

Genome-Guided Transcriptome Assembly in the Age of Next- Generation Sequencing

Liliana D. Florea and Steven L. Salzberg
IEEE/ACM Trans Comput Biol Bioinform
2013; 10(5): 1234-1240

Laura Saba, PhD
Assistant Professor
Department of Pharmaceutical Sciences
Skaggs School of Pharmacy and Pharmaceutical Sciences
University of Colorado Anschutz Medical Campus
Laura.Saba@ucdenver.edu

Outline

- My Motivation
- Paper
 1. Introduction
 2. Overview of the RNA-Seq Data Analysis Process
 3. Algorithmic techniques in transcript assembly
 4. Algorithm design considerations
 5. Conclusions
- Extensions
- Remaining Questions

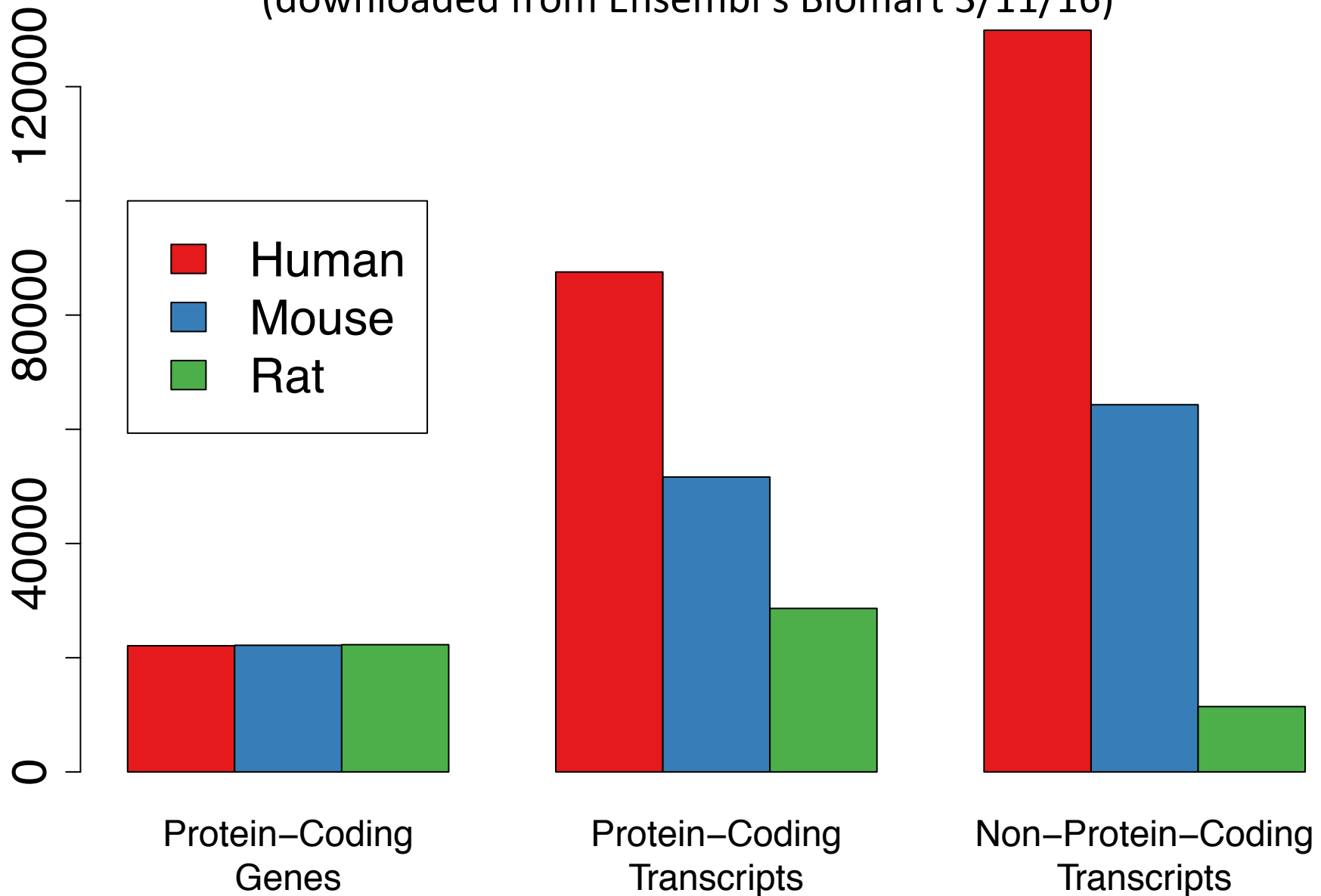
Motivation

- We really like rats:
 - Rich behavioral repertoire
 - Larger than mice for detailed physiologic measures
 - Disease genes identified in rats shown to play role in human disease
 - Amenable to invasive or terminal procedures
 - TALEN and CRISPR-Cas9 technology for genetic editing available



