מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

# RNA-Seq Transcript Level Analysis

Dena Leshkowitz,

Course: Introduction to Deep-Sequencing Data Analysis 2016

Bioinformatics Unit, WIS

# Main Topics

- Analysing RNA-Seq data transcript level analysis
- RNA-Seq pipeline:

  Tophat-Cufflinks-Cuffdiff

# SAM Alignment Quality

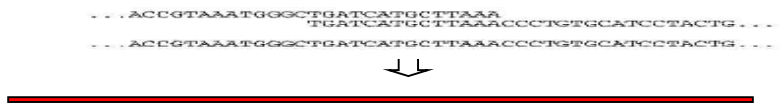
### 1.4 The alignment section: mandatory fields

In the SAM format, each alignment line typically represents the linear alignment of a segment. Each line has 11 mandatory fields. These fields always appear in the same order and must be present, but their values can be '0' or '*' (depending on the field) if the corresponding information is unavailable. The following table gives an overview of the mandatory fields in the SAM format:

| Col | Field | Type | Regexp/Range | Brief description |
|---|---|---|---|---|
| 1 | QNAME | String | `[!-?A-~]{1,254}` | Query template NAME |
| 2 | FLAG | Int | $[0,2^{16}-1]$ | bitwise FLAG |
| 3 | RNAME | String | `\*|[!-()+-<>-~][!-~]*` | Reference sequence NAME |
| 4 | POS | Int | $[0,2^{31}-1]$ | 1-based leftmost mapping POSition |
| 5 | MAPQ | Int | $[0,2^{8}-1]$ | MAPping Quality |
| 6 | CIGAR | String | `\*|([0-9]+[MIDNSHPX=])+` | CIGAR string |
| 7 | RNEXT | String | `\*|=|[!-()+-<>-~][!-~]*` | Ref. name of the mate/next read |
| 8 | PNEXT | Int | $[0,2^{31}-1]$ | Position of the mate/next read |
| 9 | TLEN | Int | $[-2^{31}+1,2^{31}-1]$ | observed Template LENgth |
| 10 | SEQ | String | `\*|[A-Za-z=.]+` | segment SEQuence |
| 11 | QUAL | String | `[!-~]+` | ASCII of Phred-scaled base QUALity+33 |

- **MAPQ: MAPping Quality.**
- **−10 log10 Pr{mapping position is wrong}** rounded to the nearest integer.
- We say an alignment is unique if it has a much higher alignment score than all the other possible alignments. The bigger the gap between the best alignment's score and the second-best alignment's score, the more unique the best alignment, and the higher its mapping quality should be
- **10 is a common threshold that means that there is 1 to ten chance that the read originated form somewhere else.**
- **A value 255 indicates that the mapping quality is not available.**

