M U N I

Cufflinks

Peter Guman

Cufflinks

- Open-source software developed by laboratories at UC Berkeley and Johns Hopkins University
- Cufflinks may be:
  - **name of the RNA-Seq workflow:** together with read mapper TopHat, tools Cuffmerge, Cuffdiff or Cuffnorm
  - **name of the standalone tool** used for the assembly of mapped reads to assemble transcriptomes and evaluate the expression
  - developed mainly in C++ and C languages allowing to have low-level control and optimized performance

What does Cufflinks do?

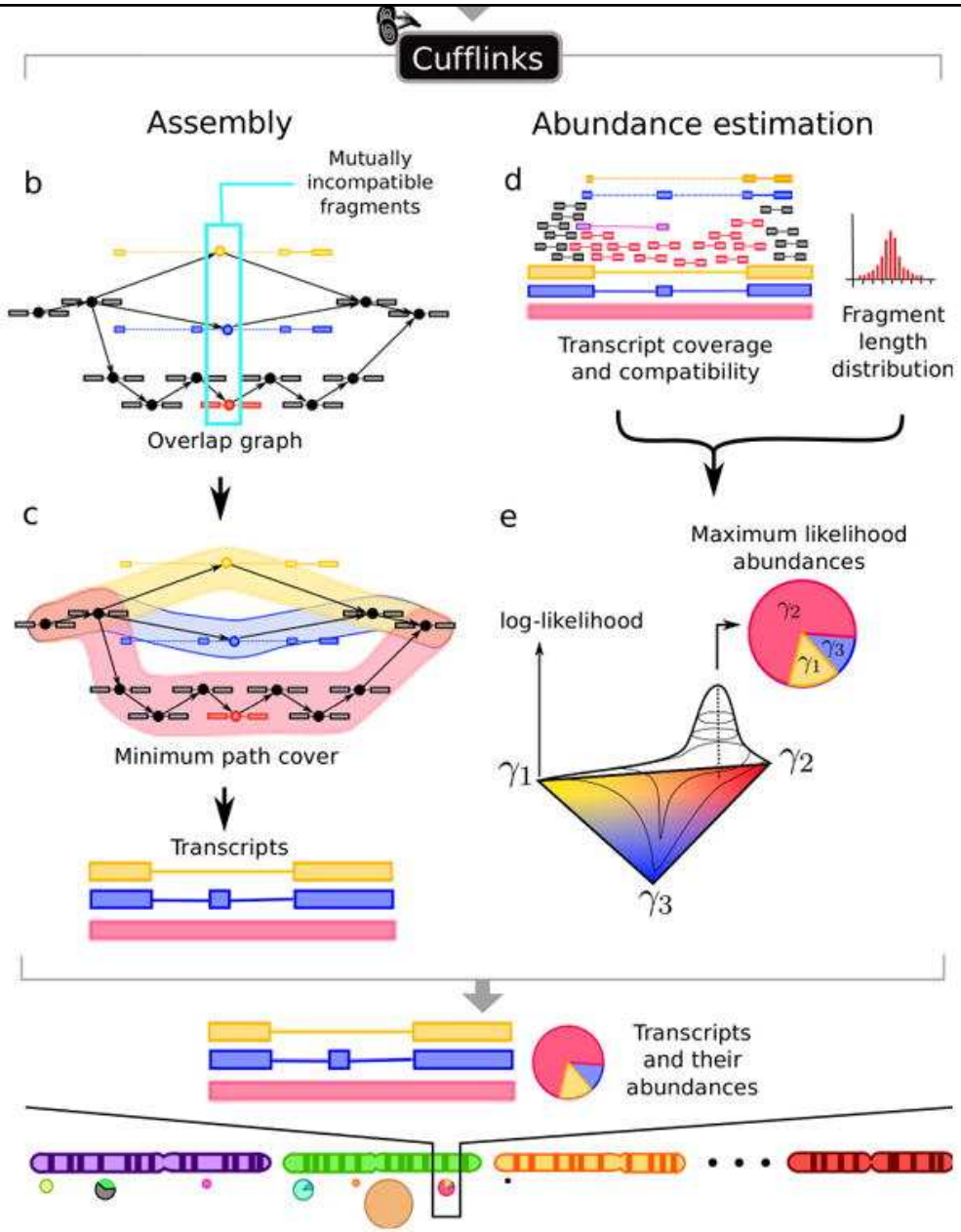
- After the mapping phase of RNA-Seq workflow: assembly of different transcriptions/alternative splicings
- Takes **input:** *aligned reads* in SAM/BAM file format
- Produces **output:**
  1. **Transcriptome assembly** in GTF/GFF format. It has isoforms produced by Cufflink with 7 standard GTF columns with optional information about genes and transcripts.
  2. **Transcript-level expression** in FPKM Tracking file format contains isoform-level expressions.
  3. **Gene-level expression** in FPKM Tracking file format on the gene level.
- Optional: give the Cufflinks tool additional *reference annotation*. This way, Cufflinks can improve already known references.

