Galaxy Workshop

Jeremy Goecks
The Galaxy Team, Emory University

Agenda

Galaxy Basics

Tophat (RNA-seq read mapping)

- algorithm discussion
- tool description, inputs, and output visualization

Cufflinks (Isoform assembly)

- algorithm discussion
- run Cufflinks, visualize output
- change parameters and rerun

Filter tools

- to remove assembly artifacts

Workflows

- creating and editing a workflow
- running a workflow on another sample

Cuffmerge

- algorithm discussion
- run CuffMerge

Cuffdiff

- algorithm discussion
- run Cuffdiff

Galaxy 101

Galaxy Basics

Create an account

Shared Data --> Published Histories

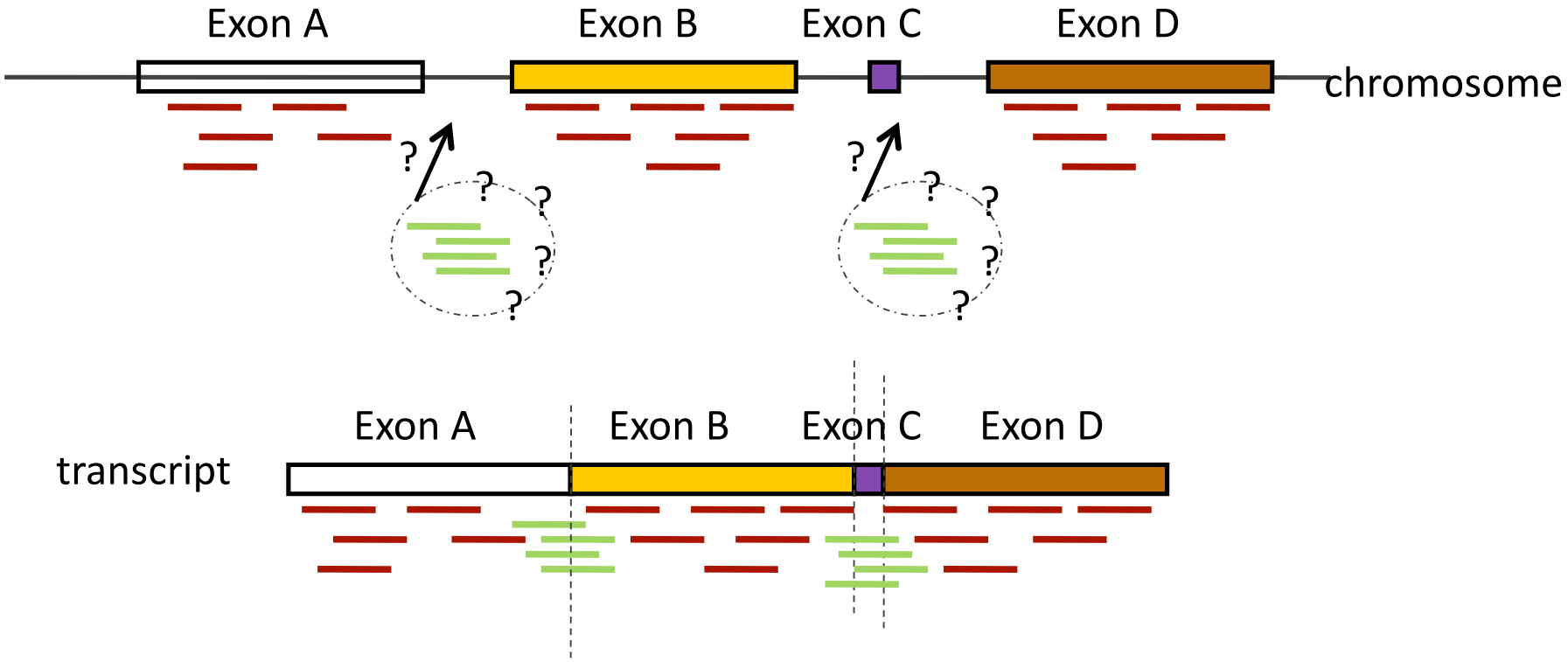
- copy histories into your account

Use UCSC to fetch hg19 gene annotation in BED format for chromosome 22

Read Alignment

Challenges? Options?

Reads in RNA-seq



Sequence **A**lignment/**M**ap (SAM/BAM)

Popular format for storing mapped reads

SAM is text, BAM is binary

Galaxy makes it easy to convert between SAM and BAM (and convert to interval)

```
ERR030889.57680506 256 chr11 1548912 1 75M * 0 0
GTTGCAGTGAGCTGAGATCGCACCACTGCACTCCAGCCTGGGCAACAGAGTG
AGACTCCGTCTCAAGAAAAATAA
HHHHHFFF*FFFFFFHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHGH
HHIHHHHHIFGHIIHFHHEBEBC
NM:i:2 NH:i:3 CC:Z:chr22 CP:i:29194605 HI:i:0
```

RNA-seq mapping

BIOINFORMATICS **ORIGINAL PAPER** *Vol. 25 no. 9 2009, pages 1105–1111*
doi:10.1093/bioinformatics/btp120

Sequence analysis

TopHat: discovering splice junctions with RNA-Seq

Cole Trapnell[1,*], Lior Pachter[2] and Steven L. Salzberg[1]

[1]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742 and [2]Department of Mathematics, University of California, Berkeley, CA 94720, USA

Received on October 23, 2008; revised on February 24, 2009; accepted on February 26, 2009

Advance Access publication March 16, 2009

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: A new protocol for sequencing the messenger RNA in a cell, known as RNA-Seq, generates millions of short sequence fragments in a single run. These fragments, or 'reads', can be used to measure levels of gene expression and to identify novel splice variants of genes. However, current software for aligning RNA-Seq data to a genome relies on known splice junctions and cannot identify novel ones. TopHat is an efficient read-mapping algorithm designed to align reads from an RNA-Seq experiment to a reference genome without relying on known splice sites.

