Identifying Genetic Interactions Associated with Late-Onset Alzheimer's Disease

Charalampos S. Floudas, M.D, Ph.D.[1§], Nara Um, M.D, M.S.[1], M. Ilyas Kamboh, Ph.D.[2], Michael M. Barmada, Ph.D.[2], Shyam Visweswaran, M.D, Ph.D.[1,3]

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

[2]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA

[3]The Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA

[§]Corresponding author

Email addresses:

CSF: chfloudas@gmail.com
SV: shv3@pitt.edu

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Abstract

Background

Identifying genetic interactions in data obtained from genome-wide association studies (GWASs) can help in understanding the genetic basis of complex diseases. The large number of single nucleotide polymorphisms (SNPs) in GWASs however makes the identification of genetic interactions computationally challenging. We developed the Bayesian Combinatorial Method (BCM) that can identify pairs of SNPs that in combination have high statistical association with disease.

Results

We applied BCM to two late-onset Alzheimer's disease (LOAD) GWAS datasets to identify SNP-SNP interactions between a set of known SNP associations and the dataset SNPs. For evaluation we compared our results with those from logistic regression, as implemented in PLINK. Gene Ontology analysis of genes from the top 200 dataset SNPs for both GWAS datasets showed overrepresentation of LOAD-related terms. Four genes were common to both datasets: APOE and APOC1, which have well established associations with LOAD, and CAMK1D and FBXL13, not previously linked to LOAD but having evidence of involvement in LOAD. Supporting evidence was also found for additional genes from the top 30 dataset SNPs.

Conclusion

BCM performed well in identifying several SNPs having evidence of involvement in the pathogenesis of LOAD that would not have been identified by univariate analysis due to small main effect. These results provide support for applying BCM to identify potential genetic variants such as SNPs from high dimensional GWAS datasets.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Introduction

Elucidating the genetic basis of common diseases will lead to understanding of the biological mechanisms that underlie such diseases and can help in risk assessment, diagnosis, prognosis and development of new therapies. During the past several decades genetic linkage studies have been effective in mapping genetic loci responsible for many Mendelian diseases that are caused by a single genetic variant (Hardy & Singleton, 2009). More recently, genetic studies have indicated that most common diseases are likely to be polygenic where multiple genetic variants acting singly and in combination underlie the expression of disease (Thornton-Wells, Moore & Haines, 2004).

The commonest type of genetic variation is the single nucleotide polymorphism (SNP) that results when a single nucleotide is replaced by another in the genome sequence. The development of high-throughput genotyping technologies has led to a flurry of genome-wide association studies (GWASs) with the aim of discovering SNPs that are associated with common diseases. GWASs have been moderately successful in identifying SNPs associated with common diseases and traits. However, in most cases the identified SNPs have small effect sizes, and the proportion of heritability explained is quite modest. One view is that SNPs may interact in subtle ways that lead to substantially greater effects than the effect due to any single SNP. Another view is that common diseases may be due to rare and usually deleterious SNPs that cause disease in individual patients and that in different individuals or subpopulations the disease is caused by different deleterious SNPs.

This paper addresses the challenge of identifying interacting SNPs that may have small effects and describes a Bayesian combinatorial method (BCM) for identifying such interactions that are associated with disease. This method has been shown empirically to perform well on low

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

dimensional synthetic data (Balding, 2006). However, to our knowledge BCM has not been applied to a disease dataset with a large number of SNPs. In this paper we apply BCM to an Alzheimer's disease GWAS dataset to identify SNPs that interact with known Alzheimer associated SNPs.

As background, we provide brief summaries about GWASs, genetic interactions, and Alzheimer's disease in the following sections.

Genome-wide Association Studies

The development of high-throughput genotyping technologies that assay hundreds of thousands of SNPs or more, along with the identification of SNPs in the human genome by the International HapMap Project led to the emergence of GWASs. GWASs are typically case-control studies aimed at discovering SNPs – either as disease causing variants or as markers of disease – that are associated with a common disease or trait. The success of GWASs is based in large part on the common disease - common variant hypothesis. This hypothesis posits that common diseases in most individuals are caused by relatively common genetic variants that have low penetrance and hence have small to moderate influence in causing disease. An alternative hypothesis is the common disease - rare variant hypothesis, which posits that many rare variants underlie common diseases and each variant causes disease in relatively few individuals with high penetrance. Both these hypotheses likely contribute to common diseases with genetic variants may range from rare to the common SNPs.

GWAS data is typically analyzed for univariate associations between SNPs and the disease of interest; the statistical tests used include the Pearson's chi-square test, the Fisher's exact test, the Cochran-Armitage trend test, and odds ratios (Cordell, 2009). SNPs identified as

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

significant by univariate analyses may be further examined for interactions among them using methods such as logistic regression.

Genetic Interactions

Genetic interactions, also known as epistasis, can be defined biologically as well as statistically. Biologically, epistasis refers to gene-gene interaction when the action of one gene is modified by one or several other genes. Statistically, epistasis refers to interaction between variants at multiple loci in which the total effect of the combination of variants at the different loci may differ considerably from a linear combination of the effects of individual loci. The detection of statistical epistasis has the potential to indicate genetic loci that have a biological interaction (Hahn, Ritchie & Moore, 2003).

Statistical methods for identifying genetic interactions can be broadly divided into exhaustive and non-exhaustive methods. Exhaustive methods examine all possible SNP-subsets and examples include Multifactor Dimensionality Reduction (MDR) (Moore et al, 2006) and the BCM (Visweswaran, Wong & Barmada, 2009) that we describe in the next section. Examples of non-exhaustive methods include BOolean Operation-based Screening (BOOST), SNPHarvester and SNPRuler. We briefly describe these methods below.

The software package PLINK that is used widely for the analysis of GWAS datasets also implements logistic regression for the detection of SNP-SNP interactions and offers the option to test either all or specific sets of SNPs in a dataset (Purcell et al, 2007).

MDR exhaustively evaluates all 1-,2-,3-,..*n*-SNP subsets where *n* is specified by the user. It combines the variables in a SNP subset to construct a single binary variable and uses classification accuracy of the binary variable to evaluate a SNP-subset. Since MDR does not

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

scale up beyond a few hundred SNPs, for high dimensional data a multivariate filtering algorithm called ReliefF is applied to reduce the number of SNPs to a few hundred (M. D. Ritchie et al, 2001; Hahn, Ritchie & Moore, 2003; Moore et al, 2006; Moore & White, 2007).

BOOST uses a two-step procedure (Wan et al, 2010a). In the screening step, it uses an approximate likelihood ratio statistic that is computationally efficient and computes it for all pairs of SNPs. Only those SNPs that pass a threshold in the first step are examined for significant interaction effect using the classical likelihood ratio test that is computationally more expensive.

SNPHarvester is a stochastic search algorithm that uses a two-step procedure to identify epistatic interactions (Yang et al, 2009). In the first step it identifies 40–50 significant SNP groups using a stochastic search strategy, and in the second step, it fits a penalized logistic regression model to each group.