MUNI

Cufflinks algorithm

- Creation of tree: Compatible fragments are connected when they are compatible*, and they have overlap in the genome
- 'incompatible' fragments must have come from differently spliced isoforms
- isoform transcripts are generated, and an estimation of abundance follows. Estimation is based on the assumption that the probability of observing each function is linear to the abundance of transcripts of their origin.

*according the Dilworth's Theorem


Cufflinks
Assembly
Abundance estimation
b
Mutually incompatible fragments
Overlap graph
d
Transcript coverage and compatibility
Fragment length distribution
c
Minimum path cover
Transcripts
e
Maximum likelihood abundances
log-likelihood
γ_1
γ_2
γ_3
Transcripts and their abundances

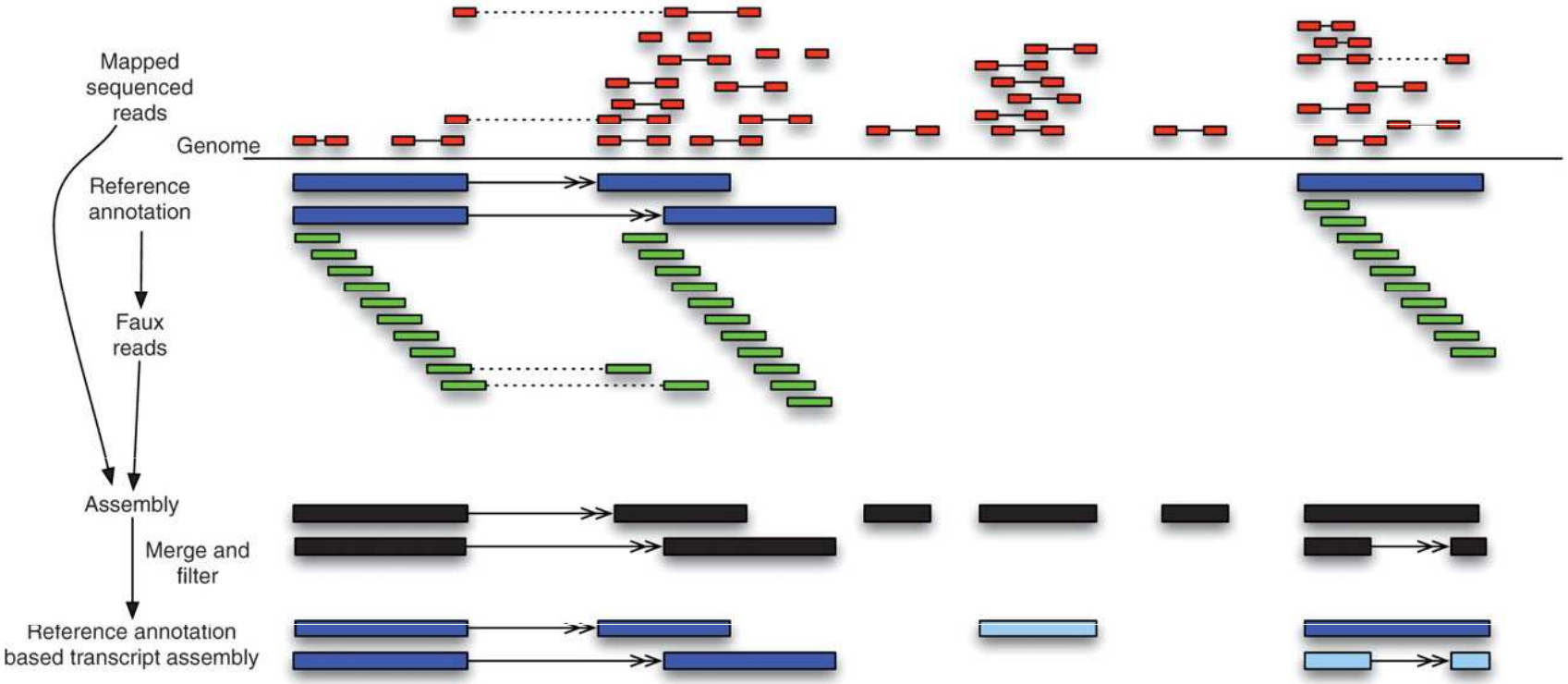
MUNI

Assembler

- Basic Cufflinks assembler creates Directed Acyclic Graph from the reads and finds the shortest path between them (in a maximum parsimony fashion). Does not need a prior reference annotation.

- *reference annotation-based transcript (RABT) assembly* is an improved algorithm. It requires additional annotation of the already known genome.

Fig. 2. An overview of our RABT assembly method. First paired-end reads (mates shown connected by solid lines) are ...



Bioinformatics, Volume 27, Issue 17, 1 September 2011, Pages 2325–2329, https://doi.org/10.1093/bioinformatics/btr355

Recources

- Adam Roberts, Harold Pimentel, Cole Trapnell, Lior Pachter, Identification of novel transcripts in annotated genomes using RNA-Seq, *Bioinformatics*, Volume 27, Issue 17, 1 September 2011, Pages 2325–2329, https://doi.org/10.1093/bioinformatics/btr355
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5. doi: 10.1038/nbt.1621. Epub 2010 May 2. PMID: 20436464; PMCID: PMC3146043.
- official site: cole-trapnell-lab.github.io/cufflinks/

MUNI