2014-12-01

Role of Epistasis in Alzheimer's Disease Genetics

Mark T. Ebbert
Brigham Young University - Provo

Role of Epistasis in Alzheimer's Disease Genetics

Mark T. W. Ebbert

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

John S. K. Kauwe, Chair
Perry G. Ridge
Seth M. Bybee
Mark J. Clement
Chris D. Corcoran
Stephen R. Piccolo

Department of Biology

Brigham Young University

December 2014

ABSTRACT

Role of Epistasis in Alzheimer's Disease Genetics

Mark T. W. Ebbert
Department of Biology, BYU
Doctor of Philosophy

Alzheimer's disease is a complex neurodegenerative disease whose basic etiology and genetic structure remains elusive, despite decades of intensive investigation. To date, the significant genetic markers identified have no obvious functional effects, and are unlikely to play a role in Alzheimer's disease etiology, themselves. These markers are likely linked to other genetic variations, rare or common. Regardless of what causal mutations are found, research has demonstrated that no single gene determines Alzheimer's disease development and progression. It is clear that Alzheimer's disease development and progression are based on a set of interactions between genes and environmental variables. This dissertation focuses on gene-gene interactions (epistasis) and their effects on Alzheimer's disease case-control status.

We genotyped the top Alzheimer's disease genetic markers as found on AlzGene.org (accessed 2014), and tested for interactions that were associated with Alzheimer's disease case-control status. We identified two potential gene-gene interactions between rs11136000 (*CLU*) and rs670139 (*MS4A4E*) (synergy factor = 3.81; p = 0.016), and rs3865444 (*CD33*) and rs670139 (*MS4A4E*) (synergy factor = 5.31; p = 0.003). Based on one data set alone, however, it is difficult to know whether the interactions are real. We replicated the *CLU-MS4A4E* interaction in an independent data set from the Alzheimer's Disease Genetics Consortium (synergy factor = 2.37, p = 0.007) using a meta-analysis. We also identified potential dosage (synergy factor = 2.98, p = 0.05) and *APOE ε4* effects (synergy factor = 4.75, p = 0.005) in Cache County that did not replicate independently. The *APOE ε4* effect is an association with Alzheimer's disease case-control status in *APOE ε4* negative individuals. There is minor evidence both the dosage (synergy factor = 1.73, p = 0.02) and *APOE ε4* (synergy factor = 2.08, p = 0.004) effects are real, however, because they replicate when including the Cache County data in the meta-analysis. These results demonstrate the importance of understanding the role of epistasis in Alzheimer's disease.

During this research, we also developed a novel tool known as the Variant Tool Chest. The Variant Tool Chest has played an integral part in this research and other projects, and was developed to fill numerous gaps in next-generation sequence data analysis. Critical features include advanced, genotype-aware set operations on single- or multi-sample variant call format (VCF) files. These features are critical for genetics studies using next-generation sequencing data, and were used to perform important analyses in the third study of this dissertation.

By understanding the role of epistasis in Alzheimer's disease, researchers will begin to untangle the complex nature of Alzheimer's disease etiology. With this information, therapies and diagnostics will be possible, alleviating millions of patients, their families and caregivers of the painful experience Alzheimer's disease inflicts upon them.

Keywords: Alzheimer's disease, epistasis, *MS4A4E*, *CLU*, *CD33*

ACKNOWLEDGMENTS

During the course of my graduate work in the Department of Biology I received support, encouragement, and guidance from numerous individuals, who made my success possible. My experience has been enlightening and educational, and I would like to specifically acknowledge those to whom I am indebted.

I first acknowledge my committee comprised of Dr. John Kauwe (Keoni), Dr. Perry Ridge, Dr. Seth Bybee, Dr. Chris Corcoran, Dr. Stephen Piccolo, and Dr. Mark Clement. Each member contributed valuable insight that taught me and fortified my research. It has been an honor to work with each of them. I particularly acknowledge Dr. Kauwe and Dr. Ridge. Dr. Kauwe has been an amazing mentor academically and professionally, and has provided invaluable life lessons. I admire his ability and zeal as a scientist. Dr. Ridge has been a great support through various challenges during my Ph.D., and provided timely, critical guidance at times.

I also acknowledge the other faculty and staff of the Biology Department who have offered their time to help me. Specifically I would like to acknowledge Dr. Byron Adams whose contagious excitement in all aspects inspires me. Christina George and Gentri Glaittli, staff within the department, have also been especially helpful throughout my schooling. Of course, many other faculty and staff contributed to my education through coursework and other ways, of which I am not even aware.

Several undergraduate students were also incredibly helpful, and it was an honor to work with them throughout my time as a student. I particularly acknowledge Kevin Boehme who contributed substantially to the work presented in this dissertation. Kevin provided valuable

insights and was always willing to help in any way. He is a good friend and colleague. I wish him luck in his pursuits.

My parents have been a fount of inspiration throughout my life and education. They continue to teach me precious life lessons as they cheer me on. While I have acknowledged the following before, I must do so again: years ago while I was in grade school, my mother wondered whether she would ever get me through high school successfully, since my educational interests were somewhat lacking. Throughout those unsettling years, my parents showed extraordinary patience by continually encouraging me to perform my best and not to settle for less. My educational interests awoke later in life, though I struggled to develop intellectually. I found strength in a principle my father taught: “persistence will prevail.” That is a valuable lesson that persistence can overcome nearly any obstacle.

No one deserves as much praise and acknowledgement as my beloved wife Cheri, who has stood by me and supported me during the greatest challenges of my life. She is a remarkable woman, wife, friend, and mother. Our children and I are among the luckiest in the world. I couldn’t imagine a more compassionate, supportive, and Christ-like companion. During the most difficult moments, when I questioned my own resolve, she showed complete support and confidence that I would succeed. She fought for me when I could not fight for myself. Her faith in me gave me the courage to confront obstacles that were larger than I believed possible to overcome.

I also want to acknowledge my amazing children, Juliette, Mark-Tyler (Tiger), Colbin, and Sadie. They always welcome me home with excitement and love. They shower me with hugs when I leave, and beg me to stay home. They tug on my heart strings every time I have to leave them. They bring life, love, and happiness in my heart that I never knew possible before.

Finally, I express gratitude to my Heavenly Father and my Savior, Jesus Christ. Life has no shortage of challenges. These challenges can be terrifyingly bitter, but are meant to help us become more like Christ, if we choose to look to Him throughout the good times and the difficult times. Some experiences have challenged me to the core, but I know my Savior supports me through them. I know the Lord has also expanded my intellect. I owe everything to the Lord. I will dedicate myself, and everything He gives me, to loving and serving His children.

To my dear wife Cheri, our amazing children (our baby birds and caterpillars), and my loving parents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Background

Epistasis involves multiple genes contributing to a single phenotype, but understanding the nature of an epistatic interaction is not always clear. Epistatic interactions are generally discovered in two ways: (1) statistically; and (2) biologically. Statistical epistasis is deviation from additive effects between factors in the model[1], while biological epistasis is a physical interaction between two or more biological components. Both statistical and biological epistasis affect a single phenotype, however.

Bridging the gap between statistical and biological epistasis is a challenging, but necessary task for understanding genetics at its roots. Most phenotypes involve epistasis in complex organisms. Experiments to discover biological epistasis are challenging to carry out and limited in the interactions that they can identify. Identifying statistical epistasis also results in unique challenges. Specifically, discovering that two biological molecules interact provides crucial pathway and functional information, but the implications across phenotypes are often less obvious. Furthermore, just because proteins from two genes don't physically interact does not mean they do not both affect the same phenotype; the two proteins may be involved in the same pathway and cause different cascading events, or a given phenotype may be determined by multiple pathways. The possibilities seem endless. This limitation of understanding biological epistasis is where statistical epistasis excels. Using statistics, we can explore whether multiple genetic factors have a non-additive effect on a phenotype. If so, these genetic factors may be co-involved in the phenotype's presentation. Limitations of statistically derived epistasis, however, involve a certain level of uncertainty in the results because of: (1) false-positive and false-negative results; and (2) biological uncertainty. False-positive results are rampant when testing

numerous hypotheses, while false-negatives are likely because of poor statistical power. Regarding biological uncertainty, any statistically positive result may leave researchers questioning whether the interaction is real because the biology may not be obvious. In some cases, little or no information is available about a given gene. By focusing efforts to bridge the gap between statistical and biological epistasis, researchers will be able to leverage the complementary strengths of these two approaches and understand genetics at is roots.

Methods to Identify Statistical Epistasis: Merits and Limitations

Identifying statistical epistasis is the most common and cost-effective approach to discovering gene-gene interactions, but most studies of genetics in human disease focus on single genetic loci—likely an oversimplification of the underlying biology. To advance our genetic understanding of all phenotypes, we must understand the underlying epistatic relationships. Some analysis methods have been developed specifically to identify gene-gene interactions. Multifactor dimensionality reduction[2–17] and logistic regression[18–30] are the two most common methods. Synergy factors are an extension of logistic regression, and for the purposes of this discussion are included in that group. Multifactor dimensionality reduction is a nonparametric approach while logistic regression is parametric. Each method has disadvantages that limit their ability to identify interactions.

Logistic regression has several drawbacks when detecting epistasis according to He et al[15]: (1) interaction terms grow exponentially as the number of main effects included in the model increase; and (2) parameter estimates have large standard errors because the data is high-dimensional—decreasing power to detect the interactions. Another limitation according to Combarros et al. is that logistic regression is generally only valid for binary interactions because of limited sample size[31]. Park et al. proposed penalized logistic regression as a method to

overcome the limitations and showed that penalized logistic regression performs better than multifactor dimensionality reduction in some situations[32].

Many studies have demonstrated the utility of multifactor dimensionality reduction[33–37]. Advantages of multifactor dimensionality reduction include increased power[15,38] and superior ability to identify high-order interactions even when main effects are statistically insignificant[32]. Limitations, however, are that it is incapable of identifying additive main effects[32] and it struggles with missing values in high-dimensional data[39].

Given that the strengths and limitations of logistic regression and multifactor dimensionality reduction complement each other, combining them may be a powerful option. Multifactor dimensionality reduction could be used to discover complex interactions while logistic regression can be used for main effects.

There are other issues to consider that apply to all available methods such as potential false positives. According to Page et al.[40], there are four reasons an allele or interaction between alleles can be associated with a complex disease: (1) it is actually causative; (2) the association is by random chance; (3) a single allele is in disequilibrium with the causative allele; and (4) the association is due to a systematic bias in some portion of the study. Because of the high-dimensionality and small sample size of many studies, there is an increased likelihood of false positives for reasons stated by Page et al.; however, there is another potential cause of false positives known as "overfitting". Overfitting happens when a complex model is fit to data and is not generalizable beyond the population from which the sample was derived[41]. The cause has commonly been attributed to either genetic and environmental heterogeneity[42] or due to epistasis[1,43].

There are many approaches designed to prevent false positives and overfitting when studying predictive alleles in a given disease, but they are not fool proof. For instance, protocol when performing multiple comparisons—thousands in the case of Genome Wide Association Studies (GWAS)—involves adjusting p-values to limit the number of false positives due to chance. Similar methods exist to prevent overfitting statistical models to data. Although these methods are useful, researchers mistakenly report false associations.

Even though weak associations are often reported, this practice is not completely wrong. Statistical analyses are limited by the available data, and data is limited because of external restraints such as financial support, limited patient availability, genetic material, and even ethical restrictions. Given the various challenges researchers face to produce data, it is no wonder weak associations are reported. The key to separating true and false associations will be testing in independent data sets if they are large enough, or using meta-analyses across many smaller data sets to determine if the signal is consistent and significant. If a signal is replicable, researchers will then need to test associations biologically in cell lines or model organisms.

Epistasis in LOAD

Numerous studies have identified statistical epistasis in Alzheimer's disease using logistic regression[18–30] and multifactor dimensionality reduction[2–14]. Here we describe studies where results have been replicated in at least two independent samples. .

In 2004 Robson et al. identified statistical epistasis between the transferrin (TF) C2 allele and the haemochromatosis (HFE) C282Y allele using logistic regression and synergy factor analysis[21]. These genes were targeted because of previous evidence of iron buildup in Alzheimer's patients, which both of these genes play a role in metabolizing[44–46]. In 2009, Kauwe et al. replicated the findings from Robson et al. in a separate cohort[22]. There is strong evidence of

a biological cascading effect for this statistical interaction, as suggested by Kauwe et al.[22]. HFE binds with transferrin receptor 1 (TfR1), but the C282Y allele has a lesser affinity, allowing TfR1 to bind TF more easily[22,47]. It was hypothesized that more aggressive binding of TF may cause over absorption of dietary iron, leading to iron deposits in various tissues[22,48]. Additionally, Giunta et al. suggested wild-type TF plays an important role in iron transport and limits amyloid aggregation[22,49]. All of this information supports hypotheses by Robson et al.[21] and Lehmann et al.[50] that this interaction increases LOAD risk through increased redox-active iron and oxidative Stress.

Likewise, in 2004 Infante et al. identified statistical epistasis between interleukin-6 (IL-6) and interleukin-10 (IL-10) associated with decreased risk for Alzheimer's disease based on previous evidence that patients with Alzheimer's disease produce more pro-inflammatory interleukin-6 and less anti-inflammatory interleukin-10[51]. In 2009 Combarros et al. replicated the statistical interaction in a separate cohort[18]. This interaction may play a critical role in LOAD because Remarque et al. demonstrated that Alzheimer's disease patients have a pro-inflammatory phenotype and that Alzheimer's disease patients produce more IL-6 (pro-inflammatory) and less IL-10 (anti-inflammatory) when compared to controls[52]. It is difficult to determine, however, whether this inflammation is contributing to Alzheimer's disease, or is simply another side effect of the underlying cause.

In 2009, Combarros et al. performed a comprehensive analysis of over 100 reports of statistical epistasis, using and introducing their own synergy factor statistic. This study highlights the innate challenges in discovering statistical epistasis. The authors were only able to support 27 of the originally reported gene-gene interactions using their synergy factor analysis. The challenge with epistatic replication is that there are many factors that influence whether the

interaction can be detected in a given data set. Sample size, heterogeneity, and environmental factors are likely the most influential for detecting a real interaction.

In 2014, Gusareva et al. published the first replicable interaction associated with LOAD using an exhaustive, genome-wide screening approach[53]. They identified an interaction between KHDRBS2 (rs6455128) and CRYL1 (rs7989332) using a cohort from France including 2,259 cases and 6,017 controls. The interaction was then replicated in a cohort from Germany including 555 cases and 824 controls. The interaction was further supported by a meta-analysis using five more independent LOAD cohorts. Transcriptome analysis showed decreased expression for both genes in the temporal cortex and cerebellum brain regions. Gusareva et al. hypothesized a biological link between KHDRBS2 and CRYL1 through a potential association with heat-shock proteins and LOAD. KHDRBS2 is believed to affect transcription of heat-shock proteins because of studies in it's homologue Slm1 in *Saccharomyces cerevisiae*[53,54]. Slm1 was shown to interact with and activate TORC2[55], a kinase complex part of the TOR pathway, which Pierce et al. demonstrated affects amyloid β and cognitive function in Alzheimer's disease mouse models[56]. Pierce et al. hypothesized the reason inhibiting the TOR pathway affects amyloid β and cognition because of upregulated heat-shock proteins. This study in particular, represents the next step in discovering and describing functional repercussions of epistasis.

Epistasis Among Top LOAD Genes

Most epistasis studies in LOAD involve candidate genes, but to date, no study has addressed possible interactions between the top LOAD genes as found on AlzGene.org (accessed December 2014). These genes include the following: *APOE*, *BIN1*, *ABCA7*, *CR1*, *MS4A4E*, *CD2AP*, *PICALM*, *MS4A6A*, *CD33*, and *CLU*. *BIN1* (rs744373), *ABCA7* (rs3764650), *CR1* (rs3818361), *MS4A4E* (rs670139), and *CD2AP* (rs9349407) are associated with increased risk

for LOAD while *PICALM* (rs3851179), *MS4A6A* (rs610932), *CD33* (rs3865444), and *CLU* (rs11136000) are associated with decreased risk (6-10). Only one study to date, by Verhaaren et al., has examined the contribution of these nine risk alleles to LOAD status prediction (11). Verhaaren et al. calculated an additive genetic risk score and compared LOAD status prediction performance of age, gender, and *APOE ε4* genotype using logistic regression with and without the additive genetic risk score. The genetic risk score did not improve prediction performance significantly, suggesting that the nine alleles may not be diagnostically useful when constrained to an additive relationship. The assumption of additive relationships between risk loci is common but is likely to be an oversimplification of the underlying biology for LOAD and other complex diseases (12-14). In fact, there may be underlying gene-gene interactions not examined in the Verhaaren et al. study or others that improve LOAD status prediction performance.

In this dissertation we evaluate the possible interactions between these variants and their effects on Alzheimer's disease in several large, independent datasets and develop software to facilitate follow-up of genetic findings using whole genome sequence data. The first chapter describes my efforts to explore the effects of interactions on the diagnostic capabilities of known AD risk markers. Briefly, we genotyped each locus in 2,419 subjects from the Cache County Study on Memory Health and Aging and verified results by Verhaaren et al., but also explored statistical epistasis among the loci to determine if any interactions are informative to the model in the presence of the main (individual) allele affects. Two interactions were significant in the model: an interaction between *CD33* and *MS4A4E* ($p < 0.003$; SF 5.31, 95% CI 1.79 - 15.77), and between *CLU* and *MS4A4E* ($p < 0.016$; SF 3.81, 95% CI 1.28 - 11.32).

In subsequent chapters we describe novel software and our efforts to replicate these gene-gene interactions by performing an independent meta-analysis of datasets from the Alzheimer's

Disease Genetics Consortium (ADGC), followed by a combined meta-analysis including the original Cache County data. This work includes evaluation of dosage effects in both interactions and an *APOE ε4* effect as well as a permutation experiment to test robustness of results that had a significant p-value in the independent analysis. Finally, we explored possible causal variants that underlie this interaction using whole-genome sequence data from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

Future Directions

Many researchers are focusing their efforts on epistasis and the community is beginning to discover epistatic interactions that play a role in LOAD. The work outlined in this dissertation, which leveraged the use of markers known to show association with AD risk, supports an interaction between *CLU* and *MS4A4E* and is an important piece in understanding LOAD etiology. Each of the top candidate genes has a consistent and strong signal across numerous data sets, making it a reasonable hypothesis that there are interactions between them. It is not reasonable, however, to assume that the most critical interactions are only between loci with main effects. As such, researchers must approach epistasis in LOAD with even larger data sets using exhaustive, genome-wide approaches as demonstrated by the exciting study by Gusareva et al.

The International Genomics of Alzheimer's Project (IGAP) has a data set of over 74,000 cases and controls[57]—a massive data set by today's standards. Given the success by Gusareve et al., a similar agnostic (hypothesis-free) approach in such a large data set will likely result in more, stable interactions associated with LOAD case-control status, thus leading to potentially useful approaches for both diagnostics and therapeutics. IGAP also discovered several more

alleles with main effects in a recent study[57]. Rerunning our analysis across the top loci including IGAP's newly discovered loci may uncover new interactions.

Ultimately, however, we must bridge the gap between statistical and biological epistasis. Biological experiments demonstrating tangible effects on known or novel LOAD pathology will be essential to understanding the underlying etiology. These gene-gene interactions may involve physical interactions between proteins, or they may be indirect where they affect a downstream product.

