MICB405 BIOINFORMATICS

MIDTERM

October 10th, 2019

DO NOT START, until you are informed that you can start.

- You have **80 mins** to complete this closed book exam.
- Please put your name and student number on the cover page.
- Please ensure that your student number is on every page of this exam in case the pages are separated .
- There are 5 double-sided pages to this exam (including this cover page). Check that you have both sides of all question pages before you begin.
- To receive full marks, please ensure that you write legibly and in pen. We have to be able to read your answer to mark it.
- This exam is closed book and closed neighbour. Notes, books, or other materials are not allowed. Candidates guilty of any of the following, or similar dishonest practices, shall be liable to disciplinary action:
    i. Making use of any books, papers, or memoranda, calculators or computers, audio or visual players, or other memory aid devices, other than those authorized by the examiners.
    ii. Speaking or communicating with other candidates
    iii. Purposely exposing written papers to the view of other candidates. The plea of accident or forgetfulness shall not be received.
- If you have any questions during the exam, raise your hand.

GOOD LUCK!

Name: _____

Student Number:_____

This exam is marked out of a TOTAL **70 MARKS**

1. Define the following terms:

a. Chastity Filtering (**1 mark**)

Illumina sequencers performs an internal quality filtering on the first 25 bases called to detect polyclonal clusters, with one allowed failed, where (Brightness intensity)/(brightest position + 2nd brightest position) >/= 0.6 .This step flags polyclonal clusters.

b. The Kernel (**1 mark**)

The kernel of Unix is the hub of the OS. Allocates time and memory to programs. It handles the file store and communications in response to the system calls.

c. Mapping Quality (**1 mark**)

A mapping quality is assigned to the read to indicate how confident the aligner is with respect to the read mapping to its position. The negative log transformed probability that the read alignment is incorrect.

2. Name **two** types of experimental measurements for which explicit controls are absent (**2 marks**). For one measurement describe 2 ways that the results can be assessed. (**2 marks**)

Ex: Biophysical methods: DNA sequencing , X-ray crystallography,
Experimental results assessment done by internal statistical scores, reproducibility and expected outcomes. ie. For DNA sequencing, we expect 4 bases, GC content etc…

3. **You are a microbiologist working at the Centre for Disease Control and have been sent to the respiratory ward of the General Hospital to investigate a bacterial infection outbreak in the patients. A phlegm sample collected from the lungs of an infected patient was used to generate a sequencing library that was subsequently sequenced on an Illumina HiSeq platform using paired-end chemistry.**

a) Describe (a figure might help) the computational steps in the correct order that you would perform to align the resulting fastq file(s) to a reference genome to generate an indexed bam file. (**4 marks**)

align reads, using bwa mem or bwa aln need to convert sai to sam for bwa, convert to bam, sort and index:

- bwa mem indexed_genome r1.fq r2.fq | samtools sort -n | samtools view -b > out.bam
- samtools index out.bam
- OR
- Bwa aln indexed_genome r1.fq > r1.sai ; bwa aln index_genome r2.fq > r2.sai
- Bwa sampe <prefix> <in1.sai> <in2.sai> <in1.fq> <in2.fq> > out.sam

b) The sequencing team that generated your sequencing data reports that the lane had an elevated phasing rate.

Name and describe the two types of phasing that can occur during Illumina SBS? (**4 marks**)

Phasing: a fragment being sequenced in a cluster runs behind of the current sequencing by synthesis cycle.

Prephasing: a fragment being sequenced in a cluster runs ahead of the current sequencing by synthesis cycle.

Can you think of a reason for why this run might have an increased phasing? (**1 mark**)

Flow cell was overloaded and clusters to close. Any reasonable reagent based answer would be acceptable.

c) Following the run you download the resulting fastq files to your computer. Describe the format of the fastq file. (**2 marks**) Assuming you did not index your library, how many fastq files were generated for this run? (**1 mark**)

@ sequence header
sequence
+
quality scores, ascii 32
Two files generated

d) To begin your analysis you need to know how many sequences are present in a resulting fastq file named 'F01.fastq'. Assuming you have access to a unix bash shell describe how you could calculate this value. (**2 marks**)

wc -l / 4 (many ways to to this, but counting lines and dividing by four is acceptable)

e) You next decide to look at the overall quality of the fastq files. What tool could you use to perform this analysis? (**1 mark**)

fastqc

f) Below is a quality string for one sequence in your fastq file that has passed chastity filtering. Calculate the mean base quality for this very short read. Show your work for full marks (**2 marks**).