# From Sequences to Transcriptome Analysis




No Genome or Transcriptome

*De novo* assembly

RNA-Seq

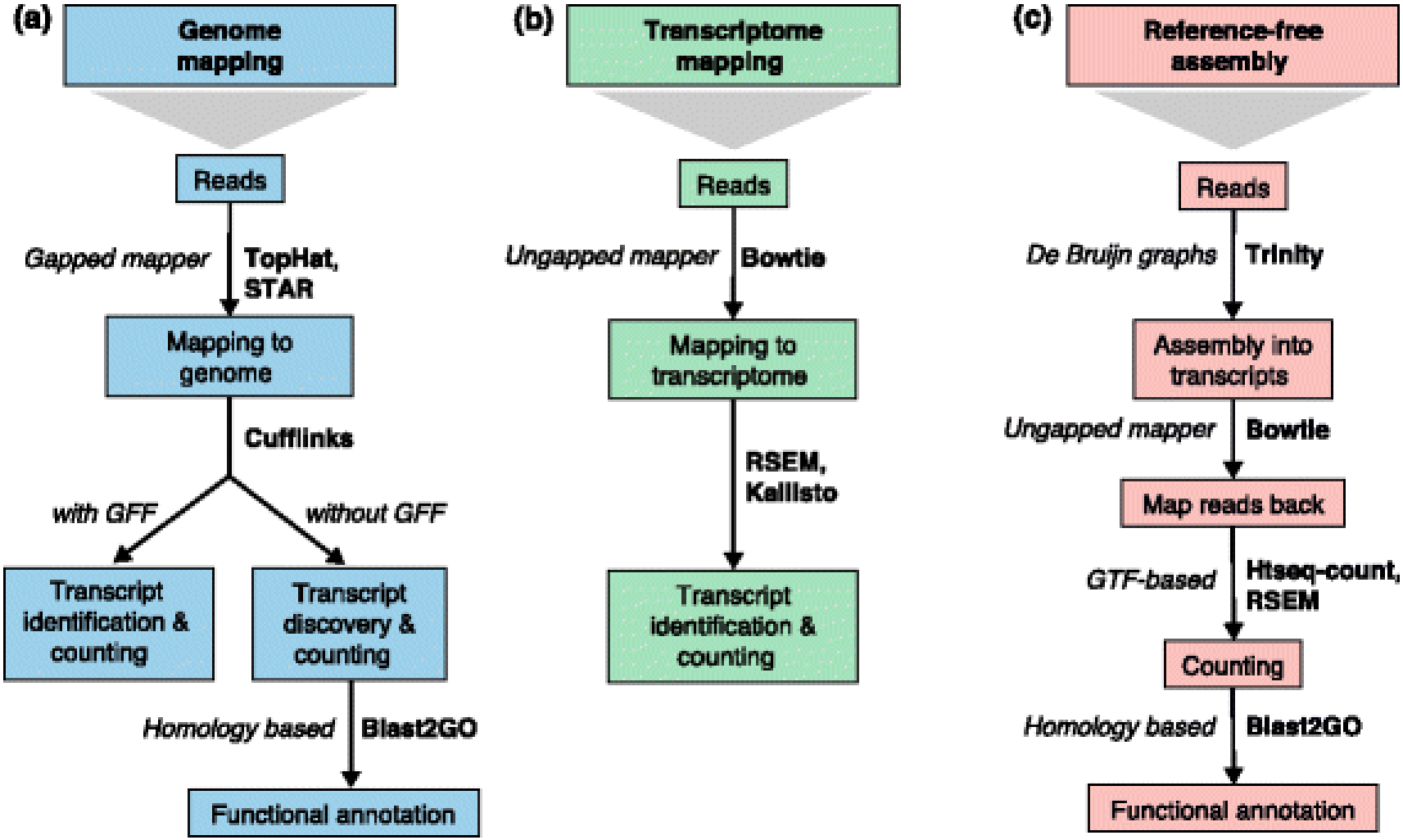
Mapping

Genome

Assemble transcripts

Transcriptome

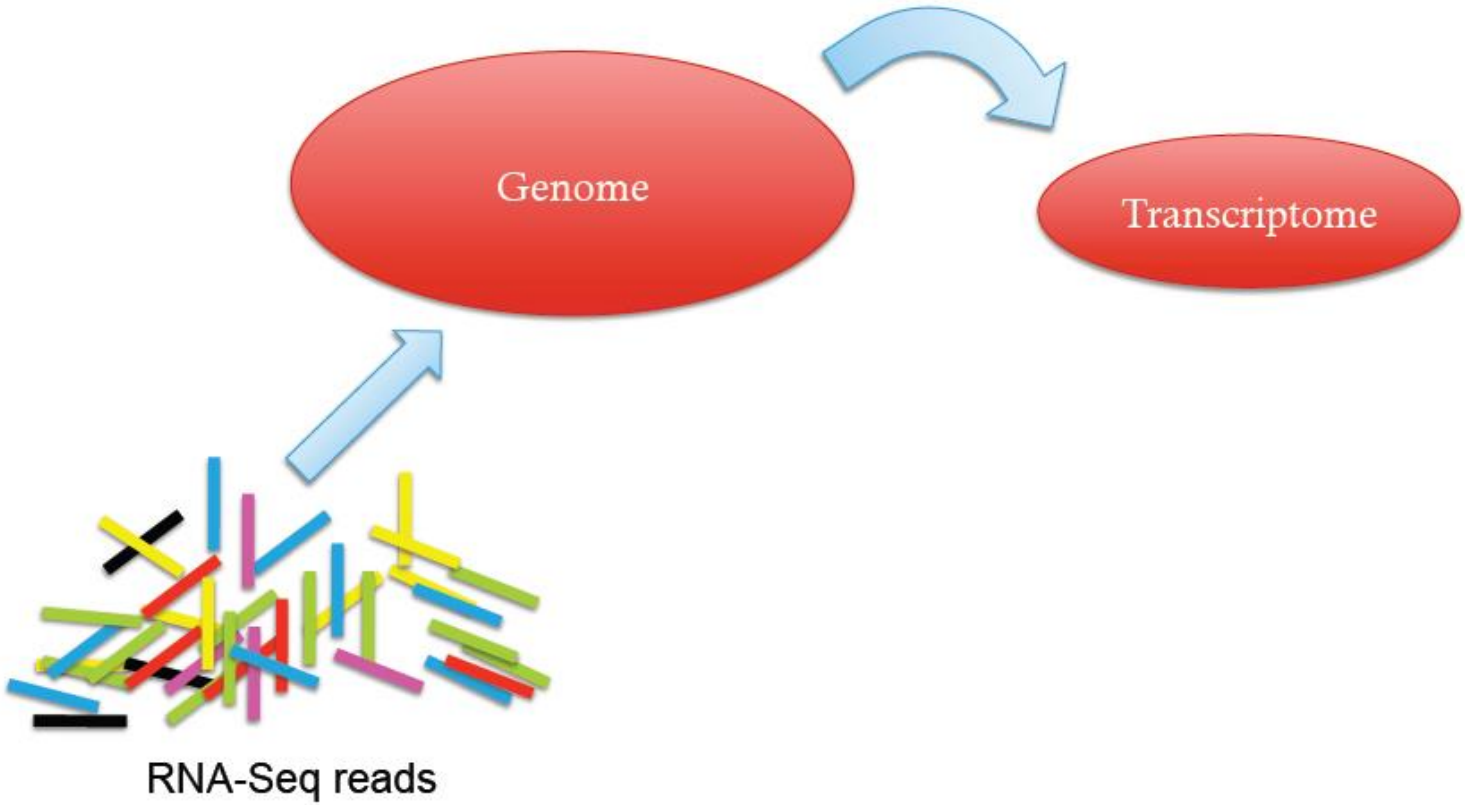
# Read mapping and transcript identification strategies
## Three basic strategies for regular RNA-Seq analysis




*Genome Biology* 2016 **17**:13

# RNA-Seq mapping with TopHat

Goal: **identify** all transcripts and estimate relative amounts from RNA-Seq data



# RNA-Seq Analysis Approaches

- *Align the reads to the genome*
- *Annotate-then-identify*
  - Use the known gene structure database to quantify the genes and transcripts
- *Assemble-then-identify*
  - Allow the aligned reads to identify novel exons and gene structures

# The Tuxedo Tools


*NATURE BIOTECHNOLOGY* | RESEARCH | LETTER

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

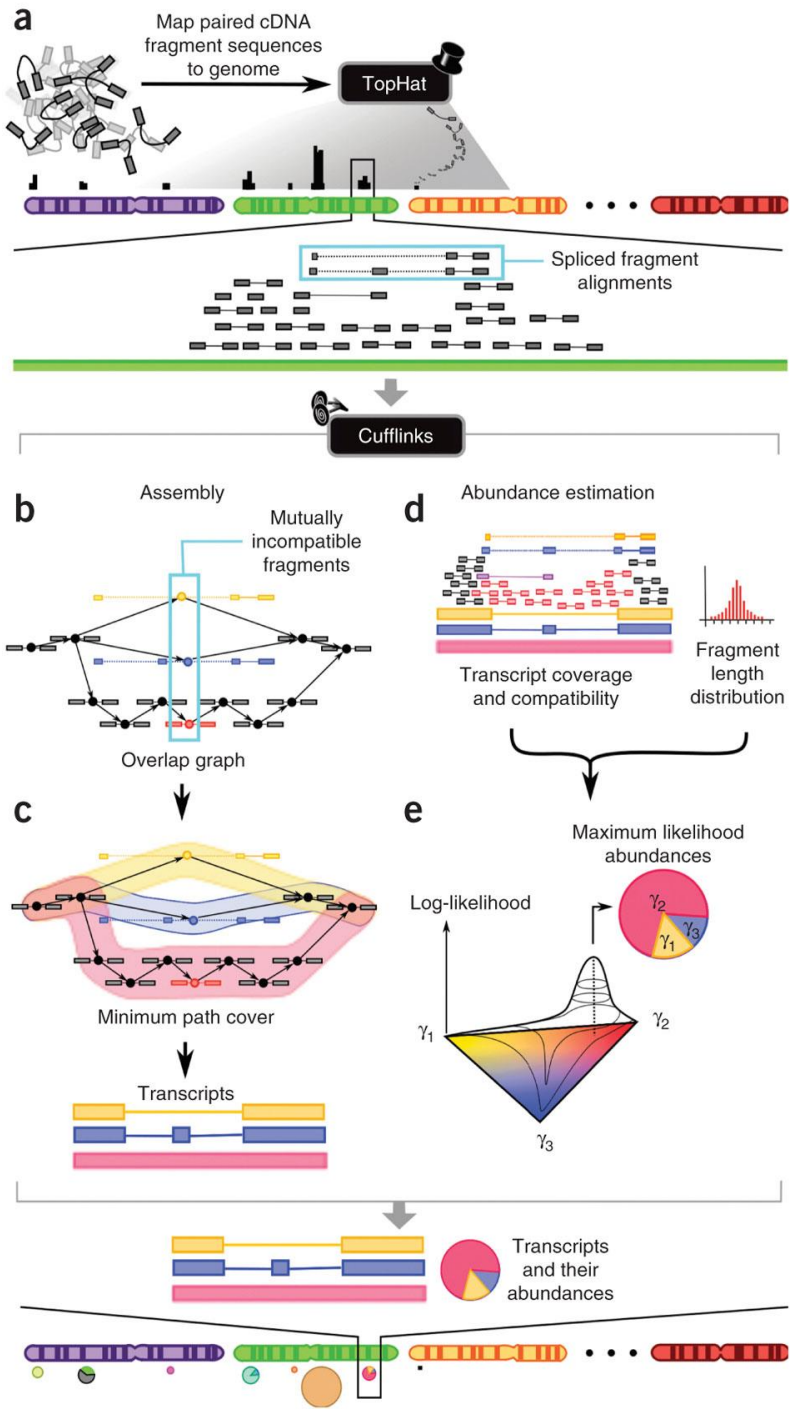
**Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold & Lior Pachter**

