Analysis of RNA-seq data

Lecture outline

- Mapping short-read transcript sequences to a reference genome
- Estimating transcript abundance

Mapping

- Typically the first stage (post QC) in RNA-seq analysis
- Several methods have been developed. We will consider one method (Bowtie/TopHat) in detail.

Bowtie 1

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Langmead *et al.* Genome Biology, 2009

- For aligning sequences without gaps
- Uses near exact matches between 5’ ends (seed) of sequence reads and the reference genome
- Seed region has length 28bp
- Exact matching can be memory-intensive, Bowtie uses data compression techniques to run on ~ desktop machine
- Aligns 25 million reads per CPU hour

Bowtie 2

Fast gapped-read alignment with Bowtie 2

Langmead & Salzberg, *Nature Methods* 2012

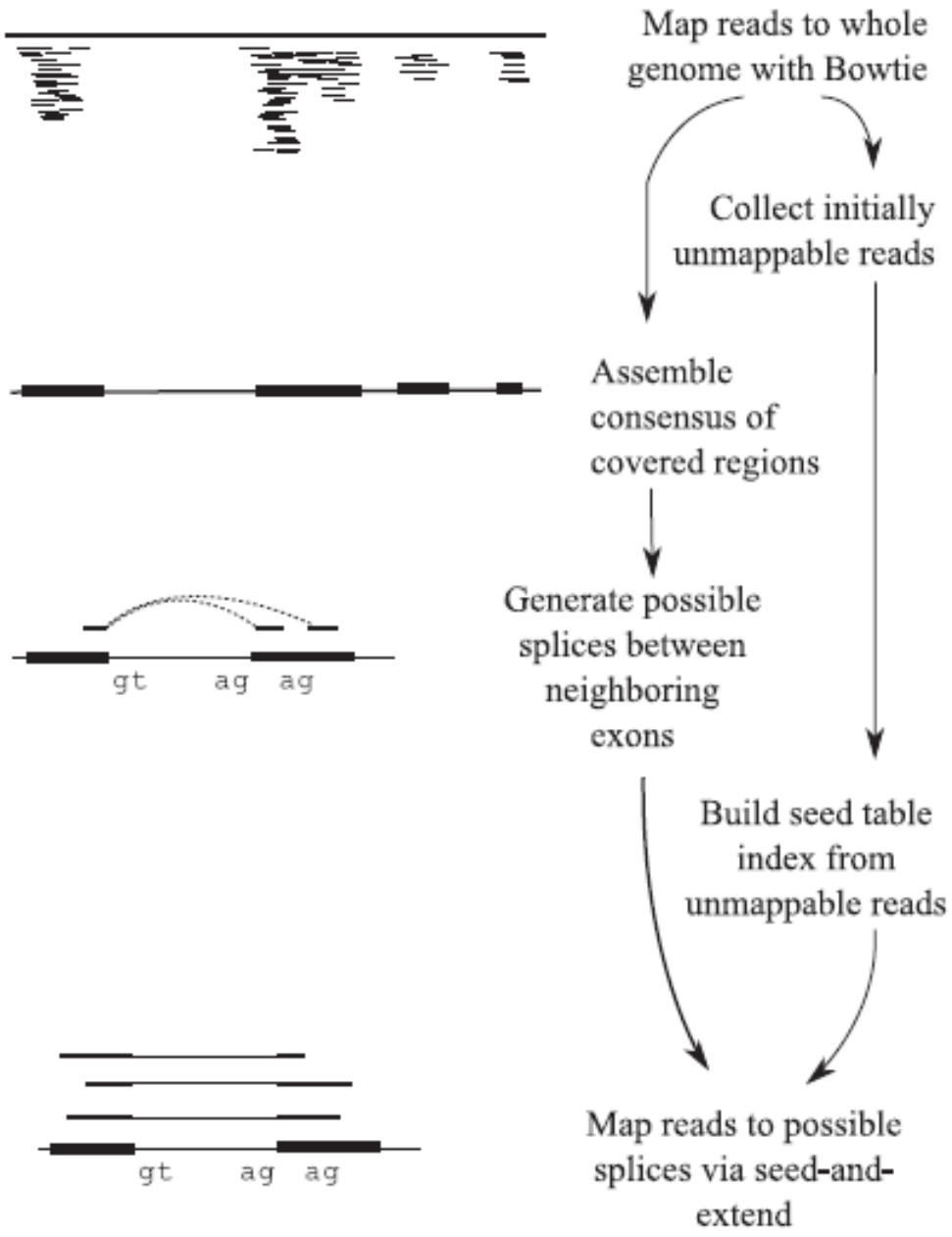
- Bowtie2 allows gapped alignments (with affine gap penalties)
- Algorithm consists of two stages
  - ungapped exact/near exact seed stage
  - extension using parallel implementation of dynamic programming based alignment algorithm

