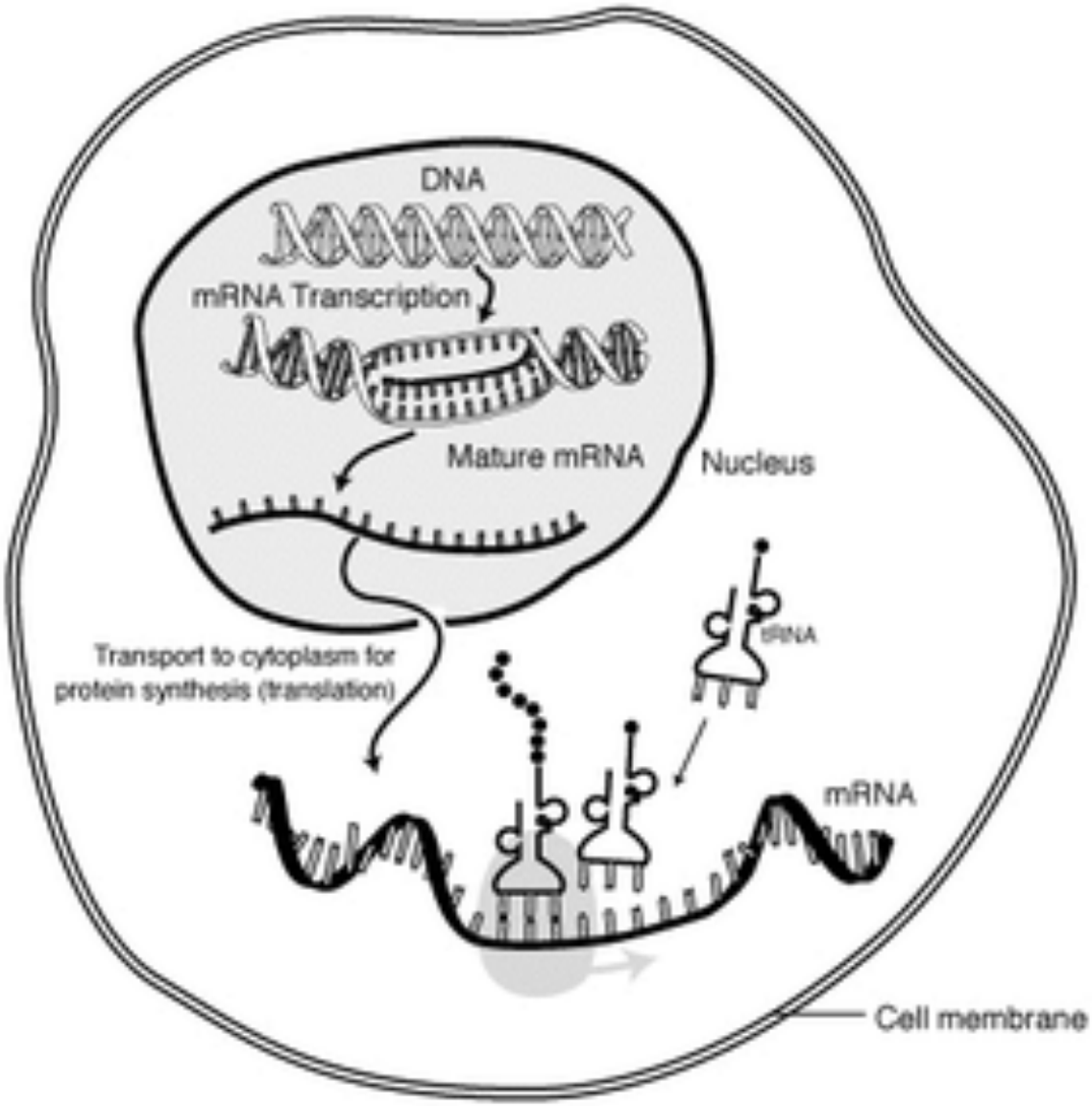
Analysis of RNA-seq Data

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: "Hellllooooooo! Where are we?"