Results: We mapped the RNA-Seq reads from a recent mammalian RNA-Seq experiment and recovered more than 72% of the splice junctions reported by the annotation-based software from that study, along with nearly 20 000 previously unreported junctions. The TopHat pipeline is much faster than previous systems, mapping nearly 2.2 million reads per CPU hour, which is sufficient to process an entire RNA-Seq experiment in less than a day on a standard desktop computer. We describe several challenges unique to *ab initio* splice site discovery from RNA-Seq reads that will require further algorithm development.

measurements of expression at comparable cost (Marioni *et al.*, 2008).

The major drawback of RNA-Seq over conventional EST sequencing is that the sequences themselves are much shorter, typically 25–50 nt versus several hundred nucleotides with older technologies. One of the critical steps in an RNA-Seq experiment is that of mapping the NGS 'reads' to the reference transcriptome. However, because the transcriptomes are incomplete even for well-studied species such as human and mouse, RNA-Seq analyses are forced to map to the reference genome as a proxy for the transcriptome. Mapping to the genome achieves two major objectives of RNA-Seq experiments:

(1) Identification of novel transcripts from the locations of regions covered in the mapping.

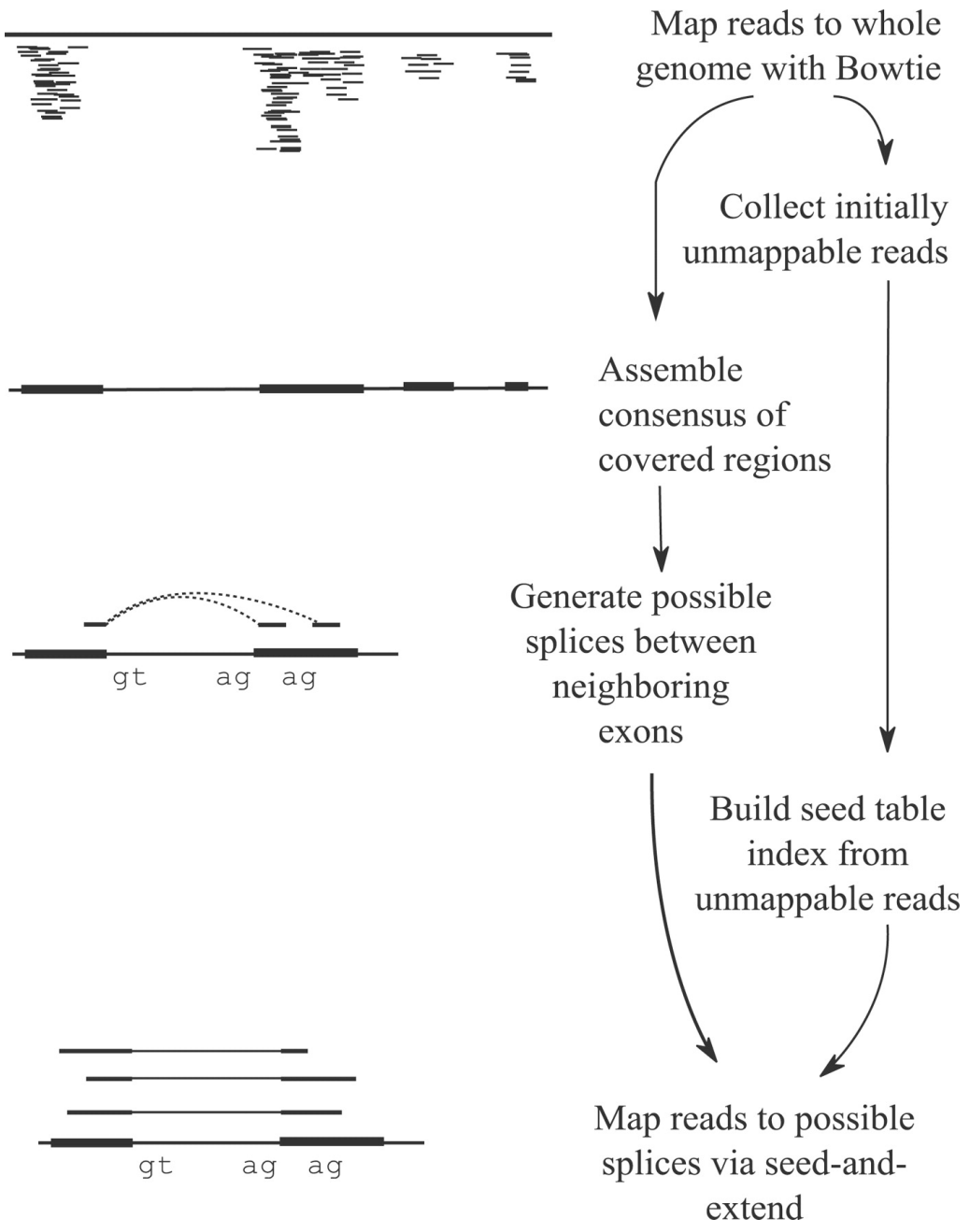
(2) Estimation of the abundance of the transcripts from their depth of coverage in the mapping.

Because RNA-Seq reads are short, the first task is challenging. Current mapping strategies (e.g. Cloonan *et al.*, 2008; Marioni *et al.*,

