Journal of Alzheimer's Disease 25 (2011) 635–644
DOI 10.3233/JAD-2011-110197
IOS Press

Investigating Statistical Epistasis in Complex Disorders

James C. Turton[a,1], James Bullock[a,1], Christopher Medway[a], Hui Shi[a], Kristelle Brown[a], Olivia Belbin[a], Noor Kalsheker[a], Minerva M. Carrasquillo[b], Dennis W. Dickson[b], Neill R. Graff-Radford[b,c], Ronald C. Petersen[d,e], Steven G. Younkin[b] and Kevin Morgan[a,*]

[a]*Human Genetics, School of Molecular Medical Sciences, Queens Medical Centre, University of Nottingham, Nottingham, UK*

[b]*Department of Neuroscience, Mayo Clinic College of Medicine, Jacksonville, FL, USA*

[c]*Department of Neurology, Mayo Clinic College of Medicine, Jacksonville, FL, USA*

[d]*Department of Neurology, Mayo Clinic College of Medicine, Rochester, MN, USA*

[e]*Mayo Alzheimer Disease Research Centre, Mayo Clinic College of Medicine, Rochester, MN, USA*

Handling Associate Editor: Francesco Panza

Accepted 4 March 2011

Abstract. The missing heritability exhibited by late-onset Alzheimer's disease is unexplained and has been partly attributed to epistatic interaction. Methods available to explore this are often based on logistic regression and allow for determination of deviation from an expected outcome as a result of statistical epistasis. Three such methodologies including Synergy Factor and the PLINK modules, –epistasis and –fast-epistasis, were applied to study an epistatic interaction between interleukin-6 and interleukin-10. The models analyzed consisted of two synergistic interactions (SF $\approx$ 4.2 and 1.6) and two antagonistic interactions (SF $\approx$ 0.9 and 0.6). As with any statistical test, power to detect association is paramount; and most studies will be underpowered for the task. However, the availability of large sample sizes through genome-wide association studies make it feasible to examine approaches for determining epistatic interactions. This study documents the sample sizes needed to achieve a statistically significant outcome from each of the methods examined and discusses the limitations/advantages of the chosen approaches.

Keywords: Complex disorders, epistasis, LOAD, modeling, PLINK, synergy factor

INTRODUCTION

Studies into complex disorders

Characterizing the genetic basis of many diseases is proving difficult, despite their relative commonality and familial clustering. This information is needed in order to create or perfect potential screening technologies and develop effective treatments. This can then be further exploited in terms of pharmacogenetics to enable identification of individuals who would benefit most from a tailored drug regime [1]. Complex diseases are influenced in a variety of ways, by both genetic and environmental factors. Genetic risk factors are frequently proving to be of large number and small effect size creating a cumulative effect, which contributes to the overall. This is in contrast to classical Mendelian disorders (such as familial forms of breast and colon cancer, maturity onset diabetes of the

[1] These authors contributed equally to this work.

*Correspondence to: Professor Kevin Morgan, Professor of Human Genomics and Molecular Genetics, Institute of Genetics, School of Molecular Medical Sciences, A Floor, West Block, Room 1306, Queens Medical Centre, Nottingham NG7 2UH, UK. Tel.: 0115 8230724; E-mail: kevin.morgan@nottingham.ac.uk.

young (MODY), and early-onset familial Alzheimer's disease) which are caused by known, fully penetrant genetic factors.

In an attempt to uncover these factors, genome-wide association studies (GWAS) have examined the genomes of thousands of individuals to a depth of several hundred thousand SNPs distributed across the genome, and succeeded in reproducibly identifying a number of novel gene candidates. The AlzGene database is a comprehensive catalogue of published findings in LOAD [2]. However, the reported effect sizes are generally small (odds ratios of 1.5 or less) and therefore cannot explain the overall heritability of these diseases [3]. Despite the contribution made by GWAS in identifying new genes in complex disorders, there is still a sizable component of heritability that remains to be explained. For example, even with the discovery of new genes in late-onset Alzheimer's disease (LOAD) no more than 40–50% of the genetic component is accounted for at present. The current hypothesis is that such disorders are both genetically complex; no simple mode of inheritance, and heterogeneous; a result of mutations and polymorphisms in multiple genes that interact together and with non-genetic factors [4].

