Research Project 1B Task 2

**Question 1.1**

AGT-A

**Question 1.2**

Sample 1 = AG-TGA

Sample 2 = AGT-GA

Samples 1 and 2 are zero SNPS apart however both samples contain one deletion when compared to the reference genome.

**Question 1.3**

1. Sample 1 = AGT---A, Sample 2 = AG---TA
2. There are no differences between the genomes of sample 1 and sample 2…
3. However, when comparing the VCF files for the two samples we can see they arose by different deletions when comparing them to the reference. Sample 1 came around by the deletions of the nucleotides (ACT) at positions 4,5, and 6 from the reference strand, whilst sample 2 arose from the deletions of the nucleotides (TAC at positions 3,4, and 5.

**Question 1.4**

2 GACT G

**Question 1.5**

VCF1

AGTC---T (AGTCT)

VCF2

AGT---CT (AGTCT)

VCF3

AG---TCT (AGTCT)

Each of the genomes produced by these three VCFs (as seen above) are exactly the same. Although the deletions occur in different places within the genome, when they are removed each of the three sequences is identical (AGTCT)

**Question 1.6**

VCF1

7 GACT G

VCF2

6 TGACT TG

VCF3

VCF4

10 TACT T

VCF5

1 AGTCTTGACTACTACTACTGGG AGTCTTGACTACTACTGGG

VCF6

13 TACT T

VCF7

16 TACT T

Above are just a few examples of VCF representations that you could give to define the sample when compared to the reference. The issue with this is, as best highlighted by VCF5, there could potentially be hundreds, or even thousands of different combinations to describe one sample relative to its reference genome. To overcome this the VCF file should be normalised, as demonstrated in a paper by Tan et al., (2015). Normalisation of a variant representation consists of two parts, left alignment and parsimony. Left alignment means that its base position is the smallest possible among all potential VCF entries of the same length and for the same variant. To be parsimonious the VCF entry must have the shortest allele length possible to represent the variant. For the example in this task the correctly normalised VCF is shown in my answers as VCF1 which is underlined.

**Question 2**

1.

Firstly, for each sample, the reads were all aligned with the following command (with the files change appropriately for each sample.

Then the variants were called and saved to a .vcf file using the following command:

2.

I believe that sample 1 is different from the other two. The reason for this is that every single one of the variants called in samples 2 and 3 are exclusively single nucleotide polymorphisms. On the other hand. In sample 1, there are quite a few indels that are called throughout the sample, a few of which can be seen in the figure below.

A number in a row

Description automatically generated

3. The indels in sample one all fall between position 1,245,282 and 1,255,000. If we look at the average coverage between these positions in sample one, we can see it is roughly double that of samples two and three, a sampling of these positions is shown in the figures below.

A number of numbers and arrows

Description automatically generated with medium confidence

Sample 1

A number of numbers and symbols

Description automatically generated with medium confidence

Sample 2

A number of numbers and symbols

Description automatically generated with medium confidence

Sample 3

It is highly unlikely that it is a sequencing error because the duplicated section of the genome spans thousands of consecutive bases, whilst illumina only sequences reads of a few hundred bases at a time. This means that a gene duplication event must have occurred, which explains the high number of indels in sample 1, compared to the complete lack of them in samples 2 and 3. This is because, when an organism has two copies of the same section of DNA one of the copies can mutate freely, without detriment to the organism, as the other copy exists almost as a ‘back-up’ to preserve the initial function of the gene.

Reference:

Tan, A., Abecasis, G.R. and Kang, H.M., 2015. Unified representation of genetic variants. *Bioinformatics* [Online], 31(13). Available from: https://doi.org/10.1093/bioinformatics/btv112.