TopHat: discovering splice junctions with RNA-seq

Trapnell *et al.* , Bioinformatics, 2009

- Begins by using Bowtie to align sequences
- Reads classed as mapped or initially unmapabble (IUM reads)
- Mapped reads define putative exons
- Join putative exons separated by gaps too small to be introns (default 6bp, but could be up to 70bp, because introns are generally > 70bp)
- Identify possible splice junctions on the basis of the reads mapped with Bowtie
    - searches for pairs of canonical splice sites between the islands of reads identified with Bowtie (must be > 70bp and < 20Kbp apart, by default)
    - extends the read islands by 45 bp (by default) to allow for missed exon ends
- Use short sequences on either side of splice junction to match to IUM reads

Trapnell et al.



Trapnell et al.

TopHat: matching to splice junctions

- Use seed and extend strategy
    - looks for exact matches to a ‘seed’ region either side of the splice junction in the IUM reads
    - considers only the first 28 bp on 5’ end of the reads (highest quality part) by default (but consider whether this is sufficient in your data…)
- Once a match is found to the seed, extend the alignment
- Typically only splice junctions between read islands can be found. What about splice junctions within islands (e.g. resulting from alternative splicing/intron retention)?
    - searches for splice junctions within islands with very high coverage

Cufflinks

Paper - Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Trapnell *et al*. Nature Biotechnology, 2010

-Uses TopHat for alignment

- Identifies mutually incompatible fragments

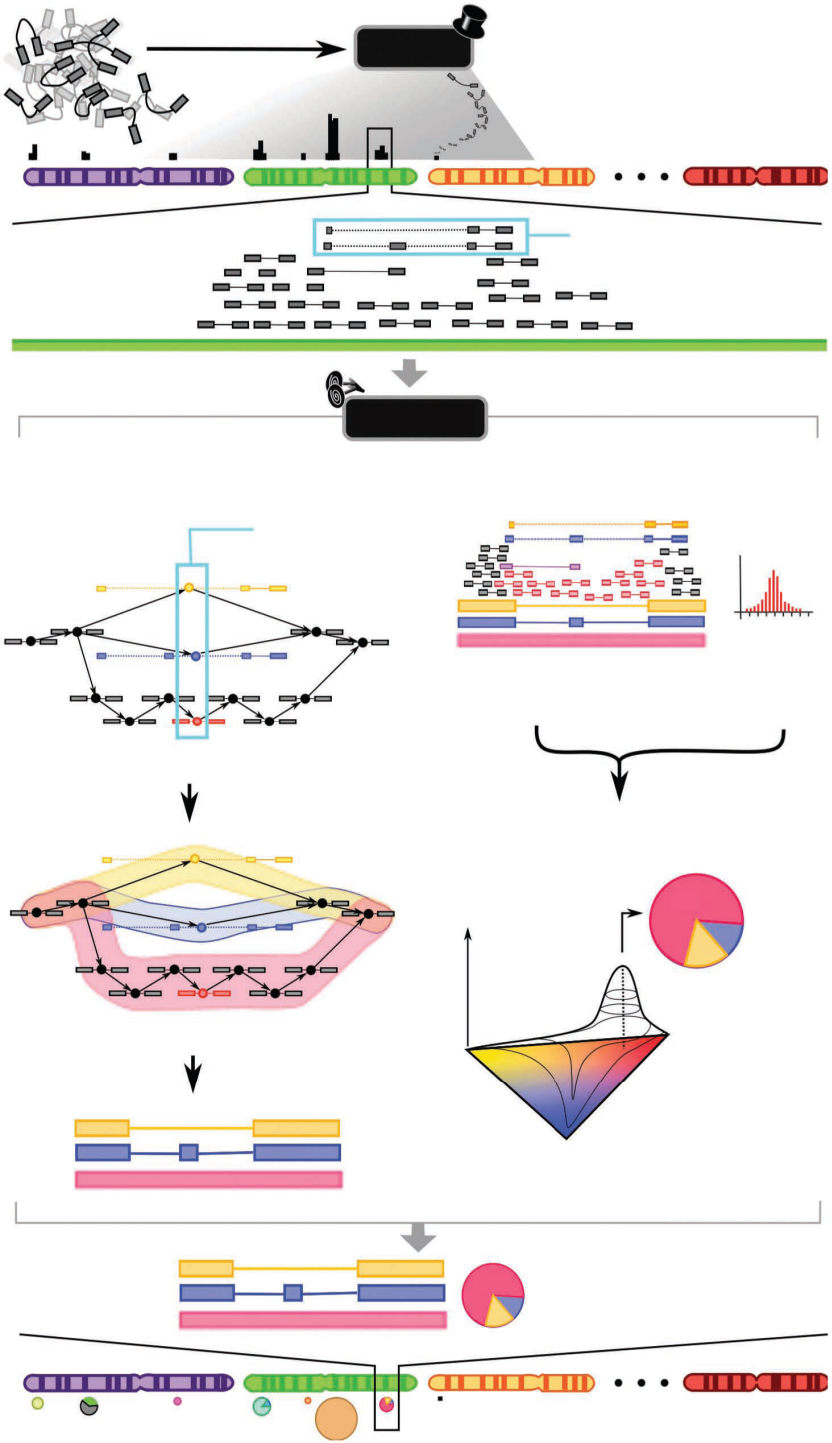
- Uses ‘Dilworth’s Theorem’ to find a minimal set of transcripts that cover all the fragments