SNPRuler searches in the space of SNP rules and uses a branch-and-bound strategy to prune the huge number of possible rules in GWAS data (Wan et al, 2010b). An example of a rule is $X_1 = 0 \wedge X_2 = 2 \Rightarrow Z = 1$ (X_1 and X_2 are SNPs, the three genotypes that a SNP can take are coded as 0, 1 and 2 and Z is a binary outcome variable). The quality of a rule is evaluated with the chi-square statistic.

Alzheimer's Disease

Alzheimer's disease (AD) is the commonest neurodegenerative disease associated with aging and the commonest cause of dementia (Goedert & Spillantini, 2006). AD affects about 3% of all people between ages 65 and 74, about 19% of those between 75 and 84, and about 47% of those over 85. AD is characterized by adult onset of progressive dementia that typically begins with

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

subtle memory failure and progresses to a slew of cognitive deficits like confusion, language disturbance and poor judgment (Bertram, Lill & Tanzi, 2010).

AD is typically divided into early-onset Alzheimer's disease (EOAD) in which the onset of disease is before 60 years of age and late-onset Alzheimer's disease (LOAD) in which the onset is at or after 60 years of age. EOAD is rare and exhibits an autosomal dominant mode of inheritance. The genetic basis of EOAD is well established, and mutations in one of three genes (amyloid precursor protein gene, presenelin 1, or presenelin 2) account for most cases of EOAD (Avramopoulos, 2009).

LOAD is widespread and is estimated to strike almost half of all people over the age of 85. LOAD is believed to be a disease with both genetic and environmental influences, and elucidating the role of genetic factors in the pathogenesis and development of LOAD has been a major focus of research for more than a decade. One genetic risk factor for LOAD that has been consistently replicated is the apolipoprotein E (APOE) locus determined by the combined genotypes at the loci rs429358 (APOE*4) at codon 112 and rs7412 (APOE*2) at codon 158 (Holtzman, Morris & Goate, 2011). Because the two SNPs are in LD their combination determines six genotypes (2-2, 2-3, 3-3, 3-4, 2-4, 4-4) and the well-established protein polymorphism as one locus with three alleles (E*2, E*3, E*4). In the past few years, GWASs have identified several additional genetic loci associated with LOAD (Reiman et al, 2007; Hollingworth et al, 2011; Hu et al, 2011; Wijsman et al, 2011; Kamboh et al, 2012).

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Bayesian Combinatorial Method

BCM uses a Bayesian network (BN) to model a set of SNPs and interactions among them and their association with disease, and the model is evaluated with a Bayesian score. It then exhaustively searches a space of all possible models to identify high scoring models.

Bayesian network model and score. For a dataset D that contains a set of n SNPs $\{X_1, X_2, \ldots, X_n\}$ and a binary outcome variable Z (e.g, disease or phenotype) on N individuals, BCM's goal is to identify a set of SNPs that together are most predictive of Z in D. We model the effects of SNPs on Z with a BN that has n SNP-nodes and an additional node for Z. In this BN, which we call a SNP-BN, a subset of the n SNPs is modeled to have an effect on Z and every node in that subset has an arc to Z and every node not in the subset does not have an arc to Z. Also, there are no arcs between the SNP-nodes since we do not model the relations among the SNPs. Figure 1 gives an example of a SNP-BN where SNPs X_2 and X_3 are modeled to have a joint effect on Z (as shown by the arcs connecting them to Z) and the remaining SNPs do not have an effect on Z.

We evaluate the goodness of fit of a SNP-BN to data using an efficiently computable Bayesian score that computes the posterior probability of the BN given the data. In particular, we compute the BDeu (Bayesian Dirichlet equivalence uniform) score described in (Heckerman, Geiger & Chickering, 1995) which is commonly used in BN learning from data. This score is computed efficiently in closed form as follows:

$$P(M \mid D) = P(M) \prod_{i=1}^{n+1} \prod_{j=1}^{J_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(n_{ij} + \alpha_{ij})} \prod_{k=1}^{K_i} \frac{\Gamma(n_{ijk} + \alpha_{ijk})}{\Gamma(\alpha_{ijk})} \qquad (1)$$

where, $\Gamma(\cdot)$ is the gamma function, M is a SNP-BN, $P(D \mid M)$ is the posterior probability of M given D, $P(M)$ is the prior probability of M, K_i is the number of states of variable X_i represented by node i, J_i is the number of joint states of the parents of node i, n_{ijk} is the number of times in

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

the data that node i is in state k given parent state j, α_{ijk} are the parameter priors in a Dirichlet distribution which define the prior probability over the BN parameters. Also, $n_{ij} = \sum_{k=1}^{K_i} n_{ijk}$, $\alpha_{ij} = \sum_{k=1}^{K_i} \alpha_{ijk}$, and $\alpha_{ijk} = \frac{\alpha}{J_i \cdot K_i}$, where α is a single user-defined parameter prior. The n_{ijk} are obtained from the data and stored in a counts table that is associated with each node (an example of a counts table for node Z is shown in Figure 1). We make the following assumptions and simplifications: (1) model the prior probability $P(M)$ as a constant, i.e, *a priori* we consider all models to be equally plausible, (2) set $\alpha = 1$ which is a commonly used non-informative parameter prior, (3) use the logarithmic form to simplify computations when dealing with very small numbers, and (4) assign the score for a SNP-BN model to be the BDeu score attributable to just node Z (Visweswaran & Wong, 2009). The reason for assumption (4) is as follows. The BDeu score decomposes over the nodes in the BN and each node makes an independent contribution to the overall score. In the space of SNP-BNs, the score contributions of the SNP-nodes is a constant since they have no incoming arcs, and hence variation in the scores for distinct SNP-BNs is due only to the score attributable to Z. Thus, the score we use for a SNP-BN is given by the following expression (index i is absent since there is only one node under consideration, namely, Z, and $K = 2$ since Z is binary):

$$score(M) = \sum_{j=1}^{J} \log \frac{\Gamma(\alpha_j)}{\Gamma(n_j + \alpha_j)} + \sum_{k=1}^{2} \log \frac{\Gamma(n_{jk} + \alpha_{jk})}{\Gamma(\alpha_{jk})} \quad (2)$$

We have evaluated the BCM score in low dimensional synthetic data and found that in such data it has significantly greater power and is computed more efficiently than MDR (Visweswaran, Wong & Barmada, 2009; Jiang, Barmada & Visweswaran, 2010; Jiang et al, 2011).

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Materials & Methods

This section describes the GWAS datasets, the experimental methods, and previously identified LOAD SNPs.

GWAS Dataset

We used two different LOAD GWAS datasets in our experiments. The first dataset was part of the University of Pittsburgh Alzheimer's Disease Research Center (ADRC) that is described elsewhere (Kamboh et al, 2012). This dataset consists of 2245 individuals, of which 1290 had LOAD and 955 did not. For each individual, the genotype data consists of 682,685 SNPs on autosomal chromosomes.

The second dataset was collected by the Translational Genomics Research Institute (TGen) (Reiman et al, 2007). This dataset consists of 1411 individuals, of which. 861 had LOAD and 550 did not. For each individual, the imputed genotype data consists of 234,665 SNPs on autosomal chromosomes. For each individual, the genotype data consists of 502,627 SNPs; the original investigators analyzed 312,316 SNPs after applying quality controls. We used those 312,316 SNPs, plus two additional APOE SNPs from the same study namely, rs429358 and rs7412.

