Thesis Report

Paul

1. *Does the twin method for heritability account for the proportion of twins, siblings and unrelated in a population, and if not then should it? I.e. In some countries there are typically large families, so many more close relatives – everything else being equal, should heritability estimates be higher, lower, or the same in such countries?*
   1. Except for very small population, the family size of the population should not affect the heritability estimates. However, for very small population, the large family size might reduce the genetic variation, thus reduces the heritability estimate, yet this is unlikely for the size of human population. This has been addressed on Pg10 under section 1.2.5 Twin Studies: However, if assortative mating occurs in the population, the additive genetic factors shared between the DZ twins might be higher than 1/2, thus leading to an underestimation of the heritability.
2. *Pg.9/10 In adoption studies why can prenatal influences such as alcohol abuse and malnutrition during pregnancy confound results – in what way do they \*confound\* results? Could you explain exactly what confound means?*
   1. This has been addressed on Pg9 under section 1.2.5 Twin Studies: Moreover, adoption studies cannot control the prenatal environment such as alcohol abuse and malnutrition during pregnancy. For example, a high comorbidity of alcohol use disorders and schizophrenia was observed (Boyd et al., 1984; Drake et al., 1990). High consumption level of alcohol during pregnancy may bring about long-term alterations in nervous system functioning of the offspring (Garrett, 2014). This increased risk of schizophrenia in the offspring might therefore be an artifact of the alcohol abuse of the mother.
3. *Pg.13 You describe the multiple testing problem in GWAS. If we perform an extremely simple study by selecting a SNP at random and then testing its association against a phenotype of interest, then what P-value threshold should use to test it? (we have only performed one test after all).*
   1. This has been addressed on Pg12 under section 1.3.2 Genome Wide association Studies
4. *Pgs 23-25 You describe that there is evidence indicating that CNVs and rare variants may disrupt the same pathways in causing SCZ as common variants – but isn’t this entirely unsurprising? Can you think of any reasons CNVs/rare variants would affect \*different\* pathways from common variants?*
   1. It is possible for the CNVs and rare variants to affect other crucial pathways in schizophrenia with more severe consequences, leading to higher survival burden to the individual. Therefore, these rare variants and CNVs are rare in the population due to active selection against these phenotypes.
5. *Pg.25 Can G\*E interactions really explain some of the ‘missing’ heritability? Wouldn’t they contribute as much to the environment portion of variance explained as the genetic?*
   1. This has now been address in pg25 under section 1.3.4.2 Rare Single Nucleotide Mutation.
6. *Pg26. ‘However, by and large, the prenatal infection is the largest environmental risk factor for SCZ’ – this sounds like a big claim and it’s unreferenced. Figure 1.4 suggests that this is probably not the case? Surely this is presently only a hypothesised risk factor, rather than established as being the biggest environmental cause – there’s a big difference and it is important to be more reasoned and cautious in scientific writing.*
   1. The statement is now corrected to better represent the fact.
7. *Pg.35 – In the summary you bring up a new topic of investigation – testing the effect of n-3 PUFA rich diet in the gene expression pattern in the brains of MIA samples – which is rather unexpected at this point. Given that this will form one of the important hypotheses tested in your thesis then there should be a subsection in the introduction chapter devoted to n-3 PUFA rich diet, providing a background that motivates the hypothesis that it may alter gene expression of MIA samples (inc. summarising the study referred to by Q. Li, Leung et al).*
   1. A small discussion is now included in under section 1.4.1 Prenatal Infection
8. *Pg38. Bulik-Sullivan et al 2015 state that LDSC has a large Cis when trait is non-polygenic… but isn’t it more about the density of causal variants rather than polygenicity per se? (if polygenic but independent SNPs then presumably Cis for LDSC would also be high). So high density of causal variants -> good for LDSC – but can you explain why this is?*
   1. This is now included in the discussion of Chapter 2
9. *Pg39. The maths is incorrect here (e.g. should be h2 = var(x)/var(y))*
   1. The math was corrected accordingly
10. *On pgs 39-51 the maths underlying the SHREK method introduced here are outlined. However, it would be useful to the reader if this was made explicit at the beginning, and if the maths outlined was put in greater context by referring to that which is shared by GCTA and LDSC, and that which is specific to the SHREK method. It would be good to make clear what differentiates SHREK from the other methods and how you might expect that to produce different/similar results For example, is the use of the effective number ( e), bottom of pg.45, unique to SHREK? And does only SHREK exploit direction of effect information, whereas LDSC doesn’t? Moreover, it would be good to provide some more intuitive explanation on the motivation, concept and key features of SHREK (e.g. compared to the other methods), rather than only describe the methods in terms of maths (much like the intuitive explanation that is often presented alongside the maths underlying LDSC).*