TopHat: island / cluster

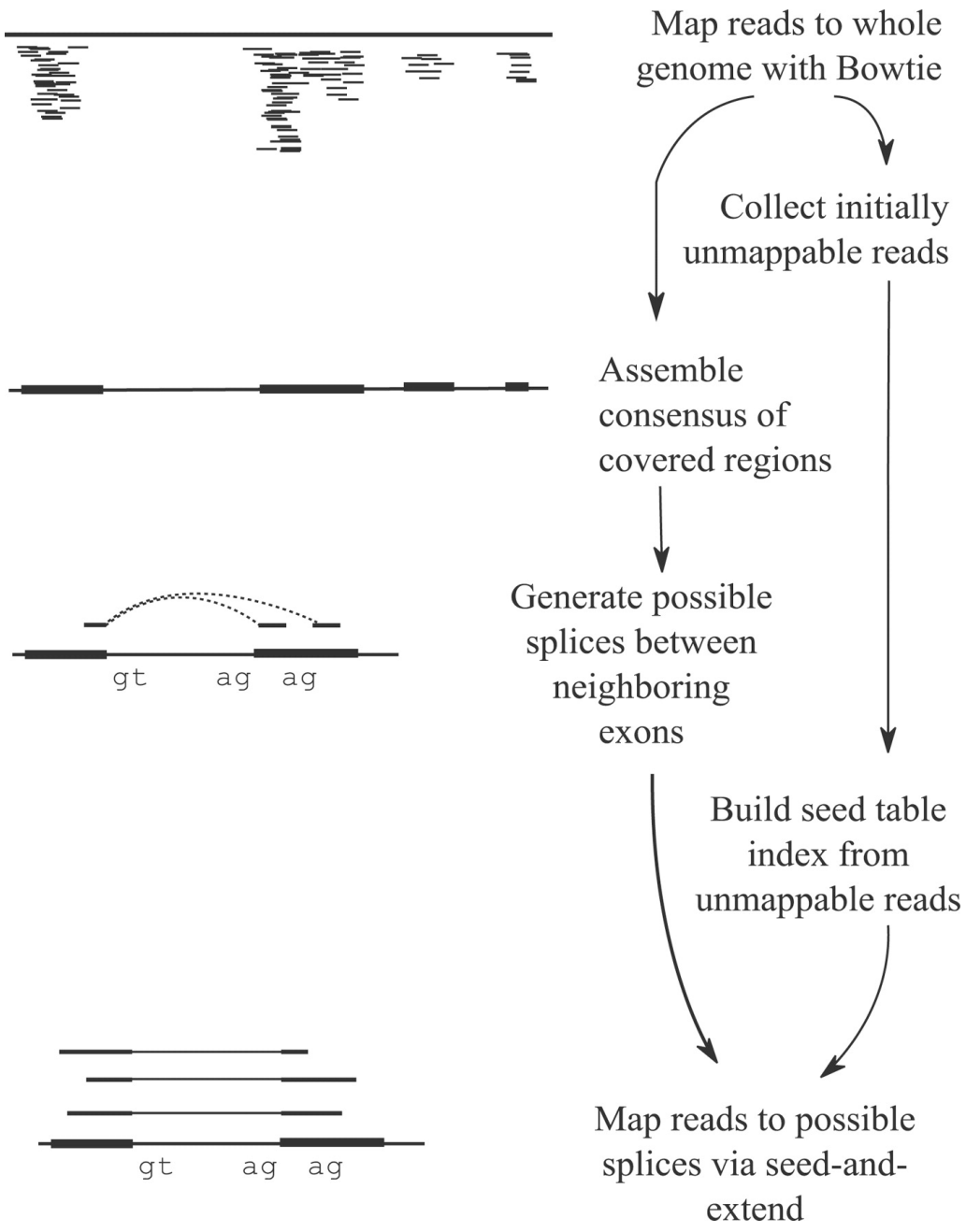
1. Map reads to genome/ transcriptome with no gaps to generate “exonic” islands

2. Map initially unmappable reads across islands, creating gapped, mapped reads



TopHat: island / cluster

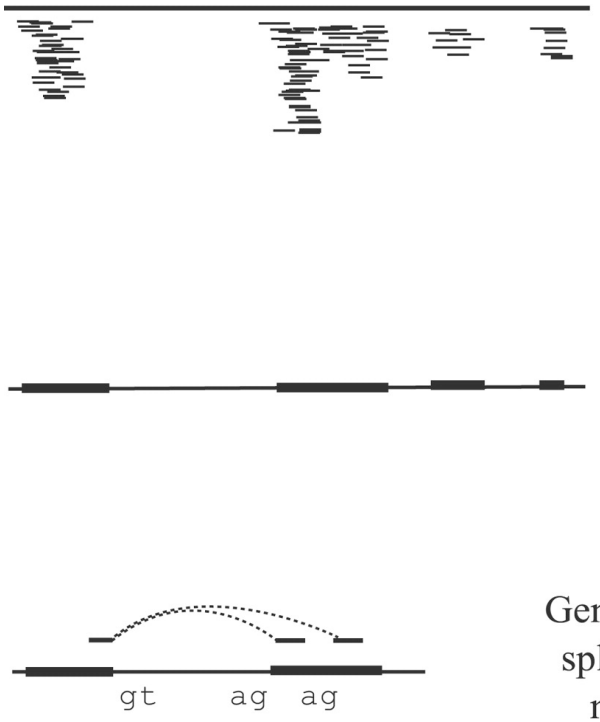
To prevent psedo-gaps of low-expressed genes, merge islands within 70bp of each other (Introns > 70bp)



TopHat: splice junctions

Find GT-AG pairing sites between neighboring (not adjacent) islands

The distance between two sites should > 70bp and <20k bp, as intron length lies within this range



Map reads to whole genome with Bowtie

Collect initially unmappable reads

Assemble consensus of covered regions

Generate possible splices between neighboring exons

Build seed table index from unmappable reads

Map reads to possible splices via seed-and-extend

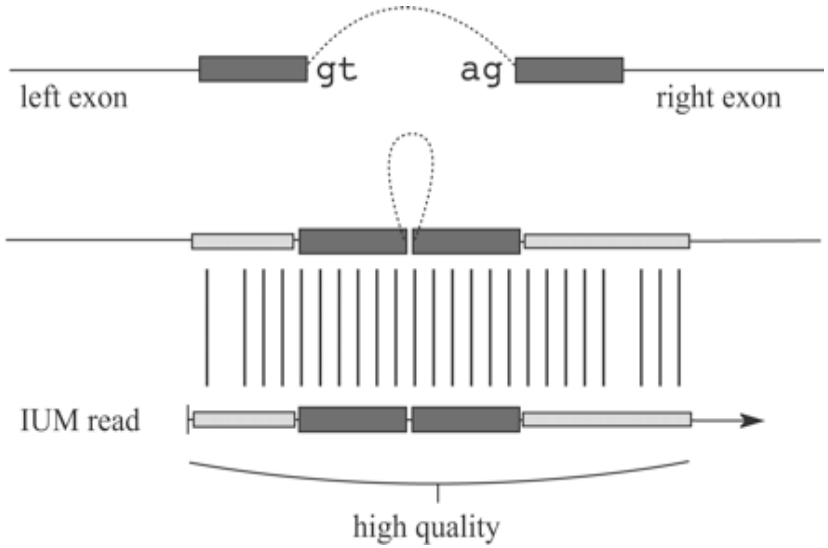
TopHat: IUM

Seed-and-extend strategy:

