Analysis of RNA-seq Data
A physicist and an engineer are in a hot-air balloon. Soon, they find themselves lost in a canyon somewhere. They yell out for help: "Hellllloooooo! Where are we?"

- 15 minutes later, they hear an echoing voice: "Hellllloooooo! You're in a hot-air balloon!!"
- The physicist says, "That must have been a mathematician."
- The engineer asks, "Why do you say that?"
- The physicist replies: "The answer was absolutely correct, and it was utterly useless."
Introduction
What is RNA-seq?

- RNA-seq refers to the method of using Next Generation Sequencing (NGS) technology to measure RNA levels.
- Is used to evaluate the “expression level” of a gene (or “gene expression”).
- Many events can control the expression level of a gene so simply looking at the genome and annotating a gene is not enough information.
Item to be sequenced:
1. Extract all RNA.
2. Prepare a library of fragments.
3. Sequence fragments.
4. Analysis, analysis, analysis.
Splicing

• A very important modification of eukaryotic pre-mRNA is splicing.
• The majority of eukaryotic pre-mRNAs consist of alternating segments called **exons** and **introns**.
• During splicing, an RNA-protein complex called a **spliceosome** will remove an intron and splice together the neighboring exon regions.
• The spliced together exons create the code that will be translated into proteins.
Alternative Splicing

• Some introns or exons can be either removed or retained in mature mRNA.
• This is referred to as alternative splicing and it creates a series of different transcripts from a single gene.
• These different transcripts can be potentially translated into different proteins, splicing extends the complexity of eukaryotic gene expression.
Alternative Splicing

Isoform 1

Isoform 2

Isoform 3
Alignment of RNA-seq Reads
Splicing Junction

- The consensus sequence within the intron region creates a splicing junction that is more easily identifiable from a computational perspective.
- Referred to as “canonical splicing forms”.
- GU-AG is the most common canonical form but there are others.
Alignment of RNA-seq Reads

Whenever a RNA-seq read spans an exon boundary, part of the read will not map contiguously to the reference, which often causes the mapping procedure to fail for that read.
Alignment of RNA-seq Reads

- Previous methods solve this problem by concatenating known adjacent exons and then creating synthetic sequence fragments from these spliced transcripts
RNA-Seq Alignment Programs

- **GSNAP (Genomic Short-read Nucleotide Alignment Program):** aligns both single- and paired-end reads. Uses a probabilistic model or a database of known splice sites.
- **MicroRazerS:** aligns short RNA-seq reads.
- **Others:** BWA, Bowtie, OSA, RUM, PALMapper, many more.
TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

TopHat is a collaborative effort between the Institute of Genetic Medicine at Johns Hopkins University, the Departments of Mathematics and Molecular and Cell Biology at the University of California, Berkeley and the Department of Stem Cell and Regenerative Biology at Harvard University.

TopHat 2.0.4 release 6/21/2012

Version 2.0.4 is a maintenance release addressing some issues found in the 2.0.3 release:

- Fixed a bug that caused the last stage of TopHat (tophat_reports) to occasionally crash for large data sets.
- For paired reads found to be incorrectly paired in the input files TopHat now outputs a warning message instead of terminating with an error.
- Alignments of paired reads mapped discordantly (e.g. on different chromosomes) are now reported by default. To disable this behavior, --no-discordant option can be used. Also please check --no-mixed option in the manual, which we borrow from Bowtie2 options.
- --fusion-search with Bowtie2 is still in developmental stage, it may require much memory space and produce many spurious fusions. You may want to try a combination of --bowtie1 and --fusion-search if it does not work.
- Environment variables such as BOWTIE_INDEXES and BOWTIE2_INDEXES are handled properly - please refer to the Bowtie website for more details about the variables.
- Prebuilt transcriptome indexes built by older versions of TopHat may not be compatible with this version due to some internal changes in parsing gtf files. It is strongly recommended to build a new transcriptome index.

TopHat 2.0.3 release 5/28/2012

Site Map

Home
Getting started
Manual
Index and annotation downloads
FAQ
Protocol
Fusion mapping

News and updates

New releases and related tools will be announced through the mailing list

Getting Help

Questions about TopHat should be sent to tophat.cufflinks@gmail.com. Please do not email technical questions to TopHat contributors directly.
TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner and then analyzes the mapping results to identify splice junctions between exons.
Shortcomings of Existing Tools

Existing programs fail to detect splice junctions for a variety of reasons, including:

• Very low sequencing coverage, in which case there might not be any read that straddles the junction with sufficient sequence on each side.

• Junctions spanning very long introns.

• Junctions with non-canonical forms.
Transcript Assembly and Quantification
De Novo vs. References Guided Transcript Assembly

• **De novo transcript assembly**: assembly of transcripts where there exists no reference genome

• **Reference guided transcript assembly**: significantly easier than de novo assembly
  – Map to the reference (using the methods discussed from last time) and use the alignment to guide assembly
Cufflinks (Trapnell et al)

• **Cufflinks**: an algorithm that identifies complete novel transcripts and probabilistically assigns reads to isoforms.