((( )))

(= (40 – 33) = 7
) = (41 -33) = 8
3(7) + 3(8) / 6 = **7.5**

g. Being satisfied with the overall quality of the fastq file you use **BWA aln** and **sampe** to align the fastq files to a reference and generate a SAM file. Below is an excerpt from the SAM file.

```
@SQ     SN:NZ_CP012076.1  LN:4912977
@PG     ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:bwa sampe GCF_001078275.1_ASM107827v1_genomic.fna F01_R1_1M.sai F01_R2_1M.sai /
projects/micb405/data/bordetella/F01_R1_1M.fastq /projects/micb405/data/bordetella//F01_R2_1M.fastq
M01783:4:000000000-A4CKG:1:1101:1673:14600      83      NZ_CP012076.1   366929  29      250M    =       366729  -450    GT
GTCGAAAGTATGGCTGATGGCCCGGGCCAGGGGCGGACTGTCGATGACCAGCCCGAGCTCGGTGTTCAGGTGGGCCGAGCGCGGGTCGAAATTGAAGGAGCCCACGAACACGCGGTGGTCGT
CCACGGCGAAGGTCTTGGCATGCAGGCTGGAGCCCGAGCTGCCGAAGGGGCCCAGGCCGCGGTGGCGCTGGACCTCGTCGCCGGCCCGGCGCATCTCGAATAGCTGCACGCCGCTGGCCAGC
AAGG    ?BFEFFFFFFBFAB:/FFFB@@@@-EAFFFF@@F@BFFB-B@FFFFFFFB?@@;@BFB?;<?FFFFFFFFFFF@FA-=B@@@@;@@EEFFFFBFFFFFFFFA;?EFFFFF?@
@=-EE?@@@AFB?@A-GBEFGGGGGGHFGFBHGHGHGHGFEC-C??CGCC@CC?EFHHCGGF/AEGCBCCCAGC/EE/E1FGFE//C@E@/>/?>A///EA/A1/C020GFGDCF1000A0A
1AF1CB>11D@1A?AA        XT:A:U  NM:i:0  SM:i:29 AM:i:29 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:250
M01783:4:000000000-A4CKG:1:1101:1673:14600      163     NZ_CP012076.1   366729  29      200M50S =       366929  450     GT
TATATGGATATGGGCCGGGGCCGCTAGAGCAGCGGCTCGAGGGGCAGCAGCGACAGGAACCACACCGACAGGCGCTGCCAGAGGCTGGCGCCCGGTTCGCTGTCGTGGCGGATGAACCCGCC
CTCGGGCGTGTGCTCCACCCCATAGAGCCGGCCGTGTTCGTCCCGCCGCACCTGGTAGGCGTCGGGCGGGCTGCGGGTGTCGCAAGGATGGCTGATTGCCCGGGCCCGGGGCGGCCTTTCGG
TGAC    1>A1A3@311D3F311AA1000AE0AE//11D1FGC?EEE@E//>/>/>E00FEEE??ECBGEFFEGGGGGGCAC@CCGGHGHEGFCGFCGGGCG?C@?C-@<.<GC..CGC
?;?-/C0CE--@--9-.9------//9;/-9A---/;//;;-;-9---99/9-9B--9-9-99-//A/;/-/-:9-----99-9----9---/9;;B---//:-;A9//9/9;:---;9---
;--;-9-///--;--9        XT:A:M  NM:i:9  SM:i:29 AM:i:29 XM:i:9  XO:i:0  XG:i:0  MD:Z:124G9G6G2A22A2A4A13C4A5
```

i) What is the length of the reference sequence used in this alignment? (**1 mark**)

ref length: 4912977

ii) Are the two reads shown paired – provide an explanation for your answer. (**2 marks**)

