Khan Inan
Transcriptomics 7653
midterm

Next Generation Sequence Analysis Midterm Exam

This is a take home exam for BI7653 Next Generation Sequence Analysis for Spring 2023.

Answers can be found directly in the course lectures, reading material, in the homework exercises, were discussed in the recorded video webinar sessions, or can be derived from concepts introduced in those materials.

Please show your work for any answers requiring calculations as this can be used to assign partial credit.

You are asked to not discuss the exam with other students.

As always, your short answers will be checked for plagiarism. You must write your answers in your own words.

The exam is due at midnight on Monday, March 20. No points will be awarded if the exam is uploaded after this time.

When you are finished, please go to the Assignments section of NYU Brightspace and upload your answers in a single attached text file. Please include your name both in the filename and inside the document itself.

The exam is worth 40 points.

Good luck!

Q1. Answer the following questions related to the sequencing by synthesis process on an Illumina HiSeq sequencer. For this question, you should refer primarily to the Illumina Sequencing Technology lecture video (Module 1) and other course materials.

Q1.1. Adapters are an essential component of sequencing-by-synthesis process whose design must be tailored to each sequencing experiment. Which of the following are components of adapter in a typical short read (Illumina) sequencing experiment of genomic DNA. Select all that apply (1 point)

(a) priming sites for PCR

(b) restriction enzyme target site for molecular cloning experiments

(c) priming sites for sequencing by synthesis

(d) short segments of DNA sequence from the genome to be sequenced

(e) barcodes for demultiplexing

Q1.2. Which of the following are true about clusters in an Illumina sequencing-by-synthesis run. Select all that apply (1 point)

(a) The templates molecules in a cluster represent two alleles from a diploid individual

(b) Template molecules include sequences corresponding to one end, but not both ends, of a paired-end sequencing run

(c) Template molecules in a cluster are covalently bound to the flow cell surface

(d) The number of template molecules in a cluster is directly related to the read depth that will be achieved once reads are aligned to the reference

(e) In a four-channel sequencer, a single cluster will ideally emit four wavelengths of light each sequencing-by-synthesis cycle

Q1.3 Select all that are true about Illumina reversible terminators or reversible terminator-bound dNTPs (1 point)

- **(a) reversible terminators are labeled with a fluorescent label**
- (b) reversible terminators are intended to ensure that at least two nucleotide are added to an elongating DNA molecule per cycle
- **(c) G nucleotides do not have a fluorescent label in 2-channel systems**
- **(d) A nucleotides do not have a fluorescent label in 2-channel systems**
- (e) C nucleotides do not have a fluorescent label in 2-channel systems
- **(f) A, T, G, C nucleotides each have a different fluorescent label in 4-channel systems**

Q1.4. Which of the following is true of PacBio sequencing. Select all that apply. (1 point)

- **(a) PacBio uses high molecular weight DNA sequencing templates**
- **(b) PacBio generates long read sequences**
- (c) PacBio DNA sequencing error rates are traditionally lower than Illumina
- (d) PacBio DNA sequencing is expected to suffer from a “dephasing” problem

Q.1.5. Sequencing-by-synthesis suffers from a “dephasing” problem. Please provide a detailed explanation of the cause of the problem, its expected effect on the output sequences and discuss how and why it impacts base qualities in the way that it does. (3 points)

The cause of dephasing is essentially differences in reaction times between the strand of DNA and its clone. If the strands reaction occurs too quickly (leading), or if the strand reacts too slowly (lagging), this would mean that the template will either not be extended or have too many extra nucleotides added. Having the templates being sequenced equally is called making the extension synchronous and this is done using dephasing algorithms. The way it affects base quality is due to contamination which prevents the reactions from occurring correctly and on time.

Q2. Sequence depth is an important concept in NGS. Please answer the following questions related to sequencing depth.

Q2.1.