1. Find IUM span junctions at least k bases on each side

2. 2k-mer 'seed' is constructed by concatenating the k bases on left and right islands

3. Mismatches are allowed except seed regions



TopHat

Map RNA-seq reads to a reference genome

- gapped/spliced mapper

Outputs (4)

- BAM file of mapped reads
- BED files for splice junctions, insertions, deletions

Goal: Understand Mapped Reads

Visualize Tophat Reads

Visualization --> New Visualization

- build: hg19

Add datasets to Visualization

- add hg19 gene annotation
- add all Tophat datasets
- go to chr22:29189371-29196677

Questions to answer:

- how many splice junctions do you see? insertions? deletions?
- do you see any reads that cross *XBP1's* small intron?

Things to try:

- change track options
- show more data (click on icons)

Assembling & Quantifying Expression

Expression values can be tabulated for individual gene loci, transcripts, exons and splice junctions

Gene expression values typically reported in FPKM

- Number of reads (paired read = 2 fragments) per kb of exonic bases per million reads in the library
- ***Why?***

Various software available

- ERANGE (Mortazavi et al, 2008. PMID: 18516045)
- DEGseq package (Wang et al, 2010. PMID: 19855105)
- ALEXA-Seq (Griffith et al, in revision)
- Cufflinks (Trapnell et al, 2010. PMID: 20436464), probably supplants ERANGE

nature biotechnology

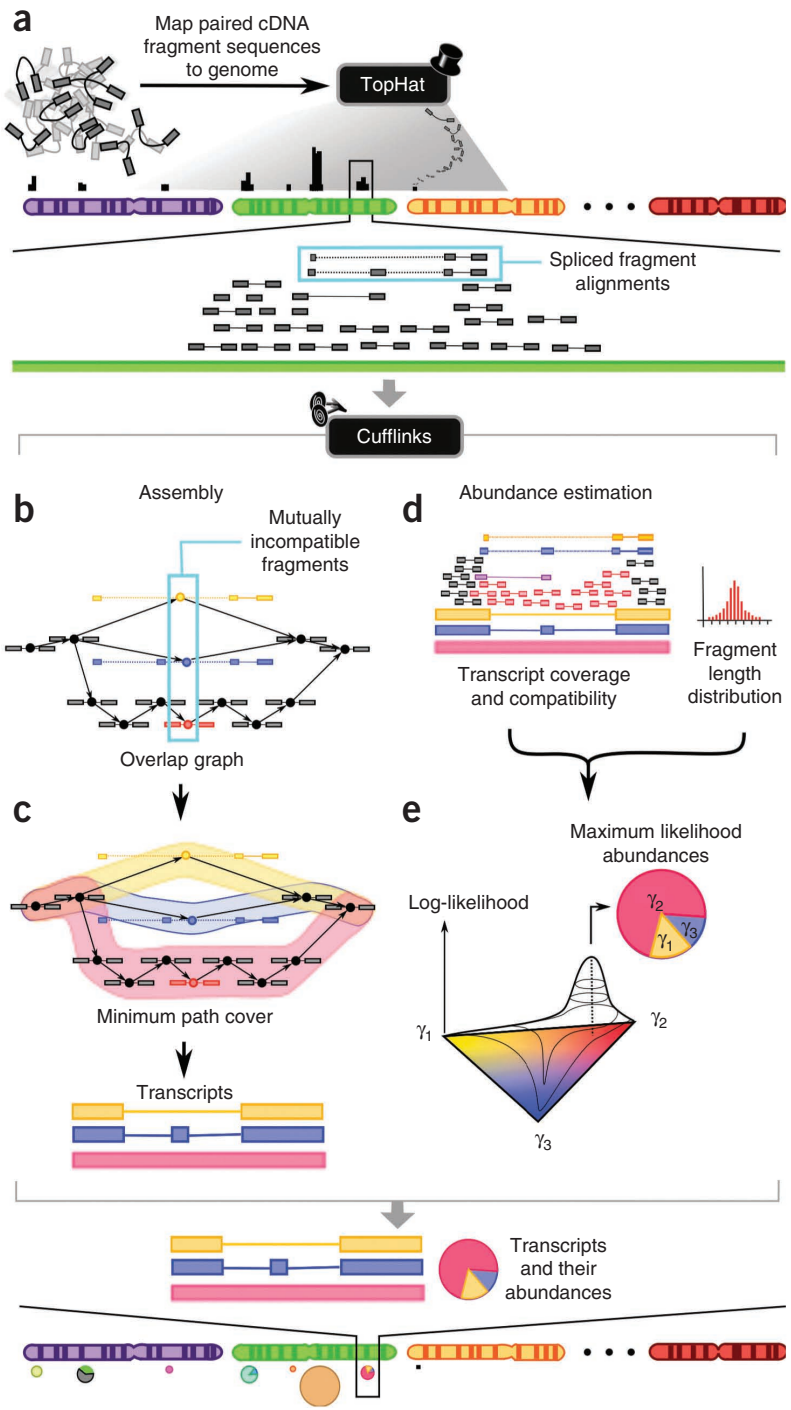
Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

