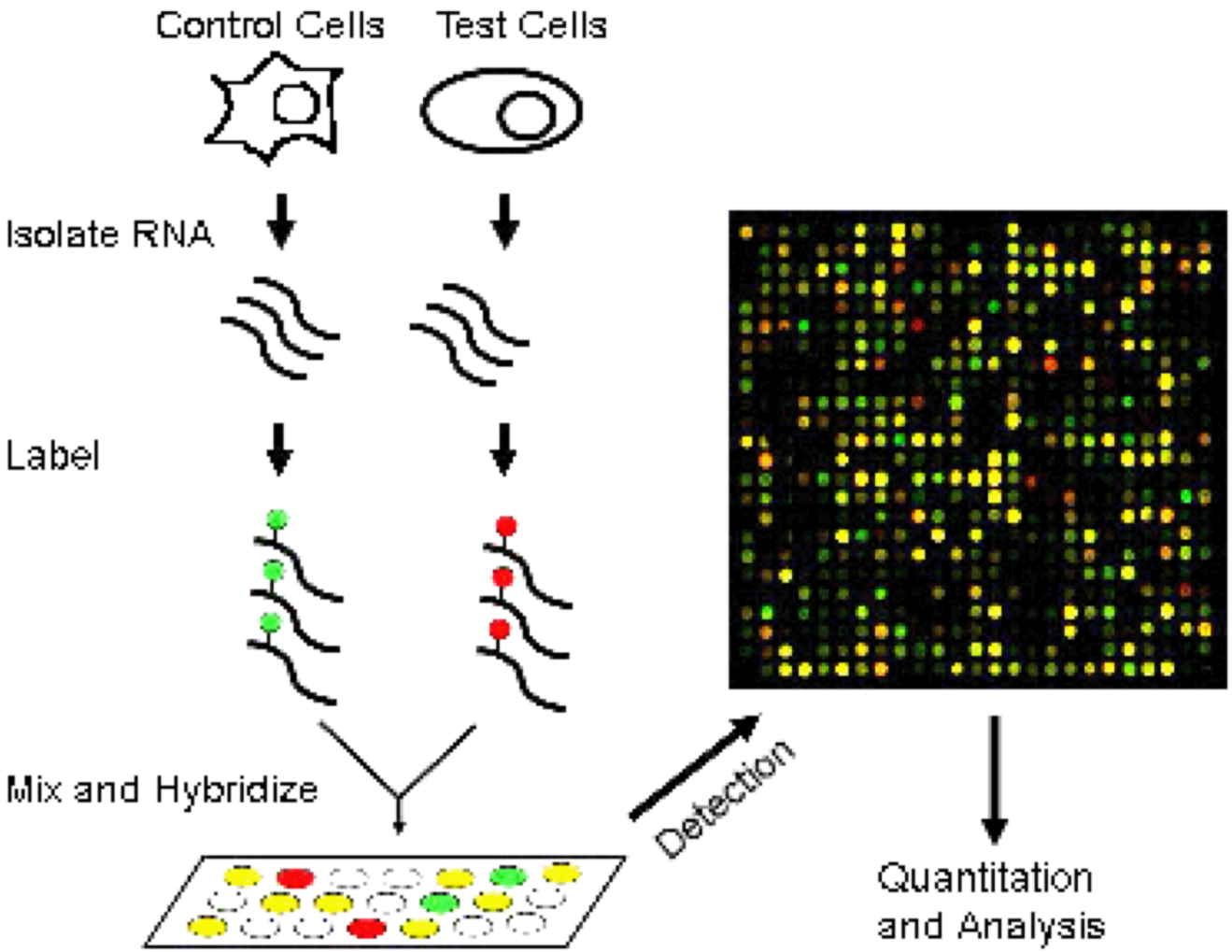
CSEP 527
Computational Biology

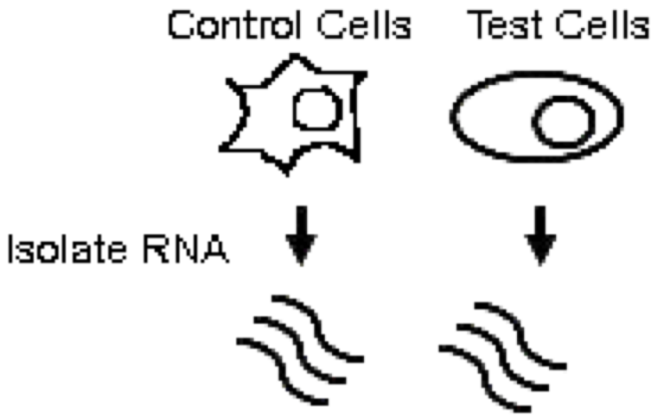
Gene Expression Analysis

Assaying Gene Expression

Microarrays



RNAseq



DNA Sequencer → Millions of reads, say, 100 bp each

map to genome, analyze

Goals of RNAseq

#1: Which genes are being expressed?

