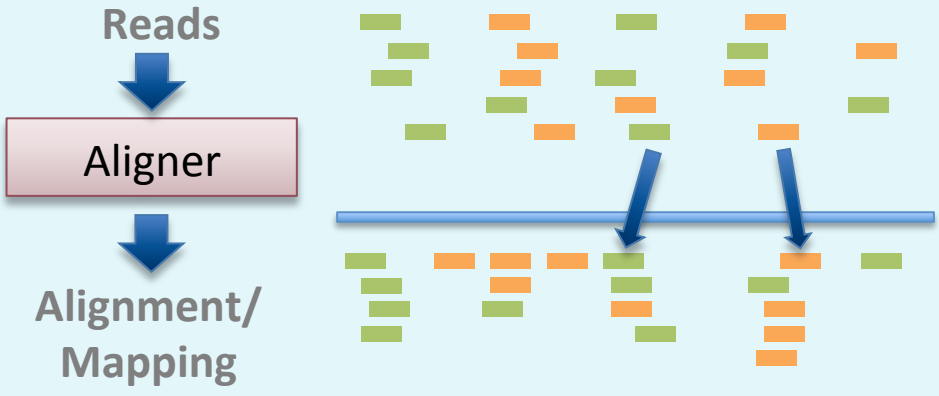
Next-Generation Sequencing Read Alignment

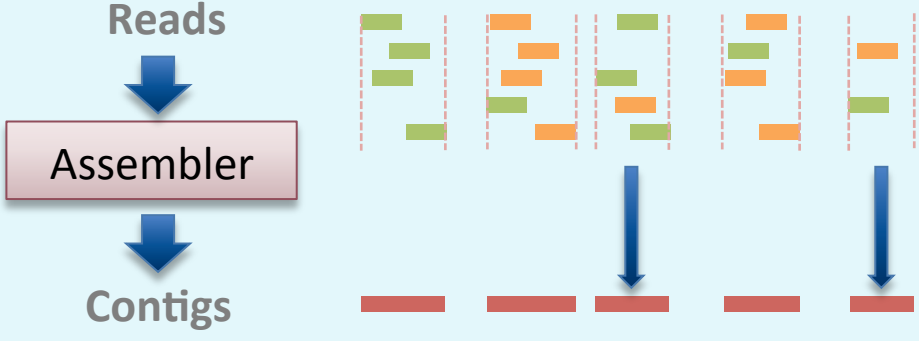
Dr. Jung Soh
MOL.923 SS 2018

Read alignment vs. *de novo* assembly

Read alignment:
reads aligned to
reference



De novo assembly:
reads assembled
into contigs



Mapping/Aligning reads

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

Reasons for aligning reads

- 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 reads

NGS read alignment tools

- Many alignment tools
  - Bowtie, Bowtie2, SOAP, BWA, SHRiMP, mrFAST, mrsFAST, ZOOM, SSAHA2, Mosaik
- Mapping result
  - Mostly in SAM (sequence alignment/map) format
    - Binary version: BAM
  - Samtools commonly used 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
- Collection of commands
  - Say **`samtools`**, give one command, provide options, supply input SAM/BAM
  - Examples:
    - **`samtools view -b align.sam`**
    - **`samtools depth -r REF:FROM-TO align.sorted.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
```

Bowtie 2

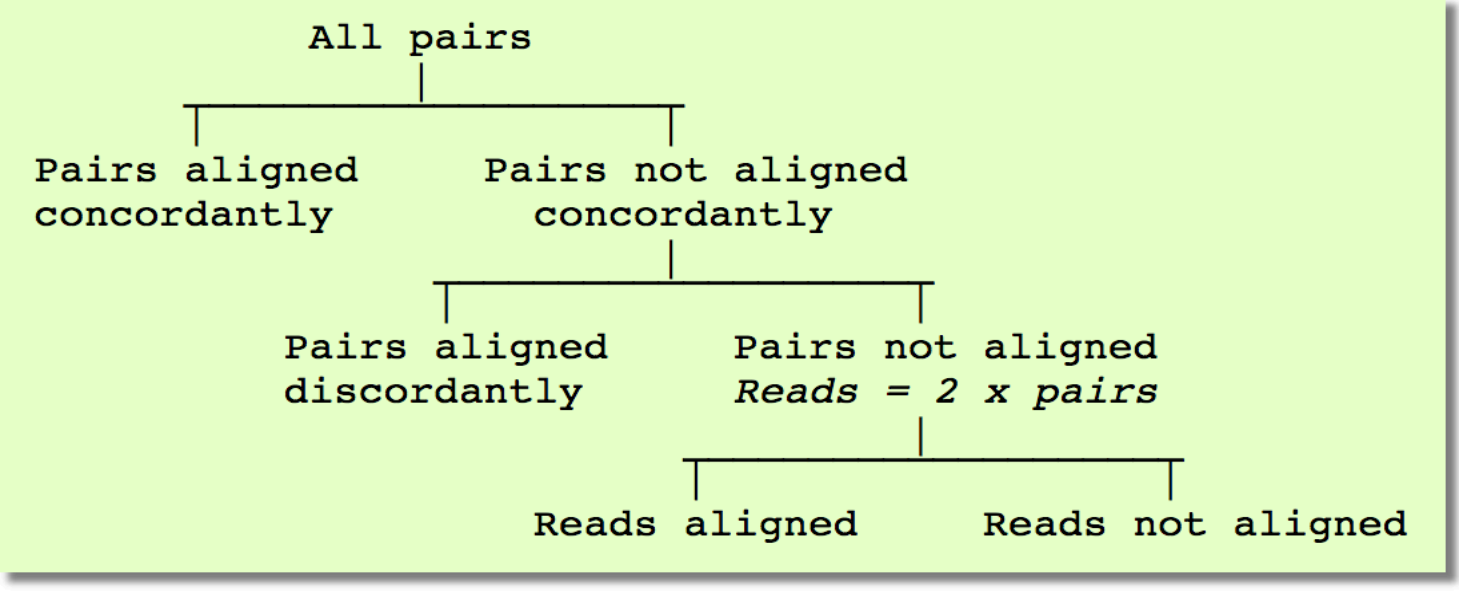
- A widely used read alignment tool
- Two steps
  - Build index (database) from sequence files
    - Files can represent a genome, chromosome, or your own set of sequences
  - Align reads to the index
- Bowtie
  - Good for reads shorter than 50 bp
  - Only ungapped, end-to-end alignments
  - Read length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

Bowtie 2 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

- Given as alignment rate:
  - (Reads aligned)/(Total reads)
- Calculation can be complicated for paired-end reads
  - Especially with inconsistent terms used by alignment tools



Alignment visualization

- Similar to assembly visualization
  - Reference sequences treated as contigs
  - Reference sequences often annotated (additional information)
- Many tools
  - IGV, SeqMonk, Tablet, BamView, Samtools

Lab overview