yes, same read name (same x y coordinate) . reciprocal read positions etc

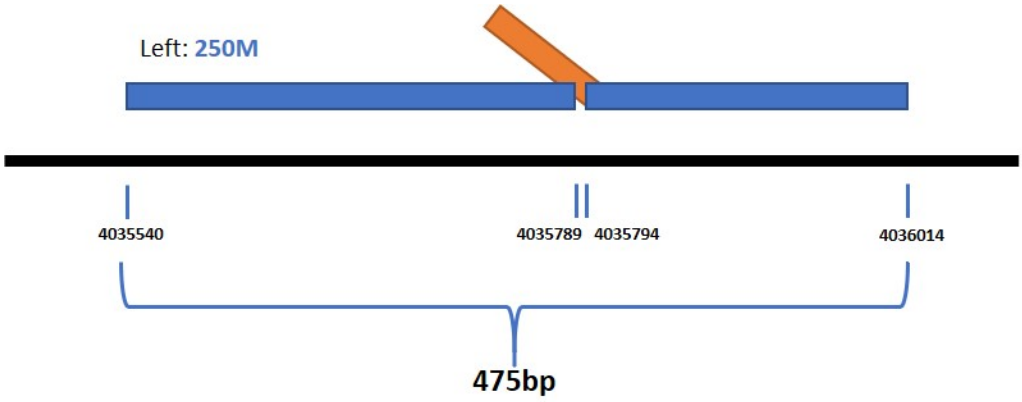
iii) What is the observed template length for the read pairs shown below? Show your reasoning for full marks (**4 marks**)

```
M01783:4:000000000-A4CKG:1:1101:1685:16112      83      NZ_CP012076.1   4035794 29      29S221M =       4035540 [redacted]      C
CAGCGCTTTGACGCCATGCTGCGGGCGGGACTGGTGGGCGGGGTGCGCCGCCTGCATGCCCGCGGCGACCTGCCCCCGGGCCTGCTTTGGGTGCGGTGCGTGGGCTACCGCCGGATATGGG
CCACCTCGATGGGGAAATCGGGCTGGAACAGGCGGGGGAGCGGGGCATCGCCGCGACCCGCCAGCTCGCCAAGCGGCAGCTTACCTGGCTGCGCGCCCAGCCCGGCCGCCTCATCGTCGAT
GCCT    ----;--/;--@-=F9/9--;----@-;/BBFFE:;;-9----@==-999--?B/F?B;@-9-@-A9-;-9--9--9--FE;-99:/-;--@-9@-9-@9-9-BB9:;---
--/BB/B-@9---9A@-;-9--;///;-9--BEB;;00;;..@--..9.9-9-.GC..-:-<--</<//C?EBBEE>>/B//E>>/GE1FGDF@1AB0HE@E>EA/////A?//EA00A0
1AA11>?1FF?AA111        XT:A:M  NM:i:13 SM:i:29 AM:i:29 XM:i:13 XO:i:0  XG:i:0  MD:Z:8C1A2A32A11C2C6C16A21C11C9C1C4A84
M01783:4:000000000-A4CKG:1:1101:1685:16112      163     NZ_CP012076.1   4035540 29      250M    =       4035794 [redacted]      C
CCGGCCCTGCACGCCGAACTGGCGCGGCACGATCCCGTGACGGCGGCGCGCCTCTCGCCCAATGACAGCCAGCGCATCCAGCGCGCGCTGGAGATCTGCCTGCTCTCCGGCCAGCCTATGT
GGCCCTGCTCACCGGCCAGGACCGCCGCGCCCGACACCGGCCATCGTTTCGTGACCATCAGCCTCGCACCCGCGGACCGCGCCGCGCTGCATGCGCGCATCGCCCAGCGCTTTTCCGCCATG
TGCC    111>11>>A10>0F0AA00AEG10B/B//AEEE?GHGG/?/FF?/E//>/>/BB?C1?/FBFCEG1DGE1?GA0F>@@ACH10C?@-A@@:;?..C/FB0FGFBFFFGGG.
;?@.;EFFFFFF-@?B@F?EBBFFF-99>=--AEF-=@@@;@@=-;=@F-@@@;/B-99EF?9-/AB/;F/;B/BF--9@@F?@;-@@;-9-@@--9;9-9F/;/AA-=---;9AA-9A;@
9-//;:AB@-/99//-        XT:A:U  NM:i:6  SM:i:29 AM:i:29 X0:i:1  X1:i:0  XM:i:6  XO:i:0  XG:i:0  MD:Z:0G188A17G28G0A11A0
```

TLEN = (LEFPOS – RIGHPOS) + (LEFTMOSTCIGARSTRING from first starting match))

TLEN= 4035794 – 4035540 + 221 =475. Tlen is 475

iv) Draw the alignment of the sequence reads relative to the reference (shown as line below) indicating as much information as possible from the SAM file entry directly above (**4 marks**)



h. To save space you convert the SAM file into a BAM file and sort the file by reference position. What tool could you use to do these two steps? (**2 marks**)

samtools sort
samtools view

samtools sort out.sam | samtools view -b > out.sorted.bam
many other combinations acceptable

4. Define batch effects (**1 mark**), what can cause batch effects (**1 mark**) and two ways to minimize batch effects in experimental design (**2 marks**).