    1. The whole section has been rewritten to include the outline and greater context of SHREK.
11. *Pg.45 – Should not simply state ‘after some tedious algebra’- do you mean that you have left several steps out? These need to be documented if so. Either here or else in an Appendix.*
    1. The full derivation are now included without skipping any steps.
12. *Pg.47 Increase which summary statistics? Do you mean the X2 statistic for association? Or heritability estimation? (or both?)*
    1. It is now addressed in pg46 under section 2.2.5 Extreme Phenotype Selection
13. *Pg. 50 To test whether LD matrix has a well-defined rank, you selected 1000 SNPs at random from chromosome 22 HapMap data and simulated using HAPGEN2. However, surely those 1000 SNPs would have typically lower LD than SNPs in the same region and thus won’t properly test whether tSVD will be appropriate for actual SNPs (since a true LD matrix will contain manly highly correlated SNPs). It would be good to state what the subsequent SNP density is in all these scenarios, since LDSC at least is surely highly dependent on the SNP density.*
    1. We now perform the test on the 1,000 genome European samples instead. Thus the simulated SNPs should be representative of the density observed in the real data. Detail of the test can be found in pg49 under section 2.2.6 Inverse of Linkage Disequilibrium Matrix
14. *Pg.53 Why only consider SNPs with MAF > 0.1 here?*
    1. We have now repeated the simulation with MAF >0.05 such that it is more consistent with later simulations
15. *Pg.53 Why only simulate sample size of 1000? Isn’t it limited to consider only a single sample size and one that is much smaller than most large GWAS?*
    1. This is explained in detail under section 2.2.10.1 Sample Size.
16. *Pg.53 Why only repeat steps 5-6 50 times? Surely you want to vary the reference panel too in order to rigorously test the reliability of LD sample correction?*
    1. We now repeated the simulation such that we repeat all steps 100 times such that we can rigorously test the reliability of LD sample correction.
17. *Pg.54 Won’t the sample sizes simulated here produce results that are not reflective of those produced when these methods are applied to large GWAS? It is not so important what the median N is for GWAS, since these methods will typically be applied to the latest large-scale meta-analysis GWAS (which far exceed the median GWAS N\_. I think it would have been worth repeating at least one of the main simulation scenarios with a large N, or else with a larger N to see if any change in result occur.*
    1. This is now discussed in the discussion and included as future tests.
18. *Pg.55 I find it unusual that the number of SNPs chosen (randomly) from each chromosome is what is controlled, rather than the subsequent SNP density, since it is surely the latter that is important. The density could be chosen to match 500k, 1Mb or imputation-level data.*
    1. This is now discussed together with question 8 in the discussion section of chapter 2
19. *Pg. 55 Why are the simulation parameters different here to the study describe on pgs 51-53? (MAF > 0.05 here, 50k SNPs on chromosome 1 etc).* 
    1. This is now explained in both section 2.2.10.2 and section 2.2.9
20. *It would be interesting to know how many significant results are produced from these simulated data/ effect sizes. Would have been good to have shown a corresponding Manhattan plot to get a better idea into the impact of the simulation parameter settings.*
    1. We did not keep track of the files due to storage concerns and at the current stage, we are unable to generate the corresponding Manhattan plots / QQ plots for the simulated data unless we repeat the simulations.
