# **Chapter 4 Supplementary Note**

**A comparison of proxy Alzheimer’s disease/dementia phenotype definitions in the UK Biobank**

Table of Contents

[Chapter 4 Supplementary Note 1](#_Toc138599632)

[Supplementary Results 3](#_Toc138599633)

[MAGMA gene and gene-set analysis 3](#_Toc138599634)

[Local genetic correlation 3](#_Toc138599635)

[Calculated effect sizes comparison to clinical AD effect sizes 4](#_Toc138599636)

[Inclusion of genotyped individual case status in case-control phenotypes 4](#_Toc138599637)

[LT-FH and LT-FH++ heritability specification 6](#_Toc138599638)

[Comparison to PGCALZ2 results 6](#_Toc138599639)

[Leave-one-out rare variant analyses 6](#_Toc138599640)

[Supplementary Methods 7](#_Toc138599641)

[Local genetic correlation 7](#_Toc138599642)

[Calculated effect sizes comparison to clinical AD effect sizes 7](#_Toc138599643)

[Supplementary Figures 8](#_Toc138599644)

[References 19](#_Toc138599645)

## Supplementary Results

### MAGMA gene and gene-set analysis

The MAGMA1 gene-based results were similar between the different phenotype methods. Any differences in significantly associated genes between the phenotype methods were due to small differences in P-values around the significance threshold (P<2.65012x10-6) (**Supplementary Table 4.5**). In terms of which loci were identified through significant genes, the small differences in P-values led to the association of the *INPP5D* locus only in the Jansen results and no association with *KAT8* in the Jansen results, a lack of association with the HLA region in the Marioni results, an association of *FIBP* only in the LTFH69 and LTFHplus69 results and no association of *KFP3* in the LTFH69 and LTFHplus69 results, no association of *SLC24A4* in the LTFH69 results and an association of *ACE* only in the LTFH69 results, and no association of *ABCA7* in the CaseControl or Jansen results. However, these differences are likely to do with random variation and unlikely to have a large effect on interpretation. The CaseControl results had the highest inflation factor (λCaseControl=1.041; λMarioni=1.039; λJansen=1.041; λLTFH69=1.037; λLTFHplus69=1.038) but this appears to have not increased the number of significant genes. Thirteen gene-sets were Bonferroni significant in at least one of the phenotypes (0.05/15485= 3.23x10-6) (**Supplementary Table 4.5**). All thirteen of these gene-sets were significant in the LTFHplus69 analysis and only 6 were significant in the Marioni analysis. After forward conditioning, there were 3 independently associated gene-sets with all phenotypes (GO\_bp:go\_negative\_regulation\_of\_amyloid\_precursor\_protein\_catabolic\_process, GO\_cc:go\_neurofibrillary\_tangle, Curated\_gene\_sets:roversi\_glioma\_copy\_number\_up). These three gene-sets were identified as independently significantly associated with all phenotypes which suggest that all of the phenotypes captured the same key gene-set associations.

### Local genetic correlation

The common variant GWAS results from the Jansen phenotype had the lowest global genetic correlation with the results from the other phenotypes (~0.6-0.9) but had a similar genetic correlation with the PGCALZ2noUKB data as the results from the other proxy phenotypes (~0.7-0.9) (**Supplementary Figure 4.1**). We performed local genetic correlation analysis using LAVA2 to identify regions that could be driving the difference between the Jansen phenotype results and the other phenotypes. We proposed two possible mechanisms that could be causing the lower genetic correlation between Jansen and the other proxy phenotype results: first, there could be regions not associated with the Jansen phenotype that are associated with the other phenotypes; second, there could be regions associated with all phenotypes that have genetic correlations less than 1 between the Jansen results and the other phenotype results.

We investigated the first possible mechanism and identified 56 loci where the Jansen phenotype results did not have nominally significant heritability (P<0.05) but all of the other phenotype results did have nominally significant heritability (**Supplementary Figure 4.5; Supplementary Table 4.9**). It is unlikely that these 56 loci are causing the differences in global genetic correlation between the Jansen phenotype and the other phenotypes because many of the most significant heritability estimates at these loci were with the Marioni results and the Marioni results had the highest global genetic correlation with the Jansen results. We also identified 31 loci where the Jansen phenotype results did have nominally significant heritability (P<0.05) but all of the other phenotype results did not (**Supplementary Table 4.9**). It is possible that these loci may explain a fraction of the difference in global genetic correlation between the Jansen phenotype and the other phenotypes. None of these 87 loci were found to be significant in the most recent GWAS meta-analyses of AD3,4 so the differences may not be due to regions related to AD.

Next, we investigated whether there were regions with significant local genetic correlations between the Jansen and other phenotypes that could explain the lower global genetic correlation (**Supplementary Figure 4.5**). We identified 37 loci with significant local genetic correlations between Jansen and the other proxy phenotypes (CaseControl, Marioni, LTFH69, LTFHplus69), all but one of these correlation estimates were between 0.75-0.9. There was one locus (*APOE* region) with a local genetic correlation of 0.99 between the Jansen phenotype and the LTFH69 and LTFHplus69 phenotypes. These regions could partly explain the lower global genetic correlation between the Jansen phenotype and the other phenotypes. However, the local genetic correlations between the Jansen and Marioni phenotypes were similar to the local genetic correlations between the Jansen and other phenotypes, which suggests that these regions alone do not explain the difference in global genetic correlation between the Jansen phenotype and the other phenotypes.

Additionally, we were interested in whether the difference at these 37 loci could be causing a difference in how well the results from the proxy phenotypes correlated with the PGCALZ2noUKB data on the global scale. We found that genetic correlations between the proxy phenotypes and the PGCALZ2noUKB data were not different at these 37 loci (**Supplementary Figure 4.5**). This supports our assumption that the difference between the proxy phenotypes are either in regions not known to be relevant to AD or are not substantial enough to cause the results of one phenotype to be more similar to clinical AD GWAS results.

### Calculated effect sizes comparison to clinical AD effect sizes

We pairwise compared the effect sizes in the CaseControl (after doubling), Marioni, Jansen (calculated), LTFH69 (calculated) and LTFHplus69 (calculated) datasets to the effect sizes in Kunkle *et al.* (2019)5 (stage 1 results only) and PGCALZ2noUKB datasets using variants significant in both datasets (**Supplementary Figure 4.6**). We found that the effect sizes in the CaseControl and Marioni datasets were in between the Kunkle and PGCALZ2 effect sizes. The slope of the best fit line of CaseControl effect sizes against the Kunkle and PGCALZ2noUKB effect sizes suggested that the Kunkle effect sizes were about 1.13 times larger than the CaseControl effect sizes and the PGCALZ2noUKB effect sizes were 0.94 times as large as the CaseControl effect sizes. A similar estimate was obtained when comparing the Marioni effect sizes (βKunkle~1.17xβMarioni and βPGCALZ2noUKB~0.97xβMarioni). However, the calculated effect sizes of significant variants in the Jansen, LTFH69, and LTFHplus69 datasets were on average larger than both the Kunkle and PGCALZ2noUKB effect sizes (βKunkle~0.81xβJansen and βPGCALZ2noUKB~0.67xβJansen; βKunkle~0.81xβLTFH69 and βPGCALZ2noUKB~0.66xβLTFH69; βKunkle~0.81xβLTFHplus69 and βPGCALZ2noUKB~0.66xβLTFHplus69). This suggests that while the calculated effect sizes are on a more similar scale to the effect sizes from the linear GWAS analyses, they appear to be less similar to the effect sizes in the clinical AD GWAS compared to the CaseControl and Marioni GWAS results.