What is the coverage (i.e., average coverage) of the following hypothetical alignment of the hypothetical reads of length 5 bp against the reference genome? (1 point)

```
      TATATA
     CTATAT
    CCTATA
   CCCTAT
   CTCTAT
 TGCCCT
 TGCTCT
ATGCCC
ATGCCCTATATA (reference genome) # this is not a read
```

(3+4+6+5+6+5+3+2+1+1)= 36
36/10 = 3.6

The coverage of the following hypothetical alignment against the reference genome is 3.6

Q2.2 Which of the following can affect read depth in the reads of a sample genome after alignment to a reference genome? Select all that apply (1 point)

- **(a) The total number of reads sequenced for the sample**
- **(b) Chance (i.e., Poisson, or Poisson-like variability)**
- **(c) Genotype Quality**
- **(d) A duplication of a gene region in the sample relative to the reference**
- **(e) Base quality score recalibration (BQSR)**

Answer is all of the above

Q2.3 Imagine you are assisting in planning a sequencing run of a population of human genomes. The core facility manager informs that you can expect a total of 2,500 Gb of total sequence from a 2 x 150 run on a single S4 flow cell on a Novaseq 6000 sequencer. Using 3.212 Gb as the size of the human reference genome, please answer the following questions.

Q2.3a. What is the maximum number of human samples you could multiplex in one run of the sequencer and expect to obtain at least 25X coverage in each sample? Please show your algebra for full credit (1 point).

2500/(3.212 x 25)
2500/(80.3) = 31.133

The maximum number of human samples that you can multiplex in one run of the sequencer and still obtain 25X coverage is about 31

Q2.3b. If you sequenced the number of genomes from your answer in 2.3a, what is the expected the coverage depth of each sample in the experiment? (1 point)

2500/(31 x 3.212)
2500/(99.572) = 25.107

From sequencing the number of genomes from my previous answer, the expected coverage depth in each sample of the experiment would be about 25.1X

Q2.4 Now imagine you would like to conduct a separate Genomewide Association Study (GWAS) experiment by sequencing a large panel of human genomes at 4X coverage each. For Q2.4a and Q2.4b Please only use the information provided to answer the question and assume a Poisson distribution of read depth. Please show your arithmetic to receive credit.

Q2.4a. Assuming you attain your goal of 4X coverage in any one genome, what percentage of the genome do you expect to not be sequenced for each individual (and therefore have missing genotypes)? (1 point)

4/(3.212) = 1.245
e^(-1.245) = 0.28794

The percentage of the genome that I would expect not to be sequenced for each individual would be about 28.794%

Q2.4b. In any one 4X coverage genome, what percentage of the genome do you expect will be sequenced at 4X coverage or less? (1 point)

(e^(-1.245) x 1.245^0) / 0! + (e^(-1.245) x 1.245^1) / 1! + (e^(-1.245) x 1.245^2) / 2! + (e^(-1.245) x 1.245^3) / 3! +(e^(-1.245) x 1.245^4) / 4!

e^(-1.245) x (0.28792 + 0.35848 + 0.22302 + 0.09261 + 0.02882)

e^(-1.245) x (0.99085) = 0.2853

At any one 4X coverage genome, I would expect around 28.53% of the genome to be sequenced at 4X coverage or less

Q2.5. In general, low coverage Illumina sequencing is problematic because of the uncertainty in genotype calls. Which of the following are true about low coverage sequencing? Select all statements that are true (1 point)

- **(a) Genotype qualities should on average be lower than in high coverage applications**
- (b) Genotype qualities should on average be higher than in high coverage applications
- **(c) Fewer heterozygous genotypes will be called in low coverage data due to a lower probability of detecting both parental alleles**
- (d) More heterozygous genotypes will be called in low coverage data due to a lower probability of detecting both parental alleles
- **(e) Applications that call genotypes from very low coverage sequencing should be cautious about homozygote genotype calls**
- (f) Applications that call genotypes from very low coverage sequencing should be confident in homozygote genotype calls

Q2.6. Structural variant discovery with “read depth” methods have some limitations in terms of the types variants that can be discovered and the conclusions that can be drawn about the variant. Which of the following are read depth methods effective at characterizing? Select all that apply (1 point)

(a) Determine the genome location of a duplicated copy of a gene region in a sample relative to the reference genome (e.g., if it’s a tandem duplicate or if duplicate copy is on a different chromosome from the copy in the reference genome)
(b) Determine if a deletion in a sample is heterozygous or homozygous
(c) Determine the copy number of a region that is not found in the reference genome
(d) Determine if a sample has 2 copies, 3 copies, or 4 copies of a gene that is single copy in the reference genome.