- 15 minutes later, they hear an echoing voice: "Hellllooooooo! 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

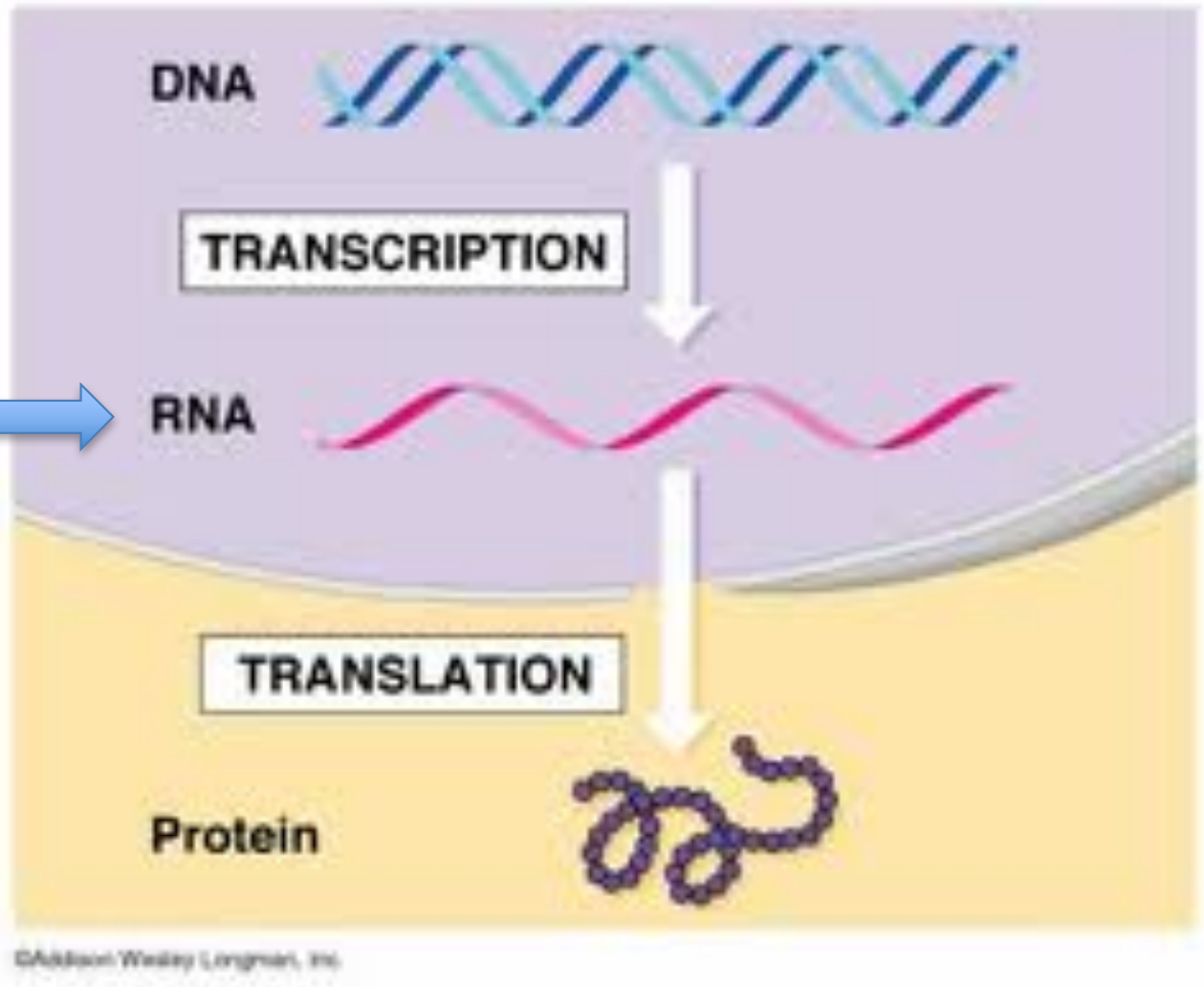

DNA
mRNA Transcription
Mature mRNA
Nucleus
Transport to cytoplasm for
protein synthesis (translation)
tRNA
mRNA
Cell membrane

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.