Experimental Methods

BCM searches exhaustively over all possible SNP-BN models in a dataset. For a GWAS dataset with half a million SNPs, the number of SNP-BN models is $2^n = 9.95 \times 10^{150514}$ and the number

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

of SNP-BN models with just 2 SNPs is $\binom{500000}{2}$ = 1.25 x 10^{11}. Thus, the search space is very large and it is computationally infeasible to evaluate every model in the space (Ritchie, 2011).

We addressed this challenge by applying BCM to a restricted space of SNP-BN models that consisted of a subset of all possible 2-SNP models. We considered only those 2-SNP models where one of the SNPs in a model is a member of a set of SNPs previously known to be associated with LOAD and the second SNP is any SNP (excluding the first SNP) in the dataset of interest. Since the number of known LOAD associated SNPs is much smaller than the number of SNPs in a dataset, it was computationally tractable to search this space of SNP-BN models. The selection of the previously identified LOAD SNPs that we used is described in the next section.

We applied BCM to each of the two GWAS datasets separately and analyzed in detail the top scoring 200 SNP-BN models. From each SNP-BN model, we extracted the SNP that was not in the set of previously identified LOAD SNPs. We mapped these SNPs to genes and considered only intragenic SNPs for further analyses. We performed the SNP to gene mapping with BioQ, a web-service which uses dbSNP build 135 and Genome Assembly GRCh37.p5 (Saccone, Quan & Jones, 2012). We performed enrichment analysis of the annotations of the associated genes in the Gene Ontology (GO) with the web-based tool GeneCoDis. For a set of genes GeneCoDis retrieves the associated GO terms, and identifies and ranks those GO terms that are significantly enriched in the set of genes (Carmona-Saez et al, 2007; Nogales-Cadenas et al, 2009). Enriched functional descriptors facilitate the interpretation of the gene set. The hierarchical nature of the GO annotations however means that the set of enriched GO terms may contain terms closely related in a parent-child relationship (Khatri & Drăghici, 2005). Such redundant terms confound

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

the interpretation. Therefore, we further examined the GO terms associated with the intragenic SNPs using the REViGo webserver. The REViGo software evaluates the semantic similarity between the enriched terms, identifies the most informative common ancestors and the related redundant GO terms and groups the latter under their ancestors (Supek et al, 2011). The resulting set facilitates simultaneous examination of the enriched GO terms at two levels: a detailed one, at the lowest level overrepresented term and a more abstract one at the highest level common ancestor of overrepresented terms. The detailed level can reveal specific genes of interest whereas the abstract level serves a compact overview of the processes, functions and cellular compartments associated with the genes in the set.

In addition to the analysis of the top scoring 200 SNP-BN models, we performed additional analyses of the top scoring 30 SNP-BN models. We analyzed the genes associated with the intragenic SNPs for differential expression in AD, through the ArrayExpress web server (Parkinson et al, 2011) and biological function analysis. Differential gene expression in relation to AD aims to integrate experimental evidence from transcriptomic analysis with those of genomic analysis. Up-regulation or down-regulation in AD of a gene in our results indicates increased biological plausibility for the reported genetic interaction. Finally, elements from the functional description of a gene (expression site, function related to the nervous system or pathways of LOAD, previous literature) were considered as supporting the biological relevance of an identified interaction.

We also compared BCM with logistic regression as implemented in PLINK for the identification of statistical genetic interactions. For this comparison we used as previous knowledge SNPs for all methods the rs429358 (APOE *4), a known LOAD risk SNP. We applied the methods to the ADRC LOAD dataset.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Previously Identified LOAD SNPs

We obtained a set of SNPs that are known to be associated with LOAD from the AlzGene website. The AlzGene website contains a regularly updated database of SNPs that have been shown to be associated with LOAD mostly in GWAS studies (Bertram et al, 2007). The curators of the AlzGene website use criteria established by the Human Genome Epidemiology Network (HuGENet) for assessing the cumulative evidence of associations of SNPs with disease (Ioannidis et al, 2008). We obtained 10 SNPs that were assessed to have sufficiently strong evidence of being associated with LOAD from the AlzGene website in March 2012. If a previously identified LOAD SNP was not present in our datasets, we selected a replacement SNP. The replacement SNP was within 500 kb, in the same gene, as the original SNP with pairwise linkage disequilibrium threshold of $r2 \geq 0.8$, using the SNAP web-based tool (Johnson et al, 2008). Using this protocol, we were unable to identify replacement SNPs in the TGen dataset for three previously identified LOAD SNPs; therefore we replaced them with SNPs from other genes, also reported as significantly associated with LOAD in the AlzGene website. Table 1 gives the list of previously identified LOAD SNPs that we used in the experiments.

Results and Discussion

This section describes the results that were obtained from applying BCM to the ADRC LOAD dataset and from applying BCM to the ADRC and the TGen GWAS datasets.

Top Scoring SNP-BN Models

Each SNP-BN model includes two SNPs of which one SNP is a previously identified LOAD SNP and the other SNP is not. We call the former SNP a *known SNP* and the latter SNP a *dataset*

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

SNP. The known and dataset SNPs from the top scoring 200 SNP-BN models are given in Table S1 (for ADRC) and Table S2 (for TGen) in the Supplemental Tables. A plot of the scores of the top scoring 200 SNP-BN models for the two datasets is shown in Figure 2.

In the ADRC dataset, the known SNP in each of the top scoring 200 SNP-BN models is rs429358 (APOE*4). In the TGen dataset, rs429358 is the known SNP in 192 of the top scoring 200 SNP-BN models, specifically models ranked 1 and 10-200, and in the 8 remaining models (ranked 2-9) the known SNP belongs to genes GAB2, MS4A6A, MS4A4E, CR1, PICALM, SORL1, TF whereas the dataset SNP is rs7412 for all 8 models. SNPs rs429358 and rs7412 are located on the APOE gene and their combined genotypes determine the APOE allelic status which is known to be the strongest genetic variant that is predictive of LOAD.

In the ADRC dataset, the dataset SNPs from the top scoring 200 models included 92 intragenic SNPs that mapped to 77 distinct genes, and the dataset SNPs from the top scoring 30 models included 18 intragenic SNPs that mapped to 15 distinct genes. In the TGen dataset, the dataset SNPs from the top scoring 200 models included 82 intragenic SNPs that mapped to 69 genes, and the dataset SNPs from the top scoring 30 models included 19 intragenic SNPs that mapped to 11 genes.

In the top scoring 200 SNP-BN models the two datasets have in common two intragenic SNPs, rs7412 (APOE gene) and rs4420638 (APOC1 gene) as well as two genes mapped from intragenic SNPs, CAMK1D (rs11257738 in ADRC and rs17151584 in TGen) and FBXL13 (rs7779121 in ADRC and rs17475512 in TGen).

