CS-E5875 High-Throughput Bioinformatics
RNA-seq analysis: alignment, assembly, and quantification

Harri Lähdesmäki

Department of Computer Science
Aalto University

November 7, 2023

Contents

Gene transcription

- A process of making an RNA copy of a gene sequence in DNA

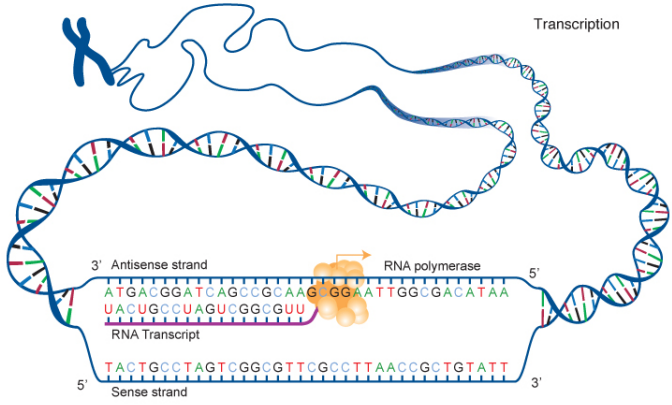

Figure from https://geneed.nlm.nih.gov/topic_subtopic.php?tid=15&sid=22

Splicing and alternative splicing

- E.g. in humans, genes consist of exons and introns
- RNA splicing: introns are generally spliced out from precursor RNA (pre-mRNA) molecules
- Alternative splicing: a process of making alternative mRNA molecules (called transcripts or isoforms) from the same precursor RNA (pre-mRNA): splice out specific exons
- In humans, $\sim$95% of multi-exonic genes are alternatively spliced

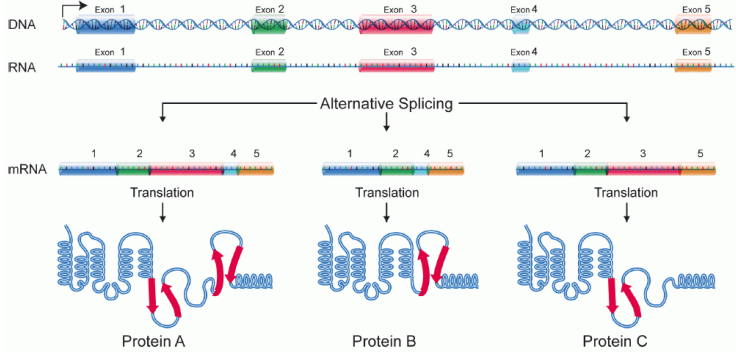

Figure from https://en.wikipedia.org/wiki/Alternative_splicing

Alternative splicing mechanisms

- Alternative splicing is largely regulated by splicing factors (proteins) that bind RNA motifs (short stretches of RNA) located in the pre-mRNA
- Alternative splicing happens co-transcriptionally

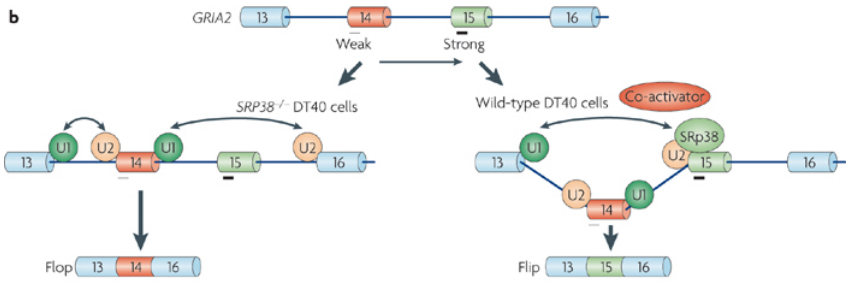

Figure from (Chen & Manley, 2009)

Different types of alternative splicing

- Basic modes of alternative splicing
- Multi-exonic genes can have complex splicing patterns

