Read Alignment

Dr. Jung Soh
657.000 SS 2017

Mapping/Aligning reads

- Possible only when we have a reference
  - Reference genome or *de novo* assembly
- Issues
  - RNA-Seq reads spanning across exon junction
  - Alternative splicing
  - Reads mapping to multiple places in genome
- Different from multiple sequence alignment
  - MSA: align a group of sequences (similar length)
  - Read alignment: align reads to a region of reference sequence(s)

Why align reads to reference

- Assemble a new genome with reference alignment
- Analyze genetic variation from reference
- Check sequencing correctness
  - Most single-organism reads should align well to reference
- Analyze taxonomy of metagenomic reads
  - Align to multiple references
- Study differential expression of transcripts
  - Possible with RNA-Seq

Read alignment tools

- Many mapping tools
  - Bowtie, Bowtie2, SOAP, BWA, SHRiMP, mrFAST, mrsFAST, ZOOM, SSAHA2, Mosaik
- Mapping result
  - Most common format: SAM (sequence alignment/map)
    - Binary version: BAM
  - Use Samtools to analyze SAM/BAM files

Samtools

- Command line tool to work with SAM/BAM files
- Good for text-based analysis
  - SAM format contains lots of (often coded) information
  - Difficult to work with Linux commands
- Collection of commands
  - Select one command
  - Possibly provide options
  - Supply input SAM/BAM

End-to-end vs. local alignment

- End-to-end: align all bases of a read

```
Read:      GACTGCGATCTCGACTTCG
Reference: TCGACTGGGCGATCTCGACTTCGAAAC
Alignment:
  Read:        GACTG--CGATCTCGACTTCG
               |||||  ||||||||||||||
  Reference: TCGACTGGGCGATCTCGACTTCGAAAC
```

- Local: some bases at ends can be unaligned (clipped)

```
Read:      ACGGTTGCGTTAATCCGCCACG
Reference: TAACTTGCGTTAAATCCGCCTGG
Alignment:
  Read:      ACGGTTGCGTTAA-TCCGCCACG
                 ||||||||| ||||||
  Reference: TAACTTGCGTTAAATCCGCCTGG
```

Bowtie2

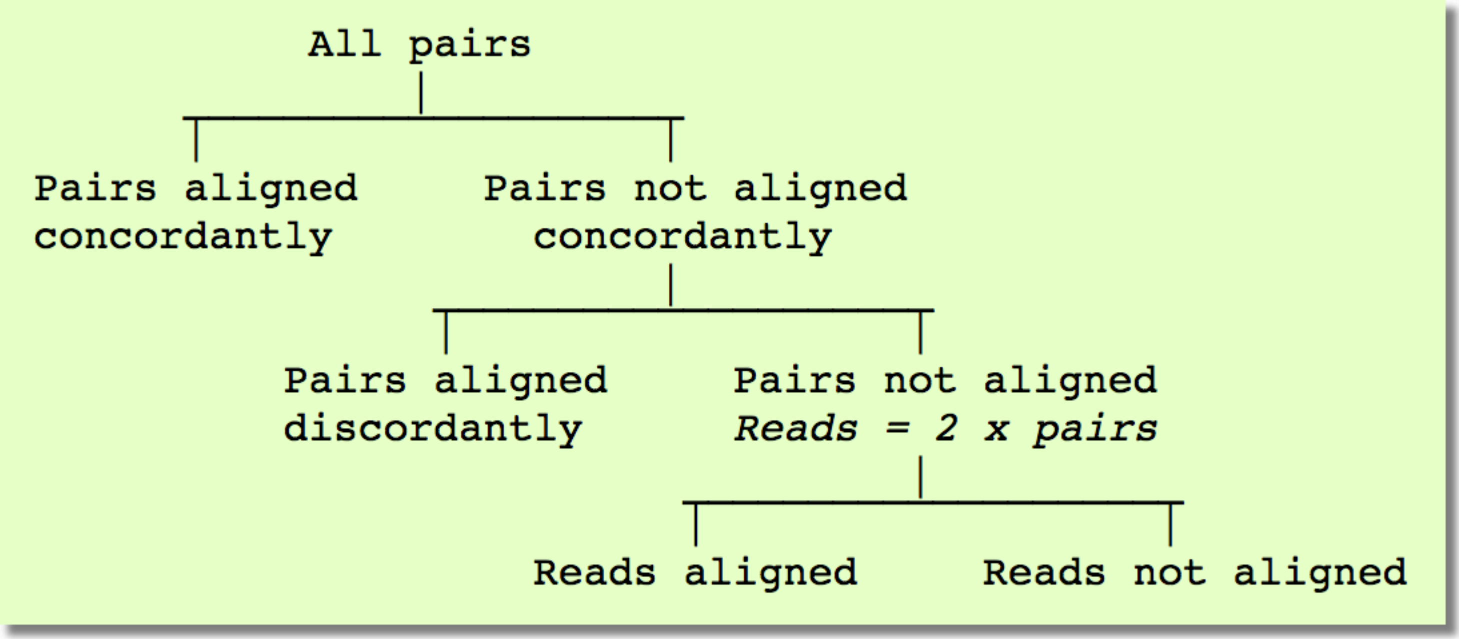
- A widely used read alignment tool
- Two steps
  - Build index (database) from sequence files
    - Files can represent a genome, chromosome, or your own set of sequences
  - Align reads to the index
- Bowtie
  - Good for reads shorter than 50 bp
  - Only ungapped, end-to-end alignments
  - Read length upper limit of around 1000 bp

Bowtie2 key options

- Alignment mode
  - End-to-end (default)
  - Local
- Reporting policy
  - Search for multiple alignments, report the best one (default)
  - Search for up to *N* multiple alignments, report each
  - Search for and report all alignments

Alignment summary

- Usually given as alignment rate
- Can be complicated for paired-end reads
  - Especially with inconsistent terms used by alignment tools



All pairs

Pairs aligned concordantly

Pairs not aligned concordantly

Pairs aligned discordantly

Pairs not aligned

Reads = 2 x pairs

Reads aligned

Reads not aligned

Alignment visualization

- Similar to assembly visualization
  - Reads aligned to reference sequences (not assembled into contigs)
  - Reference sequences often annotated
- Many tools
  - IGV, SeqMonk, Tablet, BamView, Samtools