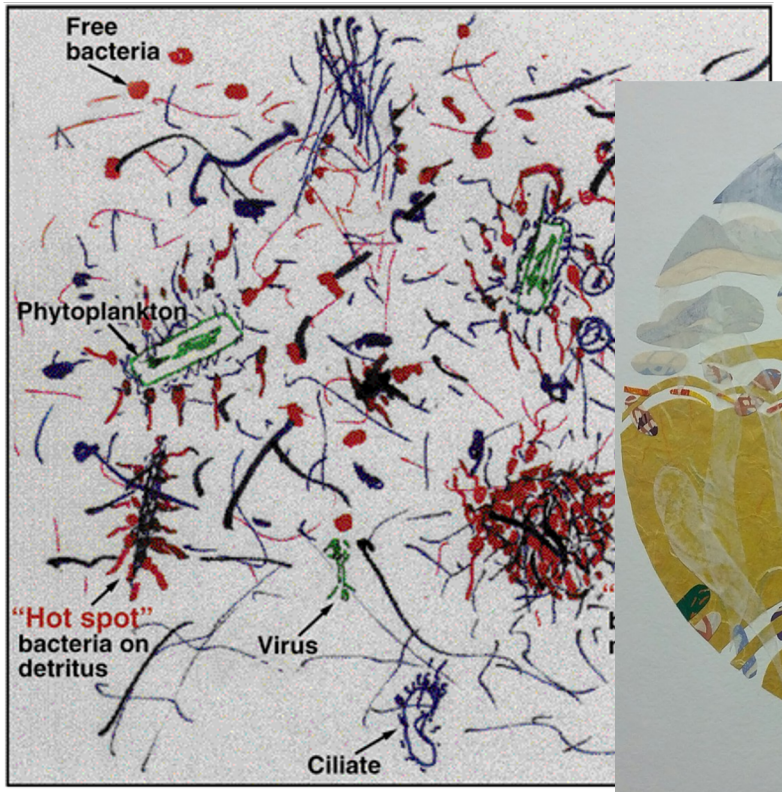
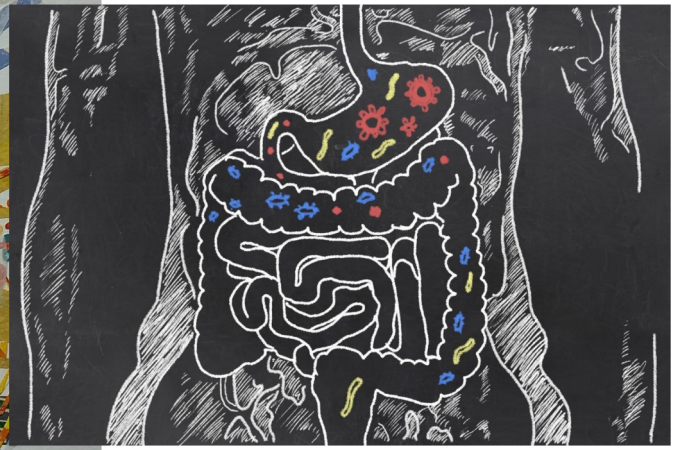
Metagenomics Lesson 4