21. *Pg57 It would be good to know what the actual simulated effect sizes are for some of these simulations (ie. Are the m chosen very different to the rest? Do they reflect kind of effect sizes reported in GWAS? Or are they much larger?)*
    1. The simulated effect sizes follow the exponential distribution with lambda equals one. However, similar to the summary files, we did not keep track of all the effect sizes due to storage concern and will have to repeat the whole simulation to obtain this information.
22. *Pg54-58 Should come under a subheading of ‘Quantitative Trait simulations’ so that it matches with the results section.*
    1. This is now corrected as suggested
23. *Pgs 59-60 Surely the extreme pheno sampling should produce essentially the same data/ results as case control? Are the extreme ends treated as 0/1 or still tested as quantitative (if the latter then there may be problems caused by non-normality of residuals in linear regression). Why is the random sampling performed here? Isn’t that just the same as the quantitative trait simulation (and if not sure this sampling reflect real data more).*
    1. We now clearly describe the simulation and indicate that a quantitative test was performed. Random Sampling was performed such that we have a clear comparison between the different sampling strategies
24. *Pgs 62/63 More information should be provided on the issue of the PGC schizophrenia GWAS having too high SNP density for the algorithm without changing the bin size to 50k. What is the density of the PGC data? At what density can SHREK run at without having to change the bin size? What is the bin size under usual SHREK settings? What is the likely inflation of the estimate that results from this change in bin size? If this is uncertain or potentially large then we will have to interpret real results from SHREK with caution.*
    1. We now describe the implementation of SHREK in details where the bin size is discussed. Usually, when the density of the SNPs is higher than 6000 SNPs within a 3 mb region, the decomposition will take indefinite amount of time because of the O(n3) complexity of the algorithm. There is a linear relationship between the bin size and the inflation in the test statistic and to handle this problem, we have updated SHREK with an alternative library such that it can now handle data with higher SNP density.
25. *Pg 64. Why is Fig2.1b so unsmooth? Too few sims? It seems that an advantage of the uncorrected (/bias) estimates are that they have lower variance. No comment was made on this or argument made that it is worth the reduction in bias to have this increase in the variance of estimates (surely not great if just trying to find out if h2 > 0 or not… and if variance is high then average distance from truth could still be further than unbiased estimate).*
    1. We have corrected the simulation as suggested in question 16. As we now resample the reference in each iteration, the variance is no longer comparable to the other simulation we’ve performed (which assumed a fixed reference). Therefore, we no longer include Figure 2.1b in our thesis.
26. Pg66 Looks like the bias in SHREK could be systematically corrected- was this considered?
    1. The systematic bias is discussed in section 2.4.1 LD Correction. However, complexity of the problem is beyond the scope of the current thesis and should be consider as an important direction of further research.
27. *Pg66 Weren’t the causal variants removed in any of these simulations? (to make the data more like real data where we assume many of the causal variants aren’t present but produce signal due to LD)*
    1. The main reason why the causal variants remains within the simulation is simply because of the simplicity for the simulation. We now include it in the discussion as something to consider in the follow ups.
28. *Pg67 Nice to see that the variance of SHREK estimates are relatively insensitive to number of causal variants (ie. Causal SNP density) but I don’t think it will be clear to many readers exactly why this is – how would you summarise how SHREK is able to be insensitive to % causal variants in a sentence or two?*
    1. This is now discussed in section 2.4.2.1 Quantitative Trait Simulation
29. Pg67 From these results, SHREK is certainly preferable if 5-10 causal SNPs is realistic and therefore it is crucial to know \*how\* realistic this is – what does it mean in practice? Does it mean only one causal SNP in each 1Mb where there is a signal? If there was some empirical data/evidence to indicate how common it might be that LDSC would have extremely high variance then that would strongly support using SHREK (but if LDSC and GCTA often give similar variance estimates, then perhaps usually sufficiently polygenic?). Good that the variance of GCTA estimates is low but this is only when the individual level data are available – it would be good to know how much larger the sample size of summary stats need to be before the variance of SHREK / LDSC is as low as GCTA. How can users know whether to use the former on large summary stats or the latter on their small raw data otherwise? Variance in LDSC estimates do seem markedly lower than SHREK.
