Alternative splicing

Mikhail Dozmorov

Spring 2018

Transcriptome quantification goals

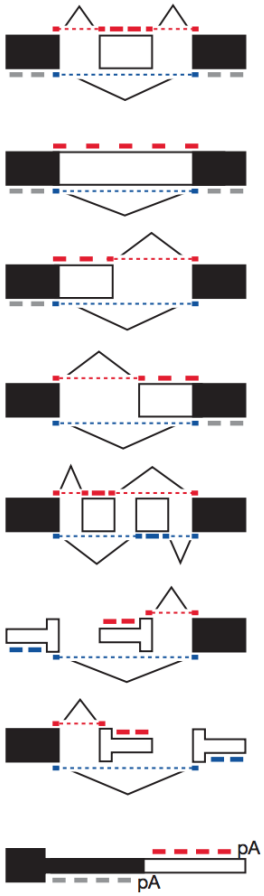
- Gene expression level estimation
  - genome-wide gene expression level estimates derived from isoform level estimates are significantly more accurate than those obtained directly from RNA-Seq data using isoform-oblivious GE methods
- Isoform expression level (abundance) estimation
- Novel isoform discovery

Alternative splicing

- **Definition:** the same pre-mRNA produces different mRNA products, through joining different exons.
- Locations where two exons join is called "junction".
- Can be detected and quantified using exon arrays, on which the probes are designed to target the junction regions.
- From RNA-seq: look at "junction reads", which are reads overlapping two exons.

Alternative splicing

Alternative transcript events	Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon	37	35	10,436	6,822	65	72
Retained intron	1	1	167	96	57	71
Alternative 5′ splice site (A5SS)	15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)	17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)	4	4	167	95	57	66
Alternative first exon (AFE)	14	13	10,281	5,311	52	63
Alternative last exon (ALE)	9	8	5,246	2,491	47	52
Tandem 3′ UTRs	7	7	5,136	3,801	74	80
Total	105	100	37,782	22,657	60	68



Constitutive exon or region — Body read — Junction read — pA Polyadenylation site

Alternative exon or extension — Inclusive/extended isoform — Exclusive isoform — Both isoforms

Wang, Eric T., Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. "Alternative Isoform Regulation in Human Tissue Transcriptomes." Nature 456, no. 7221 (November 27, 2008): 470–76. https://doi.org/10.1038/nature07509.

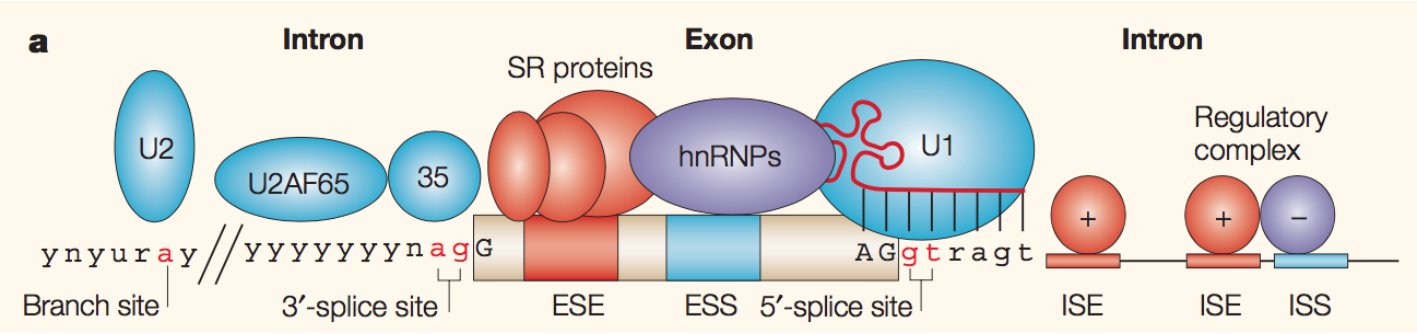
Common types of alternative transcript events

- Skipped exons and retained introns, in which a single exon or intron is alternatively included or spliced out of the mature message
- Alternative 5' splice site (A5SS) and alternative 3' splice site (A3SS) events, which are particularly difficult to interrogate by microarray analysis because the variably included region is often quite small
- Tandem 3' untranslated regions (UTRs) and alternative last exons (ALEs), in which alternative use of a pair of polyadenylation sites results in shorter or longer 3' UTR isoforms or in distinct terminal exons, respectively
- Alternative first exons (AFEs), in which alternative promoter use results in mRNA isoforms with distinct 5' UTRs.