**Affiliations** | **Contributions** | **Corresponding author**

*Nature Biotechnology* **28**, 511–515 (2010) | doi:10.1038/nbt.1621
Received 02 February 2010 | Accepted 22 March 2010 | Published online 02 May 2010

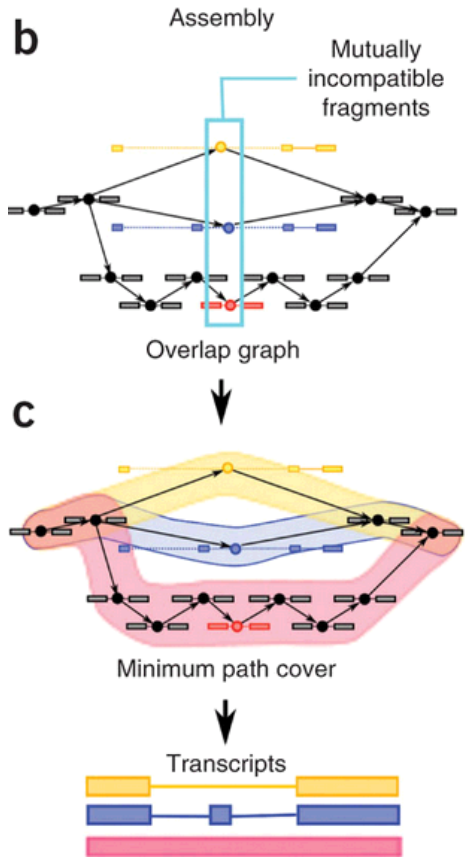
# Cufflinks Detects Novel and Known Transcripts

- "To test Cufflinks, we sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation time series. We detected 13,692 known transcripts and 3,724 previously unannotated ones, 62% of which are supported by independent expression data or by homologous genes in other species."




Nature Biotechnology 28, 511-515 (2010)

# Overview of Cufflinks




- Identify pairs of 'incompatible' fragments that must have originated from distinct spliced mRNA isoforms
- Fragments are connected in an 'overlap graph' when they are compatible and their alignments overlap in the genome
- Find minimum number of transcripts needed to 'explain' all the fragments

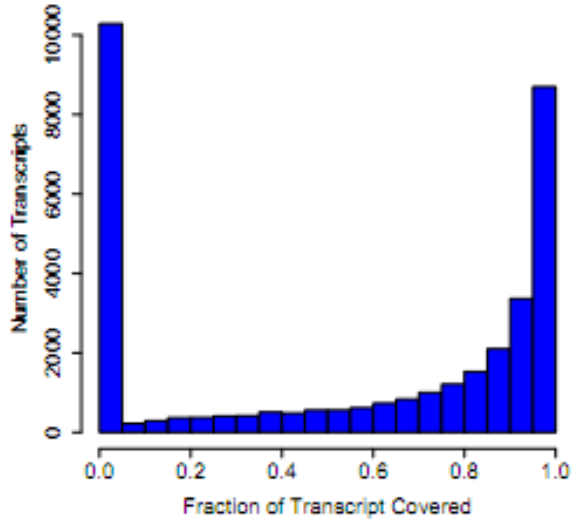

Transcripts and their abundances

*Trapnell et al. Nature Biotechnology* 28, 511-515 (2010)

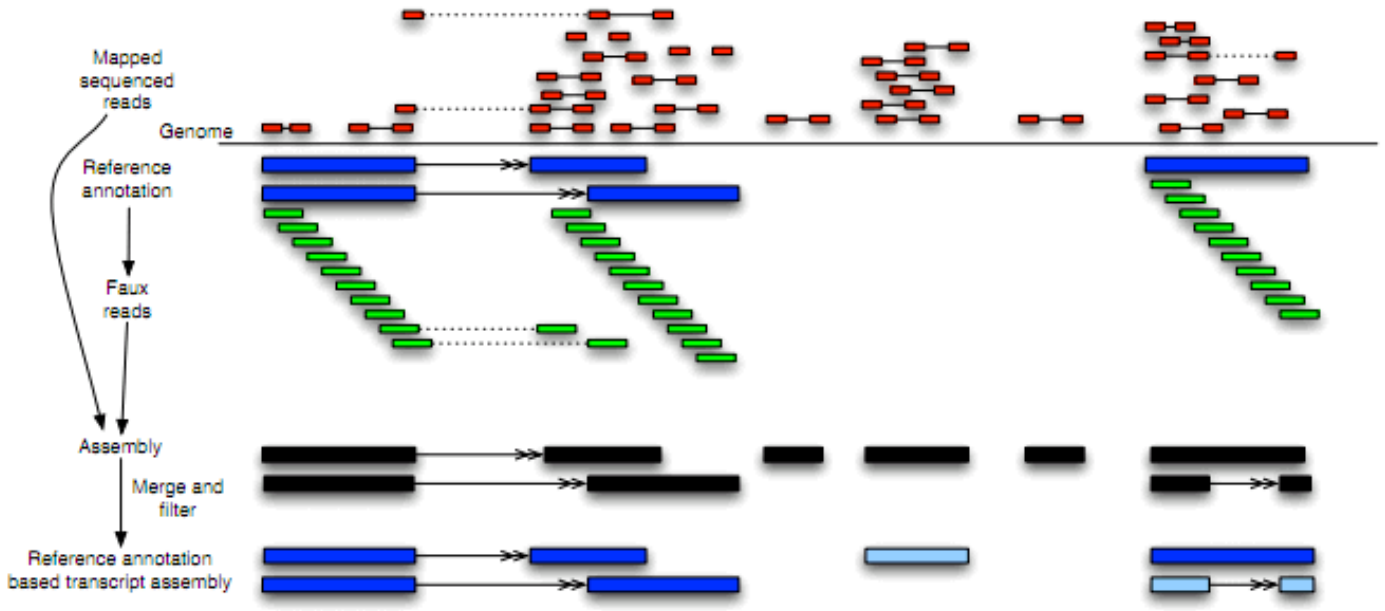
# Cufflinks -RABT

- Transcripts that are expressed in low level are represented by few reads and therefore only partially covered (64%).
- That means that naive assembly methods will fail to construct the majority of the transcripts

Roberts et al. Bioinformatics. 2011 Sep 1;27(17):2325-9.