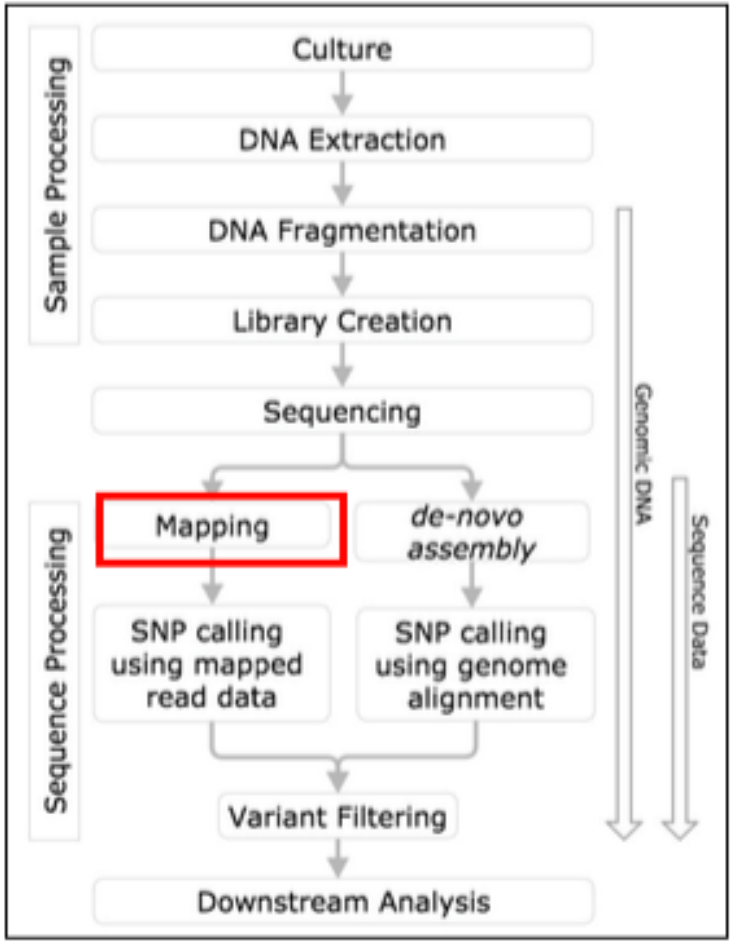




Metagenomics Lesson 3

Read mapping using bowtie2

1. What is read maping?
2. Alignment scoring
3. Different read mappers
4. Burrows-Wheeler algorithm
5. Demo: bowtie2

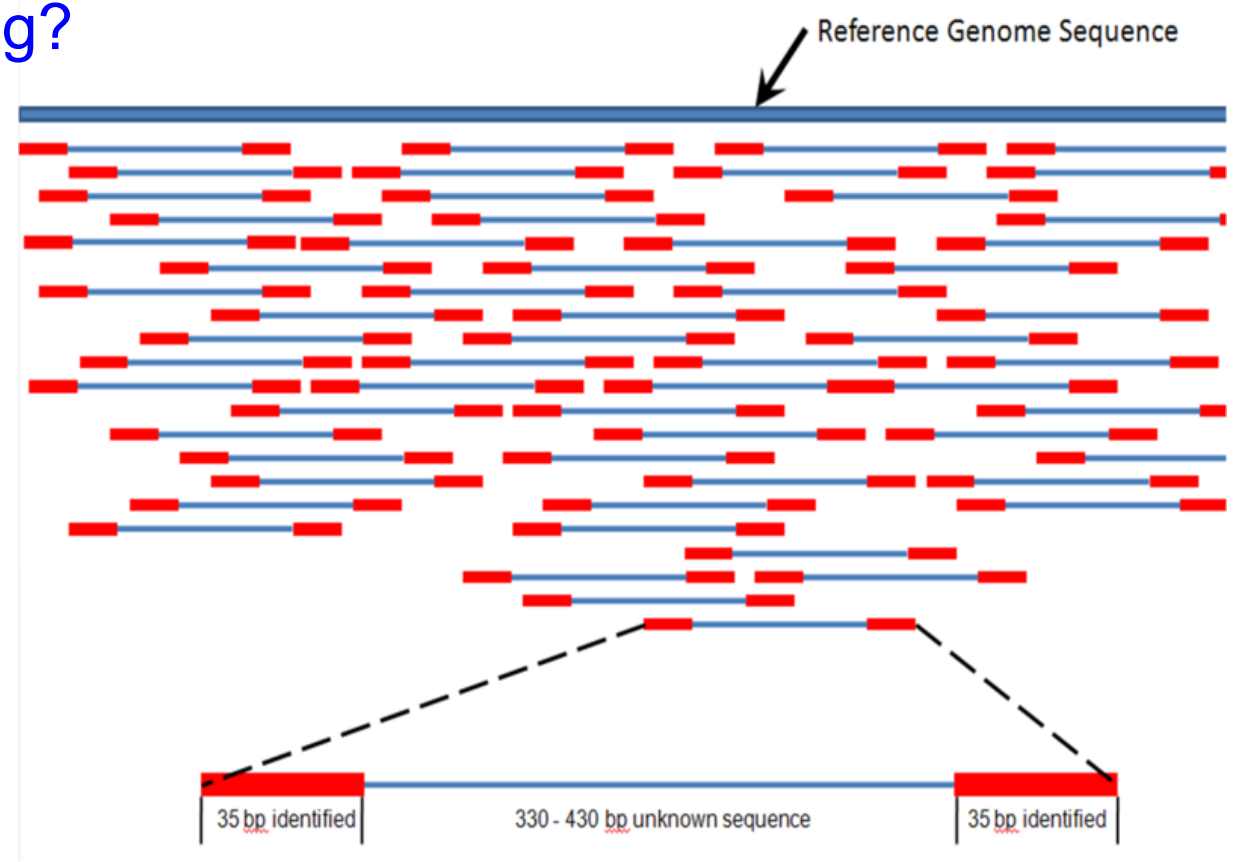


What is read mapping?

Alignment of short read to a reference genome