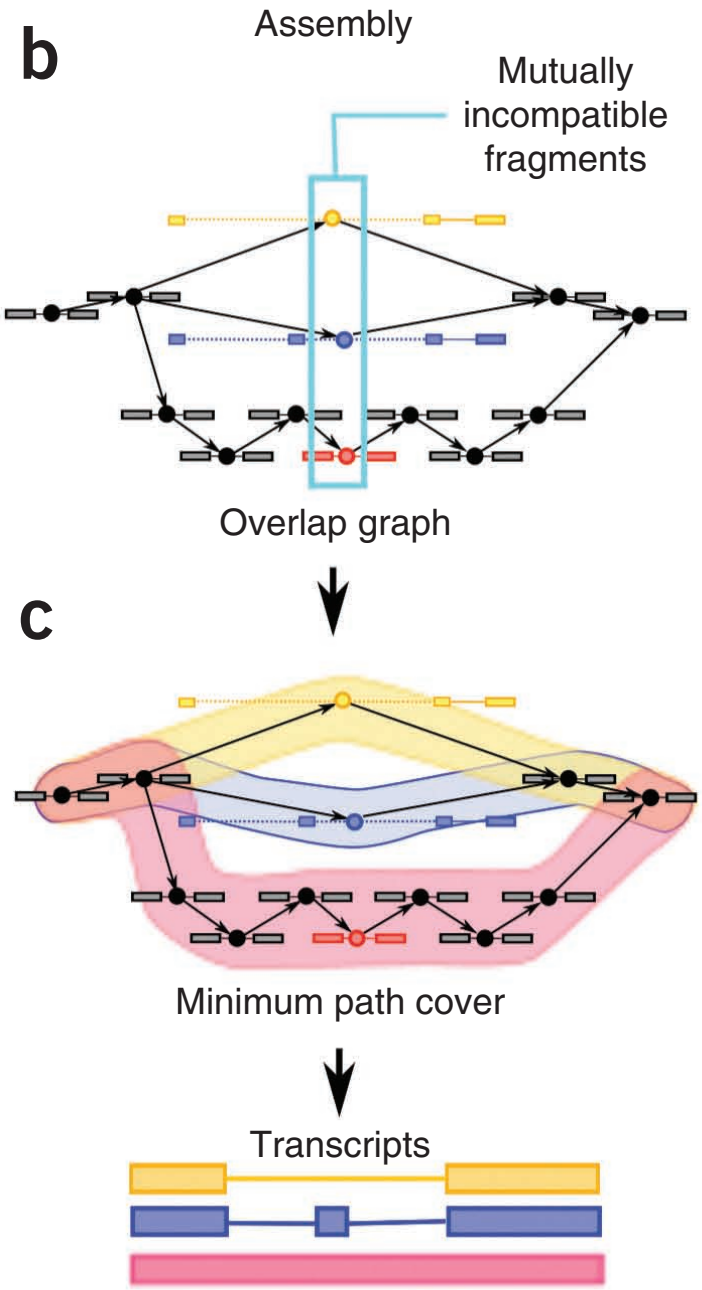
Cole Trapnell[1–3], Brian A Williams[4], Geo Pertea[2], Ali Mortazavi[4], Gordon Kwan[4], Marijke J van Baren[5], Steven L Salzberg[1,2], Barbara J Wold[4] & Lior Pachter[3,6,7]

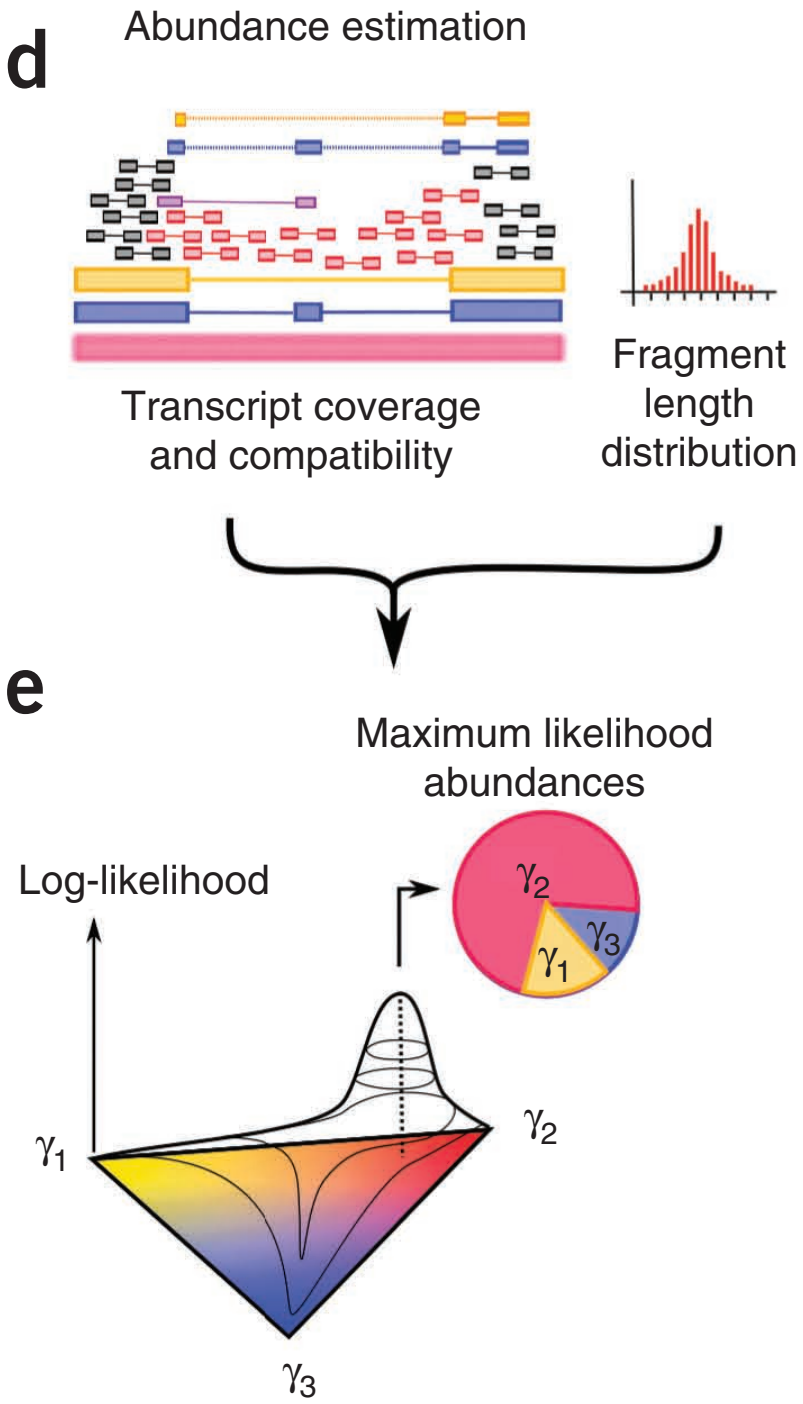
High-throughput mRNA sequencing (RNA-Seq) promises simultaneous transcript discovery and abundance estimation[1–3]. However, this would require algorithms that are not restricted by prior gene annotations and that account for alternative transcription and splicing. Here we introduce such algorithms in an open-source software program called Cufflinks. 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. Over the time series, 330 genes showed complete switches in the dominant transcription start site (TSS) or splice isoform, and we observed more subtle shifts in 1,304 other genes. These results suggest that Cufflinks can illuminate the substantial regulatory flexibility and complexity in even this well-studied model of muscle development and that it can improve transcriptome-based genome annotation.