References

1. Moore, J. H. & Williams, S. M. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays* **27,** 637–646 (2005).
2. Andrew, A. S. *et al.* Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. *Carcinogenesis* **27,** 1030–1037 (2006).
3. Briollais, L. *et al.* Methodological issues in detecting gene-gene interactions in breast cancer susceptibility: a population-based study in Ontario. *BMC Med.* **5,** 22 (2007).
4. Chen, M. *et al.* High-order interactions among genetic polymorphisms in nucleotide excision repair pathway genes and smoking in modulating bladder cancer risk. *Carcinogenesis* **28,** 2160–2165 (2007).
5. Tsai, C.-T. *et al.* Renin-angiotensin system gene polymorphisms and coronary artery disease in a large angiographic cohort: detection of high order gene-gene interaction. *Atherosclerosis* **195,** 172–180 (2007).
6. Chan, I. *et al.* Gene-gene interactions for asthma and plasma total IgE concentration in Chinese children. *J. Allergy Clin. Immunol.* **117,** 127–133 (2006).
7. Lee, J.-Y., Kwon, J.-C. & Kim, J.-J. Multifactor Dimensionality Reduction (MDR) Analysis to Detect Single Nucleotide Polymorphisms Associated with a Carcass Trait in a Hanwoo Population. *Asian-Australas. J. Anim. Sci.* **21,** 784–788
8. Ritchie, M. D. *et al.* Drug Transporter and Metabolizing Enzyme Gene Variants and Nonnucleoside Reverse-Transcriptase Inhibitor Hepatotoxicity. *Clin. Infect. Dis.* **43,** 779–782 (2006).

9. Park, H.-W. *et al.* Multilocus analysis of atopy in Korean children using multifactor-dimensionality reduction. *Thorax* **62,** 265–269 (2007).

10. Manuguerra, M. *et al.* Multi-factor dimensionality reduction applied to a large prospective investigation on gene–gene and gene–environment interactions. *Carcinogenesis* **28,** 414–422 (2006).

11. Julià, A. *et al.* Identification of a two-loci epistatic interaction associated with susceptibility to rheumatoid arthritis through reverse engineering and multifactor dimensionality reduction. *Genomics* **90,** 6–13 (2007).

12. Edwards, T. L., Lewis, K., Velez, D. R., Dudek, S. & Ritchie, M. D. Exploring the Performance of Multifactor Dimensionality Reduction in Large Scale SNP Studies and in the Presence of Genetic Heterogeneity among Epistatic Disease Models. *Hum. Hered.* **67,** 183–192 (2009).

13. Ritchie, M. D. *et al.* Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *Am. J. Hum. Genet.* **69,** 138–147 (2001).

14. Cho, Y. M. *et al.* Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia* **47,** 549–554 (2004).

15. Ritchie, M. D., Hahn, L. W. & Moore, J. H. Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet. Epidemiol.* **24,** 150–157 (2003).

16. Hahn, L. W., Ritchie, M. D. & Moore, J. H. Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics* **19,** 376–382 (2003).

17. Coffey, C. S. *et al.* An application of conditional logistic regression and multifactor dimensionality reduction for detecting gene-gene interactions on risk of myocardial infarction: the importance of model validation. *BMC Bioinformatics* **5,** 49 (2004).

18. Combarros, O. *et al.* Replication by the Epistasis Project of the interaction between the genes for IL-6 and IL-10 in the risk of Alzheimer's disease. *J. Neuroinflammation* **6,** 22 (2009).

19. Bullock, J. M. *et al.* Discovery by the Epistasis Project of an epistatic interaction between the GSTM3 gene and the HHEX/IDE/KIF11 locus in the risk of Alzheimer's disease. *Neurobiol. Aging* doi:10.1016/j.neurobiolaging.2012.08.010

20. Rodríguez-Rodríguez, E. *et al.* Interaction between HMGCR and ABCA1 cholesterol-related genes modulates Alzheimer's disease risk. *Brain Res.* **1280,** 166–171 (2009).

21. Robson, K. J. H. *et al.* Synergy between the C2 allele of transferrin and the C282Y allele of the haemochromatosis gene (HFE) as risk factors for developing Alzheimer's disease. *J. Med. Genet.* **41,** 261–265 (2004).

22. Kauwe, J. S. K. *et al.* Suggestive synergy between genetic variants in TF and HFE as risk factors for Alzheimer's disease. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet.* **153B,** 955–959 (2010).

23. Muendlein, A. *et al.* Synergistic effects of the apolipoprotein E ε3/ε2/ε4, the cholesteryl ester transfer protein TaqIB, and the apolipoprotein C3 −482 C > T polymorphisms on their association with coronary artery disease. *Atherosclerosis* **199,** 179–186 (2008).

24. Polito, L. *et al.* The SIRT2 polymorphism rs10410544 and risk of Alzheimer's disease in two Caucasian case–control cohorts. *Alzheimers Dement.* doi:10.1016/j.jalz.2012.02.003

25. Hiltunen, M. *et al.* Butyrylcholinesterase K variant and apolipoprotein E4 genes do not act in synergy in Finnish late-onset Alzheimer's disease patients. *Neurosci. Lett.* **250,** 69–71 (1998).

26. Licastro, F. *et al.* A new promoter polymorphism in the alpha-1-antichymotrypsin gene is a disease modifier of Alzheimer's disease. *Neurobiol. Aging* **26,** 449–453 (2005).

27. Talbot, C. *et al.* Polymorphism in AACT gene may lower age of onset of Alzheimer's disease. *Neuroreport* **7,** 534–536 (1996).

28. Combarros, O. *et al.* Interaction of the H63D mutation in the hemochromatosis gene with the apolipoprotein E epsilon 4 allele modulates age at onset of Alzheimer's disease. *Dement. Geriatr. Cogn. Disord.* **15,** 151–154 (2003).

29. Kamino, K. *et al.* Deficiency in mitochondrial aldehyde dehydrogenase increases the risk for late-onset Alzheimer's disease in the Japanese population. *Biochem. Biophys. Res. Commun.* **273,** 192–196 (2000).

30. Kim, J.-M., Stewart, R., Shin, I.-S., Jung, J.-S. & Yoon, J.-S. Assessment of association between mitochondrial aldehyde dehydrogenase polymorphism and Alzheimer's disease in an older Korean population. *Neurobiol. Aging* **25,** 295–301 (2004).

31. Combarros, O., Cortina-Borja, M., Smith, A. D. & Lehmann, D. J. Epistasis in sporadic Alzheimer's disease. *Neurobiol. Aging* **30,** 1333–1349 (2009).

32. Park, M. Y. & Hastie, T. Penalized logistic regression for detecting gene interactions. *Biostatistics* **9,** 30–50 (2008).

33. Musani, S. K. *et al.* Detection of Gene × Gene Interactions in Genome-Wide Association Studies of Human Population Data. *Hum. Hered.* **63,** 67–84 (2007).

34. Velez, D. R. *et al.* A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet. Epidemiol.* **31,** 306–315 (2007).

35. Moore, J. H. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert Rev. Mol. Diagn.* **4,** 795–803 (2004).

36. Cattaert, T. *et al.* Model-Based Multifactor Dimensionality Reduction for detecting epistasis in case–control data in the presence of noise. *Ann. Hum. Genet.* **75,** 78–89 (2011).

37. Namkung, J., Elston, R. C., Yang, J.-M. & Park, T. Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method. *Genet. Epidemiol.* **33,** 646–656 (2009).

38. Moore, J. H. *et al.* A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol.* **241,** 252–261 (2006).

39. He, H., Oetting, W. S., Brott, M. J. & Basu, S. Pair-Wise Multifactor Dimensionality Reduction Method to Detect Gene-Gene Interactions in A Case-Control Study. *Hum. Hered.* **69,** 60–70 (2010).

40. Page, G. P., George, V., Go, R. C., Page, P. Z. & Allison, D. B. 'Are We There Yet?': Deciding When One Has Demonstrated Specific Genetic Causation in Complex Diseases and Quantitative Traits. *Am. J. Hum. Genet.* **73,** 711–719 (2003).

41. Howard, C. G. & Bock, P. Using a hierarchical approach to avoid over-fitting in early vision. in *, Proceedings of the 12th IAPR International Conference on Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing* **1,** 826 –829 vol.1 (1994).

42. Gorroochurn, P., Hodge, S. E., Heiman, G. A., Durner, M. & Greenberg, D. A. Non-replication of association studies: 'pseudo-failures' to replicate? *Genet. Med. Off. J. Am. Coll. Med. Genet.* **9,** 325–331 (2007).

43. Wade, M. J. Epistasis, complex traits, and mapping genes. *Genetica* **112-113,** 59–69 (2001).

44. Connor, J. R., Menzies, S. L., St Martin, S. M. & Mufson, E. J. A histochemical study of iron, transferrin, and ferritin in Alzheimer's diseased brains. *J. Neurosci. Res.* **31,** 75–83 (1992).

45. Loeffler, D. A. *et al.* Transferrin and iron in normal, Alzheimer's disease, and Parkinson's disease brain regions. *J. Neurochem.* **65,** 710–724 (1995).

46. Smith, M. A., Harris, P. L., Sayre, L. M. & Perry, G. Iron accumulation in Alzheimer disease is a source of redox-generated free radicals. *Proc. Natl. Acad. Sci. U. S. A.* **94,** 9866–9868 (1997).

47. Feder, J. N. *et al.* The hemochromatosis gene product complexes with the transferrin receptor and lowers its affinity for ligand binding. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 1472–1477 (1998).

48. Townsend, A. & Drakesmith, H. Role of HFE in iron metabolism, hereditary haemochromatosis, anaemia of chronic disease, and secondary iron overload. *Lancet* **359,** 786–790 (2002).

49. Giunta, S., Galeazzi, R., Valli, M. B., Corder, E. H. & Galeazzi, L. Transferrin neutralization of amyloid beta 25-35 cytotoxicity. *Clin. Chim. Acta Int. J. Clin. Chem.* **350,** 129–136 (2004).

50. Lehmann, D. J., Williams, J., McBroom, J. & Smith, A. D. Using meta-analysis to explain the diversity of results in genetic studies of late-onset Alzheimer's disease and to identify high-risk subgroups. *Neuroscience* **108,** 541–554 (2001).

51. Infante, J. *et al.* Gene–gene interaction between interleukin-6 and interleukin-10 reduces AD risk. *Neurology* **63,** 1135–1136 (2004).

52. Remarque, E. J. *et al.* Patients with Alzheimer's disease display a pro-inflammatory phenotype. *Exp. Gerontol.* **36,** 171–176 (2001).

53. Gusareva, E. S. *et al.* Genome-wide association interaction analysis for Alzheimer's disease. *Neurobiol. Aging* **35,** 2436–2443 (2014).

54. Dickson, R. C. Thematic review series: sphingolipids. New insights into sphingolipid metabolism and function in budding yeast. *J. Lipid Res.* **49,** 909–921 (2008).

55. Berchtold, D. *et al.* Plasma membrane stress induces relocalization of Slm proteins and activation of TORC2 to promote sphingolipid synthesis. *Nat. Cell Biol.* **14,** 542–547 (2012).

56. Pierce, A. *et al.* Over-expression of heat shock factor 1 phenocopies the effect of chronic inhibition of TOR by rapamycin and is sufficient to ameliorate Alzheimer's-like deficits in mice modeling the disease. *J. Neurochem.* **124,** 880–893 (2013).

57. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* (2013). doi:10.1038/ng.2802

Chapter 2

Population-Based Analysis of Alzheimer's Disease Risk Alleles Implicates Genetic Interactions

Mark T. W. Ebbert[1,2], Perry G. Ridge[1,2], Andrew R. Wilson[2], Aaron R. Sharp[1], Matthew Bailey[1], Maria C. Norton[3,7], JoAnn T. Tschanz[4,7], Ronald G. Munger[5,7], Christopher D. Corcoran[6,7], John S. K. Kauwe[1]

[1]Department of Biology, Brigham Young University, Provo, Utah

[2]ARUP Institute for Clinical and Experimental Pathology, Salt Lake City, Utah

[3]Department of Family Consumer and Human Development, Utah State University, Logan, Utah

[4]Department of Psychology, Utah State University, Logan, Utah

[5]Department of Nutrition, Dietetics, and Food Sciences, Utah State University, Logan, Utah

[6]Department of Mathematics and Statistics, Utah State University, Logan, Utah

[7]Center for Epidemiologic Studies, Utah State University, Logan, Utah

Corresponding Author:

John S. K. Kauwe
675 WIDB
Provo, UT 84602
Phone: 801-422-2993
email: kauwe@byu.edu

Key words: Alzheimer's disease, epistasis, genetic interactions, population attributable fraction, odds ratio, risk

Abstract

Background. Reported odds ratios and population attributable fractions (PAF) for late-onset Alzheimer's disease (LOAD) risk loci (*BIN1*, *ABCA7*, *CR1*, *MS4A4E*, *CD2AP*, *PICALM*, *MS4A6A*, *CD33*, and *CLU)* come from clinically ascertained samples. Little is known about the combined PAF for these LOAD risk alleles and the utility of these combined markers for case-control prediction. Here we evaluate these loci in a large population-based sample to estimate PAF and explore the effects of additive and non-additive interactions on LOAD status prediction performance.

Methods. 2,419 samples from the Cache County Memory Study were genotyped for *APOE* and nine LOAD risk loci from AlzGene.org. We used logistic regression and ROC analysis to assess the LOAD status prediction performance of these loci using additive and non-additive models and compared ORs and PAFs between AlzGene.org and Cache County.

Results. Odds ratios were comparable between Cache County and AlzGene.org when identical SNPs were genotyped. PAFs from AlzGene.org ranged from 2.25-37%; those from Cache County ranged from 0.05-20%. Including non-*APOE* alleles significantly improved LOAD status prediction performance (AUC = 0.80) over *APOE* alone (AUC = 0.78) when allowing allelic interactions ($p = 0.03$). We also identified potential allelic interactions (p-values uncorrected): *CD33-MS4A4E* (Synergy Factor = 5.31; $p = 0.003$) and *CLU-MS4A4E* (SF = 3.81; $p = 0.016$).

Conclusions. While non-additive interactions between loci significantly improve diagnostic ability, the improvement does not reach the desired sensitivity or specificity for clinical use. Nevertheless, these results suggest that understanding gene-gene interactions may be important in resolving the etiology of Alzheimer's disease.

Introduction

Researchers have implicated several genes associated with late-onset Alzheimer's disease (LOAD) including *APOE*. *APOE ε4* increases LOAD risk and *APOE ε2* reduces risk (1-4). According to AlzGene.org (5), nine additional genes significantly affect LOAD risk; *BIN1* (rs744373), *ABCA7* (rs3764650), *CR1* (rs3818361), *MS4A4E* (rs670139), and *CD2AP* (rs9349407) are associated with increased risk for LOAD while *PICALM* (rs3851179), *MS4A6A* (rs610932), *CD33* (rs3865444), and *CLU* (rs11136000) are associated with decreased risk (6-10). Only one study to date has examined the contribution of these nine risk alleles to LOAD status prediction (11). Verhaaren et al. calculated an additive genetic risk score and compared LOAD status prediction performance of age, gender, and *APOE ε4* genotype using logistic regression with and without the additive genetic risk score. The genetic risk score did not improve prediction performance significantly, suggesting that the nine alleles may not be diagnostically useful when constrained to an additive relationship. The assumption of additive relationships between risk loci is common but is likely to be an oversimplification of the underlying biology for LOAD and other complex diseases (12-14). In fact, there may be underlying gene-gene interactions not examined in the Verhaaren et al. study or others that improve LOAD status prediction performance.

Some of the population attributable fractions for these nine loci have been reported individually and in different combinations (6, 8, 9); however, no study to date has reported the *combined* population attributable fraction for all nine risk alleles. Furthermore, previously reported odds ratios and population attributable fractions are from clinically ascertained samples rather than a population-based sample (6-10). The latter may provide a more reliable measure of

population risk because clinically ascertained samples select for disease, enriching risk alleles in the sample.

In this study we estimated the allelic odds ratios and population attributable fractions for *APOE ε2, APOE ε4,* and the nine non-*APOE* LOAD risk alleles in a large population-based sample. We also extended the genetic risk score used by Verhaaren et al. by testing whether the nine non-*APOE* alleles contribute significantly to LOAD status prediction when interactions between loci are not constrained to additive relationships.

Methods and Materials

Sample collection. The Cache County Study on Memory Health and Aging was initiated in 1994 (15). This cohort of 5,092 individuals represented approximately 90% of the Cache County population aged 65 and older. Specific details about data collection, obtaining consent, and phenotyping individuals in the Cache County population have been reported previously (15). Briefly, case-control status was determined in four triennial waves of data collection in a multi-stage dementia screening and assessment protocol. The first stage of screening consisted of administration of the Modified Mini-Mental State Exam-Revised (3MS-R) (16). Screen positive individuals and a randomly selected 19% designated subsample were invited to complete subsequent stages of evaluation consisting of an informant interview and the next stage, a clinical assessment including neuropsychological testing. The clinical assessment results were reviewed by a geropsychiatrist and neuropsychologist and preliminary diagnoses of dementia or other cognitive disorders were assigned. Those carrying a diagnosis of dementia or its prodrome were invited to complete standard laboratory tests for dementia, an MRI scan, and a geropsychiatrist examination. Final case-control status was determined by an expert panel of clinicians including study geropsychiatrists, neuropsychologists, a neurologist and cognitive

neuroscientist. Diagnoses of AD followed NINCDS-ADRDA criteria (17), and cases included Possible or Probable AD. Controls were identified as those who were diagnosed with no dementia (per clinical assessment) or whose cognitive test result was negative at each preceding screening stage. Persons with incomplete screening results (i.e., those who were screen positive at one stage, but did not complete the subsequent stage), or missing genotype data were excluded from the analyses, leaving 2093 participants without dementia (controls) and 326 persons with LOAD (cases). All study procedures were approved by the Institutional Review Boards of Utah State, Duke and the Johns Hopkins University.

DNA from the 2,419 Cache County study participants was genotyped for the nine non-*APOE* LOAD risk alleles in the AlzGene.org "ALZGENE TOP RESULTS" list (18) using TaqMan Assays (Table 2.1). Genotyping failed for rs3764650 (*ABCA7*) and rs3818361 (*CR1*) so we selected rs3752246 and rs6656401 to represent the effects reported by *ABCA7* and *CR1* for AD risk, respectively. The *CR1* SNPs are in high linkage disequilibrium (D' = 0.995, $R^2 = 0.84$) while both *ABCA7* SNPs are within 10 kilobases of each other and rs3752246 was reported as significant by Naj et al. (9) *APOE ε2* and *APOE ε4* were previously genotyped as part of the Cache County study (15).