How? *assemble* reads (fragments of mRNAs) into (nearly) full-length mRNAs and/or *map* them to a reference genome

#2: How highly expressed are they?

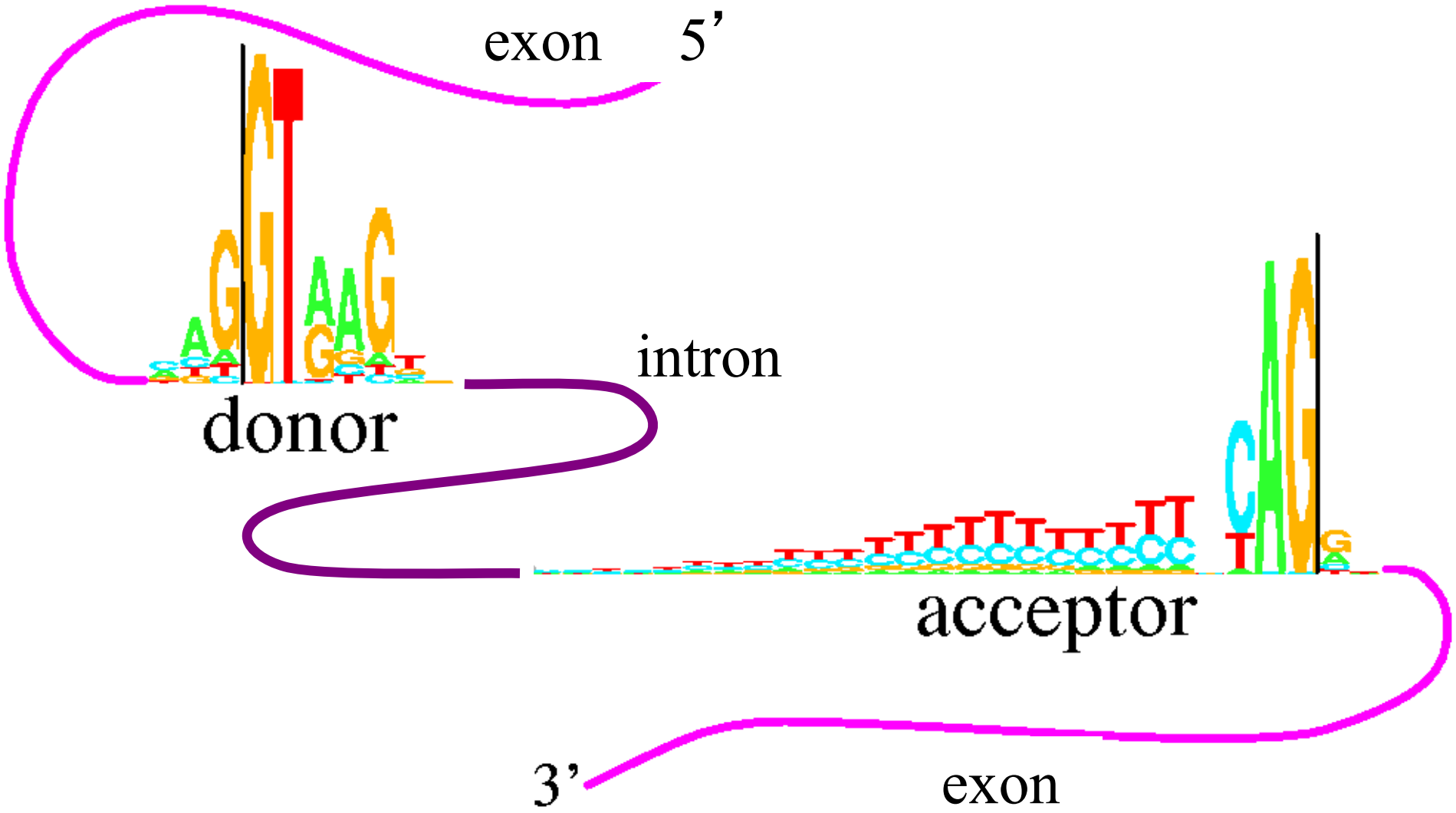
How? *count* how many fragments come from each gene–expect more highly expressed genes to yield more reads, after correcting for biases like mRNA length

#3: What's same/diff between 2 samples

E.g., tumor/normal

#4: ...

