Title
Confounding factors in association studies of neuropsychiatric disease: A case study of Bipolar Affective Disorder

Permalink
https://escholarship.org/uc/item/1rv8j36c

Author
McGrouther, Caroline

Publication Date
2014

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA, SAN DIEGO

Confounding factors in association studies of neuropsychiatric disease: A case study of Bipolar Affective Disorder

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Biomedical Sciences

by

Caroline McGrouther

Committee in charge:

Professor Nicholas Schork, Chair
Professor John Kelsoe, Co-Chair
Professor Paul Insel
Professor Walter Kaye
Professor Sanjay Nigam

2014
The Dissertation of Caroline McGrouther is approved, and it is acceptable in quality and for publication on microfilm and electronically:

Co-Chair

Chair

University of California, San Diego

2014
DEDICATION

For my supportive son, Havelock, my lovely husband, Adi, my much-loved nephews, my fun brothers, and my loving and supportive parents

Plus, all of my love to the little boy who did not get a chance to see this lovely world
Traveler, there is no path. Paths are made by walking.

*Antonio Machado*
# TABLE OF CONTENTS

Signature Page ............................................................................................................... iii
Dedication ...................................................................................................................... iv
Epigraph .......................................................................................................................... v
Table of Contents ........................................................................................................... vi
List of Figures ................................................................................................................ vii
List of Tables ................................................................................................................ viii
Acknowledgements ........................................................................................................ ix
Vita ......................................................................................................................................... xi
Abstract of the Dissertation ............................................................................................ xii
Dissertation Overview .................................................................................................... 1
Chapter 1: Background and Significance ...................................................................... 21
Chapter 2: Data Reliability ............................................................................................ 32
Chapter 3: Analytic Methods ......................................................................................... 64
Chapter 4: Phenotypic Heterogeneity ........................................................................... 93
Chapter 5: Conclusions and Future Directions ............................................................. 118
Appendix A .................................................................................................................. 124
References .................................................................................................................. 142
LIST OF FIGURES

Figure 2.1 - Raw mixed genotyped / imputed data ..............................................................58
Figure 2.2 - Performance comparison of AUC and LogR on two mixed genotyped / imputed SNPs ..................................................................................................................59
Figure 2.3 - Cross-group correlations on the Linkage Equilibrium sets .....................60
Figure 2.4 - Adjusting the various mixtures of genotyped and imputed data ..........61
Figure 2.5 - Impact of increasing IQS threshold in cross-group correlations ..........62
Figure 2.6 - Impact of increasing MAF threshold on cross-group correlations ..........62
Figure 2.7 - Two-dimensional histogram of genotyped SNP rank versus imputed SNP ranks ..................................................................................................................................63
Figure 3.1 - Histogram of Skewness vs. Imputation Quality Score vs. SNP-type ....85
Figure 3.2 - Skewness and shrinking p-value thresholds ....................................................86
Figure 3.3 - LogR Rank vs AUC Rank and cLog Rank vs cAUC Rank .........................87
Figure 3.4 - LogR Rank vs AUC Rank in Simulated Data Distributions – Skewness (low, medium, and high) vs. Signal (low, medium, and high) ................................................88
Figure 3.5 - LogR vs AUC rank (All SNPs vs. High Skewness vs. Low Skewness) ....89
Figure 3.6 - Evidence of Enrichment in SNPs with IQS > 0.3 .........................................90
Figure 3.7 - Impact of Imputation Quality Score (IQS) thresholds on the observed enrichment of more highly skewed SNPs (|skewness| > 1) .......................................................91
Figure 3.8 - High IQS and Intermediate Values ..............................................................92
Figure 4.1 - Subgroup Identification as identified by PCA (1,1), Autoencoder (1,2), and t-SNE (1,3) .........................................................................................................................115
Figure 4.2 - t-SNE Subgroups and their performance on the 650 binary DIGS questions ..................................................................................................................................115
Figure 4.3 - Subgroup 10 (S10) SNP Selection Criteria .................................................116
Figure 4.4 - Polygenic modeling of S10 ........................................................................117
LIST OF TABLES

Table 1.1 - Summary of Results from Recent BD GWAS initiatives ........................................27
Table 2.1 - Sklar et al. 2011’s studies stratified by platform and case-control status ..........56
Table 2.2 - Definitions of Group 1 and Group 2 as broken down by numbers of cases and controls and platforms ........................................................................................................57
Table 3.1 - Case / Control / Platform Details for each of the 11 studies in the PGC-BD Wave .............................................................................................................................................83
Table 3.2 - Replication Rates of cAUC vs. cLogR for various Imputation Quality Scales ..................................................................................................................................................84
Table 4.1 - DIGS Details .........................................................................................................113
Table 4.2 - More Detailed Description of Phenotypic Scales used in BiGS GWAS .............113
Table 4.3 - Participants per subgroup and Distinguishing characteristics .........................113
Table 4.4 - Recreating S10 using individual questions – ‘Psychosis’ vs. ‘Psychosis with Mania’ ........................................................................................................................................................................114
Table A.1 - Overview of Phenotypic Trait Heritability ............................................................126
Table A.2 - Landscape of Heritable Phenotypes in BD ............................................................127
Table A.3 - Bipolar Affective Disorder Subphenotype GWAS ..............................................133
Table A.4 - Regions associated with SNPs who had 0.05 > MAF and 0.95 < MAF ..........135
Table A.5 - Regions associated with SNPs who had 0.05 < MAF < 0.95 ..........................138
I would like to thank all of the clinical trial participants that made my research possible. Also, I would like to thank my committee for all of their help, support, and patience over the years. The chairs of my committee, Professor Nicholas Schork and Dr. John Kelsoe, deserve a book of praise. When all of this began, I set my mind on determining what causes bipolar affective disorder. Instead of stopping me and suggesting a smaller goal, they let me try my hand at knocking down that mountain. In the end, I may not have succeeded, but I believe I have been able to make a small contribution through my many attempts. Thank you for being so patient as we waded through a year’s worth of red tape.

I would also like to thank Erin Smith, formerly of the TSRI, now in the Fraser Lab at UCSD, School of Medicine, for answering each and every question I posed to her over the years. I would be very proud if I end up a tenth the scientist that she is. I also enjoyed my brief brushes with Nik’s other graduate students and post docs. They were very kind / welcoming and always extended themselves to help me where they could. Though they were not involved directly with my research, my MSTP and straight-MD classmates kept me sane and very close to on track.

Chapter 1, in part is currently being prepared for submission for publication of the material. McGrathier, CC; Kelsoe, JR; Schork, NJ; and Rangan, AV. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part is currently being prepared for submission for publication of the material. McGrathier, CC; Kelsoe, JR; Schork, NJ; and Rangan, AV. The dissertation author was the primary investigator and author of this paper.
Chapter 3, in part is currently being prepared for submission for publication of the material. McGrouther, CC; Kelsoe, JR; Schork, NJ; and Rangan, AV. The dissertation author was the primary investigator and author of this paper.
VITA

1996  Bachelor of Arts, University of California, Berkeley

2002  Masters of Arts, University of California, Berkeley

2014  Doctor of Philosophy, University of California, San Diego

2014  Doctor of Medicine, University of California, San Diego

Publications:


McGrouther, CC; Kelsoe, JR; Schork, NJ; and Rangan, AV. “Crucial differences between genotyped and imputed data in a 4-platform, 11-study GWAS”. Manuscript in preparation.

McGrouther, CC; Kelsoe, JR; Schork, NJ; and Rangan, AV. Corrected AUC (Area Under the receiver operating characteristic Curve): a viable, potentially more robust, alternative to logistic regression in GWAS. Manuscript in preparation.
Bipolar Affective Disorder (BD) is a highly heritable, complex, genetically heterogeneous disorder with many different, likely interacting, genetic and non-genetic determinants. Polygenic modeling and GCTA tell us that SNPs, single nucleotide polymorphisms, explain >1/3 of BD disease variance. Despite the efforts of meta-meta analyses (intended to increase power) of BD GWAS, replication failures and inconsistent findings are the norm. These problems are not unique to BD. We studied BD's
difficulties as they exemplify those of other neuropsychiatric diseases. We re-evaluated our most fundamental assumptions: i) imputed data is reliable, ii) logistic regression (LogR) is the correct tool for analyzing imputed data, and iii) the BD phenotype is defined correctly. Each assumption was found to be problematic and likely to contribute to the difficulty of identifying the SNPs responsible for causing BD.

Specifically, we found: (1) Imputed and genotyped data are not equivalent: i) imputed data was much more correlated across studies than genotyped data, ii) imputed data was at best a weak predictor of the behavior of genotyped data, iii) upwards of 90% of this cross-study correlation signal disappears with covariate-correction, meaning imputed data should not be analyzed without covariate-correction, and iv) using very strict imputation quality scores (IQS>0.98) can reduce imputed data’s cross-study correlation to close to that of genotyped data. (2) LogR is a suboptimal tool for analyzing imputed data: i) skewness, or lop-sidedness, is a prominent feature of both imputed and genotyped data, ii) skewness is influencing what SNPs are identified by LogR, iii) cAUC (corrected Area Under the receiver operating characteristic Curve) is a more appropriate tool than corrected LogR (cLogR) for imputation-reliant GWAS, e.g., far more enrichment at a p<0.05 (particularly for more skewed SNPs), cAUC replicates at a higher rate than cLogR for (IQS>0.3), and iv) cLogR’s replication rates improve as one applies stricter IQS thresholds (0.3->0.8->0.98), more-skewed SNPs being more likely to replicate than less skewed SNPs, (3) There is genetically-relevant phenotypic heterogeneity present within the BD phenotype. In short, there are many potential explanations for BD replication failures that still allow for SNPs playing a causative role in the BD disease process.
Bipolar Affective Disorder (BD) is a highly heritable, complex, genetically heterogeneous disorder with many different, likely interacting, genetic and non-genetic determinants. Over the last two decades, attempts to identify BD genetic risk variants have been plagued by replication failures and inconsistent findings. This difficulty persists even in the latest meta-meta analyses (intended to increase power) of BD GWAS\(^1\). We focused on BD for this thesis work as its difficulties exemplify those faced by other neuropsychiatric illnesses. While these replication failures impugn hope of finding a single subset of SNPs responsible for all cases of BD, recent work using common polygenic variants to assess risk in psychiatric illness suggests that causal SNPs could account for upwards of 1/3 of the variance in BD diagnosis.[1-3] How can we explain this inconsistency? There are three major assumptions being made in genome-wide association analysis: i) the genetic information is reliable, ii) appropriate analytic techniques are being used, and iii) the phenotype is defined correctly.

Let us consider each of these separately: i) **Reliable data:** Genotyping arrays have been reliably used for numerous years, but GWAS meta-analyses rely upon imputation techniques to create a common data set. There is a growing body of literature suggesting that imputed data can increase between-study heterogeneity, decrease power, and, in the case of mixed genotyped/imputed SNPs, increase type I error. ii) **Analytic techniques:** Logistic regression, corrected for study and Multi-Dimensional Scaling (MDS) components – a rough proxy for genetic ancestry, is the method of choice in GWAS. Logistic regression is designed to detect linear signals, is influenced by

\(^1\) The top SNPs in both of the two most recent meta-meta analyses failed to replicate in their replication arms, and only 15-40% of the top hits were nominally significant.
outliers, and has systematic problems dealing with non-Gaussian/non-uniform data. Despite the fact that imputed data is largely non-Gaussian and may have non-linear features both in the SNP data and the covariates, it is still analyzed as if logistic regression were an appropriate tool. iii) **Phenotype:** The definition of BD is sufficiently broad to allow for a great deal of subject heterogeneity. The criteria are subject to clinician interpretation as well as patient state at the time of the assessment, be it depressed, manic, or psychotic. The potential to conflate different phenotypic subtypes during classification of BD is further exacerbated by the movement towards meta-analyses as the tools used to assess the phenotype and the selection criteria for inclusion vary from study to study. Individuals analyzed because they were deemed to manifest BD may fall into observable clinical or subclinical phenotypic profiles that can be influenced by different sets of genetic determinants. If so, analyzing them together could obscure an otherwise strong signal. There are many potential explanations for BD replication failures which allow for SNP data to play a causative role.

In this thesis, we explored the role played by each of these three potential confounding factors in the BD GWAS. Specifically, we did the following:

i. **Data Set Reliability:** When one visually inspects data from imputed SNPs and genotyped SNPs, it is clear that imputed data is a very different animal. Imputed data is far from uniform/Gaussian and often contains outliers. We believe that imputed data may best be analyzed using the Area Under the receiver operating characteristic Curve (AUC), rather than by using logistic-regression. Specifically, the AUC can be more robust than logistic-regression when data is non-uniform/non-Gaussian, the AUC is not as sensitive to outliers, and the AUC can be approximately
corrected for both linear and non-linear components of covariates. Therefore, the AUC seems uniquely suited to evaluate imputed data.

**[Result 1]:** To that end, we broke our available data (Ncase=7,481, Ncontrol=9,250) into two groups of approximately equal size which preserved original study associations. We calculated the AUC (both uncorrected and corrected for MDS and study) for each group over a set of 282,551 SNPs that were in linkage equilibrium (I_LE) and had an imputation quality score (IQS)>0.3. Those 282,551 SNPs included purely imputed, mixed imputed/genotyped, and purely genotyped SNPs. We found that the purely and partially imputed data is much more correlated across groups than the purely genotyped data prior to correction.

a. The cross-group correlation exhibited by the uncorrected AUC across the complete set of I_LE SNPs was 0.526 (p=0), whereas the cross-group correlation exhibited by the corrected AUC (cAUC) was only 0.031 (p=4.4x10^{-61}). A comparable analysis of purely genotyped, i.e., non-imputed, data found correlations of 0.063 (p=5.8x10^{-20}) and 0.036 (p=1.5x10^{-7}) respectively. Thus, we believe that these cross-group correlations are spurious; driven by artifacts of imputation; and not disease-related. We will later refer to these cross-group correlations as ‘non-disease related signal’.

b. When we analyzed I_LE this time with logistic regression (LogR), uncorrected LogR had a correlation of 0.123 (p=0) and the corrected LogR had a correlation of
0.031 \ (p=1.0\times10^{-60}). \ A \ comparable \ analysis \ of \ the \ purely \ genotyped \ data \ found \ cross-correlations \ of \ 0.056 \ (p=6.2\times10^{-16}) \ and \ 0.037 \ (p=6.8\times10^{-8}) \ respectively. \ Note \ that \ AUC \ (see \ (a) \ above) \ detected \ a \ more \ dramatic \ non-disease \ related \ signal \ than \ did \ logistic-regression. \ Thus, \ we \ used \ the \ AUC \ and \ cAUC \ to \ measure \ and \ analyze \ cross-group \ correlations \ going \ forwards.

c. \ Now, \ using \ AUC, \ we \ examined \ the \ relationship \ between \ this \ cross-group \ correlation \ and \ imputation \ by \ starting \ with \ 100\% \ imputed \ SNPs \ and \ gradually \ adding \ in \ SNPs \ with \ increasing \ fractions \ of \ genotyped \ data. \ The \ purely \ imputed \ SNPs \ were \ correlated \ at \ a \ 0.37 \ (p=0) \ level. \ This \ correlation \ jumped \ to \ ~0.71 \ with \ the \ introduction \ of \ mixed \ imputed / genotyped \ SNPs. \ We \ continued \ to \ introduce \ SNPs \ with \ increasing \ fractions \ of \ genotyped \ data \ until \ we \ had \ the \ entire \ I_LE \ dataset \ (282,551 \ SNPs). \ The \ correlation \ decreased \ from \ ~0.71 \ to \ 0.526). \ We \ then \ started \ with \ the \ entire \ I_LE \ dataset \ and \ gradually \ removed \ those \ SNPs \ that \ were \ purely \ or \ partially \ imputed. \ The \ correlation \ decreased \ from \ 0.526 \ to \ 0.069 \ as \ the \ amount \ of \ imputed \ data \ decreased \ to \ 0.

d. \ Our \ next \ question \ was \ whether \ or \ not \ these \ imputation-associated \ correlations \ were \ driven \ primarily \ by \ imputations \ of \ low \ quality. \ To \ assess \ this, \ we \ thresholded \ our \ SNP \ set \ by \ IQS \ and \ repeated \ our \ analysis \ in \ (c). \ We
had to increase our IQS threshold from the baseline of 0.3 all the way up to 0.98 (out of 1.00) before we saw a dramatic change in the pattern of these correlations.

e. Our next question was whether or not these imputation-associated correlations were driven primarily by SNPs with small MAF. The imputation-associated correlations were again robust to adjustments in number of SNPs analyzed or MAF thresholds. [Please note: SNPs with more extreme frequencies were more highly correlated in imputed data than SNPs of less extreme frequencies. Nevertheless, even SNPs with MAF 40-60% still exhibit a much larger correlation when imputed than not.]

f. To confirm that (c) was not limited to an AUC-analysis, we repeated (c) using logistic-regression, rather than AUC. The imputation-associated correlations were still observed.

[Result 2]: We also found that the ‘top’ imputed SNPs (i.e., the imputed SNPs that were the most significant according to corrected logistic-regression) were not a good indication of the top genotyped SNPs and vice versa.

a. Specifically, there were ~106 SNPs with some genotyped data. For each SNP with imputation quality score > 0.3, we performed logistic regression, correcting for MDS components and study, for the genotyped and imputed data separately. To control for the effects of varying Ns, we focused on the ~133K SNPs with between
40-60% genotyped data. If one considers the top 0.1% (~133) of the imputed SNPs (i.e., top SNPs from the imputed data set), only 3 are in the top 0.1% (~133) genotyped SNPs, only 10 in top 1% (~1,335) of genotyped SNPs, and only 23 in top 10% (~13,353) of genotyped SNPs. If one considers the top 0.1% (~133) of genotyped SNPs, only 3 are in the top 0.1% (~133) of imputed SNPs, only 3 are in top 1% (~1,335) of imputed SNPs, only 19 in the top 1% (~13,353) of imputed SNPs. Though this is better than random (3/10/23 versus 0.13/1.3/13), it suggests that using imputed data to identify the top signals in genotyped SNPs is not particularly helpful. In our sample, imputed data predicted genotyped data slightly better than genotyped data predicted imputed data.

b. Motivated by [Result 1], we checked to see if using stricter imputation quality scores increased the predictive ability of imputed data. It did slightly improved both imputed data’s ability to predict genotyped data and vice versa, i.e., with IQS > 0.98 (N=47,464 SNPs) the overlap between the top 0.1% imputed SNPs and the top 0.1%/1%/10% genotyped SNPS was 2/4/10, respectively. The converse: the overlap between the top 0.1% genotyped SNPS and the top 0.1%/1%/10% imputed SNPs was 2/2/7 respectively. These are both better than random (i.e., 0.048/0.48/4.8).
Conclusions: 1) HapMap2-imputed data does not behave like genotyped data. 2) Mixed imputed / genotyped data is more correlated across groups than purely imputed SNPs. Purely imputed SNPs are much more correlated than purely genotyped data. 3) Using a very strict IQS threshold (0.98) almost returns the correlation observed in purely imputed data to the level observed in purely genotyped data. 4) HapMap2-imputed data is at best a weak indicator of genotyped SNP behavior. 5) HapMap2-imputation has introduced a large non-disease related signal. Logistic regression is largely blind to that signal. We suspect that AUC is better able to detect this signal than logistic regression because, unlike logistic regression, AUC is a) distribution-invariant, b) robust to outliers and c) its trial to trial variance is bounded. Given the growing body of literature on the problems with imputation, this problem may not be specific to BD and AUC may be the tool researchers need to assess both the non-linear and linear influences introduced by imputation.

Discussion points: 0) We may want to restrict our future analyses of the BD data to fully genotyped and/or fully imputed SNPs, because the mixed imputed genotyped data had the most non-disease related signal (as corroborated by the simulation studies of Li et al.) 1) We may want to use a stricter imputation-quality-threshold in our future analyses of imputed data, since we found that a very strict threshold mitigates the non-disease related signal (again, as corroborated by the literature). 2) Instead of genotyping promising imputed SNPs perhaps we should be collecting the data necessary to impute them as the success of imputation may be because it is a local measure. 3) If genotyping only a few locations is the
only option, perhaps we should rerun PLINK’s clump algorithm, which forces a greater emphasis on SNPs with more genotyped data, 4) Alternately, since 93% of the signal seems to disappear after correcting for study and MDS components, can one trust what remains to be truly disease-related? 5) It is essential that we rerun these analyses on the 1000 Genomes Imputation of the PGC-BD data to determine what, if anything, changes.

ii. **Analytic Methods:** Though genotyped and imputed data have different characteristics, we set aside that question here and focused on the data set as analyzed by the PGC-BD in Sklar et al. 2011 paper. Visual inspection of imputed data suggests that skewness, or lop-sidedness, and outliers are common. Even wholly genotyped data can be quite skewed. As was commented in the previous section, standard logistic regression (LogR) is not well-suited to analyzing skewed data, is strongly influenced by outliers, and only corrects for the linear relationships between the SNP data and its covariates. In addition to these weaknesses or perhaps because of them, logistic regression was unable to detect much of the non-disease related signal detected by AUC in the Data Reliability Section. If skewness is a defining feature of imputed/genotyped data, then we hypothesize that logistic regression may be a sub-optimal tool for detecting the disease-related signals in imputation-reliant GWAS and, therefore, its use may be undermining our ability to detect true signals in BD imputation-reliant GWAS.

In parts B and C of the Appendix, we demonstrated the following strengths of AUC: a) AUC is robust to skewness and outliers, b) its trial-
to-trial variance is bounded, c) it is well-suited to voting models (those that rely on few SNPs), and d) it can easily be corrected for linear and non-linear covariate interactions. Given that AUC is strong where LogR is weak (skewness, outliers) and it was able to detect a strong non-disease related signal in the PGC-BD imputed data, we hypothesized that Corrected AUC (cAUC) is a viable alternative to logistic regression for analyzing imputed data.

[Result 1]: We confirmed the presence of skewness in our data directly

a. We calculated the mathematical quantity, skewness, directly for each SNP. We found that highly skewed SNPs were present in every category of data, i.e., genotyped, mixed imputed/genotyped, imputed. Range of skewness values were as follows: mixed imputed/genotyped data > purely imputed >> genotyped. Using stricter imputation quality Score (IQS) thresholds narrowed the range of skewness values. Only the strictest IQS-thresholding (i.e., 0.98) was able to bring the skewness of the imputed data down to the levels observed in genotyped data. As genotyped data itself can be quite skewed, this did not eliminate the potential confound of skewness. Similarly, limiting oneself to SNP sets of varying degrees of significance, i.e., $<10^{-1}$ to $10^{-9}$, did not eliminate extreme skewness values.
b. Next, we used AUC to rank each SNP in order of significance. We did the same using LogR. Then, we compared the rank order as assigned by AUC with that assigned by LogR via a two-dimensional histogram. We did this because AUC and LogR should assign similar ranks to SNPs if their distributions are not skewed or long-tailed. (Long-tailed distribution, while not impossible, should not be a large concern with imputed data because it is bounded, i.e., within [0,2].) By examining their relationship, we sought to gain perspective on how skewness has been influencing our logistic regression results. After binning the resulting rank orders and creating our 2-dimensional plot, we found that AUC and LogR ranked the SNPs quite differently. We also observed that a very good AUC was likely to yield a consistent and very good LogR p-value whereas the opposite was not true. These findings were consistent with theoretical work and our simulations of a large signal in the presence of a high degree of skewness. This supported our supposition that AUC has the potential to outperform logistic regression in analyzing imputation-reliant GWAS data.

Of note, we compared AUC with LogR rather than cAUC with cLogR because AUC is corrected for both linear and non-linear covariates, but logR is only corrected for linear components. Even if there were no skewness
present in the data, we would expect cAUC and cLogR to rank the SNPs differently simply because of different levels of covariate correction. This expectation was borne out by the following analysis: When we did compare cAUC with cLogR, cAUC and cLogR ranked SNPs differently, suggesting either the influence of skewness or differing degrees of covariate correction. What stood out in this second analysis was the striking lack of relationship between cAUC and cLogR for all but the top SNPs. This cAUC-vs-cLogR result is most consistent with simulated data possessing high skewness and low signal.

[Result 2]: We found evidence that AUC and cAUC are much more strongly associated with enrichment at a $p < 0.05$ level than are LogR and cLogR.

a. While LogR and cLogR do demonstrate a statistically significant enrichment at a $p < 0.05$ level, the enrichment is much more pronounced for AUC and cAUC.

b. In the case of cAUC, the enrichment is driven mostly by SNPs with more skewness ($|\text{skewness}| > 1$). For cLogR, the enrichment is driven mostly by SNPs with less skewness ($|\text{skewness}| < 1$).

c. Increasing the Imputation Quality Score threshold from 0.3 to 0.8 to 0.98 seems to have very little impact on the enrichment levels observed with AUC, LogR, cAUC, and cLogR with one notable exception, i.e., cAUC enrichment
of more skewed (|skewness| >1) SNPs decreases as the IQS threshold increases. That being said, it does not return to the levels of enrichment seen with cLogR.

**[Result 3]:** cAUC was far less likely than cLogR to assign a SNP a strict p-value < 10^{-5}. In fact, the ratio of SNPs satisfying such a strict threshold for cAUC-vs-cLogR was approximately 1 : 3. This may be influenced by the fact that cAUC and cLogR correct for covariates in different ways.

**[Result 4]:** Both cAUC and cLogR were far more likely than chance to assign a less-skewed (|skewness| <1) SNP a strict p-value < 10^{-5}, i.e., cAUC vs. cLogR vs. expected - 75% vs. 60% vs. 52%.

**[Result 5]:** cAUC’s replication rates for SNPs satisfying IQS > 0.3 AND p < 10^{-5} are statistically significantly better than those of cLogR. (5.9% vs 1.9%, p=0.00040).

a. It is tempting to say that cAUC’s replication rate is driven by less-skewed (|skewness| <1) SNPs (9.3% vs.1.0%), but this would be pre-mature since we sought to replicate very few of the more-skewed (|skewness| > 1) SNPs, i.e., ~2 SNPs per A-vs-B pair. (See [result 3] and [result 4]).

b. The trend of cAUC replicating better than cLogR when the IQS threshold is shifted from 0.3 to 0.8, but in both instances (IQS=0.8, 0.98), there are simply too few SNPs being considered to make a credible comparison of replication rates between cAUC and cLogR. (recall,
however, that cAUC did reliably predict enrichment better than cLogR).

c. When one increases the IQS threshold to 0.98, 85% of the SNPs assigned a \( p < 10^{-5} \) by cAUC in Wave A are removed, leaving only \(~1\) SNPs for replication. This percent drop was roughly commensurate to that seen with cLogR (\(~60\%) drop) and makes it impossible to comment whether cAUC has bias towards assigning significance to SNPs of lower imputation quality.

[Result 6] Replication rates for cLogR improve as one successively increases the IQS threshold from 0.3 to 0.8 to 0.98 (1.9% to 3.4% to 7.2%).

a. These replication rates are mostly driven by more-skewed (|skewness| > 1) SNPs (5.4% to 8% to 19.3%) though the improvement is also present in less-skewed (|skewness| < 1) SNPs (2.6% to 3.6% to 4.9%).

Conclusions: 1) Many of the SNPs in the PGC-BD dataset are skewed, this skewness seems to be influencing the results given by logistic regression, and, in keeping with our results from the Data Reliability section, much of the signal present in the data set disappears after correction. 2) Despite this reduction in signal, both cAUC and cLogR demonstrate statistically significant enrichment at a \( p < 0.05 \) level, with cAUC showing the most dramatic enrichment. 3) cAUC’s enrichment is driven by its improvement in correctly assessing the more-skewed (|skewness| > 1) SNPs. 4) cAUC has far less enrichment than cLogR for
the less-skewed ($|\text{skewness}|<1$) SNPs ($p=1.89 \times 10^{-11}$). Since we have seen that logR and AUC behave similarly in the absence of skewness, one possible explanation for this enrichment difference is the greater degree of covariate correction for cAUC, i.e., linear + non-linear covariate relationships. 5) cAUC assigns a $p$-value $< 10^{-5}$ to $\sim 1/3$ of the SNPs assigned a $p$-value $< 10^{-5}$ by cLogR. One possible explanation for this is the increased covariate correction possible with cAUC, 6) cAUC replication rates are statistically significantly better than cLogR for SNPs with IQS $> 0.3$. It is tempting to attribute this difference to its performance on the less-skewed ($|\text{skewness}| < 1$) SNPs, but there are insufficient SNPs with $|\text{skewness}| > 1$, $p < 10^{-5}$, and IQS $> 0.3$ to credibly comment on cAUC’s replication rate for more-skewed ($|\text{skewness}| > 1$) SNPs, 7) cLogR replication rates improve as one uses increasing IQS thresholds (0.3 to 0.8 to 0.98). This is most dramatic within more-skewed ($|\text{skewness}| > 1$) SNPs, but the trend is still true for less-skewed SNPs ($|\text{skewness}| < 1$).

**Discussion Points:** 1) Given the differential performance of cAUC and cLogR, particularly since this performance seems to be influenced by skewness, it seems reasonable to assert that some of the PGC-BD’s replication difficulties stem from its reliance on cLogR without consideration of skewness. 2) It seems important to understand: a) why cAUC identifies far fewer extremely significant SNPs than cLogR, b) why cAUC is far less likely to assign a significant $p$-value to a more-skewed SNP rather than a less-skewed SNP, and c) why cLogR is more likely to replicate a more-skewed ($|\text{skewness}| > 1$) SNP than a less-skewed
(|skewness| < 1) SNP, particularly those with a very high IQS Score. Further, it seems important to inquire what role, if any, lack of signal, outliers, non-linear relationships between covariates, and minor allele frequencies are playing in the above. 3) Since risk prediction bridges enrichment and replication, it seems worthwhile to determine if there is a marked difference in the performance of linear-additive models based upon cAUC and cLogR. 4) Independent of whether cAUC is a viable alternative for analyzing the imputation-reliant PGC-BD GWAS, it seems worth questioning whether we should have faith in any signal that remains, given that the majority (i.e., 93%) of the original signal is removed after performing an approximate correction. This idea is further supported by how dramatically the histogram of rank order of AUC vs. rank order of LogR changes after correction for covariates.

iii. **Phenotypic heterogeneity:** In hopes of identifying genetically-relevant phenotypic heterogeneity, we applied multiple unsupervised machine learning methods to the raw Diagnostic Interview for Genetic Studies (DIGS) data for 4335 subject with BDI, BDII, SABD (schizoaffective bipolar disorder).