Epistasis

Several theories exist as to which factors account for this missing heritability, and it is likely to be a combination of several factors that will explain it more completely. One possible explanation is that of epistasis, a concept which has become an increasingly used generic term to mean any different way in which one area of the genome can interact with another factor [5]. In studying epistasis, a distinction between the usages of the term must be made and their differences understood so that appropriate investigations and methodologies can be applied, and pertinent conclusions drawn. This too is considered a limitation of examining epistasis and is often cited as the reason for poor reproducibility of findings. These explanations of epistasis can be roughly categorized into three broad forms which are summarized here [6]

Compositional epistasis

Of the three, this is the most similar to the original term coined around 100 years ago by William Bateson [7] to represent the effects of one allele masking those of an allele at a different loci.

Functional epistasis

This represents protein-protein interactions as opposed to strictly genetic ones, and is perhaps the least suited to be described as epistasis, but often the most commonly inferred by authors.

Statistical epistasis

Finally we have the form that will be focused on in this paper. It looks at differences between the observed and expected odds ratios when we compare the combined odds ratio (OR) of two SNPs with the additive effect of the two SNPs individually.

Alzheimer's disease

Late Onset Alzheimer's Disease (LOAD) is a neurodegenerative disease most commonly found in individuals over the age of 65. The disease is characterized by extensive cognitive decline, the deposition in the brain of extracellular senile plaques, predominantly composed of amyloid-β (Aβ) peptide and the development of intraneuronal neurofibrillary tangles formed from abnormally phosphorylated tau protein. Although there is an earlier onset form of the disease with a dominant Mendelian inheritance, the etiology of LOAD is more complex, with multiple environmental and genetic susceptibility factors being implicated.

Prior to the introduction of GWAS, only one consistently identified susceptibility loci in LOAD, apolipoprotein E (*APOE*), in particular the ε4 allele had been identified. Since 2009 and the completion of several large GWAS, strong evidence for additional candidate LOAD genes such as *PICALM*, *CLU, CR1* have been identified, thus corroborating the utility of GWAS [8, 9]. Recent meta-analyses and replication studies have provided further evidence for these genetic factors and added a previously suspected AD candidate, *BIN1*, to genome-wide significance level, again contributing to the overall heritability [10–12]. However these findings still do not account for the entire estimated genetic component. Therefore epistasis represents a potential next step in studies of AD, and finding a way to accurately and efficiently predict important interactions is essential [13].

Approaches explored

The software package PLINK (v1.07) [14] is a bio-informatics toolkit designed for whole-genome analysis, and includes several packages designed to look for evidence of epistasis. The –epistasis package uses linear or logistic regression to provide an odds ratio (OR) for the interaction of two SNPs, while the –fast-epistasis is designed as a screening tool to provide an approximation to statistical interaction between pairs of SNPs in a population. Synergy factor (SF) [15], also based on regression, is a method to identify potential interactions between pairs of SNPs and to provide information regarding the degree of that interaction.

In order to examine these methods we generated synergistic and antagonistic models based on existing studies. Combarros and colleagues [15] examined and reproduced a reported interaction between SNPs in genes for the inflammatory cytokines interleukin-6 (IL-6) and interleukin-10 (IL-10) in LOAD. Using this interaction and breaking down the tools and methods employed to achieve this, models were constructed to represent a range of synergy factors that could then also be manipulated appropriately for use in PLINK, a tool set that is being extensively used in the analysis of GWAS data [14]. Running the different approaches on the same datasets allows for a direct comparison of how effectively they handle a range of SF values and sample sizes, as well as their speed and ease of use. This in turn allows the optimization of our laboratory protocols.