(75 bp in this work versus 25 bp in our previous work) and pairs of reads from both ends of each RNA fragment can reduce uncertainty in assigning reads to alternative splice variants[12]. To produce useful transcript-level abundance estimates from paired-end RNA-Seq data, we developed a new algorithm that can identify complete novel transcripts and probabilistically assign reads to isoforms.

For our initial demonstration of Cufflinks, we performed a time course of paired-end 75-bp RNA-Seq on a well-studied model of skeletal muscle development, the C2C12 mouse myoblast cell line[13] (see Online Methods). Regulated RNA expression of key transcription factors drives myogenesis, and the execution of the differentiation process involves changes in expression of hundreds of genes[14,15]. Previous studies have not measured global transcript isoform expression; however, there are well-documented expression changes at the whole-gene level for a set of marker genes in this system. We aimed to establish the prevalence of differential promoter use and differential splicing, because such data could reveal much about the model system's regulatory behavior. A gene with isoforms that code for the same protein may be subject to complex regulation to maintain a certain level of output in the face of changes in expression of its


a
Map paired cDNA fragment sequences to genome
TopHat
Spliced fragment alignments
Cufflinks
b
Assembly
Mutually incompatible fragments
Overlap graph
c
Minimum path cover
Transcripts
d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
Transcripts and their abundances


b
Assembly
Mutually
incompatible
fragments
Overlap graph
c
Minimum path cover
Transcripts


d
Abundance estimation
Transcript coverage and compatibility
Fragment length distribution
e
Maximum likelihood abundances
Log-likelihood
γ2
γ3
γ1
γ1
γ2
γ3

Estimating Transcript Abundances

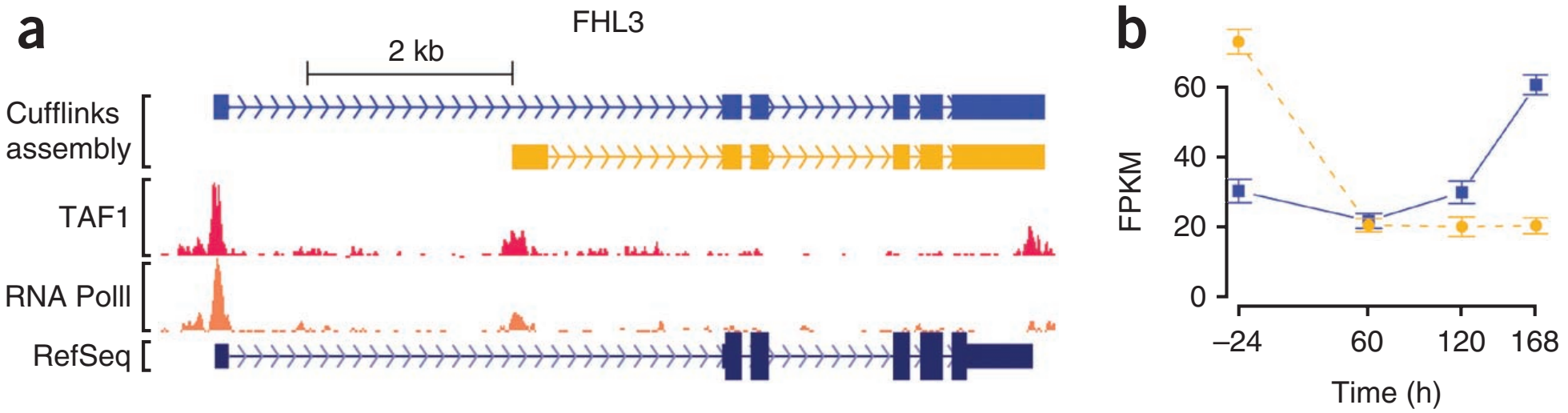
how to know which transcript a fragment belongs to?