Statistical analyses. All statistical analyses were performed in R (19). We used logistic regression and receiver operating characteristic (ROC) curve analysis to assess case-control predictive performance of the nine non-*APOE* alleles. Specifically, we tested whether the non-*APOE* alleles significantly improved LOAD status prediction performance over models excluding the non-*APOE* alleles. Two types of models were generated: additive risk profiles and genotype models to test potential additive and non-additive relationships, respectively. To assess efficacy of each model, we measured LOAD status prediction performance using the area under

Table 2.1. Summary Statistics for Significant Markers

		MAF		Odds Ratio		PAF	
SNP	Nearest Gene	AlzGene	Cache Co.	AlzGene (95% CI)	Cache Co. (95% CI)	AlzGene	Cache Co.
rs3752246*	ABCA7	0.10	0.18	1.23 (1.18 - 1.28)	0.94 (0.76 - 1.17)	2.25	4.65
rs7412	APOE2	0.06	0.09	0.62 (0.46 - 0.85)	0.89 (0.63 - 1.22)	36	10
rs429358	APOE4	0.22	0.17	3.68 (3.30 - 4.11)	2.51 (2.07 - 3.04)	37	20
rs744373	BIN1	0.29	0.30	1.17 (1.13 - 1.20)	1.02 (0.85 - 1.22)	4.61	0.54
rs9349407	CD2AP	0.29	0.28	1.12 (1.08 - 1.16)	1.03 (0.85 -1.23)	3.29	0.70
rs3865444	CD33	0.31	0.34	0.89 (0.86 - 0.92)	1.00 (0.84 - 1.19)	7.63	0.05
rs11136000	CLU	0.38	0.39	0.88 (0.86 - 0.91)	0.88 (0.74 - 1.04)	7.85	7.98
rs6656401	CR1	0.19	0.19	1.19 (1.09 - 1.30)	0.92 (0.74 -1.13)	3.49	6.84
rs670139	MS4A4E	0.41	0.41	1.08 (1.05 - 1.11)	1.0 (0.84 - 1.18)	3.14	0.05
rs610932	MS4A6A	0.42	0.43	0.90 (0.88 - 0.93)	0.89 (0.76 -1.06)	5.81	6.33
rs3851179	PICALM	0.35	0.38	0.88 (0.86 - 0.91)	0.85 (0.72 - 1.01)	8.19	9.69
				Combined PAF (All Alleles)		75	51
				Combined PAF (Excluding APOE)		38	32

Note. Minor allele frequencies, odds ratios, and population attributable risks were calculated for all SNPs using both data from AlzGene.org and the Cache County population-based study. Population attributable fractions are reported as percentages. For better interpretation and comparison to previous studies, the risk allele for each locus (whether the major or the minor allele) was used to calculate population attributable fractions but the minor allele was used for odds ratios. Minor allele frequencies are comparable between AlzGene.org and the Cache County data. Odds ratios are generally similar except *ABCA7* and *CR1* differ in direction. Individual population attributable fractions in Cache County varied in magnitude when compared to those calculated for AlzGene.org. Combined population attributable fractions were also lower in Cache County. As expected *APOE ε4* and *APOE ε2* have strong population effects whereas the remaining alleles have minimal individual effect. Based on AlzGene.org data, combined population attributable fractions suggest the combined effect of the nine non-*APOE* alleles is approximately equal to *APOE ε2* or *APOE ε4* alone; however, the nine non-*APOE* alleles appear to have a larger effect than either *APOE* allele in the Cache County data.

*The SNP for ABCA7 (rs3752246) was not reported on AlzGene.org, but was reported in Naj et al. as significant and was used in place of rs3764350

the curve (AUC) of the ROC curves. All models were adjusted for age and gender. A separate model using only age and gender was also generated to establish reference values.

We calculated three additive risk scores for participants in the Cache County Study to measure LOAD status prediction performance for the nine non-*APOE* LOAD risk alleles. Specifically, the following risk profiles were calculated: (1) *APOE* alone; (2) the nine LOAD risk alleles with *APOE*; and (3) the nine LOAD risk alleles without *APOE*. The risk allele (whether the major or the minor allele) and associated beta coefficient were used for each locus. We calculated additive risk scores as the sum of the risk across all alleles (equation 2.1), where β equals a previously calculated risk allele beta coefficient from odds ratios (β = ln(odds ratio)) reported by AlzGene.org (accessed February 2012), and N equals the subject's number of risk alleles. *APOE ε2* and *APOE ε4* were coded jointly into a single class variable as 22, 23, 24, 33, 34, and 44.

$$RP = \sum_{i}^{n} \beta_i N_i \tag{2.1}$$

We also tested genotype models using genotype data in place of the risk profile score. We generated the following genotype models: (1) *APOE* alone; (2) the nine LOAD risk alleles with *APOE*; and (3) an optimized model. Using genotypes does not constrain the model to an additive relationship, allowing for other genetic models within each locus. The optimized model was generated using a stepwise regression method to test if interactions between loci contribute to LOAD status prediction and was selected using Akaike's information criterion (AIC). To test for and avoid overfitting, we included three random variables while generating the optimized model. These variables were generated randomly with respect to all genotype and phenotype data in our study and were included to provide evidence that the selected variables provide meaningful

information (20). While the absence of all random variables in the model does not guarantee the model was not overfit, it does suggest the included variables provide useful diagnostic information.

Synergy factors—a statistic that measures the strength of allelic interactions in case-controls studies (13, 21)—were calculated for any statistically significant allelic interactions using logistic regression. All synergy factors were adjusted for age, gender, and *APOE ε4* by including only the main effects of the interacting alleles, the interaction term between the alleles, age, gender, and the number *APOE ε4* alleles (Status = allele1*allele2 + age + gender + *APOE*4num). Synergy factor confidence intervals were calculated using the interaction term coefficient ±1.96 * standard error of the parameter estimate of the interaction term.

Odds ratios and population attributable fractions were also calculated. Odds ratios here estimate the relative risk of Alzheimer's disease given allelic exposure while population attributable fractions estimate the proportional decrease in LOAD cases that would occur if the risk factor were removed from the population. Odds ratios were calculated only for the Cache County subjects but population attributable fractions were calculated for both Cache County subjects and the pooled AlzGene samples using published odds ratios and minor allele frequencies from AlzGene.org. We calculated population attributable fractions using equation 2.2 (9, 22), where p equals the allele frequency and OR is the odds ratio. A combined population attributable fraction was calculated for all risk factors and just the nine non-*APOE* risk factors using equation 2.3 (9, 22, 23) to estimate the proportional decrease in LOAD cases if all included risk factors were removed from the population. In this equation PAF_j represents previously calculated PAFs from equation 2.2 and n is the number of loci included in the combined PAF. For better interpretation and comparison to previous studies, the risk allele for each locus

(whether the major or the minor allele) was used to calculate population attributable fractions but the minor allele was used for odds ratios.

$$PAF = \frac{p(OR-1)}{p(OR-1)+1} \tag{2.2}$$

$$cPAF = 1 - \prod_{j=1}^{n}(1 - PAF_j) \tag{2.3}$$

Results

Sample demographics. The sample consisted of 1406 females and 1013 males. The mean age and standard deviation were 75.13 and 7.29 years, respectively. Mean age was significantly different between cases and controls (p = 2.2e-16), as were the proportion of males in each group (p = 0.04; Supplemental Table 2.1). Similarly, mean age was significantly different between participants included in the study and those excluded for reasons previously mentioned (p = 2.2e-16; Supplemental Table 2.2). The proportion of males, however, was not significantly different between included and excluded participants (p = 0.29).

Odds ratios. Odds ratios calculated for the Cache County data were generally comparable in direction and magnitude to odds ratios from AlzGene.org when identical SNPs were genotyped. *ABCA7* and *CR1* varied, but a different SNP was genotyped for *ABCA7* and the 95% confidence intervals for *CR1* overlap between AlzGene.org and Cache County results (Table 2.1). Odds ratios from meta-analyses on AlzGene.org for *ABCA7* and *CR1* are 1.23 (95% CI 1.18 – 1.28) and 1.19 (95% CI 1.09 – 1.30), respectively, while from the Cache County data were 0.94 (95% CI 0.76 – 1.17) and 0.92 (95% CI 0.74 – 1.13), respectively. No alleles deviated significantly from Hardy Weinberg equilibrium.

Population attributable fraction. Population attributable fractions as calculated from AlzGene.org data ranged from 2.25% to 37% while those from Cache County ranged from 0.05% to 20% (Table 2.1). The highest risks were attributed to *APOE ε4* (AlzGene = 37%; Cache = 20%) and lack of the *APOE ε2* (AlzGene = 36%; Cache = 10%) whereas the next highest risk was attributed to *PICALM* (AlzGene = 8.19%; Cache = 9.69%). The smallest risk for AlzGene.org was from *ABCA7* (2.2%) while the smallest for the Cache County data were *CD33* and *MS4A4E* (0.05%). Combined population attributable fractions for all LOAD risk alleles (including *APOE*) were 75% and 51% for AlzGene.org and Cache County, respectively. Using only the nine non-*APOE* alleles were 38% and 32% for AlzGene.org and Cache County, respectively.

LOAD status prediction performance. The non-*APOE* alleles combined with *APOE* (AUC = 0.782) did not improve LOAD status prediction performance over *APOE* alone (AUC = 0.783) when constrained to an additive model (Supplemental Figure 2.1), as previously reported (11); nor did the non-*APOE* alleles without *APOE* (AUC = 0.728) significantly improve LOAD status prediction performance over age and gender alone (AUC = 0.727; p = 0.2372). The model using all genotype data (full genotype model) when not constrained to an additive relationship (AUC = 0.796), however, did improve LOAD status prediction performance significantly over *APOE* alone (AUC = 0.783; p = 0.03; Figure 2.1). Moreover, the optimized model allowing for interactions between loci (AUC = 0.82) improves significantly over the full genotype model (p = 8.39e-07). All three genotype models improve prediction performance significantly over age and gender alone. None of the random variables previously mentioned were selected for the optimized model. Selected variables and interactions for the optimized model are as follows: rs3752246, rs6656401, rs11136000, rs610932, rs3865444, rs670139, Age, APOE.factor,

rs3865444:rs670139, rs11136000:rs670139, rs3752246:APOE.factor, rs3752246:rs610932, and rs670139:Age.

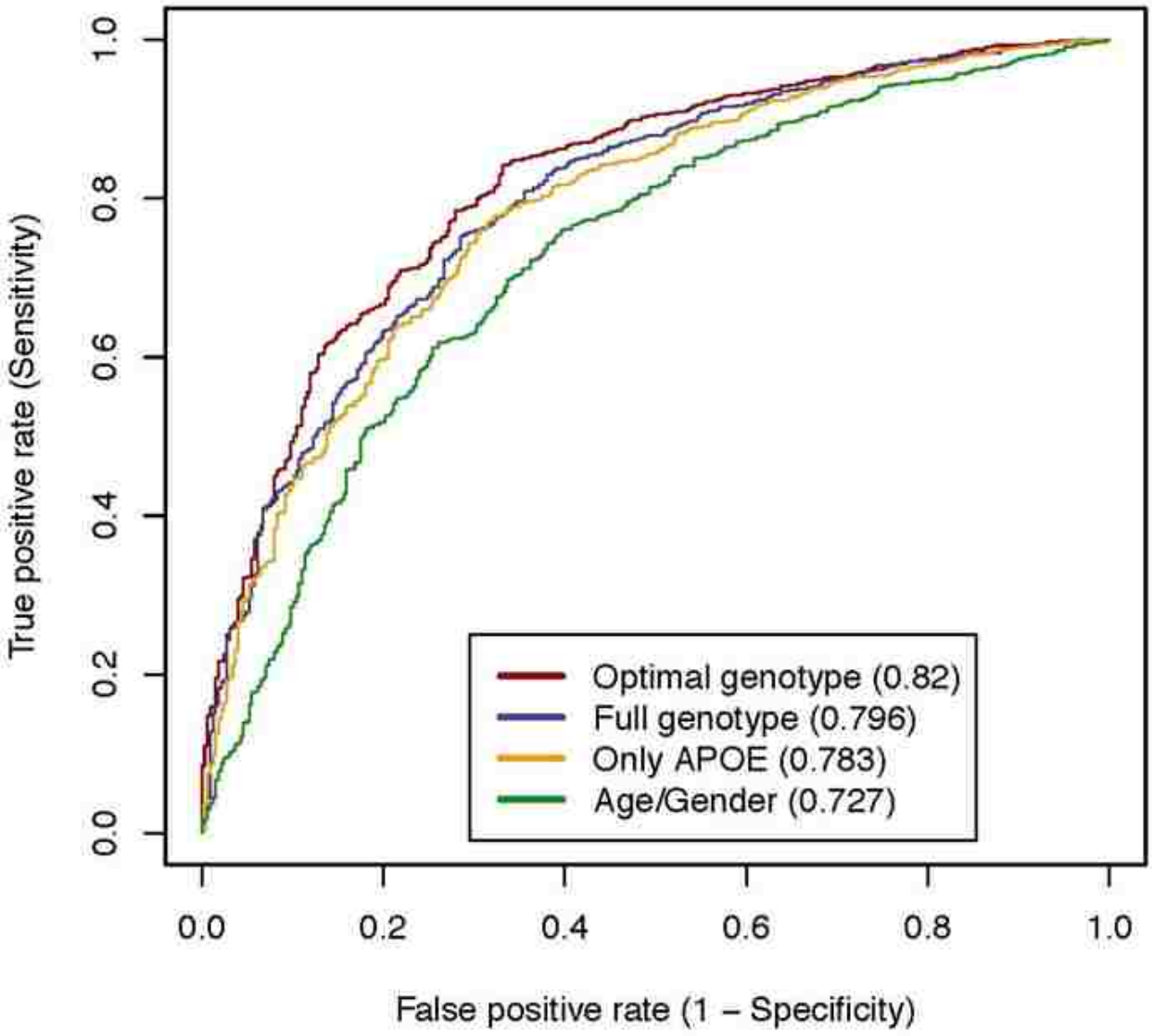


Figure 2.1. Non-*APOE* LOAD risk loci contributions to LOAD status prediction performance. Three logistic regression models based on age, gender, and genetic information for *APOE* and the non-*APOE* LOAD risk loci illustrate the contribution of the non-*APOE* LOAD risk loci in LOAD status prediction performance. The models are as follows: *APOE* alone (Only *APOE*), all loci (Full genotype), and the optimized model (Optimal genotype). A fourth model using only age and gender (Age/Gender) was also generated as a baseline. The optimized model was optimized using Akaike's Information Criterion (AIC). Comparing the full genotype model to *APOE* alone demonstrates that the LOAD risk loci contribute significantly to LOAD status prediction performance ($p = 0.03$) while the optimized model improves significantly over the full genotype model ($p = 8.39e\text{-}07$). Area under the curve (AUC) is listed in parentheses within the legend.

Locus interactions. Investigating the optimized genotype model revealed two statistically significant alleles and two significant allelic interactions, though the p-values were not corrected for multiple testing. Genotypes A/G (p = 0.02) and G/G (p = 0.03) in rs6656401 (*CR1*) were significant individually. The significant interactions were between the rs3865444 C/C (*CLU*) genotype and the rs670139 G/G (*MS4A4E*) genotype (p = 0.016; SF 3.81, 95% CI 1.28 - 11.32) and the rs11136000 C/C (*CD33*) genotype and the rs670139 G/G (*MS4A4E*) genotype (p = 0.003; SF 5.31, 95% CI 1.79 - 15.77).

Discussion

Recent research has identified several alleles that may prove useful in resolving Alzheimer's disease etiology (6-10), but until now there had not been an assessment of their population attributable fraction in a large, population-based sample. Similarly, deeper interrogation of the diagnostic utility of the Alzheimer's disease candidate genes is needed. Verhaaren et al. explored the diagnostic utility based on an additive relationship, which we replicated in this work, but they did not test locus interactions—a major aim of this research. During this process we also estimated allelic odds ratios and population attributable fractions.

The data reported in this study are generalizable to other U.S. populations of northern European descent. The Cache County population has been included in the Centre d'Etude du Polymorphisme Humain (CEPH) families that are used to represent the European sample in the HapMap project (24, 25). Utah's early pioneers were mostly unrelated and originated from various European locations (26-28), which is necessary for generalizability. The AlgGene.org data—a meta-analysis—varies between loci but is largely Caucasian-based as well. Many of the loci include populations of African, Asian, and Hispanic decent but the sample sizes for these populations are much smaller than the Caucasian populations.

Odds ratios. We compared Cache County odds ratios to those reported in the meta-analyses on AlzGene.org and found them comparable. Minor differences were observed in *ABCA7* and *CR1* where we genotyped SNPs that are not listed on AlzGene.org. Specifically, minor alleles for both *ABCA7* and *CR1* were considered risk alleles (odds ratio > 1) according to data on AlzGene.org while odds ratios in the Cache County data suggest decreased risk, although the confidence intervals from both studies are broad and overlap each other so they may not be significantly different. Possible causes include: (1) differences in sample ascertainment between clinical and population studies (e.g. the cases in clinically ascertained samples are generally younger than those in the Cache County Sample; see AlzGene.org, Supplemental Table 2.1); and (2) allelic odds ratios are not adjusted for age, gender, and other loci—nor are they adjusted for undiscovered or uncharacterized allelic interactions (13, 29-31).

Clinical and population studies differ in sample ascertainment. Clinically ascertained cases and controls are selected to minimize confounding variables and maximize contrast between the true underlying causes by minimizing known differences between the two groups except for the phenotype of interest. Population-based studies, however, are designed to represent true population characteristics such as allele frequencies, odds ratios, and population attributable fractions, as reported here. Because of the natural differences between these two study types, it is important to use them to their greatest advantage.

The complex nature of Alzheimer's disease inheritance, however, suggests that variations between studies may be exist because allelic odds ratios are not adjusted for age, gender, and other loci—nor are they adjusted for undiscovered and uncharacterized allelic interactions. Each of these factors plays a significant role in Alzheimer's disease etiology and not adjusting for them introduces error into odds ratio estimates. Allelic interactions also likely contribute to the

"missing heritability" in Alzheimer's disease. No single genetic locus characterizes Alzheimer's disease etiology. *APOE* alone is highly predictive, but the genetic loci included here also appear to influence Alzheimer's disease susceptibility, as reported in this study and others (6-10). Furthermore the effects of *APOE* vary between ethnic groups (32-36). Failure to replicate established genome-wide association study findings in some populations (13, 37) further suggests the possible influence of environmental factors, gene-environment, and gene-gene interactions.

Population attributable fractions. Cache County population attributable fractions varied in magnitude when compared to those calculated from AlzGene.org data. Combined population attributable fractions were lower in Cache County. As expected *APOE ε4* and *APOE ε2* have strong population effects whereas the remaining alleles have minimal individual effects. Based on AlzGene.org data, combined population attributable fractions suggest the combined effect of the nine non-*APOE* alleles is approximately equal to *APOE ε2* or *APOE ε4* alone; however, the combined non-*APOE* alleles appear to have a larger effect than either *APOE* allele in the Cache County data. The Cache County values are of value because they are population-based and better represent risks within populations—the purpose of the PAF statistic. Despite being more conservative than other estimates (combined), however, the population attributable fractions reported in this study may still be inflated because they are based on the unadjusted allelic odds ratios and because the exposure frequency for the genotyped SNPs may vary from the functional variants they represent. Future estimates are also likely to change as allelic interactions are discovered and incorporated into the calculations.

Diagnostic utility. Verhaaren et al. demonstrated that the nine non-*APOE* genes do not improve LOAD status prediction performance when constrained to an additive relationship,

which we confirmed in this study. When unconstrained, however, the top nine alleles improved LOAD status prediction performance significantly, demonstrating these alleles may provide more information as we better understand their epistatic relationships. The optimized model further improved LOAD status prediction performance and revealed *CLU-MS4A4E* and *CD33-MS4A4E* interactions that may prove valuable in Alzheimer's disease research. Synergy factors for both interactions suggest that being homozygous for both alleles in either interaction increases risk. Yet, although these data suggest the additional LOAD risk alleles significantly improve LOAD status prediction performance, the improvement is marginal and does not reach the desired sensitivity or specificity for clinical use.

The optimized model clearly improves LOAD status prediction performance over the full genotype model and over *APOE* alone, suggesting allelic interactions may be useful for diagnostic purposes; however, the p-values were not corrected for multiple testing. As such, these interactions need to be tested in an independent data set. It is also possible the optimized model is overfit; however, the random variables included in the model selection process were not selected for the final model, lending evidence that the final variables included provide non-random information. The revealed interactions also have strong synergy factors suggesting they may be important. Furthermore, the genotype model with all alleles improves LOAD status prediction performance over *APOE* alone, lending support for underlying relationships amongst the factors included in the model.

Implications and future directions. The results presented here offer evidence that gene-gene interactions play a role in Alzheimer's disease susceptibility; however, the reported interactions, do not appear to improve LOAD status prediction performance by an amount that is relevant in a clinical diagnostic setting. These results do suggest that to fully understand the

genetic basis of Alzheimer's disease risk we must improve our efforts to characterize gene-gene and gene-environment interactions.