MATERIALS AND METHODS

In an attempt to explore the utility of the methods described above a series of models were generated, initially based on published examples which were then extrapolated to include a variety of circumstances.

SNP selection and proxy ascertainment

To test the three approaches, we have utilized the replicated epistatic interaction between the inflammatory cytokines IL-6 (rs2069837) and IL-10 (rs1800871), for which the reported interaction gave an SF of 1.63 (95% confidence interval: 1.10–2.41, $p = 0.01$) [15].

The genotype information for model generation was taken from the samples from Mayo Clinic, Florida that were included in the published GWAS [16]. This genotyping was performed on the Illumina HumanHap300 BeadChip. Subject-level genotype data was provided in PLINK binary format, (.bed, .bim, and .fam).

In order to ensure that full coverage of the gene region was achieved, the physical location of each SNP in the GWAS was ascertained using the HapMap Genome Browser (Phase 1, 2 & 3 - Build 36, Release Date: May 2010, available at http://hapmap.ncbi.nlm.nih.gov/cgi-perl/gbrowse/hapmap3r3_B36/), producing a list that included all SNPs lying within the 20 kb 5' and 3' flanking region of the GWAS SNP. This list was then compared with a list of all those present in the GWAS. An example command string used in PLINK is as follows:

> >plink –bfile '*input file*' –chr 1 –from-bp 204993257 –to-bp 205033257 –write-snplist –out '*output file*'

Proxies for all SNPs from HapMap that were not included in the GWAS were obtained using SNP Annotation and Proxy Search (SNAP) v2.1 [17], available at http://www.broadinstitute.org/mpg/snap/ldsearch.php

Model generation

As a way to test the efficiency and accuracy of the three approaches, a range of models had to be designed. The first model utilized a strongly epistatic interaction with an SF $\approx$ 4.2. The second model, based on the Epistasis Project's attempt to replicate findings between variants in the inflammatory cytokines *interleukin-6* (*IL-6*) and *interleukin-10* (*IL-10*) [15], resulted in an SF $\approx$ 1.6. These two models would represent situations where the epistasis in the system would be synergistic, i.e., the two SNPs present together would have a larger OR than the additive effect of each if they were affecting the system independently. An SF $\approx$ 4.2 represents a strong interaction, while an SF $\approx$ 1.6 is a more moderate example, and closer to what might be expected in a biological setting.

Two additional models were also tested, the first of which was an antagonistic model of SF $\approx$ 0.9, where possession of one SNP would reduce the effect size of a second SNP. Finally, an SF $\approx$ 0.6 represents the reciprocal of the SF $\approx$ 1.6 model, and thus provides a second antagonistic model.

To model a strong SF (4.2), populations were constructed to give an appropriate SF. To increase sample numbers in each model the same population was duplicated to generate a sample number range of $n = 475$–13300 for investigation. Table 1 shows the stratified counts for each model population.

Synergy factors were calculated using the Excel spreadsheet SF calculator provided by Cortina-Borja et al. [15]. As shown in Table 2, genotype counts for a population are stratified into four groups determined by 1) possession of the protective allele at both SNPs (used as the reference group), 2) and 3) possession of at least one copy of the risk allele at either SNP (0.821 and 1.211) and 4) possession of the risk allele at both SNPs

Table 1
Basic Model population frequencies

SNP1	SNP2	Model							
		SF = ≈4.2		SF = ≈1.6		SF = ≈0.9		SF = ≈0.6	
		Controls	Cases	Controls	Cases	Controls	Cases	Controls	Cases
–	–	150	103	150	103	150	103	148	105
+	–	25	16	22	19	61	30	56	35
–	+	92	66	83	75	47	44	51	50
+	+	6	17	8	15	25	15	28	12

In each model $n = 475$. The + and – symbols relate to the combination of the presence/absence of a minor allele at each SNP.

