Lab 5 Rubric:

Section	Excellent	Good	Acceptable	Needs Improvement	Not Done
1: Data	Correctly identifies the following: 1401 individuals with 861473 SNPs. The genoBim file has a row for each SNP. The columns are the names of the SNP, the chromosome it is on, the name of the SNP again, how large the change is (in this case 0 bases as they are all single nucleotide changes), the position on the chromosome, and the minor allele followed by the major allele. The first 5 are all on chromosome 1. The alleles are: rs10458597: unknown rs12565286: G	Correctly identifies the following: 1401 individuals with 861473 SNPs. The genoBim file has a row for each SNP. Does not correctly identify what the columns are. The first 5 are all on chromosome 1. Correctly identifies at least 4 of the following SNPs. The alleles are: rs10458597: unknown rs12565286: G rs12082473: T rs3094315: C rs2286139:C. For each individual, we have if they have coronary artery disease (1 if yes 0 if no) their sex, age, and levels of	Correctly identifies the following: 1401 individuals with 861473 SNPs. The genoBim file has a row for each SNP. Does not correctly identify what the columns are. The first 5 are all on chromosome 1. Correctly identifies at least 2 of the following SNPs. The alleles are: rs10458597: unknown rs12565286: G rs12082473: T rs3094315: C rs2286139:C. Identifies that for each individual we have sex, age, tg, ldl, and hdl. Does not identify what the CAD column is doing and that tg	Correctly identifies the following: 1401 individuals with 861473 SNPs. The genoBim file has a row for each SNP. Does not correctly identify what the columns are. The first 5 are all on chromosome 1. Does not correctly identify the first 5 minor alleles. Identifies that for each individual we have sex, age, tg, ldl, and hdl. Does not identify what the CAD column is doing and that tg stands for triglycerides. There are people (like the first individual), for whom we only have CAD, sex and	Does not answer at least two of the main questions posed.

	rs12082473: T rs3094315: C rs2286139:C. For each individual, we have if they have coronary artery disease (1 if yes 0 if no) their sex, age, and levels of HDL, LDL, and TG (triglycerides). There are people (like the first individual), for whom we only have CAD, sex and age and no HDL, LDL or triglyceride levels. There are 468 clinical controls (no CAD). There are 92 individuals for whom we have no HDL data.	HDL, LDL, and TG (triglycerides). There are people (like the first individual), for whom we only have CAD, sex and age and no HDL, LDL or triglyceride levels. Correctly identifies at least one of these: There are 468 clinical controls (no CAD). There are 92 individuals for whom we have no HDL data.	stands for triglycerides. There are people (like the first individual), for whom we only have CAD, sex and age and no HDL, LDL or triglyceride levels. Correctly identifies at least one of these: There are 468 clinical controls (no CAD). There are 92 individuals for whom we have no HDL data.	age and no HDL, LDL or triglyceride levels. Does not identify how many controls and how many without hdl information.	
2: Filtering	Identified that we do not filter out any individuals in the first filtering step. Indicates that as we are using	Identified that we do not filter out any individuals in the first filtering step. Explanation for why isn't clear	Identified that we do not filter out any individuals in the first filtering step. Explanation for why isn't clear	Indicates the numbers removed at each step without explanation or discussion.	Doesn't indicate the number of individuals or SNPs filtered at each step.

