Lab 5 Rubric:

Section	Excellent	Good Acceptable		Needs	Not Done
			Improvement		
1:	Correctly identifies	Correctly identifies	Correctly identifies	Correctly identifies	Does not answer at
Data	the following:	the following:	the following:	the following:	least two of the
	1401 individuals	1401 individuals	1401 individuals	1401 individuals	main questions
	with 861473 SNPs.	with 861473 SNPs.	with 861473 SNPs.	with 861473 SNPs.	posed.
	The genoBim file	The genoBim file	The genoBim file	The genoBim file	
	has a row for each	has a row for each	has a row for each	has a row for each	
	SNP. The columns	SNP. Does not	SNP. Does not	SNP. Does not	
	are the names of	correctly identify	correctly identify	correctly identify	
	the SNP, the	what the columns	what the columns	what the columns	
	chromosome it is	are.	are.	are.	
	on, the name of the	The first 5 are all	The first 5 are all	The first 5 are all	
	SNP again, how	on chromosome 1.	on chromosome 1.	on chromosome 1.	
	large the change is	Correctly identifies	Correctly identifies	Does not correctly	
	(in this case 0	at least 4 of the	at least 2 of the	identify the first 5	
	bases as they are	following SNPs.	following SNPs.	minor alleles.	
	all single	The alleles are:	The alleles are:	Identifies that for	
	nucleotide	rs10458597:	rs10458597:	each individual we	
	changes), the	unknown	unknown	have sex, age, tg,	
	position on the	rs12565286: G	rs12565286: G	ldl, and hdl. Does	
	chromosome, and	rs12082473: T	rs12082473: T	not identify what	
	the minor allele rs3094315: C		rs3094315: C	the CAD column is	
	followed by the rs228613		rs2286139:C.	doing and that tg	
	major allele. For		Identifies that for	stands for	
	The first 5 are all	individual, we	each individual we	triglycerides.	
	on chromosome 1.	have if they have	have sex, age, tg,	There are people	
	The alleles are:	coronary artery	ldl, and hdl. Does	(like the first	
	rs10458597:	disease (1 if yes 0	not identify what	individual), for	
	unknown	if no) their sex,	the CAD column is	whom we only	
	rs12565286: G	age, and levels of	doing and that tg	have CAD, sex and	

2:	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. Identified that we	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.	Doesn't indicate
Filtering	do not filter out any individuals in the first filtering step. Indicates that as we are using	do not filter out any individuals in the first filtering step. Explanation for why isn't clear	do not filter out any individuals in the first filtering step. Explanation for why isn't clear	numbers removed at each step without explanation or discussion.	the number of individuals or SNPs filtered at each step.

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

	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 pvalues. Indicates that with Bonferroni correction, the pvalue we are looking for is 7.6E-8. That is the pvalue 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 pvalues. Does not discuss the Bonferroni correction, but references it. Indicates that None of the typed SNPs have a pvalue 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 pvalues. Does not discuss the Bonferroni correction, but references it. Indicates that None of the typed SNPs have a pvalue 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 pvalues. 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	They include	They include	They do NOT	Does not answer at
	images of the	images of the	images of the	include images of	least 2 of these
	Manhattan Plot. As	Manhattan Plot. As	Manhattan Plot. As	the Manhattan Plot	questions.
	well as the linkage	well as the linkage	well as the linkage	and the linkage	
	plot and	plot and	plot, but not from	plot.	
	information from	information from	LocusZoom.	Indicates that	
	LocusZoom.	LocusZoom.	Indicates that	there are no	
	Indicates that	Indicates that	there are no	strictly significant	
	there are no	there are no	strictly significant	SNPs, but there are	
	strictly significant	strictly significant	SNPs, but there are	three that are very	
	SNPs, but there are	SNPs, but there are	three that are very	close, rs153265,	
	three that are very	three that are very	close, rs153265,	rs1532624, and	
	close, rs153265,	close, rs153265,	rs1532624, and	rs7205804.	
	rs1532624, and	rs1532624, and	rs7205804. These	Indicates that	
	rs7205804. These	rs7205804. These	are all near 9E-8.	these are in or	
	are all near 9E-8.	are all near 9E-8.	They do not	near the gene	
	These are pointed	They do not indicate where CETP (Cholesteryl			
	out on the			ester transfer	
	Manhattan and	these SNPs are on	the graphs.	protein).	
	linkage plots.	the graphs.	Indicates that	Do no identify the	
	Indicates that	Indicates that	these are in or	name of this gene,	
	these are in or	these are in or	near the gene	or discussion of its	
	near the gene	near the gene	CETP (Cholesteryl	function.	
	CETP (Cholesteryl	CETP (Cholesteryl	ester transfer	No stand is taken	
	ester transfer	ester transfer	protein).	about the	
	protein). Discusses	protein). Discusses	Identifies the name	reasonableness of	
	that this gene is	that this gene is	of this gene, but no	the imputation.	
	involved in HDL,	involved in HDL,	discussion of its	Next steps are not	
	LDL, and	LDL, and	function.	indicated or are	
	triglyceride	triglyceride	No stand is taken	not justified by the	
	metabolism.	metabolism.	about the	data collected in	
	Given the high	Given the high	reasonableness of	this experiment.	
	degree of linkage	degree of linkage	the imputation.		

	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	in the area of SNPs, as seen the linkage diagram and recombination chart at LocusZoom, to imputations a likely trustwo (Any good justification of stand is fine but a stand must be taken). Next steps must include exploit the enzymation of CETP and/developing distortion of the function of the same and the same area.	the on rate chese are orthy. of a here, hust ight oring c role for rugs AD	Next steps are not indicated or are not justified by the data collected in this experiment.			
Overall	gene. gene. Written clearly in paragraph form. Writing has a flow. Scientific vocabulary is correctly used.		Writin	etten in paragraph form. Eting is stilted. Scientific abulary is used incorrectly.		Written in bullet form. Writing is stilted. Scientific vocabulary is used incorrectly.	
Citations			All so	urces and some tools ted.	used	No citations	