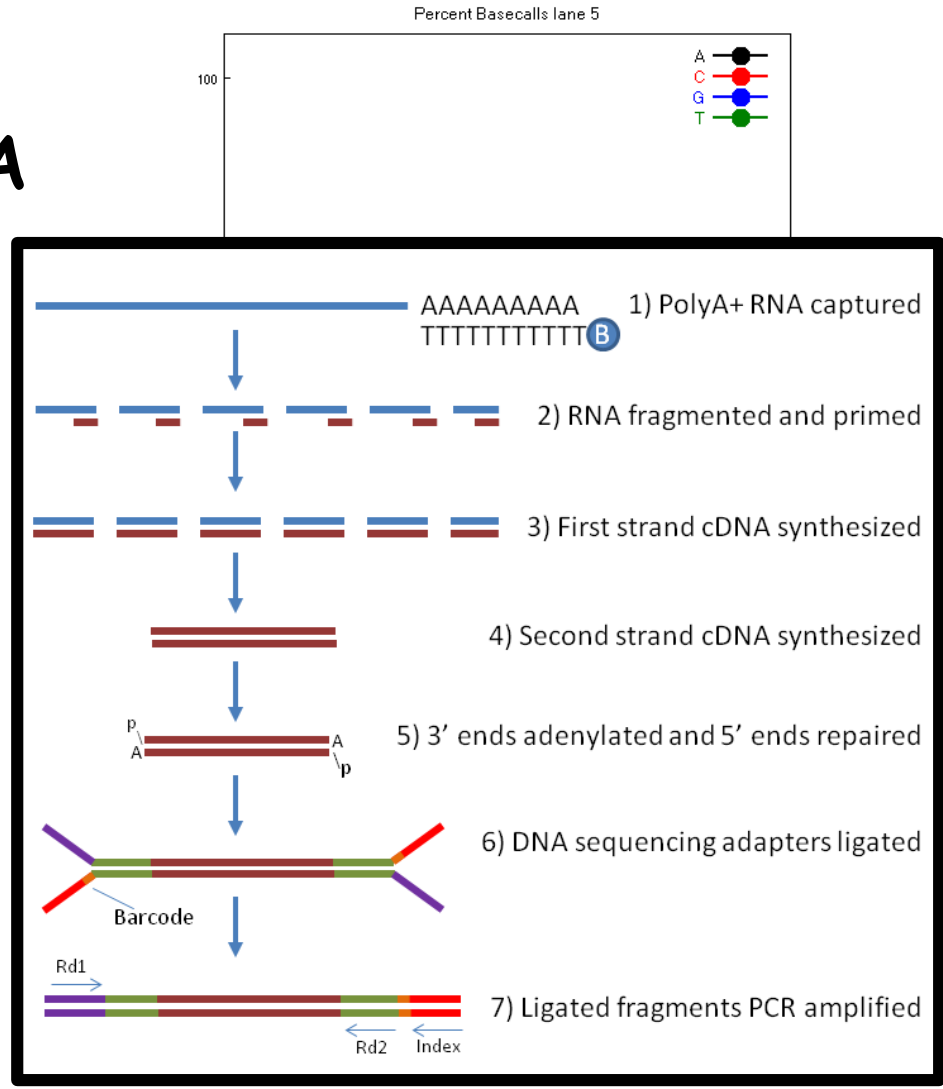
# RABT: Reference Annotation Based Assembler (-g)




Faux reads tiling the transcripts are added to the real reads by cufflinks algorithm in the process of assembly

# Cufflinks Bias Correction

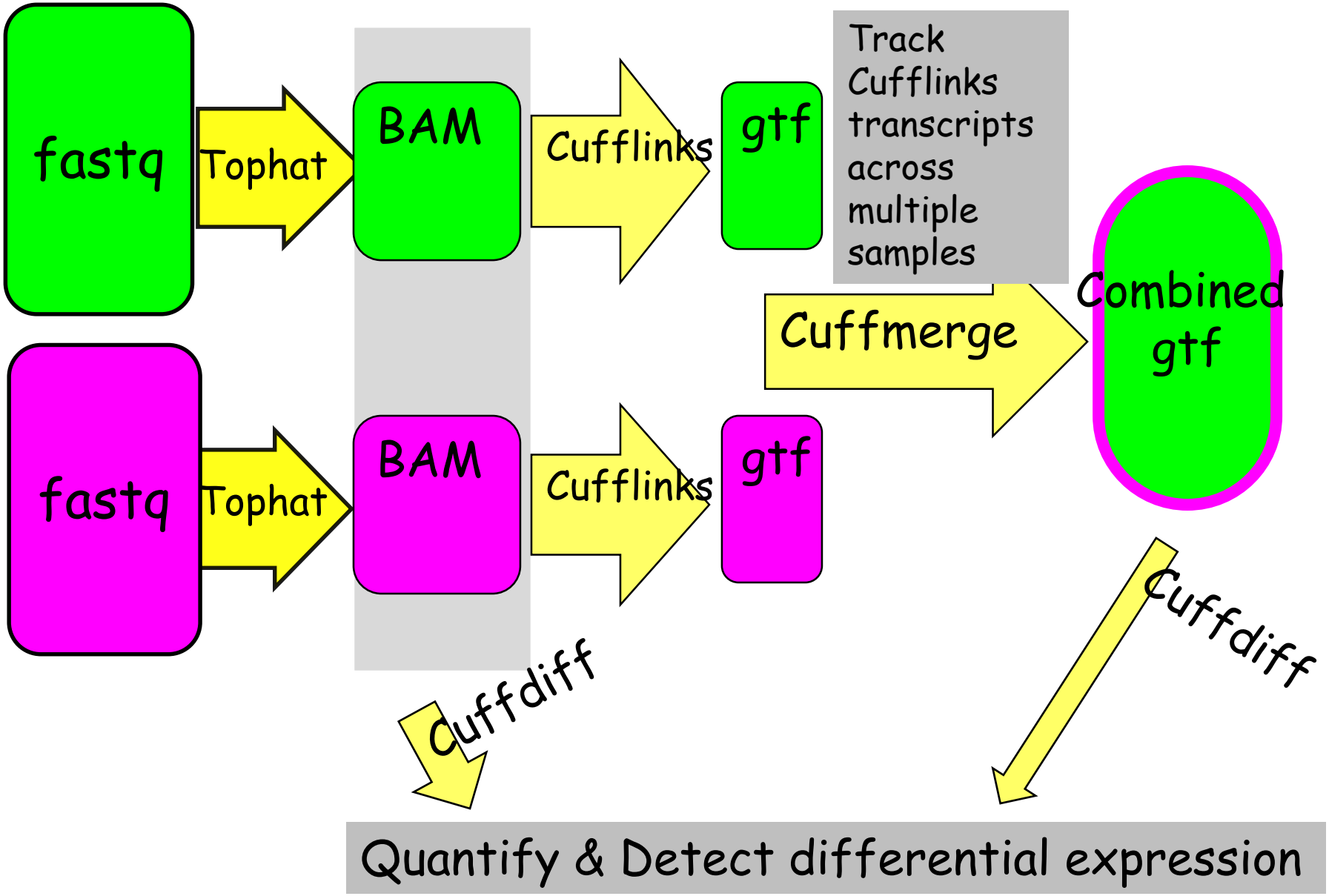
The random priming in the process of cDNA creation causes a positional preferred location for sequencing at the beginning of the transcript