[**Result 1**]: One method clearly outperformed the others; identifying five distinct subphenotype groups, defined broadly by +/- psychosis with mania, +/- panic attacks, +/-depression. The subgroup that was most phenotypically distinct was the group featuring (i) no panic attacks, (ii) no psychosis with mania, and (iii) no depression; i.e. this group corresponds to unipolar mania, without psychosis during mania or
panic attacks. This work confirmed the existence of strong phenotypic heterogeneity within the BD umbrella.

[Result 2]: We next determined whether this observed phenotypic heterogeneity was genetically-relevant. We found that the heterogeneity as quantified by Morris et al.’s Subphenotype Heterogeneity Test (SHT)[4] was inversely related to replication at a p<0.05 level in the PGC-BD dataset, i.e., the greater the heterogeneity the less likely to replicate at a p<0.05 level. We also found significant heterogeneity (PSHT<0.05) in 4 of the 32 SNPs that failed to replicate in the PGC-BD Sklar et al. 2011 paper. This was 2.5x more than we would have expected by chance. It was also worth noting that neither of the two SNPs which did replicate showed significant SHT.[5] Thus, in conclusion, it appears as though phenotypic heterogeneity does indeed permeate the diagnosis of BD, and that this phenotypic heterogeneity has a genetic signature that impacts the replicability of at least a few relevant SNPs.

[Result 3]: We then set out to determine if limiting our analysis to those BD subjects within a given subphenotype group would result in the identification of stronger genetic signature. Specifically, we focused our attention on unipolar mania with an absence of panic attacks and psychosis during mania (S10). We selected this group because it was the most phenotypically distinct subgroup identified by our machine-learning methods, and it seemed to be defined by phenotypes being collected by the PGC-BD, our intended source of replication samples.

a. First we sought to assess the significance of this group using a risk-prediction model. We found that the linear-
additive model based on S10 explained the greatest variance with a p-threshold \((p_T)\) of \(7.5 \times 10^{-4}\). This is considerably smaller than the p-threshold in schizophrenia\[1\] \((p_T=0.5)\) suggesting a much tighter link between significant SNPs and subgroup-control status. It is also worth tempering this point by noting that S10's polygenic model explained a great deal less of the disease variance (0.15% vs. ~3% in schizophrenia).

b. Next, we investigated the most promising SNPs and pathways associated with this group. Only 4 of the 52 top S_{10} SNPs had been studied extensively and had no neurological link. Also, those implicated generally had higher effect sizes than the PGC-BD top SNPs. Amongst the genes within 20kb of these SNPs were GRIN2B (glutamate receptor, ionotropic, N-methyl D-aspartate 2B) and ITPR1 (inositol 1,4,5-triphosphate receptor, type 1). These play many roles in neurophysiology. As a result, the following pathways were implicated when we ran GeneGo pathway analysis on the 33 genes within 20kb of the 52 SNPs: 1) NMDA-dependent postsynaptic long-term potentiation in CA1 hippocampal neurons \((3/80, p=9.2 \times 10^{-5})\), 2) Dopamine D2 receptor transactivation of PDGFR in CNS \((2/26, p=4.0 \times 10^{-4})\), 3) nNOS signaling in neuronal synapses \((2/29, p=5.0 \times 10^{-4})\), 4) Calcium signaling \((2/45, 1.2 \times 10^{-3})\), and 5) Alpha-2 adrenergic receptor regulation of
ion channels (2/47, p=1.3x10^{-3}). If one were to loosen the
distance criteria, CREB1, PIK3R1 (Phosphatidylinositol 3-
kinase regulatory subunit alpha), and DGKG
(diacylglycerol kinase, gamma) are each the nearest gene
to a significant region. With these additions, the signals in
NMDA-dependent postsynaptic long-term potentiation and
calcium signaling strengthen, and the lithium-sensitive
phosphatidyl inositol pathway is implicated.

Given these interesting features of S_{10}, we were disappointed
when we realized that we would not be able to create a replication set
from the available PGC-BD phenotype data. Specifically, the
‘psychosis’='No' field collected by the PGC-BD proved to be quite
different our ‘psychosis with mania’='No' field. Specifically, using their
‘psychosis’='No' ruled out a significant number of S_{10} subjects that would
have appropriately been ruled in with ‘psychosis with mania’='No'. This,
coupled with the fact that ‘panic attacks’ Y/N was barely reported, meant
that we were unable to determine if our findings were replicable.

**Conclusions:** 1) There is demonstrable phenotypic heterogeneity within
the BD umbrella. 2) This heterogeneity is likely genetically relevant. 3)
Due to difficulties creating an appropriate replication data set, we cannot
know whether or not any of our subgroups would have been associated
with replicable signals.

**Discussion:** 1) Our work, as well as the success of similar work by Meier
et al. and Kerner et al., demonstrates that phenotypic heterogeneity is
introducing noise into binary disease / no-disease BD GWAS.[6, 7] 2)
Dimensionality-reduction techniques are an efficient way to reduce a potentially overwhelming amount of phenotypic information, if one were to run a PHE-WAS per topic, to a reasonable level. 3) What limits the success of this form of subphenotype analysis is not the methods available to researchers, but actually the mainly self-reported and inconsistently-collected phenotype data available. Given that we do not yet understand the disease processes involved and by extension the relevant biomarkers, it makes the most sense to analyze the genome directly for homogeneous genetic subgroups. Biclustering methods, i.e., methods that identify subsets of SNPs and cases that are distinct from controls, are a natural fit. Based upon what we know about the differences between genotyped and imputed data, the non-disease related signal present would obscure the true signal for any biclustering method that cannot correct for covariates like MDS components and study of origin. At this time, there is no such method, but they are in the pipeline. If one had a large number of subjects genotyped on the same platform or who had undergone whole-genome sequencing, the difficulty would be that these methods are computationally expensive and a thoughtful approach to SNP selection for analysis would be appropriate. 4) One problem with this plan, regardless of the outcome is that it seems unlikely that we should have faith in any signal that remains, given that the majority (i.e., 94%) of the original signal is removed after performing an approximate correction.

Currently, BD GWAS meta-analyses suffer from replication failure. While it is tempting to conclude that there is no detectable BD signal in the GWAS data, our
findings suggest otherwise. Specifically, we propose that there is indeed a signal, albeit a heterogeneous one that is potentially confounded by (i) imputation, (ii) the reliance on logistic-regression, and (iii) a coarsely-defined phenotype. With the movement towards whole genome analysis, these problems with imputation will become a thing of the past. In the meantime we suggest reanalyzing the highest quality data available; switching from corrected logistic regression to corrected Area-Under-the-Curve (AUC). Finally, it may be possible to address phenotypic heterogeneity with the development of robust and efficient biclustering methods.
1. A General Overview of Bipolar Disorder: Bipolar Affective Disorder (BD) is a complex multi-factorial disorder with many genetic and non-genetic determinants. Identifying and characterizing each contributor to BD will be difficult – especially for genetic factors – since the effects of any one factor may be obscured by others. This complexity is seen when assessing clinical phenotypes: there are two acknowledged subtypes of BD (BDI and BDII). BD generally is characterized by periods of elevated mood, with symptoms that can include irritability or agitation, decreased need for sleep, racing thoughts, easy distractibility, inflated self-esteem, and increase in goal-directed or pleasurable activities, as well as periods of depressed mood. Unlike BDII, BDI’s mood elevation is sufficient to cause severe impairment of social or occupational functioning. It may lead to hospitalization or have psychotic features.  

Based on a 2003 survey of 85,358 subjects using the Mood Disorder Questionnaire, a validated screening instrument for BD I and BD II with sensitivity of 0.73 and specificity of 0.90, the prevalence of BD I and II was estimated at 3.7%. In another survey which involved 9,282 English-speaking Americans adults, an estimate of the lifetime prevalence was 4.4% (1.0% BDI, 1.6% BDII, 2.4% sub-threshold BD). These estimates are substantially higher than the 0.4-1.6% estimate cited in the DSMIV-TR. The prevalence and complexity of BD has huge social repercussions: In 1991, direct healthcare costs for those with BD were estimated at $7.6 billion or $12,666 per person. The treatment-related costs were $10,934 per person. The indirect costs, e.g.
lost productivity, were estimated to be $37.6 billion[13] while the lifetime costs for those with onset of BD onset during 1998 were estimated to be $24 billion dollars.[14]

BD is thought to be highly heritable. First degree relatives of those with BD have a 13.6 fold (95% CI-11.8-15.7) higher risk of developing BD.[15] In a 1977 study of 34 DZ and 37 MZ twins, Bertelsen et al. found MZ and DZ concordance rates of 62% and 8% respectively.[16] Using the DSM-IV definition of BD, McGuffin et al. found MZ and DZ concordance rates of 40% and 5.4% respectively.[17] In a review of BD twin studies (n=34,13, 25 pairs), heritability was found to be 59%, 79%, 87% respectively while a more recent meta-analysis places heritability at 85%[18, 19]. These suggest that a major genetic component drives BD and thus, may facilitate the identification of specific BD susceptibility genes. However, the mode of inheritance for BD has not yet been determined. These inconsistencies in observed inheritance patterns likely reflect BD’s genetic heterogeneity.

1.B Bipolar Affective Disorder Genetic Studies

1.B.1 Genome-Wide Linkage Studies: Over 40 BD linkage studies related to BD have been performed in the last two decades. Unfortunately, none of the genetic loci identified were consistently replicated across the studies. Hoping to account for the individual analyses that lacked sufficient statistical power to detect more subtle influences, investigators performed meta-analyses that combined the various linkage studies. The first such analysis identified 13q32 (p< 6x10^{-6}) and 22q12-13 (p<10^{-5}).[20] This finding was called into question when a larger meta-analysis (which also included unpublished data) failed to replicate these findings. This larger study in turn identified the following regions of nominal significance (p<0.05): 8q24, 9p21-22, 10q11-22, 14q24-32, and 18p-18q21.[21] The last and largest meta-analysis, included 5,179 subjects in 1067 families, found that regions in 6q and 8q achieved genome-wide significance, LOD scores of 4.19
and 3.40 respectively. There were also suggestive loci on 9p (LOD 2.04) and 20q (LOD 1.91). [22] Though a few regions have been noted, there are no regions of large effect which have consistently declared themselves. This inconsistency suggests either loci of small effect or noise in the system.

1.B.2 Candidate Gene Association Studies: Candidate gene association studies are another means of identifying genes implicated in BD. Such candidate genes are either identified by understanding the biology behind the disease or are pursued because they are located near to a region identified through a linkage study. 40+ candidate genes have been identified. As was the case with linkage studies, replication has proven difficult and no gene has been replicated in every study. The most consistent candidate genes have been SLC6A4/5-HTT (serotonin transporter), BDNF (brain-derived neurotrophic factor), COMT (catechol-O-methyltransferase), DISC1 (disrupted in schizophrenia 1, coding for a neuronal growth-related protein), DTNBP1 (dysbindin), DAOA (D-amino acid oxidase activator), and NRG1 (neuregulin 1). [23] In Seifuddin et al. meta-analysis of candidate genes BD association studies (487 papers in all) found polymorphisms in BDNF, DRD4, DAOA, and TPH1 were found to be nominally significant with a p<0.05, but they were not significant after correcting for multiple testing. [24] These meta-analyses support the idea that multiple genes each have very little effect in contributing to the disease phenotype.

1.B.3 Genome Wide Association Studies: BD GWAS seem to be fraught with problems similar to those found in the genome wide linkage studies described above as few compelling and replicated insights have resulted from their use. Table 1.1 below lists the studies, sample sizes, role of imputation, genes of interest that emerged from the studies, and overall results. One will note that each meta- or meta-meta-analysis seeks
to increase the power to detect SNPs of smaller effect, but replication of the top SNPs proves elusive.

Across the various GWAS, the following gene regions have achieved genome-wide significance: ANK3, SYNE1, CACNA1c, ODZ4, a SNP between RHEBL1 and DHH, TRPC4AP and NCAN. Since their identification, ANK3 and SYNE1 have both been replicated and failed to replicate in independent samples. [5, 25-28] Let us consider the case of ANK3. After failing to replicate in Sklar et al, it was validated in an East Asian sample making it less likely to be a false positive in Sklar et al. The fact that it only achieved $p < 0.1$ at 9 out of 11 sites included in the PGC-BD sample while still achieving genome-wide significance could reflect phenotypic recruitment differences at the sites.[5] Similarly, the SNP, rs7296288, between RHEBL1 and DHH failed to replicate in sklar et al. only to achieve genome-wide significance when additional samples were added to those original PGC-BD samples.[27] Though CACNA1C has performed fairly consistently, ODZ4 failed to replicate after Bonferroni correction in an independent sample.[27] To this author’s knowledge, replication studies have not been performed yet for TRPC4AP or NCAN, but it is worth noting that they were not previously identified in other GWAS meta-analyses as top hits. This intermittent replication suggests that the results of our GWAS depend a great deal on how we group our subjects for analysis. In essence, either we have not yet increased power sufficiently or power is not the reason why replication proves elusive. As discussed above, data quality, analytic methods, and the existence of either phenotypic or genetic sub- or bi-clusters may be responsible.

1.C Polygenic models: Since so far increasing power has had limited success in identifying replicable findings, it is tempting to say that SNPs must not play a large causative role in BD. Drawing this conclusion would be premature as both the weighted sum of the log odds ratio and Genome-wide Complex Trait Analysis (GCTA) technique
have been used to show that common SNP variants can explain upwards of a 1/3 of the variance of the bipolar diagnosis.[1, 3] Specifically, Purcell et al. scored schizophrenia cases and controls via a weighted sum of log odds ratios. They found that they could explain the most diagnosis variance when they included SNPs with p-values < 0.5 in their metric, i.e., SNPs that would have ordinarily been ignored drove the distinction they observed between cases and controls. Subsequent modeling demonstrated that causal SNPs of ~1.05 genotypic relative risk accounted for upwards of 1/3 of the variance in schizophrenia diagnosis (a psychiatric disorder that has been plagued by similar replication failures to bipolar).[1] Since Purcell et. al. published their work, the BiGS consortium ran a similar analysis on their data and found comparable results in their BD sample.[3] In 2011, Lee et al. confirmed Purcell et. al.'s modeling results by using the Genome-wide Complex Trait Analysis (GCTA) technique to demonstrate that ~38% of the variance in the bipolar phenotype is tagged by common SNPs.[2]

1.D In Summary: It seems logical to question our most fundamental assumption when one has problems replicating SNPs in a disorder where common variants are thought to explain more than a 1/3 of the disease variance. In what follows, we will systematically assess these three assumptions:

- Assumption 1: Genotyped and imputed data are equivalent, i.e., they can be analyzed together
- Assumption 2: Logistic regression is an appropriate tool for analyzing mixed genotyped / imputed data
- Assumption 3: We are defining phenotype correctly when we lump all those with bipolar (BDI, BDII, SABD) and compare them with controls
1.E Acknowledgments: Chapter 1, in part is currently being prepared for submission for publication of the material. McGrouther, CC; Kelsoe, JR; Schork, NJ; and Rangan, AV. The dissertation author was the primary investigator and author of this paper.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Markers</th>
<th>Cases/Controls</th>
<th>Loci of Interest</th>
<th>General Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTCCC[29]</td>
<td>~500,000 (not imputed)</td>
<td>1868/3000</td>
<td>16p12 (p=10^{-8} including PALB2, NDUFAB1, DCTN5); 2q31; 13q21, noted CACNA1c</td>
<td>14 SNPs p&lt;0.00005</td>
</tr>
<tr>
<td>Baum et al.[30]</td>
<td>550,235 (pooling) (not imputed)</td>
<td>461/563</td>
<td>DGHK, SORC2</td>
<td>1872 SNPs p&lt; 0.05</td>
</tr>
<tr>
<td>Sklar et al.[31]</td>
<td>372,193 (not-imputed)</td>
<td>1461/2008</td>
<td>MYOB5, TSPAN8, EGFR, noted CACNA1c</td>
<td>1 locus p&lt;5.0 x 10^{-8}</td>
</tr>
<tr>
<td>Baum et al.[32]</td>
<td>76 SNPs (Baum genotyped), tested WTCCC (imputed; unclear method)) - 27 SNPs: (Baum genotyped), repl in German, NIMH, WTCCC</td>
<td>~2300/~3500</td>
<td>JAM3, SLC39A3, BRE</td>
<td>- 76 SNPs showed enrichment - 14 of 27 SNPs signif. In 2 of 3 studies</td>
</tr>
<tr>
<td>Feirrera et al.[33]</td>
<td>331,786 SNPs</td>
<td>1,098/1,267</td>
<td>14 chrom. Regions, (1 spanning CACNA1c)</td>
<td>Lowest p value &lt;0.0008 was also identified in WTCCC + Sklar</td>
</tr>
<tr>
<td></td>
<td>3 study meta-analysis (Sklar+WTCCC+Feirrera)</td>
<td>4,387/6,209</td>
<td>Chromosome 10, 12, 15 (ANK3, CACNA1c)</td>
<td>- 39 SNPs (p&lt;10^{-5}) located in 3 regions - 18 regions with &gt;1SNP (p&lt;10^{-5})</td>
</tr>
<tr>
<td></td>
<td>325,690 SNPs; 1.8 million variants imputed with PLINK8, 60 CEU</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 1.1 - Summary of Results from Recent BD GWAS initiatives (continued)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Markers</th>
<th>Cases/Controls</th>
<th>Loci of Interest</th>
<th>General Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scott et al.[34]</td>
<td>2 study meta-analysis (NIMH/Pritzker Repository + GSK) (\sim 2.3 \times 10^6) SNPs, imputed with Mach; HapMap CEU (some overlap of subjects with BiGS study)</td>
<td>2,076/1,676</td>
<td></td>
<td>- No SNPs of genome-wide significance</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Best (2.4 \times 10^{-6})</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Some validation for ANK3 (rs10994336, (p=0.042)) in indep. GSK sample</td>
</tr>
<tr>
<td></td>
<td>3 study meta-analysis (NIMH/Pritzker Repository+ GSK + WTCCC) (\sim 2.4 \times 10^6) SNPs, imputed with Mach; HapMap CEU (some overlap of subjects with BiGS study)</td>
<td>3,683/14,507</td>
<td>1p31.1 (no known genes), 3p21 (&gt;25 known genes, including ITIH1(strongest), GNL3,NEK4,ITIH3), and 5q15 (MCTP1)</td>
<td>- No SNPs of genome-wide significance</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Best (\sim 10^{-7}) (3 regions)</td>
</tr>
<tr>
<td>Smith et al.[3] (BiGS)</td>
<td>GWAS (GAIN + TGEN): 728,187 SNPs (further imputed with Mach; HapMap2 CEU, # unspecified)</td>
<td>GAIN: 1001/1033 TGEN:1190/401</td>
<td>rs2367911 in region near the voltage-dependent calcium channel gene CACNA2D1 ((p=5.9 \times 10^{-6}))</td>
<td>- No SNPs of genome-wide significance</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- Two SNPs in ANK3 were supported at (p=0.02)</td>
</tr>
<tr>
<td>Cichon et al[35]</td>
<td>GWAS of one study: 499,494 auto / 12,484 X-chromosome SNPs (further imputed with mach, 60 HapMap CEU)</td>
<td>682 / 1300 2411 / 3613 6030 / 31749*</td>
<td>Repl I: 8 of 48 were nominally significant, including NCAN and MAD1L1, but NCAN achieved Genome-wide significance for waves 1 + 2 ((p=3.02 \times 10^{-8})) Repl II: <strong>NCAN replicated</strong> ((p=2.74 \times 10^{-4}))</td>
<td>GWAS top SNP as well as 39 other top SNPs were not replicated</td>
</tr>
<tr>
<td></td>
<td>Repl I (6 study): 48 SNPs Repl II (7 studies): 1 SNP</td>
<td></td>
<td></td>
<td>- Repl II data used in this study included WTCCC, BiGS which had not previously identified NCAN</td>
</tr>
</tbody>
</table>
Table 1.1 - Summary of Results from Recent BD GWfAS initiatives (continued)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Markers</th>
<th>Cases/Controls</th>
<th>Loci of Interest</th>
<th>General Findings</th>
</tr>
</thead>
</table>
| Sklar et al[31] (PGC-BD)      | GWAS: 2,415,422 autosomal SNPs (imputed with BEAGLE; HapMap2 CEU sample; data set based on 11 distinct studies) | 7481 / 9250 4493 / 42542*          | 18 of 34 had p-value < 0.05, including CACNA1C, ODZ4                            | - Two top hits (ANK3, SYNE1) were amongst those that were not replicated  
  - CACNA1C & ODZ4 achieved genome-wide significance in combined sample (p=1.52x10^-8, p=4.40x10^-8)  
  - Subsequently, ANK3 has been nominally replicated in an East Asian sample (p=0.048, N=2,212/2,244)[26]  
  - SYNE1 has been replicated in independent sample (p=0.0095, 1,527/1,579)[28] |
Table 1.1 - Summary of Results from Recent BD GWAS initiatives (continued)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Markers</th>
<th>Cases/Controls</th>
<th>Loci of Interest</th>
<th>General Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green et al[27]</td>
<td>3,106 SNPs with $p&lt;10^{-3}$ in Ferreira et al (2008) were genotyped (No imputation)</td>
<td>1218/2913 (completely new samples)</td>
<td>Supported CACNA1C (rs1006737, $p=4.09\times10^{-4}$) and 15q14 (rs2172835, $p=0.043$)</td>
<td>- Did not replicate ANK3 (rs10994336, $p=0.912$), SYNE1 (rs9371601, $p=0.076$), ODZ4 (rs12576775, $p=0.047$)</td>
</tr>
<tr>
<td></td>
<td>- Reduced to 569 quasi-indep SNPs by removing one SNP from each pair of SNPs with $r^2&gt;0.5$ (PGC samples were imputed using Beagle; HapMap2 CEU)</td>
<td>1218+7481/2913+9250 = PGC + new samples</td>
<td>One new region of Genome-wide significance rs3818253 ($p=3.88\times10^{-8}$, TRPC4AP, near GSS &amp; MYH7B)</td>
<td>- Continued support for CACNA1C and ODZ4 ($p=9.78\times10^{-16}, 6.20\times10^{-9}$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- SNP not previously validated in the original replication arm of the PGC-BD achieved genome-wide significance in the combined sample</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>----rs7296288 (chr 12, $p=8.97\times10^{-9}$, between RHEBL1 and DHH)</td>
</tr>
<tr>
<td>Yosifova et al[36]</td>
<td>497,732 SNPs Repl: top 100 SNPs, 84 after removal of high LD ($r^2&gt;0.8$) SNPs (no imputation)</td>
<td>188/376 Repl: 122/328</td>
<td></td>
<td>- No SNPs of genome-wide significance</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- None replicated after Bonferroni correction</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>- top 100 SNPs included prev. noted regions: CCK, GABRA1, GABRA6, CHRNA9, GRIK5</td>
</tr>
</tbody>
</table>
Table 1.1 - Summary of Results from Recent BD GWAS initiatives (continued)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Markers</th>
<th>Cases/Controls</th>
<th>Loci of Interest</th>
<th>General Findings</th>
</tr>
</thead>
</table>
| Bergen et al[37] | 745,006 genotyped SNPs, then imputed using BEAGLE, HapMap3 (Unclear if used all HapMap3 subj or a subset) | 836/2093       | genome poor region of Chr 4                 | - No genome-wide significant results  
- Best SNP (chr4, p=3.22x10^-7)                                                      |
| Djurovic et al[38] | TOP: 620,390 SNPs Repl: Top 1,000 SNPs (no imputation)                 | TOP: 194 / 336 Repl: 435 / 10,258 | DLEU2, GUCY1B2, PKIA, CCL2, CNTNAP5, DPP10, and FBN1 | - Only 35 of the top 1000 SNPs in TOP study achieved p<0.05, not corrected for multiple testing |
| Hattori et al[39]  | 100,000 SNPs Repl: 1,526 SNPs (p<0.01 in phase 1) (no imputation)       | 107/107 Repl: 395/409 | PLXNA2; SNAP25                               | - 89 of 1526 had p<0.05, none survived Bonferroni correction                      |
| Lee et al[25]     | 516,919 SNPs Repl: 15 SNPs p<10^-5, 5 survived cross-validation. These were genotyped along with the neighboring tag SNPs | 1000/1000 Repl: 409/1000 | SP8,ST8SIA2, BTF3L1/ KCTD12, CACNB2 performed best in combined analysis | - Nothing of genome-wide significance  
- Nothing replicated  
- Some support for two prev SNPs rs1938526 (ANK3, p=6.55x10^-5) and SNP nearby rs11720452 (chr3,p=1.48x10^-5) |
CHAPTER 2:

Data Reliability

**Introduction:** Of the fifteen GWAS in Table 1.1, all of which had little to no success at SNP replication, all of the meta-meta GWAS relied upon imputation to pool samples. Recent work has begun to document the failings of these imputation-dependent GWAS meta-analyses. Before we discuss this literature in detail, it is important that we have a strong understanding of imputation in general.

**2.A Imputation background and significance:**

2.A.1 Imputation Overview: Imputation is an analytic method that allows researchers to estimate untyped or missing variants using known data.[40, 41] This prediction is accomplished by leveraging patterns of linkage disequilibrium and a reference panel of subjects which has been more exhaustively genotyped than the sample one seeks to impute.[42-46]

Imputing genotyped data in this way can increase the power of studies to detect causal variants.[41, 44] This increase comes from two arenas: 1) more variants can be analyzed, i.e., the genome can be analyzed more finely, and 2) subjects from different genotype platforms can be combined for analysis.[41, 47-51] Fine-mapping is important because rare variants are increasingly suspected to play a causative role in complex diseases, e.g., psychiatric illness,[52-57] and the original genotyping platforms were designed to capture common variants.[58, 59] Merging data from different platforms is particularly important as the cost of regenotyping subjects on the same array platform can be prohibitive [54, 60], and it not yet feasible for small or medium sized laboratories to generate and analyze whole-genome sequences of thousands of individuals.[61, 62]
Uncertainty is inherent to imputed SNPS and we will discuss this in subsequent sections, but the increase in power and precision gained through imputation has still resulted in the identification of replicable and novel signals.[41, 60, 63, 64] Basically, imputation 1) boosts power, 2) allows fine-mapping, 3) makes possible meta-analysis, and 4) allows researchers to fill the occasional missing data for a given SNP that was genotyped.[40]

2.A.2 Available Imputation programs: There are 4-5 main programs in use, i.e. BEAGLE[65], MaCH+minimac[65], fastPHASE[66], IMPUTE2[67, 68], and PLINK[69]. PLINK and BEAGLE use small subsets of neighboring markers to estimate the missing genotype whereas MaCH+minimac, fastPHASE, and IMPUTE2 take into account all available markers when imputing.[62] Because they base their estimation on fewer SNPs, PLINK and BEAGLE are more computationally efficient, but potentially less accurate. Though two studies concluded that BEAGLE, MaCH, and IMPUTE performed similarly[40, 70], there were two others that found IMPUTE and Mach outperformed BEAGLE.[71, 72] Nothnagel et al. also observed that PLINK was consistently outperformed by BEAGLE, MaCH, and IMPUTE.[70]

2.A.3 Introduction of two-step imputation: In order to increase their computational efficiency, IMPUTE2 and MaCH+minimac have made it possible for the user to pre-phase their subjects’ GWAS data, i.e., estimate their haplotypes prior to imputation. Pre-phasing the subject data is computationally expensive, but only needs to be performed once per subject. If more subjects are added or if the reference panel grows, a researcher need not rephase older subjects. Since pre-phasing has already been accomplished and this makes it easier to match up subject data with the already phased reference subjects, the actual imputation is less time-consuming.[62] These two-step imputations have been observed to be both accurate and powerful methods for imputing common and rare variants in large studies using the latest reference panels.[62, 67]
2.A.4 Imputation Output: Independent of the method one selects, there are two crucial outputs per SNP upon which researchers rely: 1) a posterior probability vector and 2) the imputation quality score.

- **Posterior probability vector**: The posterior probability vector is three dimensional. Each entry reflects the probability that a particular SNP loci has a particular allele combination [aa,aA,AA] given the known genotyped information, i.e., a number between 0 and 1 is assigned to each of the values [0=aa,1=aA,2=AA] such that these three probabilities add up to 1.

- **Imputation Quality Score (IQS)**: The IQS is a number between 0 and 1 where 0 represents complete uncertainty of the genotype and 1 represents complete certainty. MaCH, BEAGLE, and IMPUTE all use slightly different means of assigning the IQS, but they are highly correlated with MaCH and IMPUTE being the most highly correlated. An unofficial guideline for considering the quality score is: an IQS equal to alpha on a sample of N individuals indicates that the amount of data at the imputed SNP is approximately equivalent to a set of perfectly observed genotype data in a sample size of alpha*N". [67] Researchers rely on the IQS to determine what SNPs to analyze, i.e., they analyze only those SNPs whose IQS is above a certain threshold. The reason for this is that it has been shown that imputation accuracy directly affects downstream analysis [64, 73] Selecting this threshold varies somewhat from paper to paper, but there seems to be a consensus not to analyze data from SNPs with imputation quality scores < 0.3.[48, 62] It is worth noting that there is a documented case where the use of thresholds resulted in the failure to detect APOE in Late-Onset Alzheimer's Disease GWAS[74], leading another group to suggest not using them at all and letting replication sort out true from false positives[75].
The most commonly used IQS are functions of allele frequency.[40] It will be discussed below in greater detail (Section C.1.f.1), but those with smaller MAF are much more difficult to impute. It has been suggested that cutoffs should not be predetermined and fixed, but rather selected based upon minor allele frequency (MAF), genotyped categories (aa, aA, AA), and an assessment of how your own data performs by the use of masking within these categories.[71]

2.A.5 Reference Panels Overview: As noted in the previous section, accuracy of imputation is very important. In addition to minor allele frequency, there are other influencing factors: 1) haplotypic diversity of the population being studied, 2) reference panel size, and 3) genetic similarity of the reference and population being analyzed is very important.[63, 65, 76-79] Reference panels play such a crucial role in the imputation process because they define the haplotype landscape which the various imputation algorithms use to fill the holes in the genotyped data.

