MICB405 BIOINFORMATICS

MIDTERM

September 27th, 2018

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

- You have 1.5 hours 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. Mapping Quality (**1 mark**)

Negative Log transformed probability that the read alignment is incorrect

b. Base Quality (**1 mark**)

Negative Log transformed probability that the base is called incorrectly

c. Somatic Variant (**1 mark**)

A sequence difference from the reference that is not present in the germline: i.e. one that arose following fertilization

2. List the two main divisions of the computer processing unit and describe what action(s) they control. (**4 marks**)

Arithmetic Logic Unit (1 mark)

- Where all arithmetic and logic operations take place (0.5)

Control Unit (1 mark)

- Computers 'nerve centre' (0.5)
- Controls order of operation (0.5)
- Accesses, interprets and directs instructions (0.5)

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. You have been given a phlegm sample collected from the lungs of an infected patient.**

a) Describe (a figure might help) the molecular steps that you would perform to generate a library from the genomic DNA sample suitable for sequencing on an Illumina sequencing platform. (**4 marks**)

Shear DNA to ~300bp mean (anything in that range is OK for full marks) (1 mark)
End repair (0.5); A-tail (0.5)
Ligate adapters (1.0)
PCR amplify with primers that extend the adapters to allow for cluster generation (1.0)

b) You decide to sequence the resulting library using paired-end sequencing chemistry on an Illumina sequencer and the sequencing team reports that the run has a high number of reads that have failed chastity.

How are the chastity values calculated and what do they indicate? (**3 marks**)

Brightest intensity/ (Brightest intensity + second brightest intensity) >= 0.6 (1 mark)

Over first 25 bases, 1 allowed failure (1 mark)

Flags polyclonal clusters (1 mark)

Can you think of a reason for why this run might have an increased number of chastity failed reads? (**2 marks**)

Flow cell was overloaded, so many clusters were too close.

c) Following the run you download the resulting fastq files to your computer. Describe the format of the fastq file. (**2 marks**) How many fastq files were generated for this run? (**1 mark**)

Four line file
@mysequence name
ATCACTCAACA
+
Base qualities encoded in ASCII base 33

d) Write a UNIX command that you could use to write the first 1000 sequences from a fastq file called ‘F01.fastq’ to a file named ‘sequence.check.fastq’. (**2 marks**)

```
head -4000 F01.fastq >sequence.check.fastq
```

many other solutions. (0.5 marks off for each error e.g. -1000)

e) To begin your bioinformatic analysis you decide to look at the overall quality of the fastq files. What tool could you use to perform this analysis? (**1 mark**) Name two features and their expected values/ranges that are produced from this tool. (**2 marks**)

Fast QC

Any 2 of the list below – with reasonable ranges.

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✗ Adapter Content

- ❌Kmer Content

f) Below is a quality string for one sequence in your fastq file that has passed chastity filtering. Would you include this sequence in your analysis? Explain your answer for full marks (**2 marks**).

AABBCCDD

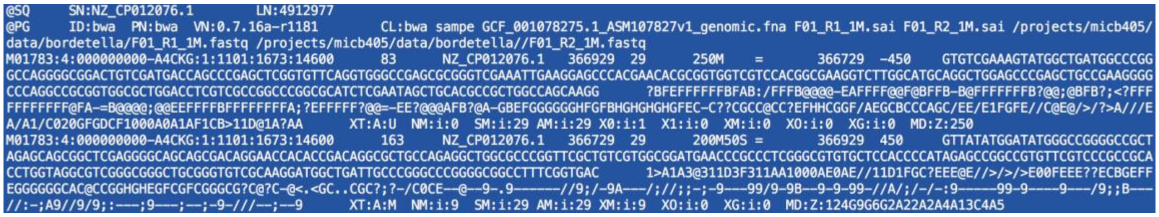
A = 65-33 = 32
B = 66-33 = 33
C =67-33 = 34
D = 68-33 = 35

64 + 66 + 68 + 70 = 33.5

Average base quality is 33.5, an error probability of ~1/1000. (1 mark for some portion of this)

This is within the average range of qualities form the illumina sequencer thus I would use (or some other reasonable explanation). (1 mark).