Additionally, environmental factors have not received as much attention as genetic factors in Alzheimer's disease research and should be thoroughly investigated (12). Although the *CLU-MS4A4E* and *CD33-MS4A4E* interactions appear to have strong effects in the Cache County study, there may be unmeasured environmental factors that increase the effect of these interactions in the Cache County population. Other research has shown that only 30% of Alzheimer's disease is explained by known genes, demonstrating that environmental effects and gene by environment interactions will be essential in future studies (38).

The *CLU-MS4A4E* and *CD33-MS4A4E* interactions have not been previously reported leaving the biological foundation in question. Using IPA (Ingenuity® Systems, www.ingenuity.com), we explored possible interactions between each pair and found that, while no information is available for *MS4A4E* specifically, both *CLU* and *CD33* interact indirectly with *MS4A2* (Supplemental Figures 2.2 and 2.3). According to IPA, both thioacetamide and *TGFB1* act indirectly on both *CLU* and *MS4A2* (Supplemental Figure 2.2). *CLU* also binds to *BCL2L1*, which is acted upon by *MS4A2*. Likewise, *CD33* acts on *PTPN6,* which binds to *MS4A2* and *CD33* binds to *CBL*, which then acts on *MS4A2* (Supplemental Figure 2.3). Both *MS4A4E* and *MS4A2* are members of the membrane-spanning 4-domain gene family. A complete IPA legend is available in Ingenuity's website (http://ingenuity.force.com/ipa/articles/Feature_Description/Legend).

Overall, the results presented in this paper suggest that gene-gene interactions (epistasis) may play an important role in Alzheimer's disease etiology. While discovering and characterizing epistatic interactions is a non-trivial task, researchers and consortiums must heed

the plentiful evidence that Alzheimer's disease is driven by complex gene-gene and gene-environment interactions.

Acknowledgements

This work was supported by grants from NIH (R01AG11380, R01AG21136, R01AG31272, R01AG042611), the Alzheimer's Association (MNIRG-11-205368) and the Utah Science, Technology, and Research initiative (USTAR), the Utah State University Agricultural Experiment Station, and the Brigham Young University Gerontology Program. The authors thank the participants and staff of the Dementia Progression Study, the Utah Population Database, and the Cache County Study on Memory Health and Aging for their important contributions to this work. Additionally, the authors acknowledge the assistance of Drs. David Ward and Ned Weinshenker. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Financial Disclosures

No authors report conflicts of interest, financial or otherwise.

References

1. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, et al. (1993): Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*. 261:921-923.
2. Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PH, Pericak-Vance MA, Joo SH, et al. (1993): Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. *Neurology*. 43:1467-1472.
3. St Clair D, Rennie M, Slorach E, Norrman J, Yates C, Carothers A (1995): Apolipoprotein E epsilon 4 allele is a risk factor for familial and sporadic presenile Alzheimer's disease in both homozygote and heterozygote carriers. *J Med Genet*. 32:642-644.
4. Strittmatter WJ, Saunders AM, Schmechel D, Pericak-Vance M, Enghild J, Salvesen GS, et al. (1993): Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci U S A*. 90:1977-1981.
5. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE (2007): Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet*. 39:17-23.
6. Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, et al. (2009): Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet*. 41:1088-1093.

7. Hollingworth P, Harold D, Sims R, Gerrish A, Lambert JC, Carrasquillo MM, et al. (2011): Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat Genet*. 43:429-+.

8. Lambert JC, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, et al. (2009): Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet*. 41:1094-1099.

9. Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, et al. (2011): Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet*. 43:436-+.

10. Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, et al. (2010): Genome-wide analysis of genetic loci associated with Alzheimer disease. *JAMA*. 303:1832-1840.

11. Verhaaren BF, Vernooij MW, Koudstaal PJ, Uitterlinden AG, Duijn CM, Hofman A, et al. (2012): Alzheimer's Disease Genes and Cognition in the Nondemented General Population. *Biol Psychiatry*.

12. Bullock JM, Medway C, Cortina-Borja M, Turton JC, Prince JA, Ibrahim-Verbaas CA, et al. Discovery by the Epistasis Project of an epistatic interaction between the GSTM3 gene and the HHEX/IDE/KIF11 locus in the risk of Alzheimer's disease. *Neurobiol Aging*.

13. Combarros O, Cortina-Borja M, Smith AD, Lehmann DJ (2009): Epistasis in sporadic Alzheimer's disease. *Neurobiol Aging*. 30:1333-1349.

14. Moore JH, Williams SM (2005): Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays*. 27:637-646.

15. Breitner JC, Wyse BW, Anthony JC, Welsh-Bohmer KA, Steffens DC, Norton MC, et al. (1999): APOE-epsilon4 count predicts age when prevalence of AD increases, then declines: the Cache County Study. *Neurology*. 53:321-331.

16. Tschanz JT, Welsh-Bohmer KA, Plassman BL, Norton MC, Wyse BW, Breitner JC, et al. (2002): An adaptation of the modified mini-mental state examination: analysis of demographic influences and normative data: the cache county study. *Neuropsychiatry Neuropsychol Behav Neurol*. 15:28-38.

17. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984): Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*. 34:939-944.

18. Bertram L MM, Mullin K, Blacker D, Tanzi R (2007): The AlzGene Database. Alzheimer Research Forum.

19. R Core Team (2012): R: A language and environment for statistical computing. 2.15.1 ed. Vienna, Austria: R Foundation for Statistical Computing.

20. Gonzalez NA, Swain NR, Obregon O, Buehler BD, Williams GP, Nelson EJ, et al. (2012): Spatial and Temporal Statistical Analysis of Water Quality Patterns in a Small Temperate Supply Reservoir. American Society of Civil Engineers, pp 1982-1992.

21. Cortina-Borja M, Smith AD, Combarros O, Lehmann D (2009): The synergy factor: a statistic to measure interactions in complex diseases. *BMC Research Notes*. 2.

22. (2011): Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet*. 377:641-649.

23. Slatkin M (2008): Exchangeable models of complex inherited diseases. *Genetics*. 179:2253-2261.

24. International HapMap C (2003): The International HapMap Project. *Nature*. 426:789-796.

25. Ridge PG, Maxwell TJ, Corcoran CD, Norton MC, Tschanz JT, O'Brien E, et al. (2012): Mitochondrial genomic analysis of late onset Alzheimer's disease reveals protective haplogroups H6A1A/H6A1B: the Cache County Study on Memory in Aging. *PLoS One*. 7:e45134.

26. O'Brien E, Rogers AR, Beesley J, Jorde LB (1994): Genetic structure of the Utah Mormons: comparison of results based on RFLPs, blood groups, migration matrices, isonymy, and pedigrees. *Hum Biol*. 66:743-759.

27. Jorde LB, Morgan K (1987): Genetic structure of the Utah Mormons: isonymy analysis. *Am J Phys Anthropol*. 72:403-412.

28. Jorde LB (1982): The genetic structure of the Utah Mormons: migration analysis. *Hum Biol*. 54:583-597.

29. Moore JH, Williams SM (2005): Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays*. 27:637-646.

30. Cordell HJ (2002): Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*. 11:2463-2468.

31. Moore JH (2003): The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered*. 56:73-82.

32. Tang MX, Maestre G, Tsai WY, Liu XH, Feng L, Chung WY, et al. (1996): Relative risk of Alzheimer disease and age-at-onset distributions, based on APOE genotypes among elderly African Americans, Caucasians, and Hispanics in New York City. *Am J Hum Genet*. 58:574-584.

33. Murrell JR, Price B, Lane KA, Baiyewu O, Gureje O, Ogunniyi A, et al. (2006): Association of apolipoprotein E genotype and Alzheimer disease in African Americans. *Arch Neurol*. 63:431-434.

34. Mayeux R (2003): Apolipoprotein E, Alzheimer disease, and African Americans. *Arch Neurol*. 60:161-163.

35. Desai PP, Hendrie HC, Evans RM, Murrell JR, DeKosky ST, Kamboh MI (2003): Genetic variation in apolipoprotein D affects the risk of Alzheimer disease in African-Americans. *Am J Med Genet B Neuropsychiatr Genet*. 116B:98-101.

36. Maestre G, Ottman R, Stern Y, Gurland B, Chun M, Tang MX, et al. (1995): Apolipoprotein E and Alzheimer's disease: ethnic variation in genotypic risks. *Ann Neurol*. 37:254-259.

37. Healy DG (2006): Case-control studies in the genomic era: a clinician's guide. *Lancet neurology*. 5:701-707.

38. Lee SH, Harold D, Nyholt DR, Consortium AN, International Endogene C, Genetic, et al. (2013): Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis. *Hum Mol Genet*. 22:832-841.

Supplemental Table 2.1. Demographic Comparison between Cases and Controls Included in the Study Analysis

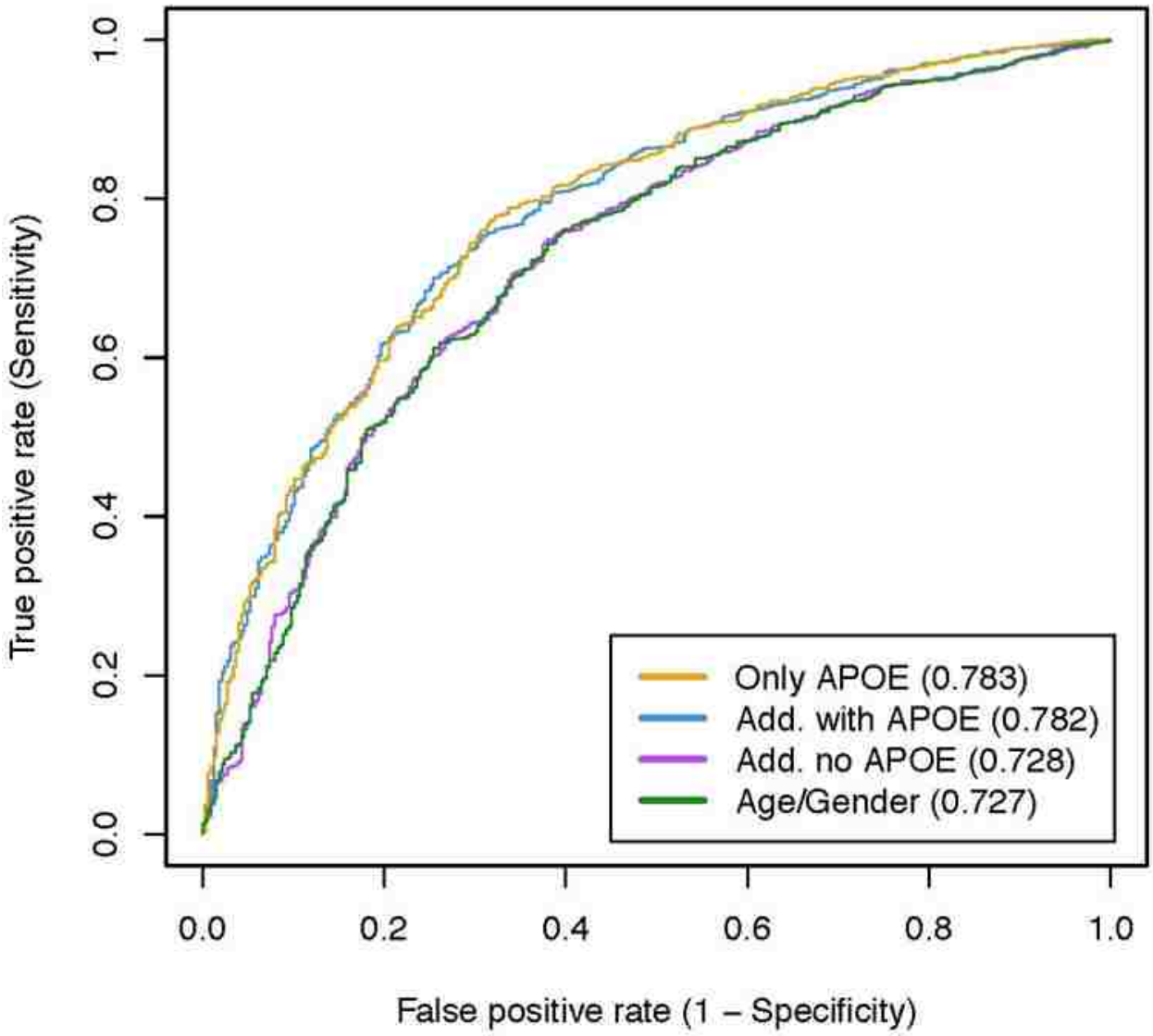
	Age		Gender			
	Mean	Standard Deviation	Male	Female	n	Proportion of Females
Cases	80.17	7.24	119	207	326	0.63
Controls	74.34	6.68	894	1199	2093	0.57
n			1013	1406	2419	
p-value	= 2.2e-16					= 0.04

Note. The mean age between cases and controls included in the study were significantly different as are the differences in the proportion of females.

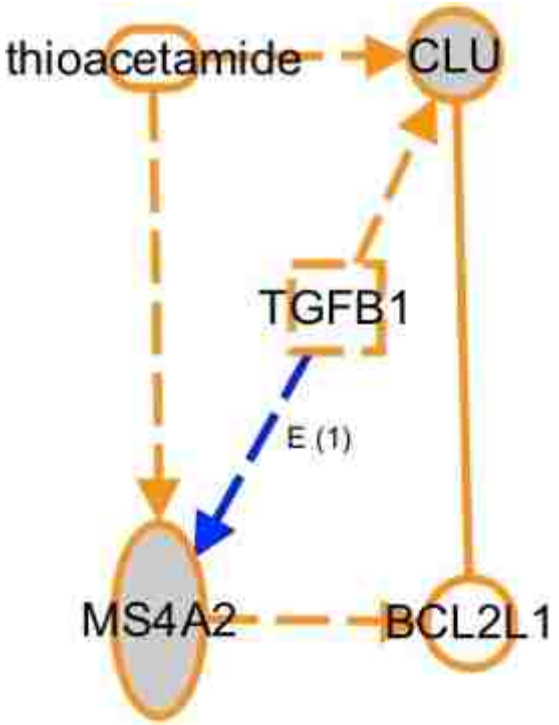
Supplemental Table 2.2. Demographic Comparison between Participants Included and Excluded in the Analysis

	Age		Gender			
	Mean	Standard Deviation	Male	Female	n	Proportion of Females
Included	75.13	6.92	1013	1406	2419	0.58
Excluded	77.33	7.48	1074	1399	2473	0.57
n			2087	2805	4892	
p-value	= 2.2e-16					= 0.29

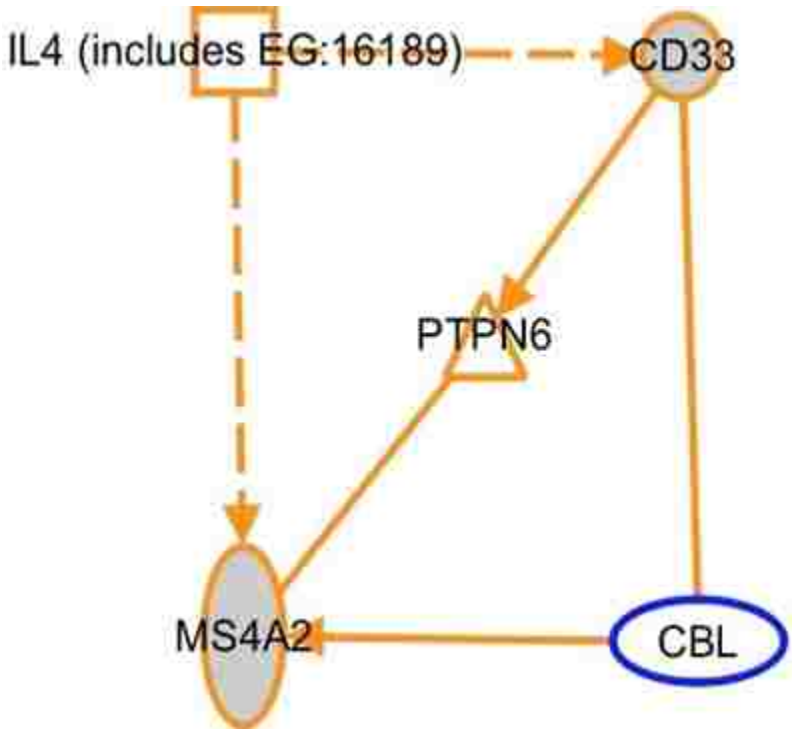
Note. The mean age between participants included and those excluded were significantly different, but the proportion of females was not. One possible cause of this difference is that samples excluded for missing genotype data were significantly older than those that were included. This is likely because the majority of DNA samples come from the original buccal swabs. These samples have lower call rates than the blood DNA that was collected at later waves of assessment. As a result, the individuals who were oldest at the start of the study have higher genotype missing rates. This results in the slightly higher age of excluded samples over included samples. However, unless there is a loss of individuals who go on to develop AD vs. those who remain non-demented this unlikely to bias our results. There is no evidence for such a bias.



Supplemental Figure 2.1. Non-*APOE* LOAD risk loci contributions to LOAD status prediction performance under additive constraints. The non-*APOE* alleles combined with *APOE* did not improve LOAD status prediction performance over *APOE* alone when constrained to an additive model; nor did the non-*APOE* alleles without *APOE* significantly improve LOAD status prediction performance over age and gender alone ($p = 0.2372$). Area under the curve (AUC) is listed in parentheses within the legend.



Supplemental Figure 2.2. *CLU-MS4A4E* pathway analysis. Pathway analysis using Ingenuity's IPA demonstrates evidence that both *CLU* and *CD33* interact indirectly with *MS4A2*, a member of the membrane-spanning 4-domain gene family, as is *MS4A4E*. Both thioacetamide and *TGFB1* act indirectly on both *CLU* and *MS4A2* (Supplemental Figure 2). *CLU* also binds to *BCL2L1*, which is acted upon by *MS4A2*. Likewise, *CD33* acts on *PTPN6,* which binds to *MS4A2* and *CD33* binds to *CBL*, which then acts on *MS4A2* (Supplemental Figure 3). No information regarding *MS4A4E* specifically was available in IPA. An exhaustive legend describing the molecules and interactions are available on Ingenuity's website (http://ingenuity.force.com/ipa/articles/Feature_Description/Legend).



Supplemental Figure 2.3. *CD33-MS4A4E* pathway analysis. Pathway analysis using Ingenuity's IPA demonstrates evidence that both *CLU* and *CD33* interact indirectly with *MS4A2*, a member of the membrane-spanning 4-domain gene family, as is *MS4A4E*. Both thioacetamide and *TGFB1* act indirectly on both *CLU* and *MS4A2* (Supplemental Figure 2). *CLU* also binds to *BCL2L1*, which is acted upon by *MS4A2*. Likewise, *CD33* acts on *PTPN6,* which binds to *MS4A2* and *CD33* binds to *CBL*, which then acts on *MS4A2* (Supplemental Figure 3). No information regarding *MS4A4E* specifically was available in IPA. An exhaustive legend describing the molecules and interactions are available on Ingenuity's website (http://ingenuity.force.com/ipa/articles/Feature_Description/Legend).

Chapter 3

Variant Tool Chest: An Improved Tool to Analyze and Manipulate Variant Call Format (VCF) Files

Mark T.W. Ebbert[1,2], Mark E. Wadsworth[1], Kevin L. Boehme[1], Kaitlyn L. Hoyt[1], Aaron R. Sharp[1], Brendan D. O'Fallon[2], John S.K. Kauwe[1], Perry G. Ridge[1*]

[1] Department of Biology, Brigham Young University, Provo, Utah, USA; [2] ARUP Institute for Clinical and Experimental Pathology, Salt Lake City, Utah, USA; [*] To whom correspondence should be addressed

Emails:

Mark Ebbert: me.mark@gmail.com

Mark Wadsworth: mew225@gmail.com

Kevin Boehme: kevinlboehme@gmail.com

Kaitlyn Hoyt: kbell037@gmail.com

Aaron Sharp: sharp.aaron.r@gmail.com

Brendan O'Fallon: brendan.ofallon@aruplab.com

John Kauwe: Kauwe@byu.edu

Perry Ridge: perry.ridge@byu.edu

Keywords: VCF, next-generation sequencing, variants, variant analysis

Abstract

Background. Since the advent of next-generation sequencing many previously untestable hypotheses have been realized. Next-generation sequencing has been used for a wide range of studies in diverse fields such as population and medical genetics, phylogenetics, microbiology, and others. However, this novel technology has created unanticipated challenges such as the large numbers of genetic variants. Each Caucasian genome has more than 4 million single nucleotide variants, insertions and deletions, copy number variants, and structural variants. Several formats have been suggested for storing these variants; however, the variant call format (VCF) has become the community standard.