30. *Pg69 A little more interpretation here would have been good (rather than just description of the plots) e.g. the underestimates in variance of SHREK are concerning because they are anti-conservative whereas the overestimates from LDSC are conservative and so less concerning.*
    1. Slightly more interpretation has been made in section 2.3.2.1 Quantitative Trait Simulation
31. *Pg70-73 These results appear to be suggesting that it’s only best to use SHREK when you only have summary stats and there’s only 1 SNP with a very large effect – but is it really about a SNP with very large effect? If you added a large effect SNP to a present GWAS would that suddenly make SHREK best? Isn’t it just an artefact of your simulations, that when there is only SNP of very large effect then the others subsequently have tiny effect, and thus it becomes similar to a genetic architecture with 1 or a few causal variants? (ie it’s not the large effect SNP itself that causes the issue…).*
    1. This is discussed in detail under section 2.4.2.2 Extreme Effect Size
32. *Pg74-78 (case-control). Great that SHREK performs better for case-control data and I would think an excellent ‘selling point’ in terms of the practical use of SHREK over (or as well as) LDSC – on that point has SHREK been submitted for publication? But isn’t it possible that the better performance of SHREK compared to LDSC here is due to the lower density of SNPs simulated for the case-control data? (LDSC performs better in high density / more polygenic scenarios).*
    1. SHREK has not been submitted for publication. An extended discussion regarding the density of SNPs is not included in the discussion section of chapter 2.
33. *Pg80-83 (Extreme Pheno Sampling). Why aren’t the extreme phenotype sampling results almost the same as the case control simulations and why aren’t the random sampling results almost the same as the quantitative trait results (bar some parameter setting differences) – they should give similar results shouldn’t they? Overall they do, but then there are a few notable differences: better performance of LDSC with extreme pheno sampling than case-control, and extremely high errors in estimating of variance by GCTA and LDSC (hugely conservative estimates). Why would the latter occur – looks like an error, or if not then wouldn’t that suggest that the estimates of variance by GCTA and LDSC on the quantitative data be incorrect?*
    1. The phenotypes simulated in the extreme phenotype sampling simulation was quantitative (we did not convert the phenotypes into 0/1), therefore this simulation is different from the case control simulation. Additionally, in the extreme phenotype sampling simulation, the samples were sampled from the extreme whereas the control from the case control simulation was randomly selected from the control population. This might also lead to the difference in performance of the tools in the two situation. Finally, we have not figure out why the estimation of variance by GCTA and LDSC differ so much from the quantitative trait simulation. Additional simulations might be required to understand the reason behind such discrepancy.
34. *Pg84-85 (Real data application). It’s a real concern that the SHREK estimates on real data seem to be inflated by pop structure. Isn’t this a severe problem for its use in practice? Do you have any possible solutions/adjustments for this? Do you have any strong arguments for applying SHREK rather than LDSC when applied to real data, which is likely to contain a reasonable amount of pop structure?*
    1. This is now discussed in the discussion of chapter 2.
35. *Pg87 wouldn’t it be good to at least repeat the quantitative trait simulations to see if SHREK without sampling correction now performs better than LDSC etc?*
    1. Yes, however, with the current computation resources (the server now forbid jobs to run longer than 48 hours. But our simulation might take more than 2 days to complete), it is not possible to repeat the simulation without the sample correction. However, we do include it as something to do in the future.