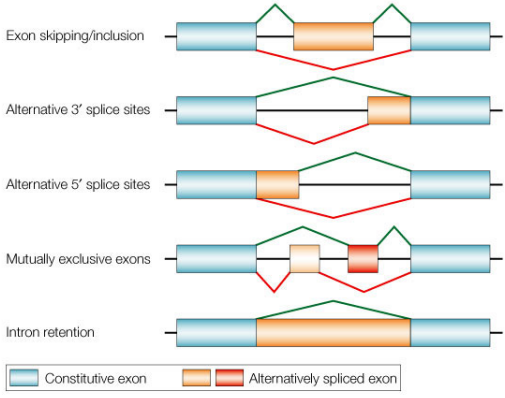

Figure from (Cartegni et al., 2002)

RNA-seq

- Types of RNA molecules in a cell

ribosomal RNA (ribosome, protein synthesis)	rRNA	∼85-90%
transfer RNA (adaptor, protein synthesis)	tRNA	∼10%
mRNA messenger (protein coding)	RNA	∼1-5%
micro RNA and other non-coding	miRNA, piRNA, etc.	rest

- High-throughput sequencing of RNA provides a comprehensive picture of the transcriptome
- RNA molecules are often enriched for mRNAs (poly-A tail) or non-coding RNAs (size selection)

RNA-seq: basic experimental protocol

1. RNA molecules are extracted from a (large) collection of cells: RNA molecules are pooled without knowing from which cell each RNA molecule comes from (bulk RNA sequencing)
2. RNA population is converted to a library of cDNA molecules, fragmented, and adaptors are attached to one or both ends
3. High-throughput sequencing for the cDNA fragment library (single-end or paired-end), read length $\sim$30-400 bp
4. Computational and statistical analysis: alignment against reference genome or transcriptome, transcriptome reconstruction, expression quantification, etc.

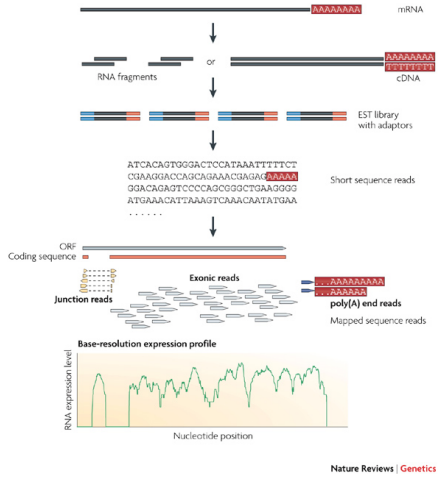

Figure from (Wang et al., 2009)

What can we do with RNA-seq data?

- Characterization of genes: location in genome, beginning/end, exons
- Transcript assembly
  - Identify transcript variants by constructing full-length spliced transcripts from the RNA-seq data (either with or without the knowledge of the reference genome)
- Transcript quantification
  - Given transcript sequence annotations (reference), estimate
    - Gene expression or
    - Abundances of all different transcripts
- Differential expression
  - Statistical inference for differential gene expression or alternative splicing
- And more!

Contents

RNA-seq read alignment with transcript or exon reference

- If full-length transcript annotations are known (see "Processed mRNA" below), then reads can be aligned exactly as aligning DNA sequence reads against a reference genome
  - Use transcripts in place of reference genome
- If transcript annotations are not known (but exons are known), still similar approaches as for aligning DNA sequence reads will work with some modifications
  - Transcriptomic reads can span exon junctions: alignments with gaps (or align against all possible exon configurations)
  - Transcriptomic reads can contain poly(A) ends (from post-transcriptional RNA processing)

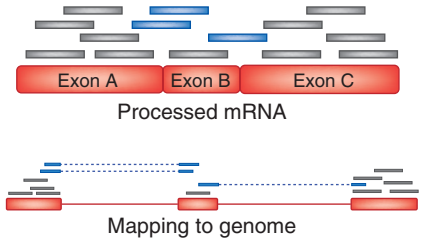

Figure from (Trapnell & Salzberg, 2009)

TopHat pipeline

- We will look at TopHat (Trapnell et al., 2009), a commonly used tool for RNA-seq alignment without transcript or exon reference
- All reads are mapped to the reference genome using Bowtie
  - These are sequencing reads that originate from individual exons, i.e., do not span exon-exon boundaries
- Reads that do not map to the genome are set aside as "initially unmapped reads" (IUM reads)
  - These are sequencing reads that potentially originate from a part of a transcript that connects two (or more) exons, i.e., span exon-exon boundaries