### Inclusion of genotyped individual case status in case-control phenotypes

Marioni *et al*. (2018)6 performed a proxy GWAS by meta-analysing the results from two parental phenotype GWAS. Initially, they created a maternal phenotype by assigning any individual with a mother with AD/dementia as a case and those without as controls. They then performed a GWAS on this phenotype. They repeated this process except using individuals with a father with AD/dementia as cases to create a paternal phenotype GWAS. They meta-analysed the results from these two GWAS using an inverse-variance weighted method without adjusting for overlapping individuals. They suggested that this was similar to a GWAS of a phenotype based on the number of affected parents (0,1,2)6. In our analysis, the relatively low intercept (Intercept=1.0426, SE=0.0101) from LDSC suggests that the meta-analysis of overlapping individuals led to relatively low confounding (**Supplementary Table 4.3**). We were interested in how this process would be affected by including true cases in the parental phenotype definitions. We found that the including true cases in the maternal and paternal phenotype definitions led to increased inflation factors and number of significant loci, without a large difference in LDSC intercept (**Supplementary Table 4.3**). However, when the GWAS from the parental phenotypes including true cases were meta-analysed together, the genomic inflation factor (1.15) and intercept (Intercept=1.11, SE=0.0108) were considerably higher. This suggests that meta-analysing the parental GWAS results together when case status of the genotyped individuals is included in the phenotype definition will lead to increased bias.

We also looked at whether including the case status of the genotyped individual in the CaseControl phenotype would affect the GWAS results. In our sample, 1999 genotyped individuals were AD cases and including them in the CaseControl phenotype increased the number of cases from 49,937 to 51,435, an increase of 1498 cases. There were 501 individuals diagnosed with AD that had at least one parent with AD/dementia. This relatively small change in case control numbers had a small effect on the GWAS results, in terms of number of significant variants, number of significant loci, inflation factor and LDSC intercept (**Supplementary Table 4.10**). It also had a small effect on the scaling of the effect estimates and standard errors, with them being approximately 1.06 times larger when the case status of the genotyped individual was included (based on the slope of the line of best fit of effect sizes of significant variants). Since these effect sizes and standard errors were scaled by a factor of two before the meta-analysis with the clinical AD GWAS data, we also compared the effect sizes and standard errors of the proxy CaseControl data (with and without genotyped individual case status included in the phenotype) to the PGCALZ2noUKB data using significant variants. We found that when the case status of the genotyped individual was ignored, the effect sizes of the PGCALZ2 data were approximately 1.99 times larger than the proxy data, and 1.87 times larger than the proxy data when genotyped individual case status was not ignored (**Supplementary Figure 4.7**). This suggests that the effect sizes are less attenuated when genotyped individual case status is included in the proxy phenotype and that the suggested 2 times increase in effect sizes and standard errors before meta-analysis may be an over-estimate.

We were interested to explore whether this attenuation was linear so we performed a series of GWAS where the number of true cases varied from 0 to 1999. We randomly selected a number of true cases to have a phenotype value of 1 starting from 0 true cases to 1999 (all of the true cases) at intervals of 200 (0,200,400,600,800,1000,1200,1400,1600,1800,1999). We performed a GWAS using variants which were significantly associated with AD in the PGCALZ2noUKB data and then estimated the attenuation factor (slope of the best fit line) from the relationship between the effect sizes in the PGCALZ2noUKB data to the effect sizes in the UKB data. We only included variants which were significant in both datasets. We repeated this for all of the different numbers of true cases and fit a line of best fit for the relationship between the attenuation factors and the proportion of cases being true cases (**Supplementary Figure 4.7**; **Supplementary Table 4.11**). We found the intercept to be approximately 2 (1.989) and the slope to be -2.81. This fit line implies that a dataset of 100% cases would have an attenuation factor of -0.81, where one would expect that the attenuation factor would be 1. This slope either implies that the relationship is not linear or that our estimates of effect size dilution are imperfect. The limitation of this analysis is that we are comparing effect sizes to the PGCALZ2noUKB data which may not represent the true effect size for each variant. Additionally, our simple model fit does not account for the error around the effect sizes and the error around the slope estimates. We also do not account for the correlation structure between the significant variants. We can only observe the attenuation between ~0-5% of cases being true cases without down sampling proxy cases and controls. In samples where there are larger numbers of true cases it may be beneficial to either analyse the true cases separately from the parental cases and then meta-analyse the results together or generate a phenotype with a method that accounts for the true and parental cases (LT-FH and LT-FH++). Future analyses using other methods7 to estimate effect size dilution that account for the correlation between variants and the standard errors of the effect sizes would be valuable in estimating how much including true cases attenuates the dilution of effect sizes.

### LT-FH and LT-FH++ heritability specification

We were interested in how specifying a different heritability estimate in the LT-FH and LT-FH++ phenotypes would impact the GWAS results. In our primary analysis, we assumed the heritability of AD to be 69%, based on twin studies8, when creating the liability threshold phenotypes. As a form of sensitivity analysis, we repeated the GWAS for these phenotypes except we specified the heritability of AD to be 5%, a common heritability estimate when applying LDSC to AD GWAS summary statistics9. We found that with both methods, increasing the heritability from 5% to 69% caused roughly a 12.5x increase in phenotype values for each person; however, the increase was not linear (**Supplementary Figure 4.8**). While the phenotype values change, specifying the lower heritability (5%) does not change the GWAS results considerably **(Supplementary Table 4.10**). Both of the methods show very similar number of significant variants, inflation factors, and intercepts. The only change in GWAS results is the scale of effect sizes and standard errors, which increase by a factor of 12.6 for the LT-FH method and 13.02 for the LT-FH++ method (**Supplementary Figure 4.8**). We transformed these effect sizes and standard errors for the meta-analysis and the transformed effect sizes and standard errors are almost identical between the results from the different heritability specifications (**Supplementary Figure 4.8**). For the purposes explored in our study, it appears that the results are robust to different heritability specifications.

### Comparison to PGCALZ2 results

In this study, we performed a meta-analysis using all of the data included in Wightman *et al.* (2021)3 except the 23andMe data. However, our approach for meta-analysis was a two-step analysis using inverse-variance weighted meta-analysis and their approach was to use a one-step sample-size weighted meta-analysis. Of the 38 significant loci in the PGCALZ2 analysis, 32 were significant, 4 were suggestive (*TSPOAP1/TSPOAP1-AS1,* *TNIP1*, *HAVCR2*, *APP*), and 2 were not even suggestive (*AGRN*, *NCK2*) in the PGC+Jansen meta-analysis results using the inverse-variance weighted method. In the PGC+Jansen data, the lead variant (rs113020870) in the *AGRN* locus from PGCALZ2 had a P-value of 0.0035 and the lead variant (rs115186657) in the *NCK2* locus was not present in the data. The results from the inverse-variance weighted meta-analysis largely replicate the results from the Wightman *et al.* (2021)3 meta-analysis.

### Leave-one-out rare variant analyses

The rare-variant aggregation analyses highlighted 7 genes associated with at least one of the proxy phenotypes and one gene borderline significantly associated (*ATP8B4*) that was associated in a previous rare variant aggregation analysis of AD10. We were interested in which variants contributed most to these associations and whether it was the same variant driving the gene association with all phenotypes. To investigate this, we performed leave-one-out analyses where the significantly associated genes were analysed again but after leaving out one of the variants mapped to this gene. This process was repeated for each of the variants and the gene level P-value was examined to determine the impact of each variant on the gene association. We found that in all cases it was the same variants driving the gene associations in all phenotypes and all genes had a single variant that contributed a large amount to the gene association (**Supplementary Table 4.7**; **Supplementary Figure 4.9**).