Q3. Library preparation, sequencing, data processing and analysis introduces a number of technical, machine, and bioinformatic/computational artifacts that influence the accuracy of snp-calling and genotyping

Q3.1 The homopolymer runs of Gs (polyG) are common artifacts produced by some Illumina sequencing platforms. Which of the following platforms suffer from the polyG problem? Which of the following is true of the polyG problem. Select all that apply. (1 point)

(a) The problem is apparent on HiSeq 2000/2500 platforms
(b) The problem is caused by spontaneous cytosine deamination in vitro
(c) The problem is apparent on NextSeq/NovaSeq platforms
(d) The problem is observed on 2-channel Illumina sequencers, not 4-channel
(e) The problem is mostly observed in the interior bases of reads
(f) The problem is caused by difficulty sequencing through short (< 5 bp) polyG tracts

Q3.2. Short read sequencing applications typically require trimming of adapter sequences using a tool like CutAdapt or fastp. Trimming of low quality bases from the 3' end of reads is more controversial. Why can trimming low quality bases be a problem for applications requiring short read alignment to a reference genome? What precaution(s) should be taken to ensure that over-trimming (either for adapters or base quality) does not impact downstream analysis? (2 points)

The main reason that trimming low quality bases can be a problem is because although trimming can lower the error rates, which is desirable, it can also reduce the number of reads and the lengths that are obtained. The best method for quality control in this case to prevent downstream analysis from being affect is to manipulate the quality threshold so that there is no over-trimming or "under-trimming" and the right amount of bases are removed.

Q3.3. In addition to issues raised above, many NGS sequencing applications suffer from the following artifacts. Please provide a detailed description of each of the artifacts listed below including why it occurs and how you might reduce the impact of the artifact on your analysis either via a software tool or change in experimental workflow (6 points)

Q3.3a PCR duplicates
Note: students should receive full credit for mentioning either digital counting applications or re-sequencing.

PCR duplicates are when you have two copies of the base molecule spread onto different locations/lanes on a flow cell. When you have a high rate of PCR duplicates, it either means that there is a large variance in fragment size or there is not enough starting material. These problems result in needing to over-amplify what little sample there is and thus also results in over-representation of smaller fragments (since they are easier to

amplify.) The most common method to deal with the PCR duplicates artifact is to remove all the duplicate sequences except one.

Q3.3b Read alignment error

Read alignment error occurs essentially when there are very similar or identical regions on the same genome and it confuses the alignment tool. The reason for this lies in how much the alignment tool thinks that the read is uniquely mapped, and if it overestimates its "uniqueness" value, then that's when this artifact can arise. By using an alignment algorithm that is suited to this kind of application, and by specifying the settings that apply to sensitivity, we can alleviate this issue. Additionally it is good practice to map according to the available reference genome.

Q3.3c Index swapping on patterned flow cell type Illumina sequencers

Index swapping occurs mainly during multiplexing and it is the result of mistakes in the assignment of libraries from the correct index to a different one. Index swapping can cause reads to be assigned to the wrong index during demultiplexing and also it can lead to poor data quality and misalignments. Regardless of precautions, some level of index hopping is typical for any application and it is usually around 2% or less. The best practices to alleviate index hopping artifacts on a software level is to use noise filtering methods and remove free adapters, and on a physical level it is best to store libraries/samples properly and pool them before sequencing takes place.

Q4. SAM format is a primary means by which sequences aligned to a reference genome are recorded. Use the following example of a SAM record to answer the questions below (8 points).

```
SRR062634.11560830    83    20    60015  60    100M  =    59999  -115
AGAGGAAGGAAGCTTGGAACCCTATAGAGTTGCTGAGTGCCAGGACCAGATCCTGGCCCTAAACAGGTGGTAAGGAAGGAGAGAGTG
AAGGAACTGCCAG
DADD=AEAAC@)BBFAA4BFAFBIDJHDBJGIAKHHFFBHGHGDDGGHEEEGDEHGHHKPPPPQRQPLRPPPRQQPRRQRPRQRPRMPRQPO
OPQPNLL0   X0:i:1  X1:i:0  MD:Z:100      RG:Z:SRR062634  AM:i:37 NM:i:0  SM:i:37 MQ:i:60 XT:A:U
BQ:Z:@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
```

Q4.1 What is the mapping quality of this alignment? (1 point)

The mapping quality of this alignment is 60

Q4.2 What is the probability that the read errantly aligned? (1 point)

The probability that this read is errantly aligned is most likely 0.000001 based one the "phred quality/errant probability" chart