There are two main sources of data: 1) The International HapMap Project, and 2) the 1000 Genomes Project. Each new reference panel release has aimed to increase the size/diversity of the reference panel and the coverage of SNPs with MAF<5%. Through the use of HapMap’s Phases 1 to 3 (Phase 1+2: 270 samples from 4 populations, directly genotyped at 3.8 million SNPs, Phase 3: 1,092 samples from 11 populations, directly genotyped at 1.6 million SNPs), a researcher can impute approximately 2.5 million SNPs in a GWAS.[41, 77, 78, 80, 81] Instead of genotyping on an array, the 1000 Genomes Project (1000G) has performed whole-genome sequencing of 1,092 individuals (181 (Admixed American), 246 (African), 286 (East Asian), and 379 (European ancestry)). 1000G covers validated haplotype map of 38 million SNPs, 1.4 million short insertions and deletions, and 14,000+ larger deletions.[62, 82] Their sequencing based approach combined with the large number of samples means that
there are many low frequency (1% < MAF < 5%) and rare (MAF<1%) variants in their data set.[41]

2.A.6 How single-study imputation data is analyzed: As noted in the imputation output section, imputation yields 1) a three-dimensional posterior probability vector and 2) an imputation quality score for each SNP. Assuming the imputed SNP data is of sufficient quality (IQS > threshold – discussed above), researchers have three main routes to choose for analyzing this data: 1) they can threshold the posterior probability data to assign the subject the allele combination that is most likely, 2) they can analyze the three posterior probabilities directly, or 3) they can analyze a one-dimensional representation of the two-dimensional posterior probability vector.

Route 1 was popular initially, but fell out of favor when it became clear that it can result in the loss of too much information when a discrete, possibly incorrect, genotype is assigned. Further, analyzing data that incorporates the imputation uncertainty has been shown to have more power than analyzing incorrect calls.[62-64, 83] Route 2, i.e., analyzing the posterior probabilities directly, is computationally expensive and, in simulations, improves power negligibly over Route 3, the direct analysis of the weighted mean (dosage) of the three posterior probabilities.[64, 84] Since Route 3 is much less-computationally expensive than Route 2 without a corresponding loss of power, it is most commonly used. Subsequently, Liu et al. were able to show theoretically and through simulations that there is no-better one-dimensional representation of the three-dimensional posterior-probability vector than the dosage for analysis by linear and non-linear models.[84] Therefore, analyzing the one-dimensional dosage data (Route 3), while controlling for relevant Multi-Dimensional Scaling (MDS) or principal components, is increasingly used. Some researchers will go back after identifying SNPs by Route 3 and analyze directly the posterior probability vector, Route 2.
2.A.7 Missing Heritability of Complex-Disease: The candidate loci identified by GWAS have yet to explain a large portion of the heritability of complex disease.[52-54, 57, 59, 85, 86] Possible explanations for this include 1) heritability is over-estimated[59, 87, 88] 2) epistasis[89], 3) the cumulative contribution of non-genome-wide significant variants[57, 59, 90], 4) attempts to minimize false positives resulting in a large number of false negatives[59, 91, 92], and 5) the importance of uncommon (1–5% minor allele frequency [MAF]) and rare(< 1% MAF) SNPs which are not adequately captured by genotyping arrays is under-estimated[55, 56, 59].

2.A.7.1 Imputation of uncommon (MAF<5%) and rare alleles (MAF<1%): Uncommon alleles is a possible explanation for why GWAS have yet to explain the heritability of complex diseases.[52-57, 59, 85, 86] Certainly, the empirical evidence supporting the role of rare variants in explaining the missing heritability is increasing.[93] Since most genotyping platforms were designed to target common variants (MAF>5%), imputation makes it possible to identify some of these uncommon variants without further costly genotyping.[55, 56, 59, 94]

Unfortunately, SNPs with low MAF (MAF< 5%) are difficult to impute[71, 94-96] with those rare variants (MAF < 1%) being the most difficult.[70, 94, 97, 98] Rare variants prove particularly elusive because of poorer coverage by GWAS SNPs, lower degree of LD, and more challenging haplotype reconstruction.[99]

Increasing the array density of the genotyping platform improves imputation quality with the greatest corresponding increase in accuracy being seen in uncommon (1%<MAF<5%) alleles.[94, 100] Increasing the number of subjects does an even better job than increasing array density of improving accuracy of uncommon (1%<MAF<5%) alleles.[59, 94, 96] Unfortunately, neither increasing density nor sample sized improved the accuracy of rare (MAF<1%) alleles.
Using the 1000G reference panel also improves our ability to detect uncommon (1%<MAF<5%) variants, i.e. increases power, because its sequencing-based approach captures many more uncommon and rare variants than HapMap’s array-based approach.[42, 62, 101, 102] In a simulation of the 1000G sample size, Li et al. observed an average IQS increase from 74% to 93% for SNPs with MAF < 5% when they enlarged the reference panel from 60 haplotypes to 1,000 haplotypes.[99, 103] When Germaine et al. reanalyzed a venous thrombosis dataset using 1000G reference panel, they were able identify a rare variant that could be validated.[104] Similarly, Holm et al. found a rare variant (MAF = 0.38%) associated with a 50% lifetime risk of sick sinus syndrome compared with 6% risk for non-carriers.[105] That being said, Nelson et al. found power remained low for rare (MAF<1%) variants even with the densest available array data, very large sample sizes, and 1000G reference panel. Basically, 1000G allows for the imputation of many uncommon and rare variants, but that many of these imputations are still of very low quality.[94]

If the ancestry of the study population is not adequately represented in the reference panel, then the imputation accuracy for uncommon SNPs will be reduced and can confound study results.[60] As a result, another means by which to improve the imputation quality of uncommon and, notably, rare variants is to sequence a subset of the study sample and use it as part of the imputation reference panel.[60, 106, 107] If one uses the most genetically diverse internal group to supplement the reference panel, then the accuracy improvement is even more dramatic, especially for rare variants.[108] Including samples from an ethnically distinct reference population is another means of increasing imputation accuracy. Huang et al. and Jostins et al. both found that incorporating a mixture of Hapmap samples improved imputation accuracy, with Jostins et al. showing the improvement most in uncommon and rare SNPs.[77, 79]
2.A.7.2 False positives: False positives are another possible explanation for why GWAS have yet to explain a large degree of the heritability of complex diseases.[59, 91, 92] One problem is that it is difficult to distinguish causal SNPs from incidentally-correlated SNPs in traditional association studies.[109] Faye et al. were able to demonstrate through simulations that tag SNPs\textsuperscript{2} and SNPs in higher LD with the tag SNP than the causal variant are more likely to be identified by traditional association analyses than causal SNPs. This idea that the association of a tag SNP will be magnified when it is selected because of its small p-value is known as the winner’s curse.[109] Ioannidis, Thomas, and Daly’s 2009 review article described the role of the winner’s curse as follows: “The magnitude of the winner’s curse is inversely related to the power of the study. In typical circumstances, for 10% power, the inflation of an additive effect could be approximately 60%; however, for 60% power, the inflation would be only 10%. For small effects, even large meta-analyses could be grossly under-powered and emerging associations could be considerably inflated. For rare variants, the power can be <1%, and therefore associations that are discovered for rare variants will have extremely inflated effects and the true effect size should await further replication.”[110]

Depending upon allele frequency and effect size, Udler et al. estimated that 1-4x the original sample size would be required to distinguish between the causal SNP and its correlated variants.[111] Faye et al made a similar estimate in their analyses.[109] As this amount of resequencing / recruitment is often prohibitive, multiple methods have been put forth for distinguishing causal from correlated SNPs after the initial association analysis. Faye et al. offer up a re-ranking system for the region around significant SNPs leveraging imputation accuracy, differential sequencing, and linkage disequilibrium.[109] Zaitlen et al had success genotyping significant SNPs in different subject populations

\textsuperscript{2}Tag SNP is a SNP in high Linkage Disequilibrium (LD) with group of SNPs or haplotype
and leveraging their distinct LD pattern to distinguish causal from correlated variants.[112] Another study had a similar level of success by ranking SNPs in multiple subject populations and focusing their replication studies on genes/SNPs common to multiple populations.[113] Others have had some success placing constraints on the regression coefficients of multivariate regression of SNPs in a given region.[114] There is no consensus yet on the best approach as minimizing/eliminating false positives is an active field of research.

2.A.7.3 False negatives: Another possible explanation for the missing heritability is false negatives. Strict Imputation Quality Score (IQS) thresholds and Bonferonni correction / comparable methods for accounting for multiple comparisons result in a great number of SNPs/genes never being considered.[59, 74, 91, 92] Further, polygenic models have demonstrated that a large portion of the disease-variance in complex diseases can be explained by non-significant/false negative SNPs.[1, 2]

2.A.8 Multiple Dataset Imputation Analysis

2.A.8.1 Methods: As noted above, imputation makes it possible to merge genomic data from various platforms. As was true for single-platform imputations (Section C.1.f), analyzing the one-dimensional dosage data is the preferred method. There are two-basic approaches used for this type of analysis: 1) meta-analysis of site-specific test statistics, e.g., p-values, z-scores, or effect sizes (beta), and 2) en masse analysis of the dosage data while controlling for study of origin and the relevant Multi-Dimensional Scaling (MDS) or Principal Components.

2.A.8.2 Problems with Multiple Dataset Imputation or imputation-reliant meta-analyses: Depending on the site-specific test statistics available (p-values, effect sizes, or z-scores), there is a meta-analysis method available that adjusts for imputation uncertainty.[115-117] These methods are important/were developed because they can
recover some of the power lost when imputation accuracy is not accounted for.[115-117]

These methods also improve study power by making it possible to analyze SNPs of lower imputation quality which would have originally been removed during quality control (QC) in a traditional meta-analyses. For instance, the APOE signal in Late-Onset Alzheimer’s Disease (LOAD) is in a region of low LD. When performing a traditional two-study meta-analysis, Beecham et al. did not detect the APOE signal because the relevant SNPs were eliminated during QC because of poor imputation quality. When they switched to an imputation quality-aware meta-analysis, they were able to recover some of this LOAD APOE signal.[74]

These ‘improvement in power’ methods papers demonstrated their effect by focusing on the analysis of SNPs that were wholly imputed. To the best of our knowledge, only one paper has systematically compared wholly genotyped, wholly imputed, and mixed genotyped/imputed data, i.e., SNPs which were directly genotyped in some studies and imputed in the rest. In this paper, Li et al. used de Bakker et al.’s imputation-accuracy aware meta-analysis of Z-scores to analyze mixed genotyped/imputed data from a 3-study/platform GWAS meta-analysis with a simulated phenotype. Li et al. used IMPUTE (ver 0.50) and the 60 CEU HapMap2 reference samples. Despite using an imputation-aware analytic method, Li et al. still found the following: 1) power: genotyped > imputed > mixed genotyped/imputed, 2) heterogeneity: mixed genotyped/imputed > imputed uncommon (<5%) MAF > imputed common MAF is approx. = genotyped, and 3) type I error: inflated in mixed genotyped/imputed data when relied on ‘fixed-effects’ rather than ‘random-effects’ model. Li et al. noted that the power of such a meta-analysis can actually be below that of an individual study in the meta-analysis even when performing an imputation-quality aware z-score analysis.[118]
Sinott et al., Uh et al., and Johnson et al. studied the consequences of cases and controls being genotyped on different arrays prior to imputation.[75, 119, 120] Each found a dramatic increase in false positives. Sinott et al and Uh et al. needed very strict IQS thresholds (IQS>=~0.98) to eliminate the false positives. Sinott et al. went so far as to advocate genotyping every significant SNP identified under these circumstances on another platform to make sure that it is not simply an artifact.[120] Johnson et al. found they were able to eliminate the false positives by determining the intersection SNP set of the two platforms and using this set as the basis of imputation.[75] In each of these papers, SNPs with low LD and small MAF, i.e., SNPs with the least available information for imputation to base its prediction, were the most likely to manifest false positives.

In all instances, these problems with power, heterogeneity, and false positives were exacerbated for SNPs with smaller MAF and lower LD, i.e., for which genetic information to base imputation prediction is less.

2.B Rationale: There are many reasons why it important that we question the assumption that genotyped data and imputed data can be used interchangeably. 1) An imputation of even a single study has its problems, i.e., uncommon (MAF<5%) and rare alleles (MAF<1%), false positives, and false negatives. At its most fundamental level, imputed data will always require monitoring some sort of imputation-quality metric. 2) Imputed data as a category becomes even more problematic when one combines data from multiple genotyping platforms. Li et al. documented reduced power, increased heterogeneity, and increased false positives when one imputes and combines data from three studies genotyped on different platforms.[118] 3) If three studies from different platforms is problematic imagine what happens in meta-analyses with more than 3 studies. In Sklar et al. 2011’s paper, the PGC-BD’s wave 1 data was drawn from 11 different studies genotyped on four different platforms (see Table 2.1A), and their
replication, or wave 2, dataset was drawn from 7 different studies genotyped on an unspecified, though likely numerous, number of platforms. 4) The Sklar et al. paper is by no means atypical in BD. There are six other published imputation-dependent BD GWAS meta-analyses of studies genotyped on different platforms. (See Table AAA) 5) Imputation-reliant meta-analyses are common practice in all psychiatric disorders. Two examples being: a) the PGC-SZ (Schizophrenia) group latest meta-analysis was of 54 studies genotyped on 13 platforms[121], and b) the PGC-MDD (Major Depressive Disorder) meta-analysis was of 8 studies genotyped on 8 platforms.[122] 6) Similar practices are employed in meta-analyses in non-psychiatric illness.[54] 7) This imputed data, with all of its potential problems, is highly represented in the top SNPs selected in the Sklar et al. 2011 paper for replication. Only 6 of the 38 SNPs the PGC-BD sought to replicate had > 2/3 genotyped data. Certainly, the two SNPs that achieved genome-wide significance were 100% imputed (rs4765913 (CACNA1C)) and 66% imputed (rs12576775 (ODZ4)). 8) Top SNPs being highly imputed is unlikely to be isolated to this one paper since most genotyping platforms only have between 300,000-1,000,000 SNPs and the datasets generated by imputing with hapmap2/3 or 1000G are ~2.5 million and ~38 million SNPs respectively. PGC's Cross Disorder Groups top SNPs would appear to support this. Within the Smoller et al. 2013 Cross-Disorder Group meta-analysis results, 3 of the top 4 CDG SNPs are more than 2/3rds imputed within the BPD sample. Specifically, the top SNP rs2535629 (p= 2.54×10⁻¹², ITIH3+many) is 68% imputed in the BD sample, the second best SNP rs11191454(p=1.39×10⁻⁸, AS3MT+many) is 30% imputed in the BD sample, the 3rd best SNP rs1024582 (1.87×10⁻⁸, CACNA1C) is 100% imputed in the BD sample, rs2799573 (p= 4.29×10⁻⁸, CACNB2) is 67% imputed in the BD sample. 9) Replication continues to be elusive in BD and we have been seeking to replicate SNPs with a great deal of imputed data. 10) Supposing there is a difference
between genotyped and imputed data, then some thought would need to given to this when determining which SNPs to replicate (rather just LD and p-values) and what type of data to replicate them with.

In summary, BD/psychiatric GWAS/non-psychiatric meta-analyses (i) depend upon imputation to combine data genotyped on a variety of platforms, (ii) treat genotyped and imputed data as interchangeable, (iii) often seek to replicate partially/wholly imputed SNPs, and (iv) are prone to replication failures. Li et al.’s work is the only paper we are aware of which documents what can happen when data genotyped on different platforms is merged. What they found merging data from three studies is concerning. To make their point, Li et al. relied on a simulated phenotype and, by construction, somewhat artificial circumstances. Most psychiatric meta-analyses involve many more studies and platforms with PGC-SZ’s 54 studies on 13 platforms being an extreme example.[121] We seek to quantify the differences, if any, between genotyped and imputed data within a real world example by analyzing the data in Sklar et al.’s 2011 PGC-BD paper. Specifically, we will determine

(1) If case-control differences in imputed data are more spuriously correlated than those in genotyped data on a per SNP basis, and
(2) Whether top imputed SNPs are good predictors of top genotyped SNPs and vice versa.

The motivation behind point (2) is fairly obvious. In imputation-reliant GWAS, researchers assume that imputed data predicts genotyped data. They give imputed and genotyped SNPs equal weight when determining which SNPs to replicate. If this assumption is flawed, it would be important to realize so that researchers can adjust their efforts accordingly.
Point (1)’s motivation is slightly more subtle. As we noted above, increased reference panel diversity results in higher quality imputations. This makes a great deal of sense since imputation can only fill holes in the genome based on the patterns present in the reference panel. If the subject’s haplotype is not represented in the reference panel, how can we expect imputation to guess correctly? There were 16,731 subjects in Sklar et al.’s PGC-BD sample. They were imputed using only the 120 reference haplotypes in the CEU (Utah residents with ancestry from northern and western Europe) sample. Given that so many subjects were effectively forced into so few holes, we hypothesized that this might result in SNP case-control differences being more correlated.

In point (1), we intend to consider the correlation between case-control differences under two conditions: (a) covariate-corrected and (b) covariate-uncorrected. As justification for (a): it is common for GWAS meta-analyses, even those that do not need to rely on imputation, to use study of origin and MDS components as covariates in their logistic regressions. As justification for (b): We imagine that much of the signal introduced by imputation is driven by study (imputation is performed on a per study basis) and MDS components (a proxy of genetic similarity between subjects previously demonstrated to reflect geography[123]). If we were to only consider the covariate-corrected relationships, likely some fraction of the non-disease-related signal introduced through imputation would already have been removed prior to our analysis. This would be fine if covariate-correction was perfect and if people never attempted to analyze the mixed imputed / genotyped data with methods that cannot be covariate corrected. Unfortunately, covariate-correction is far from perfect, e.g., logistic regression can only correct for linear relationships, and researchers do seek to analyze the imputed data in ways that are not covariate corrected, e.g., data-mining.[124]
2.C Research Design and Methods

2.C.1 Dataset Description: Through the PGC-BD, in conjunction with dbGAP, NIMH, and The Wellcome Trust, we had access to 5989 BD genotyped cases and 7998 genotyped controls.[5] A complete description of the genotyped dataset is in Sklar et al.[3] As noted above, the cases and controls were drawn from 11 different sources and were genotyped on 4 different platforms (The per study details and their genotyping platforms are listed in Table 2.1A) Case diagnoses were as follows: BDI (bipolar disorder, type 1) - 84%, BD2 (bipolar disorder, type 2) - 11%, SABP (schizoaffective bipolar disorder) - 4%, and other bipolar diagnosis - 1%. There were only 46,234 SNPs genotyped in all 11 studies. There were 1,016,924 SNPs genotyped within at least two groups.

We have access to one imputation of the genotyped data used in Sklar et al.’s 2011 paper. Specifically, HapMap2 - Imputation was performed using BEAGLE[125], leveraging the HapMap phase 2 CEU reference sample (N=60, 120 haplotypes); the resulting dataset contained 2,415,422 autosomal SNPs with a minor allele frequency ≥1% and imputation quality score IQS > 0.3

2.D Methods

2.D.1 Correlation between case-control differences in imputed vs. genotyped data: To achieve our results we: 1) split the available PGC-BD data into two approximately equal groups, 2) Create two linkage equilibrium datasets from the genotyped and mixed imputed / genotyped data, 3) calculate the case-control differences for each SNP individually, both uncorrected and corrected for study and MDS components, and 4) determine how correlated the results are across the two groups.

2.D.1.a Creating two waves from the PGC-BD Datasets: We set out to divide the 5989 BD genotyped cases and 7998 genotyped controls in a manner that preserves study
lines, has a reasonable balance of case and controls, and guarantees that each of the four genotyping platforms were represented in both waves. (See Table 2.2) Since there was only one study genotyped on the Illumina HumanHap 550, we couldn’t achieve all of these goals.

2.D.1.b Datasets for Analysis: Our goal was to generate two linkage equilibrium datasets: genotyped and mixed imputed/genotyped. We will refer to Set G (for genotyped) consisting of the N=46,234 SNPs that are 100% genotyped, i.e., genotyped in all 11 studies. We will refer to Set I (for imputed) as the entire ~2.4 million mixed imputed/genotyped SNPs PGC-BD HapMap2 Imputation. To create a LE version of set G: We analyzed the WTCCC control samples (N=2,931) with PLINK’s ‘indep-pairwise’ with the following inputs: window size: 50, shift window by 5 each time, and LD threshold of 0.05. This left us with a set of 21,088 SNPs (Set G_LE). To create a LE version of set I: We analyzed HapMap3’s CEU (N=180, Utah residents with Northern and Western European ancestry from the CEPH collection) and TSI (N=100, Tuscans in Italy) subjects with PLINK’s ‘—indep-pairwise 50 5 0.5’. This set contained 299,079 SNPs with varying ratios of genotyped/imputed data. After thresholding for the imputation quality score (IQS> 0.3) and 0.01<MAF< 0.99, we were left with 282,551 SNPs in Set I_LE.

2.D.1.c Calculating per SNP case-control differences: We used two different methods to calculate the case-control differences for the SNPs in Set G_LE and Set I_LE: 1) Logistic regression (LogR) and 2) the Area Under the receiver-operating Characteristic (AUC). We selected logistic regression because using it to analyze the 1-dimensional dosage data is the means by which PGC-BD and many others search for case-control differences on a per-SNP basis. If you recall, dosage data is the weighted-mean of the 3-Dimensional posterior probabilities for [aa,aA,AA]. In this case, dosage=2*P(aa) +
1*P(aA). We also decided to use the AUC because it has features which may make it a better tool than logistic regression for analyzing mixed imputed / genotyped data.

In a HapMap2/3-reliant imputation, Taub et al. observed that the imputed SNP distributions were skewed, i.e., lop-sided.[63] We observe the same in the PGC-BD HapMap2 data. In Figure 2.1, we bin the dosage data from 4 SNPs with different amounts of genotyped data (A - 10% genotyped, B - 34% genotyped, C – 51% genotyped, D – 86% genotyped). There are two graphs per SNP. The top graph is a histogram of the dosage data. The bottom graph also a histogram, but with the amount per bin expressed on a log scale. We made use of the log scale so that outliers and sparsely represented dosage would be visible. Clearly, not only is the data skewed, but it also has outliers; these features are typical for many of the SNPs in this study. We draw the reader’s attention to this because logistic regression is 1) not well-suited to the analysis of skewed data, 2) is sensitive to outliers, and 3) its trial-to-trial variance can be unbounded. On the other hand, AUC is 1) not adversely affected by skewness, 2) robust to outliers, and 3) has a bounded trial-to-trial variance. (For detailed explanation of this see Appendix, specifically Part B, Sections 3-5). In Figure 2.2, as further empirical evidence for using AUC, we demonstrate AUC’s ability to identify a more consistent signal than logistic regression in the presence of skewness and outliers. We took two SNPs of mixed imputed/genotyped data, 56% and 83% genotyped data respectively. In frame 1, we plotted the log histogram of the data just as we did in Figure 2.1. We then calculated the AUC and odds ratio (OR) for every possible combination of the 11 studies in the PGC-BD data. In Frame 2, we plotted the log(OR) for the entire dataset in the first graph and the log(OR) for each of the other 2^11 study combinations on the 2nd graph. In Frame 3, we plotted the histogram of the results for the 2^11 log(AUC/(1-AUC))’s and log(OR). In both instances outliers / skewness influence the conclusion drawn by logistic
regression. In the first instance, AUC sees a signal where LogR cannot, and, in the second, AUC does not find a spurious signal that logistic regression erroneously identifies because of the impact of a few outliers.

Despite the empirical evidence and positive theoretical properties, using AUC to analyze data with covariates was not previously an option because, prior to our work, there was no practical means by which to correct AUC for covariates. Now, not only are we able to correct AUC for linear relationships between covariates, but also for non-linear relationships (Appendix Part C, Sections 6-11, particularly 9). The quick uncorrected AUC code is in C.12.4. and the corrected AUC code is in C.12.6.

For each SNP in Set G_LE and Set I_LE, we calculated the uncorrected odds ratio (OR) / p-value and corrected odds ratio (cOR) / p-value using logistic regression on Groups 1 and 2. In addition, we calculated both uncorrected AUC (AUC) and corrected AUC (cAUC) on Groups 1 and 2 using the code described in Section C.12.4 and C.12.6. We present this analysis below.

2.D.1.d Correlation: Using matlab’s CORR function, we calculated the correlation between group 1 and group 2, i.e., Corr(OR₁, OR₂), Corr(cOR₁, cOR₂), Corr(AUC₁, AUC₂), and Corr(cAUC₁, cAUC₂). Since the mixed imputed/genotyped data was more correlated than the purely genotyped SNPs, we then examined whether imputation was driving this correlation. To do this, we plotted the correlation starting with the 100% imputed SNPs and observed how this correlation changed as we added back in SNPs with increasing fractions of genotyped data until the entire I_LE data set was included. After this, we plotted the correlations again, this time starting from the entire I_LE Set and removing the largely imputed SNPs until we are left with the 100% genotyped SNPs. In hopes of determining the impact of MAF or imputation quality (IQS) on these correlations, we recalculated these graphs for various MAF and IQS thresholds.
2.D.2 Imputed predictive capacity: We took the following steps: a) identified all SNPs with a fraction of genotyped data (N=1,016,924), b) subselected those SNPs with between 40-60% genotyped data so as to adjust for differences in N between the genotyped and imputed analyses, c) performed a logistic regression, correcting for the top 5 MDS components and study affiliation, for the genotyped data and imputed data separately, and d) determined whether there is a strong relationship between top imputed SNPs and top real SNPs. Top SNPs being defined as the most significant 100, 1000, or 10,000 SNPs. We also plotted the histogram of the rank order of the SNPs in hopes of identifying trends in significance.

2.E Results

2.E.1 Correlation between case-control differences in imputed vs. genotyped data:

Figure 2.3 is a 2x2 array of the cross-group correlation measured in G.LE (100% Genotyped linkage equilibrium set) and I.LE (mixed imputed / genotyped linkage equilibrium set) vs. Uncorrected AUC and Corrected AUC. Specifically, spot (1,1) is a scatter plot of AUC$_1$ vs. AUC$_2$ for SNP set G.LE, spot (1,2) is a scatter plot of cAUC$_1$ vs. cAUC$_2$ for SNP set G.LE, spot (2,1) is scatter plot of AUC$_1$ vs. AUC$_2$ for SNP set I.LE, and spot (2,2) is scatter plot of cAUC$_1$ vs. cAUC$_2$ for SNP set I.LE. The uncorrected AUC is correlated across the mixed imputed/genotyped LE set of SNPs (I.LE) at a 0.526 (p=0) level whereas the corrected AUC (cAUC) on I.LE is correlated at 0.031 (p=4.4x10$^{-61}$). A comparable analysis of purely genotyped, i.e., non-imputed, data found correlations of 0.063 (p=5.8x10$^{-20}$) and 0.036 (p=1.5x10$^{-7}$) respectively. A related analysis, this one using logistic regression to calculate the log odds ratio, found the following cross-group correlations: I.LE - 0.123 (uncorrected, p=0) and 0.031 (corrected, p=1.0x10$^{-60}$) and G.LE – 0.056 (uncorrected, p=6.2x10$^{-16}$) and 0.037 (corrected, p=6.8x10$^{-8}$). (processed data shown in Fig 2.4)
We then examined the relationship between this increased correlation and imputation by starting with 100% imputed SNPs and gradually adding in SNPs with increasing fractions of genotyped data. (See Figure 2.4) Purely imputed SNPs were correlated at a 0.37 (p=0) level. This correlation jumped to ~0.71 with the introduction of mixed imputed / genotyped SNPs. As we continue to introduce SNPs with increasing fraction of genotyped data until we had the entire I_LE dataset (282,551 SNPs), the correlation decreases (~0.71->0.526). The 2nd half of Figure 2.4 started with the entire I_LE dataset and gradually removed those SNPs that were purely or partially imputed. The correlation decreases as the fraction of imputed data decreases (0.526->0.069). Though the signal is weaker, logistic regression (log of the odds ratio) demonstrates a similar relationship.(See Figure 2.4)

Further, as shown in Fig. 2.5, we had to increase the IQS threshold to 0.98 before we see a dramatic change in the pattern of these correlations. With this IQS threshold, the corr(AUC$_1$, AUC$_2$) of the purely imputed SNPs almost returns to that of the purely genotyped SNPs (0.077 / 0.060).

SNPs with more extreme MAF frequencies are more highly correlated when purely or partially imputed than SNPs of less extreme frequencies. Nevertheless, purely or partially imputed SNPs with MAF 40-60% still exhibit a much larger correlation than purely genotyped SNPs in the same MAF range. (See Figure 2.6). The trend (shape of the correlation curves) is the same if one limits oneself to SNPs with 0.20<=freq<=0.80, though the max correlation does decrease in magnitude (becomes ~0.40 rather than ~0.71). Similarly, if one limits oneself to SNPs with 0.4<=freq<=.60, the shape remains the same, but the max correlation continues to drop (~0.28 rather than ~0.71).

In general, correlation goes as follows: mixed imputed / genotyped SNPs > purely imputed >> purely genotyped. These relationships are robust to all, but the
strictest IQS thresholding (IQS>0.98). The more extreme a SNP’s MAF, the more correlated, but even with the most strict thresholding, the purely and partially imputed data are more correlated than purely genotyped data.