Lack of Annotation in Rat

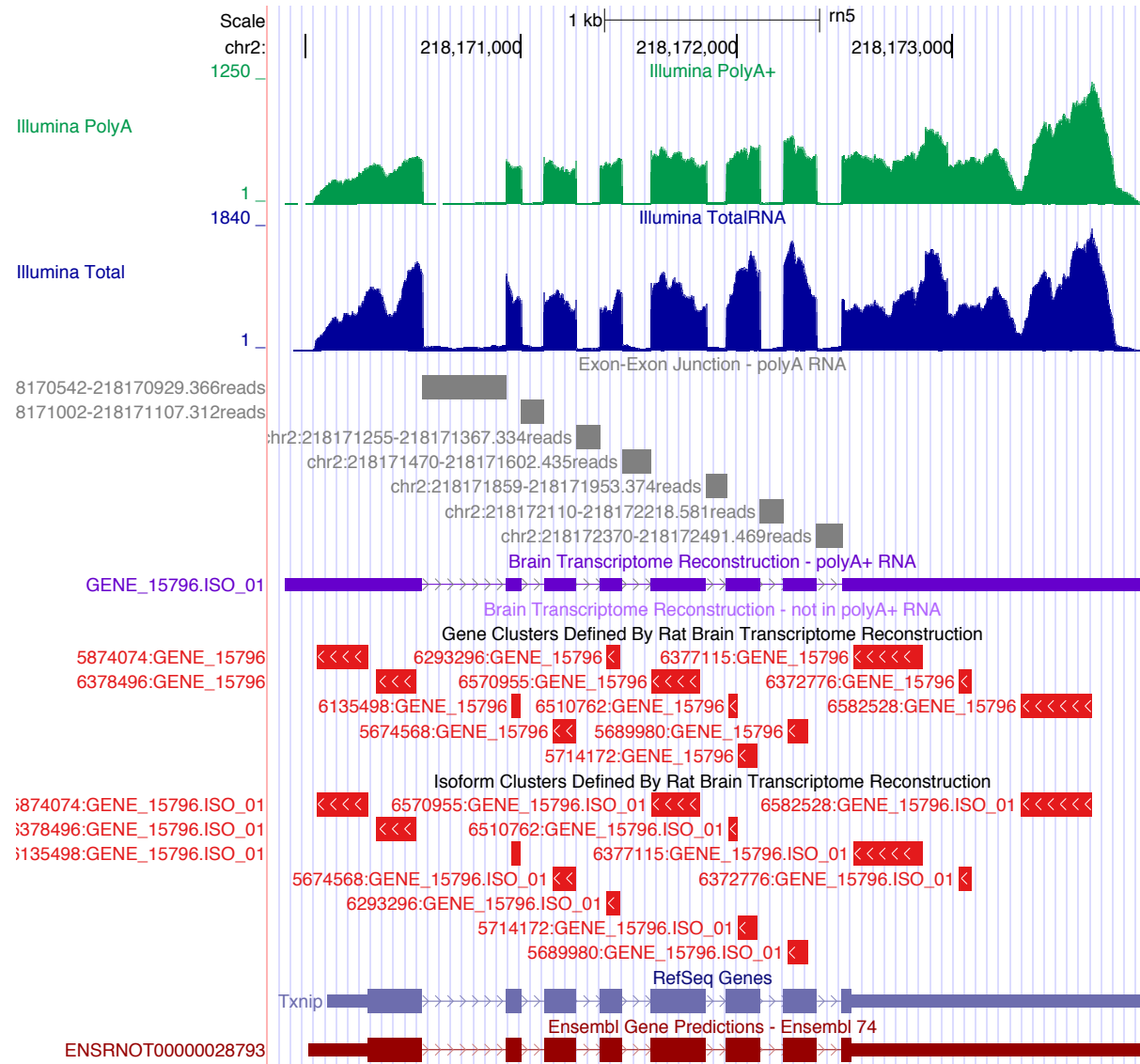
(downloaded from Ensembl's Biomart 3/11/16)