Splicing factor motifs

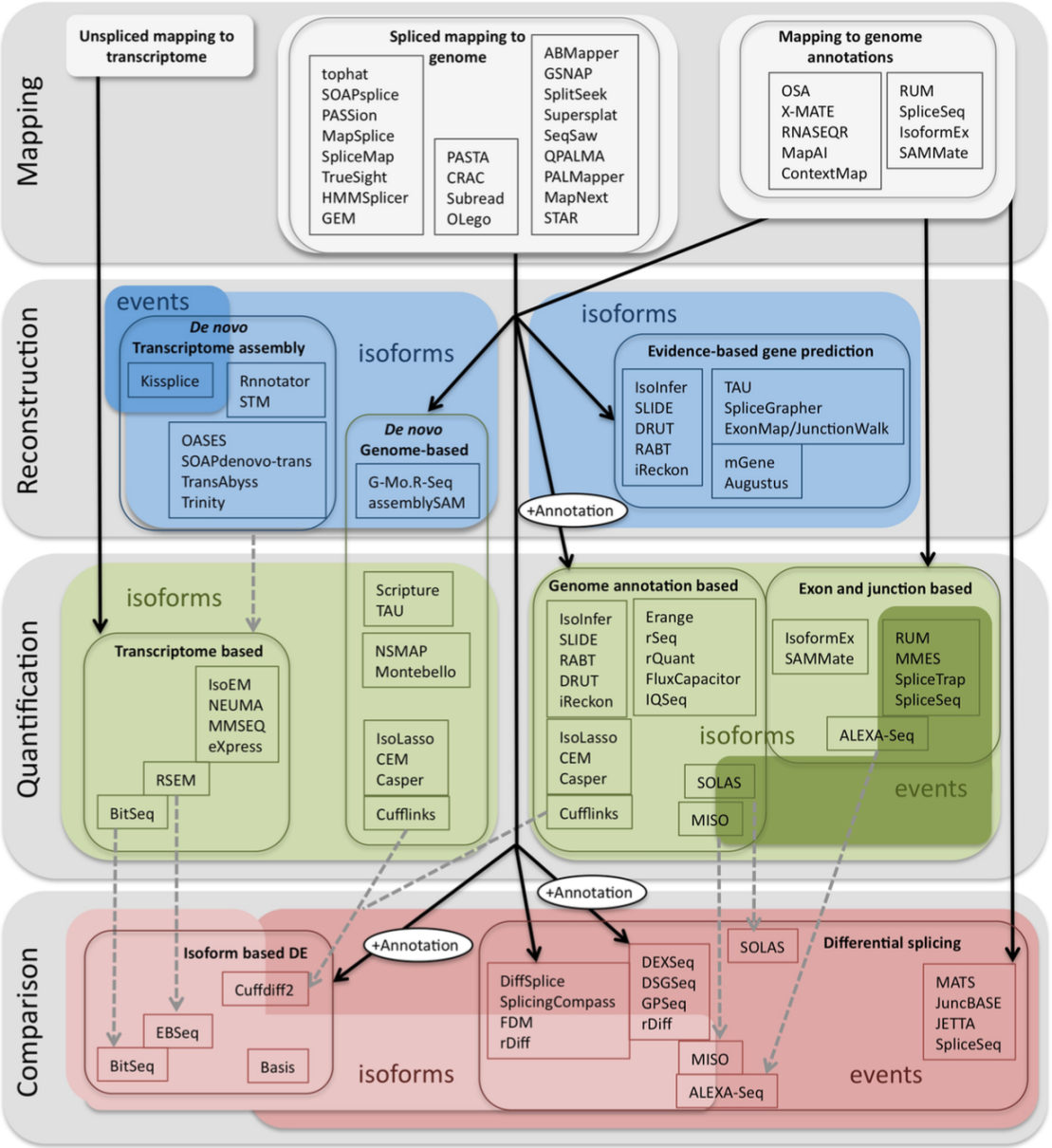
- During splicing, a protein complex known as the spliceosome assembles on the pre-mRNA to remove introns and join exons.
- This process is guided by short consensus motifs at the ends of introns called splice sites.
- Cis-acting elements can function as silencers or enhancers and are found in the vicinity of splice sites in introns and exons.
- In general, alternative splicing is determined by the combined effect of multiple positively and negatively acting elements, and the fate of cassette exons is decided by the presence and arrangement of surrounding motifs as well as the condition- specific ratio and modification status of splicing factor proteins

Splicing elements



- The GU and the AG dinucleotides that directly flank the exon (at the 3' and 5' ends, respectively) and the branch-point adenosine (all in red) are always conserved.
- In most cases, there is also a polypyrimidine tract of variable length (the consensus symbol 'y' represents a pyrimidine base — cytosine or thymine) upstream of the 3'-splice site.
- The branch point is typically located 18–40 nucleotides upstream from the polypyrimidine tract.
- Exon/Intron Splicing Enhancer (ESE/ISE), Exon/Intron Splicing Silencer (ESS/ISS) allow the correct splice sites to be distinguished