Find the area in the genome where the read matches best

Requires millions of string searches in a long string.



Why map reads?

Determine presence/absence of a genome in a sample

Calculate coverage as a measure of relative abundance

Determine variants like SNPs

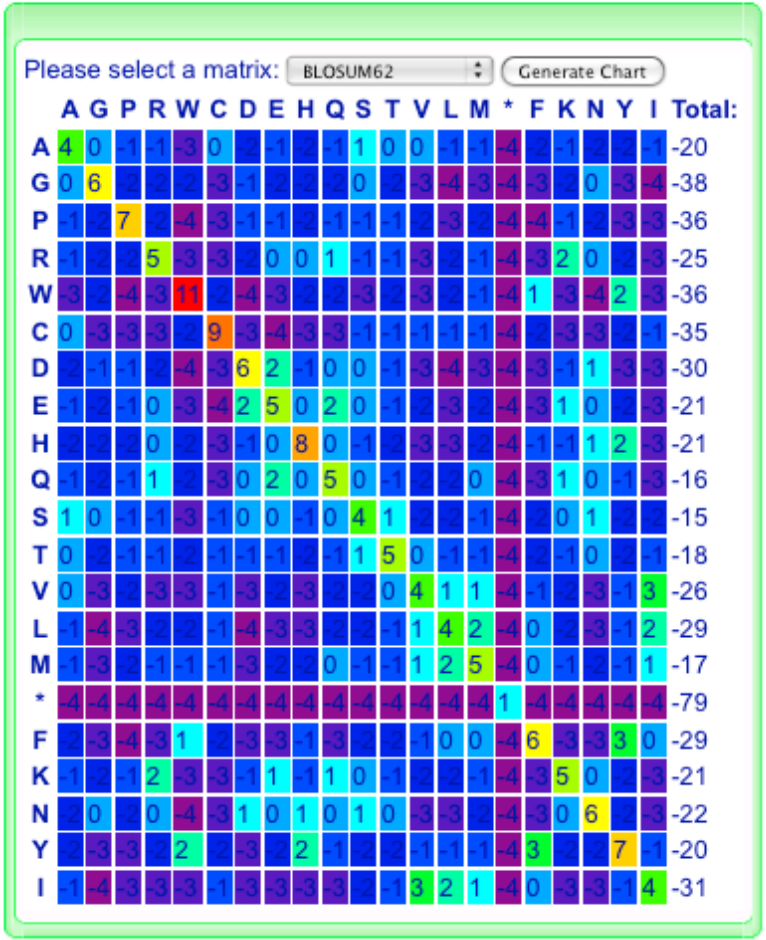
Use coverage for binning of MAGs

How does alignment work?