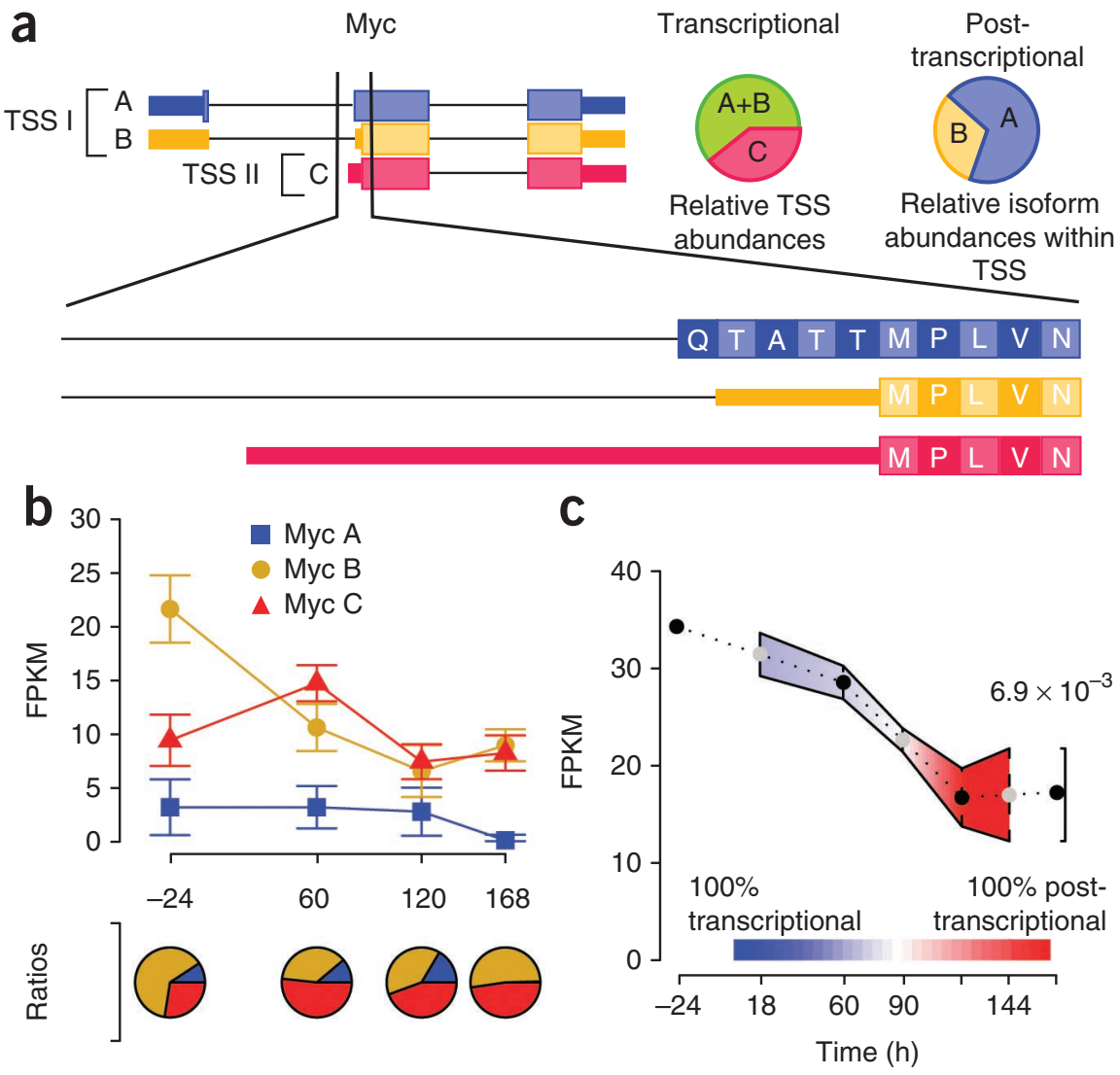
read lengths used to create a (multidimensional) linear model of transcript abundances