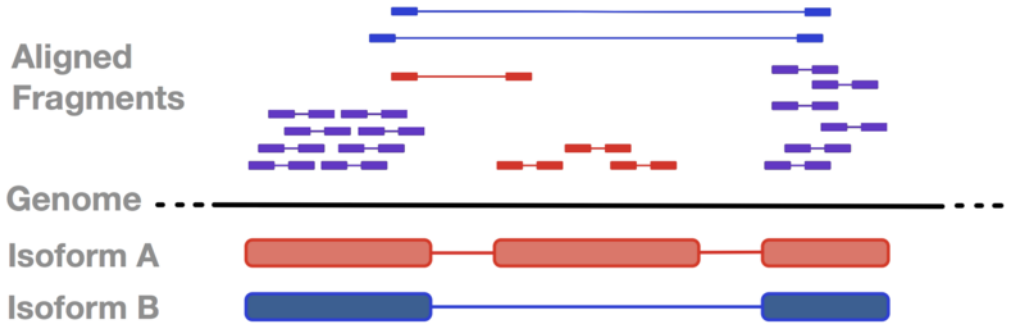
# Cuffcompare - Cuffmerge

- Compare your assembled transcripts to a reference annotation
- Track Cufflinks transcripts across multiple experiments - samples

# Tophat → Cufflinks → Cuffmerge → Cuffdiff



# Align to Transcriptome



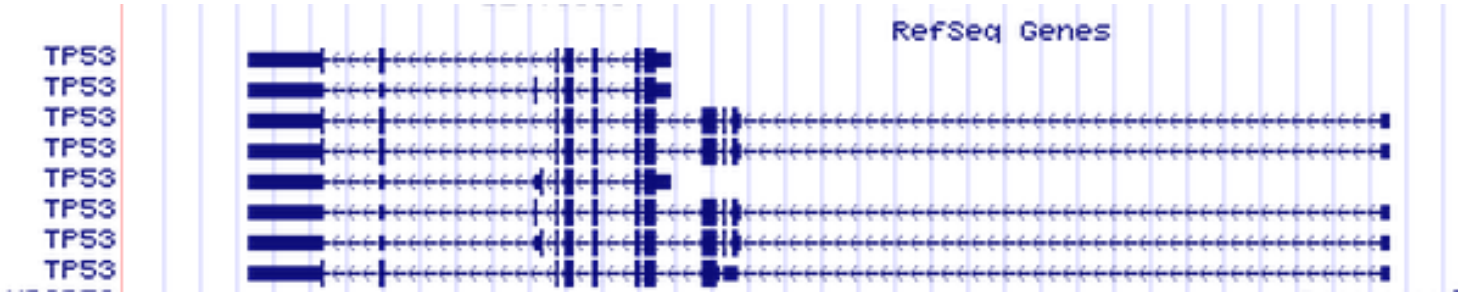

Trapnell et al. Nat Biotechnol. 2010 May;28(5):511-5.

From which transcript did the purple reads originate?

# Align to Transcriptome Quantification Problem

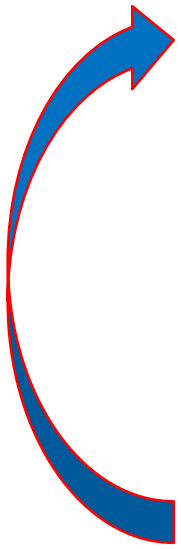
- We encounter the same problem when we align to a trasncriptome
- Counting the number of sequences that map uniquely to transcripts results in false estimates of alternatively spliced transcripts