Alternative splicing workflow



Alamancos, G. et.al. "Methods to Study Splicing from High-Throughput RNA Sequencing Data." Spliceosomal Pre-mRNA Splicing: Methods and Protocols, 2014 https://www.ncbi.nlm.nih.gov/pubmed/

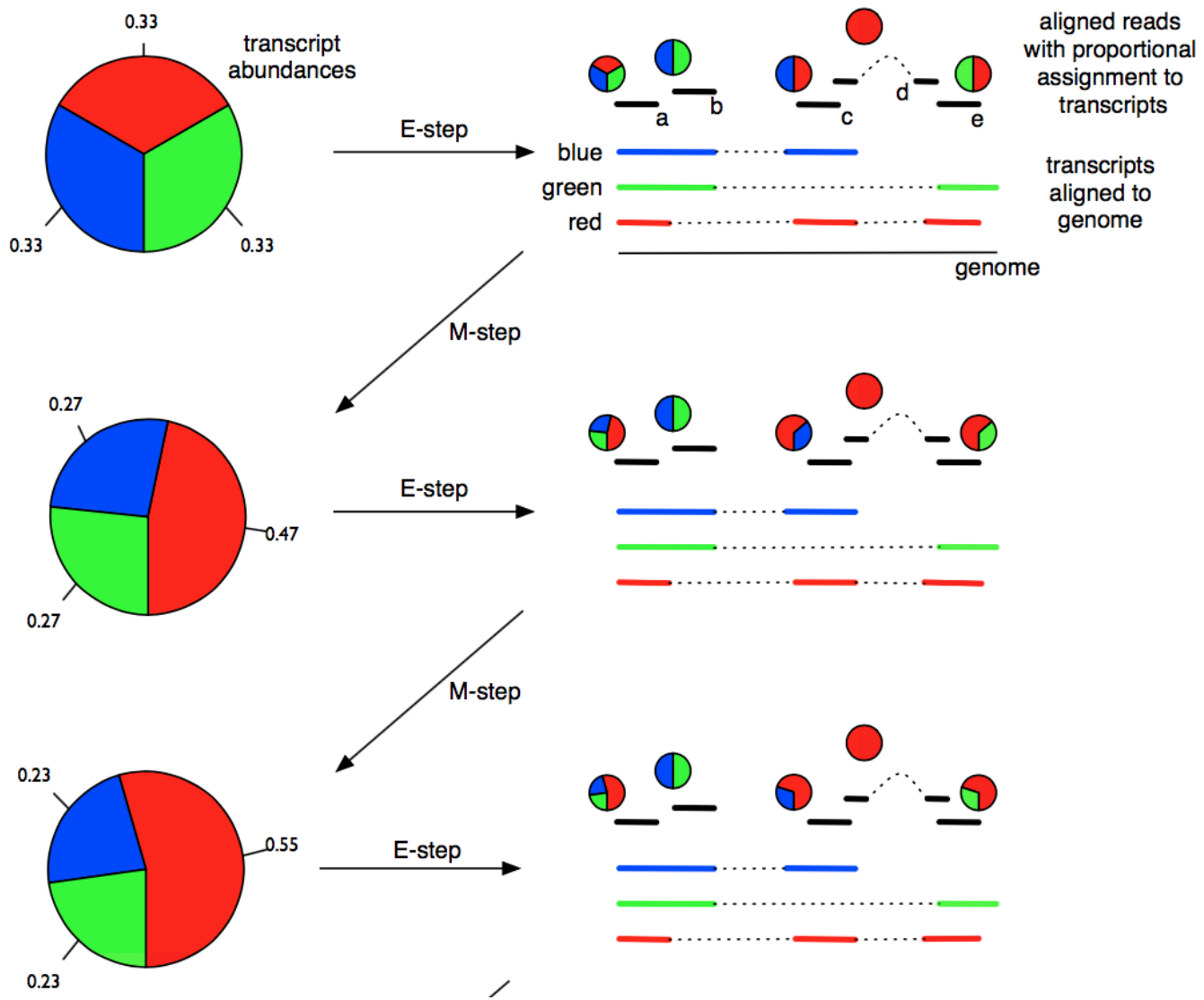
Estimate isoform expressions

- **Isoform:** different transcripts from the same gene, caused by alternative splicing.
- Different isoforms could have different expression levels.
- A toy example for a gene with 3 exons:
- It was known the gene has two isoforms: exon1+exon2, and exon1+exon3.
- The read counts from the exons are 10, 7, 5.
- What are the expression level for the two isoforms?