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

GO Term Analysis

The most informative common ancestors of the overrepresented GO terms obtained from GeneCoDis for the ADRC dataset are given in Table S3 and for the TGen dataset are given in Table S4 in the Supplemental Tables. In both sets nervous system-related terms are enriched (e.g, *regulation of dendrite development, nervous system development, regulation of axon extension, short term memory*), as well as terms related to cholesterol and lipid metabolism (e.g, *lipid metabolic process*, *chylomicron*), beta amyloid (*beta amyloid binding*) cell membranes (e.g, *integral to membrane, plasma membrane, postsynaptic , clathrin-coated endocytic vesicle*), calmodulin and intracellular calcium homeostasis (e.g, *calmodulin binding, cytosolic calcium ion transport*) and the immune system (*immunoglobulin binding*). Overrepresentation of these terms shows that the identified genes from both datasets include genes that are members of biochemical pathways involved in LOAD pathophysiology (Holtzman, Morris & Goate, 2011; Morgan, 2011).

Expression Analysis

The 15 genes corresponding to the 18 dataset intragenic SNPs from the top scoring 30 models in the ADRC dataset and the 19 genes corresponding to the 19 dataset intragenic SNPs from the top scoring 30 models in the TGen dataset were examined for relative expression in AD (Table 2 for the ADRC dataset and in Table 3 for the TGen dataset). In these tables, the second to last column gives the rank of the corresponding SNP based on the model score obtained by applying BCM to 1-SNP models. For some of the SNPs the rank based on the 1-SNP model is very low compared to the score of the corresponding 2-SNP model which implies that these SNPS would not have

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

been identified by univariate analysis. The last column gives the p values for the pairs of SNPs that were obtained from using logistic regression in PLINK.

Comparison of BCM with PLINK

Among the top 50 models ranked by BCM and PLINK respectively, there are 4 models in common, and among the top 200 models, there were 25 models in common between BCM and PLINK.

Discussion

Examining all pairs of SNPs in a GWAS dataset for identifying interacting SNP pairs is usually not computationally tractable due to the large number of SNP pairs. We addressed this challenge by examining only a subset of all pairs of SNPs where one member of the pair is drawn from a small set of previously known disease-associated SNPs and by using a Bayesian score to evaluate that statistical association of a SNP pair with the disease. We applied this strategy to two LOAD GWAS datasets and our results show that it can identify interacting SNPs of plausible biological significance. Moreover, this strategy finds SNPs that would be overlooked in a univariate analysis because they exhibit small main effects; however, they are detected when paired with another SNP due to interaction effects.

In both LOAD GWAS datasets that we examined, the previously known disease-associated SNP that was identified is either rs429358 (APOE*4) or rs7412 (APOE*2); these SNPs reside in the APOE gene which is known to be the strongest genetic determinant for LOAD. GO term enrichment analysis of the dataset SNPs identified terms that are relevant to

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

biochemical pathways implicated in the pathogenesis of LOAD such as *lipid metabolic process, calmodulin binding, nervous system development* and multiple membrane-related terms.

Gene expression analysis of the dataset SNPs showed that for each dataset studied a majority of the genes corresponding to the top 30 dataset SNPs are differentially expressed in LOAD. Functional annotations and literature evidence that are presented with the expression data in the relevant tables further support the role of these genes in the pathogenesis of LOAD.

Among the genes corresponding to the top 200 dataset SNPs, besides the APOE gene, three other genes are common to both datasets: APOC1, CAMK1D and FBXL13. Evidence supporting the interaction of APOC1 with APOE are presented in the analysis for the top 30 dataset SNPs. CAMK1D (calcium/calmodulin-dependent protein kinase ID) belongs to the family of calmodulin kinases which modulate neuronal development and plasticity (Wayman et al, 2008). It has been found to be overexpressed in AD and is expressed in the brain especially during hippocampal formation with high expression in the pyramidal cell layers (Lukk et al, 2010; Pugazhenthi et al, 2011). It encodes a member of the Ca2+/calmodulin-dependent protein kinase 1 subfamily of serine/threonine kinases (Maglott et al, 2011). CAMK1D interacts with CALM1 (calmodulin), which has been associated with AD risk (Lambert et al, 2010). The encoded protein may regulate calcium-mediated granulocyte function and activates MAPK3 (Mitogen-activated protein kinase 3 - inferred function by similarity). In vitro, it phosphorylates transcription factor CREM (cAMP responsive element binding) isoform Beta and probably CREB1 (Pugazhenthi et al, 2011). The CREB pathway has a role in memory formation and CREB phosphorylation has been proposed as a signalling pathway involved in the pathogenesis of AD (Müller et al, 2011) and also its down-regulation may have a role in exacerbations of AD (Pugazhenthi et al, 2011). Another member of the same family, neuronal CaM kinase II

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

phosphorylates tau protein on ser262, an important step in the formation of neurofibrillary tangles in AD (Yamauchi, 2005). FBXL13 (F-box and leucine-rich repeat protein 13) belongs to the F-box protein family. Members of this family have a characteristic approximately 40-amino acid F-box motif and take part in SCF (SKP1-CUL1-F-box protein) complexes that act as protein-ubiquitin ligases (Maglott et al, 2011). The ubiquitin-proteasome system is involved in protein turnover and degradation and is perturbed in AD (Riederer et al, 2011). An SCF complex of another F box protein (FBXW7) is involved in the degradation of NICD (NOTCH1 released notch intracellular domain) and probably of PSEN1 (The UniProt Consortium, 2013).

In addition to the genes corresponding to the top 30 top scoring SNP-BN models in the ADRC dataset, we found other genes in lower scoring SNP-BN models with plausible associations with LOAD. In the 80th scoring model (dataset SNP rs7793977), gene PION [pigeon homolog (Drosophila)], also known as GSAP (gamma-secretase-activating protein), is known to increase amyloid beta production (Maglott et al, 2011). In the 196th scoring model, (dataset SNP rs6534145), gene PDE5A (phosphodiesterase 5A, cGMP-specific) could be implicated to LOAD pathogenesis via two different mechanisms. PDE5A is a substrate of CASP3 (caspase 3) (Frame et al, 2001), which in turn has been shown to be involved in the early synaptic dysfunction in a mouse model of AD (D'Amelio et al, 2011). It has also been shown that inhibition of PDE5A results in a decrease in the transcription of Wnt/β-catenin (Tinsley et al, 2011). A reduction in Wnt signalling has been implicated in the amyloid beta-dependent neurodegeneration in LOAD (Inestrosa & Toledo, 2008).

While BCM has been applied to low dimensional synthetic data with good results (Visweswaran & Wong, 2009), in this paper we have applied it to GWAS datasets. BCM has several advantages. It is computationally more efficient than the widely used MDR

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

(Visweswaran, Wong & Barmada, 2009). Since BCM uses the Bayesian paradigm, the BCM score represents a coherent way to combine knowledge with data. Biological knowledge or results from analyses of earlier studies can be encoded as a prior distribution over the models that can then be used in Equation 1. Use of informative priors is becoming common in the analysis of microarray expression studies, and a similar strategy can be employed for genomic data.

A limitation of our study is the use of GWAS datasets related to a single disease, although it is an important disease. In future research, we plan to apply and investigate the utility of BCM on GWAS datasets related to additional diseases. Another limitation is the use just 10 previously known LOAD-associated SNPs. In future work, we plan to explore the use of a larger set of known LOAD associated SNPs that will include SNPs with weaker evidence of being associated with LOAD. In addition, we plan to study the effect of excluding the APOE SNPs rs429358 and rs7412 which are present in every SNP pair we examined for biological plausibility. Another limitation is that we did not use informative prior probabilities for encoding prior knowledge from the literature and previous GWASs. BCM can be extended easily to allow the incorporation of informative priors and inclusion of informative priors in the analysis is an interesting area for study.

