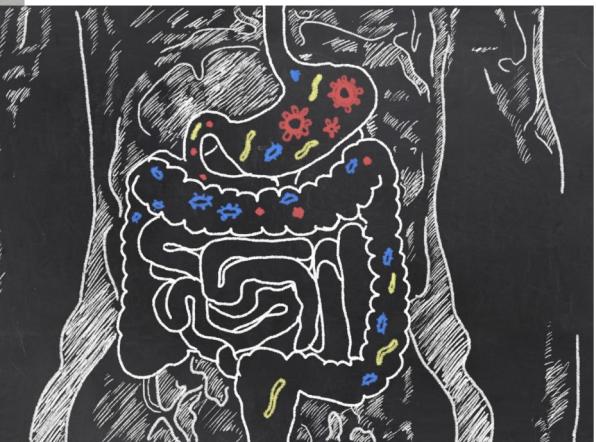
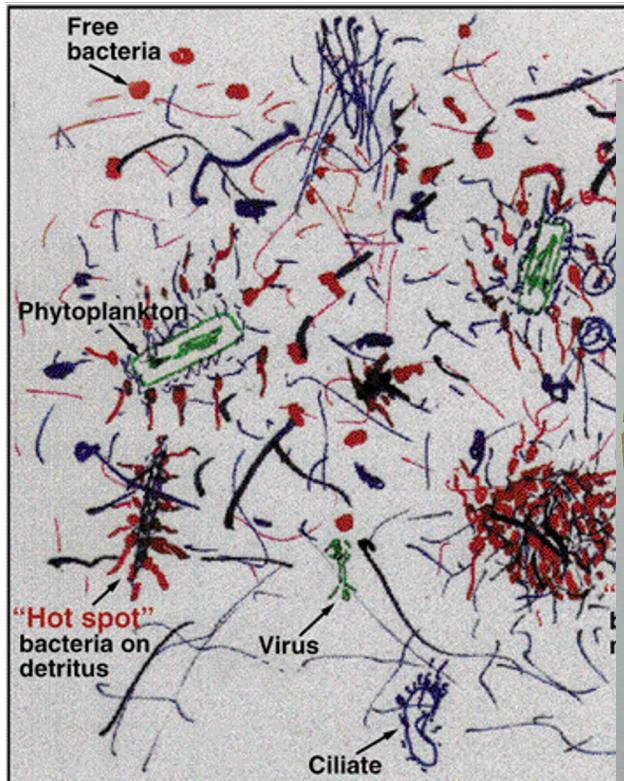