	previously used clinical data, the data was previously screened. Indicates that we screen out 203287 SNPs due to low MAF or call rate. This leaves 658186 SNPs in out data. In the inbreeding screen we do not lose any data, again because this data was already screened for these kinds of errors before being published. In step 4, we are looking at the population substructure. In this step again, we remove no individuals. Should include plot and describe that there is still some substructure to the population, despite	or complete. Indicates that we screen out 203287 SNPs due to low MAF or call rate. This leaves 658186 SNPs in out data. In the inbreeding screen we do not lose any data. Does NOT give an explanation for why. In step 4, we are looking at the population substructure. In this step again, we remove no individuals. Includes plot but doesn't indicate that there is still some substructure to the population, despite the individuals all being European. In step 5 another 1296 SNPs are removed due to being out of HWE.	or complete. Indicates that we screen out 203287 SNPs due to low MAF or call rate. Doesn't indicate how many SNPs remain. In the inbreeding screen we do not lose any data. Does NOT give an explanation for why. In step 4, we are looking at the population substructure. In this step again, we remove no individuals. Does NOT include the plot. Doesn't indicate that there is still some substructure to the population, despite the individuals all being European. In step 5 another 1296 SNPs are removed due to		

	the individuals all being European. In step 5 another 1296 SNPs are removed due to being out of HWE in the controls, indicating something going on at those locations. We end with 1401 individuals and 656890 SNPs.	No explanation is given as to why. We end with 1401 individuals and 656890 SNPs.	being out of HWE. No explanation is given as to why.		
3: GWAS calculations and Results	Indicates that we successfully calculated 10 PCAs for the data during feature selection. Indicates that SNPs on chromosome 16 were imputed based on the known linkage patterns determined by the 1000 genome project. We use the data for the CEU populations to do this. Indicate that this was done to	Indicates that we successfully calculated 10 PCAs for the data during feature selection. Indicates that SNPs on chromosome 16 were imputed based on the known linkage patterns determined by the 1000 genome project. Does not describe the data used for this. Indicate that this was done to	Doesn't indicate that we successfully calculated 10 PCAs for the data during feature selection. Indicates that SNPs on chromosome 16 were imputed based on the known linkage patterns determined by the 1000 genome project. Does not describe the data used for this. Does not indicate	Doesn't indicate that we successfully calculated 10 PCAs for the data during feature selection. Indicates that SNPs on chromosome 16 were imputed. Does not describe the how or the data used for this. Does not indicate that this was done to increase the number of SNPs that we can try to correlate to the clinical	Does not answer at least 3 of the main questions posed.

	increase the number of SNPs that we can try to correlate to the clinical characteristics of interest. Indicates that 162565 SNPs were imputed on chromosome 16. Describes that the data for p-value calculations were restricted to chromosomes 15-17. This allowed for a quicker analysis ~<10 minutes. Indicates how long it took to determine the p-values. Indicates that with Bonferroni correction, the p-value we are looking for is 7.6E-8. That is the p-value of 0.05 that us usually used for significance	increase the number of SNPs that we can try to correlate to the clinical characteristics of interest. Indicates that 162565 SNPs were imputed on chromosome 16. Describes that the data for p-value calculations were restricted to chromosomes 15-17. This allowed for a quicker analysis ~<10 minutes. Indicates how long it took to determine the p-values. Does not discuss the Bonferroni correction, but references it. Indicates that None of the typed SNPs have a p-value lower than the Bonferroni	that this was done to increase the number of SNPs that we can try to correlate to the clinical characteristics of interest. Indicates that 162565 SNPs were imputed on chromosome 16. Describes that the data for p-value calculations were restricted to chromosomes 15-17. Does not indicate why. Indicates how long it took to determine the p-values. Does not discuss the Bonferroni correction, but references it. Indicates that None of the typed SNPs have a p-value lower than the Bonferroni correction. The	characteristics of interest. Indicates that 162565 SNPs were imputed on chromosome 16. Describes that the data for p-value calculations were restricted to chromosomes 15-17. Does not indicate why. Indicates how long it took to determine the p-values. Does not discuss the Bonferroni correction, or reference it. Indicates that there are SNPs that are statistically significant as they have p-values lower than 0.05. Indicates that there are 77 SNPs that we identified in the CETP gene, 7 of them are typed and 70 are	