Conclusion

We applied BCM to two LOAD GWAS datasets to identify pairs of SNPs that in combination have high statistical association with development of LOAD. To reduce the large search space of all possible parts of SNPs in a GWAS dataset we restricted BCM to evaluate those SNP pairs where one of the SNP was drawn from a set of 10 previously known LOAD associated SNPs. Our results identified several SNPs that have biological evidence of being involved in the

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

pathogenesis of LOAD that would not have been identified by univariate analysis alone due to small main effect but were identified in conjunction with another SNP. These results provide support for applying BCM to identify potential genetic variants such as SNPs from high dimensional GWAs datasets.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Funding

CSF was supported in part by an award from the Gerondelis Foundation, SV was supported in part by NLM grant HHSN276201000030C, and M. Ilyas Kamboh was supported by National Institutes of Health grants AG030653, AG005133 and AG041718.

Acknowledgements

We thank Mr. Kevin Bui for his help in data preparation and in implementing the BCM algorithm in software.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

References

Avramopoulos D. 2009. Genetics of Alzheimer's disease: recent advances. *Genome Medicine* 1:34.

Balding DJ. 2006. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7:781–791.

Bernstein HG & Müller M. 1999. The cellular localization of the L-ornithine decarboxylase/polyamine system in normal and diseased central nervous systems. *Progress in Neurobiology* 57:485–505.

Bertram L, Lill CM & Tanzi RE. 2010. The genetics of Alzheimer disease: back to the future. *Neuron* 68:270–281.

Bertram L, McQueen MB, Mullin K, Blacker D & Tanzi RE. 2007. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nature Genetics* 39:17–23.

Blacker D, Bertram L, Saunders AJ, Moscarillo TJ, Albert MS, Wiener H, Perry RT, Collins JS, Harrell LE, Go RCP et al. 2003. Results of a high-resolution genome screen of 437 Alzheimer's disease families. *Human Molecular Genetics* 12:23–32.

Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM & Pascual-Montano A. 2007. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biology* 8:R3.

Cordell HJ. 2009. Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics* 10:392–404.

Cramer PE, Cirrito JR, Wesson DW, Lee CYD, Karlo JC, Zinn AE, Casali BT, Restivo JL, Goebel WD, James MJ et al. 2012. ApoE-directed therapeutics rapidly clear β-amyloid and reverse deficits in AD mouse models. *Science* 335:1503–1506.

Crews L, Adame A, Patrick C, Delaney A, Pham E, Rockenstein E, Hansen L & Masliah E. 2010. Increased BMP6 levels in the brains of Alzheimer's disease patients and APP transgenic mice are accompanied by impaired neurogenesis. *The Journal of Neuroscience* 30:12252–12262.

D'Amelio M, Cavallucci V, Middei S, Marchetti C, Pacioni S, Ferri A, Diamantini A, De Zio D, Carrara P, Battistini L et al. 2011. Caspase-3 triggers early synaptic dysfunction in a mouse model of Alzheimer's disease. *Nature Neuroscience* 14:69–76.

Dunn CD, Sulis ML, Ferrando AA & Greenwald I. 2010. A conserved tetraspanin subfamily promotes Notch signaling in Caenorhabditis elegans and in human cells. *Proceedings of the National Academy of Sciences of the United States of America* 107:5907–5912.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Frame M, Wan KF, Tate R, Vandenabeele P & Pyne NJ. 2001. The gamma subunit of the rod photoreceptor cGMP phosphodiesterase can modulate the proteolysis of two cGMP binding cGMP-specific phosphodiesterases (PDE6 and PDE5) by caspase-3. *Cellular Signalling* 13:735–741.

Frykman S, Teranishi Y, Hur J.-Y, Sandebring A, Goto Yamamoto N, Ancarcrona M, Nishimura T, Winblad B, Bogdanovic N, Schedin-Weiss S et al. 2012. Identification of two novel synaptic γ-secretase associated proteins that affect amyloid β-peptide levels without altering Notch processing. *Neurochemistry International* 61:108–118.

Goedert M & Spillantini MG. 2006. A century of Alzheimer's disease. *Science* 314:777–781.

Hahn LW, Ritchie MD & Moore JH. 2003. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 19:376–382.

Hardy J & Singleton A. 2009. Genomewide association studies and human disease. New England Journal of Medicine 360:1759–1768.

Heckerman D, Geiger D & Chickering DM. 1995. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *MACHINE LEARNING* 20:197–243.

Hollingworth P, Harold D, Sims R, Gerrish A, Lambert J.-C, Carrasquillo MM, Abraham R, Hamshere ML, Pahwa JS, Moskvina V et al. 2011. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nature Genetics* 43:429–435.

Holtzman DM, Morris JC & Goate AM. 2011. Alzheimer's disease: the challenge of the second century. *Science Translational Medicine* 3:77sr1.

Hu X, Pickering E, Liu YC, Hall S, Fournier H, Katz E, Dechairo B, John S, Van Eerdewegh P & Soares H. 2011. Meta-analysis for genome-wide association study identifies multiple variants at the BIN1 locus associated with late-onset Alzheimer's disease. *PLoS One* 6:e16616.

Inestrosa NC & Toledo EM. 2008. The role of Wnt signaling in neuronal dysfunction in Alzheimer's Disease. *Molecular neurodegeneration* 3:9.

Ioannidis JPA, Boffetta P, Little J, O'Brien TR, Uitterlinden AG, Vineis P, Balding DJ, Chokkalingam A, Dolan SM, Flanders WD et al. 2008. Assessment of cumulative evidence on genetic associations: interim guidelines. *International Journal of Epidemiology* 37:120–132.

Jiang X, Barmada MM & Visweswaran S. 2010. Identifying genetic interactions in genome-wide data using Bayesian networks. *Genetic Epidemiology* 34:575–581.

Jiang X, Neapolitan RE, Barmada MM & Visweswaran S. 2011. Learning genetic epistasis using Bayesian network scoring criteria. *BMC Bioinformatics* 12:89.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ & de Bakker PIW. 2008. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24:2938–2939.

Kamboh MI, Demirci FY, Wang X, Minster RL, Carrasquillo MM, Pankratz VS, Younkin SG, Saykin AJ, Jun G, Baldwin C et al. 2012. Genome-wide association study of Alzheimer's disease. *Translational Psychiatry* 2:e117.

Khatri P & Drăghici S. 2005. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21:3587–3595.

Lambert J-C, Grenier-Boley B, Chouraki V, Heath S, Zelenika D, Fievet N, Hannequin D, Pasquier F, Hanon O, Brice A et al. 2010. Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis. *Journal of Alzheimer's Disease* 20:1107–1118.

Lipinski MM, Zheng B, Lu T, Yan Z, Py BF, Ng A, Xavier RJ, Li C, Yankner BA, Scherzer CR et al. 2010. Genome-wide analysis reveals mechanisms modulating autophagy in normal brain aging and in Alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America* 107:14164–14169.