Isoform abundance estimation

- We have an unobserved variable (expression) that we wish to estimate
  - Set up a model and estimate it using the expectation-maximization (EM) algorithm
- Step 1: (Expectation) Given some abundances, estimate the probability of each read mapping to each transcript
- Step 2: (Maximization) Update the abundances by redistributing the reads
- Step 3: Repeat until convergence

Expectation-Maximization algorithm



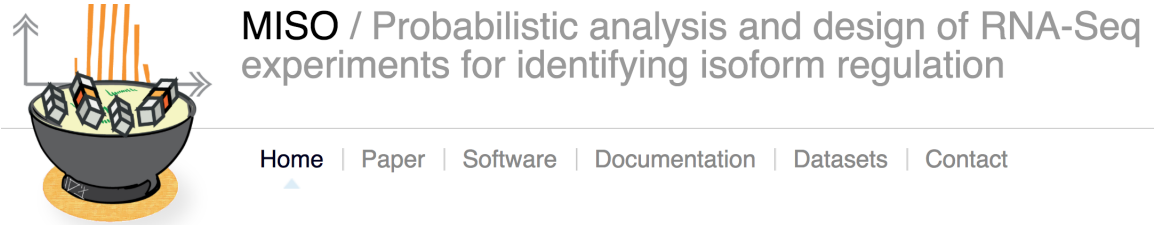
Pachter, Lior. "Models for Transcript Quantification from RNA-Seq." ArXiv Preprint ArXiv:1104.3889, 2011. https://arxiv.org/abs/1104.3889

Other EM approaches

- Underlying Poisson rate of counts is a linear combination of isoform expressions, then derive joint data likelihood.
- Compute MLE for the isoform expressions by maximizing Joint likelihood through numerical methods.

Jiang, Hui, and Wing Hung Wong. "Statistical Inferences for Isoform Expression in RNA-Seq." Bioinformatics 25, no. 8 (April 15, 2009): 1026–32. https://doi.org/10.1093/bioinformatics/btp113.

MISO



- MISO (Mixture-of-Isoforms) is a probabilistic framework that quantitates the expression level of alternatively spliced genes from RNA-Seq data, and identifies differentially regulated isoforms or exons across samples.
- By modeling the generative process by which reads are produced from isoforms in RNA-Seq, the MISO model uses Bayesian inference to compute the probability that a read originated from a particular isoform.
- MISO treats the expression level of a set of isoforms as a random variable and estimates a distribution over the values of this variable.
- The estimation algorithm is based on sampling, and falls in the family of techniques known as Markov Chain Monte Carlo ("MCMC")

MISO

- Estimates of isoform expression (ψ values, for "Percent Spliced In" or "Percent Spliced Isoform") and differential isoform expression for single-end or paired-end RNA-Seq data
- Expression estimates at the alternative splicing event level ("exon-centric" analysis) or at the whole mRNA isoform-level ("isoform-centric" analysis)
- Confidence intervals for expression estimates and quantitative measures of differential expression ("Bayes factors")

http://genes.mit.edu/burgelab/miso/index.html

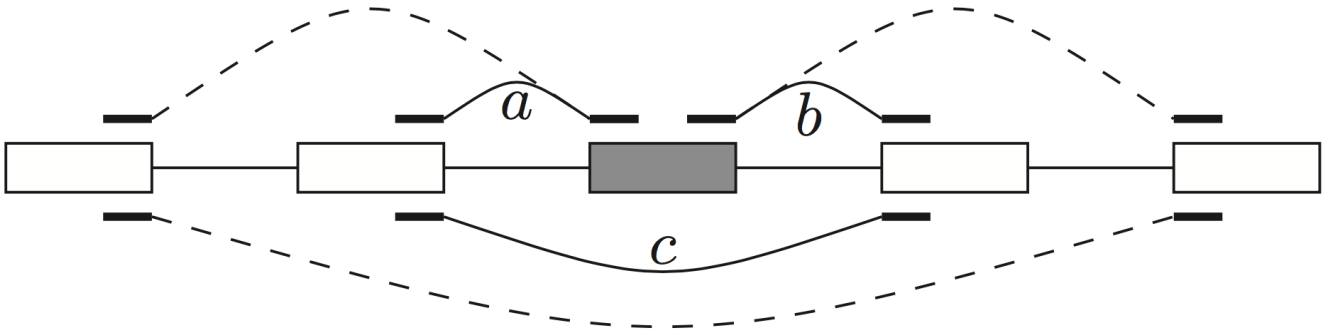
https://miso.readthedocs.io/en/fastmiso/index.html

Katz, Yarden, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. "Analysis and Design of RNA Sequencing Experiments for Identifying Isoform Regulation." Nature Methods 7, no. 12 (December 2010): 1009–15. https://doi.org/10.1038/nmeth.1528.

