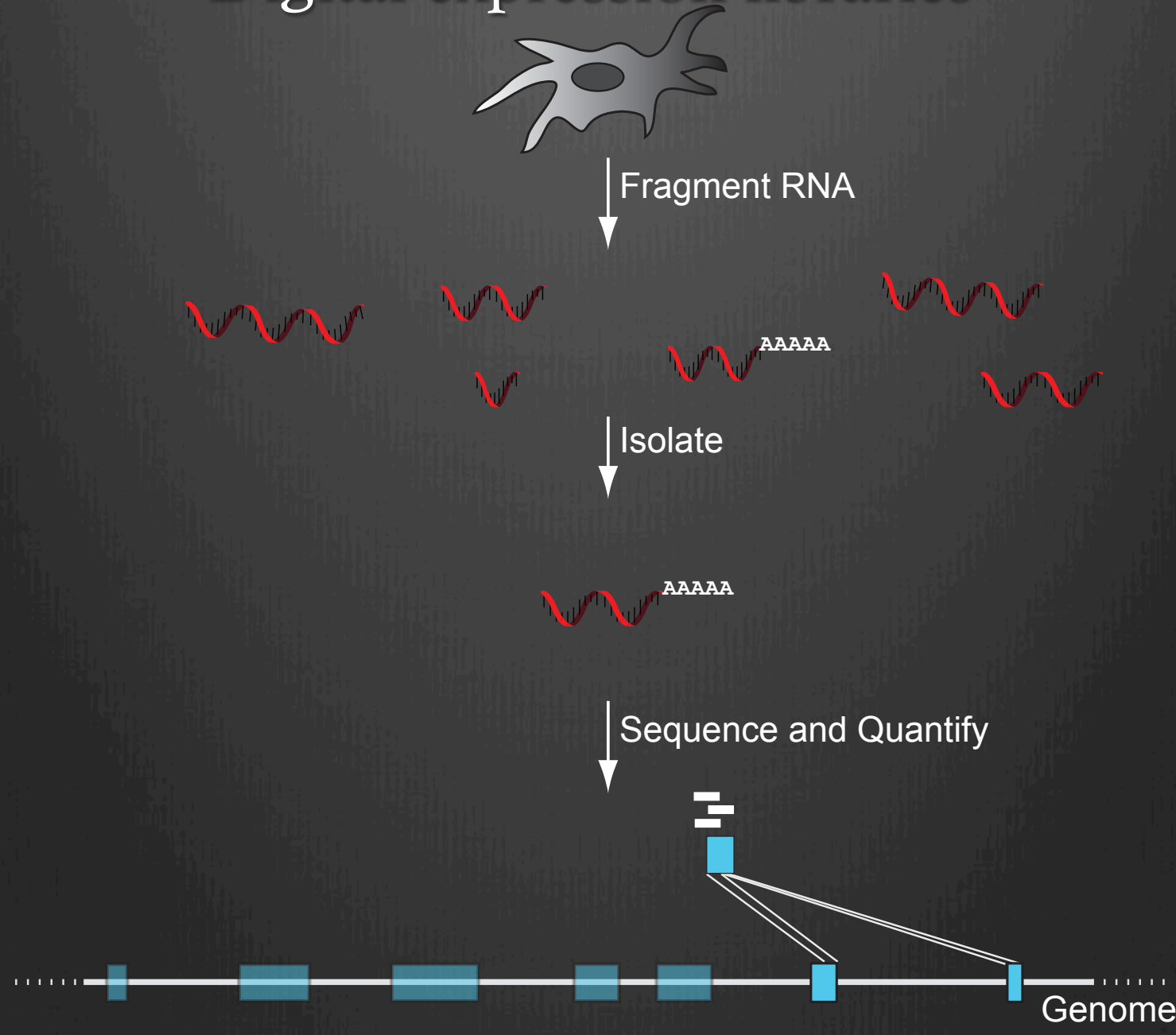
| | Underlying model | Notes |
|---|---|---|
| DegSeq | Normal. Mean and variance estimated from replicates | Works directly from reference transcriptome and read alignment |
| EdgeR | Negative Bionomial | Gene read counts table |
| DESeq | Poisson | Gene read counts table |
| Myrna | Empirical | Sequence reads and reference transcriptome |