Liu F, Arias-Vásquez A, Sleegers K, Aulchenko YS, Kayser M, Sanchez-Juan P, Feng B-J, Bertoli-Avella AM, van Swieten J, Axenovich TI et al. 2007. A genomewide screen for late-onset Alzheimer disease in a genetically isolated Dutch population. *American Journal of Human Genetics* 81:17–31.

Lourenço FC, Galvan V, Fombonne J, Corset V, Llambi F, Müller U, Bredesen DE & Mehlen P. 2009. Netrin-1 interacts with amyloid precursor protein and regulates amyloid-beta production. *Cell Death & Differentiation* 16:655–663.

Lukk M, Kapushesky M, Nikkilä J, Parkinson H, Goncalves A, Huber W, Ukkonen E & Brazma A. 2010. A global map of human gene expression. *Nature Biotechnology* 28:322–324.

Maglott D, Ostell J, Pruitt KD & Tatusova T. 2011. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* 39(Database issue):D52–57.

Moore JH, Gilbert JC, Tsai C-T, Chiang F-T, Holden T, Barney N & White B. C. 2006. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology* 241:252–261.

Moore J & White B. 2007. Tuning ReliefF for genome-wide genetic analysis. *Evolutionary computation, machine learning and data mining in bioinformatics* 4447:166–175.

Morgan K. 2011. The three new pathways leading to Alzheimer's disease. *Neuropathology and Applied Neurobiology* 37:353–357.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Müller M, Cárdenas C, Mei L, Cheung K-H & Foskett J. K. 2011. Constitutive cAMP response element binding protein (CREB) activation by Alzheimer's disease presenilin-driven inositol trisphosphate receptor (InsP3R) Ca2+ signaling. *Proceedings of the National Academy of Sciences of the United States of America* 108:13293–13298.

Nilsson T, Bogdanovic N, Volkman I, Winblad B, Folkesson R & Benedikz E. 2006. Altered subcellular localization of ornithine decarboxylase in Alzheimer's disease brain. *Biochemical and Biophysical Research Communications* 344:640–646.

Nogales-Cadenas R, Carmona-Saez P, Vazquez M, Vicente C, Yang X, Tirado F, Carazo JM & Pascual-Montano A. 2009. GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Research* 37(Web Server issue):W317–322.

Parkinson H, Sarkans U, Kolesnikov N. Abeygunawardena N, Burdett T, Dylag M, Emam I, Farne A, Hastings E, Holloway E et al. 2011. ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Research* 39(Database issue):D1002–1004.

Perry T & Greig NH. 2005. Enhancing central nervous system endogenous GLP-1 receptor pathways for intervention in Alzheimer's disease. *Current Alzheimer Research* 2:377–385.

Pugazhenthi S, Wang M, Pham S, Sze C-I & Eckman CB. 2011. Downregulation of CREB expression in Alzheimer's brain and in Aβ-treated rat hippocampal neurons. *Molecular Neurodegeneration* 6:60.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *American Journal of Human Genetics* 81:559–575.

Reiman EM, Webster JA, Myers AJ, Hardy J, Dunckley T, Zismann VL, Joshipura KD, Pearson JV, Hu-Lince D, Huentelman MJ et al. 2007. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. *Neuron* 54:713–720.

Riederer BM, Leuba G, Vernay A & Riederer IM. 2011. The role of the ubiquitin proteasome system in Alzheimer's disease. *Experimental Biology and Medicine (Maywood)* 236:268–276.

Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF & Moore JH. 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Human Genetics* 69:138–417.

Ritchie MD. 2011. Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. *Annals of Human Genetics* 75:172–182.

Saccone SF, Quan J & Jones PL. 2012. BioQ: tracing experimental origins in public genomic databases using a novel data provenance model. *Bioinformatics* 28:1189–1191.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Supek F, Bošnjak M, Škunca N & Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800.

Tesseur I, Zou K, Esposito L, Bard F, Berber E, Can JV, Lin AH, Crews L, Tremblay P, Mathews P et al. 2006. Deficiency in neuronal TGF-beta signaling promotes neurodegeneration and Alzheimer's pathology. *Journal of Clinical Investigation* 116:3060–3069.

The UniProt Consortium. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Research* 41:D43–47.

Thornton-Wells TA, Moore JH & Haines JL. 2004. Genetics, statistics and human disease: analytical retooling for complexity. *Trends in Genetics* 20:640–647.

Tinsley HN, Gary BD, Keeton AB, Lu W, Li Y & Piazza GA. 2011. Inhibition of PDE5 by sulindac sulfide selectively induces apoptosis and attenuates oncogenic Wnt/β-catenin-mediated transcription in human breast tumor cells. *Cancer Prevention Research (Philadelphia*) 4:1275–1284.

Tzschach A, Bisgaard A-M, Kirchhoff M, Graul-Neumann LM, Neitzel H, Page S, Ahmed A, Müller I, Erdogan F, Ropers H-H et al. 2010. Chromosome aberrations involving 10q22: report of three overlapping interstitial deletions and a balanced translocation disrupting C10orf11. *European Journal of Human Genetics* 18:291–295.

Visweswaran S & Wong A-KI. 2009. Bayesian combinatorial partitioning for detecting interactions among genetic variants. *Summit on Translational Bioinformatics* 2009:133.

Visweswaran S, Wong A-KI & Barmada MM. 2009. A Bayesian method for identifying genetic interactions. *AMIA Annual Symposium Proceedings* 2009:673–677.

Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NLS & Yu W. 2010. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *American Journal of Human Genetics* 87:325–340.

Wan X, Yang C, Yang Q, Xue H, Tang NLS & Yu W. 2010. Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics* 26:30–37.

Wayman GA, Lee Y-S, Tokumitsu H, Silva AJ & Soderling TR. 2008. Calmodulin-kinases: modulators of neuronal development and plasticity. *Neuron* 59:914–931.

Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P, Rohrer K, Zhao A, Marlowe L, Kaleem M et al. 2009. Genetic control of human brain transcript expression in Alzheimer disease. *American Journal of Human Genetics* 84:445–458.

Wijsman E. M, Pankratz ND, Choi Y, Rothstein JH, Faber KM, Cheng R, Lee JH, Bird TD, Bennett DA, Diaz-Arrastia R et al. 2011. Genome-wide association of familial late-onset Alzheimer's disease replicates BIN1 and CLU and nominates CUGBP2 in interaction with APOE. *PLoS Genetics* 7:e1001308.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

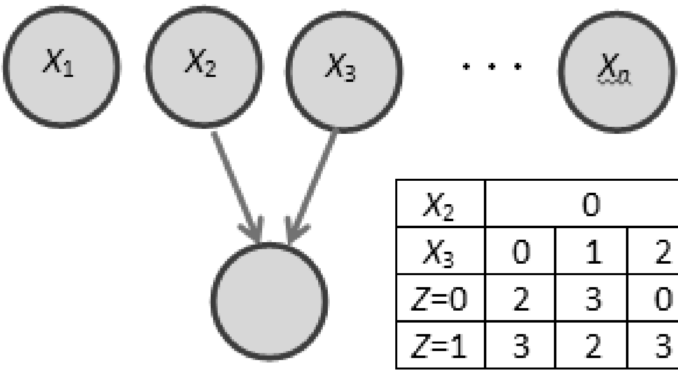
Yamauchi T. 2005. Neuronal Ca2+/calmodulin-dependent protein kinase II--discovery, progress in a quarter of a century, and perspective: implication for learning and memory. *Biological & Pharmaceutical Bulletin* 28:1342–1354.