Q4.3 Was the read in the alignment shown reverse complemented by the alignment software (1 point)

(a) TRUE
(b) FALSE

Q4.4a What chromosome (or chromosome identifier) and position did the read in the alignment above map to? (1 point)

The RNAME column (column 3) contains the chromosome which it is mapped to which in this case is 20

Q4.4b. What chromosome (or chromosome identifier) and position did the paired read for this read alignment map to? (1 point)

The chromosome/position for paired end reads should be the TLEN field (column 9) which in this case is -115

Q4.5 True or false, this read is marked as a PCR duplicate? (1 point)

I dont see any evidence of this read being marked as a PCR duplicate because it doesn't contain the duplicate flag so False

Q4.6a Is the aligned read either soft- or hard-clipped? Please explain how you arrived at your answer (1 point)

The aligned read seems to be soft clipped because the cigar string is 100M, and after the header portion of the sam record we can see a long string, which implies that the sequence was retained and thus this is an example of soft-clipping

Q4.6b What is the difference between a soft- or hard-clipped read. (1 point)

Soft clipped sequences are retained with the cigar string while hard-clipped strings are not retained

Q5. The VCF format is a feature-rich format used to record genome variation including snps, indels and structural variants. Please use the VCF snippet below to answer the following questions about the VCF format and the variant at Chr1, position 47 (8 points).

```
##fileformat=VCFv4.1
##FILTER=<ID=lQ,Description="SNP quality < 100">
##FILTER=<ID=lDP,Description="Depth < 40 across samples">
##FILTER=<ID=hDP,Description="Depth > 100 across samples”>
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
#CHROM  POS ID REF ALT  QUAL  FILTER INFO          FORMAT       sample_1          sample_2.          sample_3
Chr1    47  . T   A   50    lDP    AC=3;DP=37 GT:AD:DP:GQ:PL 0/0:13,1:14:9:0,9,44  1/1:0,15:15:12:36,12,0  0/1:5,3:8:19:19,0,24
```

Q5.1 Is the variant a SNP, indel, or structural variant?

This variant is an SNP

Q5.2. What is the probability that the variant is an error. Please assume that the PHRED-scaled variant quality is accurate?

The Quality value at the bottom is 50, so the probability that the variant is an error is most likely 0.00001

Q5.3. What filter(s) did the variant not pass? What is the description of this filter(s)?

It did not seem to pass the “lQ” and “hDP” filters. The lQ filter is shown in the header and it is responsible for filtering out SNPs that have a quality score of less than 100. From what I understand the lDP filter did pass because there was apparently depth<40 across the samples and it is shown at the bottom under FILTER INFO. hDP, however, did not pass because there was no depth>100 across samples

Q5.4 What are the predicted genotypes of sample_1, sample_2 and sample_3? Please provide your answer in terms of nucleotides (A,T,G,C). If the genotype is missing data, please indicate so.

Sample_1 is homozygous for its reference allele of 0/0, since the value is 0/0 for the “GT” field at the bottom

sample_2 is homozygous for its alternate allele since it has a 1/1 in it’s GT field at the bottom

lastly sample_3 is heterozygous for its alternate allelesince it has a 0/1 for its GT field at the bottom

Q5.5 What are the genotype qualities of each of the 3 samples? Which genotype has the lowest probability of being called in error?

The genotype qualities for sample 1,2 and 3 are 9, 12 and 19 respectively based on the values shown on the bottom. Since a higher PHRED scaled score means that there is more confidence in the genotype quality, sample 3 has the lowest probability in being called in error

Q5.6 What are the sample depths for each of the samples?

Based on the approximate read depths and not the allelic depths, the depths for the samples are 14 for sample_1, 15 for sample_2 and 8 for sample_3

Q5.7. How many reads support the alternate base in sample_3?

Based on the AD field for sample_3, 5 reads support the reference base, and 3 reads support the alternate base

Q5.8 The values for many tags in the INFO column of a vcf depend on the sample fields. If samples are added or removed, the tags in the INFO column must be dynamically updated. If you were to remove the sample_3 column from the VCF, what should the updated values of the AC, and DP INFO tags be?

Since sample 3 has an AC value of 0/1 and a DP value of 8. The AC value would change from 3 to 2 and the DP would change from 37 to 29 after sample 3 is removed from the VCF