Percent spliced-in metric

- The percent-spliced-in (PSI, ψ) metric estimates the incidence of single-exon–skipping events and can be computed directly by counting reads that align to known or predicted splice junctions.
- ψ metric is defined as the number of reads supporting exon inclusion (a + b) as the fraction of the combined number of reads supporting inclusion and exclusion (c).

$$\psi = \frac{a+b}{a+b+2c}$$



Intron-centric estimation of alternative splicing

- The majority of human splicing events are more complex than single-exon skipping
- Split the value of ψ into two indices, ψ_5 and ψ_3, measuring the rate of splicing at the 5' and 3' end of the intron, respectively
- Each intron is defined uniquely by the combination of its 5'-splice site (D, donor) and 3'-splice site (A, acceptor)

Intron-centric estimation of alternative splicing

- $n(D, A)$ the number of reads aligning to the splice junction spanning from D to A

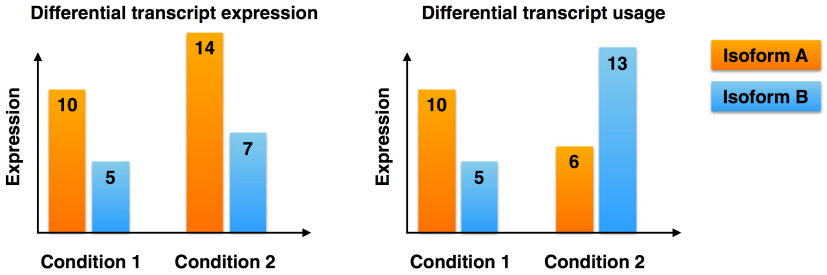
$$\psi_5(D, A) = \frac{n(D, A)}{\sum_{A'} n(D, A')}, \quad \psi_3(D, A) = \frac{n(D, A)}{\sum_{D'} n(D', A)}$$

- D' and A' run over all donor and acceptor sites, respectively, within the given genomic annotation set

Pervouchine, Dmitri D., David G. Knowles, and Roderic Guigó. "Intron-Centric Estimation of Alternative Splicing from RNA-Seq Data." Bioinformatics (Oxford, England) 29, no. 2 (January 15, 2013): 273–74. https://doi.org/10.1093/bioinformatics/bts678.

Differential Transcript Usage (DTU)

DTU considers changes in the proportions of the isoforms of a gene that are expressed as opposed to changes of the individual transcript levels.



- DTE implies that we can observe expression changes for at least one transcript between condition 1 and condition 2. However, the expression proportion of each transcript (as a percentage of the total expression of all transcripts of the same gene) does not necessarily change between conditions, and thus DTE does not necessarily imply DTU
- In DTU, on the other hand, the relative expression of the isoforms of a gene changes between the conditions, whereas the total expression of the gene may or may not remain constant
- Since at least one isoform must change expression in DTU, it also implies DTE

Methods to detect DTU

1. The assembly-based (or isoform deconvolution) methods (e.g., `cufflinks/cuffdiff`) reconstruct and quantify the expression of a set of transcripts that best explain the observed reads.
2. The second class of methods focuses on specific types of alternative splicing (e.g., retained introns or alternative exons) and identifies the number of observed reads that unambiguously support the presence or absence of each splicing event (e.g., `rMATS`)
3. The third type of DTU detection methods do not directly quantify the transcript expression, but rather use differential exon usage as a surrogate to infer DTU (`DEXSeq2`)