Yang C, He Z, Wan X, Yang Q, Xue H & Yu, W. 2009. SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics* 25:504–511.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Figures

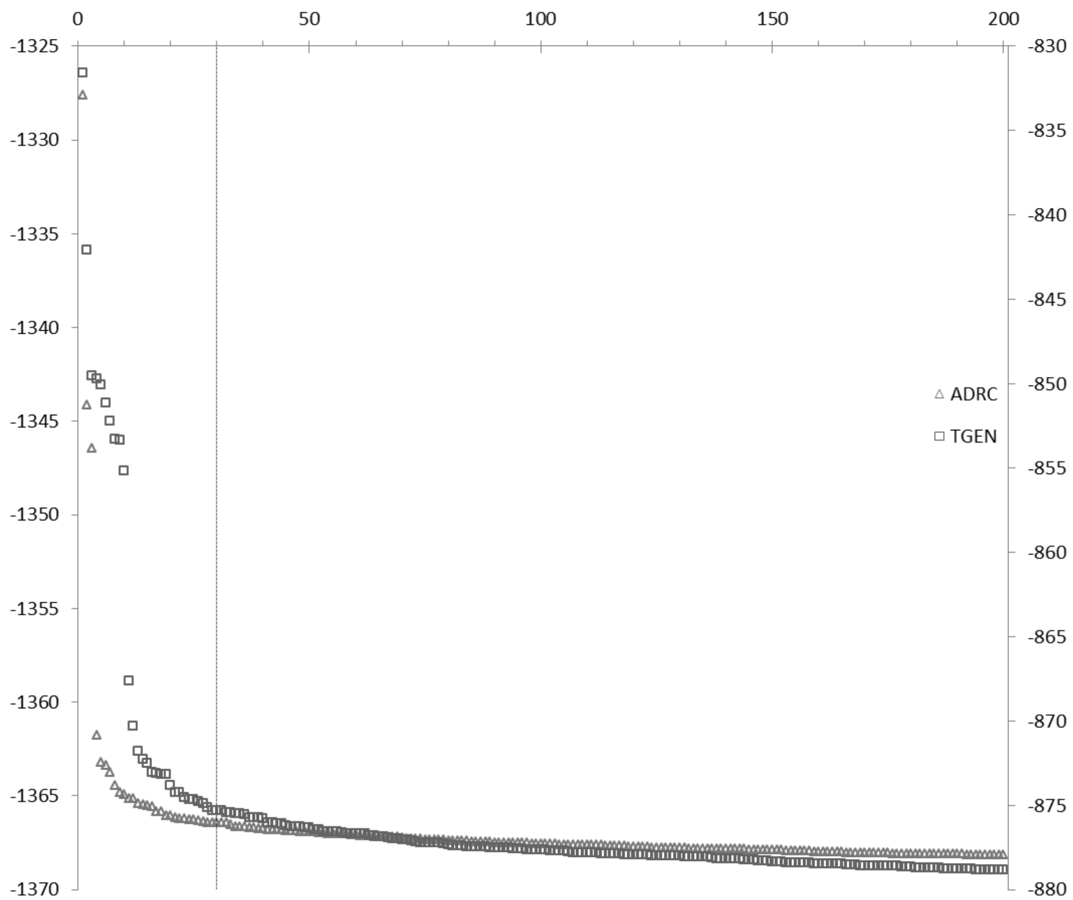
Figure 1. A SNP-BN model where SNPs *X*2 and *X*3 have an effect on *Z* and the remaining SNPs do not have an effect on *Z*. The table gives counts for the states of *Z* conditioned on the joint states of *X*2 and *X*3.



X_2	0			1			2		
X_3	0	1	2	0	1	2	0	1	2
Z=0	2	3	0	4	5	6	1	0	4
Z=1	3	2	3	3	6	3	2	0	3

Figure 2. Plots of the distribution of BCM model scores for the top ranked 200 SNP-BN models for the two datasets, ADRC and TGen. The scores for the ADRC dataset (blue points) correspond to the left hand Y axis, while those for the TGen dataset correspond to the right hand Y axis. The dotted vertical line marks the top ranked 200 SNP-BN models.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2



PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Tables

Table 1 Previously identified LOAD SNPs.

#	Gene	AlzGene SNP	Odds ratio (95% CI)	p value	ADRC SNP	r^2	TGen SNP	r^2
1	**APOE 4**	rs429358	3.685 (3.30-4.12)	<1E-50	Same	-	Same	-
2	**CR1**	rs3818361	1.174 (1.14-1.21)	4.72E-21	Same	-	rs6656401	0.840
3	**PICALM**	rs3851179	0.879 (0.86-0.9)	2.85E-20	Same	-	rs7110631	0.841
4	**MS4A6A**	rs610932	0.904 (0.88-0.93)	1.81E-11	Same	-	rs574695	0.935
5	**CD33**	rs3865444	0.893 (0.86-0.93)	2.04E-10	Same	-	Same	-
6	**MS4A4E**	rs670139	1.079 (1.05-1.11)	9.51E-10	rs600550	1	rs676309	1
7	**CD2AP**	rs9349407	1.117 (1.08-1.16)	2.75E-09	rs9296559	1	rs9296558	1
8	**GAB2**	rs2373115	0.85 (0.76-0.94)		Same	-	Same	-
9	**SORL1**	rs2282649	1.10 (1.03-1.17)		rs726601	0.922	rs726601	0.922
10	**TF**	rs1049296	1.18 (1.06-1.31)		Same	-	Same	-

AlzGene SNP: the SNP in the AlzGene meta-analysis, along with the relevant odds ratios and p values (the latter for those SNPs with p values <0.00001); *ADRC SNP*: the corresponding SNP in the ADRC dataset, along with the r2 scores; *TGen SNP*: the corresponding SNP in the TGen dataset, along with the r2 scores for linkage disequilibrium.

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Table 2. Functional description and expression of genes associated with the top 30 dataset SNPs in the ADRC dataset.