# RNA-Seq for traditional gene expression

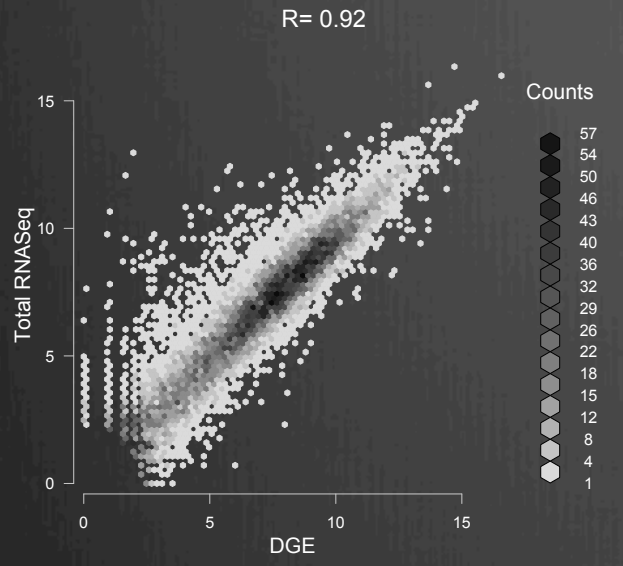


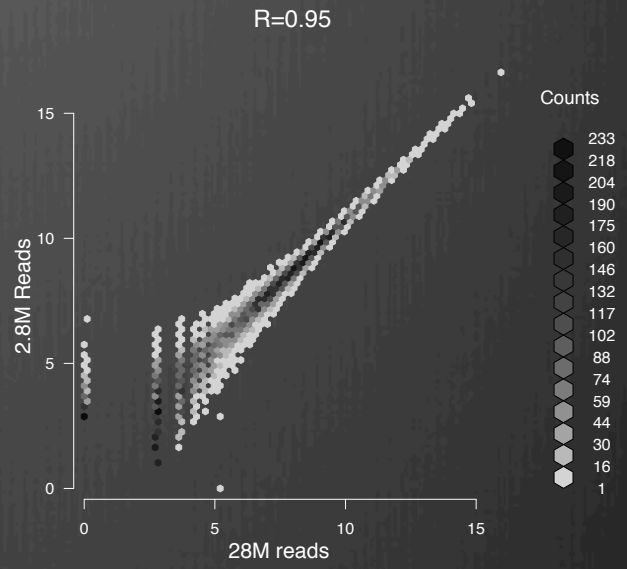
RNASeq is too expensive
for expression assays!

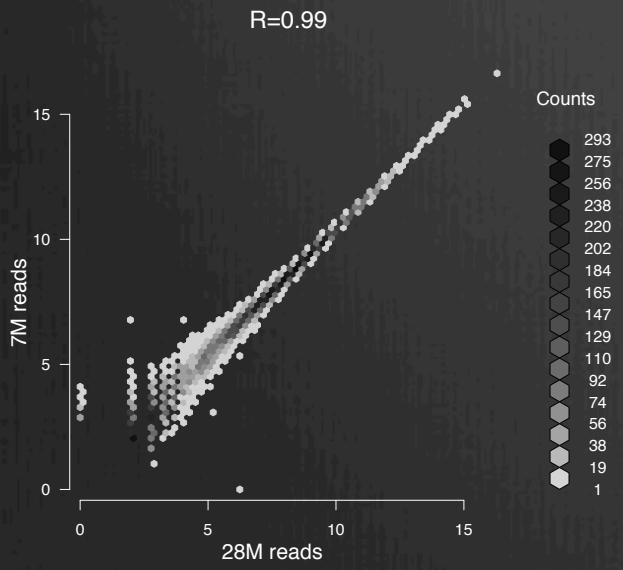
# Digital expression libraries

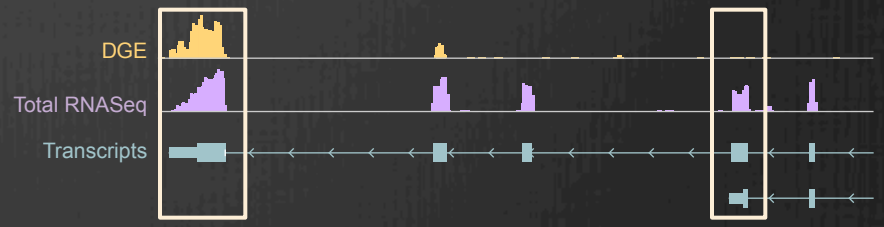


# DGE measure expression very well



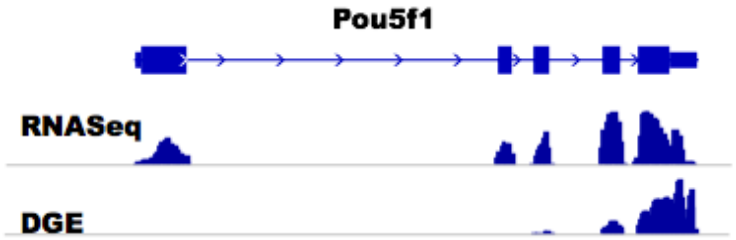

DGE captures 3' alternative ussage

# Digital gene expression

If all you want is the expression level

Easy

- Fragment RNA (heat)
- PolyA select -> RT -> 2nd strand
- Amplify
- Sequence




Cheap

RNASeq requires 100 mill reads.
DGE requires ~6-10 mill reads.

No size bias

**Replicates will be natural and analysis standard**

# Acknowledgements

**Mitchell Guttman**

**Ido Amit**
**Weizmann Institute**


**Mitchell Guttman**

New Contributors:
**Moran Cabili**
**Hayden Metsky**

RNA-Seq:
**Cole Trapnel**
**Manfred Grabher**
**Max Artyomov**
**Sebastian Kadener**

Dendritic Cells:
**Ido Amit**
Nir Yosef
Raktima Raychowdhury

Could not do it without the support of:
**Eric Lander**
**Aviv Regev**