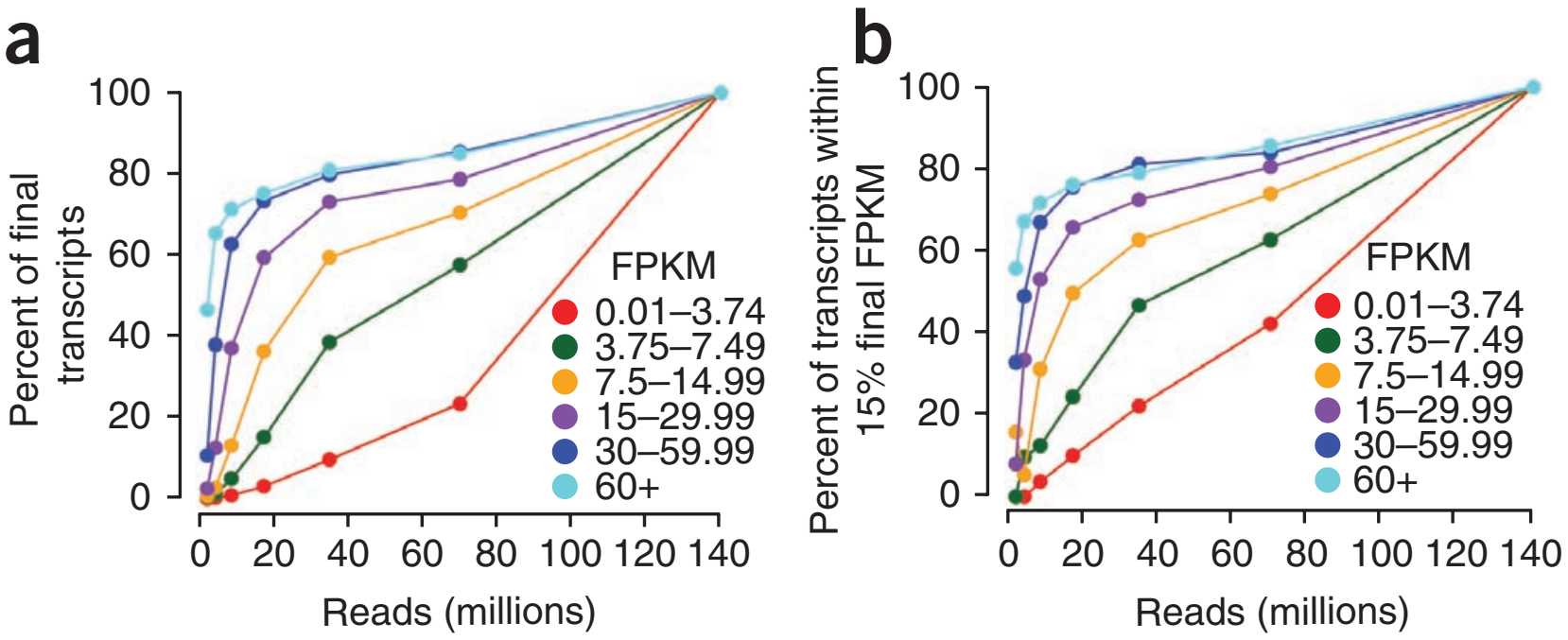
- read length model is critical

The likelihood function is given by

$$
\begin{aligned}
(16) \qquad & L(\rho|R) = \prod_{r\in R} Pr(rd.\ aln. = r) \\
(17) \ = \ & \prod_{r\in R}\sum_{g\in G} Pr(rd.\ aln. = r|locus = g)Pr(locus = g) \\
(18) \ = \ & \prod_{r\in R} \frac{\sigma_{g_r}\tilde{l}(g_r)}{\sum_{g\in G}\sigma_g\tilde{l}(g)} Pr(rd.\ aln. = r|locus = g_r) \\
(19) \ = \ & \prod_{r\in R} \beta_{g_r} \sum_{t\in g_r} Pr(rd.\ aln. = r|locus = g_r, trans. = t)Pr(trans. = t|locus = g_r) \\
(20) \ = \ & \prod_{r\in R} \beta_{g_r} \sum_{t\in g_r} \frac{\tau_t\tilde{l}(t)}{\sum_{u\in g_r}\tau_u\tilde{l}(u)} Pr(rd.\ aln. = r|locus = g_r, trans. = t) \\
(21) \ = \ & \left(\prod_{r\in R}\beta_{g_r}\right)\left(\prod_{r\in R}\sum_{t\in g}\gamma_t \cdot Pr(rd.\ aln. = r|locus = g_r, trans. = t)\right) \\
(22) \ = \ & \left(\prod_{r\in R}\beta_{g_r}\right)\left(\prod_{r\in R}\sum_{t\in g}\gamma_t \cdot \frac{F(I_t(r))}{l(t) - I_t(r) + 1}\right) \\
(23) \ = \ & \left(\prod_{g\in G}\beta_g^{X_g}\right)\left(\prod_{g\in G}\left(\prod_{r\in R: r\in g}\sum_{t\in g}\gamma_t \cdot \frac{F(I_t(r))}{l(t) - I_t(r) + 1}\right)\right).
\end{aligned}
$$


a
FHL3
2 kb
Cufflinks assembly
TAF1
RNA PolII
RefSeq
b
FPKM
60
40
20
0
–24
60
120
168
Time (h)


a
Myc
Transcriptional
Post-transcriptional
TSS I
A
B
TSS II
C
A+B
C
B
A
Relative TSS abundances
Relative isoform abundances within TSS
Q T A T T M P L V N
M P L V N
M P L V N
b
Myc A
Myc B
Myc C
FPKM
30
25
20
15
10
5
0
–24
60
120
168
Ratios
c
FPKM
40
30
20
10
0
6.9 × 10−3
100% transcriptional
100% post-transcriptional
–24
18
60
90
144
Time (h)


a
Percent of final transcripts
100
80
60
40
20
0
0 20 40 60 80 100 120 140
Reads (millions)
FPKM
0.01–3.74
3.75–7.49
7.5–14.99
15–29.99
30–59.99
60+
b
Percent of transcripts within 15% final FPKM
100
80
60
40
20
0
0 20 40 60 80 100 120 140
Reads (millions)
FPKM
0.01–3.74
3.75–7.49
7.5–14.99
15–29.99
30–59.99
60+

Cufflinks

Assembles transcripts and quantifies them

Input: aligned RNA-Seq reads, usually from TopHat, but can be from Bowtie or BWA

Outputs

- assembled transcripts (GTF)
- genes' and transcripts' coordinates, expression levels

General **F**eature **F**ormat (GFF)
Gene **T**ransfer **F**ormat (GTF)

Generic format for specifying connected intervals on the genome

- interval can be transcript, exon, TSS, UTR, etc.
- feature spread across multiple lines
- attributes at the end of line

```
chr22 Cufflinks transcript 29189653 29192232 1000 - .
gene_id "CUFF.1"; transcript_id "CUFF.1.1"; FPKM
"296772.2359645074"; frac "0.660032"; conf_lo "289221.149392";
conf_hi "304323.322538"; cov "289.709057";
```

Goal: Create Reasonable Transcript Assembly for *XBP1*

Cufflinks

Run Cufflinks with default parameters

- add assembled transcripts to visualization

Try reference-guided option

- add to visualization

Try lowering minimum isoform fraction and/or pre-mrna fractions

- add to visualization

Goal: Filter to Remove Assembly Artifacts

Filtering

Developing filter criteria

- use visualization
- mouse over to see information and/or use dynamic filters
- can “run on complete dataset”

Galaxy tools for filtering

- Filter via column using expression
- Filter GFF data by attribute

Goal: Use Workflow to Automate Transcript Assembly and Filtering

Workflows

Create a workflow with the following steps:

- **Cufflinks** with reference-guided option
- **Filter** with condition *Score >= 224*
- **Filter by GFF attribute** with condition *FPKM > 3*

Run workflow on mapped reads from brain

Cuffmerge/compare

Goals

- generate complete list of transcripts for a set of transcripts
- compare assembled transcripts to a reference annotation

Inputs: assembled transcripts from Cufflinks

Outputs:

- Transcripts Combined File
- Transcripts Accuracy File
- Transcripts Tracking Files

Cuffmerge

Run and visualize output :)

Cuffdiff

Goals

- differential expression testing
- transcript quantification

Inputs

- Combined set of transcripts
- mapped reads from 2+ samples

Outputs

- differential expression tests for transcripts, genes, splicing, promoters, CDS
- quantification values for most elements

Cuffdiff

Run it and start reading documentation :)

Learning Galaxy

http://wiki.g2.bx.psu.edu/Learn

Galaxy 101 http://usegalaxy.org/galaxy101
Screencasts
Shared Pages, Histories & Workflows
Other Tutorials
Datasets
Tools
Visualization
User Accounts

Search Galaxy content: http://galaxy.psu.edu/search/
Mailing lists: http://wiki.g2.bx.psu.edu/Mailing%20Lists

- galaxy-user for analysis questions