Table 2
The Synergy Factor Calculator

SNP1	SNP2	Control	Case	Odds ratio	ln(OR)	var ln(OR)	se ln(OR)			lower	upper	alpha
–	–	152	96	Ref				SF=	5.097	1.442	18.019	0.050
+	–	27	14	0.821	−0.197	0.125	0.354	se(ln(SF))=	0.644			
–	+	85	65	1.211	0.191	0.044	0.210	ln(SF)=	1.629			
+	+	5	16	5.067	1.623	0.279	0.529	Z=	2.528			
								p=	0.011			

Calculator includes formulas for calculating p-value and 95% confidence interval, constructed in Microsoft Excel 2007 provided by Cortina-Borja et al. [15], supplementary material. This is the specified example showing a strong synergistic interaction, SF = 5.1 (95% CI 1.442–18.019, $p = 0.011$). The count of cases and controls is populated by the user.

(5.067). Odds ratios (ORs) are subsequently calculated for each stratified group. The predicted OR assuming that there is no interaction between the SNPs is calculated by multiplying the two ORs for possession of each SNP (0.821*1.211 = 0.994). The actual combined OR is that given for possession of the minor allele at both SNPs (5.067). The SF is calculated as the ratio of actual combined OR: predicted OR (5.067/0.994 = 5.098).

PLINK: Epistasis

Once appropriate model populations had been generated, epistasis was calculated for each using the –epistasis function of PLINK using the following command line:

```
> plink –file 'input filename' –epistasis –epi1 1 –out 'output filename'
```

The use of –epi1 1 ensures that the output generated is not limited to just interactions with a p value less than 1×10^{-4} (PLINK default cut off) as we are looking for the point at which interactions become significant.

PLINK: Fast Epistasis

Epistasis was subsequently calculated for the same populations using the –fast-epistasis function of PLINK:

```
> plink –file 'input filename' –fast-epistasis –epi1 1 –out 'output filename'
```

RESULTS

Proxy

The genotyping chip used for the Mayo dataset included rs2069837 (IL-6, chr7 :22734552) but not rs1800871 (IL-10, chr1 :205013257) and therefore a proxy SNP was required to provide the same genotype information.

After application of the bioinformatics tool SNAP Proxy [17], two potential proxy SNPs in strong linkage disequilibrium ($r^2 > 0.8$) were identified:

rs3024490 - chr1 :205011934 , $r^2 = 0.951$ (1323 bp 3' rs1800871)
rs1518111 - chr1 :205011268 , $r^2 = 0.951$ (1989 bp 3' rs1800871)

Since rs3024490 had been genotyped in the Mayo dataset this SNP was chosen as a suitable proxy for rs1800871 (IL-10).

Modeling

Table 3 summarizes the minimum number of samples required to achieve significance using each method for the modeled datasets. The synergy factor interaction for the ≈4.2 model (SF ≈ 4.2, 95% CI = 1.29–13.97, $p = 0.018$) was corroborated by the PLINK modules also showing a strong statistical interaction in the same direction (OR of interaction = 2.79,

Table 3
Minimum number of samples required to achieve statistical significance

Model	Method	Total Samples	Synergy Factor	Odds Ratio of Interaction	χ 2 (O-E)	Odds Ratio for Difference in Association	95% Confidence Interval	*p* Value
≈4.2	Synergy Factor	475	4.237	–	–	–	1.286–13.968	0.018
	–epistasis	475	–	2.788	3.861	–	–	0.049
	–fast-epistasis	950	–	–	–	6.993	–	0.008
≈1.6	Synergy Factor	2850	1.650	–	–	–	1.041–2.615	0.033
	–epistasis	13300	–	1.202	4.005	–	–	0.045
	–fast-epistasis	15200	–	–	–	4.021	–	0.045
≈0.9	Synergy Factor	33000	0.89	–	–	–	0.799–0.995	0.04
	–epistasis	3800	–	0.734	5.060	–	–	0.024
	–fast-epistasis	1900	–	–	–	4.167	–	0.041
≈0.6	Synergy Factor	1900	0.620	–	–	–	0.389–0.989	0.045
	–epistasis	950	–	0.573	3.952	–	–	0.047
	–fast-epistasis	950	–	–	–	4.758	–	0.029

The $p > 0.05$ value for each model (≈4.2, ≈1.6, ≈0.9 and ≈0.6) for assessed methodologies (synergy factor, –epistasis and –fast-epistasis).