Results. We developed new software called the Variant Tool Chest (VTC) to provide much needed tools to work with VCF files. VTC provides a variety of tools for manipulating, comparing, and analyzing VCF files beyond the functionality of existing tools. In addition, VTC was written to be easily extended with new tools.

Conclusions. Variant Tool Chest brings new and important functionality that complements and integrates well with existing software. VTC is available at https://github.com/mebbert/VariantToolChest

Background

The variant call format (VCF) has become the standard format for storing variants identified in next-generation sequencing (NGS) and other studies. VCF files are flexible with eight fixed fields including chromosome (CHROM), position (POS), known variant IDs such as dbSNP identifications (ID), reference allele (REF), alternate allele(s) (ALT), variant quality score (QUAL), filter information summarizing why a variant was or was not considered valid by

the variant calling software (FILTER), and an information field (INFO). Additional fields containing genotypes for one or more samples may also be present. Each row of the file contains information about observed variants at the given position and chromosome, may have information about how the variant(s) was/were identified (allele frequency, depth, strand bias, genotype likelihoods, etc.), and biological annotations (gene, variant frequency, 1000 Genomes membership, mRNA and protein positions, etc.). The last columns of a VCF file contain genotype information specifying whether the individual is heterozygous, homozygous reference or variant, or whether it is unknown (missing). Finally, VCF files can contain information for a single or multiple samples. Alternatively, summary VCF files containing minimal information (chromosome, position, reference allele, variant allele, and genotypes) can be used. VCF files are used to store all variant types including single nucleotide variants, insertions and deletions, copy number variants, and structural variants. The VCF has become an important format in modern biology and is the only widely used format for variant storage.

Several programs exist for manipulating and comparing VCF files: VCF tools [1], BedTools [2], BcfTools, and the Genome Analysis Toolkit (GATK) [3, 4]. Each of these softwares is flexible and powerful, but missing certain essential features. In this manuscript we describe a novel program, the Variant Tool Chest (VTC). The Variant Tool Chest complements existing softwares by extending their capabilities without replicating existing solutions for working with VCF files. We also provide suggestions for building upon the VTC rather than building new tools from scratch. VTC can be downloaded at https://github.com/mebbert/VariantToolChest.

Results and Discussion

Novel features. ***Multi-sample VCF support.*** As next-generation sequencing continues to gain momentum, researchers need the ability to compile many samples into a single VCF or analyze variants from multiple VCF files. VTC was built to work with a combination of multi- and single-sample VCF files. Existing softwares are only capable of handling either a single VCF file, or one multi-sample VCF file. VTC can handle a mix of single and multi-sample VCF files, with the user defining which sample(s) to use from each of the VCF files.

Genotype set operations. VTC contains a powerful set operation tool named "SetOperator" designed to perform simple or complex set operations using VCF files, including intersects, complements, and unions. While various tools exist to perform set operations on VCF files, VTC improves existing solutions in two ways. First, existing software performs set operations based only chromosome and base pair position. This means that if one individual is heterozygous and another homozygous, the resulting operations would assume that these two individuals have the same genotype. Second, existing tools work on only a collection of single sample VCF files. In contrast, VTC can perform set operations on a single multi-sample VCF file, or a combination of multi- and single sample VCF files. Furthermore, the user can choose to only perform the operations based on certain individuals from each multi-sample VCF file. These abilities save researchers time by not forcing the user to extract all samples of interest into a collection of single sample VCF files, and allow more efficient storage of genotypes in multi-sample VCF files. For example, it is helpful and makes sense for a researcher to store all genotypes for a single family in a single VCF file; however, the researcher may have interest in performing set operations across multiple families (VCF files), such as performing an intersection of variants from all affected individuals from all families.

VTC has several operation-specific settings for intersects and complements that allow researchers to specify genotype-level requirements. For intersects, VTC currently has five genotype-level intersect methods and two record-level (i.e., ignore genotypes) intersect methods. The genotype-level intersect methods are as follows: (1) heterozygous; (2) homozygous variant; (3) heterozygous or homozygous variant; (4) homozygous reference; and (5) match sample exactly across variant pools. The record-level intersect methods are: (1) variant; and (2) position.

The genotype-level intersect methods require that all sample genotypes involved in the intersect fall into the specified category. One caveat is that the heterozygous genotype requires the sample to have a reference allele. So if a sample's genotype has two different variant alleles (i.e. a tri-allelic position), though technically a heterozygote, will not be considered as such. This distinction is made assuming that researchers interested in identifying heterozygotes will assume the samples have a reference allele. This also greatly simplifies several corner cases when dealing with multiple variants at a single location.

The record-level intersect methods ignore genotypes and only consider whether the variant pools included in the analysis contain the variant. The "position" method only considers chromosome, position, and the reference allele, while the "variant" method also includes the alternate allele(s). For the "variant" method, records with multiple alternates are considered to intersect if at least one of the alternates matches.

There are currently three complement methods: (1) heterozygous or homozygous variant; (2) exact genotype matches; and (3) variant. When performing a complement, the "heterozygous or homozygous variant" method requires that all sample genotypes in both variant pool be either a heterozygous or homozygous variant in order to be removed from the variant pool being subtracted from. The "exact genotype" method requires that all samples across both variant pools

have the same genotype, whatever it may be. The "variant" method ignores genotypes and only subtracts if the chromosome, position, reference, and alternate match between the two variant pools.

Unions combine all variants and specified samples into a single VCF file regardless of genotype. Samples missing variants will have a "no call" genotype ("./.").

Detailed set operation syntax. The Set Operator tool in the VTC empowers researchers to define set operations with a powerful, simple syntax. This simple syntax has several advantages: (1) researchers may specify any number of input files (variant pools) to perform operations; (2) researchers may specify specific samples within a given variant pool to include in the operation; and (3) each operation is assigned an identification value (ID) automatically by VTC or specified by the user, so that it can be used in subsequent operations. The general syntax structure for a single operation is as follows (no spaces):

oId=operator[input_id1[sample_id1,sample_id2,etc.]:input_id2[sample_id3,sample_id4,etc.]:etc.]

where oId is a user-specified ID for the operation (may be omitted), operator is the operation of interest (i, c, or u for intersect, complement, or union), input_id is the variant pool ID, and sample_id is a sample ID for a sample within the given variant pool. If sample IDs are omitted, Set Operator will use all samples within the variant pool. For example, the following intersect operation will perform an intersect on all samples within the variant pools named "file1" and "file2": myOP=i[file1:file2].

Operation stringing. As previously mentioned, the set operation syntax allows resulting variant pools to be used in subsequent operations. This feature allows researchers to obtain final results with a single command in most circumstances. Continuing with the previous example,

"myOP" may be specified in a subsequent operation as follows: "myOP=i[file1:file2] myOP2=c[myOP:file3]".

Intermediate files. When performing complex set operations, researchers may want all intermediate operation results to be printed to a file. Otherwise, the researcher would be required to perform separate commands. As such, a simple option named "--intermediate-files" will print each operation result to a file named according to the specified "oId" previously mentioned.

Header repair. VCF files can be complex, and maintaining a valid VCF header can be challenging. Since VTC is built on the code that defines VCFs, it is possible to detect invalid VCF headers and repair them. VTC will automatically add missing required header information such as the "GT" header line when genotypes are being printed. There are many useful (unrequired) header lines that cannot be anticipated, however. This feature is still under active development.

Add/remove "chr." Chromosome numbers in VCF files may be prefixed by "chr" or may simply be the chromosome ID (e.g., chrX or X). Many next-generation sequencing softwares are incapable of handling VCF files that do not use the same convention simultaneously. For example, if one file includes "chr" and another does not, current tools will reject the files. And some tools require the VCF files to have the same chromosome ID as the reference sequenced used in the original analysis. VTC will either prepend or remove "chr" from all variant records seamlessly according to the user's specifications by simply including (or omitting) the "--add-chr" flag.

Summary information. Several tools exist that will provide high- or low-level detail on a variant pool, but they can be cumbersome. VTC has a tool named VarStats that will provide a quick summary of the variant pool, or a detailed variant-by-variant summary. High-level

summary metrics include total number of variants, total number of single nucleotide variants (SNVs), insertions and deletions, structural variants, and variants with multiple alternates. The summary also includes summary depth and quality values. The variant-by-variant summary includes allelic counts and the minimum, maximum, and average read depth and quality scores for each variant.

Compare operation. Many analyses require researchers to perform several set operations to identify all variants in common between VCFs, those that are unique to a given VCF, and researchers may also need the combined set. Researchers are generally not satisfied knowing only the number of variants that fall into each group, such as would be represented by a Venn diagram. To obtain all of this information a researcher would perform four set operations: an intersect (common variants), two complements (unique variants), and a union (combined set). Set Operator has a compare operation ("--compare") that will automatically perform all four operations, print the results to their respective files, and print a summary of each resulting variant pool to the console. This option currently is limited to two input files.

VCF association analysis. Association analyses are common using genomic data, but we are not aware of any available tools to perform such analyses on VCF files. The VarStats tool in VTC will perform association analyses on all variants in a variant pool if a phenotype file is provided. Results, including odds ratios and p-values for each variant are printed to a file. If there are multiple alternates at a given location, VarStats will perform the analysis on each alternate and print results on a separate line. This option does not currently provide p-value correction such as multiple test correction, but will be implemented in a future release. These corrections can be easily performed in statistical software.

Future Directions

Filter tool. Next-generation sequencing variants are often filtered on various values including quality scores and depth. Several tools already exist that, when combined, satisfy most needs for filtering variants. Ideally, a single tool would incorporate all of this functionality along with new features for simplicity.

File formats. While VCFs are the most common format for next-generation sequencing variants, there are other file formats that will be incorporated into VTC including Plink (ped/map or bim/bam/fam) and comma-separated value (CSV) files. Plink is particularly important since there are many existing large-datasets in Plink format. In order to compare or combine data in Plink format to those in VCF format, there must be a tool to handle this. VTC will enable researchers to read in variant data from multiple formats and perform all of the same analyses seamlessly. This is especially pertinent as a common QC measure of single nucleotide variants identified in NGS studies is to compare NGS variants to variants genotyped on a SNP array. Array data is most often reported in Plink format.

Enhanced compare. As different technologies are compared, there is a need to determine concordance between samples tested on multiple technologies. VTC will implement an "Enhanced Compare" option that will report genotypes that are perfect matches, imperfect matches (heterozygous variant observed on one technology and homozygous variant observed from the other), and no matches for the same samples on different technologies.

Additional SetOperator options. Anticipating all possible uses and hypotheses is difficult with any new tool, especially with data as complex as genomic variants. Responding to these needs is important and will likely involve updated SetOperator options. A few options we plan to implement are to accommodate specialized union operations, similar to those for intersect

and complement. Specifically, users may need to union only heterozygotes, heterozygotes or homozygous variant, only homozygous variant, or only homozygous reference.

Incorporate new and existing tools. Building useful computational tools that interface well together benefits researchers across all disciplines. New tools, while generally valuable to the research community, often do not integrate well with other tools used within a discipline, causing end users grief. There are likely many reasons for this fragmentation, but we would like to address two major sources: (1) contributing to an existing project can be costly (in time and money) and difficult; and (2) computational researchers need to publish their work to demonstrate academic productivity.

While object-oriented programming mitigates much of the difficulty, contributing to an existing project is still difficult because of the time and effort required to become familiar with existing source code. Many projects have hundreds of classes with complex interactions that make adding new functionality daunting. In many cases, a researcher may opt to write a separate tool simply because it is more feasible. Unfortunately, this causes fragmentation between tools. To promote well-integrated tools, VTC was written specifically to facilitate contribution with its easily extensible code structure. Any computational researcher can begin a new tool without needing to familiarize him/herself with other complex code.

Contributing to existing source code can be challenging, but publishing requirements also present a challenge to computational researchers, since publications are an essential measure of academic productivity. If a computational researcher adds a novel algorithm to an existing tool, s/he may forfeit the opportunity to publish the algorithm and get feedback from the community. Because VTC is simply a collection of useful tools, however, researchers can contribute an independent tool or algorithm with an independent name and publish it independently.

As we mentioned above, it is not possible to predict all possible operations and uses for software like VTC and we anticipate the need for additional functionality. To this end, we invite all computational researchers to contribute independent tools associated with variant analysis to VTC. This will benefit researchers by promoting tool integration within a simple, intuitive framework.

Conclusions

VCF files are the standard format for storing variants identified in next-generation sequencing (NGS) and other studies, but working with them can be challenging. In this manuscript we describe a novel program, the Variant Tool Chest (VTC). The Variant Tool Chest is easily extendable and complements existing softwares by extending their capabilities without replicating existing solutions for working with VCF files. VTC is available at https://github.com/mebbert/VariantToolChest

Methods

Variant tool chest overview. The Variant Tool Chest (VTC) is a collection of tools to analyze variants from next-generation sequencing (NGS) and other studies, and is intended to become a tool chest to accommodate most analysis needs. It is written in Java (version 1.7) for speed and portability. Two tools currently exist in the tool chest named SetOperator and VarStats. Set Operator performs set operations such as intersects, complements, and unions on variant sets termed variant pools. VarStats performs statistical operations including association analyses and summaries on variant pools. Since there are numerous other tools necessary for analyzing variant pools, VTC was written with an emphasis on extensibility.

Extensibility. To make VTC easily extensible, each tool is written independently and is self-contained within a single Java package. Researchers can add tools without being forced to familiarize and integrate with other complicated code. A single class named VTCEngine is the main entry for all tools. VTCEngine receives user input and executes the appropriate tool(s). Most arguments are passed to, and handled by the tool of interest. Each tool uses a simple argument-parsing library named Argparse4j [5] to define and handle all arguments. All tools use the same variant and sample data structures known as VariantPool and SamplePool, respectively.

VariantPool is built on the open source public application programming interfaces (APIs) distributed by the Broad Institute that define the Variant Call Format (VCF) file structure. Specifically, the VTC is built on the Picard [6], SAMTools [7], tribble, and variant APIs. Tribble provides necessary utilities for creating and working with various data file types, including VCF indexes. All three libraries are essential components incorporated into the Genome Analysis Toolkit (GATK) [3, 4]. As such, VTC is capable of reading and writing valid VCF files, dependably. For generalizability, data structure classes are contained within the main vtc.datastructures Java package. Any future classes generally applicable across multiple tools should also be defined within the vtc.datastructures package. Likewise, a class named UtilityBelt was created for methods that are generally applicable. The file structure of VTC can be seen in Figure 3.1.

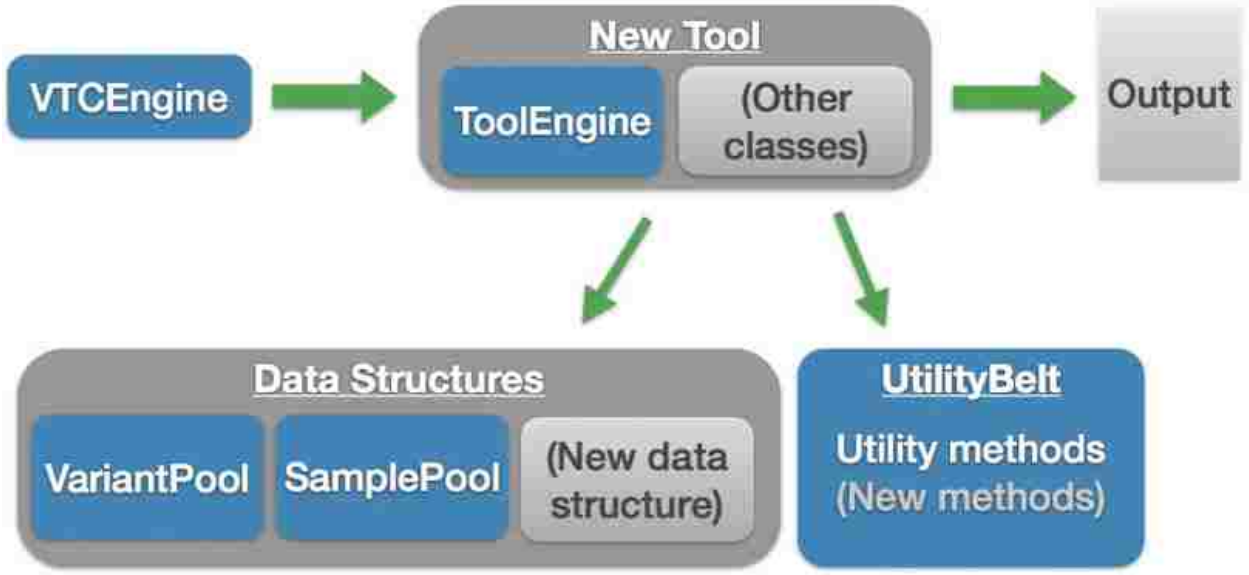


Figure 3.1. Variant tool chest (VTC). The VTC was built to be extensible. Each new tool only needs to interface with a few simple classes and is otherwise completely independent. All tools should be self-contained within a single parent Java package. The main driver class for VTC is VTCEngine. Any new tool should have its own Engine class and be instantiated from VTCEngine. All generally applicable data structures such as VariantPool and SamplePool are placed within the vtc.datastructure Java package. Any new generally applicable data structures should also be placed in vtc.datastructure. Otherwise the data structure should be housed within the tool's package. Likewise, any generally applicable methods should be placed in the UtilityBelt class.

Competing interests

All authors declare they have no competing interests.

Authors' contributions

ME participated in concept and software design, software writing, and manuscript writing; MW participated in software design and writing; KB participated in software design and writing; KH participated in software writing; AS participated in software design and writing; BO participated in software design; JK participated on concept design; PR conceived the concept and participated in concept design and manuscript writing. All authors read and approved the final manuscript.

Acknowledgments

We graciously acknowledge the resources provided for this work by grants from the NIH (R01AG042611), the Alzheimer's Association (MNIRG-11-205368), and startup funds from Brigham Young University.

References

1. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group: **The variant call format and VCFtools**. *Bioinforma. Oxf. Engl.* 2011, **27**:2156–2158.
2. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinformatics* 2010, **26**:841–842.
3. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res.* 2010, **20**:1297–1303.
4. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ: **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. *Nat. Genet.* 2011, **43**:491–498.
5. Tsujikawa T: *Argparse4j*. 2013.
6. *Picard*. 2013.
7. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup: **The Sequence Alignment/Map format and SAMtools**. *Bioinforma. Oxf. Engl.* 2009, **25**:2078–2079.

Chapter 4

Interaction between Genetic Variants in CLU and MS4A4E Modulates Risk for Alzheimer's Disease

Mark T.W. Ebbert[1], Kevin L. Boehme[1], Mark E. Wadsworth[1], ADGC, John S.K. Kauwe[1]

[1] Department of Biology, Brigham Young University, Provo, Utah

Corresponding Author:
John S. K. Kauwe
4102 LSB
Brigham Young University
Provo, UT 84602
Phone: 801-422-2993
Email: kauwe@byu.edu

Keywords: Alzheimer's disease; Epistasis; MS4A4E; CLU; CD33; Meta-Analysis

Abstract

Background. Ebbert et al. recently reported two potential gene-gene interactions between rs11136000 (*CLU*) and rs670139 (*MS4A4E*) (SF=3.81, p=.016), and rs3865444 (*CD33*) and rs670139 (*MS4A4E*) (SF=5.31, p=.003) using the Cache County data. Here, we evaluate those interactions in a large, independent dataset.