Very generally:

1. Find options for where a read might map using kmers
2. Score the sites by # matches, # mismatches and # indels
3. Pick the best hit

There are many scoring matrices that differ in their penalty for mismatches, gaps and gap extensions



Read mapping tools

- BWA
- Bowtie2
- BBmap
- Smalt
- SOAP2
- TopHat
- And many more…

Comparison between popular mappers:
http://merenlab.org/2015/06/23/comparing-different-mapping-software/

Bowtie2: Build reference genome index

- Typically start with genome sequence file (in fasta format)
- Burrows-Wheeler Transformation + FM index

```
>Y.pestis gi|16120353|ref|NC_003143.1| Yersinia pestis CO92
GATCTTTTTATTTAAACGATCTCTTTATTAGATCTCTTATTAGGATCATGATCCTCTGTGGATAAGTGAT
TATTCACATGGCAGATCATATAATTAAGGAGGATCGTTTGTTGTGAGTGACCGGTGATCGTATTGCGTAT
AAGCTGGGATCTAAATGGCATGTTATGCACAGTCACTCGGCAGAATCAAGGTTGTTATGTGGATATCTAC
TGGTTTTACCCTGCTTTTAAGCATAGTTATACACATTCGTTCGCGCGATCTTTGAGCTAATTAGAGTAAA
TTAATCCAATCTTTGACCCAAATCTCTGCTGGATCCTCTGGTATTTCATGTTGGATGACGTCAATTTCTA
ATATTTCACCCAACCGTTGAGCACCTTGTGCGATCAATTGTTGATCCAGTTTTATGATTGCACCGCAGAA
AGTGTCATATTCTGAGCTGCCTAAACCAACCGCCCCAAAGCGTACTTGGGATAAATCAGGCTTTTGTTGT
TCGATCTGTTCTAATAATGGCTGCAAGTTATCAGGTAGATCCCCGGCACCATGAGTGGATGTCACGATTA
ACCACAGGCCATTCAGCGTAAGTTCGTCCAACTCTGGGCCATGAAGTATTTCTGTAGAAAACCCAGCTTC
TTCTAATTTATCCGCTAAATGTTCAGCAACATATTCAGCACTACCAAGCGTACTGCCACTTATCAACGTT
ATGTCAGCCATTCAAGAACCCAACTGAAGTAAAGAGCTGGCATTGTACTCTGTGAATCAGCTGGGATCTA
```

Indexed genome file:
used for aligning
sequencing reads

Bowtie2: Aligning sequencing reads to reference genome index: Many parameter options!

End to end alignment:
Uses all bases in read

```
Read:      GACTGGGCGATCTCGACTTCG
Reference: GACTGCGATCTCGACATCG

Alignment:
  Read:      GACTGGGCGATCTCGACTTCG
             |||||  |||||||||| |||
  Reference: GACTG--CGATCTCGACATCG
```

