COMP90016 - Assignment 1

Author: Chris Rode
Student Number: 146637

Task 1 - Heterogenous Variant Calling

Running the bin/snv_heterozygous.py script across the three samples, and then comparing the example using the bin/snv_outfile_stats.py script I get the data listed at the end of this section

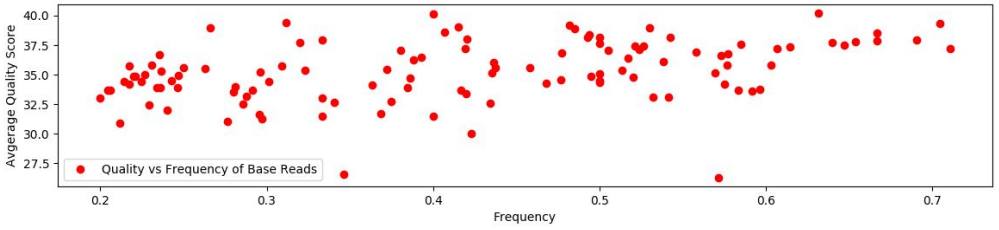
For Sample 1 and 2 adding the filter doesn't affect the results too much. It does cause a small amount of low quality variant calls to be dropped, and increases the overall average of base quality slightly. This would be expected of largely high quality data, with a few small quality outliers

Sample 3 however has a larger portion of reads that are lower quality, so the addition of the filter affects both the number of called variances and base quality more than the other two samples

As an interesting aside; the graphs for Sample 2 and 3 show marked clustering along the frequency lines. This suggests either that the data was generated, or that there was bias in the analysis

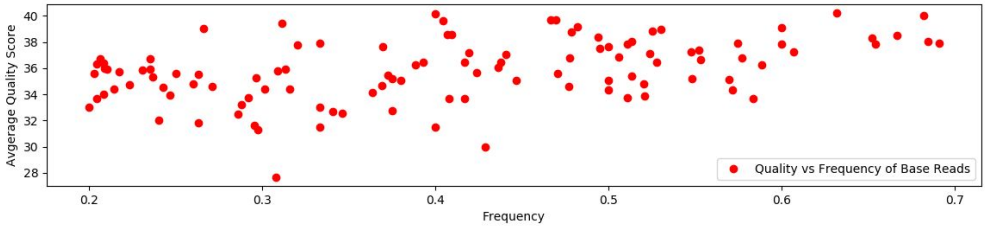
Sample1 - No Filter

Filename: sample1_hetero.txt
Heterozygote Positions = 57
Avg variant base quality = 35.2893206395



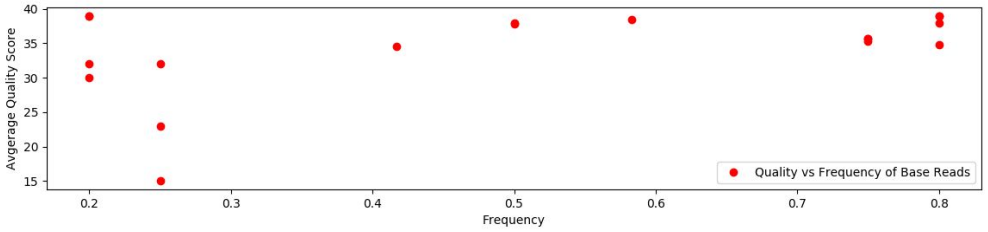
Sample1 - Filter

Filename: sample1_hetero.txt
Heterozygote Positions = 55
Avg variant base quality = 35.8215392815



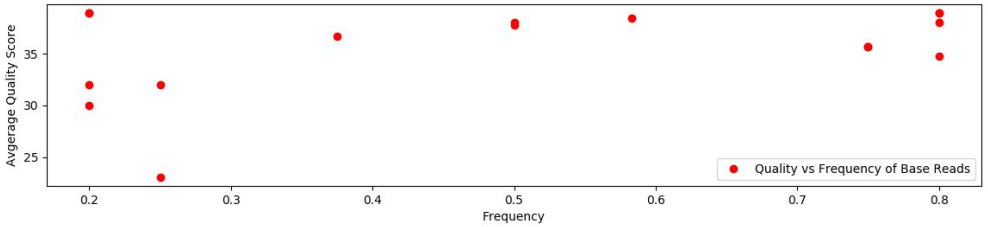
Sample2 - No Filter

Filename: sample2_hetero.txt
Heterozygote Positions = 9
Avg variant base quality = 34.2314814815



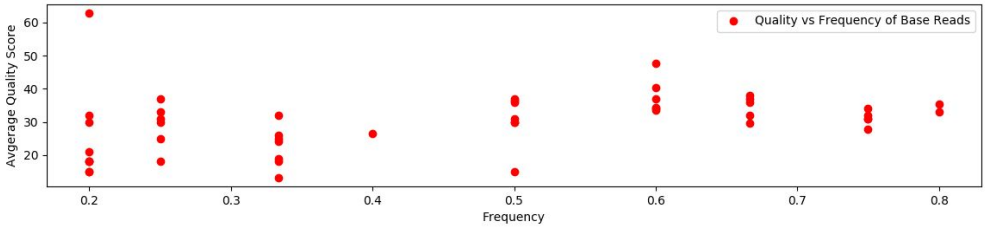
Sample2 - Filter

Filename: sample2_hetero.txt
Heterozygote Positions = 8
Avg variant base quality = 35.5