Methods. Using 32 independent data sets from the Alzheimer's Disease Genetics Consortium (ADGC), we tested each interaction, controlling for age, gender, and *APOE ε4* dose. We then performed two meta-analyses per interaction (ADGC only and with Cache) using METAL, and performed 10,000 permutations to obtain empirical p-values. We repeated the meta-analyses in *APOE ε4* carrier and non-carrier strata, estimated the combined population attributable fraction (cPAF) for both, and explored causal variants.

Results. Our results support the *CLU-MS4A4E* interaction (ADGC: SF=2.37, p=0.007; with Cache: SF=2.71, p=0.0004) and found a potential dosage effect using the ADGC data between rs11136000:C/C and rs670139:G/T (with Cache: SF=1.73, p=0.02). Empirical p-values obtained from permutations support the main interaction (ADGC: p=0.03; with Cache: p=0.002). The *CD33-MS4A4E* interaction did not replicate (ADGC: SF=1.16, p=0.78). We found an association for the *CLU-MS4A4E* interaction in Cache County for *APOE ε4* negative individuals (SF=4.75, p=0.005). This association only replicates including Cache (ADGC: SF=1.28, p=0.15; with Cache: SF=2.08, p=0.004). The estimated cPAF for *CLU* and *MS4A4E* is 8.0. We found no obvious causal variants.

Conclusions. We replicated the main *CLU-MS4A4E* interaction and provide evidence of a possible dosage and APOE *ε4* effect. We also estimate an 8% decrease in Alzheimer's disease incidence if the *CLU-MS4A4E* risk alleles were removed from the population.

Introduction

Alzheimer's disease (AD) is a common and complex neurodegenerative disease. It is the most common cause of dementia and is characterized by the accumulation of amyloid plaques and neurofibrillary tangles. To date, many genetic loci have been found that modify AD risk, but collectively, they explain only a fraction of the heritability of the disease (1) and are not diagnostically useful (2). It is hypothesized that rare variants with large effects as well as epistatic interactions account for much of the unexplained heritability in AD and have been largely hidden due to limitations in traditional GWAS studies. As such, rare variant and epistatic effects are poorly understood. Recent studies, however, have demonstrated that gene-gene interactions play a critical role in the etiology and progression of AD (2–5).

A recent study by Ebbert et al. (2) found evidence of two gene-gene interactions among three known AD genes that increase AD risk: *CLU-MS4A4E* and *CD33-MS4A4E*. Specifically, Ebbert et al. reported interactions between rs11136000 C/C (*CLU*) and rs670139 G/G (*MS4A4E*) genotypes (synergy factor = 3.81; p = .016), and the rs3865444 C/C (*CD33*) and rs670139 G/G (*MS4A4E*) genotypes (synergy factor = 5.31; p = .003). All three genes are on the "AlzGene Top Results" list, which summarizes the most established genes associated with AD to date.

In this study, we attempted to replicate these gene-gene interactions by performing an independent meta-analysis of data sets from the Alzheimer's Disease Genetics Consortium (ADGC), followed by a combined meta-analysis including the original Cache County data. We also tested for dosage effects in both interactions and an *APOE ε4* effect. We then performed a rigorous permutation experiment to test robustness of results that had a significant p-value in the independent analysis. We also explored possible causal variants using whole-genome sequence data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The main *CLU-MS4A4E*

interaction replicates in both the independent and combined meta-analysis, with minor evidence of a dosage effect. There is also minor evidence of an association between the *CLU-MS4A4E* interaction and case-control status in *APOE ε4* negative individuals. The *CD33-MS4A4E* interaction failed to replicate.

Methods

SNP data preparation and statistical analysis. We used SNP microarray data from ADGC in this study, which consists of 32 studies over two phases. More information about this dataset can be found in a previous report by Naj et al. (6) and the ADGC data preparation description (7).

Since gene-gene interactions are challenging to identify and replicate, we used only the highest quality data possible. For each ADGC data set, we filtered SNPs imputed with low information (info < .5) and then converted the IMPUTE2/SNPTEST format files to PLINK format, using PLINK v1.90b2i (8). We used the default PLINK uncertainty cutoff of .1, meaning any imputed call with uncertainty greater than .1 was treated as missing. We included SNPs with a missing genotype rate less than 0.05 (PLINK command '--geno 0.05'). After cleaning SNPs, we included only individuals with a missing rate less than 0.01 (PLINK option '--mind 0.1') to select only the samples with high genotyping rates. We then extracted the three SNPs of interest: rs3865444 (*CD33*), rs670139 (*MS4A4E*), and rs11136000 (*CLU*) and tested Hardy-Weinberg equilibrium (9; 10). Using R v3.1.1 (11), we then excluded all samples that did not have complete data for all covariates including age, gender, cohort, case-control status, *APOE ε4* dose, and the two SNPs being tested in the corresponding interaction. Any data sets missing the respective SNPs or covariates after data cleaning were excluded from further analysis.

Following data preparation, we tested the individual interactions in each data set using logistic regression. We performed logistic regressions in R using the covariates previously mentioned. We defined the R models as "case_control ~ rs3865444*rs670139 + apoe4dose + age + sex" and "case_control ~ rs11136000*rs670139 + apoe4dose + age + sex" for the *CD33-MS4A4E* and *CLU-MS4A4E* interactions, respectively, which include the main and interaction effects in the models. All analyses in this study used each gene's homozygous minor allele as the reference group.

Using results from each study, we performed a meta-analysis to test significance across the ADGC data sets using the METAL version released on 2011-03-25 (12), and performed a second meta-analysis including the Cache County results. We tested the originally reported interactions, along with heterozygous interactions (rs11136000 C/C interacting with rs670139 G/T and rs3865444 C/C interacting with rs670139 G/T) to check for potential dosage effects, based on suggestive evidence found in the original Cache County study (Supplemental Table 4.1). We also stratified the Cache County data by *APOE ε4* status and tested for an association with case-control status. Based on those results, we then tested for the same association in the ADGC data.

Following the meta-analyses, we performed a permutation analysis with 10,000 permutations for interactions that replicated in the independent data set. For each ADGC data set, we randomly permuted case-control status across all individuals, tested the interaction by logistic regression, and reran the meta-analysis. We stored the p-values from each of the 10,000 meta-analyses. We then calculated the empirical p-value by finding the original p-value's rank in the distribution of p-values divided by the number of permutations.

Our results are represented as synergy factors (4; 13) and their associated 95% confidence intervals and p-values. Synergy factors measure whether the effect size of two interacting genetic variants is greater than the sum. Similar to odds ratios, synergy factors less than one and greater than one suggest decreased and increased risk in case-control studies, respectively, as long as the appropriate reference group is used.

We calculated each synergy factor's 95% confidence interval for each meta-analysis, but omitted the ADC1 cohort, which had only a single case, inclusion of which made the 95% confidence interval for the summary synergy factor from 0 - ∞.

Exploring causal mutations. We explored causal mutations for confirmed interactions using 809 ADNI whole genomes that were sequenced, aligned to hg19, and variants called by Illumina using their internal analysis procedure. We used linkage disequilibrium, Regulome DB (accessed November 2014) (14), and functional annotations from wAnnovar (15) to isolate SNPs of interest. We first extracted all SNPs within approximately 50 kilobases of each SNP of interest, calculated linkage disequilibrium using Haploview (16), and retained all SNPs with a D' $\geq$ 0.99. Using Regulome DB and wAnnovar, we annotated each remaining SNP for: (1) known regulation and functional effects; (2) minor allele frequencies from the 1000 Genomes Project (17), 6500 Exomes Project (18), and the ADNI data set; and (3) corresponding MutationTaster predictions (19). We retained all SNPs with a Regulome DB score less than 4, and all SNPs located in untranslated (UTRs) or exonic regions (if nonsynonymous). For each retained SNP, we tested individual associations with case-control status and subsequently tested their interaction with all SNPs in the other interacting gene.

Results

Sample and data set demographics. Sample demographics and minor allele frequencies for each SNP are presented for each data set (Table 4.1). Nine of the 32 data sets passed quality controls for the *CD33-MS4A4E* interaction while seven passed for *CLU-MS4A4E*. The remaining data sets were either missing required SNP(s), missing a covariate, or consisted of only controls and could not be included in the analysis. All SNPs passed Hardy-Weinberg equilibrium in all data sets for both cases and controls.

Interaction and dosage meta-analysis results. The originally reported *CLU-MS4A4E* interaction between the rs11136000 C/C (*CLU*) and rs670139 G/G (*MS4A4E*) genotypes replicates in both the independent (synergy factor = 2.37, p = 0.007; Figure 4.1b) and combined (synergy factor = 2.71, p = 0.0004; Figure 4.1b) meta-analyses (Supplemental Table 4.1), with minor evidence for a dosage effect in the combined meta-analysis (synergy factor = 1.73, p = 0.02; Figure 4.1a). Empirical p-values obtained from permutations support the main interaction (ADGC: p = 0.03; with Cache: p = 0.002). We found an association with case-control status in people with no *APOE ε4* alleles in the Cache County data alone (synergy factor = 4.75, p = 0.005; Supplemental Table 4.2) that did not exist in people with one or more *APOE ε4* alleles (synergy factor = 1.22, p = 0.74; Figure 4.2b; Supplemental Table 4.2). The association in *APOE ε4* negative subjects did not replicate in the meta-analysis across the ADGC (synergy factor = 1.28, p = 0.15; Figure 4.2b; Supplemental Table 4.2), though it was significant when including the Cache County data (synergy factor = 2.08, p = 0.004; Figure 4.2b; Supplemental Table 4.2). The *CD33-MS4A4E* interaction failed to replicate in either the independent (synergy factor = 1.16, p = 0.78; Figures 4.3a and 4.3b) or combined (synergy factor = 1.63, p = 0.24; Figures 4.3a and 4.3b) meta-analyses.

Table 4.1. Sample Demographics by Data Set

Study	N	Cases (%)	Females (%)	Age	APOE 4 + (%)	rs670139 MAF (T)	rs3865444 MAF (A)	rs11136000 MAF (T)
ACT1	1858	487 (26.2)	1068 (57.5)	82.28	526 (28.3)	0.41	0.33	NA
ADC1	388	1 (00.3)	237 (61.1)	74.99	129 (33.2)	0.41	0.33	0.39
ADC2	681	566 (83.1)	365 (53.6)	79.38	394 (57.9)	0.42	0.33	0.39
ADNI	371	230 (62.0)	157 (42.3)	77.82	201 (54.2)	0.45	0.31	0.37
LOAD	2965	1515 (51.1)	1882 (63.5)	78.22	1667 (56.2)	0.43	0.31	0.38
TARC1	388	244 (62.9)	244 (62.9)	78.96	189 (48.7)	0.43	0.32	0.41
UMVUMSSM_A	1058	450 (42.5)	676 (63.9)	75.48	451 (42.6)	0.43	0.31	0.38
UMVUMSSM_B	390	135 (34.6)	236 (60.5)	73.99	118 (30.3)	0.41	0.33	0.38
UMVUMSSM_C	271	210 (77.5)	160 (59.0)	74.77	167 (61.6)	0.42	0.29	NA

Note. For each dataset the following information is provided: percent cases and females, age, *APOE ε4* positive percentage, and minor allele frequencies for rs670139, rs3865444, and rs11136000.

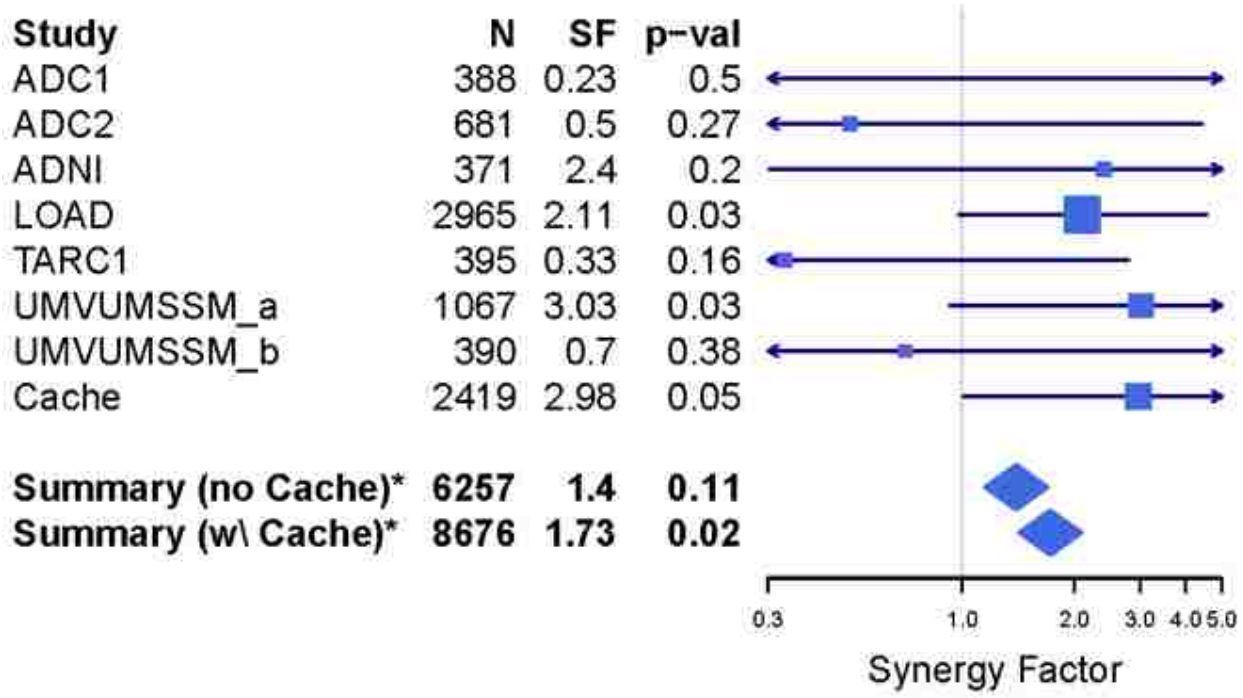


Figure 4.1a. Forest plot showing *CLU-MS4A4E* interaction replication with potential dosage effect: Original interaction test. We tested the original interaction, which replicated in both the independent and combined meta-analyses (figure b). We also tested for a dosage effect, which did not replicate independently, but does in the combined (figure a). The ADC1 data set was omitted when calculating the 95% confidence interval for the meta-analysis synergy factor because the data set only had 1 case, giving a standard error from 0 - ∞.

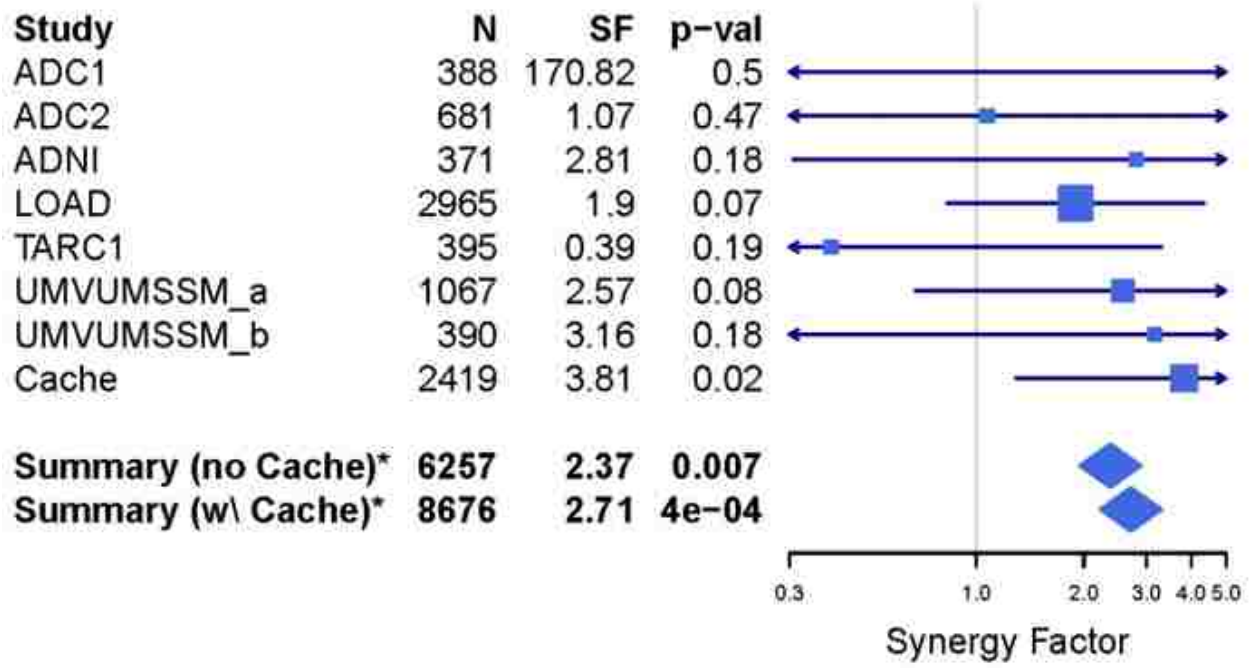


Figure 4.1b. Forest plot showing *CLU-MS4A4E* interaction replication with potential dosage effect: Dosage effect test. We tested the original interaction, which replicated in both the independent and combined meta-analyses (figure b). We also tested for a dosage effect, which did not replicate independently, but does in the combined (figure a). The ADC1 data set was omitted when calculating the 95% confidence interval for the meta-analysis synergy factor because the data set only had 1 case, giving a standard error from 0 - ∞.

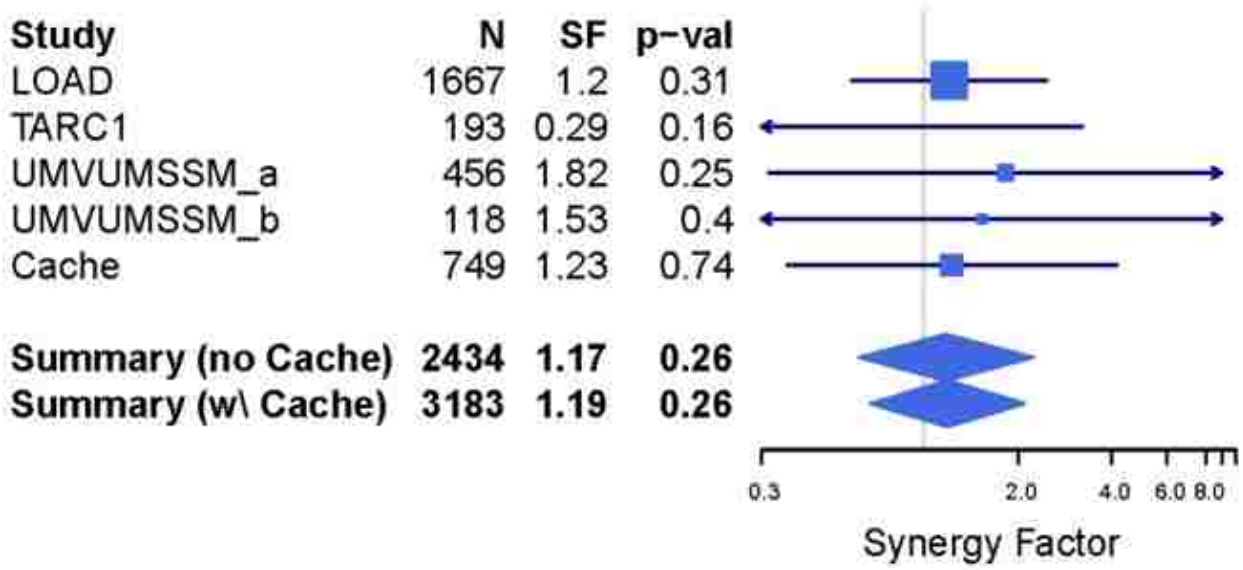


Figure 4.2a. Forest plot showing *APOE ε4* negative association with Alzheimer's disease case-control status: Independent meta-analysis. We stratified the Cache County data by *APOE ε4* status and tested for an association with Alzheimer's disease case-control status. We found an association in the *APOE ε4* negative stratum in Cache County that did not replicate in the independent meta-analysis, but did in the combined analysis (figure b). There was no association in the *APOE ε4* positive stratum.