Metagenomics Lesson 4

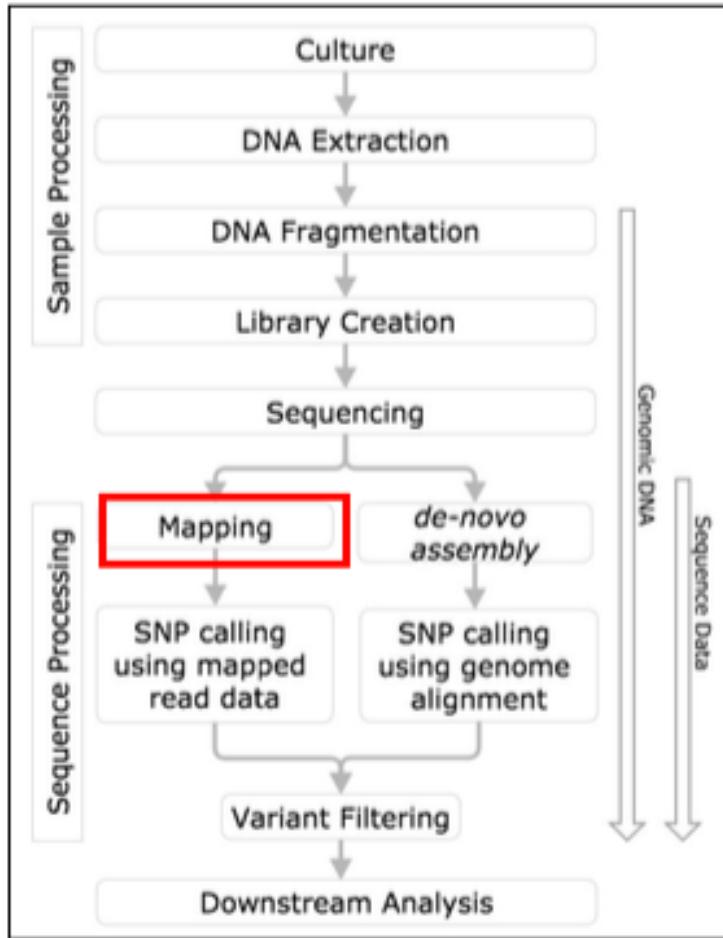


Gail.Priday.

Metagenomics Lesson 3

Read mapping using bowtie2

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

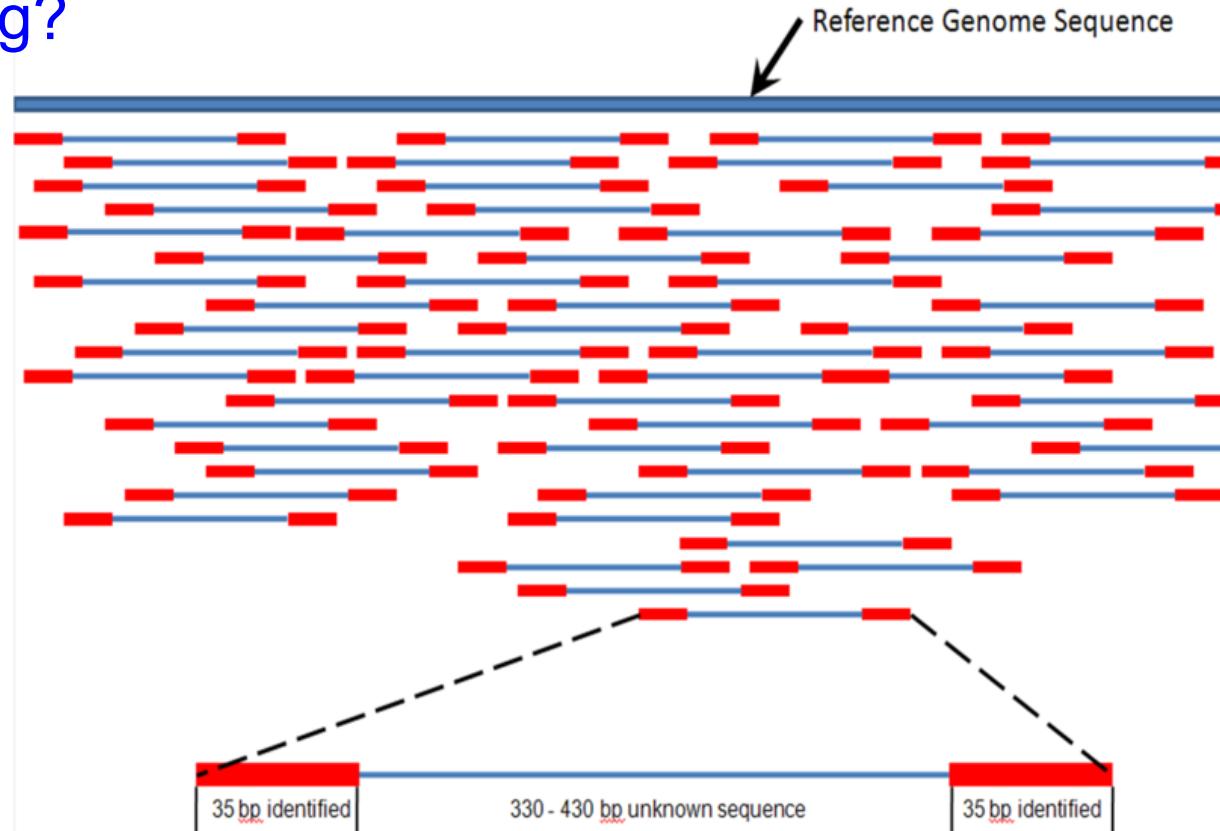


What is read mapping?