	divided by the number of hypotheses tested, one for each of the 656890 SNPs we used. None of the typed SNPs have a p-value that low. The closest is rs1532625 with a p-value of 8.45E-8. The imputed P-values, which are imputed based on the p-vales of the typed SNPs that correlate, have the lowest p-value of 9.81E-8, which is also not below the Bonferroni correction standard. Indicates that there are 77 SNPs that we identified in the CETP gene, 7 of them are typed and 70 are imputed.	correction. The closest is rs1532625 with a p-value of 8.45E-8. The imputed P-values, which are imputed based on the p-vales of the typed SNPs that correlate, have the lowest p-value of 9.81E-8, which is also not below the Bonferroni correction standard. Indicates that there are 77 SNPs that we identified in the CETP gene, 7 of them are typed and 70 are imputed.	closest is rs1532625 with a p-value of 8.45E-8. The imputed P-values, which are imputed based on the p-vales of the typed SNPs that correlate, have the lowest p-value of 9.81E-8, which is also not below the Bonferroni correction standard. Indicates that there are 77 SNPs that we identified in the CETP gene, 7 of them are typed and 70 are imputed.	imputed.	

4: Analysis	They include images of the Manhattan Plot. As well as the linkage plot and information from LocusZoom. Indicates that there are no strictly significant SNPs, but there are three that are very close, rs153265, rs1532624, and rs7205804. These are all near 9E-8. These are pointed out on the Manhattan and linkage plots. Indicates that these are in or near the gene CETP (Cholesteryl ester transfer protein). Discusses that this gene is involved in HDL, LDL, and triglyceride metabolism. Given the high degree of linkage	They include images of the Manhattan Plot. As well as the linkage plot and information from LocusZoom. Indicates that there are no strictly significant SNPs, but there are three that are very close, rs153265, rs1532624, and rs7205804. These are all near 9E-8. They do not indicate where these SNPs are on the graphs. Indicates that these are in or near the gene CETP (Cholesteryl ester transfer protein). Discusses that this gene is involved in HDL, LDL, and triglyceride metabolism. Given the high degree of linkage	They include images of the Manhattan Plot. As well as the linkage plot, but not from LocusZoom. Indicates that there are no strictly significant SNPs, but there are three that are very close, rs153265, rs1532624, and rs7205804. These are all near 9E-8. They do not indicate where these SNPs are on the graphs. Indicates that these are in or near the gene CETP (Cholesteryl ester transfer protein). Identifies the name of this gene, but no discussion of its function. No stand is taken about the reasonableness of the imputation.	They do NOT include images of the Manhattan Plot and the linkage plot. Indicates that there are no strictly significant SNPs, but there are three that are very close, rs153265, rs1532624, and rs7205804. Indicates that these are in or near the gene CETP (Cholesteryl ester transfer protein). Do no identify the name of this gene, or discussion of its function. No stand is taken about the reasonableness of the imputation. Next steps are not indicated or are not justified by the data collected in this experiment.	Does not answer at least 2 of these questions.

	in the area of these SNPs, as seen in the linkage diagram and the recombination rate chart at LocusZoom, these imputations are likely trustworthy. (Any good justification of a stand is fine here, but a stand must be taken). Next steps might include exploring the enzymatic role of CETP and/or developing drugs to decrease CAD based on the function of this gene.	in the area of these SNPs, as seen in the linkage diagram and the recombination rate chart at LocusZoom, these imputations are likely trustworthy. (Any good justification of a stand is fine here, but a stand must be taken). Next steps might include exploring the enzymatic role of CETP and/or developing drugs to decrease CAD based on the function of this gene.	Next steps are not indicated or are not justified by the data collected in this experiment.		
Overall	Written clearly in paragraph form. Writing has a flow. Scientific vocabulary is correctly used.		Written in paragraph form. Writing is stilted. Scientific vocabulary is used incorrectly.		Written in bullet form. Writing is stilted. Scientific vocabulary is used incorrectly.
Citations	All tools and sources used are properly cited.		All sources and some tools used are cited.		No citations