36. *Pg90-91 I think rather than focusing on how difficult it is to simulate pop structure it would be better to acknowledge that if there is substantial pop structure then SHREK is likely to be outperformed by LDSC-In, but that SHREK may perform best in the ‘special cases’ when there is very little pop structure in the GWAS if when it has already been well-adjusted for (e.g. by Mixed Model methods… Or perhaps suggest how SHREK might be modified to estimate pop structure (or account for) like LDSC.*
    1. This is now incorporated into the discussion of chapter 2
37. *Pg93 So do you think a nested case-control study, within a cohort, wouldn’t have these case-control or extreme phenotype biases?*
    1. If I understand correctly, our current simulation of the case control study follows the nested case-control study design, where a population was simulated and we select all the cases from the population. The controls were then randomly selected from the remaining control population, which is similar to the nested case-control study design. Therefore I would expect the same bias to be observed.
38. *Pg95 Isn’t the point about what M should be in the LDSC formula of critical importance? If the estimate is sensitive to M, and M has large uncertainty, then surely LDSC estimates have very large uncertainty and are often far from the true h2? (0.555 to 0.135 is a huge difference). Seems a major concern in terms of LDSC which could mean that overall SHREK estimates are more reliable (assuming it doesn’t suffer from a similar problem?)*
    1. This is now discussed in detail under section 2.2.4 SNP-heritability of Schizophrenia
39. *Pg97 The fact that SHREK/LDSC have poor (and even negative) estimates when most causal variants are rare is a concern…. Won’t this likely mean that when causal variants are a mix of rare and common, then the estimate will be very biased downwards? Does GCTA suffer to the same extent if there are many rare causal variants? If not then the congruence of estimates from the different methods would at least provide some reassurance.*
    1. We have not tested the performance of SHREK when handling rare causal variants. It is also unclear how GCTA preform when there are many rare causal variants. Further study might therefore be required to understand the effect of the rare variant to the performance of the tools.
40. *There is no specific recommendation in the Conclusion section regarding whether the strategy going forward for SCZ GWAS should be to increase sample sizes and stick with current genotyping strategy or whether the field should invest more in sequencing and epigenetics etc instead?*
    1. This is now addressed in the discussion of chapter 2
41. *Pg114 Which ‘unsupervised clustering’ method/approach was performed here?*
    1. The unsupervised clustering was performed using functions provided by DESeq2.
42. *Pg115 1. Tests for the effect of diet, but why not do that across saline and PolyIC samples – wouldn’t that increase power? Saline/ polyIC n-3 Vs Saline/PolyIC n-6*
    1. The main goal of the study is to study the effect of diet in relation to MIA. Therefore, although comparing across saline and PolyIC samples might increase the power, we did not perform the analysis.
43. Pg116 Can you explain exactly how this Wilcoxon Rank Sum test was performed on these data – more explanation required. Why Rank Sum test? (not multinomial test). Why only consider significant pathways in terms of the rank sum test? (Aren’t highly non-significant pathways interesting too?).
44. *Pg118-120 Given the deflation of the statistic, some of the top results in Fig3.2 look interesting despite not reaching the usual significance level. I think it would be interesting to note what those genes are too, and especially whether there is any overlap between the top genes – and e.g. is the top gene from O6vs O3 diet (3.2c) in fact just one of the top ones from the corresponding saline testing (i.e. 3.2b)… this would suggest that it’s not the impact of the diet in response to MIA in particular, it’s just an additive effect of the difference in diet.*
    1. This is now discussed in the discussion of chapter 3
45. *Pg121-122 Great that 6/7 gene sets are enriched for MIA effect (and 1 for diet) – but wouldn’t it have been good to choose a few control pathways in order to ensure that these kind of effects are not seen for unexpected/neutral pathways too?*
    1. We have performed permutation analysis to ensure these effects are not seen for unexpected pathways too. This is not correctly stated in the methodology of chapter 3.