$p = 0.049$). Modeling a SF of 4.2 required the least number of samples to achieve significance; as can be seen from Table 3, both synergy factor and –epistasis required 475 samples and –fast-epistasis, 950 samples. However for synergy factor testing, sample numbers greater than 5700 produced a *p* value too small for display in Microsoft Excel, and instead displays as $p = 0.00E+00$.

All three tests attained a significant p value with the sample numbers shown for the model of a synergy factor ≈ 0.6; with both PLINK modules requiring half the samples than the SF calculator (950 samples versus 1900).

The SF ≈ 1.6 model generally required more samples to detect an interaction than any other model. Significant synergy factor was detectable in a model containing 2850 samples (SF ≈ 1.6, 95% CI = 1.04–2.62, $p = 0.033$), although to detect a similar level of interaction in PLINK (–epistasis) required approximately 13300 samples producing an OR of interaction of 1.20, $p = 0.045$. The –fast-epistasis module failed to achieve a significant interaction within the sample number range. However by extrapolating it was determined that a significant p value was obtainable with 15200 samples, –fast-epistasis; $n = 15200$, $p = 0.045$.

In the case of the SF ≈ 0.9 model, both –epistasis and –fast-epistasis achieved significance with 3800 and 1900 samples respectively. Synergy factor required significantly more samples than were provided within the boundaries of this investigation, however an exercise to provide completeness to the models produced these results: $n = 33000$, SF ≈ 0.9, 95% CI = 0.799–0.995, $p = 0.04$.

As shown in Table 4, significance of the test statistics increased with greater sample numbers used as well as a contraction of the 95% confidence interval.

Graphical representation of the modeling assessment (Fig. 1) was obtained by plotting -log *p* values against sample numbers to provide a visual comparison of each of the three methodologies being explored for each model.

For the synergistic models (SF > 1, Fig. 1A and 1B) the SF test achieved each significance level with fewer sample numbers than the PLINK modules. Conversely for the antagonistic models (SF < 1, Fig. 1C and 1D) the PLINK modules required fewer sample numbers to achieve significance than synergy factor assessment.

DISCUSSION

The generation of model datasets has provided a platform from which to explore statistical epistatic interactions, although this study cannot be seen to be exhaustive.

The overall ethos of this modeling is an interaction between risk factors (OR > 1) for LOAD. This fact directs the interpretation of the synergy factor. In this instance, where a SF > 1 is obtained the model is synergistic; conversely a SF < 1 is termed antagonistic. In the case of protective alleles the reverse is true. Here we have assessed a strongly synergistic model (SF ≈ 4.2) that easily achieved significance within the sample numbers explored by the defined methods. A limitation of the synergy factor method can be observed in that the spreadsheet was unable to display a p value less than 2.22×10^{-16} with sample numbers > 5700; possibly as a result of an inherent limitation of the Microsoft Excel program, requiring the user to obtain the true *p*-value from an external source using significance tables.

The strongly antagonistic model (SF ≈ 0.6) did not appear to be limited, though further exploration of a

Table 4
95% Confidence Intervals and corresponding *p* values for each of the four models generated when testing using synergy factor