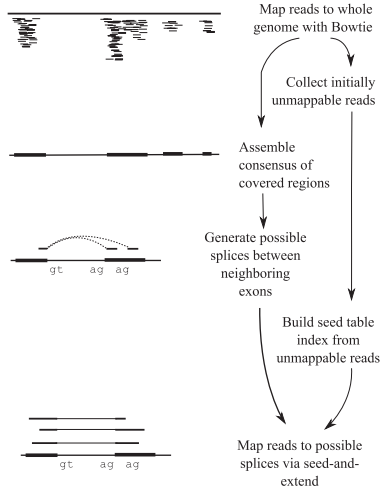

Figure from (Trapnell et al., 2009)

TopHat pipeline: consensus assembly

- Consensus assembly of initially mapped reads with Maq assembler
  - Similarly as in de novo assembly, partly overlapping short sequencing reads define the assembly (i.e., exons)
  - Note that in this case the overlaps between short reads have been found by aligning against the known reference
  - For low-quality or low-coverage positions, use reference genome to call the base
  - Consensus exons are likely missing some amount of sequence at ends

  → TopHat considers flanking sequences from reference genome (default=45bp)
- Merge neighboring exons with very short gap to a single exon

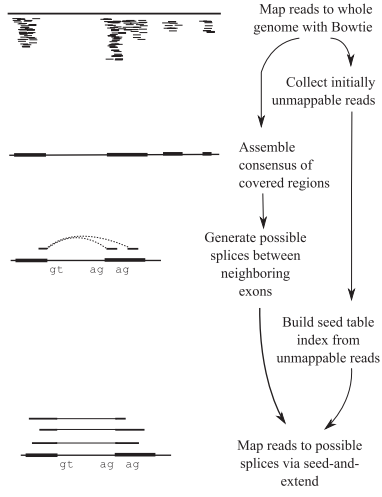

Figure from (Trapnell et al., 2009)

TopHat pipeline: splice junctions

- To map reads to splice junctions:
  - Enumerate all canonical donor and acceptor splicing sites (GT, AG, etc. di-nucleotides) between consecutive exons
  - Consider all possible pairings between donor-acceptor sites (allowed intron length is an adjustable parameter)
  - For each candidate splice junction, find initially unmapped reads that span them: seed-and-extend approach

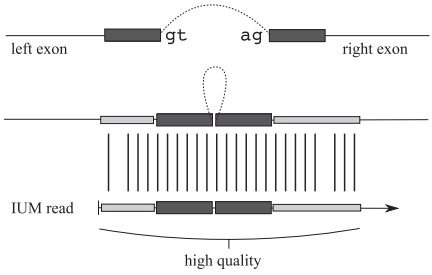

Figure from (Trapnell et al., 2009)

TopHat pipeline: seed-and-extend

- Seed-and-extend:
  - Pre-compute an index of reads: a lookup table based on partly overlapping $2k$-mer keys ($=2k$ long nucleotide subsequences) in the middle of their high-quality region (default $k = 5$)
  - For candidate splice junction, concatenate the k bases downstream of the acceptor to the k bases upstream
  - Query this $2k$-mer against the read index (exact seed match, no mismatch allowed)
  - Align remaining part of read left and right of the exact match (allowing fixed number of mismatches)

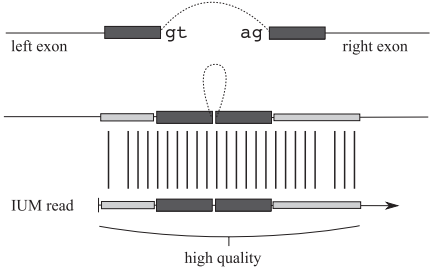

Figure from (Trapnell et al., 2009)