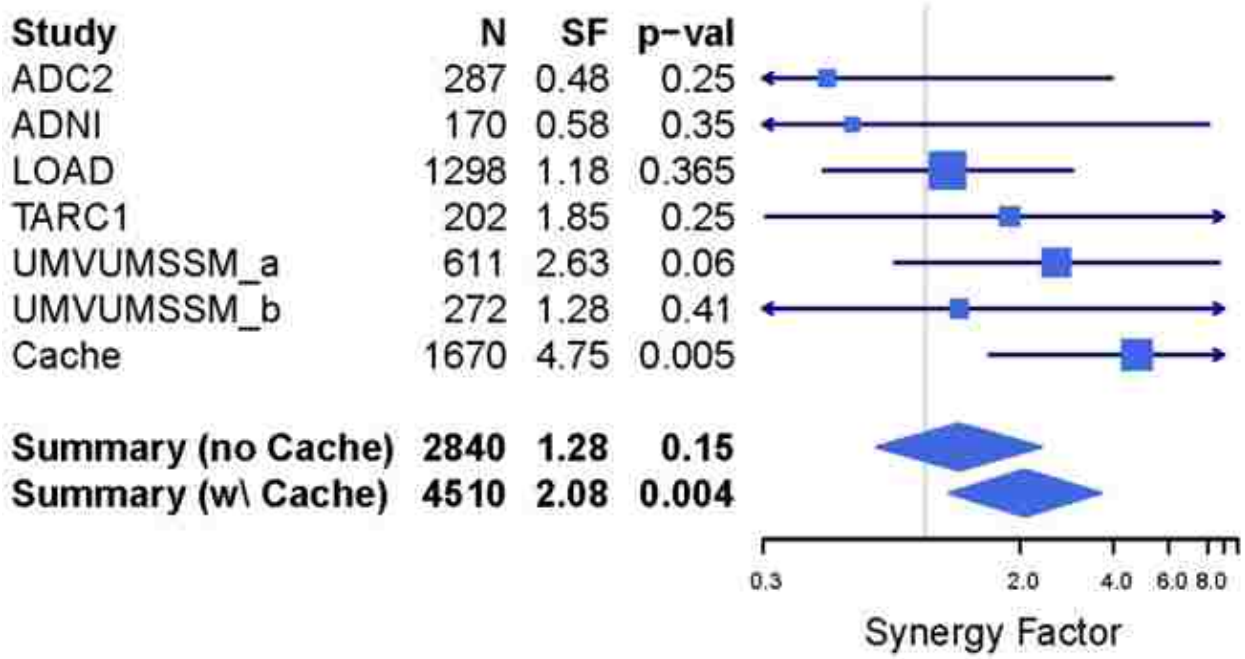


Figure 4.2b. Forest plot showing *APOE ε4* negative association with Alzheimer's disease case-control status: Combined analysis. We stratified the Cache County data by *APOE ε4* status and tested for an association with Alzheimer's disease case-control status. We found an association in the *APOE ε4* negative stratum in Cache County that did not replicate in the independent meta-analysis, but did in the combined analysis (figure b). There was no association in the *APOE ε4* positive stratum.

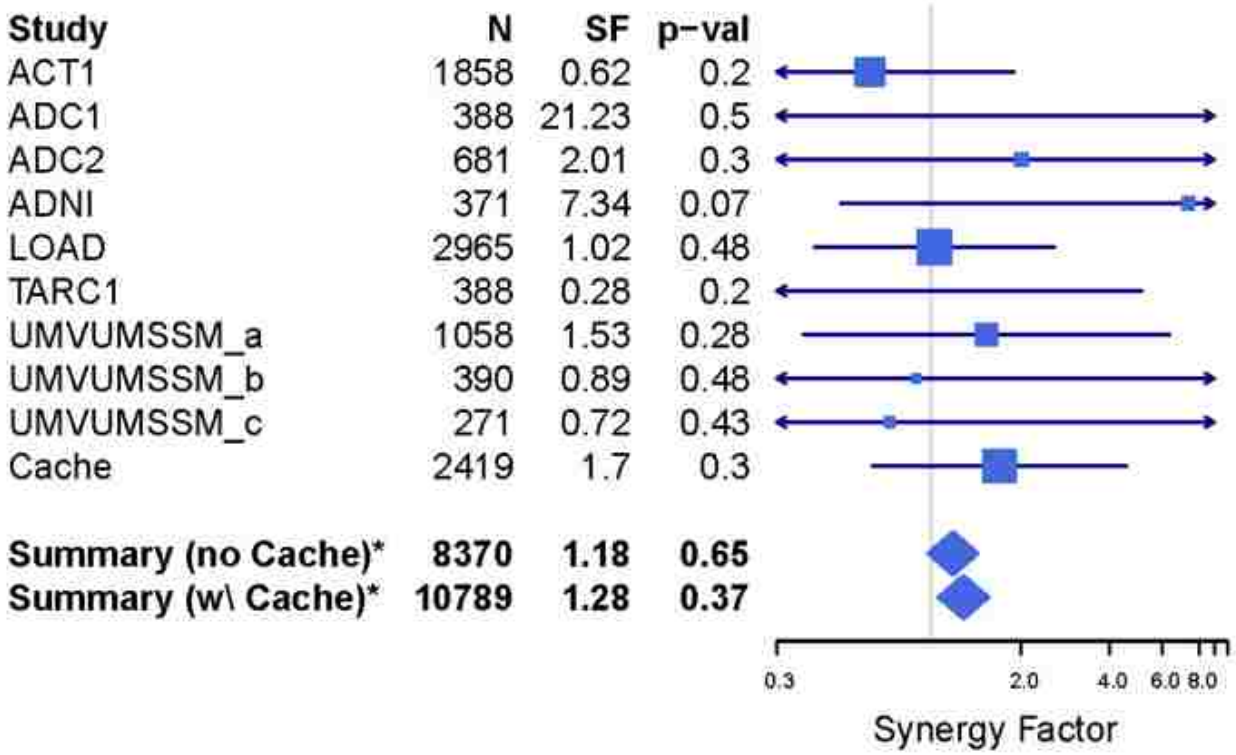


Figure 4.3a. Forest plot showing *CD33-MS4A4E* failed replication of interaction and dosage effect: Independent meta-analysis. We tested the original interaction, which did not replicate in either the independent or combined meta-analyses (figure b). We also tested for a dosage effect, which also did not exist. The ADC1 data set was omitted when calculating the 95% confidence interval for the meta-analysis synergy factor because the data set only had 1 case, giving a standard error from 0 - ∞.

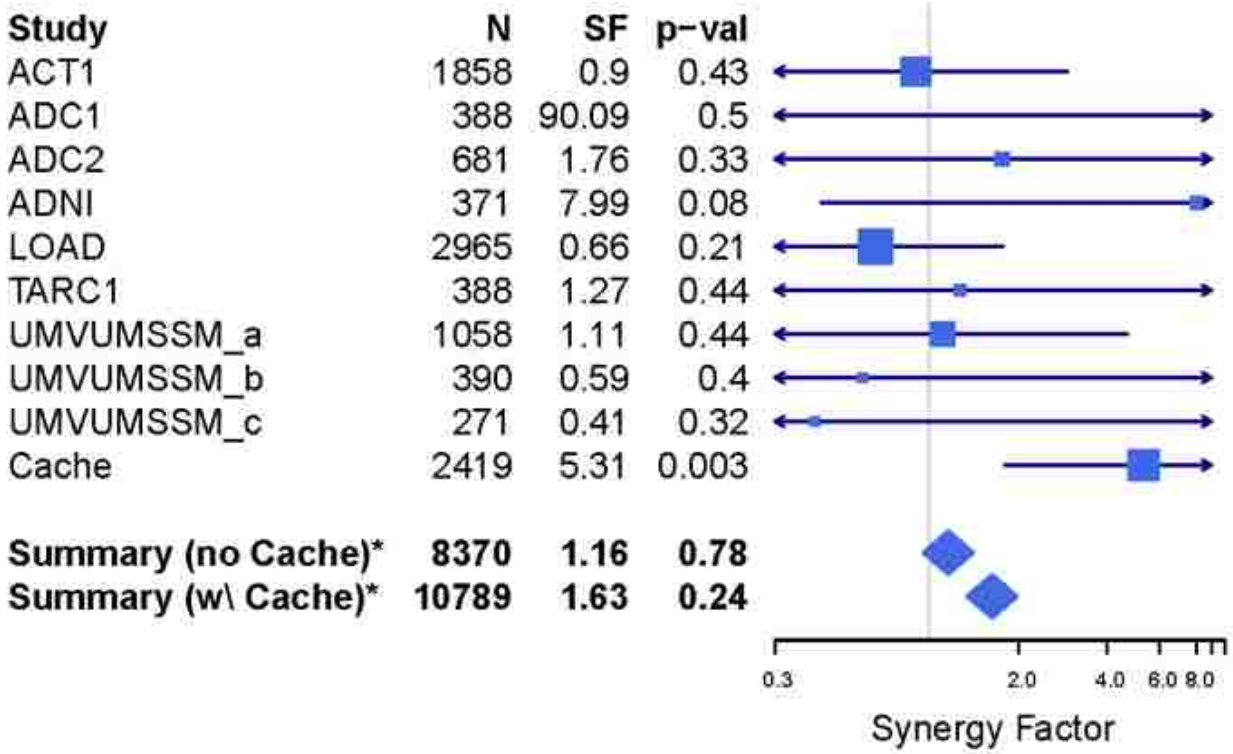


Figure 4.3b. Forest plot showing *CD33-MS4A4E* failed replication of interaction and dosage effect: Combined analysis. We tested the original interaction, which did not replicate in either the independent or combined meta-analyses (figure b). We also tested for a dosage effect, which also did not exist. The ADC1 data set was omitted when calculating the 95% confidence interval for the meta-analysis synergy factor because the data set only had 1 case, giving a standard error from 0 - ∞.

Exploring causal mutations. Since only the *CLU-MS4A4E* interaction replicated, we only explored causal SNPs within these genes. There were 36 and 32 SNPs that fit the inclusion criteria previously described for SNPs in the regions of rs11136000 and rs670139, respectively (Supplemental Tables 4.3 and 4.4). Most of the SNPs are rare (MAF < 0.01) according to the 1000 Genomes, 6500 Exomes, and ADNI data sets. None of the SNPs were significantly associated with case-control status individually. The pairwise interaction association tests between all included SNPs near and including rs11136000 (*CLU*) and rs670139 (*MS4A4E*) revealed an interaction between rs670139 and rs1532278 (synergy factor = 1.83, p = 0.01 unadjusted). The SNP rs1532278 was previously identified by Naj et al. (6) as being associated with case-control status. There were three suggestive interactions between rs9331931 (*CLU*, intronic) and the following: (1) rs7926344 (synergy factor = 0.53, p = 0.06); (2) rs2081547 (synergy factor = 0.53, p = 0.06); and (3) rs11230180 (synergy factor = 0.54, p = 0.07). SNPs rs2081547 and rs11230180 are interesting because they have been shown to modify expression of *MS4A4A* (20), the gene upstream from *MS4A4E*. They also have a Regulome DB score of '1f', meaning they are known to modify expression and are known DNase and transcription factor binding sites.

Discussion

In this study we attempted to replicate two gene-gene interactions and their association with Alzheimer's disease case-control status. The *CD33-MS4A4E* interaction failed to replicate, and may have resulted from over-fitting in the Cache County data. The CD33 protein interacts indirectly with a protein related to MS4A4E known as MS4A2 by physically interacting with the CBL protein that interacts with MS4A2. Both *MS4A4E* and *MS4A2* are members of the membrane-spanning 4-domain gene family, giving credence to an interaction between CD33 and

MS4A4E. Statistical evidence for this interaction is lacking, however, and more analyses may be necessary to draw more definitive conclusions.

We replicated the *CLU-MS4A4E* interaction, and further demonstrated some evidence of a dosage effect for *MS4A4E* along with a potential association in *APOE ε4* negative subjects. The interaction between the rs11136000 (*CLU*) C/C and rs670139 (*MS4A4E*) G/G genotypes is significant in both the independent meta-analysis (synergy factor = 2.37, p = 0.007) using only the ADGC data sets and the combined meta-analysis (synergy factor = 2.71, p = 0.0004) including the Cache County data, suggesting it may be valid. There is, however, a distinction to be made regarding statistical epistasis and biological epistasis. While there is evidence that CLU, like CD33, interacts indirectly with MS4A2, little is known about MS4A4E itself and we do not know whether it biologically interacts with CLU. MS4A2 indirectly modifies BCL2L1 activation or expression (2), which physically interacts with CLU. Research suggests CLU prevents amyloid fibrils and other protein aggregation events (Yerbury et al 2007) while MS4A4E may facilitate aggregation as a membrane-spanning protein. Membrane-spanning proteins play diverse roles in cell activity including transport and signaling. Experiments will be required to determine whether there is biological epistasis between CLU and MS4A4E, and whether the interaction affects amyloid fibril formation. These results indicate further investigative efforts in gene-gene interactions (and protein-protein interactions) may be important to resolve Alzheimer's disease etiology.

We tested for evidence of an *APOE ε4* effect in the Cache County data and found a significant effect in *APOE ε4* negative subjects and no significant effect in *APOE ε4* positive subjects. Subsequent meta-analyses with the ADGC data suggest this effect may be valid, though it only replicates when including the original Cache County result.

Since all analyses in this study used each gene's homozygous minor allele as the reference group, the interactions between major alleles are framed as a risk factor, meaning the interaction between the minor alleles is protective. The minor allele for *CLU* is protective as is being *APOE ε4* negative, while the minor allele for *MS4A4E* increases risk. The interaction between *CLU* and *MS4A4E* from the minor allele perspective is protective.

We found no obvious causal variants linked to rs11136000 or rs670139 with a D' $\geq$ 0.99 in the ADNI whole-genome data, though we believe further analysis of both rs11230180 (*MS4A4E*) and rs2081547 (*MS4A4E*) are warranted given their known expression effect on *MS4A4A*. SNP rs9331931 (*CLU*) has minimal regulome evidence, but is also worth further investigation.

A major gap in Alzheimer's disease literature to date is the lack of known causal variants. Several SNPs have repeatedly turned up in genome-wide association studies, but the tagSNPs themselves are unlikely to play a direct role in Alzheimer's disease etiology. What is more likely is that the tagSNPs are in close linkage disequilibrium with one or more causal variants. We believe there are two explanations: (1) the SNPs are linked to multiple rare variants that drive Alzheimer's disease development and progression; or (2) there is another common variant in the region with functional effects that remain unknown. In either case, given the biological complexity of Alzheimer's disease and results presented in this study, we believe epistasis plays a critical role in Alzheimer's disease etiology. As such, the community must continue to identify and vet these and other interactions that are supported in the literature.

Acknowledgements

This work was supported by grants from NIH (R01AG11380, R01AG21136, R01AG31272, R01AG042611, R01 AG 042437), the Alzheimer's Association (MNIRG-11-

205368) and the Utah Science, Technology, and Research initiative (USTAR), the Utah State University Agricultural Experiment Station, and the Brigham Young University Gerontology Program. The authors thank the participants and staff of the many centers that were involved in data collection for their important contributions to this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Financial Disclosures

Authors report no conflicts of interest, financial or otherwise.

References

1. Ridge PG, Mukherjee S, Crane PK, Kauwe JSK, Alzheimer's Disease Genetics Consortium (2013): Alzheimer's Disease: Analyzing the Missing Heritability. *PLoS ONE* 8: e79771.

2. Ebbert MTW, Ridge PG, Wilson AR, Sharp AR, Bailey M, Norton MC, *et al.* (2013): Population-based Analysis of Alzheimer's Disease Risk Alleles Implicates Genetic Interactions. *Biol Psychiatry*. doi: 10.1016/j.biopsych.2013.07.008.

3. Bullock JM, Medway C, Cortina-Borja M, Turton JC, Prince JA, Ibrahim-Verbaas CA, *et al.* (n.d.).: Discovery by the Epistasis Project of an epistatic interaction between the GSTM3 gene and the HHEX/IDE/KIF11 locus in the risk of Alzheimer's disease. *Neurobiol Aging*. doi: 10.1016/j.neurobiolaging.2012.08.010.

4. Combarros O, Cortina-Borja M, Smith AD, Lehmann DJ (2009): Epistasis in sporadic Alzheimer's disease. *Neurobiol Aging* 30: 1333–1349.

5. Kauwe JSK, Bertelsen S, Mayo K, Cruchaga C, Abraham R, Hollingworth P, *et al.* (2010): Suggestive synergy between genetic variants in TF and HFE as risk factors for Alzheimer's disease. *Am J Med Genet Part B Neuropsychiatr Genet Off Publ Int Soc Psychiatr Genet* 153B: 955–959.

6. Naj AC, Jun G, Beecham GW, Wang L-S, Vardarajan BN, Buros J, *et al.* (2011): Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet* 43: 436–441.

7. Kevin L Boehme, Shubhabrata Mukherjee, Paul K Crane, John S K Kauwe (2014, September): ADGC 1000 Genomes combined data workflow. Retrieved from kauwelab.byu.edu/Portals/22/adgc_combined_1000G_09192014.pdf.

8. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* (2007): PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* 81: 559–575.

9. Wigginton JE, Cutler DJ, Abecasis GR (2005): A Note on Exact Tests of Hardy-Weinberg Equilibrium. *Am J Hum Genet* 76: 887–893.

10. Graffelman J, Moreno V (2013): The mid p-value in exact tests for Hardy-Weinberg equilibrium. *Stat Appl Genet Mol Biol* 12: 433–448.

11. R Development Core Team (2011): *R: A Language and Environment for Statistical Computing*. . Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

12. Willer CJ, Li Y, Abecasis GR (2010): METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26: 2190–2191.

13. Cortina-Borja M, Smith AD, Combarros O, Lehmann D (2009): The synergy factor: a statistic to measure interactions in complex diseases. *BMC Res Notes* 2: 105.

14. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, *et al.* (2012): Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 22: 1790–1797.

15. Chang X, Wang K (2012): wANNOVAR: annotating genetic variants for personal genomes via the web. *J Med Genet* 49: 433–436.

16. Barrett JC, Fry B, Maller J, Daly MJ (2005): Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21: 263–265.

17. Consortium T 1000 GP (2012): An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56–65.

18. NHLBI GO Exome Sequencing Project (ESP) (2014, November): . Retrieved from http://evs.gs.washington.edu/EVS/.

19. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D (2010): MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 7: 575–576.

20. Zeller T, Wild P, Szymczak S, Rotival M, Schillert A, Castagne R, *et al.* (2010): Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PloS One* 5: e10693.

