Student number: _____

The University of Melbourne
School of Computing and Information Systems
COMP90016 Computational Genomics
Final Examination, Semester 1, 2017

Reading time: 15 minutes.
Writing Time: 3 hours.
This paper has 16 pages including this cover page and including extra blank working pages. There are 7 questions in the paper, for a total of 60 marks, making up 60% of the total assessment for the subject.

- All answers should be written on this exam paper.
- There are extra blank pages provided at the back of this exam paper if you need extra working space. If you use this extra space to answer a question, you must clearly indicate this, and make clear, which work you wish to be considered as the answer to which question.
- Your writing should be clear; illegible answers will not be marked.
- You should show all working unless otherwise indicated.

Authorised Materials: No materials are authorised.
Calculators: Calculators are permitted.
Library: This paper is to be held by the Baillieu Library.

Examiner's use only:

1	2	3	4	6	7	Total

Page left empty intentionally.

Question 1: General Knowledge. Marks: 6 total, 1 for a-f each.

In the context of cells, explain the function of:

a) DNA

b) Genes

c) RNA

d) Proteins

Explain in a few sentences:

e) What is the purpose of sequencing?

f) What is paired-end sequencing and what is it useful for?

Question 2: Alignment. Marks: 7 total, 1 for a, and 3 for b-c each

a) What is sequence alignment for?

b) Explain the difference between the Needleman-Wunsch and Smith-Waterman algorithms in a few sentences.

c) In our lectures we have seen dynamic programming as a generally useful technique. Name one general advantage and one disadvantage of this technique. Discuss these in reference to the specific task of sequence alignment.

Question 3: SNPs and SNVs. Marks: 8 total, 1 for a, b, d, 2 for d, 3 for e

a) What is a SNP?

_____
_____
_____
_____
_____

b) What is a SNV?

_____
_____
_____
_____

c) How are SNVs discovered from sequencing data?

_____
_____
_____
_____
_____

d) What is a somatic variant?

_____
_____
_____

e) Figure 1 shows counts of reads supporting different alleles (alleles 'A' and 'T') at different loci.

- How do the different patterns (clusters) in the plot arise?

  _____
  _____
  _____
  _____

- What might the four different colours represent?

  _____
  _____
  _____

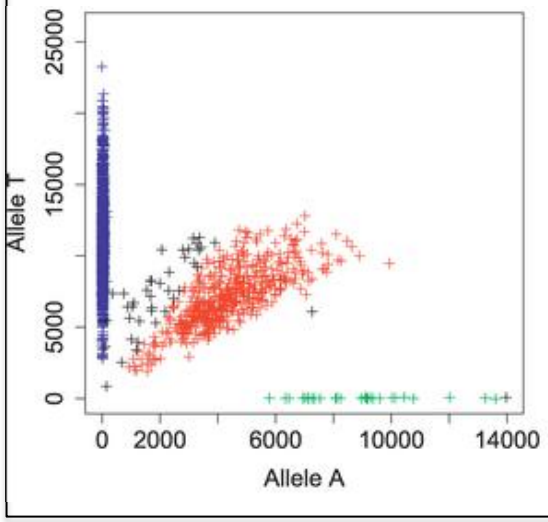

Figure 1: Read counts of reference and tumour alleles.
Note that there are four colours in this plot. Please inform the lecturer during reading time in case you have trouble distinguishing these.

Question 4: Hidden Markov Models, Marks: 10 total, 2 for a-c each, 4 for d

Suppose there has been the following discovery: There is a genetic feature that determines the outcome of students' exams. The feature is an inexact pattern of three instances of AT di-nucleotides with 3-4 bases between the first and second AT, and 6-8 bases between the second and third. None of the filling bases (between the AT pairs) is A or T. For example, such a feature could be represented by the sequence ATCCCATGCCGGGCAT.
Your task is to design a Hidden Markov Model (HMM) to detect such a feature in genomic sequences (that is, where in the input sequence are such patterns?) Draw a HMM to detect the patterns described above. Use the next page to present your solution. Make sure to include the following information:

a) Label states in a meaningful way.
b) Connect the states in the HMM with arrows, wherever a **non-zero** transition probability exists.
c) Annotate the transition arrows with the transition probability where possible.
d) Explain the design choices of your model below. If you do not have enough information to assert a transition probability, argue what the probability relates to in general.

Additional space to present HMM.