## Supplementary Methods

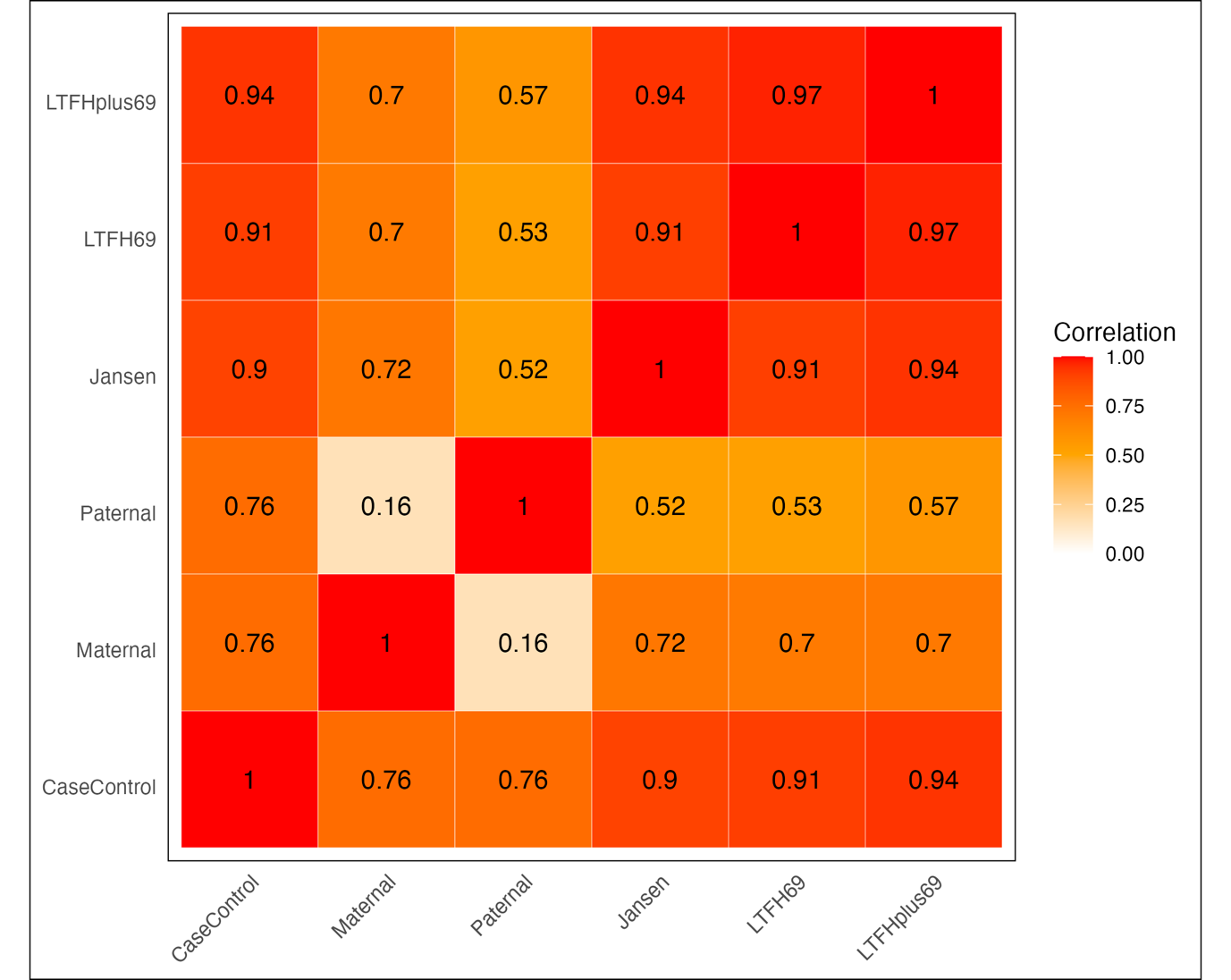
### Local genetic correlation

We performed local genetic correlation analysis using LAVA2 (v0.1.0; GitHub commit 3ad57c85) to investigate regions of the genome that could contribute to the differences between the common variant GWAS results from the different phenotypes. The analyses used a sample overlap estimate (genetic covariance intercept) from the LDSC analyses, all of which were >0.9 between the proxy phenotypes and <0.07 between the proxy phenotypes and PGCALZ2noUKB data. Sample size was defined as N/4 for the proxy phenotypes and the effective sample size was used for the PCGALZ2noUKB data (4/(1/Ncases)+(1/Ncontrols)). The loci boundaries were the same as defined in the original LAVA manuscript2 (<https://github.com/cadeleeuw/lava-partitioning/blob/main/LAVA_s2500_m25_f1_w200.blocks>). To obtain loci with significant local genetic correlations with the Jansen results, we only looked at loci with nominal (P<0.05) heritability in the Jansen results. Then we identified 3437 loci-phenotype pairs that had nominal heritability which could be used to calculate genetic correlations. Genetic correlations were considered significant if the P-values were lower than the Bonferroni correction threshold after correcting for 3437 tests (0.05/3437 =1.45x10-5). The 37 loci with significant local genetic correlations between the Jansen and other phenotypes were then used for local genetic correlation analysis between the proxy data and the PGCALZ2noUKB data.

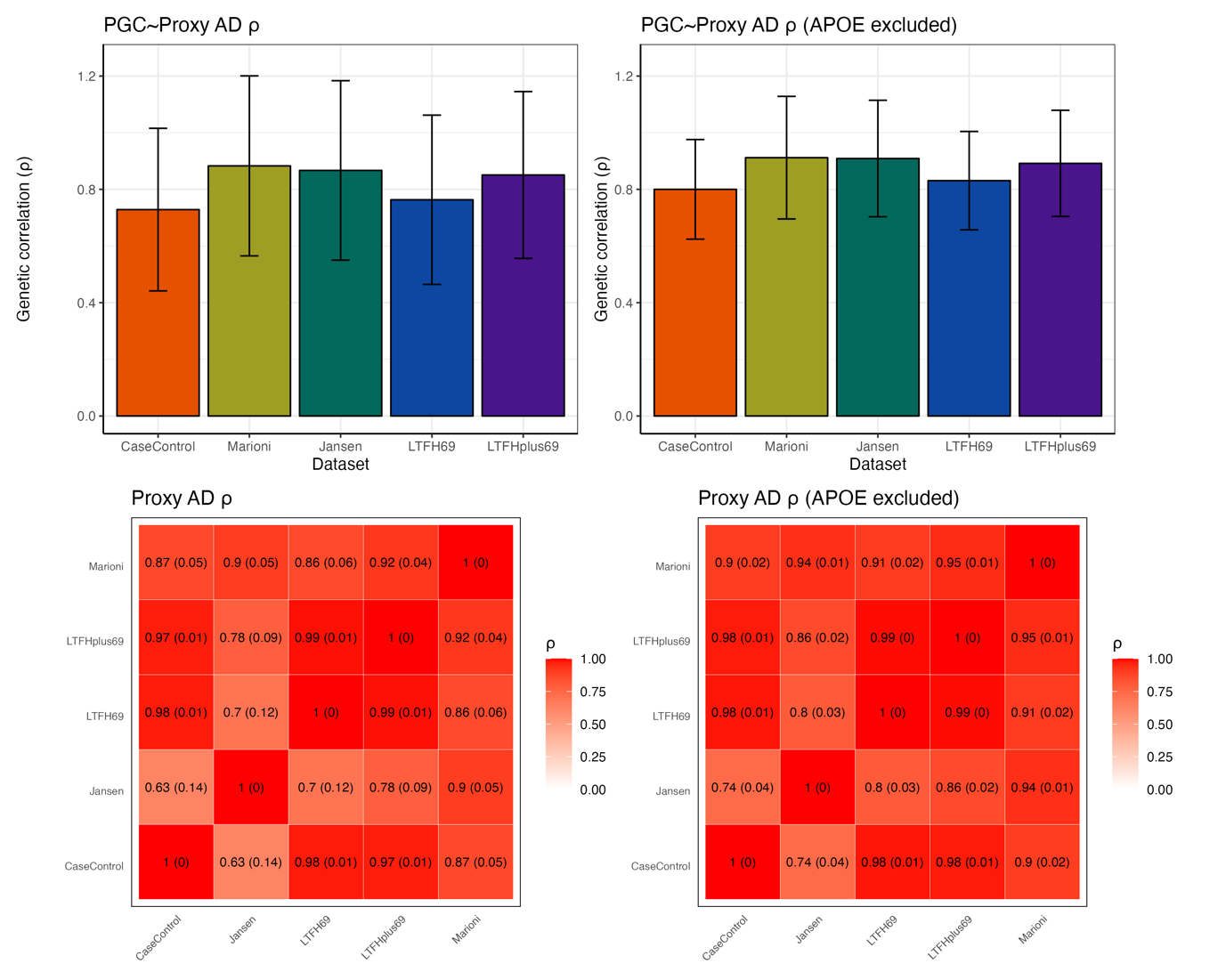
### Calculated effect sizes comparison to clinical AD effect sizes