Supplemental Table 4.1. Independent and Combined Meta-Analyses Replicate *CLU-MS4A4E* Interaction, but *CD33-MS4A4E* Fails to Replicate

Gene interaction (rs)	Genotype	Study (direction)	Synergy factor (95% CI)	p-value	N
CLU:MS4A4E (rs11136000:rs670139)	CC:GT	ADC1 (-)	0.23 (0 - ∞)	0.50	388
		ADC2 (-)	0.50 (0.06 - 4.42)	0.27	681
		ADNI (+)	2.40 (0.30 - 19.19)	0.20	371
		LOAD (+)	2.11 (0.98 - 4.55)	0.03	2965
		TARC1 (-)	0.33 (0.04 - 2.80)	0.16	395
		UMVUMSSM_a (+)	3.03 (0.92 - 9.93)	0.03	1067
		UMVUMSSM_b (-)	0.70 (0.07 - 7.34)	0.38	390
		Cache County (+)	2.98 (1.003 - 8.86)	0.05	2419
		Meta (no Cache)	1.40 (0 - ∞)	0.11	6257
		Meta (with Cache)	1.73 (0 - ∞)	0.02	8676
	CC:GG	ADC1 (+)	170.82 (0 - ∞)	0.50	388
		ADC2 (+)	1.07 (0.13 - 8.83)	0.47	681
		ADNI (+)	2.81 (0.30 - 25.70)	0.18	371
		LOAD (+)	1.90 (0.83 - 4.35)	0.07	2965
		TARC1 (-)	0.39 (0.05 - 3.30)	0.19	395
		UMVUMSSM_a (+)	2.57 (0.67 - 9.83)	0.08	1067
		UMVUMSSM_b (+)	3.16 (0.27 - 36.74)	0.18	390
		Cache County (+)	3.81 (1.28 - 11.32)	0.02	2419
		Meta (no Cache)	2.37 (0 - ∞)	0.007	6257
		Meta (with Cache)	2.71 (0 - ∞)	0.0004	8676
CD33:MS4A4E (rs3865444:rs670139)	CC:GT	ACT1 (-)	0.62 (0.20 - 1.90)	0.20	1858
		ADC1 (+)	21.23 (0 - ∞)	0.50	388
		ADC2 (+)	2.01 (0.16 - 25.46)	0.30	681
		ADNI (+)	7.34 (0.49 - 109.57)	0.07	371
		LOAD (+)	1.02 (0.40 - 2.60)	0.48	2965
		TARC1 (-)	0.28 (0.02 - 5.14)	0.20	388
		UMVUMSSM_a (+)	1.53 (0.37 - 6.37)	0.28	1058
		UMVUMSSM_b (-)	0.89 (0.02 - 42.43)	0.48	390
		UMVUMSSM_c (-)	0.72 (0.02 - 23.49)	0.43	271

Gene interaction (rs)	Genotype	Study (direction)	Synergy factor (95% CI)	p-value	N
		Cache County (+)	1.70 (0.63 - 4.58)	0.30	2419
		Meta (no Cache)	1.18 (0 - ∞)	0.65	8370
		Meta (with Cache)	1.28 (0 - ∞)	0.37	10789
	CC:GG	ACT1 (-)	0.90 (0.28 - 2.93)	0.43	1858
		ADC1 (+)	90.09 (0 - ∞)	0.50	388
		ADC2 (+)	1.76 (0.13 - 23.77)	0.33	681
		ADNI (+)	8.00 (0.43 - 148.42)	0.08	371
		LOAD (-)	0.66 (0.24 - 1.78)	0.21	2965
		TARC1 (+)	1.27 (0.05 - 30.14)	0.44	388
		UMVUMSSM_a (+)	1.11 (0.26 - 4.69)	0.44	1058
		UMVUMSSM_b (-)	0.59 (0.01 - 29.50)	0.40	390
		UMVUMSSM_c (-)	0.41 (0.01 - 19.21)	0.32	271
		Cache County (+)	5.31 (1.79 - 15.77)	0.003	2419
		Meta (no Cache)	1.16 (0 - ∞)	0.78	8370
		Meta (with Cache)	1.63 (0 - ∞)	0.24	10789

Note. Logistic regression results from each data set and both meta-analyses are shown for the main CLU-MS4A4E and CD33-MS4A4E interactions and their respective dosage analyses. The CLU-MS4A4E interaction replicates in both independent and combined meta-analyses. There is also evidence of a dosage effect when including the Cache County data in the meta-analysis. The CD33-MS4A4E interaction fails to replicate.

Supplemental Table 4.2. Minor Evidence of an Association with Alzheimer's Disease Case-Control Status in *APOE ε4* Negative Individuals

APOE ε4 status	Study (direction)	Synergy factor (95% CI)	p-value	N
+	ADC1 (NA)	NA	NA	129
	ADC2 (NA)	NA	NA	394
	ADNI (NA)	NA	NA	201
	LOAD (+)	1.20 (0.59 -	0.31	1667
	TARC1 (-)	0.29 (0.03 -	0.16	193
	UMVUMSSM_a (+)	1.82 (0.32 -	0.25	456
	UMVUMSSM_b (+)	1.53 (0.06 -	0.40	118
	Cache County (+)	1.22 (0.36 -	0.74	749
	Meta (no Cache)	1.17 (0.62 -	0.26	3158
	Meta (with Cache)	1.19 (0.67 -	0.26	3907
-	ADC1 (NA)	NA	NA	259
	ADC2 (-)	0.48 (0.06 -	0.25	287
	ADNI (-)	0.58 (0.04 -	0.35	170
	LOAD (+)	1.18 (0.47 -	0.37	1298
	TARC1 (+)	1.85 (0.31 -	0.25	202
	UMVUMSSM_a (+)	2.63 (0.79 -	0.06	611
	UMVUMSSM_b (+)	1.28 (0.18 -	0.41	272
	Cache County (+)	4.75 (1.59 -	0.005	1670
	Meta (no Cache)	1.28 (0.70 -	0.15	3099
	Meta (with Cache)	2.08 (1.19 –	0.004	4769

Note. We found a significant association between Alzheimer's disease case-control status and *APOE ε4* negative individuals in the Cache County data set, while no association existed in *APOE ε4* positive individuals. The association in *APOE ε4* negative individuals did not replicate independently, but was significant when including the Cache County data in the meta-analysis.

Supplemental Table 4.3. Top Variants in Linkage Disequilibrium with rs11136000 (*CLU*) that Have a Regulome DB Score Less than 4, or Are Located in UTR or Exonic Regions

SNP	SNP	Ref	Alt	Position	Gene	Regulome DB	Function	Function consequence	MAF (ADNI)	MAF (1000G)	MAF (esp6500)	Mutation taster score (prediction)
rs11136000 (CLU)	chr8:27441327	C	T	27441327	EPHX2, CLU	2a	intergenic		0.001	0.004		
	rs9331945	A	G	27454957	CLU	2b	UTR3		0.014	0.004		
	rs9331892	G	A	27468005	CLU	2b	exonic	synonymous	0.007	0.05	0.06	
	chr8:27445942	G	A	27445942	EPHX2, CLU	2b	intergenic		0.006	0.003		
	chr8:27449609	G	A	27449609	EPHX2, CLU	2b	intergenic		0.006	0.004		
	chr8:27452473	C	T	27452473	EPHX2, CLU	2b	intergenic		0.001			
	rs1532278	T	C	27466315	CLU	2b	intronic		0.62	0.72		
	chr8:27468411	T	C	27468411	CLU	2b	intronic		0.001	0.003		
	rs56121659	C	T	27469064	CLU	2b	intronic		0.001	0.01		
	rs9331886	C	T	27469066	CLU	2b	intronic		0.003	0.02		
	chr8:27471977	G	C	27471977	CLU	2b	intronic		0.001			
	rs77336101	G	A	27474871	CLU, SCARA3	2b	intergenic		0.02	0.02		
	chr8:27475208	C	G	27475208	CLU, SCARA3	2b	intergenic		0.002	0.0005		
	rs73560231	C	T	27478302	CLU, SCARA3	2b	intergenic		0.002	0.003		
	chr8:27491389	G	A	27491389	SCARA3	2b	upstream		0.002			
	chr8:27494300	G	C	27494300	SCARA3	2b	intronic		0.001			
	rs73679246	G	A	27463156	CLU	2c	intronic		0.009	0.07		
	rs73558162	G	A	27423389	EPHX2, CLU	3a	intergenic		0.001	0.01		
	rs78590228	G	T	27442119	EPHX2, CLU	3a	intergenic		0.001	0.01		
	chr8:27445866	C	T	27445866	EPHX2, CLU	3a	intergenic		0.009	0.0009		
	chr8:27448407	A	G	27448407	EPHX2, CLU	3a	intergenic		0.003	0.0005		
	chr8:27452662	A	G	27452662	EPHX2, CLU	3a	intergenic		0.001			
	rs9331931	G	C	27458104	CLU	3a	intronic		0.28	0.15		
	chr8:27461286	C	A	27461286	CLU	3a	intronic		0.001			
	chr8:27483098	C	T	27483098	CLU, SCARA3	3a	intergenic		0.001			
	rs56276902	A	G	27511118	SCARA3	3a	intronic		0.002			
	chr8:27472251	G	T	27472251	CLU	4	UTR5		0.003	0.003	0.004	0.997 (N)
	chr8:27491676	C	T	27491676	SCARA3	4	UTR5		0.001	0.0009		
	chr8:27507233	G	A	27507233	SCARA3	5	exonic	non-synonymous	0.001		0.0002	0.881 (N)
	rs9331938	C	T	27457479	CLU	7	exonic	non-synonymous	0.002	0.005	0.01	1 (N)
	rs9331936	T	G	27457512	CLU	7	exonic	non-synonymous	0.008	0.06	0.07	0 (P)
	chr8:27462461	G	A	27462461	CLU	7	exonic	non-	0.001	0.0005	0.001	1 (N)

SNP	SNP	Ref	Alt	Position	Gene	Regulome DB	Function	Function consequence	MAF (ADNI)	MAF (1000G)	MAF (esp6500)	Mutation taster score (prediction)
	rs41276297	G	A	27462662	CLU	7	exonic	synonymous non-synonymous	0.001	0.001	0.003	1 (N)
	chr8:27516016	C	T	27516016	SCARA3	7	exonic	non-synonymous	0.001			1 (N)
	chr8:27516732	C	T	27516732	SCARA3	7	exonic	non-synonymous	0.001		0.00008	1 (D)
	rs3735754	G	A	27516955	SCARA3	7	exonic	non-synonymous	0.002	0.04	0.001	0.997 (D)

Note. Using the ADNI whole-genome data, we used linkage disequilibrium, Regulome DB (accessed November 2014), and functional annotations from wAnnovar to isolate SNPs of interest. We first extracted all SNPs within approximately 50 kilobases of each SNP of interest, calculated linkage disequilibrium using Haploview, and retained all SNPs with a D' $\geq$ 0.99. Using Regulome DB and wAnnovar, we annotated each remaining SNP for: (1) known regulation and functional effects; (2) minor allele frequencies from the 1000 Genomes Project, 6500 Exomes Project, and the ADNI data set; and (3) corresponding MutationTaster predictions. We retained all SNPs with a Regulome DB score less than 4, and all SNPs located in untranslated (UTRs) or exonic regions (if nonsynonymous). No SNPs were significantly associated with case-control status in the ADNI whole-genome data.

Supplemental Table 4.4. Top Variants in Linkage Disequilibrium with rs670139 (*MS4A4E*) that Have a Regulome DB Score Less than 4, or Are Located in UTR or Exonic Regions

SNP	SNP	Ref	Alt	Position	Gene	Regulome DB	Function	Function consequence	MAF (ADNI)	MAF (1000G)	MAF (esp6500)	Mutation taster score (prediction)
rs670139	rs11230180	G	T	59961486	MS4A6A, AB231731	1f	intergenic		0.36	0.27		
(MS4A4E)	rs2081547	C	T	59989430	AB231729, AB231731	1f	ncRNA_intronic		0.37	0.31		
	chr11:59940500	C	T	59940500	MS4A6A, MS4A6A	2b	exonic;splicing	non-synonymous	0.004	0.005	0.003	0.996 (N)
	chr11:59936960	G	A	59936960	MS4A2, MS4A6A	2b	intergenic		0.001			
	chr11:59961500	A	G	59961500	MS4A6A, AB231731	2b	intergenic		0.003			
	rs79917136	T	G	59961877	MS4A6A, AB231731	2b	intergenic		0.006	0.01		
	chr11:59962069	T	A	59962069	MS4A6A, AB231731	2b	intergenic		0.02	0.01		
	rs76834915	G	T	59936781	MS4A2, MS4A6A	2c	intergenic		0.002	0.01		
	rs79315507	T	G	59923606	MS4A2, MS4A6A	3a	intergenic		0.001	0.01		
	chr11:59926285	A	G	59926285	MS4A2, MS4A6A	3a	intergenic		0.001			
	rs12285212	G	T	59927523	MS4A2, MS4A6A	3a	intergenic		0.002	0.01		
	chr11:59929465	C	T	59929465	MS4A2, MS4A6A	3a	intergenic		0.002	0.003		
	chr11:59936007	G	A	59936007	MS4A2, MS4A6A	3a	intergenic		0.01	0.005		
	chr11:59936023	C	G	59936023	MS4A2, MS4A6A	3a	intergenic		0.002	0.01		
	rs683892	A	T	59938266	MS4A6A	3a	downstream		0.002			
	chr11:59943683	C	T	59943683	MS4A6A	3a	intronic		0.005			
	chr11:59960287	C	T	59960287	MS4A6A, AB231731	3a	intergenic		0.002	0.002		
	chr11:59961597	G	T	59961597	MS4A6A, AB231731	3a	intergenic		0.001			
	rs7926344	G	A	59962166	MS4A6A, AB231731	3a	intergenic		0.36	0.27		
	rs71488445	T	C	59962240	MS4A6A, AB231731	3a	intergenic		0.07	0.03		
	chr11:59998347	A	T	59998347	AB231729, AB231731	3a	upstream		0.002	0.01		

SNP	SNP	Ref	Alt	Position	Gene	Regulome DB	Function	Function consequence	MAF (ADNI)	MAF (1000G)	MAF (esp6500)	Mutation taster score (prediction)
	chr11:60009095	C	T	60009095	AB231731, MS4A4A	3a	intergenic		0.003			
	chr11:59940217	A	T	59940217	MS4A6A	4	UTR3		0.002	0.01		
	chr11:59940271	A	C	59940271	MS4A6A	4	UTR3		0.002	0.004		
	chr11:59950687	C	G	59950687	MS4A6A	4	UTR5		0.002	0.004		
	chr11:59940074	G	A	59940074	MS4A6A	5	UTR3		0.002	0.0005		
	chr11:59980590	C	T	59980590	AB231729, AB231731	5	ncRNA_exonic		0.006			
	rs7929057	C	T	59980598	AB231729, AB231731	5	ncRNA_exonic		0.13	0.13		
	chr11:59980750	C	T	59980750	MS4A4E	6	exonic	non-synonymous	0.001	0.0005		1 (N)
	chr11:59939123	C	T	59939123	MS4A6A	7	UTR3		0.002	0.0005		
	chr11:59939286	T	C	59939286	MS4A6A	7	UTR3		0.002	0.004		
	rs61742546	A	G	59949058	MS4A6A	7	exonic	non-synonymous	0.02	0.01	0.02	1 (N)

Note. Using the ADNI whole-genome data, we used linkage disequilibrium, Regulome DB (accessed November 2014), and functional annotations from wAnnovar to isolate SNPs of interest. We first extracted all SNPs within approximately 50 kilobases of each SNP of interest, calculated linkage disequilibrium using Haploview, and retained all SNPs with a D' $\geq$ 0.99. Using Regulome DB and wAnnovar, we annotated each remaining SNP for: (1) known regulation and functional effects; (2) minor allele frequencies from the 1000 Genomes Project, 6500 Exomes Project, and the ADNI data set; and (3) corresponding MutationTaster predictions. We retained all SNPs with a Regulome DB score less than 4, and all SNPs located in untranslated (UTRs) or exonic regions (if nonsynonymous). No SNPs were significantly associated with case-control status in the ADNI whole-genome data.

Chapter 5

Future Directions

Despite decades of research, a major gap in Alzheimer's disease etiology persists, and no effective treatments exist. A major contributor to this gap is the void of known causal variants, or even a clear understanding of which pathways drive development and progression of Alzheimer's disease clinical symptoms. There are many avenues to pursue in understanding Alzheimer's disease, but discovering the genetic basis for the disease is a critical aspect that researchers must accomplish to understand its etiology. Understanding the pleiotropic and epistatic nature of involved genes, and specific mutations, may be critical. Several SNPs have repeatedly turned up in genome-wide association studies, but the SNPs themselves generally do not have obvious functional effects, and are unlikely to play a role in Alzheimer's disease etiology. What is more likely is that the SNPs are linked to one or more causal variants. We see two possible reasons the non-causal SNPs are robust across data sets, yet causal variants remain elusive: (1) the SNPs are linked to multiple rare variants that drive Alzheimer's disease development and progression; or (2) there is another common variant in the region that is either unobserved in large studies, or has misunderstood functional effects (e.g., gene regulation). Examples of common and rare functional variants with significant effects on Alzheimer's disease development and progression include the common *APOE ε4* and *APOE ε2* alleles, and the rare *TREM2* rs75932628 (R47H) variant and the *PLD3* rs145999145 (V232M) variant. Whether the causal variants are rare or common, given the biological complexity of Alzheimer's disease and results presented in this study, we believe epistasis plays a critical role in Alzheimer's disease etiology. As such, the community must continue to identify and vet these interactions.

In this research, we identified an interaction between rs11136000 (*CLU*) and rs670139 (*MS4A4E*) in the Cache County data that later replicated in a meta-analysis across the ADGC data. We also identified potential dosage and *APOE ε4* effects. To further understand this interaction's nature, researchers need to do the following: (1) test the interaction in more data sets to verify its veracity; (2) identify causal variants in the region; and (3) test causal variants *in vitro* within cell lines or *in vivo* within mice.

Identifying and verifying epistatic interactions is challenging, largely because of the statistical power limits. There are $\frac{n(n-1)}{2}$ possible interactions amongst covariates, where n is the number of genotypes. Given the large number of variables that can be included, the number of hypothesis tests grows quickly, draining valuable statistical power. Thus, managing false positives and negatives becomes an uphill battle. Once an interaction is identified, testing replication is still challenging because having data sets large enough to contain sufficient numbers of each genotype is not trivial. The *CLU-MS4A4E* interaction we identified replicated in ADGC, but epistatic interactions must be vetted to prevent costly, unsuccessful biological experiments. With further evidence to support this interaction, future experiments to understand the biological nature will be warranted.

Along with further replication, researchers need to identify causal variants in the region. At present, the two most likely explanations are multiple rare causal variants linked with rs11136000 and rs670139, and more common SNPs that weren't measured in genome-wide association data sets. Next-generation sequencing across large cohorts in these areas should provide the necessary clarity to identify the causal variants, whether rare or common. We identified a list of variants for both rs11136000 and rs670139 (Supplemental Tables 4.3 and 4.4) with a D' $\geq$ 0.99 using next-generation sequencing and annotated them with their known

regulome and exonic functions. Most were rare variants with no known functional effect, but some are worth investigating because of known function. Genotyping these alleles in a large cohort such as the Cache County data will make it possible to explore association with Alzheimer's disease development.

Any vetted variants that demonstrate statistical verification and have reasonable biological support need to be tested *in vitro* or *in vivo* to verify function. Research suggests *CLU* prevents amyloid fibrils and other protein aggregation events. Any variants known to modify *CLU* expression or function should be tested for correlation with protein aggregation *in vitro*. Little is known about *MS4A4E*, but given the statistical interaction with *CLU*, testing suspect functional variants with protein aggregation may be the most logical choice.

We have presented valuable information regarding epistasis in Alzheimer's disease in this research, including a novel gene-gene interaction between *CLU* and *MS4A4E* that modulates risk for Alzheimer's disease. Alzheimer's disease is a complex neurodegenerative disease whose genetic structure remains elusive, but this research provides convincing evidence that epistasis plays an important role in disease etiology and must be thoroughly explored.