Human p53 locus



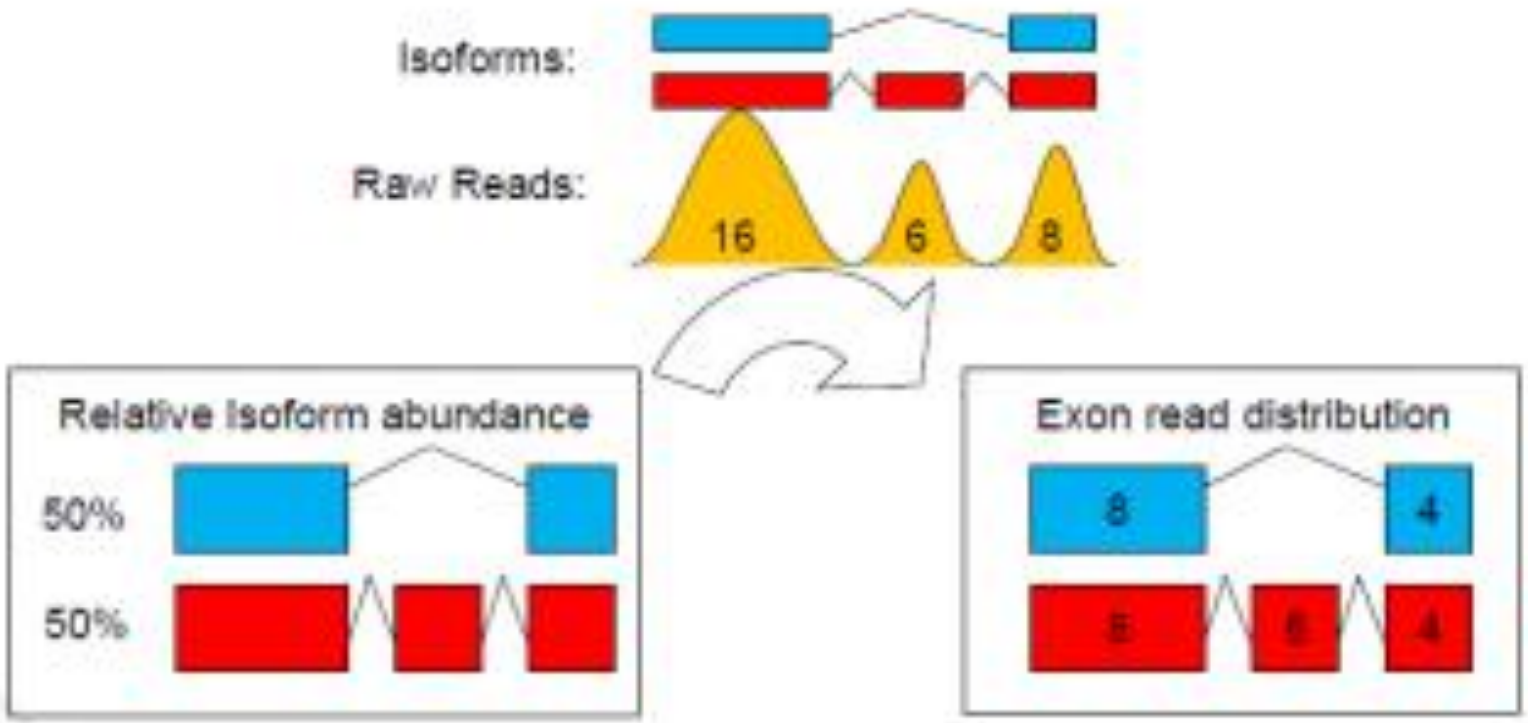

- Computational challenge- use reads that map ambiguously between isoforms and genes (EM algorithm)

# Isoform Expression Quantification Expectation Maximization Algorithm

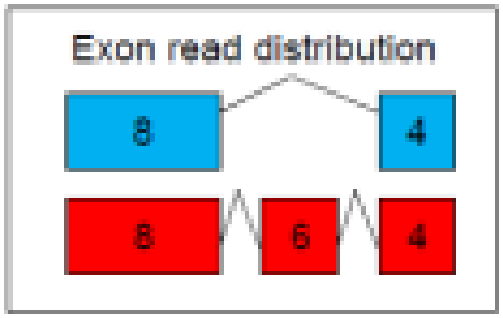


- Step 1- Assume isoforms are equally abundant
- Step 2 - Distribute the reads to the isoforms based on the abundance
- Step 3 - Recalculate the isoforms abundance based on the reads counts and isoforms length
- Step 4- If abundance has changed go back to step 2 otherwise stop

# 1st step E/M algorithm



# Calculating Abundance after 1st EM Cycle




| Total reads | Length |
|---|---|
| 12 | 300 |
| 18 | 400 |

Exon length (bp) 200 100 100

The red transcript abundance after the first cycle:

$$p_{red} = \frac{counts_{red} / length_{red}}{counts_{red} / length_{red} + counts_{blue} / length_{blue}}$$

$p_{red}$ = 18/400 / (12/300 + 18/400) = 0.53
$p_{blue}$ = 12/300 / (12/300 + 18/400) = 0.47

# EM Calculation: 100 Iterations

| | starting relative proportion (p) | read counts | New proportion after iteration (p) | | starting relative proportion (p) | read counts | New proportion after iteration (p) | | Iteration # |
|---|---|---|---|---|---|---|---|---|---|
| Blue | 0.5 | 12 | 0.470588 | Red | 0.5 | 18 | 0.529412 | | 1 |
| | | 11.29412 | 0.445993 | | | 18.70588 | 0.554007 | | 2 |
| | | 10.70383 | 0.425161 | | | 19.29617 | 0.574839 | | 3 |
| | | 10.20386 | 0.407324 | | | 19.79614 | 0.592676 | | 4 |
| | | 9.775778 | 0.39191 | | | 20.22422 | 0.60809 | | 5 |
| | | 9.405837 | 0.378482 | | | 20.59416 | 0.621518 | | 6 |
| | | 9.083574 | 0.366704 | | | 20.91643 | 0.633296 | | 7 |
| | | 8.800885 | 0.356308 | | | 21.19911 | 0.643692 | | 8 |
| | | 8.551391 | 0.347084 | | | 21.44861 | 0.652916 | | 9 |
| | | 8.330004 | 0.338859 | | | 21.67 | 0.661141 | | 10 |
| | | 6.00743 | 0.25029 | | | 23.99257 | 0.74971 | | 94 |
| | | 6.006965 | 0.250272 | | | 23.99303 | 0.749728 | | 95 |
| | | 6.006529 | 0.250255 | | | 23.99347 | 0.749745 | | 96 |
| | | 6.006121 | 0.250239 | | | 23.99388 | 0.749761 | | 97 |
| | | 6.005738 | 0.250224 | | | 23.99426 | 0.749776 | | 98 |
| | | 6.005379 | 0.25021 | | | 23.99462 | 0.74979 | | 99 |
| | | 6.005042 | **0.2502** | | | 23.99496 | **0.7498** | | 100 |

Blue 25%

Red 75%

# Normalized Expression Values

**Fragments (Reads) Per Kilobase of exon per Million mapped fragments**

**Nat Methods. 2008, Mapping and quantifying mammalian transcriptomes by RNA-Seq. Mortazavi A et al.**

$$FPKMi = 10^6 \times 10^3 \times \frac{Ci}{NLi}$$

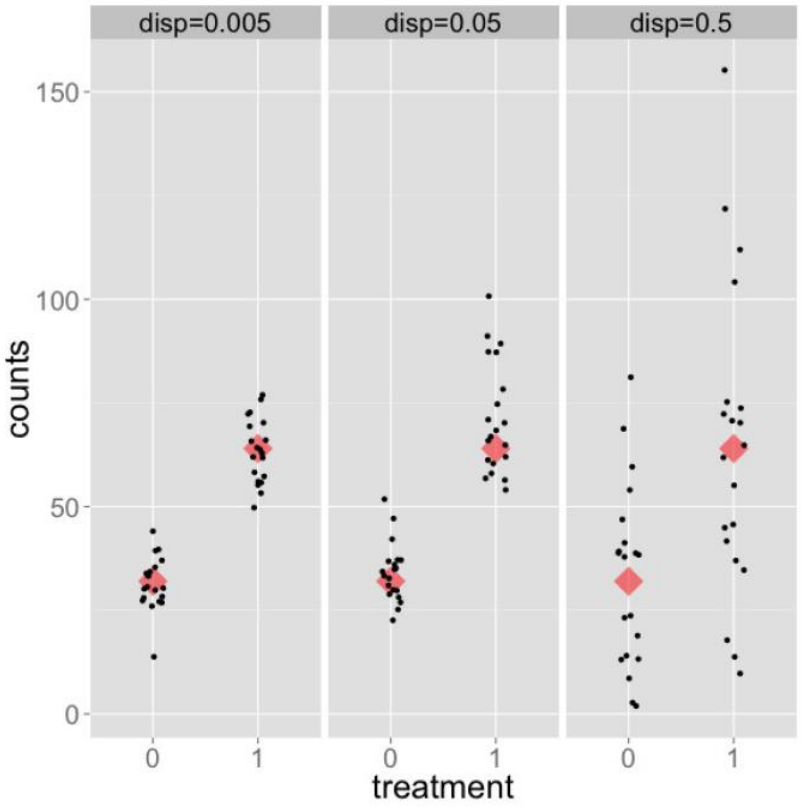
C= The number of fragments mapped onto the transcript exons

N= Total number of (mapped) fragments in the experiment

L= The length of the transcript (sum of exons)

# Determining Differentially Expressed Genes and Transcripts

Discover transcripts showing different average expression levels across two groups




http://www.molgen.mpg.de/1242892/rnaseq.pdf

The statistical model for finding differential expressed transcripts or genes depends on whether we have biological replicates. The advantage of having many replicates allows to learn about the biological variation within the conditions tested.