Gene Symbol (SNP)	**Name**	**Description**	**Expression in AD**	**1-SNP model rank**	**p value of pair from PLINK**
APOC1 (rs4420638)	Apolipoprotein C-I	Appears to modulate the interaction of APOE with beta-migrating VLDL. Binds free fatty acids.	Overexpressed (Lukk et al, 2010)	2	0.3326
TOMM40 (rs157582)	Translocase of outer mitochondrial membrane 40 homolog	Channel-forming subunit of the translocase of the mitochondrial outer membrane (TOM) complex, essential for protein import into mitochondria.	Underexpressed (Lukk et al, 2010)	3	0.4139
APOE (rs7412)	Apolipoprotein E	ApoE is essential for the normal catabolism of triglyceride-rich lipoprotein constituents. Known risk factor for LOAD.	Overexpressed (Lukk et al, 2010)	5	0.8172
SNTG1 (rs16914489)	Gamma-1-syntrophin	Specifically expressed in the brain, highly expressed in the cortex. Organizes the subcellular localization of a variety of proteins.	Overexpressed (Lukk et al, 2010)	24906	0.004546
TMEM217 (rs9470543)	Transmembrane protein 217	Expressed in the brain	-	4584	0.004643
SMAD6 (rs3934907)	Mothers against DPP homolog 6	Negative regulation of BMP and TGF-beta/activin-signaling. BMP-6 is increased in AD brains and leads to impaired neurogenesis (Crews et al, 2010). Reduced TGF-beta signaling is involved in neurodegeneration and promotes AD like changes in mice (Tesseur et al, 2006).	Underexpressed (Lukk et al, 2010)	41282	0.0000998
NPAS3 (rs4981180)	Neuronal PAS domain protein 3.	Transcription factor. May regulate genes involved in neurogenesis. Associated with schizophrenia and mental retardation	Overexpressed (Lukk et al, 2010)	1086	0.1225
NTM (rs11222692)	Neurotrimin	May promote neurite outgrowth and adhesion. NTM lies at locus 11q25, which has been associated	Overexpressed (Lukk et al, 2010)	12209	0.1422

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

		with AD (Blacker et al, 2003; Liu et al, 2007).			
PPAPDC1A (rs4752432)	Phosphatidic acid phosphatase type 2 domain containing 1A	-	-	6852	0.3963
NPFF (rs8192593)	Neuropeptide FF-amide peptide precursor	Modulation of morphine-induced antinociception.	-	3981	0.1251
SLC25A21 (rs7140725)	Solute carrier family 25	Known also as ornithine decarboxylase (ODC). Mitochondrial oxoadipate carrier, part of polyamine synthesis pathway.	Overexpressed (Bernstein & Müller, 1999; Nilsson et al, 2006)	444	0.06767
RAB23 (rs182662)	Member RAS oncogene family	Intracellular protein transportation. Regulated by miRNA155, which also regulates PICALM (a known AD association).	Underexpressed (Lukk et al, 2010)	96	0.08251
UNC5D (rs4577954)	unc-5 homolog D (C. elegans)	Netrin receptor: netrins are secreted proteins that direct axon extension and cell migration during neural development. APP also binds Netrin-1 and in transgenic mice this suppresses amyloid beta peptide production (Lourenço et al, 2009).	-	63972	0.6731
CHD9 (rs3852742)	Chromodomain helicase DNA binding protein 9, PPARA -interacting complex 320 kDa protein	Transcriptional co-activator for PPARA. The APOE gene promoter has a binding site for PPAR alpha. Low CHD9 activity could reduce apoE levels. Increase in APOE transcription has been shown to clear amyloid beta in AD mouse models (Cramer et al, 2012).	Overexpressed (Lukk et al, 2010)	1061	0.04696
CNTN4 (rs9819935)	Contactin 4, Brain-derived immunoglobulin superfamily protein 2	Mainly expressed in brain. Neuronal membrane protein that may play a role in the formation of axon connections in the developing nervous system. Associated with	-	2149	0.002386

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

		Spinocerebellar Ataxia, Amyotrophic Lateral Sclerosis, 3p deletion syndrome.			

1-SNP model rank: rank of the corresponding SNP in terms of univariate 1-SNP model score

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

Table 3. Functional description and expression of genes associated with the top 30 dataset SNPs in the TGen dataset.

Gene Symbol (SNP)	**Name**	**Description**	**Differential Expression in AD**	**1-SNP model rank**	**p value of pair from PLINK**
APOE 2 (rs7412)	Apolipoprotein E	ApoE is essential for the normal catabolism of triglyceride-rich lipoprotein constituents. Known risk factor for LOAD.	Overexpressed (Lukk et al, 2010)	1	0.05993
APOC1 (rs4420638)	Apolipoprotein C-1	Appears to modulate the interaction of APOE with beta-migrating VLDL. Binds free fatty acids.	Overexpressed (Lukk et al, 2010)	3	0.705
C10orf11 (rs7079348)	Chromosome 10 open reading frame 11	A brain-expressed gene. Haploinsufficiency of C10orf11 contributes to the cognitive defects in 10q22 syndrome (Tzschach et al, 2010).	-	4	0.009623
VWC2 (rs10499687)	von Willebrand factor C domain-containing protein 2 (Brorin, Brain-specific chordin-like protein)	Encodes a secreted bone morphogenic protein (BMP) antagonist. The encoded protein is possibly involved in neural function and development and may have a role in cell adhesion. BMP-6 is increased in AD brains and leads to impaired neurogenesis (Crews et al, 2010).	Underexpressed (Webster et al, 2009)	12	0.7698
PSD3 (rs17126808)	Pleckstrin and Sec7 domain containing 3	Guanine nucleotide exchange factor for ARF6 that contributes to the regulation of dendritic branching (The UniProt Consortium, 2013).	Overexpressed (Lukk et al, 2010)	34	0.001623
GXYLT2 (rs3732443)	Glucoside xylosyltransferase 2	Elongates the O-linked glucose attached to EGF-like repeats in the extracellular domain of Notch proteins (The UniProt Consortium, 2013), which are substrates of γ-secretase, the enzyme involved in amyloid beta production (Frykman et al,	Underexpressed in a murine AD model (D'Amelio et al, 2011)	6	0.211

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2

		2012).			
GABBR2 (rs2779550)	Gamma-aminobutyric acid (GABA) B receptor, 2	Target for autophagy regulation in neurodegenerative diseases (Lipinski et al, 2010).	Overexpressed (Lukk et al, 2010)	391	0.0002945
ENPP2 (rs16892852)	Ectonucleotide pyrophosphatase/phosphodiesterase 2	Hydrolyzes lysophospholipids to produce lysophosphatidic acid (LPA) in extracellular fluids. Predominantly expressed in brain, placenta, ovary, and small intestine. Secreted by most body fluids including serum and cerebrospinal fluid (The UniProt Consortium, 2013).	Overexpressed (Lukk et al, 2010)	92	0.04851
GLP1R (rs910171)	Glucagon-like peptide 1 receptor	Member of the glucagon receptor family (also includes glucagon, GLP-2, secretin, GHRH and GIP receptors).In the brain located in hypothalamus and brainstem. Protective against amyloid beta accumulation in rats (Perry & Greig, 2005).	Overexpressed (Lukk et al, 2010)	193	0.01462
MOSC1 (rs746767)	MOCO sulphurase C-terminal domain containing 1	A mitochondrial oxidoreductase, cofactor: molybdenum, is expressed in the brain. MOSC1 is a target for miR-129-5p, like GABBR2, and miR-155, like PICALM.	-	66	0.04507
TM4SF20 (rs4408717)	Transmembrane 4 L six family member 20	Tetraspannin superfamily member. Tetraspanins are often thought to act as scaffolding proteins, anchoring multiple proteins to one area of the cell membrane. Other tetraspanin superfamily members have been implicated in Notch signaling and g-secretase activity modulation (Dunn et al, 2010).	-	95	0.004495

1-SNP model rank: rank of the corresponding SNP in terms of univariate 1-SNP model score

PeerJ PrePrints | https://peerj.com/preprints/123v2/ | v2 received: 11 Dec 2013, published: 11 Dec 2013, doi: 10.7287/peerj.preprints.123v2