Sample Numbers	≈4.2		≈1.6		≈0.9		≈0.6	
	95% CI	*P* Value	95% CI	*P* Value	95% CI	*P* Value	95% CI	*p* Value
475	1.2885–13.9683	0.018	0.5337–5.0995	0.385	0.3593–2.2287	0.811	0.2441–1.5760	0.315
950	1.8231–9.8495	0.001	0.7428–3.6642	0.219	0.4694–1.7060	0.736	0.3208–1.1994	0.156
1900	2.3339–7.6935	<0.0001	0.9384–2.9006	0.082	0.5670–1.4122	0.633	0.3891–0.9887	0.045
2850	2.6039–6.8960	<0.0001	1.0408–2.6153	0.033	0.6165–1.2988	0.559	0.4239–0.9076	0.014
3800	2.7794–6.4604	<0.0001	1.1070–2.4587	0.014	0.6481–1.2356	0.500	0.4461–0.8625	0.005
4750	2.9060–6.1791	<0.0001	1.1547–2.3573	0.006	0.6705–1.1942	0.451	0.4619–0.8330	0.002
5700	3.0031–5.9793	<0.0001	1.1911–2.2851	0.003	0.6876–1.1645	0.408	0.4739–0.8119	0.001
6650	3.0031–5.9793	<0.0001	1.2203–2.2306	0.001	0.7012–1.1420	0.372	0.4834–0.7958	<0.0001
7600	3.1448–5.7097	<0.0001	1.2443–2.1875	0.001	0.7123–1.1242	0.340	0.4913–0.7831	<0.0001
8550	3.1989–5.6132	<0.0001	1.2645–2.1525	<0.0001	0.7217–1.1096	0.311	0.4979–0.7727	<0.0001
9500	3.2454–5.5328	<0.0001	1.2819–2.1234	<0.0001	0.7297–1.0974	0.286	0.5035–0.7641	<0.0001
10450	3.2860–5.4645	<0.0001	1.2970–2.0986	<0.0001	0.7366–1.0870	0.263	0.5084–0.7567	<0.0001
11400	3.3217–5.4057	<0.0001	1.3104–2.0772	<0.0001	0.7428–1.0781	0.242	0.5128–0.7503	<0.0001
12350	3.3536–5.3543	<0.0001	1.3223–2.0585	<0.0001	0.7482–1.0702	0.224	0.5166–0.7447	<0.0001
13300	3.3823–5.3089	<0.0001	1.3329–2.0420	<0.0001	0.7531–1.0633	0.207	0.5201–0.7398	<0.0001