# Model Used Negative Binomial

- Count data follows a Poisson distribution. Poisson assumes that the mean equals the variance . However, in RNA-Seq data, genes with larger mean cou er variance, which is de overdispersion probl
- **Negative binomial m DESeq, cuffdiff) ac term in the model**

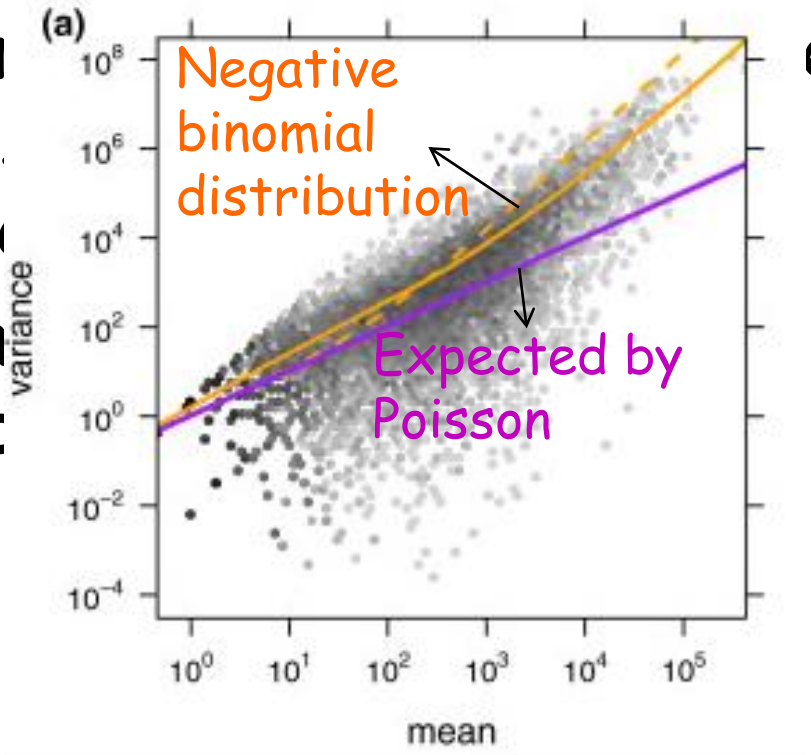
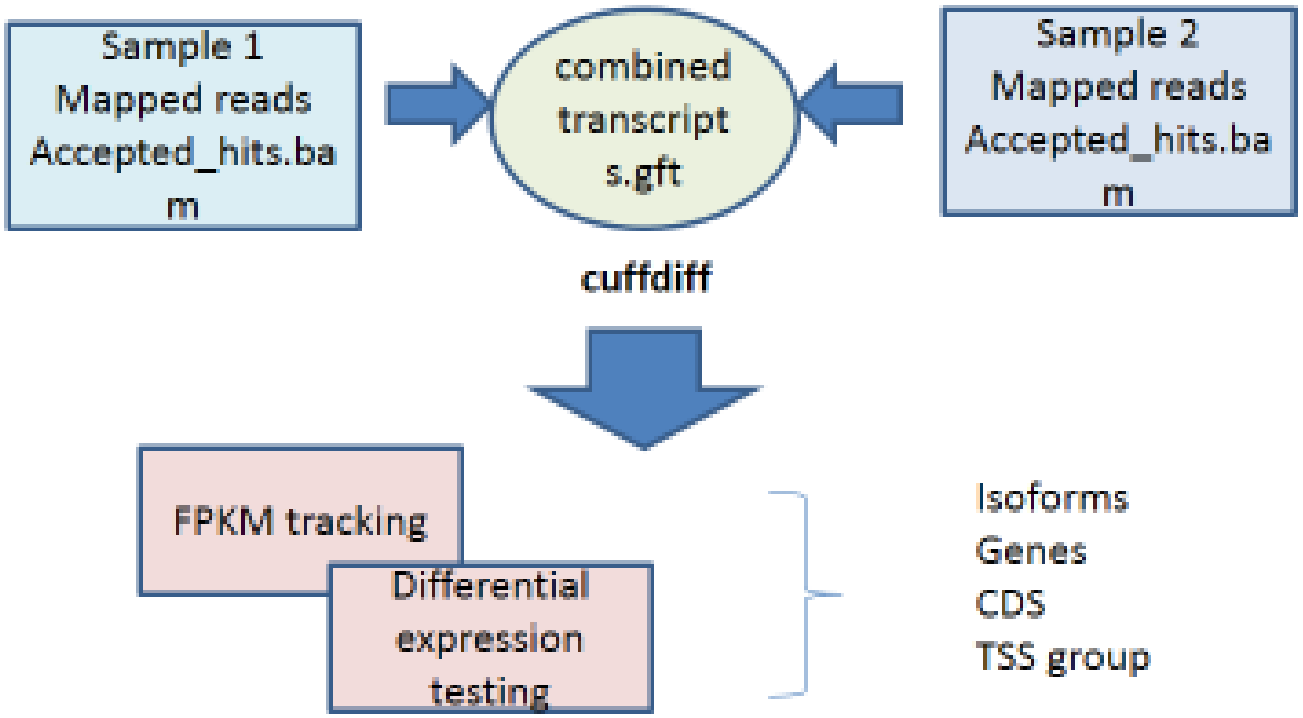



Fig. 1 from Anders & Huber, 2010: Dependence

# Cuffdiff

- Quantifies transcripts and finds significant changes in transcript expression



# The Benefit of Longer and PE Reads

- Reads mapping to junctions
  - With longer reads we will have more of these reads

- Paired end reads

  Knowing both ends of a fragment it is easier to determine from which isoform this fragment originated