46. *Should include more ideas / more hypotheses and descriptions of studies that you would do to follow-up on this project (given more time, resources etc). Should give the reader the feeling that you could create further avenues for investigation that would follow up this research in a sensible direction in future, now that you have accrued all this knowledge and expertise on the topic. Any ideas now?*
    1. We now include a new section 3.4.4 Designing the Replication Study to discuss about ideas of future studies.

Dr Pang

1. *In the future studies proposed, what ideas do you have about combination of multiple dimensions?*
   1. For the MIA study, it will be interesting to also consider the effect of sex and strain in the future studies. Therefore we can better understand the effect of MIA and diet.
2. *Have you considered how you would extend this to admixed populations?*
   1. This is now included in the discussion section under chapter 2
3. *Why do extreme phenotype sampling?*
   1. This is discussed in detail under 2.2.10.7 Extreme Phenotype Sampling
4. *What motivated you to carry out this research?*
   1. The motivation are now clearly stated in the introduction.
5. *What were the crucial research decisions you made?*
   1. The batch effect is one of the biggest confounding factor in a RNA Sequencing study. Therefore we have tried our best to balance out the batches before we conduct the study such that we can minimize the effect of the batches.
6. *What have you learned from the process of doing your PhD?*
   1. Matrix algebra, multithreading computer programming, analysis of next generation sequencing data

Dr Song

1. *There are typos, e.g. “oligogenetic” in abstract and other sections*
   1. The typos are corrected
2. *LD is complex. Some SNPs are in LD with many other SNPs; Some SNPs are in LD with no other SNPs; SNPs between a pair of associated SNPs are not necessarily associated with the flanking SNPs. In 2.4.1 LD correction: how did you make LD correction?*
   1. This is now discussed in detail under section 2.4.1 LD correction and section 2.2.9 Comparing Different LD correction Algorithms
3. *Although you study only focused on common variants, how do you think rare variants contribute the risk for schizophrenia?*
   1. This is now included in Chapter 1 Introduction
4. *MHC is the most significant locus in GWAS and your MIA model also implicated the immune system involved in schizophrenia, why did you remove variants in this locus in your simulation and how this will affect your result?*
   1. We did not perform simulation on the MHC variants. However, we did filtered out the MHC region from the real data analysis. The main reason behind such decision is because the MHC region contains unusual LD structure which might have unpredictable effect to the estimates from LDSC and SHREK. Therefore, we have to remove the variants from this region. However, shall we include this region in the estimation, we do expect to observe a relatively higher heritability for all disorders, especially for schizophrenia considering the significance of the MHC region.
5. *Why did you select this gestation Day (GD) 9 in MIA model?*
   1. It is now included in the introduction of chapter 3: GD 9 is when the neural tube close and previous studies suggested that this is the critical period where MIA during this period have a larger effect than MIA in mid-gestation period. Therefore it was selected as the MIA time point in our study.
6. *Why do you think to use different strains of mice in MIA model?*
   1. Different strains have different genetic background, therefore might react differently to the MIA insult. As our lab has only been working on the C57BL6/N mice, we are uncertain whether if the same model works if another strain of mice were used. Therefore, we only focused on the C57BL6/N. Shall there be more resources and time, then it will definitely be interesting to also investigate the effect of MIA using other mouse strains.
7. *Page 117, Section 3.2.8 “the partitioning of heritability was performed using LDSC (B.K. Bulike-Sullivan et al., 2015)”, what is the reason that you didn’t use your own SHREK.*
   1. This is now explained in the introduction of chapter 3: As SHREK has not tested for its performance in the partitioning of heritability, LDSC, which has been tested, is used.
8. *Page 130, rt-PCR -> RT-PCR. As the candidature stated that the expression change of Sgk1 detected by RNA Seq need to be confirmed by RT-PCR and it would be nice to include this result in the revised version.*
   1. This is corrected. However, although the RT-PCR result is crucial for the publication of the data, we currently don’t have sufficient manpower for the experiment, and therefore we do not have the RT-PCR result.