- Estimate the abundance of each isoform

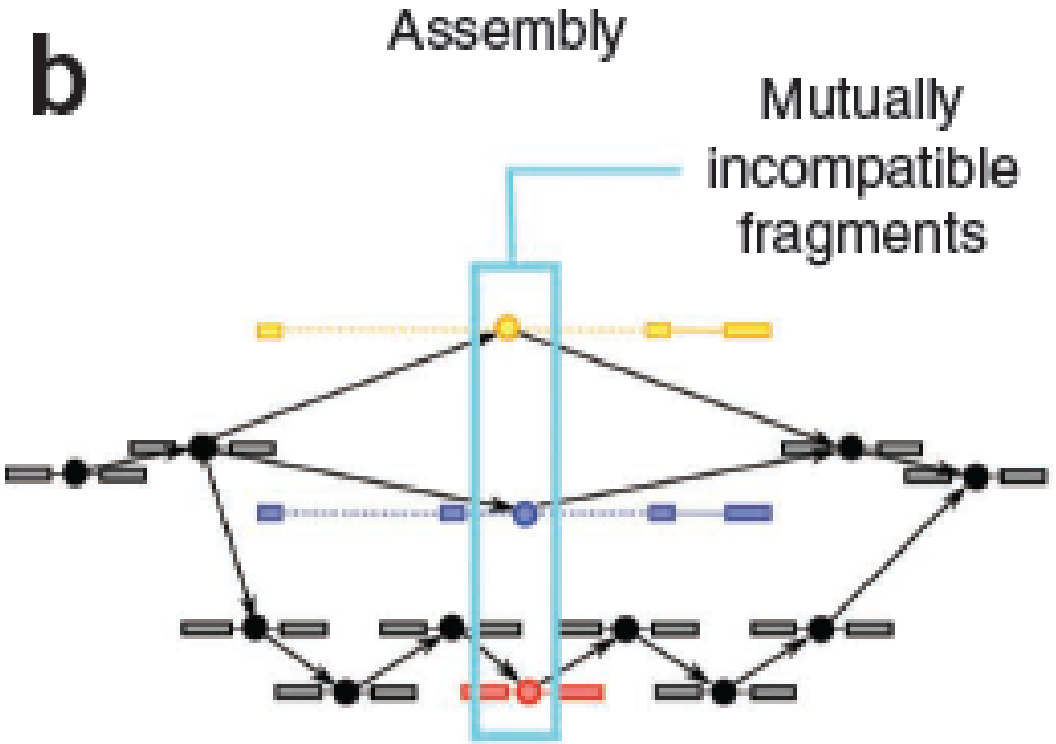
  - takes account of fragment length distribution and probabilitically assigns reads to isoforms