Results from Simple Genome-Guided Reconstruction in CuffLinks

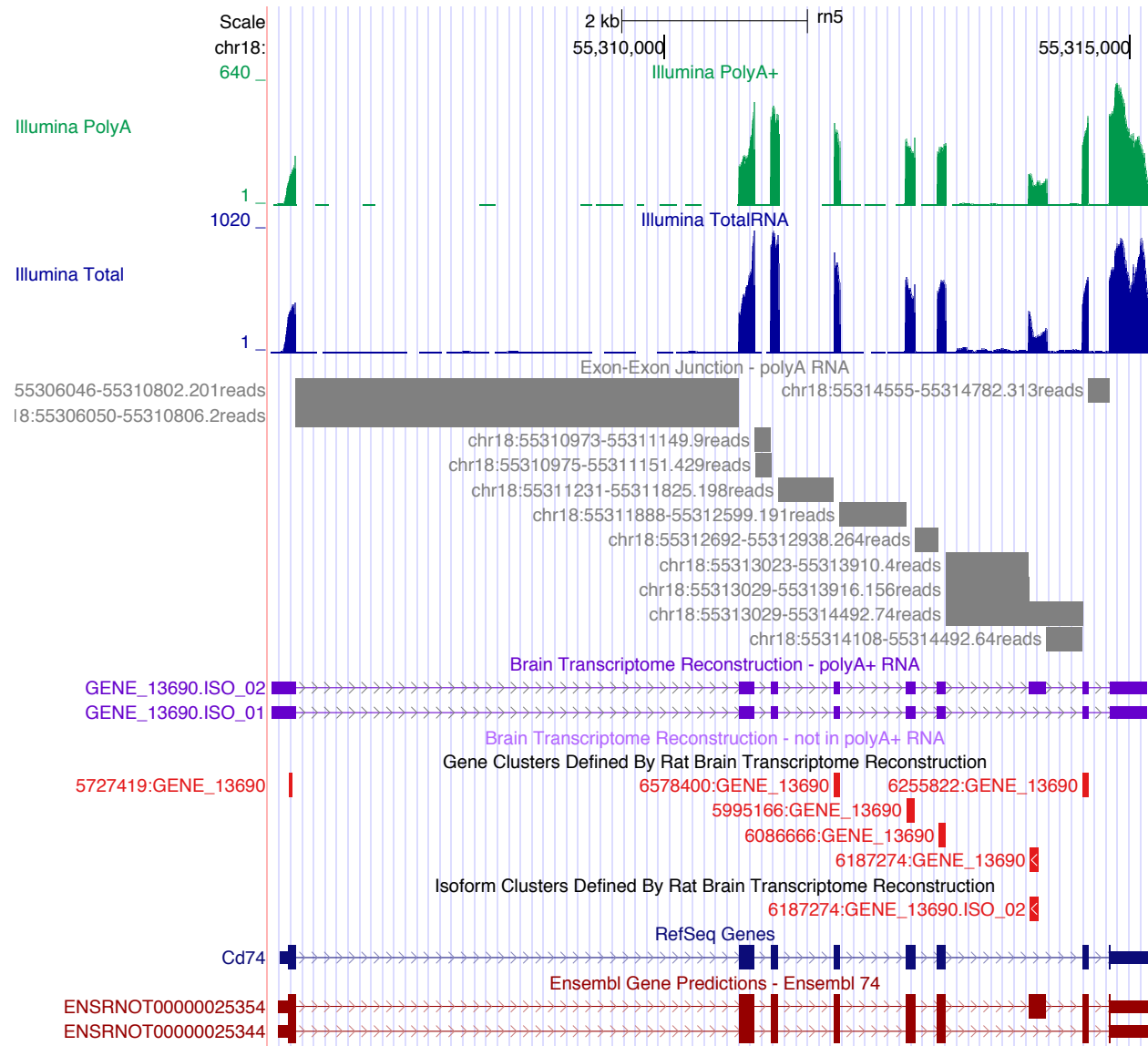
- Heart Reconstruction
 - 377M paired-end reads from SHR = 167,389 transcripts (152,030 genes)
 - 17,569 Ensembl transcripts
 - 1,617 transcripts recovered from skipped regions
 - 332M paired-end reads from BNLx = 165,182 transcripts (149,619 genes)
 - 17,500 Ensembl transcripts
 - 1,572 transcripts recovered from skipped regions

Examples From Recent Publication



Saba et al (2015). The sequenced rat brain transcriptome, its use in identifying networks predisposing alcohol consumption. FEBS J.; 282(18):3556-78.

Examples From Recent Publication



Saba et al (2015). The sequenced rat brain transcriptome, its use in identifying networks predisposing alcohol consumption. FEBS J.; 282(18):3556-78.