Batch effects are sources of variation that are "unrelated to the biological or scientific variables in a study". Technical variabilities that potentially contribute to batch effects include experiments conducted by different personnel and labs, different experimental/sample prep dates, different sample processing methods/reagents/equipment. Ways to minimize include replication, randomization, blocking independent biological replicates etc…

5. List two differences and two similarities between STAR and BWA (**4 marks**).

BWA aligner: index/hashing-> fm-index , Algorithm: Burrow-wheeler transform and smith-waterman method prefix/suffix matching algorithms. Bwa generally used/designed for DNA alignment, does not handle intron-sized gaps well.

Star aligner: Designed to specifically address challenges of RNA-seq data mapping, splice aware alignment. Algorithm: seed finding done by sequential search for a maximal mappable prefix (MMP). This done using a uncompressed suffix array, also makes the algorithm very fast. Then a Clustering, stitching and scoring algorithm is applied. Can use long cDNA reads.

6. Name the 3 parts of the UNIX operating system. (**3 marks**)

The kernel, the shell, the programs

7. Where on the sequence read is the seed region extracted for sequence alignment in **bwa** and what is the default seed length in **bwa aln**. (**2 marks**)

Extracted from 5' end of read. Default seed length is 32bp

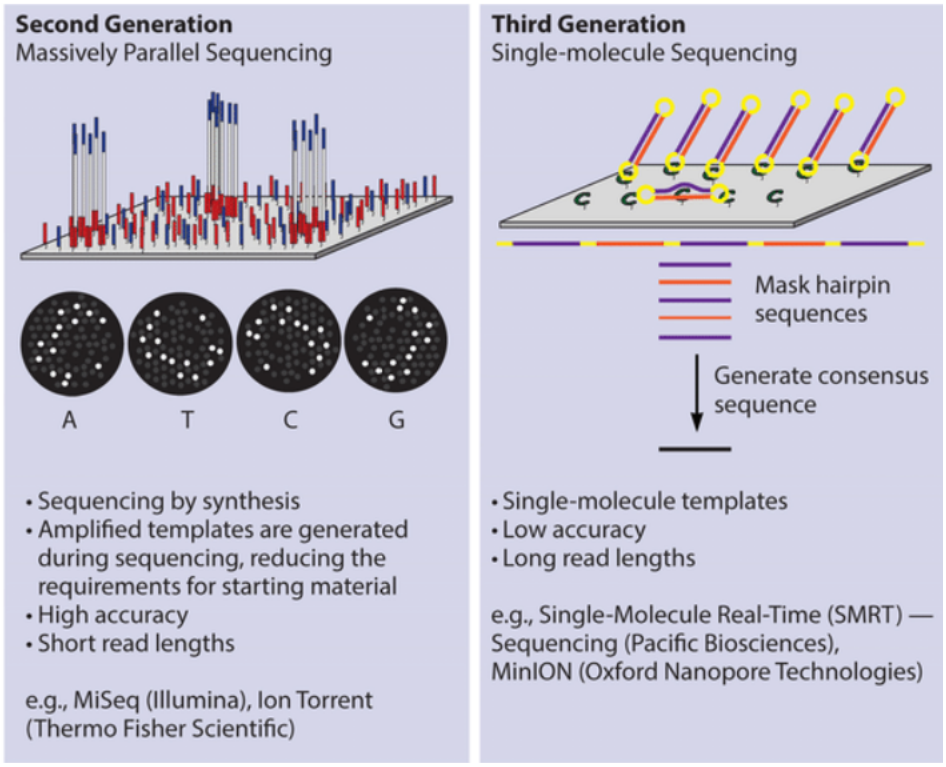
8. List two reasons for why RNA-Seq stranded library preparations are more advantageous than un-stranded library protocols (**2 marks**)? Do you need to align sequencing reads from stranded library preps differently when you use STAR (**1 mark**)? How about when you use HTSeq (**1 mark**)?

saves information of which strand the mrna is derived from (the + or – strand)

can differentiate and quantify overlapping features

you do not need to specify strandedness for the STAR aligner.