RNA-seq read alignment

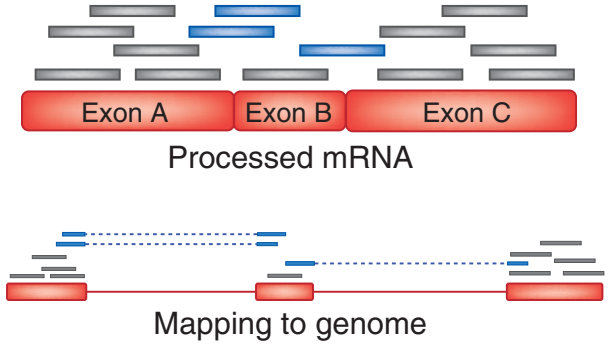

Figure from (Trapnell & Salzberg, 2009)

Contents

Simplified gene expression counting schemes

- Expression of a gene: sum of the expression of all its transcript variants / isoforms
  - Computing isoform abundances can be challenging
- Simplified counting schemes without computing isoform abundances
  - Exon union method: count sequencing reads mapped to any of the exons
  - Exon intersection method: count reads mapped to constitutive exons

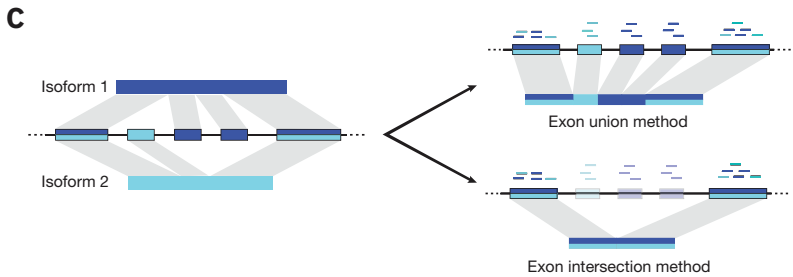

Figure from (Garber et al., 2011)

Simplified gene expression counting schemes

Disadvantages of the simplified models

- The union model tends to underestimate expression for alternatively spliced genes
  - Because it overestimates the length of isoforms: we will see the reason for this later
- The intersection can reduce statistical power for differential expression analysis
  - Because a fraction of mapped sequencing reads are ignored

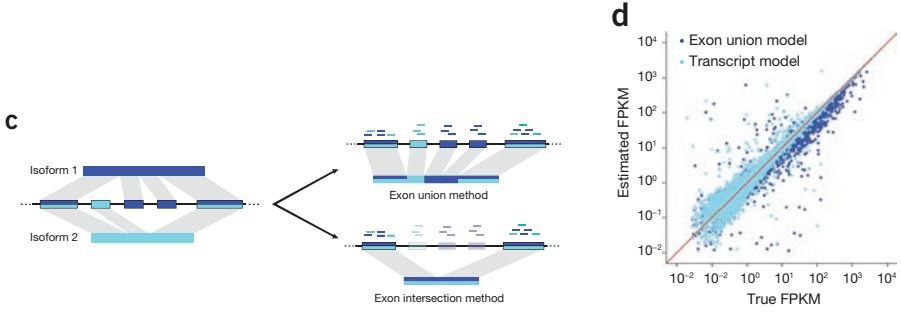

Figure from (Garber et al., 2011)

Gene expression quantification

- Basic idea: read count corresponds to the expression level
- Basic assumption

$$\theta_i = P(\text{“randomly sample a sequencing read from gene” } i) = \frac{1}{Z}\mu_i \ell_i,$$

where

- μ_i is the expression level (abundance) of gene i
- ℓ_i is the length of gene i (e.g. the total length of constitutive exons for the intersection method)
- Normalizing constant is $Z = \sum_i \mu_i \ell_i$

Gene expression quantification

- Use the so-called frequency estimator to estimate the probability that a read originates from a given gene i

$$\hat{\theta}_i = \frac{k_i}{N},$$

where

  - k_i is the number of sequencing reads mapping to gene i
  - N is the total number of mapped reads