• Extends the work of TopHat (Pachter lab).

• The RNA sequence fragments are mapped to the reference using TopHat.

• Aim is to recover the minimal set of transcripts supported by the alignments.
Map paired cDNA fragment sequences to genome

TopHat

Spliced fragment alignments

Cufflinks
Cufflinks (Trapnell et al)

- A fragment corresponds to a single cDNA molecule, which can be represented by a pair of reads from each end.
- Uses a comparative transcriptome assembly algorithm to produce the minimal set of transcripts supported by the fragment alignment.
- Reduces the transcript assembly problem to finding a maximum matching in a weighted bipartite graph.
TopHat

CuffLinks
Cufflinks (Trapnell et al)

- Takes as input cDNA fragment sequences that have been aligned to the genome by using software that is capable of doing split alignments.
- With paired-end RNA-seq, Cufflinks treats each pair of fragment reads as a single alignment.
- The alignments are then ranked.
- Only the highest ranked alignments are used.
Ranking

Let x and y be fragment alignments. x < y if:
1. x is a single alignment and y is not;
2. x crosses more splice junctions than y;
3. the reads from x map significantly farther apart and y’s do not;
4. the reads from x are significantly closer together and y’s do not;
5. x and y both span an intron region and x spans a longer one;
6. x has more mismatches than y.
The first step is to 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. Each **fragment** has one **node in the graph**, and an **edge**, directed from left to right along the genome, is placed between each pair of **compatible fragments**.
Compatible Fragments

• Two pairs of fragments, x and y, are defined to be compatible if they do not overlap, or if every implied intron in one fragment overlaps an identical intron in the other fragment.

• Sometimes it is impossible to identify whether a fragment is compatible or not (uncertain).

• Nested fragments are contained in the other.
Partially Ordered Set

- **Partially ordered set** (or **poset**) formalizes and generalizes the intuitive concept of an ordering, sequencing, or arrangement of the elements of a set.

\[
\text{Poset} = \text{Set} + \text{Binary Relation}
\]
Antichain and Poset Width

• We say two elements $a$ and $b$ of a partially ordered set are **comparable** if $a \leq b$ or $b \leq a$.

• **Chain**: set of elements every two of which are comparable.

• **Antichain**: subset of a partially ordered set such that any two elements in the subset are incomparable.

• **Width of a poset**: the cardinality of a maximum antichain.
From Fragments to Poset

A partial order (directed acyclic graph) is constructed from the fragments (uncertain and nested are omitted) as follows: \( a \leq b \) for two fragments, \( a \) and \( b \), if \( a \) begins at or before \( b \), and \( a \) and \( b \) are compatible.
Assembly

Mutually incompatible fragments

Overlap graph
Isoforms are then assembled from the overlap graph.

Paths through the graph correspond to **mutually compatible fragments** that could be merged into complete isoforms.
• A partition of P into chains yields an assembly because every chain is a totally ordered set of compatible fragments.

• The problem of finding a minimum partition P into chains is equivalent to finding a maximum antichain in P. (By Dilworth’s theorem)

• An antichain is a set of mutually incompatible fragments.
Why Dilworth’s Theorem?

• Because the problem of finding a maximum antichain in $P$ can be reduced to finding a maximum matching in a certain bipartite graph.

• The “reachability graph” is the transitive closure of the DAG, i.e. it is the graph where each fragment $x$ has nodes $L_x$ and $R_x$, where there is an edge between $L_x$ and $R_y$ when $x \leq y$ in $P$.

• Finding a maximum matching can be done in polynomial time.
Thus, we can solve the problem of finding the minimum number of paths (i.e. “isoforms”) in the graph that explain all the read alignments in polynomial time. Surprising?
Transcript Abundance is Estimated

- Fragments are matched to the transcripts from which they could have originated.
- **Transcript abundance is estimated** using a statistical model in which the probability of observing each fragment is a linear function of the abundance of the transcripts from which it could have originated.
- Because only the ends of each fragment are sequenced, the length maybe unknown.
Assigning a fragment to different isoforms often implies a length for it. Cufflinks incorporates the distribution of fragment lengths to help assign fragments to isoforms.
Lastly, Cufflinks maximizes a function that assigns a likelihood to all possible sets of relative abundances, which produces the abundance that best explains the observed fragments (shown in the pie chart).
De Novo Transcript Assemblers

**Trans-ABySS**: one of the first tools, a repurposed de Bruijn genome assembler (ABySS) that works well for viruses and bacteria.

**Oases**: is the equivalent to Trans-ABySS from the developers Velvet.
De Novo Transcript Assemblers

- **Trinity** is probably the best one in terms of results and ease of use. The original paper showed some impressive results on non-coding RNAs in mammals.

- **SOAPdenovo-Trans**: developed at BGI. Has heavy memory requirements of SOAP tools (30 GB for a RNA-seq run).