# Assignment 2
# Sample Solution

The programming sample solution for this assignment comes from Christopher, and can be accessed via Github:

https://github.com/cirode/course-COMP90016-assignment2/

It showcases some great programming practice including OO and python modules.

**Task1:**

See the LMS for a sample code.

Discussion of the results requires looking at the different metrics and discussing the possible reasons.

The following table shows Ashley's approach to the task and a fairly comprehensive overview of the change when applying a quality filter:

| | Script | Lines | Average Quality | Qual Standard Deviation | Av difference freq/variant | Av difference qual/variant |
|---|---|---|---|---|---|---|
| Sample 1 | week5.py | 114 | 35.289 | 2.533 | -0.011 | 0.485 |
| | task1.py | 110 | 35.821 | 2.357 | | |
| Sample 2 | week5.py | 18 | 34.232 | 6.144 | -0.003 | 0.135 |
| | task1.py | 16 | 35.500 | 4.272 | | |
| Sample 3 | week5.py | 48 | 29.722 | 9.148 | -0.007 | 0.375 |
| | task1.py | 28 | 33.095 | 7.776 | | |

We can observe that the number of calls gets reduced for every single sample. This makes sense, since a quality filter will prevent sequencing errors (with low qualities) presenting as SNVs.

The quality values for all three samples increase on average and their range of quality values is less variable, due to the restricted space.

Sample 3 has a significant reduction in calls. Many students have argued here that clearly sample 3 has worse sequencing quality to any of the other samples, and the average scores are indeed lower. However, this is not the driving factor here (the average was well above 20 before filtering already). Instead, observe that the average read depth for any heterozygote in sample 1 is **71.6**, sample 2 **12.5**, and sample 3 **3.9**. With such low read depth, a single erroneous base (of quality >=20) can produce a SNV call.

**TASK2:**

See code.

**TASK3:**

See code.

The approach to the haplotyping can be done in several ways.

The strategy chosen in the code is to first read through the VCF file from task 2 and only retain the het calls, which we are interested in phasing.

The program proceeds to go through pairs of het calls, which have reads overlapping both positions (retrieved from the BAM file with the fetch operation). It counts support for haplotypes by counting any of the combinations of the expected haplotypes (ie 00, 01, 10, 11). The score of a haplotype is established by its support divided by the total number of reads with *the haplotype's first base at the first variant*. If this ratio is above 90% *that haplotype* is considered phased. If two haplotypes phase like this, the two variants are considered phased. This is according to our understanding of the program, but it makes for one complication: the haplotypes 00 and 10 could be phased, even though it is contradictory to the fact that variant two has now become a homozygote (00).

Relatively few variants are phased this way for our samples:

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

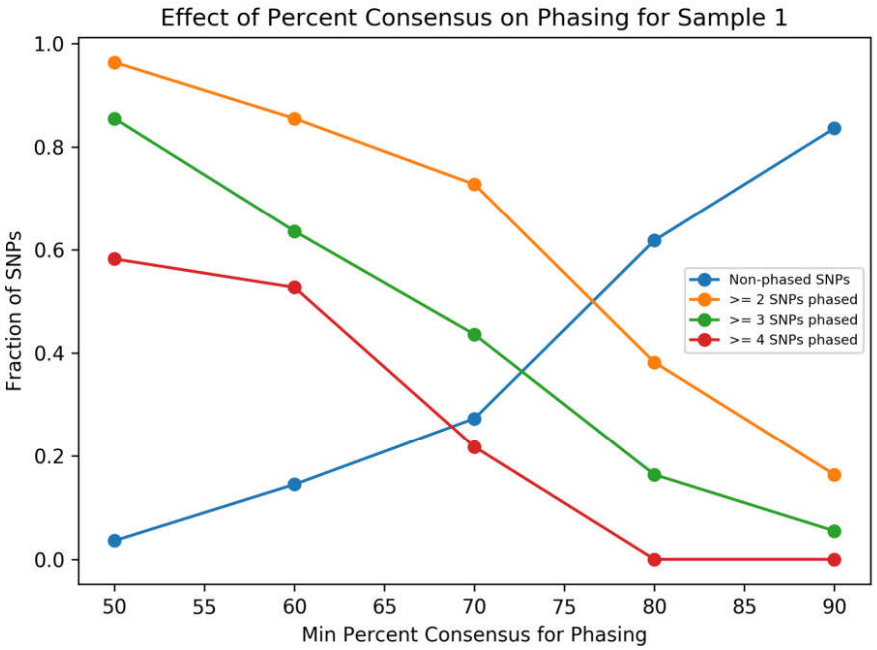
Sample2

Total Haplotype Considered: 11

Two Base Phased Haplotypes Accepted: 9 Two Base Phased Haplotypes Rejected: 2

Some variants can't be phased because no reads overlap them, but others due to lack of consensus.

Ashley went a bit further in this task and investigated the sensibility of the 90% threshold. Observe the plot below:



We can see that at the 90% threshold a very small fraction of neighbouring SNVs can be phased due to lack of consensus. In her data (different to the results above), over 80% of the variant

pairs are rejected due to lack of haplotype consensus. However, this balance changes if this threshold were relaxed to something smaller (here shown to as little as 50% consensus).
We can clearly establish that a trade-off between high-quality phasing and needlessly strict rejection of haplotypes has to be made by choosing a consensus threshold.

The haplotype phasing can be very useful for further analysing the data:

1. We can establish linkage of variants in our population. Causative variants can be narrowed down for specific phenotypes by investigating the linkage structure. Phasing could be compared between the different samples for consistency. An alternative phasing strategy might be envisioned as well, which takes knowledge of likely phasing into account, to relax the consensus support in other samples.
2. Rejected haplotypes and contradictory phasing can be used to identify wrong genotype calls. A lack of haplotype support can make us re-investigate the possibility that one of the alleles is actually observed due to error.

**Task 4:**
The genotypes for the variants of interest is as follows:

| **Sample** | **Rs12913832** | **Rs1129038** | **Diplotype** |
|---|---|---|---|
| Sample 1 | A/G | C/T | AC/GT |
| Sample 2 | A/A | C/C | AC/AC |
| Sample 3 | G/G | T/T | GT/GT |

Note that sample 2 has no entries in the VCF file, so we can deduce that it's homozygous reference.
The phasing from task 3 was not able to link the two variants onto a common haplotype (for sample 1), so we can only assume that linkage is identical to that established in the paper: the AC/GT diplotype.
Counting all occurrences of these diplotypes in the table gives us the following frequencies:

- Sample 1: 136 observations of brown eyes with this haplotype combination and 11 with blue eyes => 92.5% chance for brown eyes (as determined by rows 2-4).
- Sample 2: 170/172=98.8% chance of blue eyes (rows 1 and 7).
- Sample 3: 36/36=100% brown eyes (rows 5, 6, 10, and 11).

The probabilities indicate a fairly confident determination of eye colour of at least 92%.
However, there are some factors that complicate this assertion:

- The table only shows data for blue and brown eyed phenotypes. This does not allow to argue whether the samples are, for example, green eyed.
- The experiment in the paper has a small and biased population sample: Mainly Caucasian individuals. We cannot be certain whether the same observations can be made for other ethnicities, nor what group(s) our samples are from.
- Our own analysis relies on getting the genotypes wrong. We have seen above that this can be tricky, particularly for sample 3 with the low coverage. Sequencing errors might play a role in introducing wrong genotypes.