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

Please select a matrix: BLOSUM62

	A	G	P	R	W	C	D	E	H	Q	S	T	V	L	M	*	F	K	N	Y	I	Total:
A	4	0	-1	-3	0	-2	-1	2	1	0	0	-1	-4	-2	-1	-2	-1	-4	-2	-1	-2	-20
G	0	6	2	2	2	-3	-1	2	2	0	2	-3	-4	-3	-4	-3	-2	0	-3	-4	-3	-38
P	-1	2	7	2	-4	-3	-1	-1	2	-1	-1	1	-2	-3	-2	-4	-4	-1	-2	-3	-3	-36
R	-1	2	5	2	-3	-3	2	0	0	1	-1	1	3	2	-1	-4	-3	2	0	2	3	-25
W	3	-2	-4	3	11	2	-4	-3	2	2	3	2	-3	-2	-1	-4	1	-3	-4	2	3	-36
C	0	-3	-3	-3	2	9	-3	-4	-3	3	-1	-1	-1	-1	-1	-4	-3	-3	-2	-1	-1	-35
D	-2	-1	-1	-4	-4	3	6	2	-1	0	0	-1	3	-4	-3	-4	-3	-1	-1	-3	-3	-30
E	-1	2	-1	0	-3	-4	2	5	0	2	0	-1	-1	3	-2	-4	-3	1	0	2	3	-21
H	2	2	0	2	-3	-1	0	8	0	-1	2	-3	-3	-2	-4	-1	-1	1	2	-3	-21	
Q	1	-2	-1	1	2	-3	0	2	0	5	0	-1	-1	0	-4	-3	1	0	-1	3	-16	
S	1	0	-1	-1	3	-10	0	0	-10	4	1	-1	-1	-1	-4	-2	0	1	2	-1	-15	
T	0	2	-1	-1	-2	-1	-1	-1	2	-1	1	5	0	-1	-1	-4	-2	-1	0	-2	-1	-18
V	0	-3	-2	-3	-3	-1	-3	2	-3	2	-2	0	4	1	1	-4	-1	-2	-3	-1	3	-26
L	-1	-4	-3	-2	-2	-1	-4	-3	-3	-2	-2	-1	1	4	2	-4	0	-2	-3	-1	2	-29
M	-1	-3	-2	-1	-1	-1	-3	-2	-2	0	-1	-1	1	2	5	-4	0	-1	-2	-1	1	-17
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-1	-4	-4	-4	-4	-79
F	-2	-3	-4	-3	1	2	-3	-3	-1	3	2	-2	-1	0	0	-4	6	3	3	3	0	-29
K	-1	2	-1	2	-3	-3	-1	-1	-1	1	0	-1	-1	-1	-1	-4	-3	5	0	-1	-3	-21
N	0	0	-2	0	-4	-3	1	0	1	0	1	0	-3	-3	-4	-3	0	6	-2	3	-22	
Y	-2	-3	-3	-2	-2	-3	-2	-1	-2	-1	-1	-1	-1	-1	-4	3	2	-2	7	1	-20	
I	-1	-4	-3	-3	-3	-1	-3	-3	-3	-2	-1	-3	2	1	-4	0	-3	-3	-1	-4	-3	-31