2.E.2 Imputed predictive capacity: Here we assess whether or not the significance of SNPs calculated using imputed data can be used to predict the significance of those same SNPs when calculated using genotyped data. Specifically, there were \(-10^6\) SNPs with some genotyped data. For each SNP with imputation quality score (IQS) > 0.3, we performed logistic regression, correcting for MDS components and study, for the genotyped and imputed data separately. To control for the effects of varying Ns, we focused on the \(\sim133\)k SNPs with between 40-60% genotyped data. If one considers the top 0.1% \((\sim133)\) of the imputed SNPs (i.e., top SNPs from the imputed data set), only 3 are in the top 0.1% \((\sim133)\) genotyped SNPs, only 10 in top 1% \((\sim1,335)\) of genotyped SNPs, and only 23 in top 10% \((\sim13,353)\) of genotyped SNPs. If one considers the top 0.1% \((\sim133)\) of genotyped SNPs, only 3 are in the top 0.1% \((\sim133)\) of imputed SNPs, only 3 are in top 1% \((\sim1,335)\) of imputed SNPs, only 19 in the top 10% \((\sim13,353)\) of imputed SNPs. Though this is better than random \((0.13/1.3/13)\), it suggests that using imputed data to identify the top signals in genotyped SNPs is not particularly helpful. In our sample, imputed data predicted genotyped data slightly better than genotyped data predicted imputed data. Using stricter imputation quality scores slightly improved both imputed data’s ability to predict genotyped data and vice versa. Limiting ourselves to SNPs with IQS > 0.98 \((N=47,464\) SNPs) yielded 2/4/10 and 2/2/7 respectively (compare with the null hypothesis of 0.048/0.48/4.8).

Figure 2.7 shows the 2-dimensional histogram of genotyped SNP Rank assignment versus imputed SNP Rank assignment. We varied number of bins (10,100)
and IQS threshold (0.3, 0.8, 0.98). The relationship requires 10% bins for any relationship to become apparent.

2.F Discussion:

2.F.1 Conclusions: 1) It is very clear that HapMap2-imputed data does not behave like genotyped data. 2) Mixed imputed / genotyped data is more correlated across groups than purely imputed SNPs. Purely imputed SNPs are much more correlated than purely genotyped data. 3) Using a very strict IQS threshold (0.98) almost returns the cross-correlation observed in purely imputed data to the level observed in purely genotyped data. 4) HapMap2-imputed data is at best a weak indicator of genotyped SNP behavior. 5) HapMap2-imputation has introduced a large non-disease related signal. Logistic regression is largely blind to that signal. AUC is better able to detect this signal than logistic regression because, unlike logistic regression, AUC is a) distribution-invariant, b) robust to outliers and c) its trial to trial variance is bounded. Given the growing body of literature on the problems with imputation, this problem may not be specific to BD and AUC may be the tool researchers need to assess both the non-linear and linear influences introduced by imputation.

2.F.2 Consequences for further PGC-BD analyses: Our results line up nicely with the simulation studies of Li et al., Uh et al., and Sinott et al.[118-120] Given this, going forward: 1) Researchers may want to restrict our future analyses of the PGC-BD data to fully genotyped and/or fully imputed SNPs, because the mixed imputed genotyped data had the most non-disease related signal (as corroborated by the simulation studies of Li et al). 2) Researchers may also want to use a stricter imputation-quality-threshold (IQS > 0.98) in our future analyses of imputed data, since we found that a very strict threshold mitigates the non-disease related signal (again, as corroborated by the literature). 3) Instead of genotyping promising imputed SNPs perhaps we should be collecting the data
necessary to impute them as the success of imputation may be because it is a local measure, and 4) If genotyping only a few locations is the only option, perhaps we should rerun PLINK’s clump algorithm in a manner which forces a greater emphasis on SNPs with more genotyped data, e.g., input into clump a SNP list with more genotyped data.

2.F.3 **Word of Caution:** If a researcher chooses to continue to use weaker IQS thresholds and to include mixed genotyped/imputed and/or purely imputed SNPs, they must remember that up to 93% of the signal present in the dataset disappears with correction for MDS components and study of origin. This has three main consequences: 1) Since this is a rather substantial drop and correction is far from a perfect science, it is unclear whether the answer which remains can trusted the signal which remains, 2) Since LogR is largely blind to this signal, it may not be best suited to correcting away this non-disease related signal, and 3) If a researcher intends to analyze imputation-reliant data directly, e.g., using machine learning, they will need to make use of a method capable of covariate-correction. Otherwise, the non-disease-related signal will swamp that of the disease-related signal.

2.F.4 **Future Work:** It is absolutely essential that this work be repeated for the 1000 Genomes imputation of the PGC-BD. This work simply informs researchers who have imputed their data using HapMap2 that some of their results may be suspect. It does not give them a sense if re-imputing their data should improve the quality of their results. Further, it does not give researchers any sense for how confident they should be when imputing their dataset with the 1000G for the first time. It remains to be seen whether using a larger reference panel, e.g., the 1000 Genomes Project, will eliminate the problems such as those we have identified in the HapMap2-imputed PGC-BD GWAS.
2.G Acknowledgments: Chapter 2, in part is currently being prepared for submission for publication of the material. McGrouther, CC; Kelsoe, JR; Schork, NJ; and Rangan, AV. The dissertation author was the primary investigator and author of this paper.
Table 2.1 - Sklar et al. 2011’s studies stratified by platform and case-control status: The PGC-BD’s wave 1 data was drawn from 11 different studies genotyped on four different platforms. This table documents the number of subjects and platforms. Those highlighted in Yellow were used to define Group 1, the rest were used in group 2

<table>
<thead>
<tr>
<th>Study</th>
<th>Ancestry</th>
<th>Case (n)</th>
<th>Control (n)</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOMA-Bipolar Study, University of Bonn and CIMH Mannheim</td>
<td>German</td>
<td>675</td>
<td>1,297</td>
<td>Illumina HumanHap 550</td>
</tr>
<tr>
<td>Genetic Association Information Network (GAIN) &amp; Bipolar Genome Study (BiGS)</td>
<td>European-American</td>
<td>542</td>
<td>649</td>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
</tr>
<tr>
<td>GlaxoSmithKline (GSK)</td>
<td>British, Canadian or Scottish</td>
<td>890</td>
<td>902</td>
<td>Illumina HumanHap 550</td>
</tr>
<tr>
<td>Pritzker Neuropsychiatric Disorders Research Consortium (Pritz)</td>
<td>European-American</td>
<td>1,130</td>
<td>718</td>
<td>Illumina HumanHap 550</td>
</tr>
<tr>
<td>Systematic Treatment Enhancement Program for Bipolar Disorder (STEP1)</td>
<td>European-American</td>
<td>922</td>
<td>645</td>
<td>Affymetrix GeneChip Human Mapping 500K Array</td>
</tr>
<tr>
<td>Systematic Treatment Enhancement Program for Bipolar Disorder (STEP2)</td>
<td>European-American</td>
<td>659</td>
<td>192</td>
<td>Affymetrix Genome-Wide Human SNP Array 5.0</td>
</tr>
<tr>
<td>Thematically Organized Psychosis (TOP) Study</td>
<td>Norwegian</td>
<td>203</td>
<td>349</td>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
</tr>
<tr>
<td>Trinity College Dublin</td>
<td>Irish</td>
<td>150</td>
<td>797</td>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
</tr>
<tr>
<td>University College London (UCL)</td>
<td>British</td>
<td>457</td>
<td>495</td>
<td>Affymetrix GeneChip Human Mapping 500K Array</td>
</tr>
<tr>
<td>University of Edinburgh</td>
<td>Scottish</td>
<td>282</td>
<td>275</td>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
</tr>
<tr>
<td>Wellcome Trust Case-Control Consortium (WTCCC)</td>
<td>British</td>
<td>1,571</td>
<td>2,931</td>
<td>Affymetrix GeneChip Human Mapping 500K Array</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>7,481</td>
<td>9,250</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.2 - Definitions of Group 1 and Group 2: Group 1 and Group 2 are broken down by numbers of cases and controls and platforms

<table>
<thead>
<tr>
<th></th>
<th>Group 1 (WTCCC, Bonn, Dub)</th>
<th>Group 2 (BiGS, GSK, Pritz, Step1/2, TOP, UCL, EDI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Case</td>
<td>Control</td>
</tr>
<tr>
<td>Affymetrix GeneChip Human Mapping 500K Array</td>
<td>1571</td>
<td>2931</td>
</tr>
<tr>
<td>Illumina HumanHap 550</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
<td>150</td>
<td>797</td>
</tr>
<tr>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
<td>675</td>
<td>1297</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2396</strong></td>
<td><strong>5025</strong></td>
</tr>
</tbody>
</table>
Figure 2.1 - Raw mixed genotyped / imputed data: We binned the dosage data from 4 SNPs with different amounts of genotyped data (A - 10% genotyped, B - 34% genotyped, C – 51% genotyped, D – 86% genotyped). There are two graphs per SNP. The top graph is a histogram of the dosage data. The bottom graph also a histogram, but with the amount per bin expressed on a log scale. We made use of the log scale so that outliers and sparsely represented dosage would be visible. Clearly, not only is the data skewed, but it also has outliers; these features are typical for many of the SNPs in this study.
Figure 2.2 - Performance comparison of AUC and Logistic regression on two mixed genotyped / imputed SNPs: AUC’s ability to identify a more consistent signal than logistic regression in the presence of skewness and outliers. We took two SNPs of mixed imputed/genotyped data, 56% and 83% genotyped data respectively. In frame 1, we plotted the log histogram of the data just as we did in Figure AA1. We then calculated the AUC and odds ratio (OR) for every possible combination of the 11 studies in the PGC-BD data. In Frame 2, we plotted the log(OR) for the entire dataset in the first graph and the log(OR) for each of the other $2^{11}$ study combinations on the 2nd graph. In Frame 3, we plotted the histogram of the results for the $2^{11}$ log(AUC/(1-AUC))’s and log(OR). In both instances outliers/skewness influence the conclusion drawn by logistic regression. In the first instance, AUC sees a signal where LogR cannot, and, in the second, AUC does not find a spurious signal that logistic regression erroneously identifies because of the impact of a few outliers.
Figure 2.3 - Cross-group correlations in Linkage equilibrium Sets: This is a 2x2 array of the cross-group correlation measured in G_LE (100% Genotyped linkage equilibrium set) and I_LE (mixed imputed / genotyped linkage equilibrium set) vs. Uncorrected AUC and Corrected AUC. Specifically, spot (1,1) is a scatter plot of AUC$_1$ vs. AUC$_2$ for SNP set G_LE, spot (1,2) is a scatter plot of cAUC$_1$ vs. cAUC$_2$ for SNP set G_LE, spot (2,1) is scatter plot of AUC$_1$ vs. AUC$_2$ for SNP set I_LE, and spot (2,2) is scatter plot of cAUC$_1$ vs. cAUC$_2$ for SNP set I_LE. The uncorrected AUC is correlated across the mixed imputed/genotyped LE set of SNPs (I_LE) at a 0.526 (p=0) level whereas the corrected AUC (cAUC) on I_LE is correlated at 0.031 (p=4.4x10$^{-61}$).
Figure 2.4 - Adjusting the various mixtures of genotyped and imputed data: We then examined the relationship between this increased correlation and imputation by starting with 100% imputed SNPs and gradually adding in SNPs with increasing fractions of genotyped data. Purely imputed SNPs were correlated at a 0.37 (p=0) level. This correlation jumped to ~0.71 with the introduction of mixed imputed / genotyped SNPs. As we continue to introduce SNPs with increasing fraction of genotyped data until we had the entire I_LE dataset (282,551 SNPs), the correlation decreases (~0.71->0.526). The 2nd half of Figure 2.4 started with the entire I_LE dataset and gradually removed those SNPs that were purely or partially imputed. The correlation decreases as the fraction of imputed data decreases (0.526->0.069). Though the signal is weaker, logistic regression (log of the odds ratio) demonstrates a similar relationship.
Figure 2.5 - Impact of increasing imputation quality score thresholds on cross-group correlations: We replotted Figure 2.4 with the SNPs which remained after applying stricter Imputation Quality Score (IQS) thresholds. We had to increase the IQS threshold to 0.98 before we saw a dramatic change in the pattern of these correlations. With this IQS threshold, the corr(AUC1, AUC2) of the purely imputed SNPs almost returns to that of the purely genotyped SNPs (0.077 / 0.060).

Figure 2.6 - Impact of increasing minor allele frequency (MAF) thresholds on cross-group correlations: We replotted Figure 2.4 with the SNPs which remained after applying stricter minor allele frequency (MAF) thresholds. SNPs with more extreme MAF frequencies are more highly correlated when purely or partially imputed than SNPs of less extreme frequencies. Nevertheless, purely or partially imputed SNPs with MAF 40-60% still exhibit a much larger correlation than purely genotyped SNPs in the same MAF range.
2.7 - Two-dimensional histogram of genotyped SNP rank versus imputed SNP rank: We varied number of bins (10, 100) and IQS threshold (0.3, 0.8, 0.98). The relationship requires 10% bins for any relationship to become apparent.
3.A Introduction: In the Background and Significance Section, we established the following: 1) It is critical that we determine the cause of bipolar affective disorder (BD); 2) BD is highly heritable; 3) GWAS meta-analyses have had difficulty identifying replicable SNPs; and 4) despite the difficulties replicating SNPs, the results of polygenic modeling suggest that SNPs do play some role in causing BD.

In the Data Reliability section, we began by reviewing the imputation literature and then compared genotyped and imputed data. We found that, for this particular data set, imputed data seemed to be structured differently than genotyped data. We also found that the Area Under the receiver operating characteristic Curve (AUC) was a far better metric for detecting this structure than logistic regression (LogR).

In this section we set aside the possible complications associated with mixing imputed and genotyped data, and focus on the potential upside of using the AUC to detect case-control signals in BD-GWAS.

When analyzing our data, we have observed that (1) all but ~46K of the PGC-BD SNPs are wholly- or partially- imputed. Hence, it is possible that the majority of these SNPs are skewed +/- outliers (see Figure 2.1). In fact, (2) previous work has observed that imputed data can often be skewed [63], further hinting that our data may also exhibit this feature. Finally (3) we have seen that a substantial amount of the signal is associated with covariates, making it likely that covariates contribute nonlinearly to the disease state.
This combination of features may be responsible for our observations that AUC outperforms LogR in detecting the non-disease related signals investigated in Section 2. Specifically: LogR is 1) not well-suited to the analysis of skewed data, 2) is sensitive to outliers, 3) its trial-to-trial variance can be unbounded, and 4) can only be corrected for linear relationships between covariates whereas AUC is 1) not adversely affected by skewness, 2) is robust to outliers, 3) has a bounded trial-to-trial variance, and 4) can be corrected for BOTH linear and non-linear relationships between covariates. (For a detailed explanation of 1-3 see Appendix, specifically Part B, Sections 3-5. For a detailed explanation of covariate-correction, see Appendix Part C, Sections 6-11, particularly 9.)

We would like to accommodate for the potential confound of skewness by (3a) quantifying the amount of skewness in our data across the fully-genotyped, mixed and imputed SNPS, and (3b) using AUC – a skewness insensitive measure -- to detect the signal associated with each SNP.

One issue previously preventing us from implementing (3b) above is that AUC has only recently become an option; prior to our recent work, we were not aware of a practical means by which to correct AUC for covariates. Given that we now have the means to correct for both skewness and covariates by using AUC, it seems natural / important to now ask whether or not AUC would be a better analytic tool for analyzing our imputed data than logistic regression. (The quick uncorrected AUC code is in C.12.4, and the corrected AUC code is in C.12.6.)

3.B Research Design and Methods

3.B.1 Dataset Description: In this analysis we will be working with the same dataset we analyzed in the Data Reliability Section. To review, through the PGC-BD, in conjunction with dbGAP, NIMH, and The Wellcome Trust, we had access to 5989 BD
genotyped cases and 7998 genotyped controls.[5] A complete description of the
genotyped dataset is in Sklar et al.[31] As noted above, the cases and controls were
drawn from 11 different sources and were genotyped on 4 different platforms. (The per
study details and their genotyping platforms are listed in Table 3.1) Case diagnoses
were as follows: BDI (bipolar disorder, type 1) - 84%, BD2 (bipolar disorder, type 2) -
11%, SABP (schizoaffective bipolar disorder) - 4%, and other bipolar diagnosis - 1%.
There were only 46,234 SNPs genotyped in all 11 studies. There were 1,016,924 SNPs
genotyped in at least two groups.

3.B.2 Methods

3.B.2.a Quantifying the amount of skewness in our data across the fully-genotyped,
mixed and imputed SNPs

3.B.2.a.1 Calculating Skewness per SNP: Generally, the skewness
as the lop-sidedness of a data-distribution. More formally, it is a mathematical concept
which can be calculated using the following formula: \( \gamma = \frac{1}{\sigma^3} \left[ \langle v^3 \rangle - 3 \mu \sigma^2 - \mu^3 \right] \) where
\( \langle v^3 \rangle \) represents the mean of the cube of the dosage, and \( \mu \) and \( \sigma \) represent the mean
and standard-deviation of the dosage, respectively. Rather than relying upon something
as vague as visual inspection when dealing with ~2.4 million SNPs, we calculated
skewness for each SNP. We then plotted a histogram of the following SNP sets: 1) purely genotyped, 2) mixed imputed/genotyped, and 3) purely imputed. We did this for a
range of information quality score (IQS) thresholds (0.3, 0.8, 0.98) to see if it would
impact / reduce the degree of skewness in our data. Since our ultimate goal is to
understand whether skewness may be interfering with LogR’s ability to detect the
disease related signal in BD, we also examined whether significant SNPs are also skew.
To that end, we created histograms of the skewness collected across the following SNP
sets: SNPs with \( p < 10^{-1}, 10^{-2}, 10^{-3}, \ldots, 10^{-9} \).
3.B.2.a.2 Rank order of AUC vs. rank order of LogR: To determine if skewness impacts the conclusions drawn from either AUC or LogR-based analyses, we created a two-dimensional histogram of rank order of SNPs according to AUC versus the rank order of SNPs according to LogR. We used 64 bins vs. 64 bins. We considered SNPs with IQS > 0.3 and 0.01<MAF<0.99 as in the Sklar et al. paper. If it were indeed the case that AUC and LogR assigned similar ranks to SNPs, then we should observe that most of the mass present within this 2-d histogram would be concentrated around the ‘diagonal’; in other words, SNPs that have high or low rankings under AUC should receive comparable rankings under LogR. In fact, as shown in the Appendix, Section B, Part 3.4., we should expect such a diagonal distribution if the data were not skewed or long-tailed.

Here, we check to see if/how our data differs from this expectation. When analyzing our data, it is fair to assume that we are assessing the influence of skewness rather than long-tailedness because long-tailed distributions, though not impossible, should not occur frequently due to the bounded nature of dosage data, i.e, within [0,2]. To put our results, which did not lie along the main diagonal, into greater context, we generated comparable graphs for simulated data. We generated this simulated data to be as close to Gaussian as possible, while possessing varying degrees of skewness and signal strength (see Appendix B Figure 3 and section 3.3 for details on how we constructed this simulated data).

Before continuing, we remark that we expect to be able to draw the strongest conclusions by comparing the rankings of AUC against LogR, rather than comparing the rankings of cAUC against cLogR. This is because cAUC and cLogR are corrected to different degrees; cAUC is corrected for both linear and non-linear covariates, but cLogR is only corrected for linear components. Even if there were no skewness present in the
data, we would expect cAUC and cLogR to rank the SNPs differently simply because of different levels of covariate correction.

While keeping the caveat of different degrees of covariate-correction in mind, we nevertheless performed an analogous analysis comparing SNP rank by cAUC with SNP rank by cLogR for the sake of completeness.

3.B.2.b Can using AUC/cAUC rather than LogR/cLogR in imputation-reliant GWAS improve replication: Replication analyses require two waves: One for the initial analysis and another for replication. We had hoped to use the wave 1 and wave 2 data used in the Sklar et al. 2011 PGC-BD analysis. Unfortunately, we were unable to access this wave 2 data.

Thus, in order to complete our analysis, we created our Waves A & B from the PGC-BD Sklar et al. 2011 Wave 1. We were concerned that if we settled on a single Wave A large enough to detect true signals, our Wave B results would be too greatly affected by study-to-study heterogeneity. Therefore, we decided to analyze multiple Wave A/B combinations of the PGC-BD Wave 1 data and look for trends. Ideally, one would consider all 2^11 means of combining the 11 available studies as we did in Figure 2.2. Given the number of covariates involved, these types of analyses are rather time intensive so this was not an option. Though it is a bit arbitrary, we decided to run our analyses for the 11 choose 2 (55) ways for selecting Wave A (9 studies) and Wave B (2 studies) and to consider the results in aggregate. Again, we limited our attention to SNPs with IQS > 0.3 and 0.01<MAF< 0.99.

3.B.2.b.1 Replication: For each Wave A/B pair, we ran the following analyses: cAUC, and cLogR. For each of these two types of analyses, we defined the top SNPs to be those with p-value < 10^{-5}. As was done in the Sklar et al. 2011 paper, this top SNP Set was then further refined using PLINK’s ‘--clump’ function. PLINK’s ‘—clump’ function first
identifies a set of index SNPs which represent the SNP clumps within this set. It then clumps all of the other SNPs that are within 1 Mb of these index SNPs, in LD with the index SNP (r² > 0.2) or nominally associated with disease (P < 0.05). This refined top SNP set was then tested for replication in Wave B. cAUC calculated on Wave B being used to validate cAUC values calculated on Wave A, and cLogR calculated on Wave B being used to validate cLogR values calculated on Wave A. After determining which of these SNPs replicated after correction for multiple comparisons in each wave A/B pair, we compared the relative replication success rate of cAUC with cLogR using t-tests. We repeated this for multiple Imputation Quality Score thresholds, i.e., 0.3, 0.8, and 0.98, and considered three sets of SNPs, i.e., all SNPs, SNPs with |skewness| > 1, and SNPs with |skewness < 1|.

We had also intended to compare replication rates between AUC and LogR in a comparable manner. Because of the great disparity in signal detected AUC and LogR, there were approximately 10² times more significant SNPs identified by AUC than LogR, e.g., 51355 vs. 347. This number was so large as to be make clumping them with PLINK too computationally expensive. Therefore, we limited our analysis to cAUC and cLogR as this corrected signal would be of greatest practical interest to GWAS researchers.

3.B.2.b.2 Enrichment: For each of the 55 Wave A/B pairs, we will use logistic regression to calculate the association between our Wave A SNP results with enrichment at a p<0.05 level in Wave B. For AUC, we will be inputting abs(log(AUC/(1-AUC))) into the logistic regression. For LogR, we will be inputting abs(log(p-value)). Once we have done this for all 55 A/B pairs, we will compare the relative relationships: enrichment using AUC vs. enrichment using LogR and enrichment using cAUC vs. enrichment using cLogR.
3.C Results

3.C.1 Quantifying the amount of skewness in our data across the fully-genotyped, mixed and imputed SNPS

3.C.1.a Calculating skewness per SNP: After we calculated skewness per SNP, we divided our SNPs into the three categories we identified in the Data Reliability Section: 1) Purely genotyped, 2) Mixed imputed/genotyped, and 3) Purely imputed. We found that skewness was present in each of these categories. (See Figure 3.1) The mixed imputed / genotyped data SNPs had the widest range of skewness. This was followed closely by purely imputed. It is worth noting that skewness is present in purely genotyped SNPs as well, although the range of values is narrower.

Using stricter IQS thresholds, including the strict threshold of 0.98, narrowed the range of skewness values observed to almost that of the genotyped data. These stricter thresholds by no means eliminated skewness as genotyped data itself can be quite skewed. Similarly, we checked whether skewness can be eliminated/reduced by focusing on SNPs of increasing significance, i.e., corrected p-values <10⁻¹ to 10⁻⁹. We found that, though skewness is reduced, even the most significant SNPs demonstrated some level of skewness.(See Figure 3.2) We created a similar figure by thresholding on uncorrected p-values and found similar level of skewness.

3.C.1.b Rank order of AUC vs. rank order of LogR: Figure 3.3’s row 1 contains the two-dimensional histogram generated when one compares the rank order of SNPs according to AUC versus the rank order of SNPs according to LogR. Column 1 is the raw bin data, and Column 2 is the same data displayed on a log scale. Figure 3.3’s row 2 contains the two-dimensional histogram generated when one compares the rank order of SNPs according to cAUC versus the rank order of SNPs according to cLogR. Again, Column 1 is the raw bin data, and Column 2 is the same data displayed on a log scale.
Though some relationship between AUC SNP rank and LogR SNP rank was evident in row 1, its lack of symmetry demonstrates that they are ranking the SNPs quite differently. Interestingly, a very good AUC was likely to yield a consistent and very good LogR p-value, but the opposite was not true. There was little to no relationship between cAUC and cLogR (a very dispersed signal). For what signal there was, the relationship between cAUC and cLogR was relatively symmetric, and there was a general agreement amongst the top SNPs (in this case the top 1/64th of 2.5 million).

To put these results into context, we created data distributions with varying degrees of skewness (low, medium, and high) and signal (low, medium, and high). We then created comparable two-dimensional histograms of the simulations LogR SNP rank vs. AUC SNP rank (See Figure 3.4). We noted that the two-dimensional histogram for AUC SNP rank vs. LogR SNP rank (see Figure 3.3) is most like the high signal / high skewness simulation. We also noted that the two-dimensional histogram for cAUC SNP rank vs. cLogR SNP rank is most like the low signal / high skewness simulation. To further explore this point, we determined how the AUC SNP rank vs. LogR SNP rank histogram would change if we replotted them after dividing the SNPs into ‘High Skewness’ (|skewness|>1) and ‘Low Skewness’ (|skewness|<1). The ‘High Skewness’ graph looks even more like the high signal / high skewness simulated graph in Figure 3.4, and the ‘Low Skewness’ graph looks like the high signal / low skewness simulated graph in Figure 3.4. (See Figure 3.5)

It would appear that skewness is influencing both the uncorrected and corrected analyses. Unfortunately, it also appears that, similar to our findings in the Data Reliability section, much of the signal in this dataset disappears with correction. In the presence of high signal / high skewness, it would appear that the use of AUC could result in the identification of a more robust and, therefore, replicable signal. It is less clear, given the
general agreement of top SNPs in the low signal / high skewness scenario, if we would expect to see an improvement with cAUC.

3.C.1.c Can using AUC rather than LogR in imputation-reliant GWAS improve replication:

3.C.1.c.1 Replication: In Table 3.2, we report the replication rates of cAUC and cLogR for various Imputation Quality Score (IQS) thresholds, specifically 0.3, 0.8, and 0.98.

3.C.1.c.1.a Replication Set Trends: cAUC identified far fewer SNPs with p<10^{-5} (~1/3) than cLogR. At the IQS threshold of 0.3, i.e., the level of thresholding used in the Sklar et al. 2011 paper, we were only seeking to replicate ~8 SNPs for cAUC and ~24 SNPs for cLogR after applying PLINK’s clump function. A replication rate < 12.5% would be difficult to verify for cAUC and, similarly, a replication rate < 4.2% would be difficult to verify for cLogR. Further IQS thresholding decreases the number of SNPs we sought to replicate, i.e., IQS > 0.8 = ~6 SNPs and IQS > 0.98 = ~1 SNPs, would make it difficult to find evidence of replication rates less than 16% and 100% respectively. This is

    Less-skewed SNPs (|skewness| < 1) were far more likely to be significant than one would expect by chance, i.e., cAUC – 75%, cLogR – 60%, expected – 52%. This, combined with the fact that cAUC identified so few SNPs to replicate even at the IQS > 0.3 (~ 2 SNPs), made it impossible to meaningfully comment on how well more-skewed (|skewness| > 1) SNPs identified using cAUC would replicate.

3.C.1.c.1.b Replication Rates: First, we will discuss the results from IQS > 0.3, i.e., the IQS threshold used in Sklar et al. There were a few striking results: 1) The SNPs identified by cAUC replicate at a higher rate than cLogR (5.9% vs. 1.9%, p=0.00040), and 2) the replication rate of cAUC is mainly driven by the replication rate of the SNPs with |skewness| < 1, i.e., |skewness| < 1 - 9.3%, |skewness| > 1 - 1.0%. Though tempting, one should not conclude that cAUC performs less well on more-skewed
more-skewed (|skewness| > 1) SNPs. As we noted in the previous section, we were testing very few
more-skewed (|skewness| > 1) SNPs. Given that we were testing ~2 SNPs per A-vs-B
pair and presupposing a replication rate comparable to the less-skewed SNPs (~9.3%),
it is remarkable that we succeeded in replicating any SNPs at all. Further, the fact that
we were only testing ~6 less-skewed SNPs, but detected a 9.3% replication rate
suggests that this rate might be an underestimate. On the other hand, the replication
rate of cLogR does seem to be mainly driven by the replication rate of the SNPs with
more skewness (|skewness| > 1). cLogR’s less-skewed SNP replication rate is 2.6%,
and its more-skewed SNP replication rate of 5.4%. We can make this assertion because
far fewer of these types were being tested (~10 SNPs vs. ~15 SNPs) and yet the
replication rate was higher.

When one increases the IQS threshold from 0.3 to 0.8, the replication rates are
no longer statistically significant different between cAUC and cLogR for either the ‘All
SNP’ or ‘SNPs with |skewness| < 1’ though the trend of cAUC outperforming cLogR
continues. cLogR replication rate on SNPs with |skewness| > 1 is slightly statistically
significantly better than the rate of cAUC, but again, this should not be assigned any real
weight since so were seeking to replicate so few SNPs (~1 SNPs).

When one increases the IQS threshold from 0.8 to 0.98, we are seeking to
replicate ~10 SNPs for cLog and ~1 SNPs for cAUC. These numbers are before making
skewness distinctions. It no longer makes any sense to compare their replication rates.

Since cLogR is starting from a pool of ~24 SNPs satisfying IQS<0.3, p < 10^-5,
there are still enough SNPs standing after increasing the IQS threshold to 0.98 to have a
meaningful discussion of its replication rates. Interestingly, as one increases the IQS
threshold, i.e., focuses on higher quality SNPs, replication rates improve. This is most
marked amongst its most skewed (|skewness| > 1) SNPs (5.4%->8.0%->19.4%), but is also true amongst the less-skewed (|skewness| < 1) SNPs (2.6% -> 3.6% -> 4.9%).