The effect sizes of the significant variants in each of the proxy GWAS results were compared to the effect sizes of significant variants in the Kunkle *et al.* (2019)5 (Stage 1 results only) and PGCALZ2noUKB data. Only variants that were significant in both of the datasets in the pairwise comparison were kept. A line of best fit was estimated using the lm() function in R v4.1.311 and the slope was used to compare the effect size scaling.

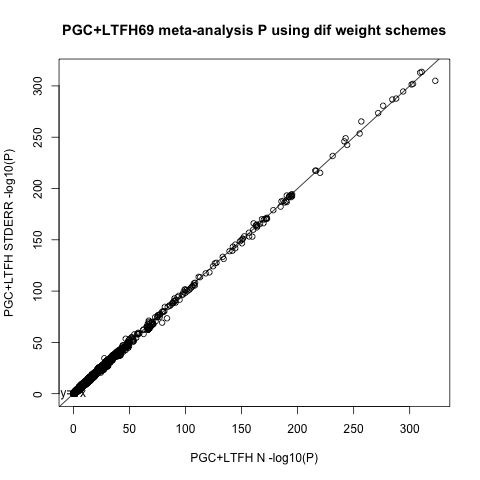
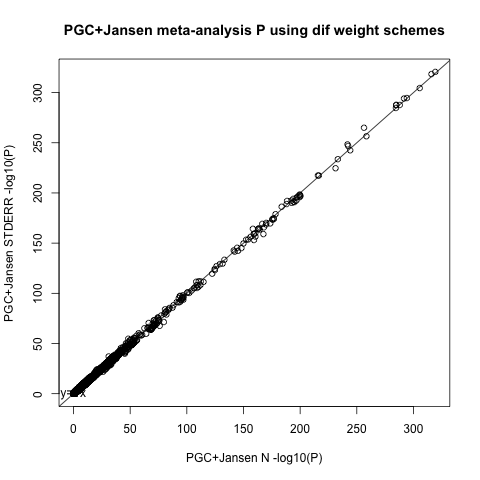
## Supplementary Figures

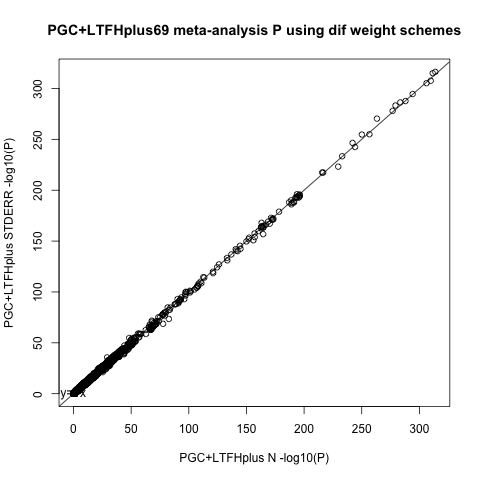


Supplementary Figure 4.1: The phenotype level correlation between the phenotypes used in this study shows that the Maternal and Paternal phenotypes are the least similar and phenotypes which used both maternal and paternal case status in the phenotype definition are very highly correlated. Correlation estimates between binary phenotypes (CaseControl, Maternal, and Paternal) were tetrachoric, correlation estimates between (pseudo) continuous phenotypes (Jansen, LTFH69, LTFHplus69) were Pearson correlations, and correlation estimates between (pseudo) continuous and binary phenotypes were biserial.