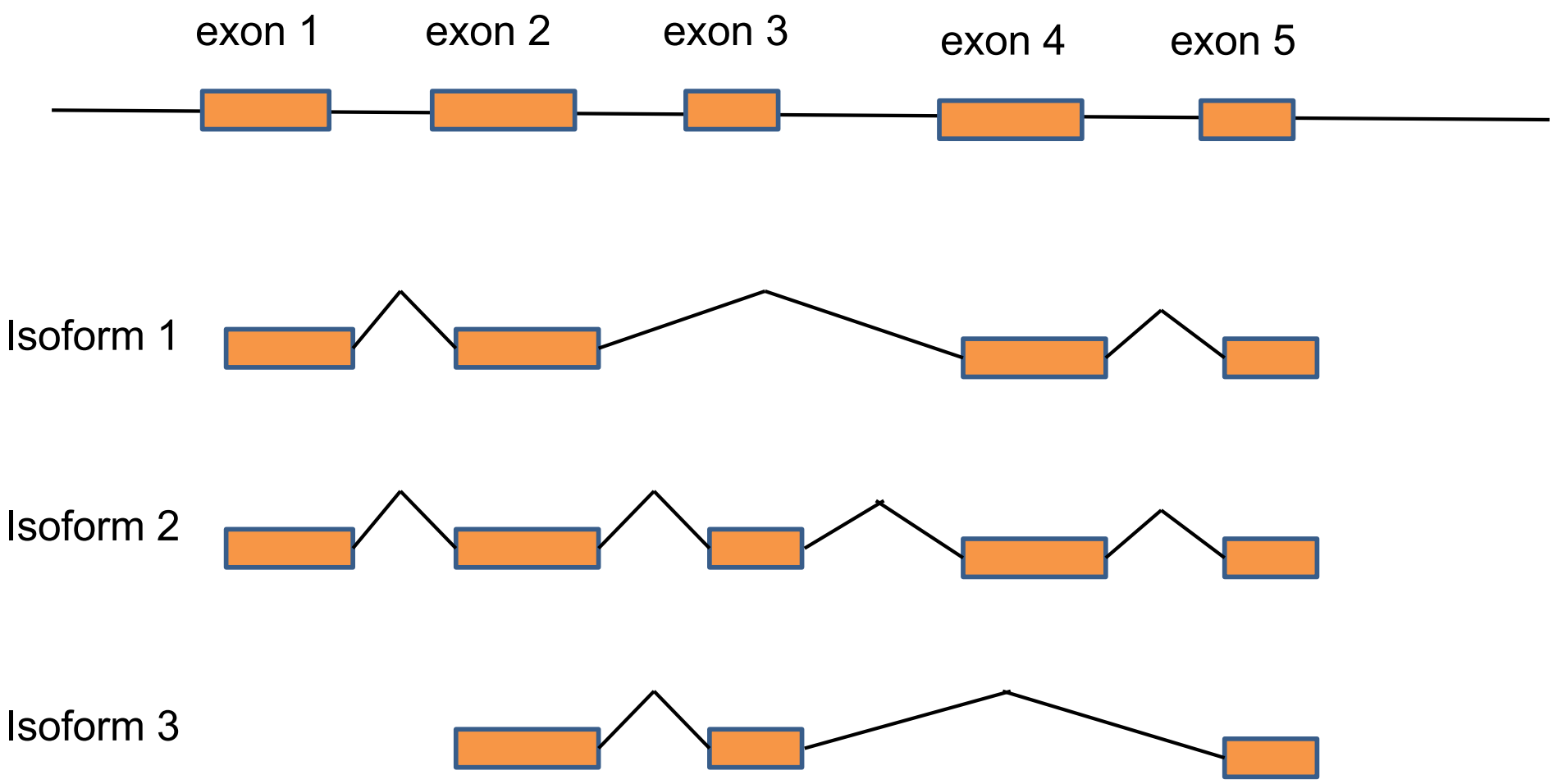
Splicing

- 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

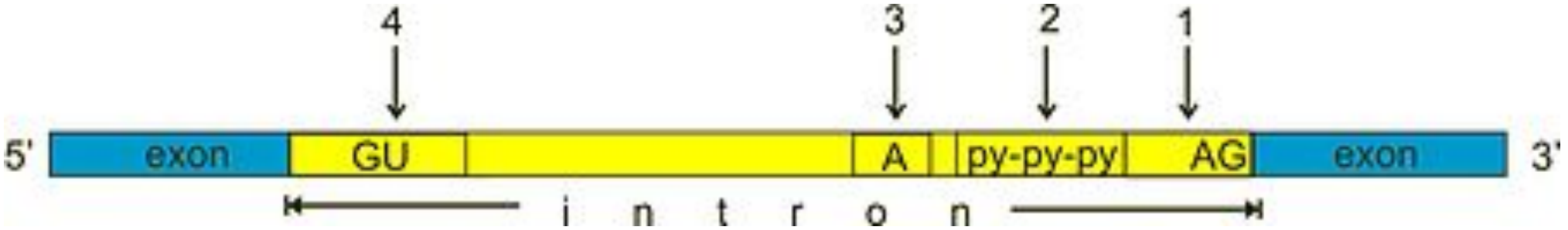
- Some introns or exons can be either removed or retained in mature mRNA.
- This is referred to as **alternative splicing** and it creates a series of different transcripts from a single gene.
- These different transcripts can be potentially translated into different proteins, splicing extends the complexity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