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 C092
GATCTTTTATTAAACGATCTTTATTAGATCTTATTAGGATCATGATCCTCTGGATAAGTGAT
TATTACACATGCCAGATCATATAATTAAAGGAGGATCGTTGTGAGTGACCGGTGATCGTATTGCGTAT
AAGCTGGGATCTAAATGGCATTTATGCACAGTCACTCGGCAGAACATCAAGGTTGTTATGTGGATATCTAC
TGGTTTACCTGCTTTAAGCATAGTTATACACATTGCTCGCCGATCTTGAGCTAATTAGAGTAAA
TTAACCAATTGGACCCAAATCTCTGCTGGATCCTCTGGTATTTCATGTTGGATGACGTCATTCTA
ATATTCACCCAACCCTTGAGCACCTTGTGCGATCAATTGTTGATCCAGTTTATGATTGACCCGAGAA
AGTGTCAATTCTGAGCTGCCAAACCAACCGCCCCAAAGCGTACTTGGGATAAATCAGGCTTTGTTGT
TCGATCTGTTAATAATGGCTGCAAGTTATCAGGTAGATCCCCGGCACCAGTGGGATGTCACGATTA
ACCACAGGCCATTAGCGTAAGTCGCAACTCTGGGCACTGAAGTATTCTGAGAAAACCCAGCTTC
TTCAATTTCAGCTAAATGTTAGCAACATATTAGCACTACCAAGCGTACTGCCACTTATCAACGTT
ATGTCAGCCATTCAAGAACCAACTGAAGTAAAGAGCTGGATTGACTCTGTGAATCAGCTGGATCTA
```



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: GACTGGCGATCTGACTTCG

Reference: GACTGCGATCTGACATCG

Alignment:

Read: GACTGGCGATCTGACTTCG

||||| ||||||| |||

Reference: GACTG--CGATCTGACATCG

Local alignment:

Can exclude bases at the ends of the read

Read: ACGGTTGCCTTAATCCGCCACG

Reference: TAACTTGCCTTAAATCCGCCTGG

Alignment:

Read: ACGGTTGCCTAA-TCCGCCACG

||||||| |||

Reference: TAACTTGCCTTAAATCCGCCTGG

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 ($Q = -10 \log_{10} p$)
 - 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

RNAME, POS,MAPQ

QNAME @HD VN:1.0 SO:coordinate
@SQ SN:chr20 LN:64444167
@PG ID:TopHat VN:2.0.14 CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-realign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714 16 chr20 190930 3 100M * 0 0
CCGTGTTAACAGGTGGATGCGGTACCTTCCCAGCTAGGCTTAGGGATTCTAGTGGCCTAGGAAATCCAGCTAGTCCTGTCTCAGTCCCCCTCT
C BBDCCDDCCDDDCDDDDCDCCCDBC?DDDDDDDDDDDDDDDDCCDCDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDBDHFFFFDC@
AS:i:-15 XM:i:3 X0:i:0 XG:i:0 MD:Z:55C20C13A9 NM:i:3 NH:i:2 CC:Z:= CP:i:55352714 HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961 16 chr20 193953 50 100M * 0 0
TGCTGGATCATCTGGTTAGTGGCTCTGACTCAGAGGACCTCGTCCCCTGGGCAGTGGACCTTCAGTGATTCCCTGACATAAGGGGCATGGACGA
G DCDDDDDEDDDDDDDCDDDDDDCCDDDCDDDEEC>DFFEJJJJJIGJJJJIHGBHHGJJJJJJJJGJJJJJJJJHJJJJJJHHHHHFFFFFCCC
AS:i:-16 XM:i:3 X0:i:0 XG:i:0 MD:Z:60G16T18T3 NM:i:3 NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030 16 chr20 270877 50 100M * 0 0
GGCTTTATTGGTAAAAAGGAATAGCAGATTAATCAGAAATTCCACCTGGCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAACCA
C DDDDDDDDDCCDDDDDDDEEEEEEEFFFEFFFEGHHHFGDJJIHJJIIJJIIIIIGFJJJIHIIIIJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
AS:i:-11 XM:i:2 X0:i:0 XG:i:0 MD:Z:0A85G13 NM:i:2 NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699 0 chr20 271218 50 50M4700N50M * 0
0 GTGGCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam

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)