Recall: splicing



RNAseq Data Analysis

De novo Assembly

- mostly deBruijn-based, but likely to change with longer reads
- more complex than genome assembly due to alt splicing, wide diffs in expression levels; e.g. often multiple “k’s” used
- pro: no ref needed (non-model orgs), novel discoveries possible, e.g. very short exons
- con: less sensitive to weakly-expressed genes

Reference-based (more later)

- pro/con: basically the reverse

Both: subsequent bias correction, quantitation, differential expression calls, fusion detection, etc.

“TopHat” (Ref based example)

BWA

- map reads to ref transcriptome (optional)
- map reads to ref genome
- unmapped reads remapped as 25mers
- novel splices = 25$_{mers}$ anchored 2 sides
- stitch original reads across these
- remap reads with minimal overlaps

- *Roughly:* 10m reads/hr, 4Gbytes
  (typical data set 100m–1b reads)

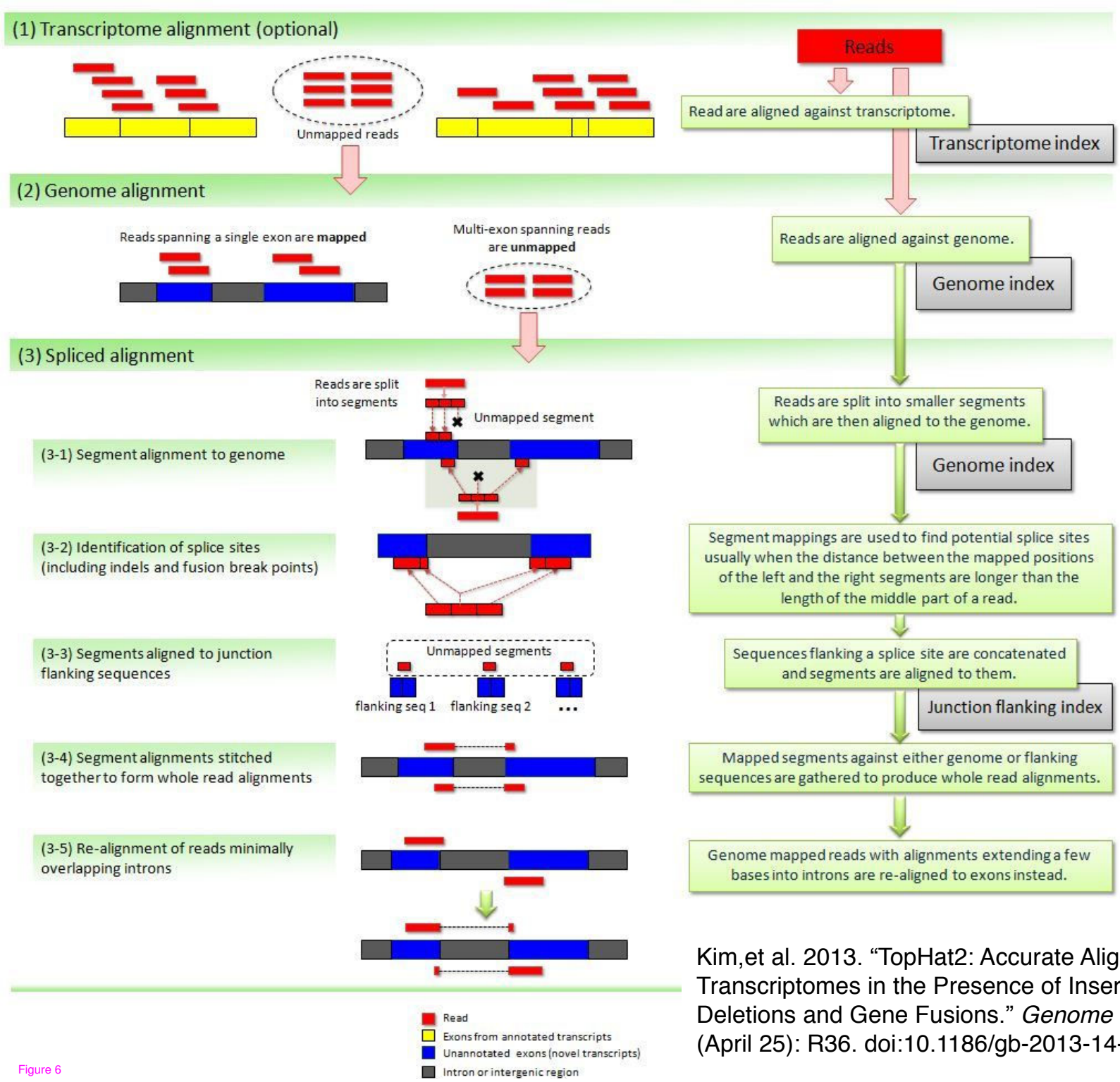
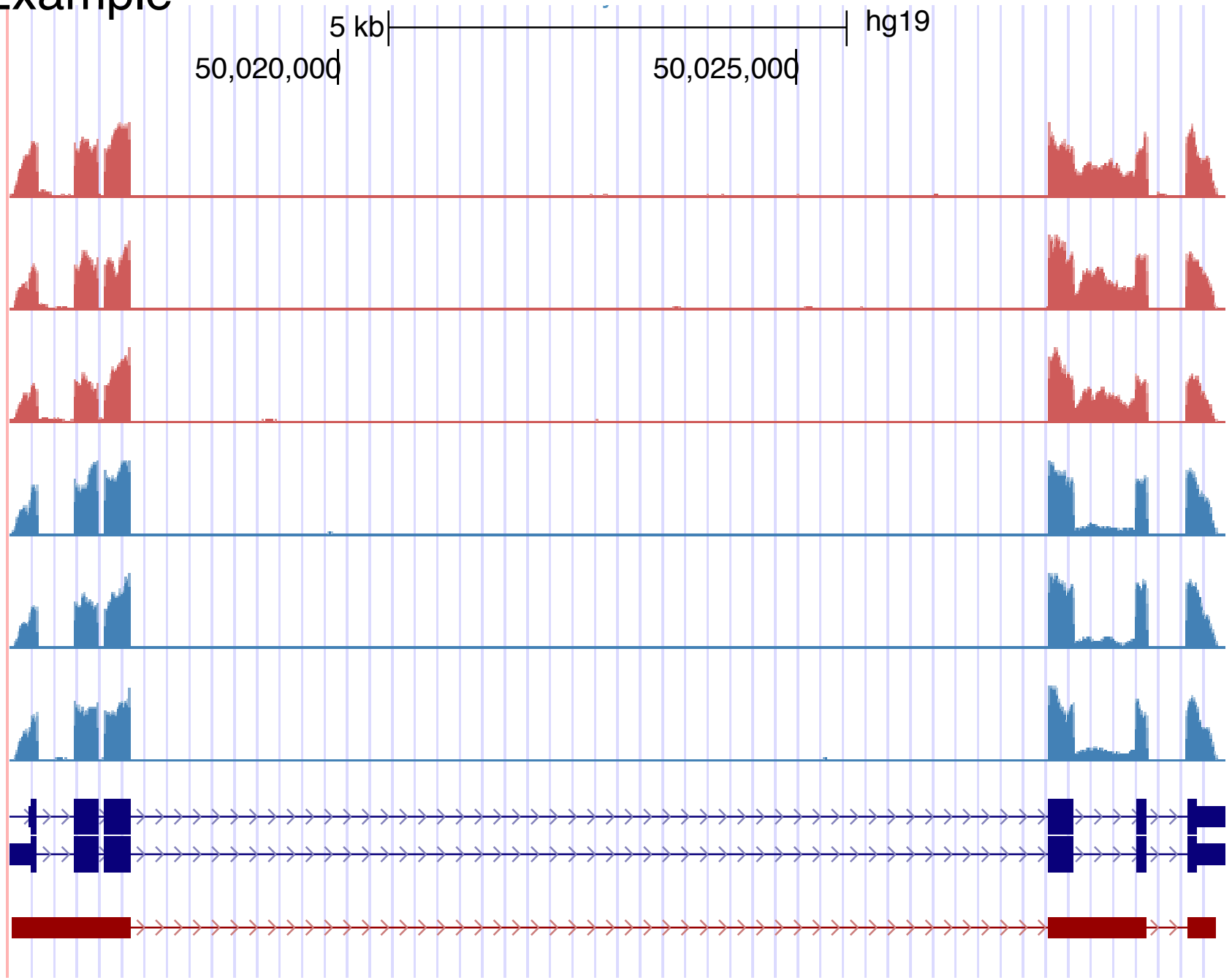


Figure 6

Kim,et al. 2013. "TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions." *Genome Biology* 14 (4) (April 25): R36. doi:10.1186/gb-2013-14-4-r36.

RNAseq Example



RNAseq protocol (approx)

Extract RNA (either polyA polyT or tot – rRNA)

Reverse-transcribe into DNA (“cDNA”)

Make double-stranded, maybe amplify

Cut into, say, ~300bp fragments

Add adaptors to each end

Sequence ~100-175bp from one or both ends

CAUTIONS: non-uniform sampling, sequence (e.g. G+C), 5’-3’, and length biases