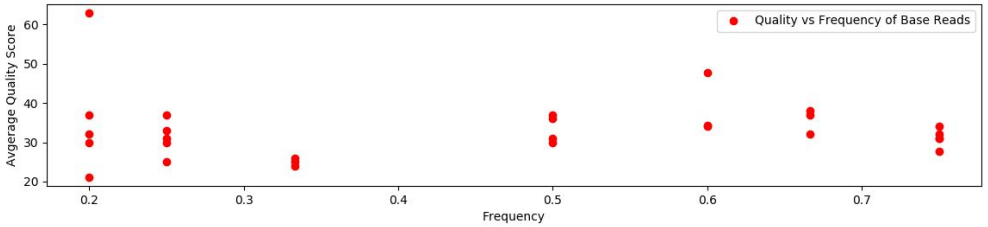
Sample3 - No Filter

Filename: sample2_hetero.txt
Heterozygote Positions = 24
Avg variant base quality = 29.7222222222



Sample3 - Filter

Filename: sample2_hetero.txt
Heterozygote Positions = 14
Avg variant base quality = 33.0952380952



Task 2 - Variant Calling using the Reference (VCF Format)

See the script in bin/snv_reference.py

NOTE: For this task I opted to keep the Quality Filtering from Task 1 as well as adding the Frequency cutoff of 0.2 as specified by this Task 2

NOTE: The output positions in the VCF file are zero-based

Task 3 - Phasing

See the script in bin/phaser.py

NOTE: The output positions for this script are zero-based

Phasing Detection

My approach for the Phasing Detection was to utilise the VCF file from Task 2 to find the positions at which there were heterozygous base-pairs, this is seen in the SAMPLE column with a genotype consisting of more than one base. eg: 0/1 as opposed to 1/1. The genotype data is then used to find all of the called bases that were seen at each position in the raw data. This allows later stages in the phasing detection to ignore any reads that were considered errors

I grouped the positions into groups of two that exist next to eachother. Eg: positions 1,3,5,6,7 would have been grouped into 1|3,3|5,5|6, & 6|7. All reads that covered both positions for each group were cycled through, finding all of the haplotypes for the position pairs in the data. Any reads that were deemed to be errors were able to be filtered at this stage utilizing the known_bases from the VCF file above.

Each phased haplotype pair was then given a frequency score based on the first base. Since these are diploid heterozygous positions, it is known that they have two unique bases. These could lead to up to two phased haplotypes for any position-pair. Utilising the first base in the pair is used as a proxy for knowing which chromosome copy the read came from, allowing for both haplotypes to be called

This lead to the lists as can be seen in the output, with any two-base haplotype having a consensus score over 90% considered to be Phased and anything below that being rejected and marked as Not Phased

As noted above, each position pair may have up to two phased-haplotypes called as can be seen for the position pair 28364366,28364367 in the sample1 output

Since the reads in the BAM file were paired reads, this information could also have been utilised for determining phasing as it's known that the paired reads are on the same strand. For example, if a position is on one read of a pair and the next position is on another, then the bases seen at these positions could be used to infer phasing. This was not done in my code

NOTE: The output generated only shows the haplotypes in the samples and does not include the reference haplotype

Sample1

Total Two-Base Haplotypes Considered: 188
Two Base Phased Haplotypes Accepted: 38
Two Base Phased Haplotypes Rejected: 150

Sample2

Total Haplotype Considered: 13
Two Base Phased Haplotypes Accepted: 7
Two Base Phased Haplotypes Rejected: 6

Sample2

Total Haplotype Considered: 11
Two Base Phased Haplotypes Accepted: 9
Two Base Phased Haplotypes Rejected: 2

Haplotype Extension

My approach for the Haplotype Extension was to take the consensus haplotypes as the starting point and rely on their Associative Nature to merge neighbouring ones

Merging occurs based on the following criteria

1. The preceding haplotype and the next must overlap by one base
2. The overlapping base must be the same in the preceding and next
3. If the preceding position has multiple haplotypes that end in the same base, then merging will not occur. This is because it's impossible to know which strand the merging should occur on

Also, I chose to allow the haplotypes for each strand to grow independent of each other in this merging process which is why the output seems quite uneven. For example Sample1 has a merged haplotype occuring at positions 28364366,28364367, & 28364391 CCA whilst the other strand did not have enough consensus to extend the corresponding to three positions leaving it at two 28364366,28364367 TT

It's also worth noting that the consensus scores given for the two-base haplotypes are a probability of association. By extending by another base, the probability of the three base extension would be the two consensus scores multiplied together. This could have been taken into account to introduce a cutoff for extension length, however this was not done here

Sample1

Haplotypes after merging: 29
Max Haplotype Size: 3
3-Base Haplotypes: 9

Sample2

Haplotypes after merging: 5
Max Haplotype Size: 4
4-Base Haplotypes: 1
3-Base Haplotypes: 0