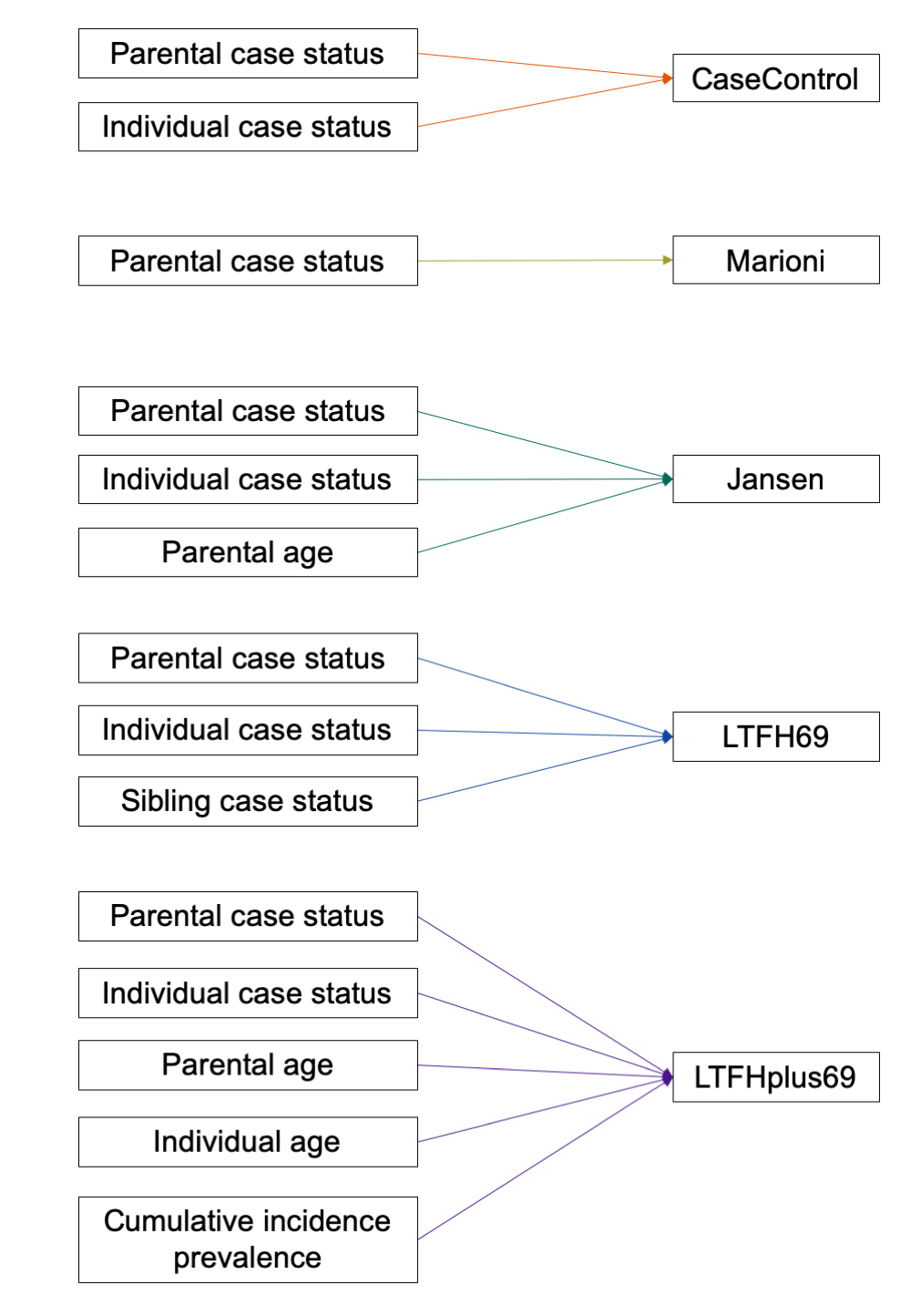


Supplementary Figure 4.2: The genetic correlations between the proxy AD/dementia GWAS results and the PGCALZ2noUKB data and the genetic correlations between the proxy AD/dementia GWAS results highlight the similarity between the proxy phenotypes and the proxy phenotypes and the clinical AD GWAS (PGCALZ2noUKB). Results with and without the larger APOE region (GRCh37: 19:40,000,000-50,000,000) are reported. The error bars represent the 95% confidence intervals.

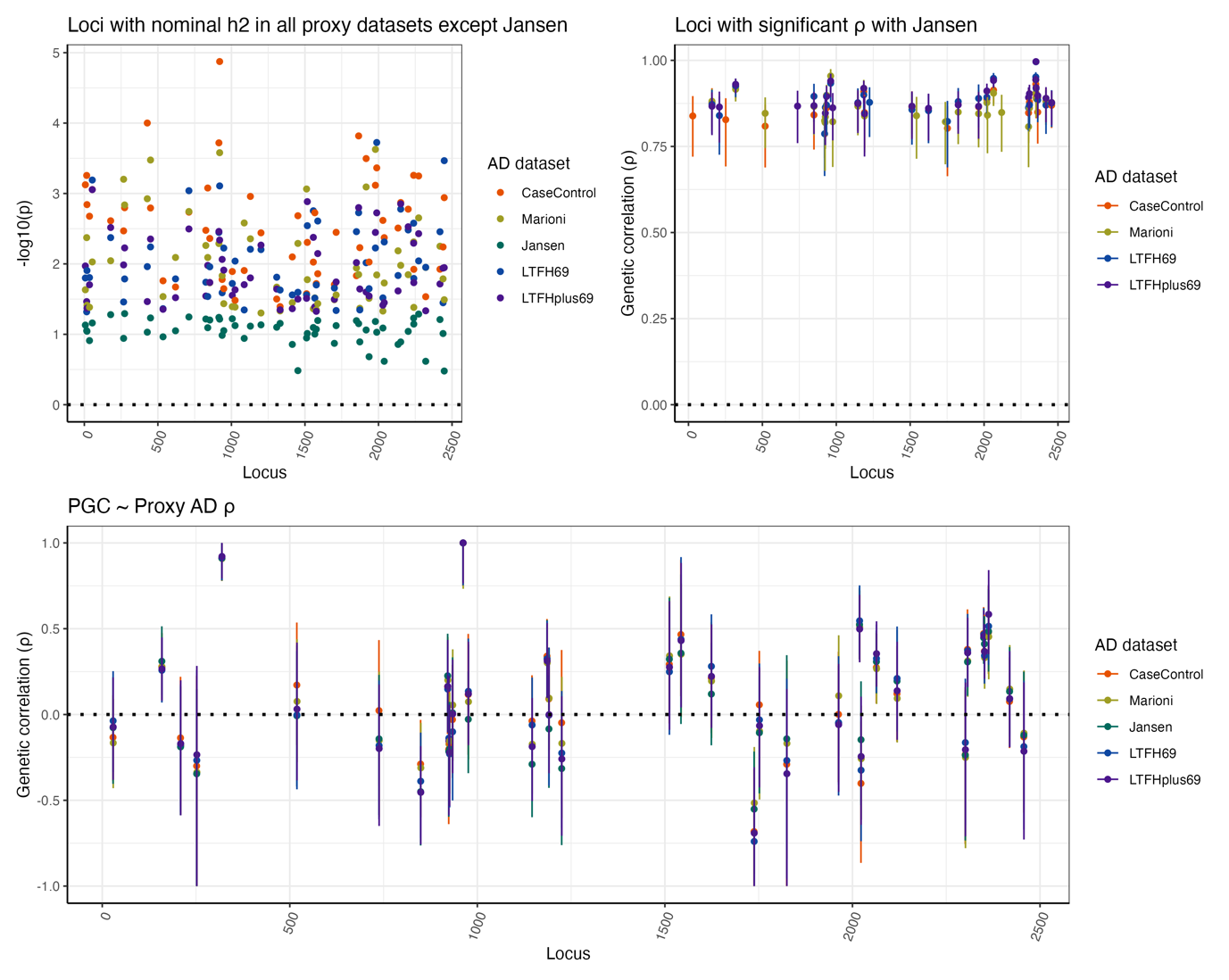




Supplementary Figure 4.3: The -log10 P-values of the inverse-variance and sample size weighted meta-analysis of the Jansen+PGC, LTFH69+PGC, and LTFHplus69+PGC meta-analyses shows almost perfect concordance between the two meta-analysis approaches. The line represents the identity line.



Supplementary Figure 4.4: The information used to create each of the five phenotypes used in this study highlights that the LTFHplus69 phenotype requires the most information.

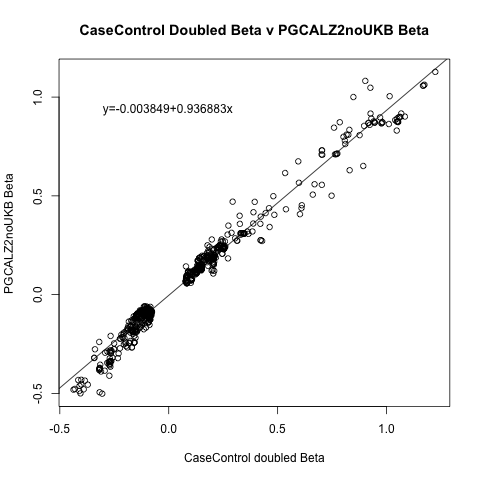
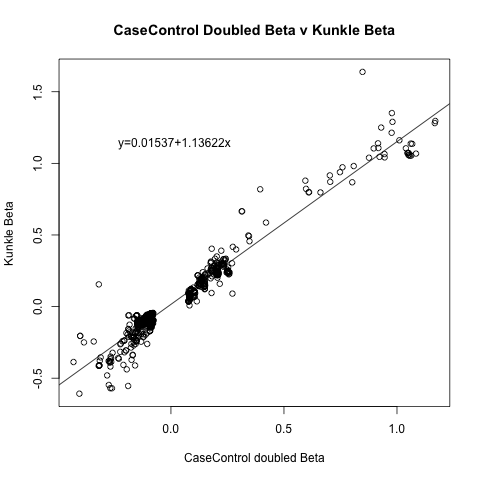
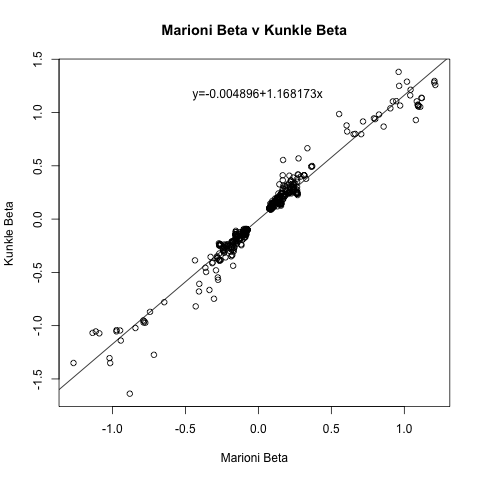
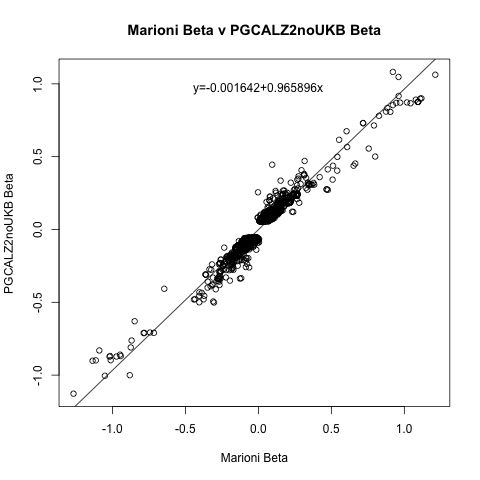