1. INTRODUCTION

RNA-Seq and the Transcriptome

- Even human genome is under-annotated
 - Missing splice variants
 - Long non-coding RNA
- Would need to deep sequence every cell type at every developmental stage to identify an exhaustive list of transcripts

Transcriptome Assembly

- Definition – “assembling reads together to reconstruct the transcripts from which they came”
- *De novo* transcriptome assembly – assembling transcripts when a genome sequence is not available
- **Genome-guided transcriptome assembly** – using the genome as a guide when assembling transcripts

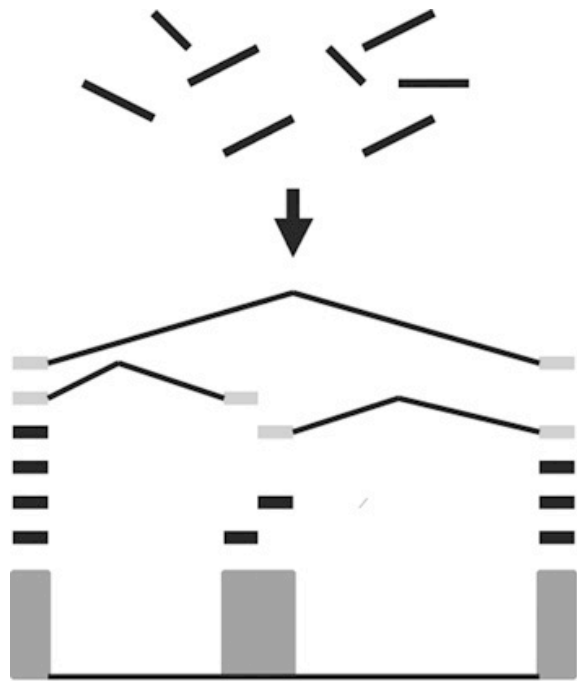
Challenges for Reconstructing Transcripts

- Fragmentation
 - i.e., reads aren't as long as transcripts
 - Some reads align to multiple places in the genome
 - Repeat regions, gene families, pseudogenes
 - Precisely aligned reads may originate from any of several transcripts at the same gene locus
 - More than 90% of human genes have multiple isoforms
 - Multiple isoforms due to alternative splicing, alternative transcription initiation and termination sites

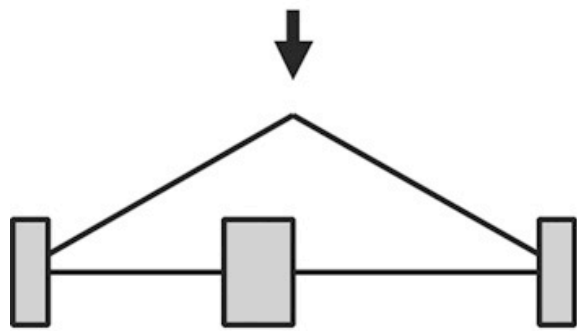
Challenges for Reconstructing Transcripts

- Variability in read coverage
 - between transcripts
 - within a transcript
- Computational efficiency

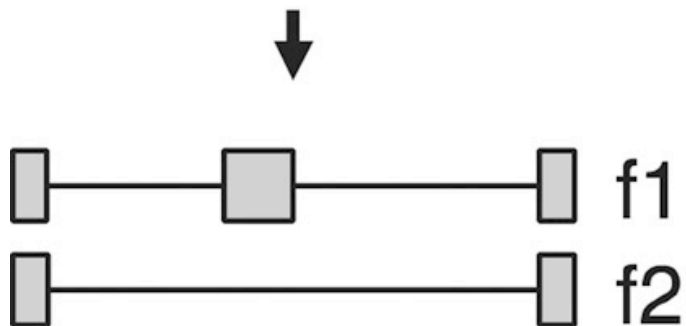
2. OVERVIEW OF THE RNA-SEQ DATA ANALYSIS PROCESS



1. RNA-seq experiment



2. Alignment



3. Transcript assembly
and selection

4. Abundance estimation

www.rna-seqblog.com

RNA-Seq Blog
Transcriptome Sequencing Research & Industry News

otogenetics
corporation

Next-Gen RNA sequencing
made easy and affordable

HOME

NEWS »

EVENTS »

JOBS »


TECHNOLOGY »

DATA ANALYSIS »

BLOG

CONTACT »

RNA-SEQ NEWS

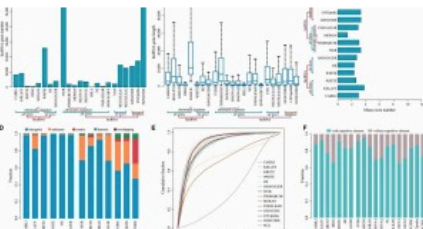


Minimally invasive genomic and transcriptomic profiling of visceral cancers by next-generation sequencing of circulating exosomes

14 hours ago 0 152 Views

The ability to perform comprehensive profiling of cancers at high resolution is essential for precision medicine. Liquid biopsies using shed exosomes provide high-quality nucleic acids to obtain molecular characterization, which ...