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:0.7.16a-r1181        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    GTGTCGAAAGTATGGCTGATGGCCCGG
GCCAGGGGCGGACTGTCGATGACCAGCCCGAGCTCGGTGTTCAGGTGGGCCGAGCGCGGGTCGAAATTGAAGGAGCCCACGAACACGCGGTGGTCGTCCACGGCGAAGGTCTTGGCATGCAGGCTGGAGCCCGAGCTGCCGAAGGGG
CCCAGGCCGCGGTGGCGCTGGACCTCGTCGCCGGCCCGGCGCATCTCGAATAGCTGCACGCCGCTGGCCAGCAAGG      ?BFEFFFFFFBFAB:/FFFB@@@@-EAFFFF@@F@BFFB-B@FFFFFFFB?@@;@BFB?;<?FFF
FFFFFFFF@FA-=B@@@@;@@EEFFFFBFFFFFFFFA;?EFFFFF?@@=-EE?@@@AFB?@A-GBEFGGGGGHFGFBHGHGHGHGFEC-C??CGCC@CC?EFHHCGGF/AEGCBCCCAGC/EE/E1FGFE//C@E@/>/?>A///E
A/A1/C020GFGDCF1000A0A1AF1CB>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     GTTATATGGATATGGGCCGGGGCCGCT
AGAGCAGCGGCTCGAGGGGCAGCAGCGACAGGAACCACACCGACAGGCGCTGCCAGAGGCTGGCGCCCGGTTCGCTGTCGTGGCGGATGAACCCGCCCTCGGGCGTGTGCTCCACCCCATAGAGCCGGCCGTGTTCGTCCCGCCGCA
CCTGGTAGGCGTCGGGCGGGCTGCGGGTGTCGCAAGGATGGCTGATTGCCCGGGCCCGGGGCGGCCTTTCGGTGAC      1>A1A3@311D3F311AA1000AE0AE//11D1FGC?EEE@E//>/>/>E00FEEE??ECBGEFF
EGGGGGGCAC@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 was the name of the fastq file(s) used in this alignment? (**1 mark**)

F01_R1_1M.fastq (0.5)
F01_R2_1M.fastq (0.5)

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

Yes (1 mark)
Bit flags, read names the same, reciprocal read postions (any of these; 1 mark)

iii) To what positions on the reference do the 5' ends of the sequences shown above align to? Show your reasoning for full marks (**4 marks**)

+ strand reports 5' position (1 mark)
- strand reports 3' position (1 mark)
Read position in 4th column

366729 (+ strand) (1 mark)
366929 + 250 (from cigar string) = 367,179 (1 mark)

367,179 – 366729 = 450

iv) Do both sequence reads **fully** match the reference - provide an explanation for your answer. (**2 marks**)

No (1 mark)

Cigar string of + strand alignments shows 50 nts were soft clipped from the 3' end (1 mark)

h. To save space and prepare for variant calling 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 (0.5) view (0.5)
Samtools (0.5) sort (0.5)

4. How is sequence indexing performed on an Illumina sequencer and what can it be used for? (**4 marks**)

Sequence indexing is performed through the addition of a sequence barcode in the adapter (2 mark). A 3rd sequence read is performed (this is 2nd in order, so will also accept this), that reads the index (1 mark). The index is associated with the sequence read(s) through the sequence name encoding lane, tile, x, y. (1 mark)

5. Describe the differences between a global and local alignment and name a program that performs each (**4 marks**)

Global optimizes the alignment of the entire sequence (1 mark) – clustal (1 mark)
Local alignment optimizes alignment of sub-strings but does not try to optimize entire string (1 mark) – BLAST or BWA (1 mark).

6. Name two types of biophysical methods for which explicit controls are absent. (**2 marks**)

DNA sequencing (1 mark)
Protein structure (1 mark)
Many others…

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**)

5' end of the read (1 mark)
32 is the seed (1 mark)

8. Name the primary nucleotide sequence databases and why they were established. (**4 marks)**

Genbank, DDBJ, ENA (1 mark each)

To share sequence datasets emerging from the human genome project. (or just the open sharing of sequence dataset; 1 mark)

9. Compare and contrast first and second-generation sequencing platforms. (**1 mark for each difference, 4 marks total**)

1st. Longer reads, analog reads derived from a pool of PCR fragments, di-deoxy terminators (ie non-reversible), relatively expensive

2nd sorter reads, sequencing clonal copies of individuals DNA fragments (ie digital sequencing), reversible terminators, less expensive, enabling genome sequencing on the population scale

3. sequencing of single DNA molecules, very long reads.

1 mark for each correct point, 0.5 mark off for each incorrect point – for a total of 4 marks.