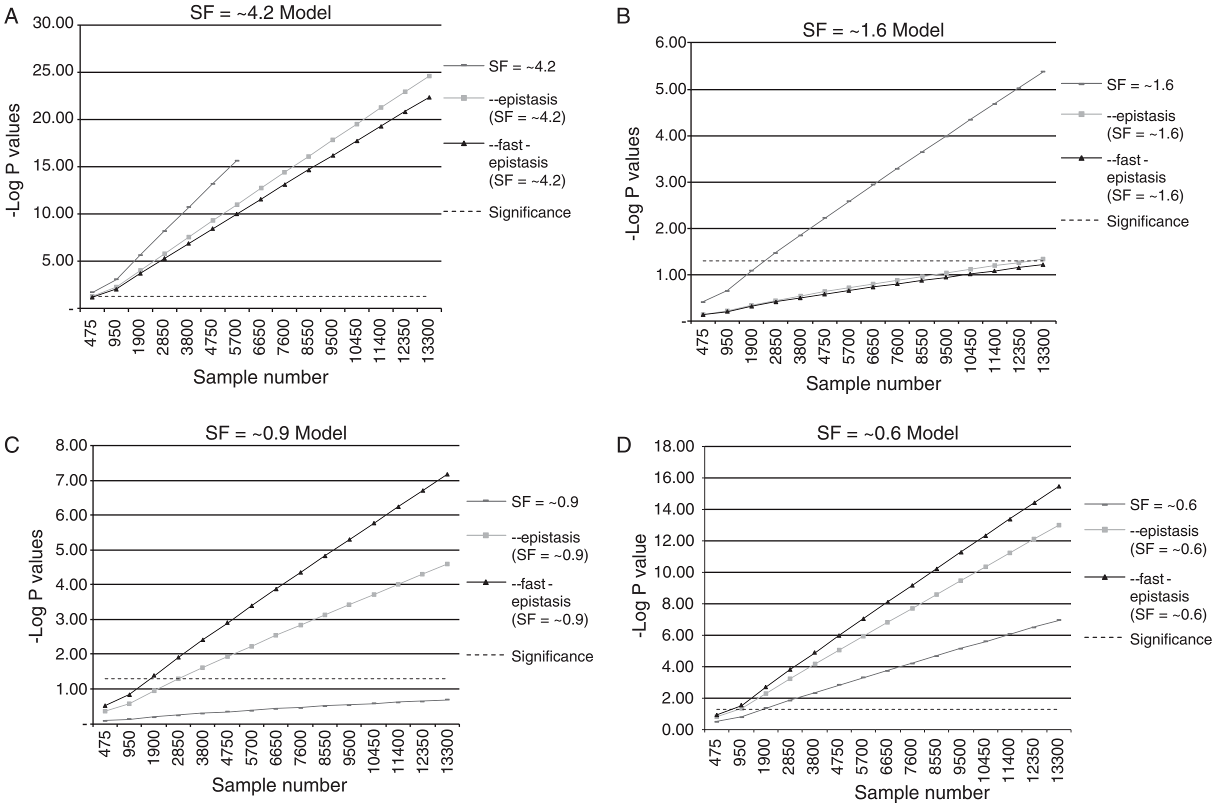

Fig. 1. Graphical representation of the performance of each method. Sample numbers (x-axis) plotted against -log *p* values (y-axis). In each case the significance threshold is reported as -log 1.30 ($p = 0.05$) represented by the horizontal dotted line. A) Model SF ≈ 4.2; B) Model SF ≈ 1.6; C) Model SF ≈ 0.9; D) Model SF ≈ 0.6.

model that is the inverse of SF ≈ 4.2 (≈0.24), and thus even more strongly antagonistic may similarly reveal an issue with *p* value generation.

Based on the models described here, as shown in Table 5, large datasets (>12350) would be required to obtain a suitably significant statistical interaction ($p < 0.05$). The requirement in this study for sample numbers over and above those initially explored to attain significance in both the ≈1.6 and ≈0.9 model is an indication of the limitation of these types of investigation. It is not beyond the realms of possibility that to detect subtle interactions between SNPs with modest effect sizes, in the order of OR < 1.2, large sample sets will be needed. Interactions not involving *APOE* in LOAD are likely to fall into this category [13].

An anomaly observed in this modeling, which can clearly be seen in the graphs in Fig. 1, is that in synergistic models the synergy factor test is more sensitive, requiring comparatively less sample numbers to achieve significance than PLINK testing. The opposite, in antagonistic models, occurs where PLINK is more sensitive than synergy factor. Further exploration of protective factor interaction (where synergy = SF < 1 and antagonism = SF > 1), risk versus protective factors and more complex interactions (3 + or inclusion of environmental factors) would be needed to enable a more complete explanatory hypothesis to be derived. It is well known that the *APOE* $\varepsilon2$ allele is protective and $\varepsilon4$ confers risk in AD [18]. Also interaction between the inflammatory cytokines IL6 (risk) and IL10 (protective) has also been recently reported [15].

Synergy Factor is based on logistic regression which freely enables the inclusion of covariates in the calculation of odds ratio (OR) statistics. In this way we can assess the deviation from an expected outcome in case-control data between factors existing together or independently. Details on the use of synergy factor, limitations and application have been discussed previously [15]. One limitation of the method is the fact that it cannot be easily used for investigations of whole genome datasets; it is more suitable for examining a finite number of suspected interactions due to the required manual user input.

The –epistasis module within PLINK is similarly based on application of logistic regression algorithms for the generation of the OR of Interaction for allele dosage calculations. Conversely –fast-epistasis is considered as a screening tool by the authors of PLINK [14] since the resultant statistic is an approximation to a statistical interaction.