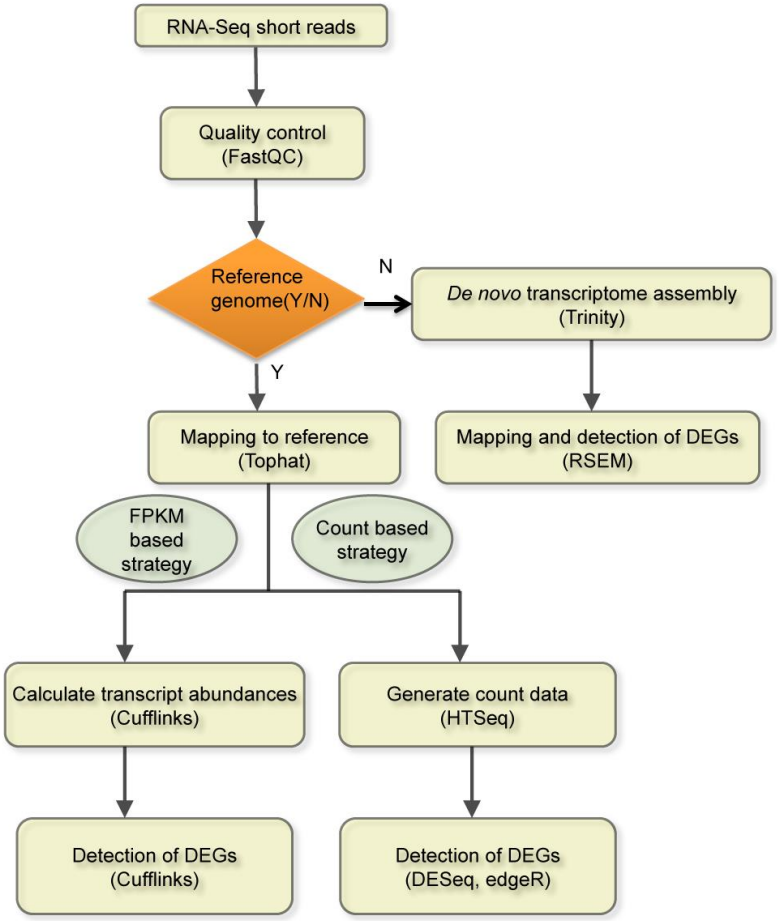
# Experimental Design

## Mammalian tissue

## Liu Y. et al., 2014; ENCODE 2011 RNA-Seq

| | | |
|---|---|---|
| Differential gene expression profiling | 10-25M | 50 base single-end |
| Alternative splicing | 50-100M | 100 base paired-end |
| Allele specific expression | 50-100M | 100 base paired-end |
| De novo assembly | >100M | 100 base paired-end |

# RNA-Seq pipelines




PLoS ONE 2014 9(8): e103207.

# Assessment of transcript reconstruction methods for RNA-seq

Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, The RGASP Consortium, Tim J Hubbard, Roderic Guigó, Jennifer Harrow & Paul Bertone

Affiliations | Contributions

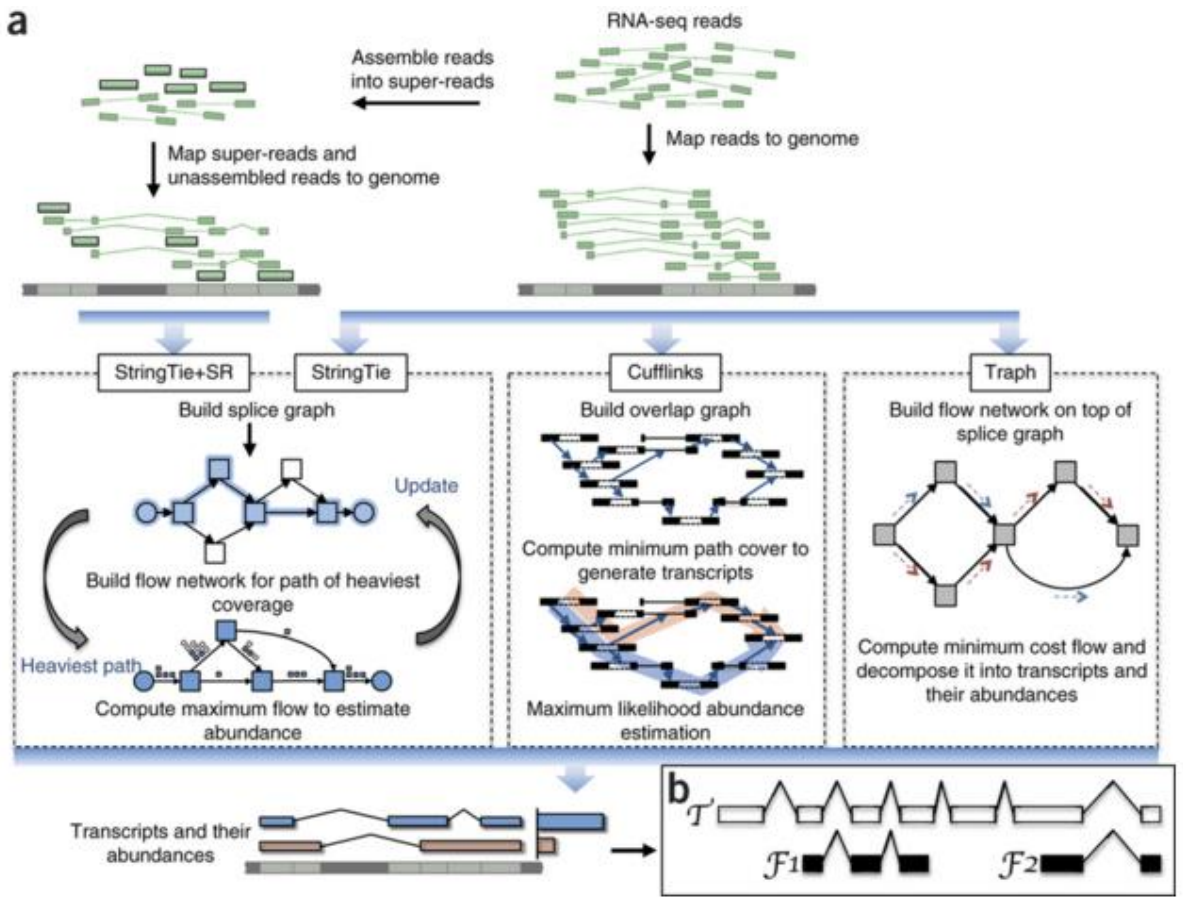
*Nature Methods* **10**, 1177–1184 (2013) | doi:10.1038/nmeth.2714

Received 31 March 2013 | Accepted 23 September 2013 | Published online 03 November 2013

Results were evaluated from methods based on genome alignments (Augustus[9], Cufflinks[3], Exonerate[10], GSTRUCT, iReckon[2], mGene[11], mTim, NextGeneid[12], SLIDE[4], Transomics, Trembly and Tromer[13]) as well as *de novo* assembly (Oases[5] and Velvet[14]).
Programs were run without genome annotation, aside from iReckon and SLIDE


No method achieved even 60% accuracy for transcript reconstruction in human

# New Tools




Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. (2015) **StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.** *Nat Biotechnol* [Epub ahead of print]. [article]

# the next-generation "Tuxedo" tools

**Bowtie2**

Fast alignment

**HISAT**

Spliced alignment

**Ballgown**

- Differential expression

**StringTie**

- Transcript assembly
- Quantitation

Steven Salzberg - Transcriptome Assembly Computational Challenges of Next Generation Sequence Data
https://www.youtube.com/watch?v=2qGiw4MRK3c

# RNA-Seq Exercises

- Observe the Tuxedo outputs
- Use Genome Browser IGV to analyse outputs

THANKS

Questions???