The Tuxedo Suite: `Bowtie`, `TopHat`, `cufflinks`, and `cuffdiff`

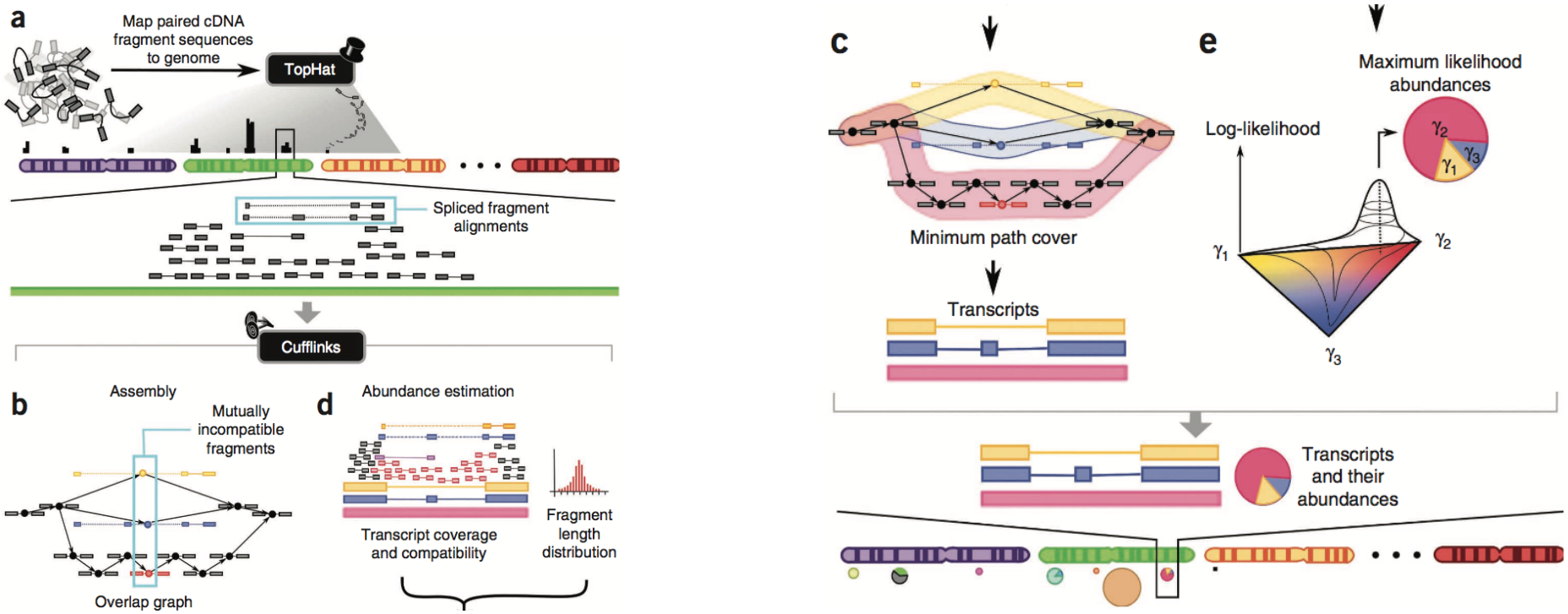
- Developed by Steven Salzberg's group at Hopkins
- `Bowtie` - alignment
- `TopHat` - alignment to exon junctions
- `cufflink` - estimate isoform expressions
- `cuffdiff` - estimate differential isoform expression
- The `cuffdiff` test for DTU within a gene is based on the Jensen–Shannon divergence, measuring the similarity between two probability distributions

TopHat: a spliced read mapper for RNA-seq

- Based on Bowtie, aligns RNA-Seq reads to a genome in order to identify exon-exon splice junctions.
- Runs on Linux and Mac OSX
- Command: `tophat -o out_dir -G known_genes.gtf --library-type fr-firststrand --mate-inner-dist 124 -p 8 --transcriptome-index bowtid_index isample1_read1.fastq sample1_read2.fastq`
- Output:
  - `accepted_hits.sam` - read alignments in SAM format.
  - `junctions.bed` - junction reads in BED format.
  - `insertions.bed` - BED track of insertions
  - `deletions.bed` - BED track of deletions

Cufflinks

A product of Bernoulli model with multivariate normal prior, then use Bayesian method to report maximum a posteriori (MAP).



Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. "Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation." Nature Biotechnology 28, no. 5 (May 2, 2010): 511–15. https://doi.org/10.1038/nbt.1621.

Use Cufflinks

- Runs on Linux or Mac OSX
- Input is alignment result from TopHat.
- Command: `cufflinks -o output_dir --library-type fr-firststrand -p 8 -G genes.gtf -b genome.fa -M rRNA.tRNA.gtf -u --compatible-hits-norm accepted_hits.bam`
- Output:
  - `transcripts.gtf`
  - `genes.fpkm_tracking`
  - `isoforms.fpkm_tracking`
  - `skipped.gtf`

