Question 1

Part A

Create the weighted sum (Madsen-Browning) scores using all variants. Test the score for association with the quantitative trait, qt, adjusting for age, sex, and the population indicator variables.

A multiple linear regression analysis was used to test whether the weighted sum scores of all variants was linearly associated with the quantitative trait, adjusting for sex, age, and population indicators. The t-score was 1.490 and resulting p-value was 0.1365. With a p-value greater than the α=0.05 significance level, the null hypothesis of there being no linear association between the quantitative trait and weights was not rejected. There is insufficient evidence suggesting that the quantitative trait and weighted sum scores are associated.

Part B

Create the weighted sum (Madsen-Browning) scores using only variants with a minor allele frequency (MAF) of less than 1%. Test the score for association with the quantitative trait, qt, adjusting for age, sex, and the population indicator variables.

A multiple linear regression analysis was used to test whether the weighted sum scores of variants with a minor allele frequency of less than 1% was linearly associated with the quantitative trait, adjusting for sex, age, and population indicators. The t-score was 1.443 and resulting p-value was 0.1490. With a p-value greater than the α=0.05 significance level, the null hypothesis of there being no linear association between the quantitative trait and weights was not rejected. There is insufficient evidence suggesting that the quantitative trait and weighted sum scores of rare variants are associated.

Part C

Compare the mean and SD of the Madsen-Browning scores computed in the previous questions. Are they similar? Why or why not?

	Mean	**Standard Deviation**
All Variants	0.1043	0.2123
Rare Variants	0.4428	0.1950

The mean of the Madsen-Browning scores was much higher for the dataset with only rare variants compared to the weights calculated using all variants, regardless of rarity. That is because the formula for calculating the weights is based on the inverse of the minor allele frequency. The rarer a variant is, the lower the allele frequency, and the larger the weight that is given to account for rarer variants having a larger phenotypic effect. Thus, the dataset with only rare variants will have much higher weights. The standard deviation decreased slightly because the common variants with much lower weights were removed, lowering the variability of the weights.

Question 2

Part A

Create CMC scores using only variants with a minor allele frequency (MAF) of less than 1%. Test the score for association with the quantitative trait, qt, adjusting for age, sex, and the population indicator variables.

A multiple linear regression analysis was used to test whether the CMC scores were linearly associated with the quantitative trait, adjusting for sex, age, and population indicators. The t-score was 1.689 and resulting p-value was 0.0915. With a p-value greater than the $\alpha=0.05$ significance level, the null hypothesis of there being no linear association between the quantitative trait and CMC scores was not rejected. There is insufficient evidence suggesting that the quantitative trait and CMC scores of the variants with a minor allele frequency of less than 1% are associated.

Part B

Create CMC scores using only nonsynonymous variants. Test the score for association with the quantitative trait, qt, adjusting for age, sex, and the population indicator variables.

A multiple linear regression analysis was used to test whether the CMC scores were linearly associated with the quantitative trait, adjusting for sex, age, and population indicators. The t-score was 1.513 and resulting p-value was 0.1307. With a p-value greater than the $\alpha=0.05$ significance level, the null hypothesis of there being no linear association between the quantitative trait and CMC scores was not rejected. There is insufficient evidence suggesting that the quantitative trait and CMC scores of the nonsynonymous variants are associated.

Part C

Create CMC scores using only nonsynonymous variants with a minor allele frequency (MAF) of less than 1%. Test the score for association with the quantitative trait, qt, adjusting for age, sex, and the population indicator variables.

A multiple linear regression analysis was used to test whether the CMC scores were linearly associated with the quantitative trait, adjusting for sex, age, and population indicators. The t-score was 2.673 and resulting p-value was 0.00765. With a p-value less than the $\alpha=0.05$ significance level, the null hypothesis of there being no linear association between the quantitative trait and CMC scores was rejected. There is evidence suggesting that the quantitative trait and CMC scores of the nonsynonymous variants with a minor allele frequency of less than 1% are associated.

Part D

Compare the effect size and p-values from the three CMC scores in the previous questions. Are they similar? Why or why not?

	Effect Size	**P-value**
Rare Variants	2.50014	0.0915
Nonsynonymous Variants	1.09550	0.1307
Rare & Nonsynonymous Variants	4.68727	0.00765

The effect sizes differed greatly between all 3 models. SNPs 1-15 were present in all three datasets. The only difference was the rare variants model included SNPs 16, 17, and 18 and the nonsynonymous model included SNP 19. The largest effect size was seen in the dataset only including rare and nonsynonymous variants, which indicates that SNPs 16-19 had either a much smaller magnitude of effect or had the opposite direction of effect than SNPs 1-15. The lowest effect size was seen in the nonsynonymous variants, so SNP 19 must be very different than SNPs 1-15 to bring the average down that much.

The p-value for the rare and nonsynonymous model was much smaller than the p-values of the other two models. That's probably because the CMC scores for SNPs 16-19 had a very different magnitude or opposite direction of effect, so removing them allowed the SNPs that were more similar to each other show an association with the CMC scores.

Question 3

Using the CMC score calculated from only nonsynonymous variants with a minor allele frequency (MAF) of less than 1%, test the score for association with the quantitative trait, qt, adjusting for age and sex.

Part A

How do your results compare to Question 2c where the model did not adjust for the population indicator variables? Compare the effect estimate and p-value for the CMC score.

Covariates	Effect Size	P-value
Age, Sex, Population	4.68727	0.00765
Age, Sex	4.76919	0.00695

The effect size and p-values were very similar in each model. The model only adjusting for age and sex had a slightly larger effect size and slightly more significant p-value.

Part B

Is there any evidence that population membership confounds the association between variants and the quantitative trait, qt?

Since the effect sizes are within 10% of each other, the population indicator variables were not confounding the data.

Question 4

Part A

Is there an association between variants in the ABCA1 gene and the quantitative trait? If so, is it more likely to be due to the rare variants, or the nonsynonymous variants?

From all the analysis ran, there is no significant association between variants in the ABCA1 gene and the quantitative trait. However, when only considering rare **AND** nonsynonymous variants, a linear association appears.

Part B

Suggest additional models or other information about the variants that would be helpful in understanding the association between this gene and the quantitative trait, qt.

Variable threshold models, where analyses are run across many minor allele frequency cutoffs, will be helpful in understanding the association between the ABCA1 gene and the quantitative trait.