9. List 3 differences between second and third-generation sequencing platforms. (**1 mark for each difference, 3 marks total**)



10. You have been accepted to a progressive graduate program that includes bioinformatics training as part of the core curriculum. You have been provided with a user name and password for the University's unix server.

a. When you ssh into the server what command(s) would you run to view the contents of the root directory? (**2 marks**)

ls /

b. What command can you run to list **the permissions** of **all** the files in the /scripts directory? (**1 mark**)

ls -al /scripts

c. You find a file called tellmethetime.sh in the /scripts directory that is not owned by you and has the following permissions:

```
-rw----rw-
```

Translate the permissions in the space below. (**2 marks**).

user: can read, can write
group: cannot do anything /no permissions
other users: can read, can write

You less into the file and confirm that the code will indeed print the current time to a file called currenttime.txt. What will happen if you ran the command given below? (**1 mark**)

```
/scripts/tellmethetime.sh
```

Nothing happens, you cannot execute the file.

11. ENTREZ uses a combination of hard links and neighbors to link entries across databases. Define and provide an example of a hard link and neighbor in ENTREZ. (**4 marks**)

Hard link: direct connections between entries in different databases. Examples: Link to paper describing a nt seq, link to taxonomy database for a protein query, link from a nt seq to its protein CDS, link from protein seq to 3d structure.

Neighbors: entry in another dataset with subjective / similar connections. Examples: sequences similar to a nucleotide or protein query, related papers, similarity between protein structures.

12. During sequence alignment using bwa aln a 100 nt sequence read has aligned to two positions in the reference. Both alignments have a single but different mismatch to the reference. In **position 1**, the base quality of the mismatched base is 10 and in **position 2** the base quality of the mismatched base is 40. Which alignment position(s) will be reported in the SAM file and why? (**3 marks)**

Position 1, would be reported. The mismatch at position 1 has a lower quality value then the mismatched base at position 2. This means the base at position 1 is more likely to be a sequencing error rather than position 2 where the base call has a higher quality call.

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

SAM file column descriptors

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	Null	32	20	Space	64	40	@	96	60	`
1	01	Start of heading	33	21	!	65	41	A	97	61	a
2	02	Start of text	34	22	"	66	42	B	98	62	b
3	03	End of text	35	23	#	67	43	C	99	63	c
4	04	End of transmit	36	24	$	68	44	D	100	64	d
5	05	Enquiry	37	25	%	69	45	E	101	65	e
6	06	Acknowledge	38	26	&	70	46	F	102	66	f
7	07	Audible bell	39	27	'	71	47	G	103	67	g
8	08	Backspace	40	28	(	72	48	H	104	68	h
9	09	Horizontal tab	41	29	)	73	49	I	105	69	i
10	0A	Line feed	42	2A	*	74	4A	J	106	6A	j
11	0B	Vertical tab	43	2B	+	75	4B	K	107	6B	k
12	0C	Form feed	44	2C	,	76	4C	L	108	6C	l
13	0D	Carriage return	45	2D	-	77	4D	M	109	6D	m
14	0E	Shift out	46	2E	.	78	4E	N	110	6E	n
15	0F	Shift in	47	2F	/	79	4F	O	111	6F	o
16	10	Data link escape	48	30	0	80	50	P	112	70	p
17	11	Device control 1	49	31	1	81	51	Q	113	71	q
18	12	Device control 2	50	32	2	82	52	R	114	72	r
19	13	Device control 3	51	33	3	83	53	S	115	73	s
20	14	Device control 4	52	34	4	84	54	T	116	74	t
21	15	Neg. acknowledge	53	35	5	85	55	U	117	75	u
22	16	Synchronous idle	54	36	6	86	56	V	118	76	v
23	17	End trans. block	55	37	7	87	57	W	119	77	w
24	18	Cancel	56	38	8	88	58	X	120	78	x
25	19	End of medium	57	39	9	89	59	Y	121	79	y
26	1A	Substitution	58	3A	:	90	5A	Z	122	7A	z
27	1B	Escape	59	3B	;	91	5B	[	123	7B	{
28	1C	File separator	60	3C	<	92	5C	\	124	7C	|
29	1D	Group separator	61	3D	=	93	5D	]	125	7D	}
30	1E	Record separator	62	3E	>	94	5E	^	126	7E	~
31	1F	Unit separator	63	3F	?	95	5F	_	127	7F	□

ASCII Table