Alternative Splicing



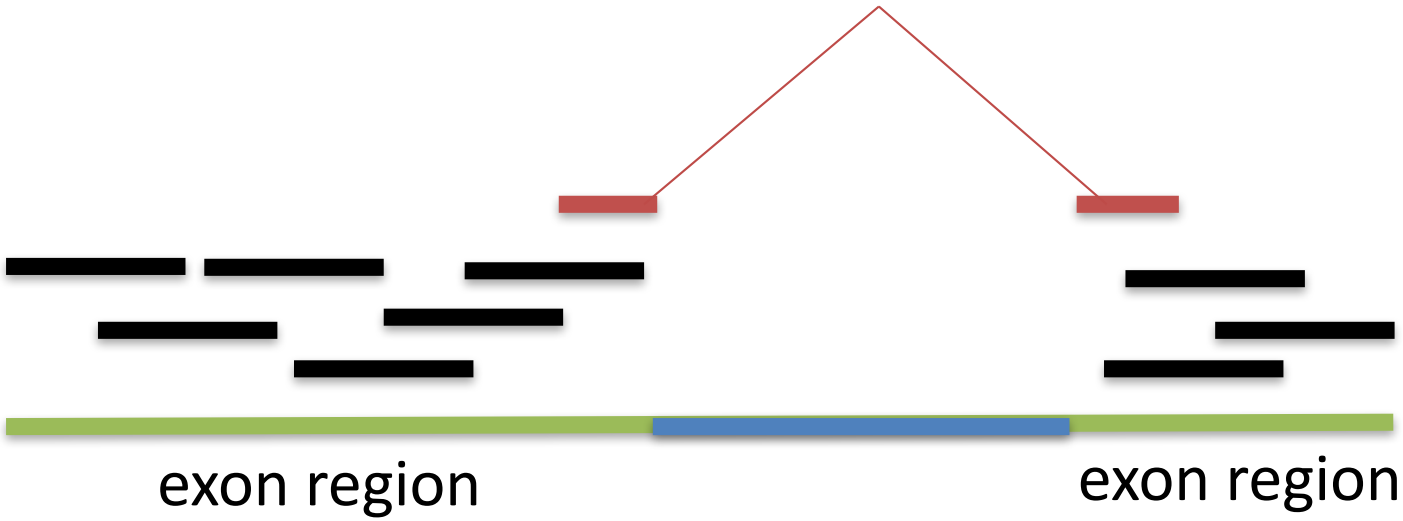
Alignment of RNA-seq Reads

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

exon region

- 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

A spliced read mapper for RNA-Seq

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.

OSI certified

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 5/26/2012

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