3.C.1.c.2 Enrichment: In Figure 3.6, we compare the results of our enrichment analyses for SNPs satisfying IQS > 0.3. We created a histogram of the absolute value of the Log base 10 of the 55 p-values generated when performing the enrichment analyses for LogR, AUC, cLogR, and cAUC under the following conditions - All SNPs, |skewness >1|, and |skewness > 1|. Column 1 contains the results from LogR, Column 2 the results from AUC, Column 3 the results from cLogR, and Column 4 the results from cAUC. Since we wanted to be able to assess the significance of the difference in the distributions using a t-test, we replaced the abs(Log10 (p=0)), i.e., Infinity, with 500, when we performed those calculations.

Strong performances by LogR and AUC in Wave A are dramatically associated with p-value enrichment in Wave B. That being said, the evidence of this enrichment is far stronger and more consistently present with AUC. Also note that unlike the cLogR and cAUC, the degree of enrichment we observe seems to be independent of the skewness of the SNPs being considered.

Though both cLogR and cAUC demonstrate enrichment, cAUC is associated with far greater enrichment \(p = 2.08 \times 10^{-13}\) than cLogR. This is driven mainly by the enrichment of the more-skewed (|skewness| > 1) SNPs \(p= 1.80 \times 10^{-40}\). cAUC shows significantly less enrichment than cLogR \(1.89 \times 10^{-11}\) with SNPs with less skewness (|skewness| < 1).

Interestingly, all but one of the trends we observed in Figure 3.6 continue if one increases the IQS threshold from 0.3 to 0.8 to 0.98. The exception being that the differences in enrichment observed between cLogR and cAUC on more-skewed (|skewness| > 1) SNPs. This difference decreases though it does not disappear as the
IQS threshold increases. The change is obvious with visual inspection. In addition, the t-test p-value drops from $1.80 \times 10^{-40}$ to $8.86 \times 10^{-33}$ to $8.46 \times 10^{-7}$ as the IQS threshold increases from 0.3 to 0.8 to 0.98. (See Figure 3.7)

3.D Discussion:

3.D.1 Skewness in our data PGC-BD: Skewness is clearly present in the PGC-BD dataset. Approximately half of the SNPs with an IQS > 0.3 have $|\text{skewness}| > 1$. It is most present in the mixed imputed/genotyped and imputed data sets, but it is still a big factor in the genotyped data. It cannot be eliminated/avoided simply by focusing on the most significant SNPs and appears to influence the way that AUC and logR rank SNPs. The fact that we found that a good AUC resulted in a good LogR, but the reverse was less true, hints that this influence has a more negative impact on LogR results rather than AUC and should not simply be ignored when analyzing the PGC-BD data.

3.D.2 Replication and Enrichment:

3.D.2.a LogR vs. AUC: As noted in Section 3.b.1.c.1, replication rates in GWAS are typically assessed by 1) identifying the SNPs in Wave A satisfying $p < X$, 2) performing some sort of thinning process which balances linkage disequilibrium and p-values, and 3) determining if those SNPs achieve p-values in Wave B which are still significant after Bonferroni correction. Because of the great disparity between AUC’s and LogR’s ability to detect the uncorrected signal present in our PGC-BD data, we were not able to directly compare their replication rates. There was though a very obvious and expected increase in the enrichment at a $p < 0.05$ level as observed with AUC when compared with LogR. (Figure 3.6) This enrichment was independent of the skewness level.

3.D.2.b cAUC outperforms cLogR for SNPs with IQS>0.3: Because of the marked decrease in signal after correction, we did not expect either method to replicate particularly strongly. Given the general lack of expected replication and the slight
relationship between the top SNPs of cAUC and cLogR, we were skeptical that there would be obvious differences in replication rate.

Polygenic models and GCTA tell us that the majority of the signal in BD is within common SNPs rather than its outliers. Given that enrichment is a more powerful metric for assessing this type of signal, we thought differences, if any, between the two methods would be more likely to manifest in this type of analysis. We were, therefore, not particularly surprised that cAUC is associated with a marked increase in enrichment. It also seemed entirely appropriate that this improvement was driven by cAUC analysis of SNPs with $|\text{skewness}| > 1$ as this is where we expected AUC’s distribution-independence to distinguish itself.

We were unsure how to interpret the fact that the enrichment of more-skewed ($|\text{skewness}| > 1$) SNPs was tempered by increasing IQS thresholds, i.e., limiting ourselves to SNPs of the best imputation quality. Since the level of skewness does not change dramatically when one increases the IQS threshold from 0.8 to 0.98 while the level of enrichment continues to decrease, changes in skewness alone are unlikely to be solely responsible. While skewness remains relatively constant, the fraction of outliers / intermediate values continues to decrease with stricter thresholding. Therefore, it is more likely that the enrichment we are observing on the more-skewed ($|\text{skewness}| > 1$) SNPs is a product of not just cAUC’s insensitivity to skewness, but also its insensitivity to outliers or, in the case of less well-imputed SNPs, intermediate values. It is also possible that some of the reduction of effect is driven by less extreme p-values being assigned when a t-test is performed on smaller-sized SNP sets. (IQS > 0.3 – 1,120,072 SNPs, IQS > 0.8 – 891,235 SNPs, IQS > 0.98 – 360,994 SNPs)

cLogR shows far greater enrichment than cAUC on the less-skewed ($|\text{skewness}| < 1$) SNPs. In the absence of skewness, we would have expected them to roughly
agree. The most-likely explanation for this disparity is that cAUC is corrected to a greater
degree than cLogR, i.e., non-linear + linear relationships vs. linear relationships. Of
course, this is concerning because GWAS, in general, and the PGC-BD GWAS, in
particular, are relying on corrected logistic regression to correct for all non-disease
related signal.

Given the dramatic drop in signal after correction, we were surprised that we
were able to detect differences in the replication rates of cAUC and cLogR, namely, that
cAUC outperformed cLogR for SNPs with IQS > 0.3. Since cAUC’s dramatic enrichment
was driven by the more-skewed SNPs (|skewness| > 1), we were initially surprised to
find that cAUC’s replication rates appeared to be driven by its performance on SNPs
with |skewness| < 1. As we discussed above, we ultimately realized that there were too
few SNPs satisfying the criteria p<10^{-5}, IQS>0.3, and |skewness| > 1 to meaningfully
comment on cAUC’s replication rate for more-skewed (|skewness|>1) SNPs.

We do not yet have an explanation for the following: 1) cAUC is far less likely
than one would expect by chance to assign a significant p-value to a more skewed
(|skewness| > 1) SNP, e.g., IQS > 0.3 – 75% vs. 52%, IQS > 0.8 – 79% vs. 54%, IQS >
0.98 – 81% vs. 57%, and 2) cAUC identifies on average a third of the number of SNPs
with p<10^{-5} that clogR identifies. The second point though may be a combination of
these factors: 1) the signal after correction is weak, 2) cAUC is being corrected for both
non-linear and linear relationships between covariates whereas cLogR is only being
corrected for linear relationships between covariates, and 3) cAUC is not sensitive to
outliers which can influence cLogR.

We were unable to meaningfully compare cAUC and cLogR for stricter IQS
thresholds because so few SNPs were identified by cAUC. That being said, the trend of
cAUC outperforming cLogR is still present with IQS > 0.8. Further, since the percent
drops in the number of SNPs we sought to replicate with cAUC or cLogR was not dramatically different with changes in IQS thresholds (0.3 to 0.8 – 37.5% vs. 20% and 0.8 to 0.98 – 85% vs. 60%), we did not feel able to comment on whether cAUC had a bias towards assigning significance to low quality SNPs.

3.D.2.c cLogR replication rates improve with stricter IQS thresholds: It is worth reiterating that increasing the IQS threshold (0.3->0.8->0.98) improves the replication rate of cLogR: (1.9%->3.4%->7.2%) – All SNPs, (5.4%->8.0%->19.4%) - the more skewed (|skewness| >1) SNPs and (2.63% to 3.56% to 4.91%) – less-skewed SNPs. As was true for cAUC, far fewer more-skewed (|skewness| >1) SNPs were being tested for replication making this greater replication rate even more meaningful. This result is a curious juxtaposition to our enrichment results wherein cLogR enrichment seems largely driven by less-skewed (|skewness| <1) SNPs.

Since the set [SNPs with IQS>0.98] is contained in the set [SNPs with IQS > 0.8] and the set [SNPs with IQS > 0.8] is contained in the set [SNPs with IQS > 0.3], SNPs of higher imputation quality are responsible for cLogR’s replication rates. This, when combined with the fact that using an IQS threshold of 0.98 almost reduces the non-diseased related signal we are able to detect with AUC to that of the level of genotyped SNPs, suggests that psychiatric GWAS could eliminate false positives, perhaps without introducing false negatives by using the much stricter IQS score of 0.98. Since cLogR is not supposed to be well-suited to analyzing skewed, non-integer data, it would be important to determine on a SNP by SNP basis whether or not the skewness or outliers / intermediate values are somehow leading cLogR to first identify and then confirm a false positive. Using an IQS threshold of 0.98 reduces the fraction of, but does not entirely eliminate, outliers / intermediate values which can erroneously influence cLogR. (See Figure 3.8)
3.D.4 In Summary: In our 11 choose 2 analyses, (1) cAUC is associated with dramatic improvement in enrichment of replication at a p< 0.05 level when compared with cLogR. This improvement in enrichment is largely driven by an improvement in more-skewed (|skewness| > 1) SNPs, (2) cLogR is also associated with enrichment of replication at a p < 0.05 level though it is less dramatic than that seen with cAUC. It seems to be driven more by its performance on less-skewed (|skewness| <1) SNPs.

When one limits oneself to p-value outliers (p<10-5 in Wave A), (1) cLogR finds the disease-related signal within the PGC-BD GWAS within the more-skewed SNPs of higher imputation quality and (2) cAUC finds it within the SNPs of lower imputation quality. cLogR’s behavior (1) is the opposite of what we would have expected given our enrichment findings and our gross understanding of the strengths and weaknesses of AUC and LogR in general. Given that the skewed SNPs that replicated where of high imputation quality, they may actually be well-suited to LogR analysis, but this should be confirmed as a high-IQS value does not guarantee that the influence of outliers / intermediate values has been eliminated. (See Figure 3.8) It is quite possible that skewness, outliers / intermediate values, and the impact of non-linear relationships between covariates may be repeatedly and incorrectly influencing LogR to replicate what is actually a false positive.

We were unable to determine if cAUC is bias towards lower imputation quality SNPs. cAUC certainly identified and replicated more SNPs of lower imputation quality across our 11 choose 2 studies than cLogR. Given the difficulty of imputing SNPs with lower minor allele frequency (MAF) and the growing support the idea that complex-disease being driven by these less-common variants, it is possible that cAUC is detecting a true signal that to which cLogR is blind. It also worth acknowledging that
neither replicates particularly strongly under the best of circumstances and this may be mainly due to the signal that remains after correction not being particularly strong.

Since psychiatric-illness is increasingly believed to be caused by many SNPs of small effect, perhaps we should focus less on the performance of p-value outliers (replication rates) and place greater emphasis on the results of our enrichment analysis. From that perspective, it would appear cAUC outperforms cLogR for more-skewed ($|\text{skewness}| > 1$) SNPs. Though it is tempting to conclude that cLogR outperforms cAUC for less-skewed ($|\text{skewness}| > 1$) SNPs, we are reluctant to take this position because cAUC is corrected for more relationships between covariates than cLogR. These correction differences leave open the possibility that cLogR apparent increase in enrichment for the less-skewed ($|\text{skewness}| < 1$) SNPs is actually the result of a still-uncorrected relationships between covariates. Given cAUC’s larger covariate-correction, when it demonstrates more enrichment than cLogR, it is much more convincing.

Given the differential performance of cAUC and cLogR, particularly since this performance seems to be influenced by skewness, it seems reasonable to assert that some of the PGC-BD’s replication difficulties stem from its reliance on cLogR without consideration of skewness.

3.D.4 Further work: It seems important to understand: 1) why cAUC identifies far fewer extremely significant SNPs than cLogR, 2) why cAUC is far less likely to assign a significant p-value to a more-skewed SNP rather than a less-skewed SNP, and 3) why cLogR is more likely to replicate a more-skewed SNP than a less-skewed SNP, particularly those with a very high IQS Score. It seems important to inquire what role, if any, a weak disease signal, outliers, non-linear relationships between covariates, and Minor Allele Frequencies are playing in the above.
Since risk prediction bridges the gap between enrichment and replication, it seems worthwhile that we determine if there is a marked difference in the performance of linear-additive models based upon cAUC and cLogR. If cAUC were able to explain more of the disease variance than cLogR using fewer SNPs, it would be a very strong indication that cAUC is bringing the relevant signal to the fore.
3.E Acknowledgments: Chapter 3, in part is currently being prepared for submission for publication of the material. McGrouther, CC; Kelsoe, JR; Schork, NJ; and Rangan, AV. The dissertation author was the primary investigator and author of this paper.
Table 3.1 - Case / Control / Platform Details for each of the 11 studies in the PGC-BD Wave: Through the PGC-BD, in conjunction with dbGAP, NIMH, and The Wellcome Trust, we had access to 5989 BD genotyped cases and 7998 genotyped controls.[5] A complete description of the genotyped dataset is in Sklar et al.[31] The cases and controls were drawn from 11 different sources and were genotyped on 4 different platforms. This table articulates the per study details and their genotyping platforms.

<table>
<thead>
<tr>
<th>Study Description</th>
<th>Ancestry</th>
<th>Case (n)</th>
<th>Control (n)</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOMA-Bipolar Study, University of Bonn and CIMH Mannheim</td>
<td>German</td>
<td>675</td>
<td>1,297</td>
<td>Illumina HumanHap 550</td>
</tr>
<tr>
<td>Genetic Association Information Network (GAIN) &amp; Bipolar Genome Study (BiGS)</td>
<td>European-American</td>
<td>542</td>
<td>649</td>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
</tr>
<tr>
<td>GlaxoSmithKline (GSK)</td>
<td>British, Canadian or Scottish</td>
<td>890</td>
<td>902</td>
<td>Illumina HumanHap 550</td>
</tr>
<tr>
<td>Pritzker Neuropsychiatric Disorders Research Consortium (Pritz)</td>
<td>European-American</td>
<td>1,130</td>
<td>718</td>
<td>Illumina HumanHap 550</td>
</tr>
<tr>
<td>Systematic Treatment Enhancement Program for Bipolar Disorder (STEP1)</td>
<td>European-American</td>
<td>922</td>
<td>645</td>
<td>Affymetrix GeneChip Human Mapping 500K Array</td>
</tr>
<tr>
<td>Systematic Treatment Enhancement Program for Bipolar Disorder (STEP2)</td>
<td>European-American</td>
<td>659</td>
<td>192</td>
<td>Affymetrix Genome-Wide Human SNP Array 5.0</td>
</tr>
<tr>
<td>Thematically Organized Psychosis (TOP) Study</td>
<td>Norwegian</td>
<td>203</td>
<td>349</td>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
</tr>
<tr>
<td>Trinity College Dublin</td>
<td>Irish</td>
<td>150</td>
<td>797</td>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
</tr>
<tr>
<td>University College London (UCL)</td>
<td>British</td>
<td>457</td>
<td>495</td>
<td>Affymetrix GeneChip Human Mapping 500K Array</td>
</tr>
<tr>
<td>University of Edinburgh</td>
<td>Scottish</td>
<td>282</td>
<td>275</td>
<td>Affymetrix Genome-Wide Human SNP Array 6.0</td>
</tr>
<tr>
<td>Wellcome Trust Case-Control Consortium (WTCCC)</td>
<td>British</td>
<td>1,571</td>
<td>2,931</td>
<td>Affymetrix GeneChip Human Mapping 500K Array</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>7,481</td>
<td>9,250</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2 - Replication Rates of cAUC vs. cLogR for various Imputation Quality Scales: Here we define replication is achieving the appropriate p-value after bonferroni correction

<table>
<thead>
<tr>
<th>Imputation Quality Score</th>
<th>Analytic method</th>
<th>All SNPs</th>
<th>Skewed SNPs: (skewness &gt; 1)</th>
<th>Non-skewed SNPs (skewness &lt; 1)</th>
<th>t-test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean # of SNPs p&lt;10^{-5} in Wave A</td>
<td>Mean # of SNPs replicating in Wave B</td>
<td>Mean # of SNPs p&lt;10^{-5} in Wave A</td>
<td>Mean # of SNPs replicating in Wave B</td>
<td>Mean # of SNPs p&lt;10^{-5} in Wave A</td>
</tr>
<tr>
<td>0.3</td>
<td>cAUC 8.4364</td>
<td>0.0586</td>
<td>2.1091</td>
<td>0.0102</td>
<td>6.3455</td>
</tr>
<tr>
<td></td>
<td>cLogR 24.2909</td>
<td>0.0193</td>
<td>10.2909</td>
<td>0.0540</td>
<td>14.9273</td>
</tr>
<tr>
<td>0.8</td>
<td>cAUC 5.7818</td>
<td>0.0486</td>
<td>1.2364</td>
<td>0.0227</td>
<td>4.5636</td>
</tr>
<tr>
<td></td>
<td>cLogR 19.9818</td>
<td>0.0346</td>
<td>8.0545</td>
<td>0.0800</td>
<td>12.4545</td>
</tr>
<tr>
<td>0.98</td>
<td>cAUC 1.2364</td>
<td>0</td>
<td>0.2364</td>
<td>0</td>
<td>1.0000</td>
</tr>
<tr>
<td></td>
<td>cLogR 10.7636</td>
<td>0.0724</td>
<td>3.7818</td>
<td>0.1935</td>
<td>7.0909</td>
</tr>
</tbody>
</table>
After we calculated skewness per SNP, we divided our SNPs into the three categories we identified in the Data Reliability Section: 1) Purely genotyped, 2) Mixed imputed/genotyped, and 3) Purely imputed. Skewness was present in each of these categories. The mixed imputed / genotyped data SNPs had the widest range of skewness. This was followed closely by purely imputed. It is worth noting that skewness is present in purely genotyped SNPs as well, although the range of values is narrower.
Figure 3.2 - Skewness and shrinking p-value thresholds: As was done in the Sklar et al. paper, we considered all SNPs of Imputation Quality Score (IQS) > 0.3. We then plotted the histogram of the skewness of the SNPs below a decreasing p-value threshold (p=10^{-1} (1,1), 10^{-2} (1,2), 10^{-3} (1,3), 10^{-4} (2,1), 10^{-5} (2,2), 10^{-6} (2,3), 10^{-7} (3,1), 10^{-8} (3,2), 10^{-9} (3,3)). Though the skewness is reduced, even the most significant SNPs demonstrated some level of skewness.
Figure 3.3 - LogR Rank vs AUC Rank and cLog Rank vs cAUC Rank: Row 1 contains the two-dimensional histogram generated when one compares the rank order of SNPs according to AUC (uncorrected) versus the rank order of SNPs according to LogR (uncorrected). Column 1 is the raw bin data, and Column 2 is the same data displayed on a log scale. Figure 3.3’s row 2 contains the two-dimensional histogram generated when one compares the rank order of SNPs according to cAUC (corrected) versus the rank order of SNPs according to cLogR (corrected). Though some relationship between AUC SNP rank and LogR SNP rank was evident in row 1, its lack of symmetry demonstrates that they are ranking the SNPs quite differently. Interestingly, a very good AUC was likely to yield a consistent and very good LogR p-value, but the opposite was not true. Curiously, there was little to no relationship between cAUC and cLogR (a very dispersed signal). For what signal there was, the relationship between cAUC and cLogR was relatively symmetric, and there was a general agreement amongst the top SNPs (in this case the top 1/64th of 2.5 million).
Figure 3.4 - LogR Rank vs AUC Rank in Simulated Data Distributions – Skewness (low, medium, and high) vs. Signal (low, medium, and high): When there is a signal and skewness is present (1, 3), AUC and LogR rank SNPs quite differently, note asymmetry and widely dispersed signal. When there is a signal, but SNPs are not very skew (1, 1), SNPs are ranked roughly equivalently, note symmetry and clustering around the main diagonal. In the cause of low signal, but highly skewed data (3, 3), the result is dispersed, but relatively symmetric. In the case of low signal and low skewness, the result is more dispersed, but still somewhat asymmetric.
Figure 3.5 - LogR vs AUC rank (All SNPs vs. High Skewness vs. Low Skewness): we determined how the AUC SNP rank vs. LogR SNP rank histogram would change if we replotted them after dividing the SNPs into 'High Skewness' (|skewness|>1) and 'Low Skewness' (|skewness|<1). The 'High Skewness' graph looks even more like the high signal / high skewness simulated graph in Figure 3.4, and the 'Low Skewness' graph looks like the high signal / low skewness simulated graph in Figure 3.4.
Figure 3.6 - Evidence of Enrichment in SNPs with IQS > 0.3: In this graph, we are comparing LogR with AUC and cLogR with cAUC under the following conditions - All SNPs, \(|\text{skewness} > 1|\), and \(|\text{skewness} > 1|\). Specifically, each individual graph bins the log10 of the 11 choose 2 p-values assigned to the relationship between X in Wave A and achieving a p-value < 0.05 in Wave B. (X being log(p-value assigned to a SNP by LogR) for LogR, log(AUC/1-AUC) for AUC, log(p-value assigned to a SNP by cLogR) for cLogR, and log(cAUC/1-cAUC) for cAUC).

Strong performances by LogR and AUC in Wave A are dramatically associated with p-value enrichment in Wave B. Visual inspection alone demonstrates that this enrichment is far stronger in AUC than in LogR. This degree of enrichment seems to be independent of the skewness of the SNPs being considered.

Though both cLogR and cAUC demonstrate enrichment, cAUC is associated with far greater enrichment \((p = 2.08 \times 10^{-13})\) than cLogR. This is driven mainly by enrichment of the SNPs with \(|\text{skewness}| > 1\). With SNPs with less skewness \((|\text{skewness}| < 1)\), cAUC shows significantly less enrichment than cLogR \((1.89 \times 10^{-11})\).
Figure 3.7 - Impact of Imputation Quality Score (IQS) thresholds on the observed enrichment of more highly skewed SNPs (|skewness| > 1): In this figure, we again bin the Log10 of the 11 choose 2 p-value results of our enrichment analysis for LogR (Column 1), AUC (Column 2), cLogR (Column 3), and cAUC (Column 4). Unlike Figure 3.6, we are only considering enrichment as seen in SNPs with |skewness| > 1. Row 1 contains the results for SNPs with |skewness| > 1 AND IQS > 0.3, Row 2 contains the results for SNPs with |skewness| > 1 AND IQS > 0.8, and Row 3 contains the results for SNPs with |skewness| > 1 AND IQS > 0.98. As the IQS threshold increases, the increased enrichment observed with cAUC when compared with cLogR decreases, though it does not entirely disappear.
Figure 3.8 - High IQS and Intermediate Values: Similar to Figure 2.1, we have binned the dosage data for two SNPs with high Imputation Quality Scores (IQS = 0.97 and IQS = 0.99, respectively). There are two graphs per SNP. The top graph is a histogram of the dosage data. The bottom graph also a histogram, but with the amount per bin expressed on a log scale. We made use of the log scale so that intermediate values / outliers and sparsely represented dosage would be visible. Clearly, a high IQS does not rule the influence of intermediate values / outliers; these features are typical for many of the SNPs in this study.
4.A Overview / Rationale: As noted in the Background and Significance Section, BD is a very heritable disease whose GWAS have been plagued with replication failures despite ever increasing power. It would be tempting to decide that SNPs do not play a causative role, but polygenic models have found that greater than 1/3 of the disease variance can be explained by common SNPs.[1-3] We have already discussed the differences between genotyped and imputed data, as well as the potential problems of relying on logistic regression to analyze the imputed data, and now we would like to consider our reliance on the binary no disease / disease model of BD.

By definition, a BD GWAS relies upon a binary disease / no-disease model to classify their subjects. In order to make this type of classification, many continuous and categorical data points are collected, synthesized, but ultimately, discarded. Even gross categories are merged in these analyses. In the Sklar et al 2011 PGC-BD paper, BDI, BDII, and SABD (schizoaffective bipolar disorder) are treated interchangeably. Now, imagine for the moment that BD is not just a single disease with a common genetic signature, but rather an umbrella comprising BD Subtypes 1 though N, with each subtype having its own distinct genetic signature. If this were indeed the case, then SNPs having a large effect when taken in the context of let us say Subtype 13 might appear to have small to no effect when analyzed en masse with the other BD Subtypes, i.e. 1 to 12 and 14 to N.

We believe that this scenario is possible, and that inherent heterogeneity within the current phenotypic classification system may be contributing to an inability of BD
genetic studies to achieve their goals. In fact, as described in detail within Appendix A, the BD literature offers up two forms of support for the existence of such genetically-relevant subphenotypes or subtypes: (1) Certain BD subphenotypes appear to be heritable, and (2) Some genome-wide significant / replicable SNPs have been identified after using a subphenotype to define the disease state. The existence of heritable subphenotypes alone would be sufficient to raise questions about the disease / no-disease model. The fact that some of these subphenotypes appear to have unique genetic signatures cemented our commitment to questioning this most fundamental assumption. Our goal is to pursue this hypothesis by searching for genetically relevant phenotypic heterogeneity.

When we began this work, a systematic, data-driven approach to uncover subgroups driven by interactions had not yet been undertaken. The closest we had found in the literature was the work by Niculescu et al., who observed much within-group heterogeneity in BD phenotype clustering, e.g. motor measures, mood scales. We sought to do what they were unable to do because they lacked genetic data.[126] It was our intention 1) to use machine-learning algorithms to systematically identify phenotypic subgroups within the population and 2) to determine which subgroups were clinically relevant through replication. The NIMH sample (N=4335) we analyzed included a broad range of phenotypic information on diagnosis, symptoms, and course of illness (Table 4.1). As a result, we were in the position to consider a wide range of phenotypes and their interactions across a large number of subjects and proceeded as follows.

There was/is not a consensus as to how best to identify phenotypic subgroups. It can be helpful to think of machine learning as a means of finding a low-dimensional representation of a high-dimensional input. Most machine learning approaches rely upon training data or previously classified data, e.g. regression analysis and decision trees.
Though we are fairly confident that heritable traits defined by the interaction of phenotypic traits likely exist in BD, it is difficult to know a priori how they are defined so it is not possible to provide training sets of phenotypic subgroups and thus we must use unsupervised machine learning methods.

There are a range of unsupervised machine learning methods. Some use a distance metric to group subjects while others are driven by how well their system is able to reconstruct the original high-dimensional solution from its low-dimensional representation. To some degree, they all require the user to decide how many significant subgroups are present in a dataset. Ultimately, the guideline researchers most often turn to is whether the identified subgroups have biological significance.

We thus proposed to: 1. Apply multiple unsupervised machine learning methods to my phenotypic sample, 2. Apply a heterogeneity test to determine which SNPs behave differently across my subgroups, i.e. deserve further inspection, 3. Compare the results of GWAS before and after correction for the subgroups, 4. Determine whether phenotypic heterogeneity is associated with replication failure, and determine whether using a homogenous phenotypic subgroup will improve the performance of polygenic modeling, and 5. Seek to replicate my findings using the PGC-BD phenotype / genotype data.

4.B Research Design and Methods

4.B.1 Dataset Description

4.B.1.a NIMH Phenotypic Data: Through the BiGS, we have access to the phenotypic data collected by the NIMH Bipolar Genetics Initiative (NIMHBGI). As of 2009, NIMHBGI had performed the Diagnostic Interview for Genetic Studies (DIGS) on 4335 subject with BDI, BDII, SABD (schizoaffective bipolar disorder). (See Table 4.1 for overview and Table 4.2 for further details)
4.B.1.b NIMH Genetic Data: Because of the efforts of the BiGS consortium as well as the Pritzker consortium, 2682 of these subjects as well as 1653 controls were genotyped (~2.4 million imputed SNPs).[3, 34] The imputation was performed using BEAGLE and CEU (Utah residents with ancestry from northern and western Europe) HapMap2.[125]

4.B.1.c Additional Subjects with phenotypic data: Through the PGC-BD we had access to an additional 5989 BD genotyped cases and 7998 genotyped controls.[5] A complete description of the genotyped dataset is in Sklar et al.[31] The phenotypic information for the PGC-BD was collected two years after we ran the first stage of our analysis. It included gross categorical phenotypic information, e.g., psychotic Y/N, rapid cycling Y/N, and suicidal Y/N, for a subset of the 7998 genotyped. The number of subjects with phenotypic information varied from topic to topic. None of these additional PGC-BD subjects underwent the Diagnostic Interview for Genetic Studies (DIGS) as part of their evaluations so a direct replication set was never an option.

4.B.2 Methods

4.B.2.a Aggregate Analyses of NIMH phenotypic Data: First, we set out to identify phenotypic subgroups within the 4335 NIMH subjects DIGS data.

4.B.2.a.1 Machine Learning / Dimension Reduction Methods for DIGS Analysis: We applied several techniques, including PCA, Autoencoder (a Deep Belief Neural Network), and t-distributed stochastic neighbor embedding (t-SNE), to the DIGS binary questions in order to identify clinical subgroups in BD. We focused our attention on the binary questions because questions like age-at-first hospitalization are problematic to analyze in those subjects who were never hospitalized. Missing data from interviewer error was randomly replaced.