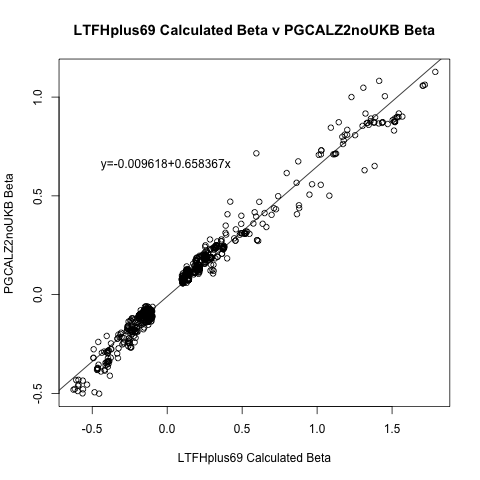
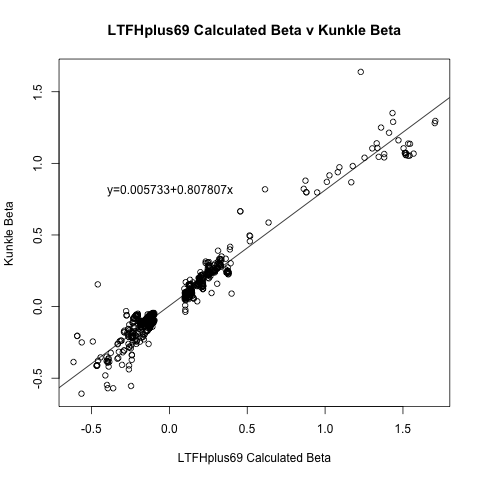
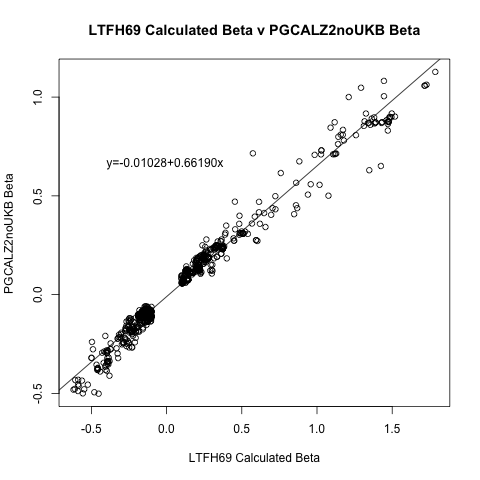
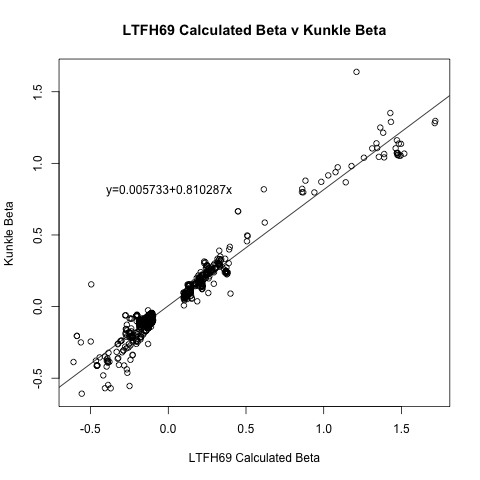
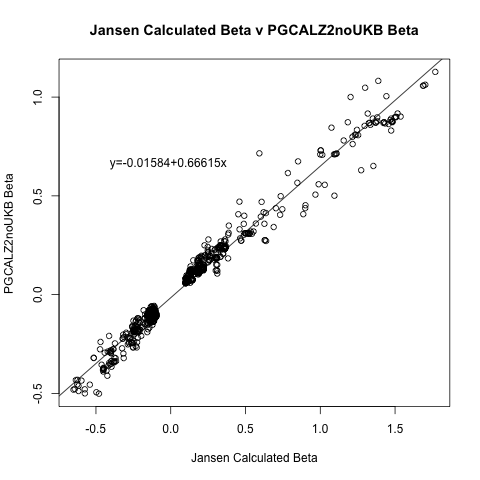
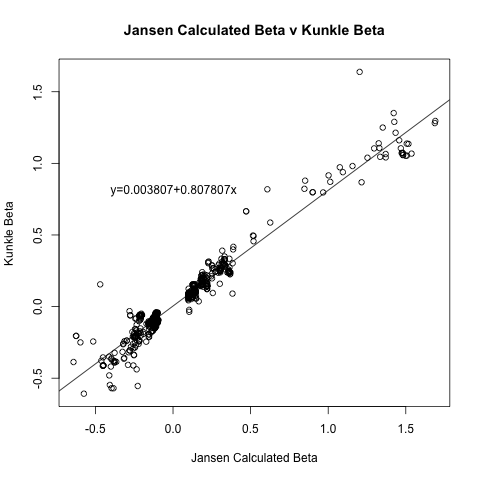
c)

b)

a)

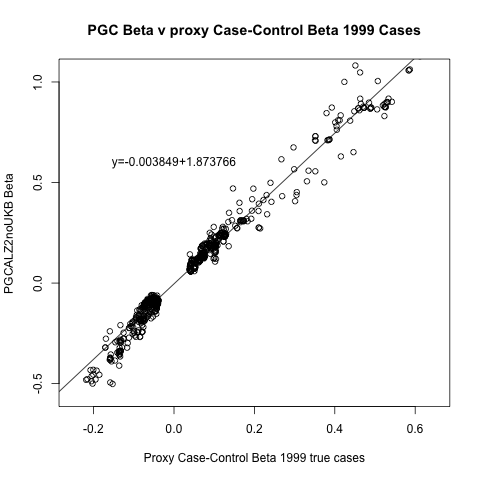
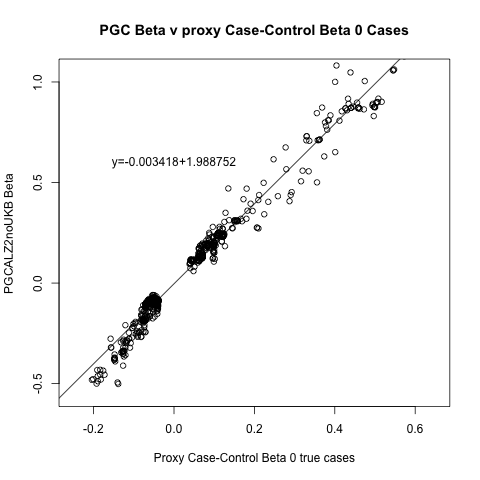
Supplementary Figure 4.5: A) The P-values of the local heritability of regions with nominal (P<0.05) heritability in all proxy datasets except the Jansen dataset highlights that the Marioni and CaseControl datasets have the highest significance in these regions. B) The genetic correlations of loci with significant local genetic correlations between the Jansen dataset and the other phenotypes shows consistent higher local genetic correlations between the Marioni dataset and the Jansen dataset. C) The local genetic correlation between the PGCALZ2noUKB data and the proxy datasets at the loci described in B shows that while differences between proxy datasets exist at these loci the differences are not large enough to affect the genetic correlation between the proxy and non-proxy GWAS data. The error bars represent the 95% confidence intervals.