Merge annotation information with `cuffmerge`

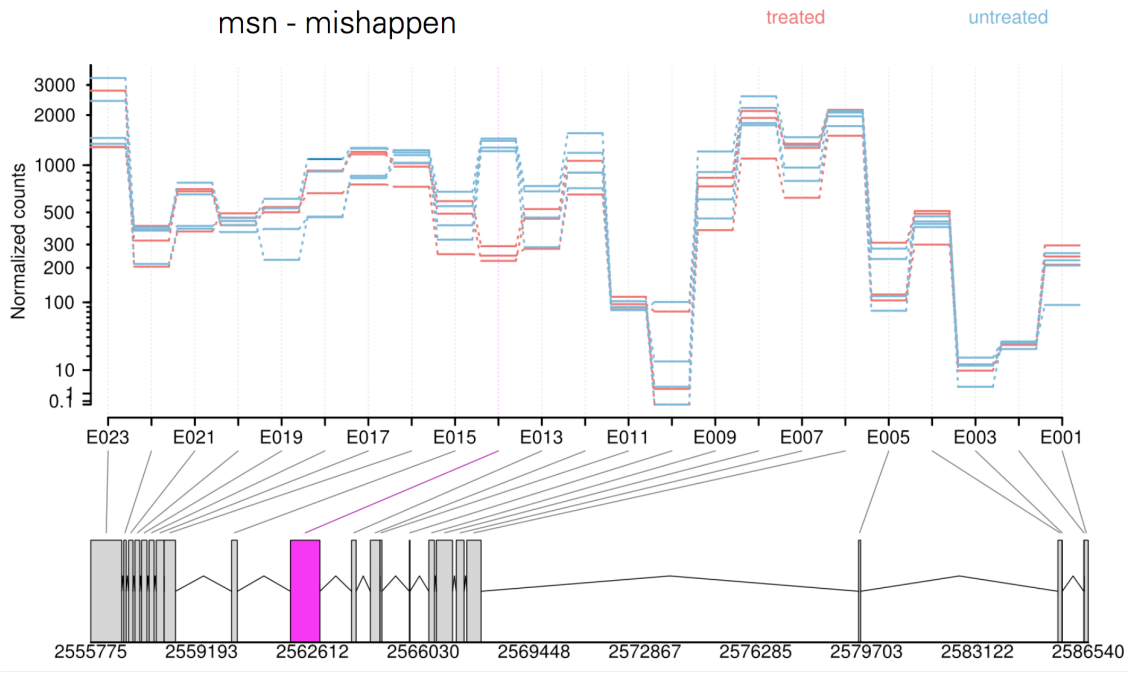
- The `cuffmerge` command takes gtf files that were generated by `cufflinks` and merges them into one combined file.
- This combined file can be used for differential expression testing in `cuffdiff`
- Create a text file with gtf file names to be merged, e.g. `mygtfs.txt`
- Command: `cuffmerge -p 8 -o merged.gtf compare -g genes.gtf -s genomefasta mygtfs`
- Output: `merged.gtf` file

Testing for differential expression with `cuffdiff`

- Command: `cuffdiff --library-type fr-firststrand -o cuffdiff_output -p 2 -b genome.fa -u -L sample1,sample2 -M rRNA.tRNA.gtf merged.gtf sample1.bam sample2.bam`
- Takes a long time (>12 hours) to run
- Output: many files, `gene_exp.diff` has p-values and q-values.

DEXSeq - differential isoform usage

- Test for changes in the (relative) usage of exons: (number of reads mapping to the exon) / (number of reads mapping to the other exons of the same gene)



Anders, S., A. Reyes, and W. Huber. "Detecting Differential Usage of Exons from RNA-Seq Data." Genome Research 22, no. 10 (October 1, 2012): 2008–17. https://doi.org/10.1101/gr.133744.111.

https://bioconductor.org/packages/release/bioc/html/DEXSeq.html

DEXSeq - differential isoform usage

- Negative binomial and generalized linear model of log mean. Cox-Reid dispersion estimator

$$K_{ijl} \sim \mathrm{NB}(s_j \mu_{ijl}, \alpha_{il})$$

counts in gene *i*, sample *j*, exon *l*

size factor

dispersion

$$\log \mu_{ijl} = \beta_i^0 + \beta_{il}^{\mathrm{E}} x_l^{\mathrm{E}} + \beta_{ij}^{\mathrm{T}} x_j^{\mathrm{T}} + \beta_{ijl}^{\mathrm{ET}} x_l^{\mathrm{E}} x_j^{\mathrm{T}}$$

expression strength in control

fraction of reads falling onto exon *l* in control

change in expression of gene *i* due to treatment

change to fraction of reads for exon *l* due to treatment

Anders, S., A. Reyes, and W. Huber. "Detecting Differential Usage of Exons from RNA-Seq Data." Genome Research 22, no. 10 (October 1, 2012): 2008–17. https://doi.org/10.1101/gr.133744.111.

https://bioconductor.org/packages/release/bioc/html/DEXSeq.html

EBseq - an empirical Bayes hierarchical model for inference in RNA-seq experiments