Sample3

Haplotypes after merging: 8
Max Haplotype Size: 3
3-Base Haplotypes: 1

Further Analysis Utilising the Phasing Information

The phasing data gives extra information as to the exact nature of the alleles that exist in each of the samples. Often a protein will need multiple alterations in order to lose its efficacy and so in these cases it's not enough to know if a person carries all of the variances, it needs to be determined if the person carries the variances on the same strand

Task 4 - Eye Colour Analysis

Based on the vcf file output of Task 2, the genotypes at the two loci rs12913832 and rs1129038 for each of the samples are as follows

	rs1129038	rs12913832		Likelihood Brown (From Table 3 in Paper)	Likelihood Blue (From Table 3 in Paper)
Sample1	A/G	C/T		92.5% [136/(136+11)]	7.5% [11/(136+11)]
Sample2	A/A	C/C		1.2% [2/(170+2)]	98.8% [170/(170+2)]
Sample3	G/G	T/T		100% [36/36]	0% [0/36]

Added to the Data from Task 2 is Data from the paper by Sturm et al. titled “A Single SNP in an Evolutionary Conserved Region within Intron 86 of the HERC2 Gene Determines Human Blue-Brown Eye Color”. Table 3 in the paper shows Phased Genotypes at four loci along with counts of the the corresponding phenotypes (Blue and Brown eyes) from their experimental data. Summing the rows corresponding to the genotypes from Task 2 gives the likelihood of each Phenotype as displayed in the Table

From this data we can see that Sample1 and Sample2 are likely to have Brown eyes whilst Sample3 is likely to exhibit Blue eyes

Unfortunately, this table does not show other colours of eye that might have been displayed with the same Phased Genotypes. In their paper Sturm et al report that whilst the overall classifcation of their data-set was “blue/gray, green/hazel, or brown”, only data from the Blue and Brown subsets was used in Table 3. Addition of data from the Green/Hazel group into the table may cause one of our samples to be reclassified into the Green/Hazel group

Thankfully, Figure 2 gives us the ability to broadly cross-check the likelihoods above as Green is introduced as a Phenotype. The rs1129038 locus however was not included in this analysis, however aking data from this figure leads to the following likelihoods

	rs12913832		Likelihood Blue (From Fig 2 in Paper)	Likelihood Green (From Fig 2 in Paper)	Likelihood Brown (From Fig 2 in Paper)

Sample1	C/T		7% [74/(74+399+590)]	37.5% [399/(74+399+590)]	55.5% [399/(74+399+590)]
Sample2	C/C		72.2% [1394/(1394+524+14)]	27.1% [524/(1394+524+14)]	0.7% [14/(1394+524+14)]
Sample3	T/T		1.1% [2/(147+24+2)]	13.9% [24/(147+24+2)]	85% [24/(147+24+2)]

Broadly speaking, this re-analysis using fig 2 data concurs with the original with Sample1 and Sample3 having a high likelihood of exhibiting Brown eyes and Sample2 having a high likelihood of Blue. Additionally, both Sample1 and Sample2 can be seen to have a reasonable likelihood of green eyes based just on the one locus.

Given the data used in these two tables, it's unclear as to what effect the rs1129038 locus may have on the Green Phenotype, however Sturm et al mention that they found the two loci "were confirmed to be in almost perfect LD" (Sturm et al. 2008). A quick check of our samples' Task 2 vcf data bears this out with Samples 2 and 3 being Homozygous in both locations and Sample1 being Heterozygous in both. Unfortunately the loci were too far away from each other for my Task 3 phasing data to validate the claim, however the addition of paired-end read information may be able to substantiate this further.

If the rs1129038 and rs12913832 do show linkage disequilibrium however, it would be unlikely that the addition of the rs1129038 loci data to Sturm et als fig2 would alter the probabilities significantly

Other things that may have an effect on the likelihoods include the distribution of data included in Sturm et als paper. They report that gender bias was tested for with roughly 50% of each gender represented (52% Female, 48% Male) and that there was "there were no significant gender differences in eye color distribution." (Sturm et al, 2008). There was however a significant skew in the amount of participants of each eye colour included in the report with "46.1% blue/gray, 27.7% green/hazel, and 26.3% brown in the total sample collection" (Sturm et al, 2008). Considering the relatively low percentage of Brown vs Blue participants this may have skewed the likelihoods in the study towards the Blue. Given the significance of the Blue-Brown split on the rs1129038-rs12913832 loci however, this bias does not seem to be material

Given the above, I would say it's likely that the Sample2 has Blue eyes and Sample3 has Brown eyes. Sample1 is unlikely to have Blue eyes, however they may exhibit either Green or Brown eyes with the higher likelihood being Brown

References

Sturm, R. A., Duffy, D. L., Zhao, Z. Z., Leite, F. P. N., Stark, M. S., Hayward, N. K., … Montgomery, G. W. (2008). A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *American Journal of Human Genetics*, *82*(2), 424–431. https://doi.org/10.1016/j.ajhg.2007.11.005