There were many algorithms in the literature in 2009. We limited ourselves to those that were well-vetted, thought to perform well, and readily available. Since PCA is
commonly used to reduce dimensionality, it was chosen to serve as a baseline by which to assess the other measures. Autoencoder was selected because (i) its code was publicly available, (ii) it had been widely used in other fields, (iii) it was well-suited to binary data, (iv) its choice of low-dimensional representation was driven by how accurately the original data can be reconstructed, (iv) it was proven to outperform PCA for many natural structured high-dimensional data sets, (v) it did not cluster the data for the user, and (vi) it could be used to produce a map between high-dimensional input-space and low-dimensional output which can then be applied to new subjects.[127] t-SNE was selected because (i) it was more successful than other methods at segregating nearby clusters of high-dimensional data, (ii) it was currently the best tool for representing a variety of high-dimensional data-sets (handwriting, etc.), (iii) unlike many other visualization tools t-SNE aimed for a unique low-dimensional representation of data, (iv) it did not perform clustering – that was done later (on the low-dimensional output) using whatever methods worked best, (v) most of the 'error' in the fit came from topological inconsistencies in closely knit groups of points, and (vi) due to the definition of 'similarity', the separation between distant groups of points was represented faithfully.[128]

4.B.2.a.2: Subgroup Assignment: PCA, t-SNE, and Autoencoder each projected every subject into a single position in 3D space. Visual inspection was used to assess the presence or absence of subgroups. For those that were clearly present, Matlab’s built-in clustering function was used to assign subjects to subgroups.

4.B.2.b: Aggregate Analysis of Genomic Data: Once our subgroups were defined using the 4335 NIMH subjects’ DIGS data, we turned to analyzing the 2682 subjects and 1653 controls with genomic data (~2.4 million mixed genotyped/imputed SNPs).
4.B.2.b.1 Heterogeneity test: We checked each of the 2.4 million SNPs to determine whether there is sufficiently heterogeneous behavior at the polymorphism to merit further analysis by applying Morris et al.’s Subphenotype Heterogeneity Test (SHT). Essentially, this metric computed multinomial regression coefficients under two paradigms, 1) BD Subgroupi vs. No-disease and 2) BD vs. No-disease and then determined if the sum of the differences over all of the SNPs was significantly different according to a Chi-squared distribution.[4] We selected our p-value threshold (0.1) once when we knew the size of our smallest subgroup (~90 subjects) because the SHT would have more trouble detecting heterogeneity driven by a small subgroup with aberrant behavior.

4.B.2.b.2: Heterogeneity ~ (1 / Replication) Analysis: We performed a logistic regression analysis to determine if the heterogeneity (using ln(SHT p-value) as a proxy for level of heterogeneity) was negatively associated with replication.

4.B.2.b.3: Assess Heterogeneity of Replication SNPs in PGC-BD: PGC-BD sought to replicate 34 SNPs. Only two were found to replicate after correcting for multiple comparison.[31] We checked whether the remaining 32 pass Morris et al. Subphenotype Heterogeneity Test (SHT) under the conditions of the various phenotypic subgroups we identified.

4.B.2.c: Subgroup Analysis: We then examined how each of the 650 binary questions behaved at the subgroup level. After identifying the phenotypic traits that defined our subgroups ((+/-) 'mania with psychosis', (+/-) 'depression', (+/-) 'panic attacks'), we focused our efforts on the most phenotypically distinct subgroup, Subgroup 10, whose definition was driven mainly by two variables, i.e., ‘depression’=No, ‘psychosis with mania’=Yes. The PGC-BD was not optimistic that panic attacks data would be captured and we wanted to pursue a subgroup we could reasonably hope to create a replication set. For S_{10}, we pursued the following:
4.B.2.c.1: Genetic Association Analyses: We treated Subgroup_{10} (S_{10}) as if it were a new form of bipolar. In all of the analyses which follow, we compared BD S_{10} with controls. For those SNPs who passed the heterogeneity test, we analyzed the observed allele counts or imputed allele dosages using logistic regression assuming an additive genetic model. We used White’s sandwich estimator to adjust the estimated variances to account for sibling relationships. For those SNPs with p-value < 10^{-4}, we used logistic regression to determine their p-values on a per study basis. Since we were hypothesizing that our efforts would yield genetically homogeneous subgroups, we set aside for replication only those SNPs with a signal in more than one study. As uncommon variants are just that, we took the following approach: 1) If the minor allele frequency (MAF) < 0.05, we considered those SNPs with two per study p-values < 0.1 and the third study was trending in the correct direction. 2) If MAF>0.05, only those SNPs who achieved p-values < 0.1 across the three studies were set aside for replication.

4.B.2.c.2: Common Variant Polygenic Modeling: We performed polygenic modeling on S_{10} using the approach Purcell et al. applied in schizophrenia and BD.[1] The model requires a set of genotyped (rather than imputed) SNPs in linkage equilibrium. Only 142,380 SNPs were originally genotyped in both the BiGS and Pritzker samples. After pruning these SNPs to linkage equilibrium (r^2 < 0.5) using the “—indep-pairwise” command in PLINK with a sliding window of 50 SNPs and a 5 SNP step, 114,009 remained. S_{10}’s samples had to be split randomly into two separate sets (Set 1, Set 2) as we did not and, ultimately, do not have a replication data set (see below) available to us. We then used a variety of p-value thresholds (P_T) to select the SNPs we used to generate the inputs required by PLINK’s —score function, i.e. SNPs with p < p_T were used. Once we had our SNP list, we used Set 1 to calculate the log odds for each SNP.
These log odds were then input into PLINK’s –score function for evaluation on Set 2. By varying $p_T$, we sought to identify which $p_T$ identified the greatest variance in diagnosis.

Purcell et al. explained the largest variance in schizophrenia diagnoses by using a $p_T=0.5$. This suggested that schizophrenia and BD are in large part driven by many SNPs of small effect. Our assumption was that if we had achieved a cleaner signal by analyzing a phenotypically homogeneous subgroup then we would need a smaller $p_T$ to explain the greatest degree of diagnosis variance.

4.B.2.c.3: Pathway Enrichment Analysis: Once we identified the relevant genes using PLINK’s annotation function, we then used GeneGo to determine the pathways implicated.

4.B.2.c.4: Replication in the independent PGC-BD Sample: At the time that we began these analyses, we were pleased that the subgroups we originally identified were defined by a few gross categories (e.g. ‘psychosis with mania’, ‘depression’, ‘panic attacks’). We then sought to replicate our findings using the coarser phenotypic information available through PGC-BD (e.g., ‘psychosis’, ‘depression’, ‘panic attacks’).

4.C Results

4.C.3.a Aggregate Analyses of NIMH phenotypic Data:

4.C.3.a.1 Machine Learning analysis of DIGS data, including subgroup assignment: PCA, Autoencoder, and t-SNE were all used to project the 4335 subjects’ 650 binary DIGS questions into a three-dimensional space. Visual inspection found that t-SNE achieved the greatest phenotypic separation. Figure 1 shows that PCA and Autoencoder were also identifying a similar signal.

Since assigning subjects to groups in 3D space manually can be difficult, we used Matlab’s (version R2010a) built-in clustering function to define the groups that were apparent through visual inspection. Of the 10 clusters Matlab identified, only five had
greater than 100 subjects, (labeled 1, 2, 5, 6, 10 in the Figure 4.1). We decided to focus on these in hopes of getting more reliable statistical outcomes. As noted in the dataset description section, only 2682 of these subjects had corresponding genetic data. This left us with one subgroup with less than 100 subjects. As noted in passing above, we then examined how each of the 650 binary questions behaved at the subgroup level (Figure 4.2). What drives these subgroups appears to be (+/-) ‘mania with psychosis’, (+/-) ‘depression’, (+/-) ‘panic attacks’ [Table 4.3].

4.C.3.b: Aggregate Analysis of Genomic Data:

4.C.3.b.1 Heterogeneity test (SHT): We used a threshold for our heterogeneity test of p-value <0.1 because some of our subgroups were quite small. ~275,000 SNPs passed the heterogeneity test which represents some enrichment as it is slightly more than the expected when analyzing 2.4 million SNPs.

4.C.3.b.2 Heterogeneity ~ (1 / Replication) Analysis: We ran a logistic regression of the natural log of the SHT p-value and replication at p<0.05 level in order to test the hypothesis that the probability of association at P < 0.05 in the ‘PGC-BD’ sample was negatively associated with the heterogeneity observed in the BiGS/Pritzker. We found that we were more likely to replicate associations at P<0.05 when there was less DIGS subgroup heterogeneity. (P=6.04x10^{-5}). As we are not analyzing SNPs individually, we do not need to correct for multiple-hypothesis testing.

4.C.3.b.3 Assess Heterogeneity of Replication SNPs in PGC-BD:

As noted above, only two of the 34 SNPs were significant after one corrects for multiple comparisons. Of those 32, only four passed the SHT (p<0.05). The two top SNPs which had achieved genome-wide significance in the first arm were not amongst these four. 4 represents 2.5x the amount one would expect by chance.
4.C.3.c: Subgroup Analysis:

4.C.3.c.1 Genetic Association Analyses of Subgroup 10: Subgroup 10 (S_{10}) is defined by the absence of depression, mania with psychosis, and panic attacks. The absence of depression means that Subgroup 10 is a subset of unipolar mania.

There were 361 SNPs with p-value < 10^{-4}. Figure 4.3 shows how many SNPs / Regions met our criteria and how they are organized within output tables (Table A.4 contains those with MAF < 0.05 or MAF >0.95. Table A.5 contains those with 0.05 < MAF < 0.95. Those who are underlined had p-values < 0.05 across all 3 studies. These Tables are in Appendix A)

Of the 22 gene / regions which satisfy the criteria 1) p-value < 10^{-4}, 2) control frequency < 0.05, 3) case frequency > control frequency, 4) p-value < 0.05 in 2+ studies, and 5) varying in the correct direction in all three studies, only two have been widely studied and appear to have no neurological link. Ten have been noted in a psychiatric context: Five in schizophrenia (XKR4, within 500kb of MAGL, GRIN2B, within 30kb of CREB1, within 300kb of NOVA-1), three in bipolar (ITPR1, GRIN2B, within 30kb of CREB1), two in autism (within 2Mb of HSD17B4, SLC9A9, and one in OCD, Tourette’s (within 2Mb of SLITRK1,*SLITRK6). Of note, one is in the lithium sensitive phosphatidyl inositol pathway (ITPR1), two are involved in circadian rhythm (within 30kb of CREB1, within 500kb of MAGL) and another is a NMDA receptor (GRIN2B). Five have a clearly articulated role in neuronal function, but have not previously been associated with a psychiatric disorder (DMXL2, within 200kb of MOCS2, within 200kb of KCNV1, within 200kb of MYCB2, RGS6, FANCD2). One is expressed in the brain, but the function is unknown. The rest are of unknown function (within 20kb of MCOLN2, CNBD1, within 200kb of C15orf59, within 500kb of ANKRD55). [Details in Table A.4]
Of the 30 gene / regions which satisfy the criteria p-value < 10^{-4}, 0.05 < control frequency < 0.95, and p-value < 0.1 across all three studies, only two have been widely studied and appear to have no neurological link. Of note, one may regulate voltage-dependent calcium channels (CACHD1), another is involved in glutamatergic transmission (ZDDHC8), two in the lithium sensitive phosphatidyl inositol pathway (PIK3R1, DGKG), another is an accessory subunit of NMDA-receptor (NETO1) and important to GRIN2A (GRIN2B identified in previous section), another may play a role in the microtubule-dependent transport of the GABA-B receptor (JAKMIP1). Eleven had been linked previously to a psychiatric disorder: Three in bipolar (within 30kb of GALNTL4, within 200kb of PIK3R1, within 2Mb of ADCY8), two have had comparable genes previously identified in bipolar (within 200kb of DGKG, within 2 Mb of KCNQ3), five in schizophrenia (SIRPA2, within 200kb of NOVA-1, within 100kb ZDDHC8 & RTN4R, within 200kb PIK3R1, within 30kb of WWC1), and two in other psych disorders (FAM155A - ADHD, depression; within 300kb of PDE2A – depression). Seven have a clearly articulated role in neuronal function, but have not previously been associated with a psychiatric disorder (CACHD1, FAM134B, JAKMIP1, NETO1, within 2Mb of SNAI2 & SNTG1, within 50kb of CLIC4, within 20kb of SHMT1). Six are expressed in the brain, but don’t have a clear neurological function. Two have been noted in non-psychiatric brain disorders. Two are not yet understood.

PGC-BD sought to replicate 34 SNPs. Only two SNPs remained significant after correcting for multiple comparisons. These SNPs were all of low effect with odds ratio ranged from 0.90-0.94 and 1.05-1.13. The odds ratios of the top subgroup 10 SNPs ranged from 0.19-0.58 and 1.8-6.53. Because subgroup 10 is small, the 95% confidence intervals are rather large. If one wishes to be conservative and use the upper and lower bound of the 95% CI, one gets 0.04-0.91 and 1.02-31.3. [Table A.4 and A.5]
Unfortunately, PGC-BD did not list their CI intervals so a direct comparison cannot be made. Despite this, one should note that the subgroup 10 SNPs typically have a large effect within this subpopulation, and even at their worst estimate these SNPs have an effect similar to the PGC-BD SNPs.

4.C.3.c.2 Common Variant Polygenic Modeling: Figure 4.4 shows the results from polygenic modeling of S₁₀. The largest percentage of diagnoses variance (0.15%) was explained with a p-threshold ($p_T$) of $7.5 \times 10^{-4}$. The associated p-value is 0.03. Note that the total variance explained in this analysis is lower than the total variance explained by Purcell et al. This may be because, unlike the Purcell et al. paper, 1) our trial/test set for $S_{10}$ are small (N=42/46) and 2) our genotyped SNP list is relatively small. We were encouraged though that a small $p_T$ is capturing the majority of the diagnoses variance. This may suggest that we have identified a phenotypically homogeneous BD subgroup with clear genetic roots.

4.C.3.c.3: Pathway Analysis: We selected one SNP per significant region for a total of 52 SNPs. 21 were not within 20kb of a gene. We ran GeneGo pathway enrichment analysis on this set of genes. GRIN2B (glutamate receptor, ionotropic, N-methyl D-aspartate 2B) and ITPR1 (inositol 1,4,5-triphosphate receptor, type 1) play many roles in neurophysiology. As a result, the following pathways were implicated: 1) NMDA-dependent postsynaptic long-term potentiation in CA1 hippocampal neurons (3/80, $p=9.2 \times 10^{-5}$), 2) Dopamine D2 receptor transactivation of PDGFR in CNS (2/26, $p=4.0 \times 10^{-4}$), 3) nNOS signaling in neuronal synapses (2/29, $p=5.0 \times 10^{-4}$), 4) Calcium signaling (2/45, $1.2 \times 10^{-3}$), and 5) Alpha-2 adrenergic receptor regulation of ion channels (2/47, $p=1.3 \times 10^{-3}$).

CREB1, PIK3R1 (Phosphatidylinositol 3-kinase regulatory subunit alpha), and DGKG (diacylglycerol kinase, gamma) are each the nearest gene to a significant region, within 30kb, 200kb, and 200kb respectively. CREB1 also plays a role in both NMDA-
dependent postsynaptic long-term potentiation and calcium signaling. PIK3RI also plays a role in NMDA-dependent postsynaptic long-term potentiation. PIK3RI and ITPR1 are important in Inositol-1,4,5-trisphosphate (IP3) signaling. DAG signaling and IP3 signaling are intertwined. DGKG regulates DAG concentrations which is a precursor of phosphatidylinositol (PI) and phosphoinositide (PIP, PIP\(_2\), PIP\(_3\)). TRAC is also a gene in this pathway so it would appear that the phosphatidylinositol pathway is also enriched.

4.C.3.c.4 Replication in the independent ‘PGC-BD’ Sample: At the time that we proposed this analysis, the phenotypic data was still being collected by the ‘PGC-BD’. As we understood it, they were collecting depressed Y/N, psychosis Y/N, panic attacks Y/N. Using what we believed to be comparably coarse categorizations (e.g. ‘depressed’ Y/N, ‘psychosis with mania’ Y/N, ‘panic attacks’ Y/N / ‘panic disorder’ Y/N), 94% / 100% of SNPs stayed significant. Given this, we were optimistic that the coarse categorical information available through the PGC-BD could be used to replicate the SNPs we identified using the DIGS instrument.

As noted above, none of the additional ‘PGC-BD’ subjects were evaluated using the DIGS. The criteria for the ‘PGC-BD’ psychosis subphenotype varied with study\(^3\). A

\[\begin{array}{|c|l|}
\hline
\text{WTCCC} & \text{Subjects were rated for the lifetime occurrence of psychosis based upon the OPCRIT item checklist and the Bipolar Affective Disorder Dimension Scale (BADDS)} \\
\hline
\text{Norway/Andreassen} & \text{Positive and Negative Syndrome Scale (PANSS) score above 3 on any of the 7 positive symptom criteria, e.g., delusions, disorganization, hallucinations} \\
\hline
\text{UCL/Gurling} & \text{If schizophrenic or bipolar psychotic symptoms noted in SADS-L or OPCRIT or case notes} \\
\hline
\text{GSK/Muglia} & \text{If psychotic symptoms were noted either during the worst depressive or the worst manic episode} \\
\hline
\text{Edinburgh} & \text{presence of delusions or hallucinations noted in SADS-L or rate symptoms recorded in hospital records} \\
\hline
\end{array}\]
uniform set of criteria were used to define the depression subphenotype variable. Unfortunately, panic attack information was not captured for the majority of our intended replication subjects.

Once we were given the data, we determined how well these two categorical variables (‘psychosis’ Y/N, ‘depression’ Y/N) could be used to reidentify our Subgroup 10 within our original sample with complete DIGS data (GAIN, TGEN, Pritzker). ‘Psychosis’ in this case based on Module K (psychosis review) of the DIGS, specifically DIGS4 K.1.i - "a psychotic symptom is present persistently for one day or intermittently for 3 days". ‘Psychosis with mania’ in this case was also based on Module K (psychosis review) of the DIGS, specifically DIGS4 K.48 - "Did (Delusions or Hallucinations) ever occur when you were feeling extremely good or high, or when you were feeling unusually irritable?". Unfortunately, there appears to be a big difference between ‘psychosis’ = ‘No’ and ‘psychosis with mania’ = ‘No’, resulting in a 100% increase in false negatives (FN). (Table 4.4)

Given the rare nature of this particular phenotype, this was an unacceptable level of false negatives. We concluded that we did not have the phenotype data through the PGC-BD necessary to create a replication dataset. Without panic attacks information, we were unable to shift our focus to any of our other subtypes. We did attempt to solicit ‘psychosis with mania’ data, but were unsuccessful in collecting that data. As a result, we were unable to determine if our signal would have replicated.

---

4: ‘Yes’ was coded if there was one or more episode(s) of depression and/or an age of onset for depression was scored and/or unipolar mania was scored as negative. If unipolar mania data was scored as positive, but one of the first two criteria indicated the presence of depression, depression was coded ‘yes’.
4.D Discussion:

4.D.4.a DIGS Heterogeneity and PGC-BD replication: It would appear that we have captured a non-negligible fraction of the genetically-relevant heterogeneity. We found that DIGS heterogeneity, as assessed by Morris et al.’s Subphenotype Heterogeneity Test (SHT), is inversely related to replication at a 0.05 level (p=6.04x10^{-5}), i.e., the more DIGS heterogeneity present, the less likely a SNP was to replicate. That being said, when we checked the specific relationship between the top 34 SNPs selected in PGC-BD Sklar et al. paper, of which 32 failed to replicate, and SHT, only four of the 34 showed significant DIGS subphenotypic heterogeneity (p_{SHT} < 0.05). We found it reassuring that the two SNPs which did ultimately replicate were not amongst these four. Given the inverse relationship between SHT and replication and the fact that this was 2.5x what one would expect by chance, the replication of some or all of those four SNPs may have been compromised by unaccounted for DIGS heterogeneity in the wave 1 and replication datasets.

If we had captured all of the genetically-relevant heterogeneity and there were no other reasons for replication failure (e.g., differences in genotype / imputation data, skewness/outliers influencing logistic regression), it might be safe to conclude that the other 28 which did not replicate were false positives. Given that both Meier et al.[125] and Kerner et al.[7] identified different forms of seemingly genetically-relevant heterogeneity (negative mood delusions, relationships to substance abuse +/- alcohol abuse), such an assertion would be premature. To further support our point, let us consider the top SNP in the first PGC-BD arm, rs10994397 (ANK3). Even though it failed to be replicated (p-value>0.1) by the PGC-BD, it has since been replicated in an independent East Asian BD sample.[26] Given this, it seems unlikely that it plays no role in the pathogenesis of bipolar disorder, but that its replication failure was a function of
noise in the system. If 1) this noise is a function of phenotypic heterogeneity, 2) all of the relevant phenotypic information was captured by the DIGS instrument, and 3) our methods performed optimally, then this SNP should have passed the SHT for our DIGS subgroups. Unfortunately, it did not. This, considered in conjunction with the work of Meier et al.[7] and Kerner et al.[6], leaves us to conclude that the DIGS data in combination with our methods is, at best, an incomplete picture of the phenotypic heterogeneity present in BD.

4.D.4.b Subgroup 10 (S_{10}) - Unipolar mania without panic attacks/psychosis with mania: Subgroup 10 (S_{10}) polygenic model’s small p-threshold, 7.5x10^{-4} compared with 0.5 in schizophrenia[1], purports a much tighter link between significant SNPs and subgroup-control status. The work of Visscher’s group, which introduced and applied polygenic modeling to complex diseases, indicates that such a result would reflect a process driven by a relatively small number of genes with at least moderate effect sizes.[57, 129, 130] This is certainly consistent with the larger odds ratios observed in the S_{10} / no disease model when compared with the disease / no disease model. This further strengthens our position that phenotypic heterogeneity may be obscuring the genetic signal responsible for BD and that using machine learning to identify phenotypically homogeneous subgroups can reduce that noise.

Previously, we discussed the difficulties involved in unsupervised machine learning subgroup selection and described how the identification of a biological basis for a subgroup would be seen as reasonable evidence of validity. Viewed in that context, S_{10} seems promising. First, two of the three traits it is defined by (psychosis, co-morbid panic disorder) are believed to be heritable.[131, 132] Second, we noted that there were many ties between the genes/regions identified and the brain. Third, and likely more convincing, we found that the top pathways flagged by GeneGo made biological sense.
They were 1) NMDA-dependent postsynaptic long-term potentiation in CA1 hippocampal neurons, 2) calcium signaling, and 3) the phosphatidyl inositol pathway. The NMDA-receptor has long been implicated in schizophrenia and psychosis. It has also been studied in the context of bipolar though it has experienced replication failures.[133, 134] Calcium channels and by extension calcium signaling have recently been garnering support in the psychiatric community. CACNA1C (L-type calcium channel) has so far been identified in the PGC-BD, PGC-schizophrenia, and PGC-Cross-disorder GWAS.[135] [5] The phosphatidyl inositol pathway has long been known to be lithium sensitive.[136] These pathways were identified only using the genes implicated by SNPs that passed the heterogeneity test. It is possible that further elements in these pathways as well as additional pathways would declare themselves if we were to consider all SNPs that meet our significance criteria for S₁₀.

4.D.4.c Benefits of Machine Learning / Dimension Reduction: Machine learning not only made it possible to identify this potentially useful group, but it did so very efficiently. The DIGS instruments 650 binary questions are divided up amongst 11 categories, e.g. depression, mania, anxiety, gambling, suicide. We noted above in the PHE-WAS section that groups defined by interactions of traits were likely important. Without the machine learning algorithms, we would have had to perform 11-choose-3 (165) genome-wide analysis in order to guarantee that we would have checked the five subgroups. Unfortunately, after running so many tests, it is very unlikely that a reviewer would have been particularly receptive to our findings.

4.D.4.d Subphenotype Analyses – Current limitations: Our genetically-relevant subgroups were defined by ‘depression’, ‘psychosis with mania’, and ‘panic attacks’. Meier et al. found a replicable SNP of genome-wide significance associated with the ‘negative mood delusions’ (poverty, guilt, nihilism) phenotype.[7] Kerner et al. found
SNPs of genome-wide significance using (1) Psychosis and/or substance abuse in the absence of alcohol dependence and (2) Alcohol dependence and other co-morbidities.[6] This suggests that there is a great deal of genetically-relevant phenotypic heterogeneity underlying the BD phenotype. Since t-sne, latent class analysis, and factor analysis all identified genetically-relevant phenotypic subgroups, it seems reasonable to conclude that, given the correct inputs, there are many ways to identify relevant subphenotypes.

Therefore, the larger problem seems to be that these methods can only find patterns within the available phenotypic data. As we found out, this information is collected inconsistently and seems to be limited to self-report instruments in the larger studies. Self-report instruments are clearly fallible, e.g., answers subject to current outlook of interviewee – mania vs. depression, interviewer-to-interviewer variability. Biomarker data, such as blood pressure or imaging data, would be more objective, but, not only did none of our various groups have access to biomarker data, but also, none of us would necessarily know a priori what biomarker data to collect.

Since each of our respective groups was hoping to identify genetically homogenous subgroups by eliminating phenotypic heterogeneity and the phenotypic information available is limited at best, it would make more sense just to seek out these genetically homogeneous subgroups directly. Biclustering methods, i.e., methods that can identify subsets of SNPs and cases which are different than controls, are a natural fit.

Unfortunately, biclustering methods are not quite up to such a task. 1) As we showed in our section on quantifying the difference between genotyped and imputed data, performing such an analysis on a mixed genotyped / imputed data is problematic because there is so much non-disease-related signal. If a biclustering method cannot be
corrected for covariates like MDS components and study of origin, it would have no means by which to find the true signal in the noise. At this time, there are no published biclustering methods with covariate correction. 2) Supposing your dataset did not rely upon imputation, i.e., one has a large dataset where each subject was genotyped on the same platform or has been fully sequenced, one will need to balance computation time against the total number of SNPs considered. Certainly, there are no programs out there which could handle a set of 20K subjects by 600K SNPs. To shorten your SNP list, the two most obvious approaches are 1) to create linkage equilibrium sets or 2) to focus on top SNPs as defined by covariate corrected association analysis. The difficulty with the first is that much of the signal may be lost depending on how dramatically one has to thin the SNP set to reduce computation time. The difficulty with the second is that you are forced to rely on a binary phenotype to thin one’s SNP list and by the very nature of biclusters the SNPs associated with any particular bicluster may not be significant under the disease / no-disease paradigm.

Covariate-corrected biclustering approaches are in the pipeline and so are approaches that can handle the big datasets which will inevitably be generated as the costs of genotyping and whole-genome-sequencing drop. At some point, we will be able to move from subjective data directly to analyzing the genome. Once we known the genotypic relationships, the phenotypic correlates may become more obvious though potentially redundant.

4.D.4.e Inability to run replication analysis: As we described in the results section, the information captured by the absence of psychotic symptoms is very different than the absence of psychotic symptoms when experiencing a manic episode. Since the PGC-BD captured the former rather than the latter, we were not able to test replication. As
our phenotypic subgroups do appear to have some relationship with the genome, we will continue to seek out a replication dataset.

4.D.4.f Failure to publish: We did not publish our initial analyses when we completed them as we thought replication was essential to make a compelling point in bipolar disorder. Given the fact that a similar paper was published without a replication arm, we may have been too conservative in our approach. Because 1) Kerner et al. (without a replication arm) and Meier et al. (with a replication arm) have used dimensionality reduction to identify SNPs of genome-wide significance, and 2) Meier et al.’s case demonstrated dimensionality reduction can be leveraged to identify SNPs which replicate, we feel we would need to demonstrate replication to justify submitting this work for publication.

4.D.4.g General: As we noted above, replication failures in BD may be because genotyped and imputed data should not be treated as interchangeable or because logistic regression is not well-suited to analyze mixed genotyped/imputed data. Based on our work and that of Meier et al.[7] and Kerner et al.[6], there is a great deal of phenotypic heterogeneity underlying the BD phenotype. Though we were not able to demonstrate it ourselves, Kernel et al. found that this heterogeneity can be leveraged to identify a replicable genetic signal.[6] Making exhaustive phenotype collection a part of any GWAS would be appropriate. Once we are capable of implementing biclustering methods which can correct for covariates and handle large numbers of SNPs, it would make the most sense to directly analyze the genome for homogeneous subgroups.
Table 4.1 - DIGS Details: The NIMH sample (N=4335) we analyzed included a broad range of phenotypic information on diagnosis, symptoms, and course of illness

<table>
<thead>
<tr>
<th>Test</th>
<th>Phenotypes</th>
</tr>
</thead>
</table>
| DIGS (Structured interview)[137, 138] | - Diagnosis  
- Course of the illness  
- rapid cycling, psychosis, history of suicidality, inter-episode functioning  
- Chronology of psychotic and mood syndromes (e.g. AAO)  
- Comorbid diagnosis & characteristics |

All variables are nominal or numerical (Integers)

Table 4.2 - More Detailed Description of Phenotypic Scales used in BiGS GWAS

<table>
<thead>
<tr>
<th>Test</th>
<th>Details</th>
</tr>
</thead>
</table>
| DIGS (Diagnostic Interview in Genetic Studies) (685296, 7944874) | Structured Clinical Interview intended to capture psychopathology  
- It has polydiagnostic capacity, enables a detailed assessment of  
----Course of the illness  
----Chronology of psychotic and mood syndromes  
----Comorbidity  
- Allows for algorithmic scoring of symptoms (Binary)  
- Test-retest reliabilities were shown to be excellent for MDD (0.94), bipolar (0.95), schizophrenia (0.75). Schizoaffective disorder has a Kappa value of 0.31, but its bipolar and depressed subtype had good specificity |

Table 4.3 - Participants per subgroup and Distinguishing characteristics: Out of the 650 DIGS questions with binary data, these three individual questions best re-identified the subgroups defined by t-SNE

<table>
<thead>
<tr>
<th>Subgroup</th>
<th># without siblings</th>
<th># with siblings</th>
<th>Depression</th>
<th>Psychosis w/ Mania</th>
<th>Panic attacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>956</td>
<td>1003</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>1084</td>
<td>1205</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>179</td>
<td>185</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>183</td>
<td>188</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>88</td>
<td>91</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 4.4 - Recreating S10 using individual questions – ‘Psychosis’ vs. ‘Psychosis with Mania’: Here we show how accurately we can recreate Subgroup 10 (S₁₀) as defined by t-SNE’s analysis of the 650 binary questions using ‘Psychosis’ vs. ‘Psychosis with Mania’. Unfortunately, using ‘Psychosis’ (the subphenotype available through the PGC-BD) rather than ‘Psychosis with Mania’, dramatically increases the FN rate.