Figure 2a)

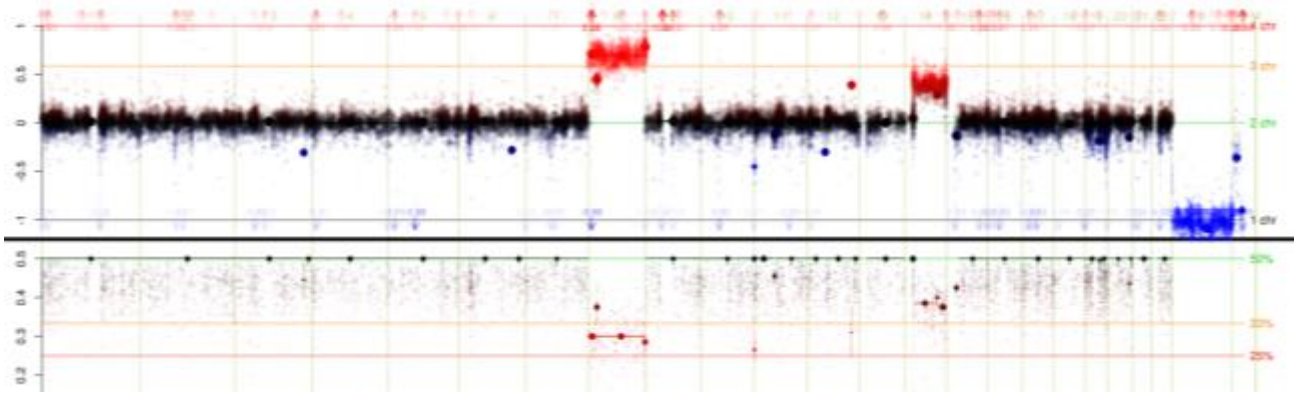

Figure 2b)

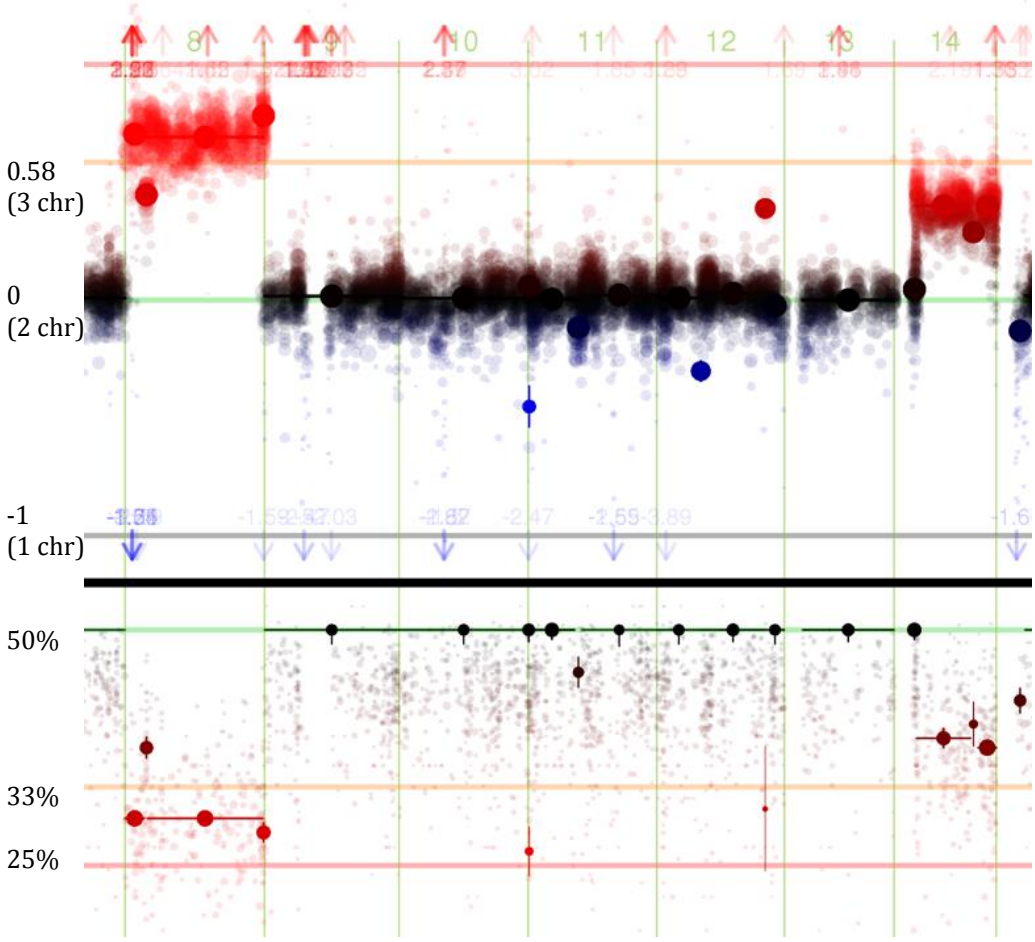

Figure 2: Read depth signal and allele frequencies of cancer exome sequencing data.
Each data point in the top panels of a) and b) represents the log2 fold-change of the read depth in an exon of the cancer sample over normal DNA.
Each point in the bottom panels represents the allele frequency of a heterozygous SNP in the cancer sample. Note that the maximum frequency shown is 50%, and frequencies higher than 50 are subtracted from 100%. For example, a SNP with 60% allele frequency would show as 40% in the plot. Further, Intervals of consistent frequencies are shown as dots (centre) and lines (width).
Part a) shows the entire genome, and part b) a zoomed in version of the same data.