[Read More »](#)




A comprehensive overview of lncRNA annotation resources

14 hours ago 0 411 Views

Long noncoding RNAs (lncRNAs) are emerging as a class of important regulators participating in various biological functions and disease processes. With the widespread application of next-generation

STAY CONNECTED



POLLS

How are you applying RNA-Seq?

☐ Still learning about RNA-Seq

☐ Basic Research

☐ Translational Research

☐ Clinical Research

☐ Clinical Care

[Vote](#)


[View Results](#)

SUBSCRIBE TO THE RNA-SEQ BLOG

[Subscribe](#)

RNA-SEQ PRODUCTS & SERVICES

Your most challenging RNA-Seq samples.



3. ALGORITHMIC TECHNIQUES IN TRANSCRIPT ASSEMBLY

3.1 Transcript Representation and Enumeration

Common approach → Build a directed acyclic graph and tranverse graph to resolve individual isoforms

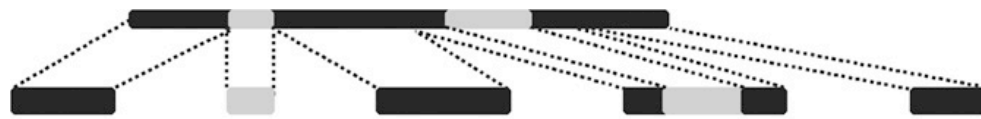
3.1.1 Overlap graph (read based)

3.1.2 Connectivity graph (nucleotide based)

3.1.3 Splice graph and subexon graph (exon/
subexon based)

3.1 Transcript Representation and Enumeration

3.1.1 Overlap graph – compactly represents the order of reads along putative transcripts



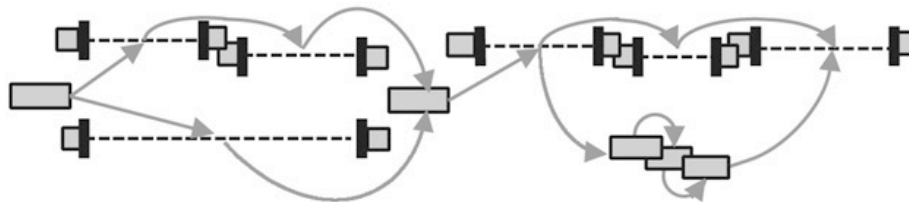
RNA molecules



Read coverage levels

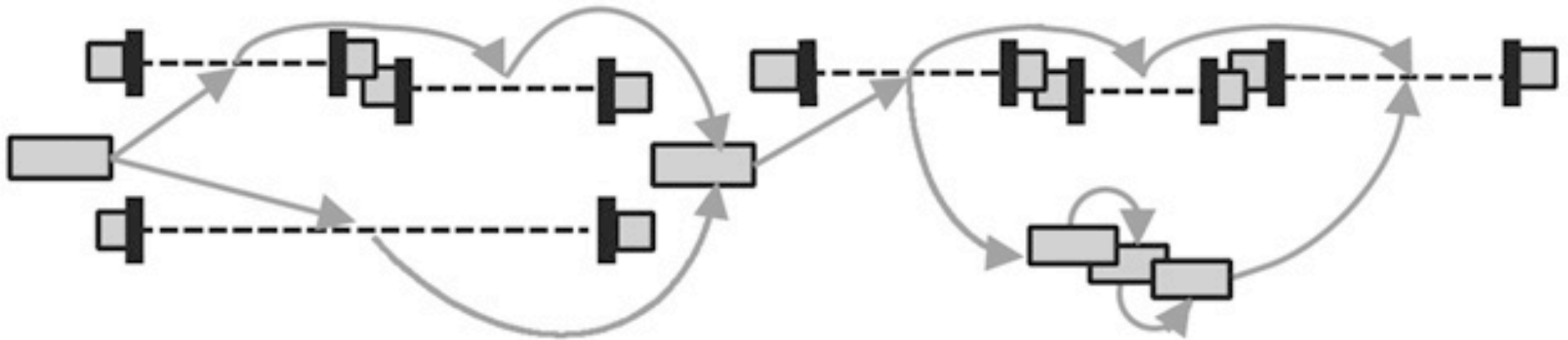


Splice junctions



Overlap graph

Overlap Graph



- Nodes – reads
- Edges – overlapping compatible reads
- Reads contained within another read are ignored for simplicity
- “uncertain” reads are ignored