- Convert the estimates into expression values by normalizing by the gene length
- Recall from the previous slide that $\theta_i = \frac{1}{Z}\mu_i\ell_i$, thus $\theta_i \propto \mu_i\ell_i$, which we can solve for μ_i

$$\hat{\mu}_i \propto \frac{\hat{\theta}_i}{\ell_i} = \frac{k_i}{N\ell_i}$$

RPKM: reads per kilobases per million reads

- The number of reads that map to a specific gene i depends on
  - The total number of mapped reads N
  - The length of the gene ℓ_i
- By normalizing with these two terms, N and ℓ_i, we obtain a common unit to quantify gene expression
  - Across different experiments that may have different N
  - Across different genes that may have different ℓ_i
- RPKM: reads per kilobases per million reads

$$\text{RPKM}_i = \frac{k_i}{\frac{\ell_i}{10^3} \cdot \frac{N}{10^6}} = 10^9 \frac{k_i}{\ell_i N} = 10^9 \hat{\mu}_i$$

- RPKM is for single-end reads
- FPMK is essentially the same as RPKM but defined for paired-end reads such that each read-pair is counted only once

Gene expression quantification: illustration

- Consider 4 transcripts with different lengths and expression levels illustrated below (left)
- The read counts normalized by the transcript length using the RPKM (or FPKM) metric (right)
  - Transcripts 2 and 4 have comparable read-counts, transcript 2 has a significantly higher normalized expression level
  - After normalization, transcripts 3 and 4 have similar expression values

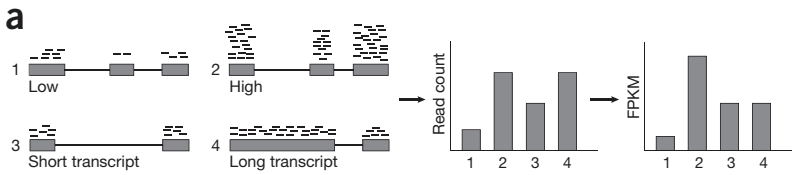

Figure from (Garber et al., 2011)

- When the same gene is compared between conditions, the read counts normalized by sequencing depth (but not by gene length) are just fine

Gene expression quantification

- The above formulation assumes that all reads can be assigned uniquely to a single gene
- That is not always true
  - E.g. genes belonging to the same gene families have similar genome/RNA sequence, which may cause mis-alignments
  - Different genes can be located in the same genomic region but on opposite DNA strands (strand specific RNA-seq resolves this issue)
- Unique alignment is typically not true for transcripts
  - Different transcript isoforms can share a large fraction of their exons

Contents

Transcriptome assembly

- TopHat pipeline can identify exons and exon-exon junctions, but does not output the full-length transcripts
- Goal: define precise map of all transcript variants / isoforms that are expressed in a particular sample
- Challenges
  - For short reads, hard to determine from which isoform they were produced, because isoforms contain the same exons and exon-exon pairs
  - Gene expression spans several orders of magnitude, with some genes represented by only few reads (lowly expressed are more difficult)
  - Majority of the sequencing reads typically originate from mature mRNA, but a (small) fraction of the reads can originate from incompletely spliced precursor RNA (with introns included)
- Two main classes of methods
  - Genome-independent (de Brujin graph, see previous lecture)
  - Genome-guided, i.e., using reads that are already aligned against a reference genome

Paired-end sequencing reads

- Paired-end sequencing technology quantifies the nucleotide content of genomic DNA or cDNA (for RNA) fragments from both ends of the fragments

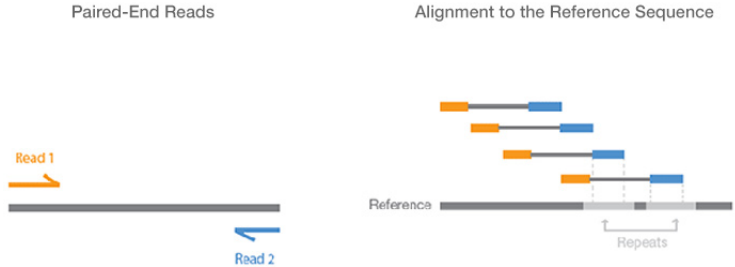

Figure from https://emea.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html

Transcriptome reconstruction with Cufflinks

- Genome-guided: takes TopHat spliced alignments as input
- With paired-end RNA-seq data, Bowtie and TopHat align both reads separately but aligned paired reads are treated together as a single alignment