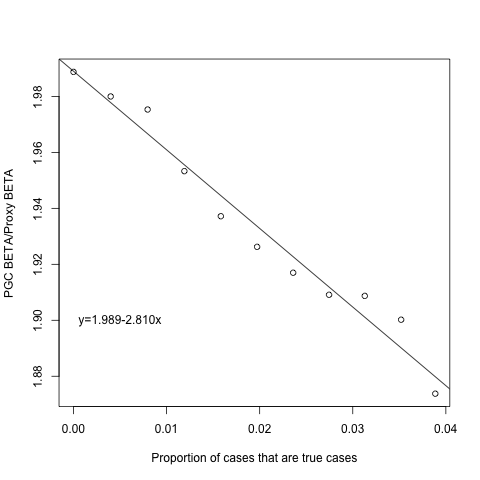
 



*Supplementary Figure 4.6: The effect sizes of significant variants in the proxy phenotype GWAS results compared to the PGCALZ2noUKB and Kunkle effect sizes highlights that the calculated effect sizes in the Jansen, LTFH69, and LTFHplus69 datasets are ~1.2-1.5 larger than the GWAS of clinical AD. The lines represent the line of best fit from a linear model.*

a)

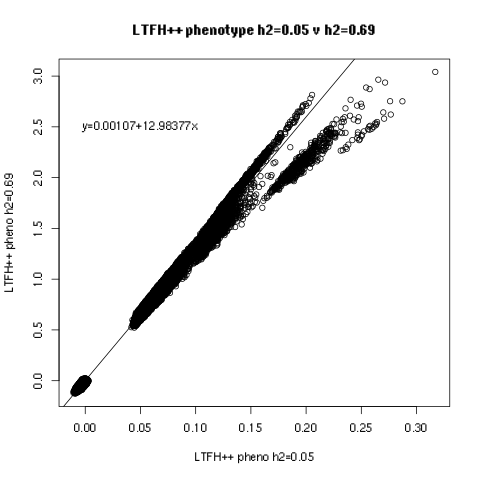
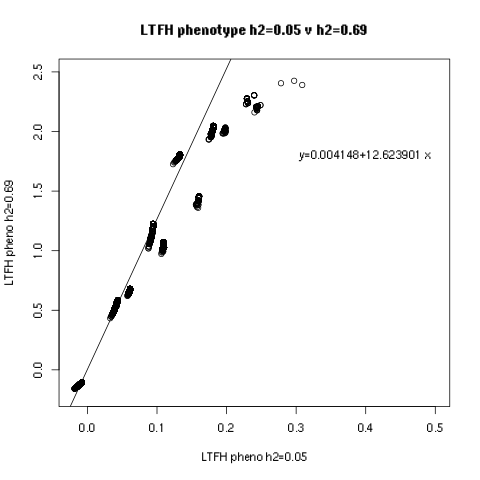




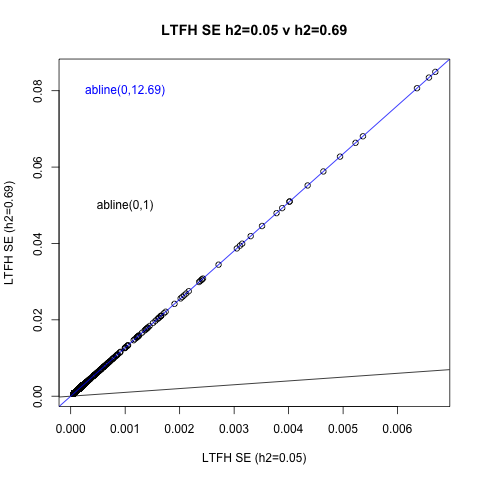
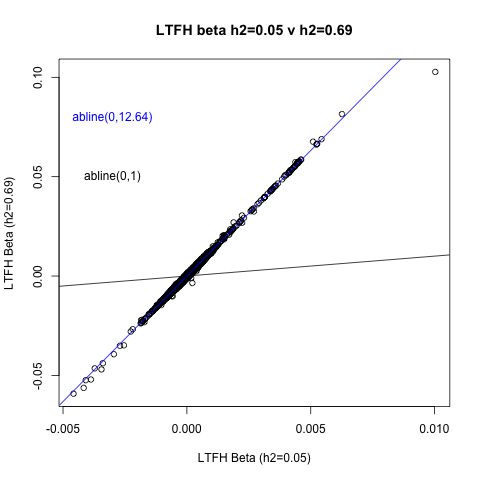
b)

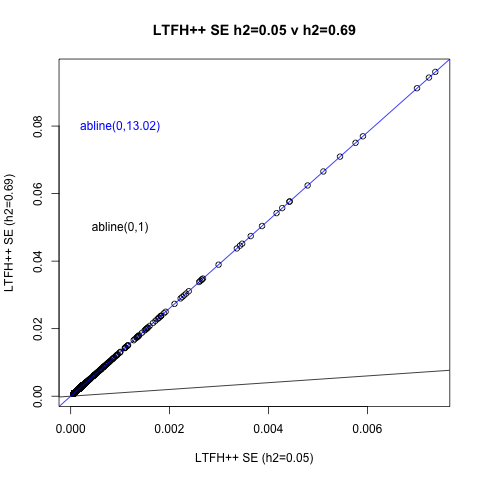
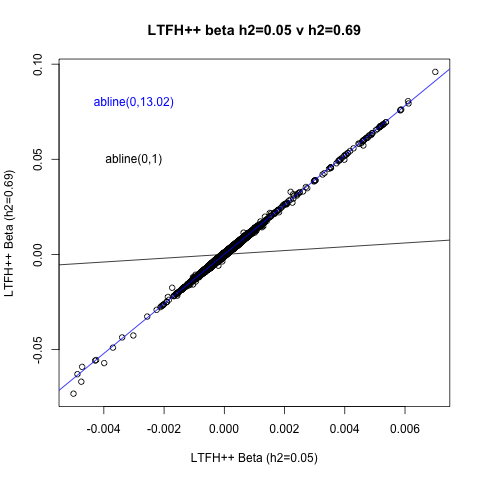
Supplementary Figure 4.7: A) The effect sizes of significant variants in the CaseControl GWAS results compared to the PGCALZ2noUKB effect sizes highlights that including genotyped case status in the CaseControl phenotype leads to less attenuated effect sizes compared to the clinical AD GWAS results (PGCALZ2noUKB). The line represents the line of best fit from a linear model. B) The proportion of cases that are true against the slope of the relationship between the CaseControl and PGCALZ2noUKB effect sizes shows that increasing the number of true cases in the phenotype definition leads to a smaller ration between the CaseControl and PGCALZ2noUKB effect sizes. The line represents the line of best fit from a linear model.