Question 5: Clonality of Changes, Marks: 11 total, 2 for a and c, 3 for d, and 4 for b

Observe the Signal in Figure 2. The top track shows the log2 ratio (log-fold change – LFC) of a cancer sample's DNA over a control from the same patient. The second track shows the allele frequencies of heterozygous SNPs from the same cancer (remember that the signal is symmetrical around 50% allele frequency).
Figure 2b) shows the same data, but enhances the area of the plot around chromosomes 8-14 for better legibility.

a) In the entire genome, which chromosomes have copy numbers other than two, and what could be the reason for the difference?

b) For each of chromosomes 8 and 14, identify a possible configuration of copy number for each allele and clonality in order to explain the change of read depth and SNP frequencies (BAF).
For reference, find below the formulas to calculate logR and BAF from known clonality (*c*), and allele copy numbers *vA* and *vB:*

$$\log R = \log 2\left(\frac{(vA+vB)cn+(1-c)2n}{2n}\right) = \log 2\left(\frac{((vA+vB)c+2-2c)}{2}\right)$$

$$BAF = \frac{vB\cdot c\cdot n+(1-c)n}{(vA+vB)cn+(1-c)2n} = \frac{vB\cdot c+1-c}{(vA+vB)c+2-2c}$$

Show all your workings to arrive at your answer. You answer does not have to match the data exactly, but should be as close as possible. You might find it practical to start estimating clonality from the BAF formula and then compare your result to the observed log-fold changes.

c) Based on your results from part b) of this question, explain whether you think if the changes to chromosomes 8 and 14 are present in separate cells, or in two populations of cells that partially overlap, or in two populations where one is a subset of the other (subclone), or in the same population.
If you did not manage to infer clones and copy number in part b), you can answer the question with the following parameters for chromosome 8: c=0.25, vA=0, vB=1, and chromosome 14: c=0.5, vA=2, vB=2.

d) Draw a representative set of normal and cancer cells of the sampled cancer. The diagram should reflect (i) observed clonalities, (ii) your inference of intersection of clones, and (iii) a model of the chromosomal gains or losses of 8 and 14 below (continue with the alternative set of values provided in part c) if need be). You don't need to draw the entire genome in each cell – focus on 8 and 14 only.

Question 6: Genomic Variants. Marks: 9 total, 1 for a-d, 2 for f, and 3 for e

Both single nucleotide mutations and structural variations of the genome can have an effect on the phenotype of an organism.

Describe how this could be the case for:

a) SNVs

_____
_____
_____
_____

b) Deletions

_____
_____
_____
_____

c) Inversions (not directly within a gene)

_____
_____
_____
_____

d) Tandem duplications

_____
_____
_____
_____

The following table shows the (simplified) results from a read depth analysis of genomics data from a human sample (specifically, 1M bases of chromosome 17, from position 42,000,000 to position 43,000,000). The read depth was established as the log2 ratio of reads in bins over the average bin size in all of the data. The width of the bins in this experiment is 1000bp.

Bin number	1	2	3	4	5	6	7	8	9	10
Log2 ratio	0	0	-0.7	-1	-1	-1	-0.27	0	0	0

The scientist analyzing this data establishes that the read depth shows evidence of genomic rearrangements in the sample.

e) What would her conclusion be? What structural variant is indicated by the data? Where in the DNA would it be, and what genotype should it have?

_____
_____
_____
_____
_____

The initial sequencing was done with single-end read data. Given the evidence above, the scientist responsible for the project decides to also do paired-end sequencing on the same sample.

f) Why would paired-end data be useful in this instance?

_____
_____
_____
_____

Question 7: RNA Analysis. Marks: 9 total, 2 for a, 3 for b, and 4 for c

a) In many experiments, RNA analysis is used as a proxy for a question of interest not related to RNA. What is this question?

b) When we perform RNA sequencing, reads have to be aligned to a reference. When using the strategy of aligning sequencing reads to the whole reference genome, what makes this a challenge in organisms, such as humans?

c) Two libraries of RNA were sequenced, aligned, and counts of genes are now ready for analysis. The first library is from a healthy individual and it contains 10M reads. The second sample is from an individual with a disease and contains 20M reads. What do we have to take into account to make genes from these two sets comparable in order to explore what the differences between the healthy and diseased samples are? Name and explain two things.

Extra page for working. Please indicate the question that your working belongs to here and on the question's sheet.

Extra page for working. Please indicate the question that your working belongs to here and on the question's sheet.

Extra page for working. Please indicate the question that your working belongs to here and on the question's sheet.

Extra page for working. Please indicate the question that your working belongs to here and on the question's sheet.