10. You have been accepted to a progressive medical school that includes bioinformatics training as part of the core curriculum. You have been provided with a user name and password for the hospital'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**)

cd / (1 mark)
ls (1 mark)

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

ls -al

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

```
-rw----r--
```

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

Owner can read and write
Group has no permissions
Everyone can read the file

2 marks for all correct, 0.5 off for each wrong point or for missing permissions

Can you view the contents of this file? (**1 mark**)

Yes

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: Direct connections between entries in different databases (1 mark)

- Examples 1 mark for any of the below, or other

- Link to a paper describing a nucleotide sequence
- Link to a taxonomy database for a protein sequence
- Link from a nucleotide sequence to protein CDS
- Link from protein sequence to 3D structure entry

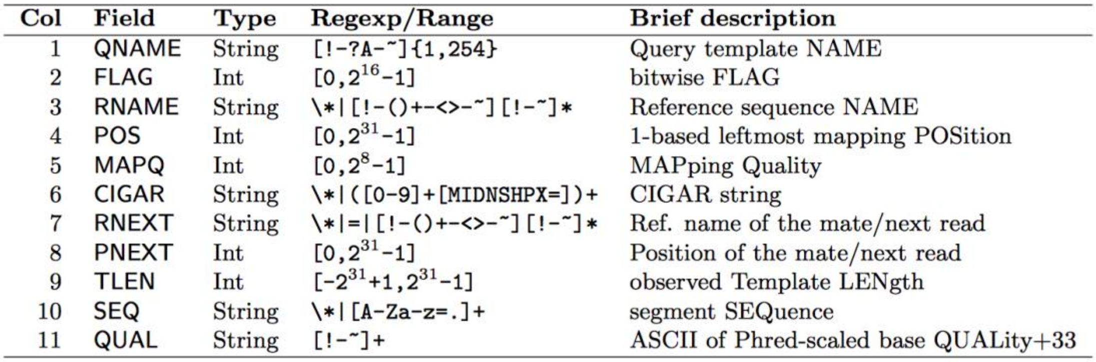
Neighbour: Subjective connections between entries. (1 mark)

- Examples 1 mark for any of the below, or other

  - Similar sequences
  - Related papers
  - Similarity in 3D structure

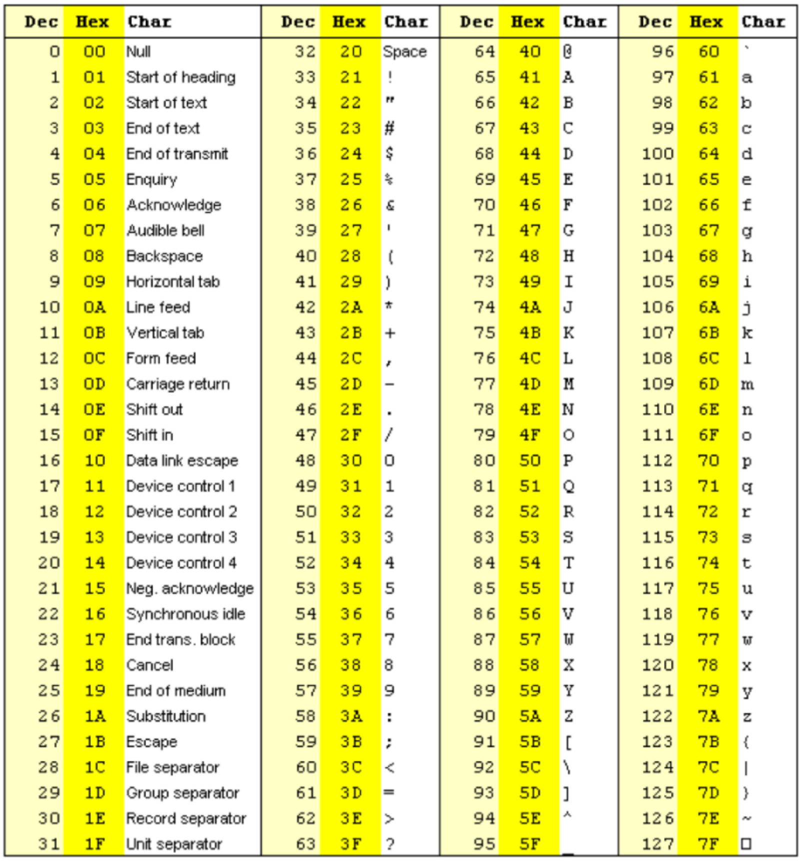
12. During sequence alignment using bwa aln a 100 nt sequence read has aligned to two positions in the reference. One sequence has a cigar string of 95M5S and the other has a cigar string of 100M. Which alignment position(s) will be reported in the SAM file and why? (**3 marks)**

100M – this is complete match and would have the highest probability of being correctly placed.



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