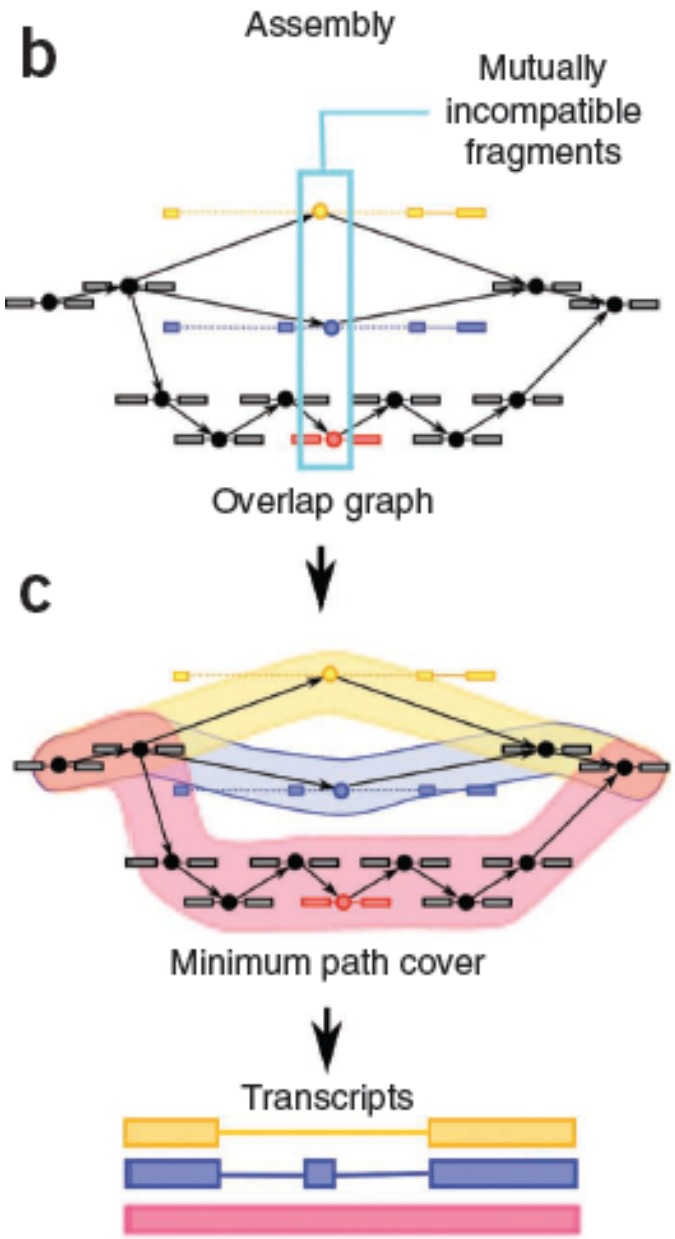
  - takes account also of the expression level of each fragment

  - uses EM algorithm to estimate simultaneously the abundance of each isoform and the source of each fragment



Trapnell et al.


b
Assembly
Mutually
incompatible
fragments


b
Assembly
Mutually
incompatible
fragments
Overlap graph
c
Minimum path cover
Transcripts

Cufflinks EM algorithm

- assume the proportions of each of the, potentially multiple, transcripts for the gene are known

- given these proportions, infer the most likely transcript from which each read comes

- update the estimate of transcript proportions

- begin again

Cufflinks tools

- Cufflinks comes with a set of programs for further analysis of an RNA-seq experiment

  - Cuffmerge: merges transcripts identified in different samples and merges all with know transcripts

  - Cuffdiff: Differential expression analysis

  - CummeRbund: manage, integrate and visualize (in R) results of Cufflinks RNA-seq analysis

See

> Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks
>
> Trapnell *et al.* Nature Protocols, 2012

Addressing biases in RNA-seq.

Model-based methods are being developed to address biases (e.g. by sequence content or position) in RNA-seq – e.g. Pachter et al. 2011 (below)

Probability of generating a read from transcript t of length l that maps to position i of the transcript is

$$P(f^{3'} = (t,i), l(f) = l_t(f)) = \alpha_t \frac{u_{(t,i)} v_{(t,i-l+1)} w_{\frac{i}{l_t}} \frac{D(l)}{\sum_{k=1}^{l-1} D(i-k)}}{\tilde{l}_t}$$

α_t	probability read comes from transcript t
u	sequence bias
v	sequence bias ($5'$ end)
w	positional bias
D	fragment length distribution
$\tilde{l}_t$	effective length of transcript t

EM algorithm used to simultaneously estimate isoform abundance and assign sequence reads to transcript isoforms.

Pachter, 2011

Assignment

1. Carry out initial quality control on all of the RNA-seq datasets in /data2/cathal/RP/Data

These unpublished paired-end RNA-seq data were generated from whole blood samples from patients with a mutation that causes retinitis pigmentosa (samples 4,5,8 and 9) and sibling controls (samples 3,6,7 and 10). Review and discuss all of the qc reports.