“Uncertain” Reads

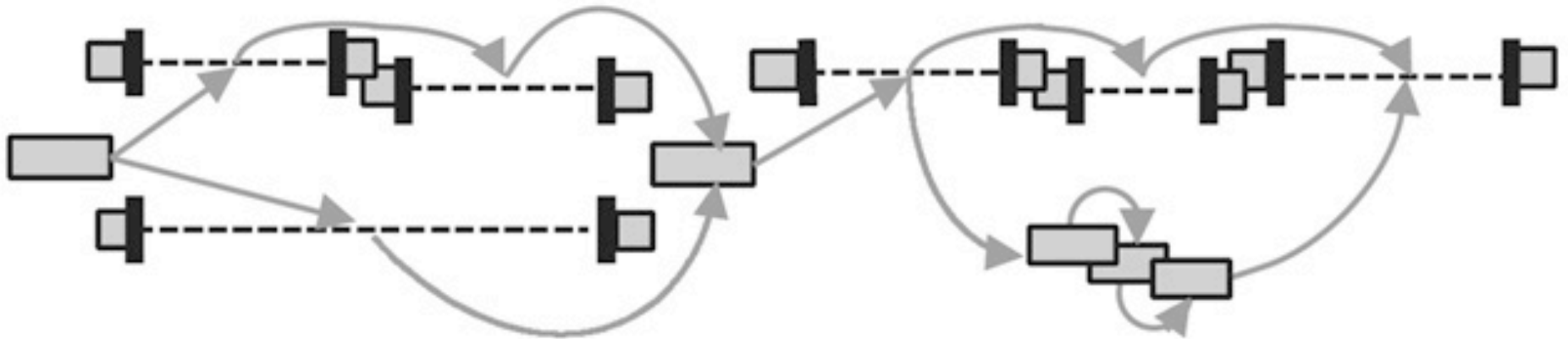
(a)



- P2 can be compatible with the both P1 and P3, but P1 and P3 cannot be from the same transcript
→ P2 is an “uncertain” read

Overlap Graph

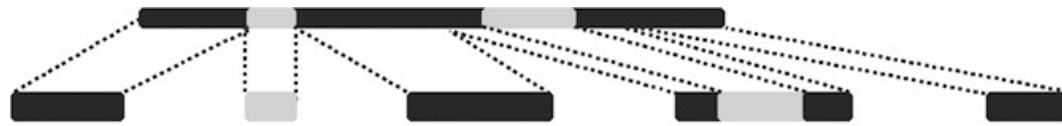
- Transverse graph to paths (putative transcripts) that cover all qualifying reads



Example of Overlap Graph

- Cufflinks
 - Infers exons and full transcripts simultaneously
 - Selects the minimum number of transcripts (isoforms) that can explain all reads
 - Using an efficient polynomial time partitioning algorithm
 - May miss some transcripts by picking ‘minimum’ number of transcripts
 - May pick wrong set to represent minimum number of transcripts

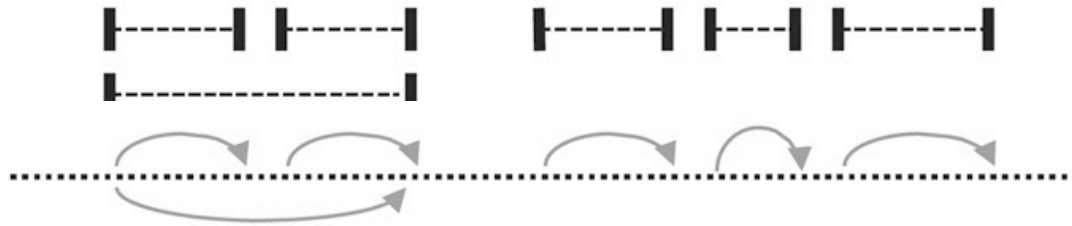
Connectivity Graph



RNA molecules



Read coverage levels



Splice junctions

Connectivity graph

- DAG using bases instead of reads

Examples of Connectivity Graphs

- Scripture
 1. Start with all bases on a chromosome as nodes and edges connect consecutive bases on the genome and the two endpoints of an intron
 2. Uses segmentation approach which incorporates read coverage to determine significant paths
 3. Significant paths → splice graph

