Mapping with Tophat 2014

Objectives

In this lab, you will explore a popular transcriptome-aware mapper called Tophat. Simulated RNA-seq data will be provided to you; the data contains paired-end reads that have been generated in silico to replicate real gene count data from Drosophila. The data simulates two biological groups with three biological replicates per group (6 samples total). The objectives of this lab is to:

1. Learn how Tophat2 works and how to use it.
2. Learn how it is different from using a mapper like BWA.

12 raw data files have been provided for all our further RNA-seq analysis:

- c1_r1, c1_r2, c1_r3 from the first biological condition
- c2_r1, c2_r2, and c2_r3 from the second biological condition

Introduction

Tophat is part of the tuxedo suite of RNA-Seq tools. Tophat does a transcriptome-aware alignment of the input sequences to a reference genome using either the Bowtie or Bowtie2 aligner (in theory it can use other aligners, but we do not recommend this).

How Tophat Works

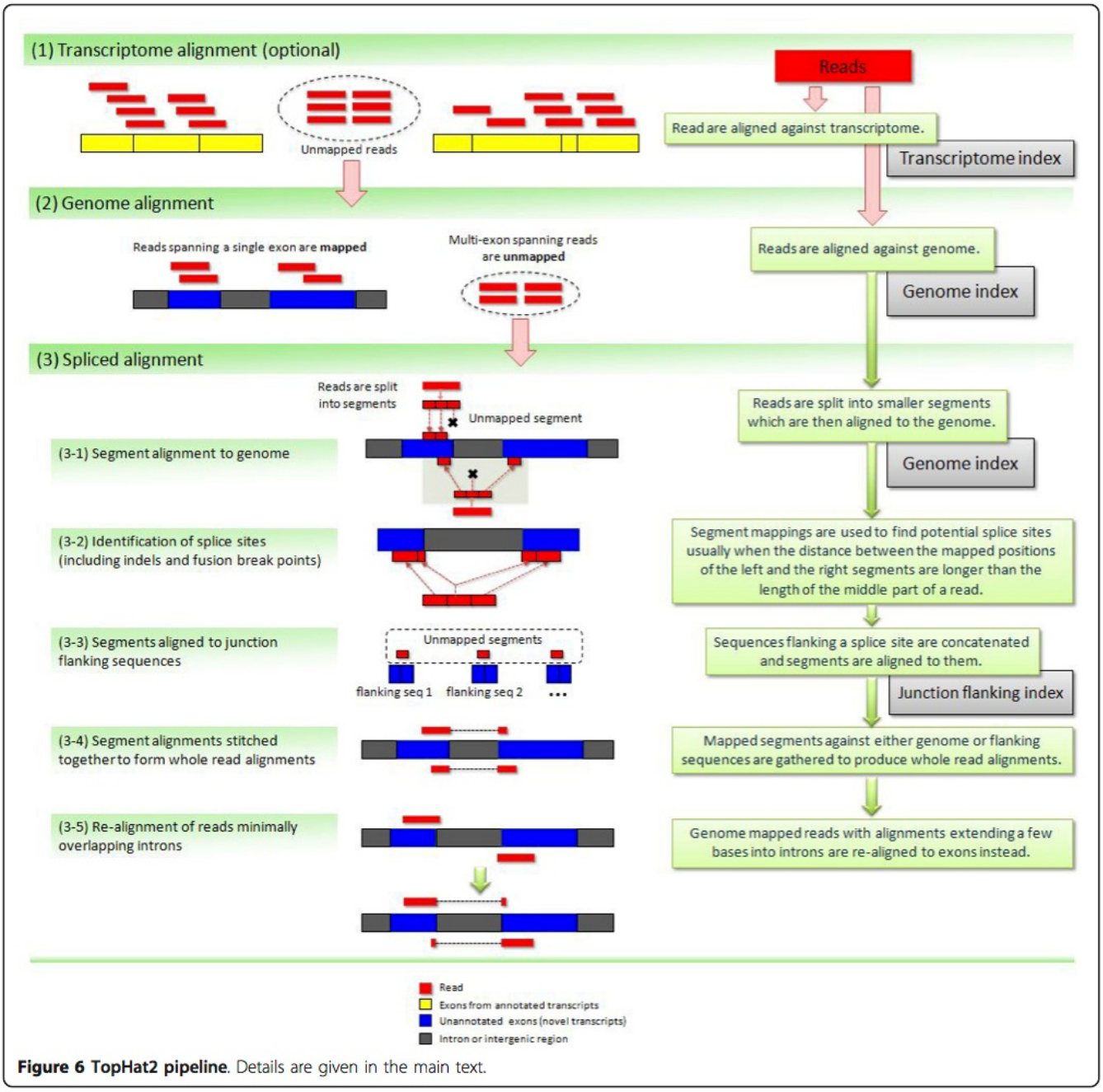

Figure 6 TopHat2 pipeline. Details are given in the main text.

`Image from:` http://genomebiology.com/2013/14/4/R36

1. The input sequences are aligned to the transcriptome for your reference genome, if you provided a GTF/GFF file.
    - sequences that align to the transcriptome are retained, and their coordinates are translated to genomic coordinates
    - sequences that do not align to the transcriptome are subjected to further analysis below
2. Remaining sequences are broken into sub-fragments of at least 25 bases, and these sub-fragments are aligned to the reference genome.
    - if two adjacent sub-fragments align to non-adjacent genomic locations, they are "trans frags" that will be used to infer splice junctions

At the end of the Tophat process, you have a BAM file describing the alignment of the input data to genomic coordinates. This file can be used as input for downstream applications like Cuffmerge-Cufflinks-Cuffdiff, which will be described in further sections. You will also have files describing the junctions found.

More documentation on tophat2 can be found here: http://tophat.cbcb.umd.edu/manual.shtml

Why splice aware/split alignment is important?

Split Read Alignment

Now on to our tophat exercises.