<table>
<thead>
<tr>
<th>Subject Description</th>
<th>TP</th>
<th>FN</th>
<th>FP</th>
<th>TN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depression=No, Psychosis=No</td>
<td>31</td>
<td>58</td>
<td>8</td>
<td>2577</td>
</tr>
<tr>
<td>Depression=No, Psychosis with mania = No</td>
<td>60</td>
<td>29</td>
<td>13</td>
<td>2572</td>
</tr>
</tbody>
</table>
Figure 4.1 - Subgroup Identification as identified by PCA (1,1), Autoencoder (1,2), and t-SNE (1,3): We used PCA (Principal Component Analysis), Autoencoder, and t-SNE to project the 4335 subjects' 650 binary DIGS questions into a three-dimensional space. Visual inspection found that t-SNE achieved the greatest phenotypic separation, but visual inspection shows that PCA and Autoencoder were also identifying a similar signal. Since assigning subjects to groups in 3D space manually can be difficult, we used Matlab's (version R2010a) built-in clustering function to define the groups that were apparent through visual inspection. Of the 10 clusters Matlab identified, only five had greater than 100 subjects. We labeled these 1,2,5,6,10.

Figure 4.2 - t-SNE Subgroups and their performance on the 650 binary DIGS questions: This is a representation of the 650 binary DIGS questions by psychiatric category. Red = percent yes, Blue = percent No. The Green fades in proportion to much the question varies across subgroups.
Figure 4.3 - Subgroup 10 (S₁₀) SNP Selection Criteria: There were 361 SNPs with p-value < $10^{-4}$. If these, 332 had an Information Quality Score (IQS) > 0.3. Of the SNPs with a minor allele frequency (MAF) greater than or equal to 0.05, 122 SNPs (30 regions) had p < 0.1 across all three studies. Further, 47 SNPs (17 regions) had a p<0.05 across all three studies. Of the SNPs with a minor allele frequency (MAF) less than 0.05, 26 SNPs (22 regions) had p < 0.05 in 2+ studies (all in the correct direction). Further, 6 SNPs (6 regions) had a p<0.05 across all three studies.
Figure 4.4 - Polygenic modeling of $S_{10}$: The largest percentage of diagnoses variance (0.15%) was explained with a p-threshold ($p_T$) of $7.5 \times 10^{-4}$. The associated p-value is 0.03. Note that the total variance explained in this analysis is lower than the total variance explained by Purcell et al. This may be because, unlike the Purcell et al. paper, 1) our trial/test set for $S_{10}$ are small (N=42/46) and 2) our genotyped SNP list is relatively small. We were encouraged though that a small $p_T$ is capturing the majority of the diagnoses variance. This may suggest that we have identified a phenotypically homogeneous BD subgroup with clear genetic roots.
CHAPTER 5:

Conclusions and Future Directions

5.A Conclusions:

5.A.1 HapMap2-imputed GWAS: Imputation using HapMap2 is problematic. While imputation may make it possible for researchers to merge GWAS datasets from various sources, imputation also introduces a large non-disease related signal. After covariate-correction, 93% of this signal disappears, implying that covariate-correction is critically important. In our opinion, this imputed data should never be directly analyzed with any method which is not covariate corrected.

5.A.2 Logistic Regression use in imputation-reliant GWAS is problematic: Skewness, or lop-sidedness, and outliers / intermediate values are big features of HapMap2-imputed data. Over half the 2.4 million SNPs we analyzed demonstrated a high degree of skewness. Logistic regression, the default analytic method used to analyze this data, is not well-suited to analyzing skewed data or data with outliers. In keeping with this, our work shows that logistic regression is largely blind to the signal in these skewed SNPs. In light of this, it is hard to judge whether or not these SNPs (~50%) have been appropriately analyzed within previous GWAS.

In this work, we did not merely run simulations. We analyzed actual multi-study, multi-platform bipolar affective disorder case-control data. The trends we observed were consistent with those of the single published multi-platform imputation simulation study: (1) The non-disease related signal is greatest in the mixed genotyped/imputed data, followed by the purely imputed data, and is smallest in the purely genotyped data. (2) Using a very strict imputation quality score threshold, i.e., IQS > 0.98, reduced the level
of noise in our purely imputed data to almost that observed in the genotyped data. This strict thresholding also improved the performance of a logistic regression analysis. (a) Using the strictest IQS threshold (IQS > 0.98) strengthened the relationship between the case-control differences observed in imputed data and the case-control differences observed in genotyped data. Unfortunately, even with this improvement in imputed data’s ability to predict the behavior of genotyped data, it was still only slightly better than chance. (b) When we analyzed imputed and genotyped data together, this strict IQS threshold also improved the degree to which significant SNPs identified by logistic regression replicated. This improvement was most dramatic for the more-skewed SNPs leaving open the possibility that outliers and skewness are playing some role in confounding replication studies.

5.A.3 Area Under the receiver operating characteristic Curve (AUC) may be a better tool: Much of our work relied upon the use of AUC (Area Under the receiver operating characteristic Curve) and Corrected AUC (cAUC) to evaluate SNP case-control differences. We found: (1) AUC was far better able to detect the non-disease related signal introduced by the HapMap2-imputation than logistic regression. (2) Corrected AUC seems to be a very promising alternative to corrected logistic regression for analyzing imputed GWAS data, particularly for skewed data.

Corrected AUC’s improved performance appears to be driven by its robustness to skewness and outliers and its increased covariate correction, i.e., it can be corrected for non-linear relationships. As we have already mentioned, Corrected AUC can see a signal in the more-skewed SNPs to which logistic regression is largely blind. The degree to which it detects this signal decreases with increasing Imputation Quality Score thresholds. Since increasing these thresholds affects the degree of skewness less than it does the fraction of intermediate or outlier values, this improvement on more-skewed
SNPs reflects a joint benefit of Corrected AUC’s ability to handle outliers / intermediate values and skewness. Another reason we suspect increased covariate correction is important comes from the fact that both logistic regression and AUC interpreted even the less-skewed SNPs differently. Specifically, corrected logistic regression detected far more signal than Corrected AUC. Since our simulation studies have demonstrated that they should behave comparably on less-skewed SNPs, this disconnect suggests that different levels of covariate correction or the influence of outliers may be at play. It would appear that Corrected AUC promises to improve GWAS analysis of both skewed and non-skewed SNPs.

5.A.4 Binary disease / no-disease phenotype in bipolar affective disorder: There is a great deal of genetically-relevant phenotypic heterogeneity in bipolar affective disorder. Three different groups, including our own, each identified a different and likely genetically-relevant complex trait. Since we each used a different dimensionality-reduction technique to identify these subphenotypes, the specific choice of method seems less important than the quality and kind of phenotype data available to be analyzed. At this time, the data available is mainly self-reported and subject to interviewer bias and interviewee state-of-mind. Biomarker data, e.g., imaging data, blood work, is not currently being collected en masse. Partly, this is because of cost, i.e., imaging is very expensive, but it is also because we do not know a priori what biomarker data we should be collecting. To add further complication, there could be multiple genetic signatures underlying a single phenotype. If the latter were the case, it would not be sufficient to sort the phenotype even if we had all of the relevant phenotypic information available to us.

Since logistic regression looks for clusters, i.e., SNPs that are different across all cases and all controls, and it has not born much fruit, it makes sense to directly analyze
the genome using covariate-corrected methods to identify biclusters, i.e., subgroups of both subjects and SNPs, where case behavior and control behavior are different. If bipolar affective disorder is truly multiple diseases lumped together rather than a single disorder, using biclustering methods to analyze the genome directly should allow us to identify these subtypes even without any additional phenotypic information.

Covariate-corrected methods have just recently become available. Without covariate-correction, the non-disease related signal introduced through imputation would have prevented non-covariate corrected methods from detecting the true, but weak disease-related signal in our data. We intend to pursue this next logical analytic step in earnest using these new methods.

5.B Advice to researchers: (1) Be wary of HapMap2-imputed data. Do not analyze it with any method that cannot be covariate corrected, (2) Focus one’s analytic energies as follows: Purely Genotyped > Purely Imputed >> Mixed genotyped / imputed data. (3) Consider re-genotyping any significant imputed SNP prior to seeking its replication. Further, rather than viewing imputed data as indicative of behavior at that specific location on the genome, consider it as indicator of potentially interesting behavior in the region. (4) Consider using Corrected AUC to analyze your GWAS data, particularly for the more-skewed SNPs (~50% in our data). (5) If one chooses to use logistic regression to analyze this data, use very strict imputation quality thresholds (IQS > 0.98). For those SNPs achieving significance, attempt to assess the influence of non-linear relationships, skewness, and outlier / intermediate values. (6) For diseases with seemingly heritable phenotypic traits, consider directly-analyzing the genome with covariate-corrected methods to identify biclusters, i.e., subgroups of subjects and SNPs.

5.C Future Work: We plan to perform the following analyses: (1) Repeat our Data Reliability (Chapter 2) analyses for the 1000 Genomes dataset. Researchers need to
know how much non-disease related signal is present in their data prior to correction. For example, one may ask if the presence of non-disease related signal is restricted to HapMap2-imputation, or is a more general problem. If the non-disease-related signal is dramatically reduced by imputing with ~2400 haplotypes rather than 120 haplotypes, there would be no benefit to a strict IQS threshold (IQS > 0.98) and no need to toss out ~50% of the SNPs. (2) Compare risk prediction models based on Corrected AUC with those based on Corrected logistic regression. We want to know if Corrected AUC does a better job of identifying the disease-related signal that polygenic models and GCTA have demonstrated exists in the SNP data. (3) We also plan to use covariate-corrected biclustering methods to identify the homogeneous genetic subgroups, if any, present in the data. We will leverage our findings about Data Reliability (see Chapter 2) to create the sets we analyze.

5. D General: Our results regarding non-disease related signal may change if one were to impute using a larger reference panel. The large amount of non-disease related signal in our HapMap2-imputation is more likely the result of our reliance on 120 haplotypes to fill the holes in our almost 17,000 subjects rather than the process of imputation itself. If so, we would expect its magnitude to decrease when we impute with the 2400 haplotypes available through the 1000 Genomes project.

Our findings regarding skewness and outliers are unlikely to change if we were to re-impute using a larger reference panel. Skewness and outliers are natural features of the imputation process, particularly for lower imputation quality SNPs. It has already been documented in the literature that using 2400 haplotypes does not guarantee a high imputation quality score. Areas of low linkage disequilibrium and low minor allele frequency remain problematic. Given this, using Corrected AUC to analyze our GWAS data will likely continue to be a good and necessary alternative to logistic regression.
When we began this work, it was our intention to use our analysis of the bipolar affective disorder HapMap2-imputation GWAS as a forum for commenting upon the larger problems of neuropsychiatric GWASs. Ultimately, only one of our findings is likely disease or disease-category specific. As we have mentioned, skewness and outliers / intermediate values are features of imputed data, independent of disease class. Non-linearity of the data or its covariates is independent of imputation and disease class. As a result, the only part of our work that seems to be disease specific is our work documenting the existence of genetically-relevant subphenotypes and this is likely true of other neuropsychiatric / complex diseases. Therefore, the problems we have discussed, as well as the pressing need to address them are by no means limited to bipolar affective disorder or neuropsychiatric disorder imputation-reliant GWAS.

Researchers are moving toward whole genome sequencing and the problematic nature of imputed data will one-day soon be a concern of the past. Even when this comes to pass, cAUC’s will remain a promising technique for analyzing this non-imputed data because of its greater capacity for covariate correction.
Literature review: The BD literature offers up two forms of support for the existence of such genetically-relevant subphenotypes or subtypes: (1) Certain BD subphenotypes appear to be heritable, and (2) Some genome-wide significant / replicable SNPs have been identified after using a subphenotype to define the disease state, i.e., defining a new more focused homogeneous phenotype. To elaborate: there have been two large studies of phenotypic heritability (172 families, 598 sibpairs respectively). Both independently identified substance abuse, psychosis, panic disorder, and suicidality as heritable traits.[131, 132] (Table A.1 for a heritability summary, Table A.2 for the heritability details) Further, some of these subphenotypes seem to have unique genetic signatures. Single-trait subphenotypes have been leveraged to varying degrees of success in linkage and GWAS studies to identify promising candidate genes. (Table A.1/A.2 contains the results for linkage analysis, Table A.3 contains the details for Single-trait subphenotype GWASs) More complex subphenotypes, i.e., those defined by the interaction of more than one trait, have proved to be even more promising in GWAS. Specifically, there have been two GWAS whose phenotypes were defined by the use of Factor Analysis and Latent Class Analysis (LCA). (Factor analysis and LCA identify latent variables reflecting joint behavior in observed variables.) In the first, factor analysis performed on the 48 clinical OPCRIT variables yielded 12 factor dimensions. The autosomal SNPs were then tested for association with those 11 binary factor dimensions using the allelic Chi^2 model and the armitage trend test. One SNP reached genome-wide significance with the ‘negative mood delusions’ (poverty, guilt, nihilism) phenotype ($N_{case}/N_{control} = 927 / 2,168$). The SNP, rs9875793, was in proximity of the
SLC2A2. The G allele was limited to 89 of the 927 subjects with ‘negative mood delusions’. This finding was validated in second similar sample (1,247/1,434).[7] In the second, Latent Class Analysis (LCA) of the 10 co-morbid conditions found to be present in at least 5% of the BD subjects identified three latent variables: (1) Psychosis and/or substance abuse in the absence of alcohol dependence, (2) Alcohol dependence and other co-morbidities, (3) Absence of co-morbidities. The first latent variable (1) led to the identification of one SNP of genome-wide significance and another one that was close to genome-wide significant. The SNP with genome-wide significance (p=1.7x10⁻⁸) is near PDE10A, a gene previously implicated in psychosis. The other SNP (p=5.9x10⁻⁸) is near MARK1 and homozygotes for its minor allele are only present in cases. The second latent variable (2) was associated with one SNP of genome-wide significance (p=3.3x10⁻⁸). This region has been found to be deleted in individuals with autistic features and is thought to play a role in the axon connection formation in the developing brain. The last group (3) was not associated with genome-wide significant variants. The second study acknowledged that their sample sizes were small, they had identified rare variants, and resequencing would be an important step in validating these SNPs.[6] The existence of heritable subphenotypes alone would be sufficient to raise questions about the disease / no-disease model. The fact that some of these subphenotypes appear to have unique genetic signatures cemented our commitment to questioning this most fundamental assumption.
Table A.1 - Overview of Phenotypic Trait Heritability

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Significant Evidence of Heritability</th>
<th>Signif. linkage loci / Signif. higher than BAD alone</th>
<th>Linkage With cand. gene studies support</th>
<th>Separate cand. gene evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at onset (AAO)</td>
<td></td>
<td></td>
<td></td>
<td>Yes[161]</td>
</tr>
<tr>
<td>Psychosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puerperal psychosis</td>
<td></td>
<td></td>
<td></td>
<td>Yes[161]</td>
</tr>
<tr>
<td>Mood-congruent psychotic features</td>
<td></td>
<td></td>
<td>Yes/Yes[162]</td>
<td></td>
</tr>
<tr>
<td>Major Mood Episodes</td>
<td></td>
<td></td>
<td>Yes[163]</td>
<td></td>
</tr>
<tr>
<td>Rapid Cycling</td>
<td></td>
<td></td>
<td>Yes[164, 165]</td>
<td></td>
</tr>
<tr>
<td>Lithium Responsiveness</td>
<td></td>
<td>Yes/Not stated[166]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyclothymic Traits</td>
<td></td>
<td>Yes/Yes[167]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Panic Disorder Co-morbidity</td>
<td>Yes[168]</td>
<td>Yes/Yes[156, 169]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Suicide attempt</td>
<td>Yes[170]</td>
<td>Yes[171, 172]</td>
<td>Yes[170]</td>
<td>Yes/Yes[156, 173-175]</td>
</tr>
<tr>
<td>Phenotype</td>
<td>Observable Heterogeneity</td>
<td>Evidence of heritability</td>
<td>Genes identified by leveraging quantitative trait</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
</tbody>
</table>
| Age at onset (AAO) | - Distribution can be represented as 3 normal distributions. Cut points do vary slightly between studies[142, 176, 177]  
- Early-onset associated with (i) higher rates of comorbidity, (ii) longer episodes, (iii) unremitting course, and (iv) suicide attempts[178, 179]  
- As observed earlier, data has suggested different pattern of inheritance for those over and under 25 y.o., i.e. non-mendelian major gene with a polygenic component vs. multifactorial model[180] | - FDR of onset <15 y.o. were 2x as likely as late-onset to have BP, i.e. 18–29.4% vs. 5–7.4%[139, 140]  
- Affected FDRs of an early-onset proband were more likely than others to have an early onset (odds ratio=4.53, 95% CI=3.09-6.64) [142]  
- Age of onset of affected siblings highly correlated (r=0.42)[141] | Linkage Studies:  
- AAO of mania or depression, cut-point 21yrs.: Linkage at 21q22.13 (early onset) & 18p11.2 (late onset) were identified using LODPAL. 21q22.13 was verified in independent sample. 18p11.2 achieved genome wide significance using OSA (n=874 subjects; 150 multiplex pedigrees). Both are sites have suffered from inconsistent linkage [145]  
- AAO of mania or depression, cutpoint 21yrs: 98 BD families,428 subjects analyzed using LODPAL and AAO covariate: 3q28 with LOD 3.49, but only suggestive in independent sample [143]  
- AAO of mania, cutpoint 20yrs: 97 BD families,540 individuals: LOD >2.0: 2.21 (6q25), 3.21 (9q34), and 2.16 (20q11) [144]  
Candidate genes:  
- 5-HTTLPR (s/s vs. s/l vs. l/l) trend of earlier onset (n=129) [149, 151, 181]  
- Per3(VNTR): 5/5-earlier age at onset, 4/5 intermediate, 4/4 later onset (n=99) [146]  
- BDNF inconsistently associated with early-onset BD [147, 148] |
<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Observable Heterogeneity</th>
<th>Evidence of heritability</th>
<th>Genes identified by leveraging quantitative trait</th>
</tr>
</thead>
</table>
| Psychosis | - 50% of BD patients reported at least one lifetime psychotic episodes, most frequently grandiose delusions\[182\]  
- ↓ levels of RELN and GAD\[67\] in post-mortem prefrontal cortex of schizophrenia and BD subjects when compared with unipolar and control subjects\[183\]  
- SELENBP1 gene expression (blood) was 14% (p=0.27) higher in psychotic BD subjects vs. controls.  
- Note: Non-psychotic BD vs. controls was not significant, but the study did not directly compare psychotic BD levels with non-psychotic BD levels\[184\] | - In two independent studies of 65-69 BD pedigrees comparing non-psychotic vs. psychotic bipolar probands  
- families with psychotic BD proband more likely to have a relative with affective d.o. & psychotic Sx (64-76% vs. 28-48%)  
- A higher percentage of the FDR with affective d.o. have psychosis if the proband has psychosis (34-35% vs. 11-22%)  
- Affective psychotic Sx and bipolar psychotic Sx revealed significant clustering in families\[152, 153\]  
- Prevalence of schizophrenia spectrum traits is as high in the relatives of affective psychotics as it is in the relatives of schizophrenics\[185\] | BD & Schizophrenia:  
- 10p12-13, 13q32, 18 p11, 2, and 22q11-13 linked to both schizophrenia and bipolar\[186\]  
- Neuregulin 1 8p12, DISC 1, 1q42, COMT 22q11  
and DAOA(G72)/G30 13q33 have been implicated in both bipolar and schizophrenia studies\[186\]  
Linkage studies using psychosis covariates:  
1) 8p and 13q, based upon early-print abstract\[154\]  
2) When reanalyzed a pedigree in terms of psychotic BD, kerner et al. found linkage at 5q31-34. This region had been previously identified in schizophrenia, but not in the prior unrestricted analysis of the BD pedigree. Other suggestive linkages were 6q21-22, 8q21.2, 13p11.1-13q13, and 13q22-32 \[155\]  
---Fine mapping of the 5q31-34 region identified multiple SNPs in GRIA1 that were each significant in a different waves of the study. None were replicable across data sets\[157\]  
f3) 154 BD multiplex families, 1060 individuals:  
gsdenome wide significance at 16p12. Suggestive at 1p13, 1p21, 2p25, 4p16, 5p15, 6p25, 8p22, 10p11, 13q32, 19p13 \[156\] |
<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Observable Heterogeneity</th>
<th>Evidence of heritability</th>
<th>Genes identified by leveraging quantitative trait</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychosis (continued)</td>
<td></td>
<td></td>
<td>4) 13q31 and 22q12 identified by Potash et al when limited pedigree to psychotic BD [187]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>---- Fine mapping of 22q12.3 yielded region straddling the gene HMG2L1 and part of TOM1.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- 3 SNP haplotype within TOM1 was associated with BD, p= 0.0011 (allele-wise), but was associated more strongly with psychotic BD p=0.00049 (allele-wise). Psychotic/non-psychotic mean odds ratio of 1.91 vs. 0.8.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>- Single SNP in HMG2L1 (p=0.037) [158]</td>
</tr>
</tbody>
</table>

Candidate Genes:
- Association studies of psychotic BP and subtypes such as mood-incongruent psychotic BP have uncovered modest positive results for several candidate schizophrenia susceptibility genes, including dysbindin, DAOA/G30, Disrupted-in-Schizophrenia-1, and neuregulin, based upon early-print abstract [154]
- The significance of two SNPs (p=0.022, 0.031) in S100B (21q22) increased to p=0.016 and P = 0.009 when limited analysis to a psychotic subset [159]
### Table A.2 - Landscape of Heritable Phenotypes in BD (continued)

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Observable Heterogeneity</th>
<th>Evidence of heritability</th>
<th>Genes identified by leveraging quantitative trait</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mood-incongruent psychotic features</td>
<td>More severe course: increased hospitalization and attempted suicide [162]</td>
<td>Mood-incongruent proband predicted more mood incongruence in BD relatives (29.8% vs. 14.4%) [162]</td>
<td>Linkage Analysis using Mood-incongruent psychosis: When limited pedigree analysis to mood-incongruent psychosis: linkage at 13q21-33 (LOD 2.99, ↑ from 0.18) and 2p11-q14 (LOD 3.14, ↑ from 2.1). - Both are still only suggestive after genome-wide correction - 13q21-33 includes G72, but 2p11-q14 has only previously been identified in schizophrenia meta-analysis[162]</td>
</tr>
<tr>
<td>Puerperal psychosis</td>
<td>Psychosis in the time after childbirth (&lt;6 weeks)</td>
<td>- Puerperal psychosis happens more frequently in women with a family history (74% 20/27 vs. 30% 38/125) - Women with a FH of puerperal psychosis were at a 6-fold greater risk - 23/27 pairs of parous sisters were concordant puerperal psychotic episodes [160]</td>
<td>Linkage Studies using puerperal psychosis: A pedigree of 36 families with puerperal psychosis probands and affecteds: 16p13 (LOD 4.07, i.e. genome-wide significant linkage), 8q24 (LOD 2.03-suggestive). No significant or suggestive linkage was observed in these regions in the original bipolar scan [161]</td>
</tr>
<tr>
<td>Major mood episodes</td>
<td></td>
<td></td>
<td>Candidate Gene: DAOA/G30 locus (n=709 schizophrenics, 706 BD I, and 1416 controls): - 3/9 tag SNPs associations with BD (p = .01-.047), but not with schizophrenia. - 4/9 tag SNPs were more significantly associated with major mood disorders (SCZ &amp; BD together) (n=818). Genome-wide p=0.009 [163]</td>
</tr>
<tr>
<td>Phenotype</td>
<td>Observable Heterogeneity</td>
<td>Evidence of heritability</td>
<td>Genes identified by leveraging quantitative trait</td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Rapid Cycling (RC)</td>
<td>- 10-20% of BD (DSMIV-TR)</td>
<td>Candidate Gene: 1) BDNF: FB association study (n=56 RC BD families, 256 Non-RC BD families: 4 of 6 BDNF SNPs (hCV11592756, Val66Met, GT(n), and rs2049045) were significant in the RC sample, but not in the rest 2) COMT: 2 studies associating low-activity COMT allele with RC BD ---Examining Val108Met in 55RC vs. 110 NRC: low-activity (Met) allele frequency higher in RC than non-RC (0.55 vs. 0.42, p=0.012); dosage-dependent increased lifetime risk (hetero OR: 1.83, homo OR: 2.80) [165] ---6UURC vs. 19 RC, but Not UURC: low activity COMT was higher in Ultra-ultra-rapid cycling than among all others (p=0.002) [164]</td>
<td></td>
</tr>
<tr>
<td>Lithium Responsiveness</td>
<td>- Stargazin (calcium channel gamma-2 subunit gene, cacng2, on 22q13.1) expression levels in post-mortem brain samples is 1.6-fold higher in subjects with BPD (P=0.000036). None of the 12 SNPs in cacng2 were associated with BD, but 3 were significantly associated with Lithium response [189]</td>
<td>Several studies have suggested that lithium-responsive patients are more likely to have relatives affected with BD than lithium nonresponders(11526471) - Segregation analysis suggests major gene effect with autosomal recessive mode of inheritance[190]</td>
<td>Linkage analysis using Rapid cycling covariates: Using 31 families w/ excellent lithium responders probands: - 15q14 (LOD=3.46, p=0.000014), - Suggestive at 7q11.2 (LOD=2.68, p= 0.00011) - Other interesting findings at chromosome 6 (LOD=2.0, p=0.00004) and 22 (LOD=1.91) [166]</td>
</tr>
</tbody>
</table>
Table A.2 - Landscape of Heritable Phenotypes in BD (continued)

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Observable Heterogeneity</th>
<th>Evidence of heritability</th>
<th>Genes identified by leveraging quantitative trait</th>
</tr>
</thead>
</table>
| **Cyclothymic Disorder**   | - Chronic, fluctuating mood disturbance involving numerous periods of hypomanic symptoms and numerous periods of depressive symptoms (DSMIV-TR)  
- 0.4-1% lifetime prevalence symptoms (DSMIV-TR)  
- 15-50% risk of eventually developing BD symptoms (DSMIV-TR) | Some personality traits have been observed to be heritable in wwin studies in range from 40% to 60%. Article suggests that cyclothymic personality traits is a promising, likely heritable, endophenotype[191] | Linkage analysis using cyclothymic temperament trait:  
Evaluated MERLIN nonparametric multipoint regression linkage for cyclothymic temperament trait in 28 BD families (158 individuals w/ temperament data):  
- found suggestive linkage (LOD 2.71, p=0.0002) on Chromosome 18  
- Additional, linkage peaks on 3 and 7 (LOD=2.07, 2.05, respectively).  
- No loci had LOD>2.1 when either a broad or narrow BD classification was used and highest peaks identified were on chromosome 17 and 20 [167] |
| **Panic Disorder co-morbidity** | 25% of patients with BD have co-morbid panic disorder                                      | Relatives of probands with BD + panic were at higher risk of having panic disorder (37% vs. 13%) [168] | Linkage analysis using panic disorder as covariate:  
1) 154 BD multiplex families, 1060 individuals:  
   - genome wide significance at 7p21 (LOD=2.85) [156]  
2) 28 BD families, 5 with comorbid panic: (Looked only at chromosome 18) identified chromosome 18q21.1 (LOD= 3.10, p=0.00048) [169] |
### Table A.3 - Bipolar Affective Disorder Subphenotype GWAS

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Number of subjects</th>
<th>PHE-WAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at onset (AAO)</td>
<td>443/1,731</td>
<td>- Follow-up study on five regions identified in prev. genome-wide linkage-scan[143, 192]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Evidence for genetic heterogeneity within the ADRB2 (beta-2-adrenoreceptor) gene region</td>
</tr>
<tr>
<td></td>
<td></td>
<td>associated with early onset form of BD, known to interact with CACNA1C</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Could not replicate as SNPs unavailable [193]</td>
</tr>
<tr>
<td></td>
<td>2,836/2,744</td>
<td>Belmont Mahon et al.[194]</td>
</tr>
<tr>
<td></td>
<td>Repl: 3,916/5,112</td>
<td>- No SNPs of genome-wide significance in stage I of GWAS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- None of the SNPs with p&lt;0.0001 replicated</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Nor did any achieve genome-wide significance in the meta-analysis of both stages</td>
</tr>
<tr>
<td>Psychosis</td>
<td>1,775/2,744</td>
<td>Belmont Mahon et al [194]</td>
</tr>
<tr>
<td></td>
<td>Repl: &lt;=3,916/5,112</td>
<td>- No SNPs of genome-wide significance in GWAS in Stage I, replication set, or combined sample</td>
</tr>
<tr>
<td>Mood-incongruent psychotic features</td>
<td>2,196/8,148</td>
<td>Close to genome-wide, the following were over-represented in MICP sample [195]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- rs1171113, PRSS35/SNAP91 gene complex, P=9.67×10⁻⁸</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- rs2333194, intron of numb homolog, p=7.03×10⁻⁷</td>
</tr>
<tr>
<td>Seasonal Patterned Mania</td>
<td>392 BD with SM / 930BD with Non-SM</td>
<td>Nothing of genome-wide significance [196]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-rs41350144 (NF1A, p=3.08×10⁻⁷)</td>
</tr>
<tr>
<td>Migraine</td>
<td>56 BD with headaches/699 BD without headaches</td>
<td>Two SNPs close to genome-wide significance in KIAA0564 [197]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- rs9566845 (p= 7.7×10⁻⁸), p= 1.4×10⁻⁵ if use fisher's exact b/c of low # of cases)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>----- Associated with migraine in an independent Norwegian sample of adult ADHD patients w/ &amp; w/o</td>
</tr>
<tr>
<td></td>
<td></td>
<td>co-morbid migraine (n=131 and n=324 respectively), p=0.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- rs9566867 (p= 8.2×10⁻⁸, p= 1.5×10⁻⁵ if use fisher's exact b/c of low # of cases)</td>
</tr>
<tr>
<td>Suicide attempt</td>
<td>1,201 BP with suicide attempts / 1,497 without suicide attempts</td>
<td>Willour et al [198]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- No SNPs survived replication, but one SNP rs300774 at genome-wide significance threshold in the</td>
</tr>
<tr>
<td></td>
<td></td>
<td>combined sample (2p25, p=5.07×10⁻⁸, LD with block containing ACP1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- ACP1 expression is elevated in BD subjects that have completed suicide</td>
</tr>
<tr>
<td>Phenotype</td>
<td>Number of subjects</td>
<td>PHE-WAS</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| **Suicide Attempt** | 1,295 BP with suicide attempts/1,822 BP without suicide attempts Repl: 1,201 BP with suicide attempts/1,497 BP without suicide attempts | **Perlis et al [199]**
|                     | - 5 loci p<10^{-5}, 1 SNP rs1466846 (no genes within 400kb, p=1.98×10^{-6}), but none replicated | - They concluded that the inherited risk of suicide is unlikely due to common variants |
Table A.4 - Regions associated with SNPs who had 0.05 > MAF and 0.95 < MAF: Color guide: Darker blue = more relevant the gene / surrounding genes. Clear = no evidence of neuronal role, green = Not much known about SNP/region [A,B / C,D]: A: # of studies w/ p < 0.1, B: # of studies with correct direction trend in study cases vs. study controls C: # of studies w/ p < 0.1, D: # of studies with correct direction trend in study cases vs. ALL controls. * Demonstrates location of SNP relative to the genes. **Values for most significant SNP in the gene / region shown