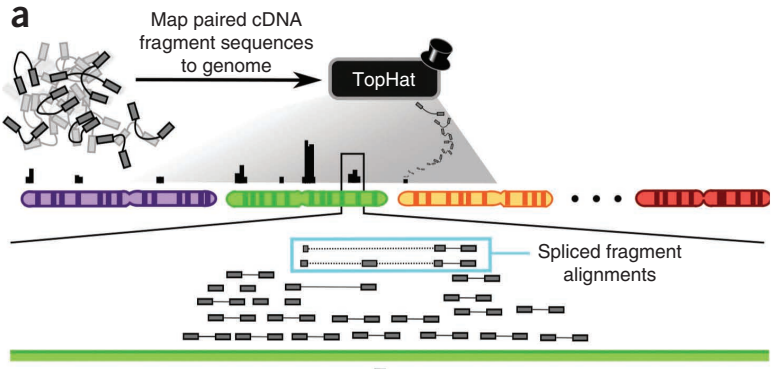

Figure from (Trapnell et al., 2010)

Transcriptome reconstruction with Cufflinks

- Connect fragments in an overlap graph
- Each fragment (read pair) corresponds to a node
- Directed edge from node x to node y if
  - The alignment for x starts at a lower coordinate than y
  - The alignments overlap in the genome, and
  - The fragments are "compatible", i.e., the fragments x and y can come from the same transcript isoform

- If two reads originate from different isoforms they are likely incompatible

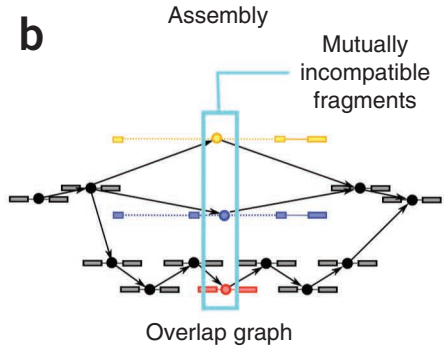

Figure from (Trapnell et al., 2010)

Transcriptome reconstruction with Cufflinks

- From the overlap graph, construct minimal set of transcript isoforms that can explain all the fragments
  - Minimum path cover problem
- Dilworth's theorem: maximum number of mutually incompatible fragments equals minimum number of paths covering the whole graph (=minimum number of transcripts needed to explain all the fragments)

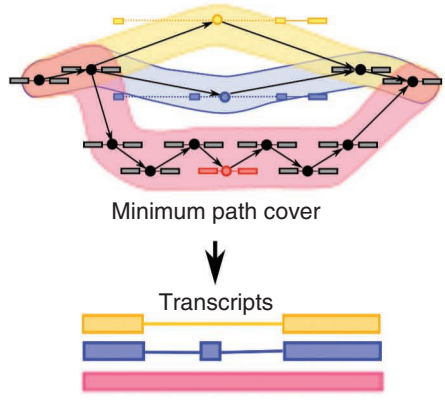

Figure from (Trapnell et al., 2010)

References

- Cartegni, L., Chew, S. L., & Krainer, A. R. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nature Reviews Genetics 3, 285-298 (2002)
- Mo Chen & James L. Manley, Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nature Reviews Molecular Cell Biology 10, 741-754, 2009.
- Manuel Garber, et al., Computational methods for transcriptome annotation and quantification using RNA-seq, Nature Methods 8, 469-477, 2011.
- Katz Y, et al., Analysis and design of rnA sequencing experiments for identifying isoform regulation, Nature Methods, 7(12):1009-15, 2010.
- Cole Trapnell, Lior Pachter and Steven L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, Bioinformatics, 25(9):1105-1111, 2009.
- Cole Trapnell & Steven L Salzberg, How to map billions of short reads onto genomes, Nature Biotechnology 27, 455-457, 2009.
- Cole Trapnell, et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nature Biotechnology 28, 511-515, 2010.
- Zhong Wang, Mark Gerstein & Michael Snyder, RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10, 57-63, 2009.
- McCarthy, D.J., Chen, Y., and Smyth, G.K., Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research 40, 4288-4297, 2012
- Chen Y, et al., edgeR: differential expression analysis of digital gene expression data, User?s Guide, 11 October 2017.