Expression quantitative trait loci mapping answer sheet

Question 1:

We can represent the individuals based on the number of A allele they carry: $[2\ 1\ 1\ 0]$ or based on the number of T: $[0\ 1\ 1\ 2]$.

Question 2:

- 1000
- 500
- *The idea behind this question is to get students to understand that the number is arbitrary.*
    - The answer depends on the seed:
        - The seed given in the practical generates a genotype of 0.
        - If we count the number of reference alleles, then the genotype is: GG
        - If we count the number of alternate alleles, then the genotype is: CC

Question3:

- Depends on the seed.
- If an allele is not represented enough in the analysis, then this variant cannot be tested.
    - This can be used as a jumping point to discuss how many genotypes need to be within the data and why (Additive effect is assumed by the linear regression, so only 0,1 is needed to perform the regression).

- The minor allele frequency represents how often we found the allele in a population.
    - Let's define p as the minor allele and q as the alternate allele with a frequency of 1- p.
    - If p = 0.7, then q=0.3 and q will be the minor allele.
    - It is, therefore, important to understand that the coding of the allele (0,1,2) is based on frequency and arbitrary.

Question 4:

	Population genotype frequency:	*1,000 individuals*	*10,000 individuals*	*100,000 individuals*
AA *Frequency:* p^2	$0.99^2 = 0.9801$	980.1	9801	98010
AT *Frequency:* $2pq$	$2 * 0.99 * 0.01 = 0.0198$	19.8	198	1980
TT: *Frequency:* q^2	$0.01^2 = 0.0001$	0.1	1	10

- The coding of the allele is 0,1,2 so the effect of the assumed beta is multiplied by either 0,1 or 2. In other words, the effect is additive, so inferring the effect between 0 and 1 is the same as between 1 and 2, allowing us to perform QTL mapping with only 2 genotype group present.
    - NOTE: Mapping is more accurate when the three genotype groups are present due to sampling variance.

- We can perform QTL mapping using either the 10,000 or 100,000 population with the second one giving a more accurate mapping due to an increase in sample size and inclusion of the third genotype group.

<u>Question 5:</u>

- 1000 genes were simulated.
- 50 QTLs were simulated
- 3 different cell type was created.
- The bulk data was created by adding the expression of the three cell types.
    - An extension to this question would be to modify the code to add different representation of cell type.
- The bulk data represent bulk-RNA-sequencing while the cell type corresponds to scRNA-seq
    - scRNA-seq technically create different cells, but QTL mapping is performed on the average expression of the cell, here we only simulated the average expression of the cells, due to computational limitation.

Question 6:

- Mean will be around 10 due to simulation.
- Students might get associated SNPs or not, depending on randomness.
- Identification should be done by testing association.

Question 6:

- The mean expression will depend on the seed.
- Log count, which is the logarithmic value of count data.

  What is the unit of gene expression?

- This question asks them to change the code to Gene5 and visually inspect the results. Any SNPs with a slope can be considered as "associated" leading to the question about how we actually assess if an association is statistically significant → Linear regression, pvalue testing whether the coefficient is significantly different from zero.

Question 7:

- SNP name will be dependent on the seed.
- Change the GeneID variable to "Gene982" and rerun the code.

Question 8:

- No significant cell type effect should be observed as gene 1 is generated with the same distribution for all cell types.

- This basically mean that the cell type previously observed is either shared between all cell type (if we observe a slope in all cell type) or only observed due to sampling variance between different cell.

Question 9:

- Cell type 1 display a positive QTL, cell type 2 a non significant, cell type 3 a negative QTL.
- This should not be a significant, we observe an average expression of the three cell type.

Question 10:

- age, sex and ancestry, look at the GTEx sample page and talk about the different effects.

Question 11:

- Investigate the effect of genetic on lipid levels, identifying parts of the genome that are strongly associated.

Question 12

- LDLR
- -.221 (mg/dL) per allele
- 209.7-2*(0.221)= 209.258

Question 13

- SMARCA4 and SLC44A2
- Esophagus – Mucosa and Muscle – Skeletal
- No eQTL between this gene and LDLR - could be horizontal pleiotropy.

Question 14:

- Muscle – skeletal/ Esophagus muscosa / skin
- Chromosome 5, (75336328-75362104), within a 3'UTR element.
- The SNP is associated with an eQTL, change in gene expression could therefore drive the observed effect on cholesterol level observed.
    - However, we only see the effect in muscle, is that enough?