<table>
<thead>
<tr>
<th>Gene / Region*</th>
<th>E(r^2)</th>
<th># of SNPs in region **</th>
<th>P ** (grp10)</th>
<th>OR (CI 95%)</th>
<th>P** (case/ctrl)</th>
<th>Case Freq **</th>
<th>Cntrl Freq **</th>
<th>λ p-value **</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>within 200kb of ITGA2, MOCS2,*</td>
<td>0.969</td>
<td>1</td>
<td>1.04E-06</td>
<td>0.26 (0.09-0.70)</td>
<td>0.059</td>
<td>0.945</td>
<td>0.985</td>
<td>0.021</td>
<td>MOCS2: molybdenum cofactor deficiency type B, *causes severe neurologic damage, ITGA2: integrin, implicated in lots of disorders, not mood though</td>
</tr>
<tr>
<td>XKR4</td>
<td>0.762</td>
<td>1</td>
<td>1.84E-06</td>
<td>3.40 (1.50-7.68)</td>
<td>0.024</td>
<td>0.081</td>
<td>0.025</td>
<td>0.002</td>
<td>Brain expressed; associated with antipsychotic iloperidone in GWAS, ADHD GWAS, schizophrenia, smoking cessation</td>
</tr>
<tr>
<td>ITPR1</td>
<td>0.329</td>
<td>1</td>
<td>2.23E-06</td>
<td>3.20 (1.02-10.0)</td>
<td>0.069</td>
<td>0.039</td>
<td>0.013</td>
<td>0.015</td>
<td>Intracellular receptor for inositol 1,4,5-trisphosphate, bipolar disorder (actions of lithium ruled in by pharmacogenetics, ruled out by a small study looking at binding site), antidepressant study in mice, ataxia and cerebellar disorders</td>
</tr>
<tr>
<td>within 200kb of KCNV1</td>
<td>0.715</td>
<td>11</td>
<td>1.42E-05</td>
<td>4.58 (1.42-14.7)</td>
<td>0.099</td>
<td>0.040</td>
<td>0.009</td>
<td>0.056</td>
<td>Potassium channel, expressed in brain, may play role in signalling both electrically excitable and non-excitable cells</td>
</tr>
<tr>
<td>DMXL2</td>
<td>0.924</td>
<td>1</td>
<td>1.69E-05</td>
<td>3.13 (1.46-6.69)</td>
<td>0.796</td>
<td>0.093</td>
<td>0.032</td>
<td>0.008</td>
<td>May serve as a scaffold protein for MADD and RAB3GA on synaptic vesicles; Expressed in brain</td>
</tr>
<tr>
<td>within 200kb of MYCB2, *SCEL</td>
<td>0.601</td>
<td>2</td>
<td>2.09E-05</td>
<td>0.39 (0.18-0.82)</td>
<td>0.547</td>
<td>0.905</td>
<td>0.960</td>
<td>0.000</td>
<td>MYCB2: May have a role during synaptogenesis, Expressed in brain, SCEL: hair, epidermis</td>
</tr>
</tbody>
</table>
Table A.4 - Regions associated with SNPs who had 0.05 > MAF and 0.95 < MAF (continued)

<table>
<thead>
<tr>
<th>Gene / Region*</th>
<th>E(r^2)</th>
<th># of SNPs in region**</th>
<th>P ** (grp 10)</th>
<th>OR (CI 95%)</th>
<th>P** (case/ctrl)</th>
<th>Case Freq**</th>
<th>Cntrl Freq**</th>
<th>P-value**</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>within 500kb of MKRN3, NDN, MAGL, in region that is imprinted improperly in prader willi (&amp;&amp;)</td>
<td>0.431</td>
<td>1 [2,3 / 2,3]</td>
<td>2.59E-05</td>
<td>0.27 (0.07-0.96)</td>
<td>0.318</td>
<td>0.967</td>
<td>0.991</td>
<td>0.001</td>
<td>MKRN3: imprinted in prader-willi syndrome, NDN: neuronal development / terminal differentiation of neurons, autism MAGL: similar to NDN, circadian rhythm; Ruled out in mood disorders and schizophrenia in a japanese group</td>
</tr>
<tr>
<td>RGS6</td>
<td>0.962</td>
<td>1 [2,3 / 2,3]</td>
<td>2.72E-05</td>
<td>3.72 (1.24-11.0)</td>
<td>0.235</td>
<td>0.044</td>
<td>0.012</td>
<td>0.002</td>
<td>Regulates G-protein signalling, interacts with SCG10 to promote neuronal differentiation, cancer, early-onset alzheimer's, smoking cessation, COPD</td>
</tr>
<tr>
<td>GRIN2B</td>
<td>0.852</td>
<td>1 [2,3 / 2,3]</td>
<td>3.31E-05</td>
<td>3.10 (1.36-7.08)</td>
<td>0.001</td>
<td>0.077</td>
<td>0.026</td>
<td>0.067</td>
<td>NMDA receptor, bipolar, schizophrenia, alcoholism, alzheimer's, parkinson's</td>
</tr>
<tr>
<td>C8orf37</td>
<td>0.972</td>
<td>2 [3,3 / 2,3]</td>
<td>3.71E-05</td>
<td>3.27 (1.49-7.18)</td>
<td>0.175</td>
<td>0.087</td>
<td>0.028</td>
<td>0.001</td>
<td>Expressed in the brain, not much known,</td>
</tr>
<tr>
<td>within 30 kb of CREB1</td>
<td>0.573</td>
<td>1 [3,3 / 2,3]</td>
<td>3.85E-05</td>
<td>0.41 (0.19-0.85)</td>
<td>0.530</td>
<td>0.902</td>
<td>0.957</td>
<td>0.008</td>
<td>Implicated in synchronization of circadian rhythmicity, bipolar, schizophrenia, cognitive deficit, drug addiction, memory disorders</td>
</tr>
<tr>
<td>within 2Mb DMXL1, TNFAIP8, HSD17B4, FAM170A*, PPR16</td>
<td>0.436</td>
<td>2 [3,3 / 3,3]</td>
<td>4.04E-05</td>
<td>0.38 (0.14-1.03)</td>
<td>0.934</td>
<td>0.948</td>
<td>0.979</td>
<td>0.000</td>
<td>DMXL1: DMXL2 also identified which plays a role in the brain, Not much known about DMXL1, TNFAIP8, HSD17B4:expressed in brain, link to autistic traits, FAM170A: Not much known, *, PRR16: Not much known</td>
</tr>
<tr>
<td>within 300kb of NOVA-1</td>
<td>0.751</td>
<td>30 [2,3 / 2,3]</td>
<td>4.11E-05</td>
<td>2.49 (1.26-4.91)</td>
<td>0.004</td>
<td>0.116</td>
<td>0.050</td>
<td>0.056</td>
<td>neuron-specific RNA-binding protein, May regulate RNA splicing or metabolism in a specific subset of developing neurons, neural development, plasticity viability GWAS of antipsychotic induced parkinson's in schizophrenia</td>
</tr>
<tr>
<td>within 20kb of MCOLN2</td>
<td>0.743</td>
<td>1 [2,3 / 2,3]</td>
<td>6.63E-05</td>
<td>2.57 (1.26-5.22)</td>
<td>0.332</td>
<td>0.106</td>
<td>0.044</td>
<td>0.000</td>
<td>cation channel proteins, smoking cessation, mucolipodosis IV, not much there</td>
</tr>
<tr>
<td>Gene / Region</td>
<td>E(r²)</td>
<td># of SNPs in region</td>
<td>P** (grp 10)</td>
<td>OR (CI 95%)</td>
<td>P** (case/ctrl)</td>
<td>Case Freq**</td>
<td>Ctntr Freq**</td>
<td>λ p-value**</td>
<td>Details</td>
</tr>
<tr>
<td>--------------</td>
<td>-------</td>
<td>---------------------</td>
<td>--------------</td>
<td>-------------</td>
<td>----------------</td>
<td>-------------</td>
<td>-------------</td>
<td>------------</td>
<td>---------</td>
</tr>
<tr>
<td>CNBD1</td>
<td>0.984</td>
<td>1 [2,3 / 2,3]</td>
<td>6.98E-05</td>
<td>0.29 (0.12-0.69)</td>
<td>0.008</td>
<td>0.926</td>
<td>0.977</td>
<td>0.001</td>
<td>not much known; smoking cessation</td>
</tr>
<tr>
<td>within 200kb of CD276, C15orf59,*</td>
<td>0.646</td>
<td>2 [2,3 / 2,3]</td>
<td>7.11E-05</td>
<td>0.19 (0.04-0.73)</td>
<td>0.276</td>
<td>0.970</td>
<td>0.994</td>
<td>0.021</td>
<td>CD276: May participate in the regulation of T-cell-mediated immune response; cancer, c15orf59: not much known</td>
</tr>
<tr>
<td>Within 500kb of ANKRD55,*</td>
<td>1.015</td>
<td>1 [2,3 / 2,3]</td>
<td>8.99E-05</td>
<td>0.39 (0.19-0.77)</td>
<td>0.192</td>
<td>0.884</td>
<td>0.951</td>
<td>0.002</td>
<td>Not much known, Rheumatoid arthritis, prostate cancer</td>
</tr>
<tr>
<td>within 2Mb of SLITRK1,*,SLITRK6</td>
<td>0.622</td>
<td>2 (other SNP within 50kb of SLITRK6) [2,3 / 2,2]</td>
<td>9.06E-05</td>
<td>0.44 (0.21-0.91)</td>
<td>0.876</td>
<td>0.899</td>
<td>0.952</td>
<td>0.000</td>
<td>SLITRKs are expressed predominantly in neural tissues and have neurite-modulating activity. SLITRK1: tourettes, OCD, tichotrillomania, SLITRK6: no diseases yet</td>
</tr>
<tr>
<td>PBMUCL1</td>
<td>0.693</td>
<td>1 [3,3 / 3,3]</td>
<td>9.12E-05</td>
<td>6.53 (1.36-31.3)</td>
<td>0.018</td>
<td>0.024</td>
<td>0.004</td>
<td>0.015</td>
<td>mucin like gene; diffuse panbronchiolitis</td>
</tr>
<tr>
<td>WWOX</td>
<td>0.860</td>
<td>1 [3,3 / 2,3]</td>
<td>9.28E-05</td>
<td>0.36 (0.17-0.80)</td>
<td>0.991</td>
<td>0.913</td>
<td>0.966</td>
<td>0.002</td>
<td>apoptosis, tumor suppression, role in bone density; no mention of anything psychiatric though much about cancer</td>
</tr>
<tr>
<td>within 20kb of CRMP1</td>
<td>0.846</td>
<td>1 [2,3 / 2,3]</td>
<td>9.52E-05</td>
<td>3.80 (1.37-10.5)</td>
<td>0.073</td>
<td>0.052</td>
<td>0.014</td>
<td>0.014</td>
<td>cytosolic phosphoproteins expressed exclusively in the nervous system. Plays a role in axon guidance, invasive growth and cell migration. Noted in a study of disease specific protein disorders in psychiatric conditions</td>
</tr>
<tr>
<td>FANCD2</td>
<td>0.747</td>
<td>[2,3 / 3,3]</td>
<td>9.75E-05</td>
<td>4.43 (1.35-14.5)</td>
<td>0.363</td>
<td>0.038</td>
<td>0.009</td>
<td>0.001</td>
<td>Required for maintenance of chromosomal stability. Mainly cancer, implicated in DNA damage in the brain pathway as well as 2 case studies in alzheimer</td>
</tr>
<tr>
<td>SLC9A9</td>
<td>1.007</td>
<td>[3,3 / 3,3]</td>
<td>9.80E-05</td>
<td>0.32 (0.14-0.72)</td>
<td>0.040</td>
<td>0.916</td>
<td>0.971</td>
<td>0.000</td>
<td>sodium hydrogen exchanger. Implicated in ADHD, autism</td>
</tr>
</tbody>
</table>

*bold indicates regions were significantly associated with the phenotype. **P-values were adjusted for multiple testing using the false discovery rate (FDR) method.
Table A.5 - Regions associated with SNPs who had 0.05 < MAF < 0.95: Color guide: **Darker blue** = more relevant the gene / surrounding genes. **Clear** = no evidence of neuronal role, **green** = Not much known about SNP/region. ***Original SNP was only significant in 2 studies, but another SNP in same gene / region was < 0.1 for all three***

<table>
<thead>
<tr>
<th>Gene</th>
<th># of SNPs in region</th>
<th>P (grp 10)</th>
<th>OR (CI 95%)</th>
<th>p (case/cntrl)</th>
<th>Case Freq</th>
<th>Cntrl Freq</th>
<th>λ p-value</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>within 2Mb of EFCAB1, SNAI2, C8orf22, *, SNTG1</td>
<td>7</td>
<td>7.86E-08</td>
<td>2.27 (1.44-3.56)</td>
<td>0.002</td>
<td>0.34</td>
<td>0.19</td>
<td>0.001</td>
<td>EFCAB1: not much known, BP/hypertension, SNAI2: Transcriptional repressor. Involved in the generation and migration of neural crest cells, cancer including brain, C8orf22: Not much known SNTG1: Syntrophins are cytoplasmic peripheral membrane proteins. Specifically expressed in brain, smoking cessation, diabetes, scoliosis</td>
</tr>
<tr>
<td>SIRPA2</td>
<td>1</td>
<td>6.26E-07</td>
<td>0.53 (0.32-0.87)</td>
<td>0.763</td>
<td>0.74</td>
<td>0.84</td>
<td>0.038</td>
<td>Supports adhesion of cerebellar neurons, neurite outgrowth and gial cell attachment. May play a key role in intracellular signaling during synaptogenesis and in synaptic function, schizophrenia proteome analysis, autoimmune, leukemia</td>
</tr>
<tr>
<td>within 200kb of NOVA-1</td>
<td>30 (5 are 3,3)</td>
<td>6.36E-06</td>
<td>0.44 (0.26-0.74)</td>
<td>0.000</td>
<td>0.77</td>
<td>0.88</td>
<td>0.015</td>
<td>(repeat of opposite effect SNPs) neuron-specific RNA-binding protein, expressed in brain, anti-psychotic induced parkinsonianism in schizophrenia, dementia</td>
</tr>
<tr>
<td>within 100kb ZDDHC8, *, RTN4R, 200kb of DGCR8, TRMT2A, RANBP1, 500kb of COMT (bipolar)</td>
<td>6 (5 are 3,3)</td>
<td>7.54E-06</td>
<td>1.94 (1.27-2.96)</td>
<td>0.123</td>
<td>0.49</td>
<td>0.33</td>
<td>0.005</td>
<td>DGCR8: biogenesis of microRNAs, diGeorge syndrome, TRMT2A: May be involved in nucleic acid metabolism and/or modifications, schizophrenia, attention deficit, RANBP1: cancer, shock, ZDDHC8: may function as a palmitoyltransferase, involved in glutamatergic transmission, susceptibility to schizophrenia, ADHD, ruled in/out of schizophrenia, *, RTN4R: Receptor mediates axonal growth inhibition and may play a role in regulating axonal regeneration and plasticity in the adult central nervous system, schizophrenia, alzheimer's</td>
</tr>
<tr>
<td>TEX10</td>
<td>15 (7 are 3,3)</td>
<td>9.31E-06</td>
<td>0.47 (0.30-0.73)</td>
<td>0.005</td>
<td>0.34</td>
<td>0.52</td>
<td>0.003</td>
<td>Not much known; cancer, neuroblastoma</td>
</tr>
<tr>
<td>within 2Mb TMEM132B, 2 theoretical proteins, within 20kb a theoretical LOC40084</td>
<td>2 (both 2,3)</td>
<td>9.33E-06</td>
<td>2.04 (1.29-3.21)</td>
<td>0.370</td>
<td>0.33</td>
<td>0.19</td>
<td>0.059</td>
<td>TMEM132B: expressed in brain, not much known, theoretical proteins are just that</td>
</tr>
<tr>
<td>Gene</td>
<td># of SNPs in region</td>
<td>P (grp 10)</td>
<td>OR (CI 95%)</td>
<td>p (case/cntrl)</td>
<td>Case Freq</td>
<td>Ctrl Freq</td>
<td>p-value</td>
<td>Details</td>
</tr>
<tr>
<td>------</td>
<td>------------------</td>
<td>----------</td>
<td>------------</td>
<td>----------------</td>
<td>----------</td>
<td>----------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>within 50kb of CLIC4, *, RUNX3</td>
<td>7 (2 are 3,3)</td>
<td>9.45E-06</td>
<td>2.08 (1.24-3.48)</td>
<td>0.041</td>
<td>0.79</td>
<td>0.65</td>
<td>0.000</td>
<td>CLIC4: chloride ion channel, regulate fundamental cellular processes including stabilization of cell membrane potential, transepithelial transport, maintenance of intracellular pH, and regulation of cell volume, <em>Binds directly to brain dynamin I in a complex containing actin, tubulin and 14-3-3 isoforms, expressed in brain, rat hippocampus, seizure, smoking cessation RUNX3: RUNT transcription factor, mainly cancers, no mood disorders noted</em></td>
</tr>
<tr>
<td>within 200kb of DGKG (many within 2Mb) ***</td>
<td>3 (2 are 2,3)</td>
<td>1.09E-05</td>
<td>2.16 (1.35-3.44)</td>
<td>0.002</td>
<td>0.72</td>
<td>0.84</td>
<td>0.017</td>
<td>diacylglycerol kinases, which are involved in lipid metabolism. <em>They play a role in the lithium pathway. In the gogli membrane, obesity, diabetes, bone density. Note that: DGKH was top hit in BAD GWAS and thought to play role in lithium pathway</em></td>
</tr>
<tr>
<td>RD3</td>
<td>4 (3 are 3,3)</td>
<td>1.44E-05</td>
<td>0.39 (0.21-0.69)</td>
<td>0.928</td>
<td>0.83</td>
<td>0.93</td>
<td>0.010</td>
<td>retinal protein; retinal degeneration</td>
</tr>
<tr>
<td>ZNRF3 ***</td>
<td>6 (2 are 2,3)</td>
<td>6.91E-06</td>
<td>0.48 (0.31-0.74)</td>
<td>0.654</td>
<td>0.40</td>
<td>0.58</td>
<td>0.000</td>
<td>Expressed in brain, function unknown; fat distribution, diabetes, ataxia</td>
</tr>
<tr>
<td>CACHD1</td>
<td>3 (1 is 3,3)</td>
<td>1.68E-05</td>
<td>0.42 (0.22-0.83)</td>
<td>0.139</td>
<td>0.88</td>
<td>0.94</td>
<td>0.010</td>
<td>May regulate voltage-dependent calcium channels; expressed in the brain. Not much known</td>
</tr>
<tr>
<td>LOC100129335, within 2kb of SOSTDC1, within 200kb of BZW2</td>
<td>2 (both 3,3)</td>
<td>2.31E-05</td>
<td>0.36 (0.19-0.69)</td>
<td>0.012</td>
<td>0.87</td>
<td>0.95</td>
<td>0.002</td>
<td>hypothetical protein, not much known, SOSTDC1: bone morphogenetic protein (BMP) antagonist; kidney, within 200kb of BZW2 (neuronal differentiation, identified in the opposite effects section) BZW2 (neuronal differentiation, identified in the opposite effects section with low E(r2))</td>
</tr>
<tr>
<td>within 30kb of GALNTL4</td>
<td>2 (both 3,3)</td>
<td>2.63E-05</td>
<td>1.80 (1.16-2.77)</td>
<td>0.687</td>
<td>0.40</td>
<td>0.27</td>
<td>0.053</td>
<td>May catalyze the initial reaction in O-linked oligosaccharide biosynthesis. One of top SNPs in the Baum et al. bipolar GWAS</td>
</tr>
<tr>
<td>within 300kb of CLPB, *, PDE2A, and ARAP1</td>
<td>8 (all 3,3)</td>
<td>2.71E-05</td>
<td>0.55 (0.35-0.85)</td>
<td>0.354</td>
<td>0.38</td>
<td>0.53</td>
<td>0.002</td>
<td>CLPB: May function as a regulatory ATPase and be related to secretion/protein trafficking process, AIDS, strep, shock, PDE2: Cyclic nucleotide phosphodiesterase, could play role in learning and memory. Failure to replicate in MDD, ARAP1: Expressed in the brain, diabetes</td>
</tr>
</tbody>
</table>
### Table A.5 - Regions associated with SNPs who had 0.05 < MAF < 0.95 (continued)

<table>
<thead>
<tr>
<th>Gene</th>
<th># of SNPs in region</th>
<th># of SNPs in group</th>
<th>P (CI 95%)</th>
<th>OR (CI 95%)</th>
<th>p (case/cntr)</th>
<th>Case Freq</th>
<th>Ctrl Freq</th>
<th>p-value</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>within 200kb * PIK3R1, but on the other side (previously noted by an opposite effect SNP)</td>
<td>4 (2 are 3,3)</td>
<td>3.09E-05</td>
<td>2.09 (1.20-3.64)</td>
<td>0.541</td>
<td>0.18</td>
<td>0.10</td>
<td>0.10</td>
<td>0.001 (repeat of a MAF &lt; 0.05 SNP which was excluded b/c had low E(r2) value) PIK3R1: Phosphatidylinositol 3-kinase, lithium thought to regulate, Decreased neuregulin 1-induced activation of the PI3K/AKT system is associated with impaired sensory gating in first-episode schizophrenia. Mainly cancer, insulin resistance</td>
<td></td>
</tr>
<tr>
<td>FAM134B</td>
<td>4 (all 3,3)</td>
<td>3.18E-05</td>
<td>0.51 (0.33-0.80)</td>
<td>0.944</td>
<td>0.64</td>
<td>0.78</td>
<td>0.000</td>
<td>Required for long-term survival of nociceptive and autonomic ganglion neurons. Defects in this gene are a cause of hereditary sensory and autonomic neuropathy type II (HSAN II).</td>
<td></td>
</tr>
<tr>
<td>NETO1</td>
<td>1</td>
<td>3.39E-05</td>
<td>0.43 (0.25-0.74)</td>
<td>0.203</td>
<td>0.80</td>
<td>0.90</td>
<td>0.001</td>
<td>Involved in the development and/or maintenance of neuronal circuitry. Accessory subunit of the neuronal N-methyl-D-aspartate receptor (NMDAR) critical for maintaining the abundance of GRIN2A-containing NMDARs in the postsynaptic density. Regulates long-term NMDA receptor-dependent synaptic plasticity and cognition, at least in the context of spatial learning and memory, expressed in the brain; one mention of diabetes. Note: GRIN2B was identified in the previous section</td>
<td></td>
</tr>
<tr>
<td>JAKMIP1</td>
<td>1</td>
<td>4.16E-05</td>
<td>0.53 (0.35-0.82)</td>
<td>0.074</td>
<td>0.52</td>
<td>0.67</td>
<td>0.004</td>
<td>Associates with microtubules and may play a role in the microtubule-dependent transport of the GABA-B receptor. May play a role in JAK1 signaling and regulate microtubule cytoskeleton rearrangements, only disease mentioned in GWAS crohn’s, GABA thought to play role in bipolar and schizophrenia</td>
<td></td>
</tr>
<tr>
<td>TOP3A, within 20kb of SHMT1 (which also had a significant hit)</td>
<td>5 (3 3,3)</td>
<td>4.57E-05</td>
<td>0.58 (0.37-0.88)</td>
<td>0.215</td>
<td>0.55</td>
<td>0.68</td>
<td>0.039</td>
<td>DNA topoisomerase, tumors, ataxia; SHMT1: cellular form of serine hydroxymethyltransferase. Interconversion of serine and glycine; Mentioned in a list of 306 genes involved in neurotransmission and neurodevelopment, cancer. Possible role in neural tube defects</td>
<td></td>
</tr>
<tr>
<td>within 30kb of WWC1, *RARS</td>
<td>2 (all 3,3)</td>
<td>4.67E-05</td>
<td>1.85 (1.20-2.85)</td>
<td>0.258</td>
<td>0.60</td>
<td>0.45</td>
<td>0.013</td>
<td>WWC1: Expressed in brain, cytoplasmic phosphoprotein that interacts with PRKC-zeta and dynein light chain-1. Alleles of this gene have been found that enhance memory in some individuals, memory, schizophrenia, depression, RARS: Aminoacyl-IRNA synthetases catalyze the aminoaoylation of IRNA by their cognate amino acid.</td>
<td></td>
</tr>
<tr>
<td>within 100kb of *PRR18, SFT2D1</td>
<td>8 (2 are 3,3)</td>
<td>4.87E-05</td>
<td>1.86 (1.18-2.92)</td>
<td>0.540</td>
<td>0.68</td>
<td>0.53</td>
<td>0.007</td>
<td>PRR18: not much known, SFT2D1: not much known</td>
<td></td>
</tr>
<tr>
<td>SNTB1</td>
<td>1</td>
<td>4.94E-05</td>
<td>0.53 (0.34-0.83)</td>
<td>0.886</td>
<td>0.38</td>
<td>0.53</td>
<td>0.003</td>
<td>Adapter protein that binds to and probably organizes the subcellular localization of a variety of membrane proteins. May link various receptors to the actin cytoskeleton and the dystrophin glycoprotein complex; expressed in neurons and glia; muscular dystrophy;</td>
<td></td>
</tr>
</tbody>
</table>
Table A.5 - Regions associated with SNPs who had 0.05 < MAF < 0.95 (continued)

<table>
<thead>
<tr>
<th>Gene</th>
<th># of SNPs</th>
<th>P (grp 10)</th>
<th>OR (CI 95%)</th>
<th>p (case/cntrl)</th>
<th>Case Freq</th>
<th>Cntrl Freq</th>
<th>p-value</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>within 2Mb of ADCY8, *, EFR3A, OC90, HHLA1, KCNQ3, HPRG1</td>
<td>2 (both 3,3)</td>
<td>6.67E-05</td>
<td>2.78 (1.31-5.91)</td>
<td>0.102</td>
<td>0.92</td>
<td>0.80</td>
<td>0.011</td>
<td></td>
</tr>
<tr>
<td>ADCY8: membrane-bound, calcium-stimulable adenylyl cyclase. May be involved in learning, in memory and in drug dependence, expressed in brain; bipolar, mood, narcolepsy, *EFR3A: differentially expressed in brainstem; hearing, OC90, HHLA1, KCNQ3 (slowly activating and deactivating potassium channel that plays a critical role in the regulation of neuronal excitability; epilepsy) – KCNQ2 previously identified in bipolar disorder, HPRG1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABCC6</td>
<td>1</td>
<td>6.81E-05</td>
<td>2.53 (1.33-4.82)</td>
<td>0.336</td>
<td>0.13</td>
<td>0.06</td>
<td>0.086</td>
<td></td>
</tr>
<tr>
<td>ABCC6: ATP-binding cassette (ABC) transporters, drug resistance, coronary artery disease, connective tissue disease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERBB2IP</td>
<td>1</td>
<td>6.87E-05</td>
<td>2.58 (1.37-4.86)</td>
<td>0.039</td>
<td>0.14</td>
<td>0.06</td>
<td>0.096</td>
<td></td>
</tr>
<tr>
<td>leucine-rich repeat and PDZ domain. Acts as an adapter for the receptor ERBB2, expressed in brain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP8A1</td>
<td>1</td>
<td>6.92E-05</td>
<td>0.37 (0.11-1.22)</td>
<td>0.853</td>
<td>0.03</td>
<td>0.08</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>P-type ATPases, most abundant in brain, heart, skeletal muscle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRAC</td>
<td>4 (all 2,3)</td>
<td>7.46E-05</td>
<td>1.90 (1.22-2.94)</td>
<td>0.816</td>
<td>0.64</td>
<td>0.48</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>T-cell receptor alpha constant, may play a role in MS, kidney troubles</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAM155A</td>
<td>2 (both 3,3)</td>
<td>8.14E-05</td>
<td>0.40 (0.21-0.75)</td>
<td>0.689</td>
<td>0.86</td>
<td>0.94</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td>Not much known, in the brain, depression GWAS, ADHD and conduct disorder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>within 30kb EEPD1</td>
<td>1</td>
<td>8.59E-05</td>
<td>0.53 (0.34-0.81)</td>
<td>0.397</td>
<td>0.39</td>
<td>0.55</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Expressed in brain: Not much known about function</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPP2</td>
<td>2 (both 3,3)</td>
<td>9.82E-05</td>
<td>0.34 (0.15-0.75)</td>
<td>0.068</td>
<td>0.08</td>
<td>0.19</td>
<td>0.003</td>
<td></td>
</tr>
<tr>
<td>mammalian peptidase that, at neutral pH, removes tripeptides from the N terminus of longer peptide, expressed in the brain, released tripeptides from several naturally occurring neuropeptides with quite broad specificity in one study, burkitt lymphoma, tumors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


160.