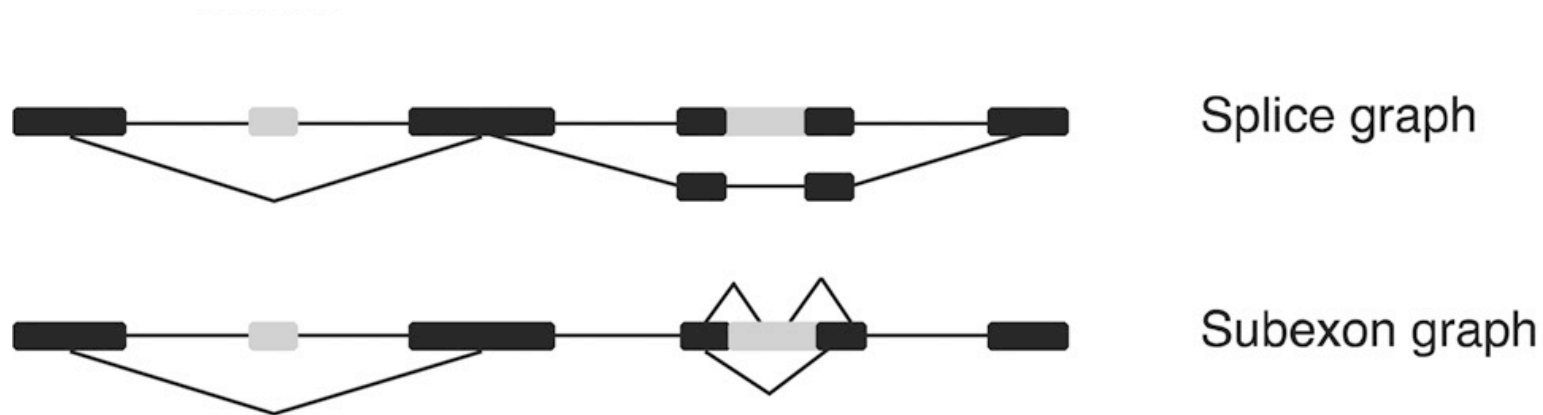


4. Enumerates all possible transcripts

Examples of Connectivity Graphs

- IsoLasso
 1. Like Scripture but only includes bases with some read coverage
 2. Enumerates all possible transcripts
 3. Later reduces the number of possible transcripts using a quadratic program

Splice Graph and Subexon Graph



- **Splice graph** – exons as nodes, connected by introns (edges); splice variants can be read from the graph as maximal paths
- **Subexon graph** – connects gene segments if they are adjacent on the genome as part of the same exon, or are connected via a spliced reads

Too Many Transcripts

- ‘Maximal paths’ can produce thousands of possible transcripts
- Need to select a subset of likely transcripts
 - e.g., greedy methods, dynamic programming, linear programming, and LASSO or expectation maximization (EM) algorithms.

3.2 Transcript Selection via Numerical Optimization

3.2.1 Quadratic programming – finds the combination of transcripts that best explains the observed read coverage levels, by minimizing the total estimation error

$$X^* = \underset{X}{\operatorname{argmin}} f(X) = \sum_{i=1}^M \left(\frac{r_i}{l_i} - \sum_{j=1}^N a_{ji} x_j \right)^2$$

M = the number of subexons in gene

r_i = number of reads aligned to subexon i

l_i = length of subexon i

a_{ji} = indicator variable for if subexon i is included in transcript j

x_j = abundance level of transcript j

N = number of possible transcripts

Examples of Quadratic Programming

- IsoLasso
 - starts from a connectivity graph
 - applies a penalty function to minimize the number of isoforms in the solution
- SLIDE
 - starts from a subexon graph constructed from known annotations

3.2 Transcript Selection via Numerical Optimization

3.2.2 Expectation Maximization

1. assigns a probability to each read/transcript pair which is based on a current abundance estimate (Expectation Step)
2. calculates the expression levels for all isoforms (Maximization Step)

Examples of EM Algorithm

- Cufflinks
 - used only to estimate abundance and does not influence the set of transcripts
- iReckon
 - used to estimate abundance and determine set of transcripts

Paired-End Reads

- All algorithms can take paired-end reads into account
- All algorithms use paired-end reads at the graph construction stage

4. ALGORITHM DESIGN CONSIDERATIONS

Condensing Strategy

- Minimum Number of Isoforms
 - may miss some isoforms (true set is bigger or minimal set is not unique)

- Exhaustive set of isoforms

NO GOLD STANDARD

- implication on precision of isoform expression estimates
- Best Fit
 - difficult to solve when starting with a large number of possible transcripts

TABLE 1**Design Characteristics of Transcript Assembly Programs**

Program	Condensing strategy	Read completeness	Transcript feasibility	Transcript representation	Transcript selection	Intronic reads context	Novel gene ends
Cufflinks	Parsimony	No	Yes	Overlap graph	Minimum partition	Exon	Partial
IsoLasso	Best fit	No	Yes	Connectivity graph	QP	Not available	Partial
Scripture	Exhaustive	Yes	No	Connectivity graph, splice graph	None	Genome	Partial
iReckon	Best fit	Yes	No	Splice graph	EM	Gene	No
SLIDE	Best fit	No	No	Subexon graph	QP	Ignore	No
SpliceGrapher	None	Yes	No	Splice graph	None	Ignore	No

QP - quadratic programming; EM - expectation maximization.

Read Completeness

- Are the all of the input reads used for the solution?