Local alignment:
Can exclude bases at the ends of the read

```
Read:      ACGGTTGCGTTAATCCGCCACG
Reference: TAACTTGCGTTAAATCCGCCTGG

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

http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

Bowtie2: Mapping quality

- Sometimes it is challenging to identify the correct location in the reference genome (especially for reads that align to repetitive regions)
- indels are challenging (sometimes require a secondary realignment step; other software may be better for this such as GATK)
- MAPQ = Mapping quality (Q = -10 log10 p)
    - Where p is the probability of an incorrect alignment (the read's true origin is somewhere else in the genome)
    - High MAPQ scores are better for more confident DNA variant calling

Bowtie2: Dealing with paired-end reads

Concordant pair: Match expectations - Orientation and spacing OK

Discordant pair: Does not match expectations! Read 2 is in an insertion element that is not in this location the reference genome

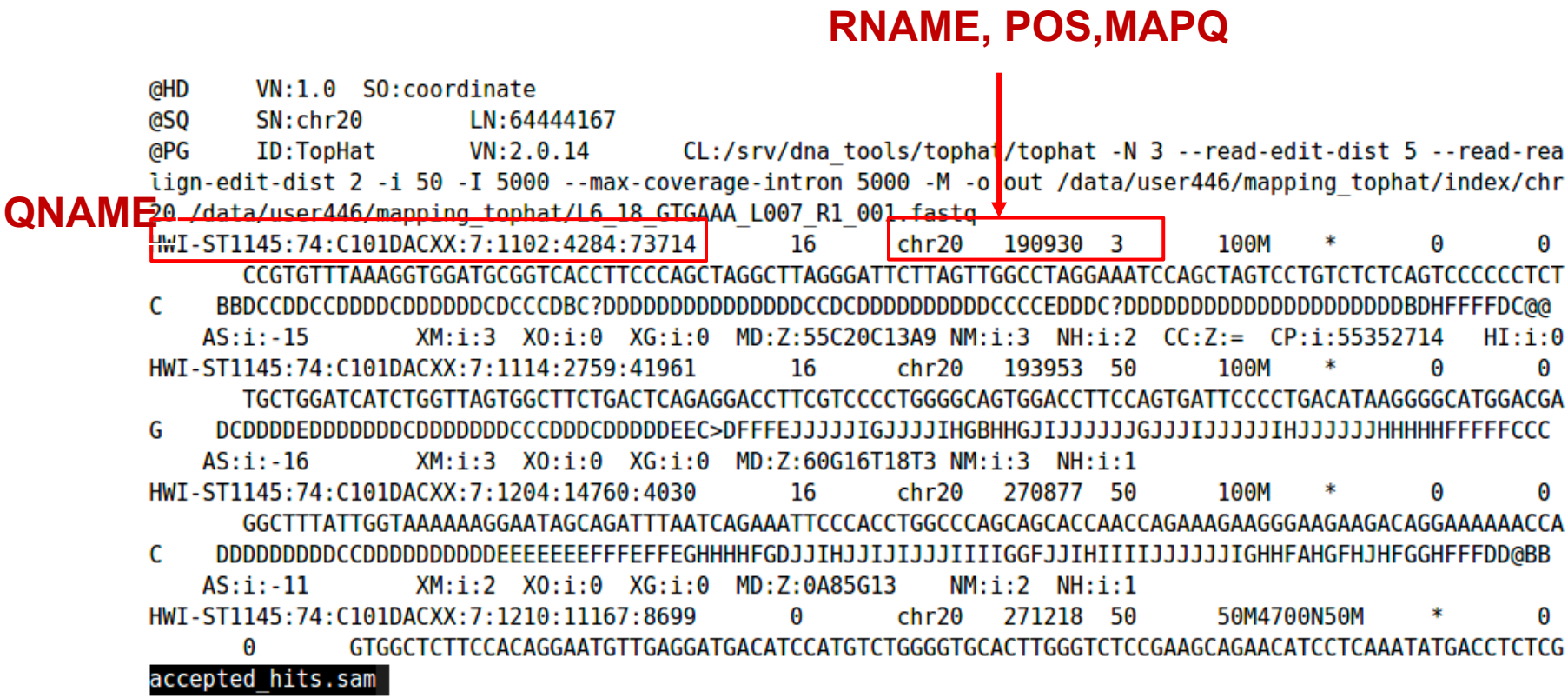
SAM/BAM files- The output of sequence alignment software (like Bowtie2)

- SAM = Sequence Alignment/Map format (Text-based format)

- BAM = Binary format of a SAM file. Better for computation.

Col	Field	Type	Brief Description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

Example SAM file



SAMtools: Processing SAM/BAM files post alignment

- **Sorting SAM/BAM reads by chromosome position** – this is done prior to variant calling

- **Remove duplicate reads** – This is often done (as these may be PCR artifacts introduced during the prep of the sequencing library)

http://samtools.sourceforge.net/