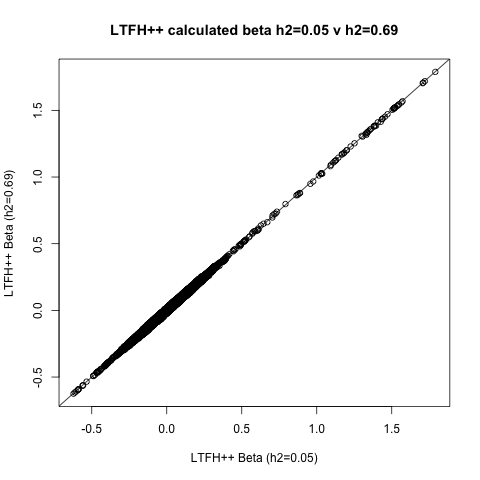
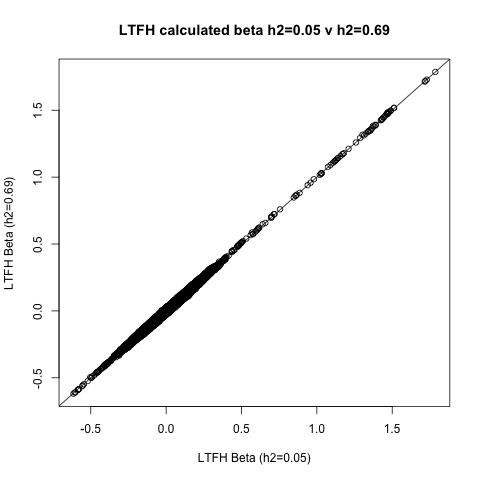
a)



b)

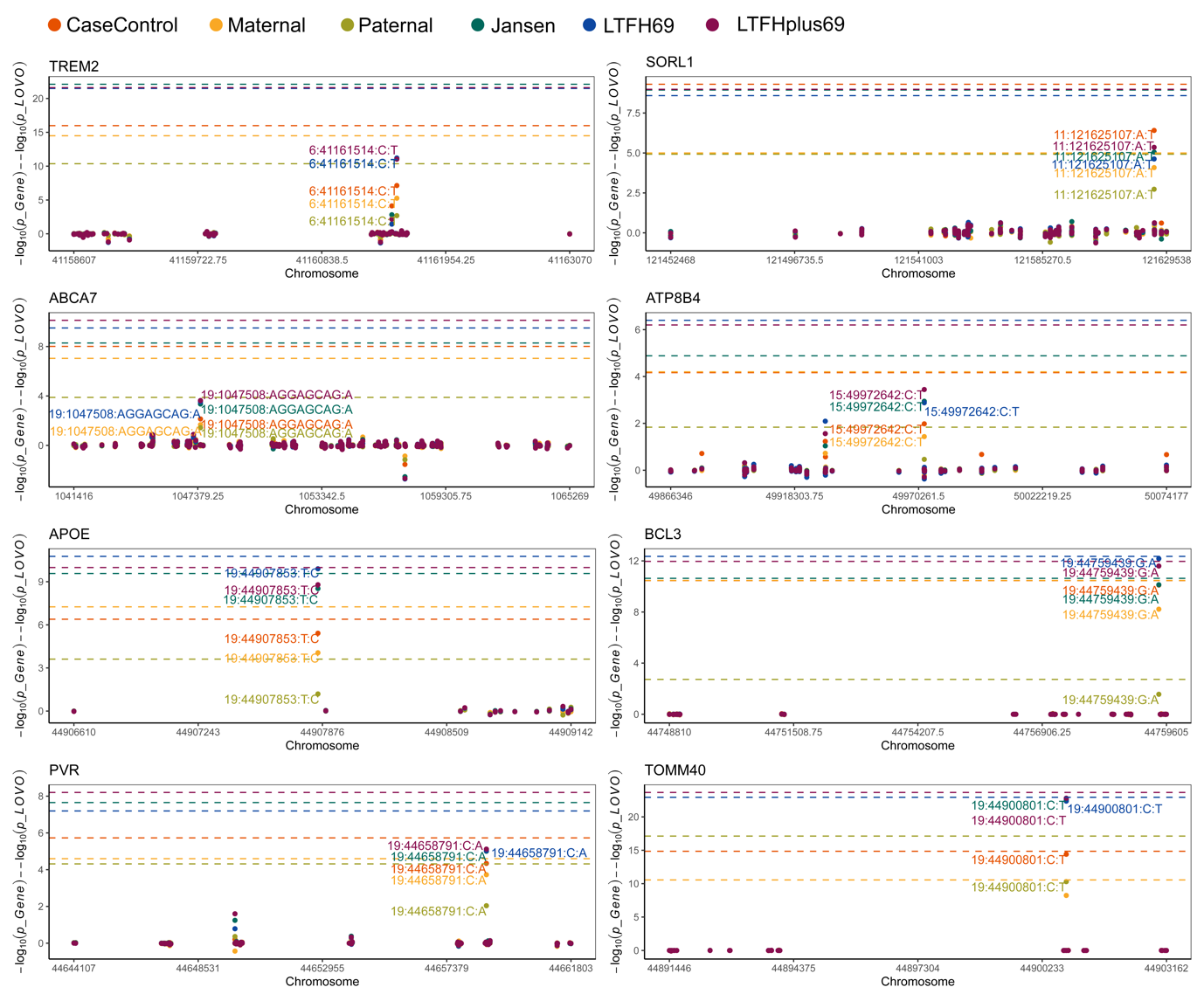






c)

*Supplementary Figure 4.8: A) The phenotype values from LT-FH and LT-FH++ compared after specifying the heritability as 0.69 and 0.05 shows an approximately 12x increase in phenotype value when specifying the larger heritability estimate. The line represents the line of best fit from a linear model. B) The effect sizes and standard errors from the GWAS from the LT-FH and LT-FH++ methods across the different heritability specifications (0.05 and 0.69) shows an approximately 12x increase in effect sizes and standard errors when specifying the larger heritability estimate. The line represents the line of best fit from a linear model. C) The comparison between the calculated effect sizes when specifying different heritabilities in the LT-FH and LT-FH++ methods shows that specifying different heritability estimates does not affect the calculated effect sizes. The line represents the identity line.*



Supplementary Figure 4.9: The amount of gene level association explained by each variant from the leave-one-out analyses shows that one variant can explain a large amount of association in all of the significantly associated genes and one borderline gene (ATP8B4). The dashed lines represent the -log10 P-value of the gene association and the y axis represents how much of that association is lost when leaving out each variant. Points closer to the dashed line represents variants which explain more of the gene-level association. The genotype build in the variant IDs is GRCh37.

## References

1. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput. Biol.* **11**, e1004219 (2015).

2. Werme, J., van der Sluis, S., Posthuma, D. & de Leeuw, C. A. An integrated framework for local genetic correlation analysis. *Nat. Genet.* **54**, 274–282 (2022).

3. Wightman, D. P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer’s disease. *Nat. Genet.* **53**, 1276–1282 (2021).

4. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer’s disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).

5. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer’s disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).

6. Marioni, R. E. *et al.* GWAS on family history of Alzheimer’s disease. *Transl. Psychiatry* **8**, 99 (2018).

7. Burstein, D. *et al.* Detecting and Adjusting for Hidden Biases due to Phenotype Misclassification in Genome-Wide Association Studies. *medRxiv* 2023.01.17.23284670 (2023) doi:10.1101/2023.01.17.23284670.

8. Gatz, M. *et al.* Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatry* **63**, 168–174 (2006).

9. Escott-Price, V. & Hardy, J. Genome-wide association studies for Alzheimer’s disease: bigger is not always better. *Brain Commun.* **4**, fcac125 (2022).

10. Holstege, H. *et al.* Exome sequencing identifies rare damaging variants in ATP8B4 and ABCA1 as risk factors for Alzheimer’s disease. *Nat. Genet.* **54**, 1786–1794 (2022).

11. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* (2017).