TABLE 1

Design Characteristics of Transcript Assembly Programs

Program	Condensing strategy	Read completeness	Transcript feasibility	Transcript representation	Transcript selection	Intronic reads context	Novel gene ends
Cufflinks	Parsimony	No	Yes	Overlap graph	Minimum partition	Exon	Partial
IsoLasso	Best fit	No	Yes	Connectivity graph	QP	Not available	Partial
Scripture	Exhaustive	Yes	No	Connectivity graph, splice graph	None	Genome	Partial
iReckon	Best fit	Yes	No	Splice graph	EM	Gene	No
SLIDE	Best fit	No	No	Subexon graph	QP	Ignore	No
SpliceGrapher	None	Yes	No	Splice graph	None	Ignore	No

QP - quadratic programming; EM - expectation maximization.

Transcript Feasibility

- Can all the transcripts produced by the method be explained by the reads?

TABLE 1

Design Characteristics of Transcript Assembly Programs

Program	Condensing strategy	Read completeness	Transcript feasibility	Transcript representation	Transcript selection	Intronic reads context	Novel gene ends
Cufflinks	Parsimony	No	Yes	Overlap graph	Minimum partition	Exon	Partial
IsoLasso	Best fit	No	Yes	Connectivity graph	QP	Not available	Partial
Scripture	Exhaustive	Yes	No	Connectivity graph, splice graph	None	Genome	Partial
iReckon	Best fit	Yes	No	Splice graph	EM	Gene	No
SLIDE	Best fit	No	No	Subexon graph	QP	Ignore	No
SpliceGrapher	None	Yes	No	Splice graph	None	Ignore	No

QP - quadratic programming; EM - expectation maximization

Intronic Reads

- Because unspliced mRNAs can be contained in the sample, intronic reads occur often
- The algorithms differ on how they determine if the intronic reads represent a retained intron or just noise

Design Characteristics of Transcript Assembly Programs

Program	Condensing strategy	Read completeness	Transcript feasibility	Transcript representation	Transcript selection	Intronic reads context	Novel gene ends
Cufflinks	Parsimony	No	Yes	Overlap graph	Minimum partition	Exon	Partial
IsoLasso	Best fit	No	Yes	Connectivity graph	QP	Not available	Partial
Scripture	Exhaustive	Yes	No	Connectivity graph, splice graph	None	Genome	Partial
iReckon	Best fit	Yes	No	Splice graph	EM	Gene	No
SLIDE	Best fit	No	No	Subexon graph	QP	Ignore	No
SpliceGrapher	None	Yes	No	Splice graph	None	Ignore	No

QP - quadratic programming; EM - expectation maximization.

Novel Gene Ends

- None of the methods handle this well.
- Some can detect alternative polyadenylation events or alternative promoter usage if the transcripts has an alternative last or first exon.

Design Characteristics of Transcript Assembly Programs

Program	Condensing strategy	Read completeness	Transcript feasibility	Transcript representation	Transcript selection	Intronic reads context	Novel gene ends
Cufflinks	Parsimony	No	Yes	Overlap graph	Minimum partition	Exon	Partial
IsoLasso	Best fit	No	Yes	Connectivity graph	QP	Not available	Partial
Scripture	Exhaustive	Yes	No	Connectivity graph, splice graph	None	Genome	Partial
iReckon	Best fit	Yes	No	Splice graph	EM	Gene	No
SLIDE	Best fit	No	No	Subexon graph	QP	Ignore	No
SpliceGrapher	None	Yes	No	Splice graph	None	Ignore	No

QP - quadratic programming; EM - expectation maximization.

5. CONCLUSIONS

We aren't there yet...

- Need work on
 - distinguishing transcriptional noise from intronic retention
 - removing read mapping artifacts
 - how to accurately assess number of transcripts
 - estimating isoform level expression
- Possible future solutions
 - longer reads that capture the entire transcript
 - public catalogs continue to grow and get closer to a 'gold standard'

Extensions

- Transcriptome and genome guided assemblies
- Mixing genome guided and de novo assemblies
- Incorporate multiple samples
 - ISP
- Alternative methods
 - StringTie
 - Traph
 - Bayesemblem
 - Astroid
 - Grit
 - ParSeq
 - ...

My Remaining Questions

- Accuracy measure for individual transcripts
- Defining alternative transcription start and stop sites
- Combining transcriptome from multiple genetic backgrounds

Shameless Plug

- Lots of RNA-Seq data
 - panel of rats with lots of other data already measured by us and other researcher around the globe
 - DNA sequence
 - Phenotypes
 - different tissue, same rat and/or strain
 - some biological replicates
 - mRNA, long non-coding RNA, small RNA including microRNA
 - endless questions
 - access to additional animals/tissue/RNA for follow-up studies