BIFX 553 – Spring 2023 - Final Exam

Student Name: J. Jedediah Smith

1. SAM and BAM files

Consider the following beginning of a SAM file:  
A picture containing text, receipt, font, white

Description automatically generated

1. What is the name of the reference sequence according to the data?

ref

1. What is the length of the reference sequence?

45

1. How many unaligned reads are shown?

4 of the 6 reads appear to be unaligned

1. What is the first position on the reference sequence where read r004 matches?

It’s first position should match, as indicated by it’s the 6M that starts its CIGAR specifier. In this case, that happens to be an A.

1. Is the SAM file sorted? Yes or no.

Yes, the SO is set to coordinate, which will sort by RNAME and then POS.

1. Read r003 matches twice in the shown part of the SAM file. Which alignment is the primary alignment, and which is the supplementary alignment indicating potential structural variations? Hint: look at the flag field. As a reminder, the flags describing an alignment are by convention:  
   A picture containing text, screenshot, number, font

   Description automatically generated

The r003 read with the 2064 flag is our supplementary alignment. It doesn’t appear to match directly with the chart, but that is because it has two flags! 2048 (Supplementary) + 16 (Reverse) = 2064. This leaves the 0 flag as the primary alignment.

1. What is the *samtools* program command line for filtering out supplementary alignments?

samtools view -f 2048 should do the job! If we wanted to filter for different things, we would just change the flag.

1. Read Alignments

Consider the following part of a read that is aligned to a reference sequence.

A picture containing text, font, white, typography

Description automatically generated

1. What is the CIGAR string representation of that alignment?

3M1I2M1D16M

1. If this read was part of a larger sequencing dataset, what tools or methods could be used to generate the given alignment between the reference and read sequences? List the names of at least 2 programs.

Samtools might be able to work, but there are probably more specialized tools that could do the job better. Picard and VariantBam are both good options.

1. Describe the potential challenges in aligning short reads to repetitive regions or regions with high sequence similarity. How might these challenges affect the interpretation of the given alignment?

In such a situation, it is likely that the short reads may match to multiple different places, making it difficult to ascertain where it actually goes.

1. Describing Genomic Polymorphisms

Another alignment between a read and a genomic DNA sequence is shown here (the genomic first position of the reference sequence is indicated as 1012):  
  
A picture containing text, font, white, screenshot

Description automatically generated

1. How many insertions and how many deletions are visible?

The read has one insertion and one deletion.

1. How many Single Nucleotide Polymorphisms (SNPs) are apparent in that alignment?

There appears to be two SNP.

1. What is the CIGAR string representation of that alignment?

6M1I2M1D12M

1. To properly describe all variations one needs to go beyond the CIGAR string representation and use a more versatile format such as VCF. Fill out below template of a portion of a VCF file that, if applicable, captions all insertions, deletions and SNPs of the sequence pair alignment shown above. Assume that the G->A polymorphism at that locus has been given the name ‘rs1234’, while the other polymorphisms have not been named yet.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHR | POS | ID | REF | ALT |
| chr7 | 1017 | . | A | AC |
| chr7 | 1019 | rs1234 | G | A |
| chr7 | 1020 | . | A | <DEL> |
| chr7 | 1026 | . | A | T |
|  |  |  |  |  |

1. GWAS
2. Describe what the acronym GWAS stands for and “unpack” the full name indicating why this is a brief description of the method.

GWAS = Genome-Wide Association Study. This is when a study looks at the whole genome of an organism to better explore the connections between certain genes or mutations and particular outcomes. A brief description does not do it justice, given how large a genome can be and the amount of data involved, not to mention things like epigenetic markers and ETS regions that may need to be taken into account.

1. Look at the GWAS results presented in form of a Manhattan plot below. The study looked at associations between genotype and a phenotype of susceptibility to infectious diseases among a cohort of British UK Biobank participants. Discuss the evidence for this association for a gene mentioned in the chart named COLQ . Discuss statistical significance as well as additional considerations that increase or decrease the merit of the evidence for this potential disease associations.A picture containing plot, diagram, screenshot, line

   Description automatically generated

There doesn’t seem to be a strong correlation with COLQ. It has only a sole datapoint that is just barely above the rest of the labeled genes. This could be more promising if there was a trail of high datapoints, but right now it seems like a chance outlier when compared to many of the other genes.

1. Suppose we are conducting a GWAS to investigate the association between a genetic variant and a disease. We genotype 2100 individuals from a population and record their genotype at the variant locus (homozygous for the variant allele, heterozygous, or homozygous for the wild type allele), as well as whether they have the disease (yes or no). The resulting contingency table is shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Genotype | Genotype | Disease | No Disease | Total |
| Homozygous- Variant Allele | 1/1 | 80 | 20 | 100 |
| Heterozygous | 0/1 | 200 | 600 | 800 |
| Homozygous Wildtype | 0/0 | 280 | 920 | 1200 |

1. Compute the odds ratio for this contingency table. What does this tell us about the strength of the association between the variant and the disease?  
   Hint: an odds ratio is a fraction of 2 fractions, therefor a story of 4 quantities. Our table has 6 quantities! To simplify, convert it back to 4 numbers assuming a simple model that the gene of interest is recessive – only the homozygous variant case will show the disease phenotype.  
   An odds ratio of (80 / 20) / (280/920) = 13.14 indicates that there is a much stronger odds of being diseased with the homozygous variant allele than the wild type.
2. Repeat the odds ratio computation but this time assume the gene of interest is dominant with respect to the phenotype (the heterozygous case will show the disease phenotype). What is the odds ratio now?  
   An odds ratio of (200 / 600) / (280/920) = 1.01 indicates that the odds of being diseased is nearly the same between heterozygous and homozygous wild type.
3. Using the chi-square statistic, calculate the P-value for this test. Is there evidence of an association between the variant and the disease?

I did a chi-square test on the data and got a p-value of 2.2e-16, which is quite small. It indicates that the association between variant and disease is a significant one.

Text

Description automatically generated

1. Now we are applying an approach alternative to the chi-square test in form of the logistic regression. In other words, we want to model the probability binary variable or a persona having or not having the disease as a function of the number of copies of the disease gene a person carries (being one of 0, 1 or 2). You can use the R function *glm* (generalized linear model) by specifying the independent and dependent variable, the parameter data for the dataframe and the parameter ‘family’ that has to be set to ‘bionomial()’. The challenge is that we cannot use the aggregate data of the contingency table directly. But we can convert back to individual-level data as shown in class. You can use R or Python (sklearn.linear\_model import LogisticRegression)  
   What are the resulting coefficients of the fitted logistic regression?  
   The intercept was -1.34 and the condition was 0.59.

Text

Description automatically generated

1. What are the predicted probabilities of having the disease for individuals with 0, 1 or 2 copies of the gene of interest?

They are as follows:

2 Copies = 80/100 = 0.80

1 Copy = 200/800 = 0.25

0 Copies = 280 / 920 = 0.34