- Differential expression of isoforms, genes
- Negative binomial distribution of the expected counts for isoform i in gene g and sample s, condition C, library size l_s - $X^C_{gi,s}|r_{gi,0}, l_s, q^C_{gi} \sim NB(r_{gi,0}, l_s, q^C_{gi})$
- Mean $\mu^C_{gi} = r_{gi,0}(1 - q^C_{gi})/q^C_{gi}$, variance $(\sigma^C_{gi})^2 = r_{gi,0}(1 - q^C_{gi})/(q^C_{gi})^2$
- A prior distribution describes fluctuations in technical and biological variation $q^C_{gi}|\alpha\beta^{lg} \sim Beta(\alpha\beta^{lg})$, where hyperparameter α is shared across isoforms and β depends on l_g accommodating the systematic differences in variability among l_g groups, obtained via EM algorithm

Leng, Ning, John A. Dawson, James A. Thomson, Victor Ruotti, Anna I. Rissman, Bart M. G. Smits, Jill D. Haag, Michael N. Gould, Ron M. Stewart, and Christina Kendziorski. "EBSeq: An Empirical Bayes Hierarchical Model for Inference in RNA-Seq Experiments." Bioinformatics (Oxford, England) 29, no. 8 (April 15, 2013): 1035–43. https://doi.org/10.1093/bioinformatics/btt087.

https://www.biostat.wisc.edu/~kendzior/EBSEQ/

EBseq - an empirical Bayes hierarchical model for inference in RNA-seq experiments

- Differential isoform expression corresponds to $\mu_{gi}^{C1} \neq \mu_{gi}^{C2}$, so $q_{gi}^{C1} \neq q_{gi}^{C2}$ since $r_{gi,0}$ is common across conditions
- Given p is the prior probability of differential expression, counts are modeled as $(1-p)f_0^{lg}(X_{gi}^{C1,C2}) + pf_1^{lg}(X_g i^{C1,C2})$, where $X_g i^{C1,C2}$ represents g_i's read counts across the two conditions, f_0 and f_1 are the predictive distributions under equal and differential expression, respectively

Leng, Ning, John A. Dawson, James A. Thomson, Victor Ruotti, Anna I. Rissman, Bart M. G. Smits, Jill D. Haag, Michael N. Gould, Ron M. Stewart, and Christina Kendziorski. "EBSeq: An Empirical Bayes Hierarchical Model for Inference in RNA-Seq Experiments." Bioinformatics (Oxford, England) 29, no. 8 (April 15, 2013): 1035–43. https://doi.org/10.1093/bioinformatics/btt087.

https://www.biostat.wisc.edu/~kendzior/EBSEQ/

EBseq - an empirical Bayes hierarchical model for inference in RNA-seq experiments

- f_0 and f_1 are the predictive distributions under equal and differential expression, respectively

$$f_0^{I_g}(X_{g_i}^{C1,C2}) = \left[\prod_{s=1}^{S}\binom{X_{g_i,s}+r_{g_i,s}-1}{X_{g_i,s}}\right] \times \frac{Beta\left(\alpha+\sum_{s=1}^{S} r_{g_i,s}, \beta^{I_g}+\sum_{s=1}^{S} X_{g_i,s}\right)}{Beta(\alpha,\beta^{I_g})}$$

$$f_1^{I_g}(X_{g_i}^{C1,C2}) = f_0^{I_g}(X_{g_i}^{C1}) f_0^{I_g}(X_{g_i}^{C2})$$

Leng, Ning, John A. Dawson, James A. Thomson, Victor Ruotti, Anna I. Rissman, Bart M. G. Smits, Jill D. Haag, Michael N. Gould, Ron M. Stewart, and Christina Kendziorski. "EBSeq: An Empirical Bayes Hierarchical Model for Inference in RNA-Seq Experiments." Bioinformatics (Oxford, England) 29, no. 8 (April 15, 2013): 1035–43. https://doi.org/10.1093/bioinformatics/btt087.

https://www.biostat.wisc.edu/~kendzior/EBSEQ/

Summary for isoform expression

- Mostly for known isoforms (the combination patterns of exons).
- Similar strategies are used for gene fusion detection
- MLE approaches for estimation.

Alternative splicing

- How to predict novel and alternative splicing events from RNA-seq data
  - https://www.biostars.org/p/68966/
  - https://www.biostars.org/p/62728/
- How to detect alternative splicing
  - https://www.biostars.org/p/65617/
  - https://www.biostars.org/p/11695/
- Identifying genes that express different isoforms in cancer vs normal RNA-seq data
  - https://www.biostars.org/p/50365/
- Visualization of alternative splicing events using RNA-seq data
  - https://www.biostars.org/p/8979/