PLINK is used extensively in the search for genetic risk factors and is not as limited as the labor intensive SF calculator. When using GWAS data, we propose the use of –fast-epistasis as a suitable screening tool to identify the top 100 GWAS ‘hits’ for further examination. We recommend this screening to be followed by the application of either –epistasis or the synergy factor calculator to then focus on potential interactions, for confirmation and replication in further sample sets.

Another limitation of these methods is in the number of interactions that can be simultaneously analyzed. Both –epistasis and –fast-epistasis in PLINK are designed to examine pair-wise combinations of SNPs only [14]. Synergy factor is slightly more versatile and to include a third interacting factor would require the user to perform two sequential SF calculations, i.e., SNP x SNP, after initial stratification based on a previous SNP. This method may then be hampered by insufficient power in smaller datasets [15].

With sufficient computer facilities both association analysis and determination of statistical epistatic interaction could technically be performed in parallel on GWAS datasets. We show here that sample numbers typically used in GWAS (5000–10000) are sufficiently large to detect even modest epistatic interactions.

There is much to be explored in the investigation of the genetic architecture of complex diseases such as LOAD - particularly regarding rare variation. This may provide a plethora of previously unconsidered potential epistatic interactions and biological pathways that further bolster evidence found for association - i.e., factors may be associated with disease because of the potential for epistatic interaction. This may in turn provide a parallel direction of further study into pharmacogenetics and enable the pinpointing of individuals for a range of different treatments. Epistasis has already been implicated in antiepileptic drug resistance from a pharmacogenetic perspective [19], and while this study highlights limitations we feel this is a point that needs to be considered in complex disorders, such as LOAD, and further enhances the utility in determining evidence for epistatic interaction.

We are only looking at statistical interactions; therefore attempting to discuss the interactions from a biological perspective is beyond the remit of this study. However, by effectively ‘screening’ for interactions we can then use this information to identify potential gene-gene interactions and make economical use of laboratory and investigator time which may help to direct studies towards potentially important biological pathways. The presence of a true interaction being observed in these model datasets, however, cannot be commented on from these results. Larger sample collections and extensive functional studies would be

Table 5

Complete set of p values obtained from the different models and methodologies

Sample Size	475	950	1900	2850	3800	4750	5700	6650	7600	8550	9500	10450	11400	12350	13300
SF = 4.2	1.770E-02	7.920E-04	2.080E-06	6.180E-09	1.940E-11	6.240E-14	2.220E-16	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001
–epistasis = 4.2	4.943E-02	5.457E-03	8.504E-05	1.487E-06	2.737E-08	5.183E-10	9.999E-12	1.954E-13	3.857E-15	7.669E-17	1.534E-18	3.082E-20	6.218E-22	1.259E-23	2.555E-25
–fast-epistasis = 4.2	6.150E-02	8.812E-03	1.842E-04	4.647E-06	1.233E-07	3.367E-09	9.366E-11	2.640E-12	7.513E-14	2.154E-15	6.212E-17	1.800E-18	5.237E-20	1.529E-21	4.475E-23
SF = 1.6	3.845E-01	2.188E-01	8.200E-02	3.320E-02	1.390E-02	5.964E-03	2.594E-03	1.140E-03	5.050E-04	2.250E-04	1.010E-04	4.530E-05	2.050E-05	9.260E-06	4.200E-06
–epistasis = 1.6	7.053E-01	5.928E-01	4.494E-01	3.543E-01	2.848E-01	2.317E-01	1.902E-01	1.571E-01	1.303E-01	1.086E-01	9.078E-02	7.608E-02	6.392E-02	5.381E-02	4.537E-02
–fast-epistasis = 1.6	7.230E-01	6.162E-01	4.784E-01	3.853E-01	3.161E-01	2.623E-01	2.195E-01	1.847E-01	1.562E-01	1.326E-01	1.129E-01	9.640E-02	8.247E-02	7.070E-02	6.070E-02
SF = 0.9	8.114E-01	7.358E-01	6.332E-01	5.589E-01	4.997E-01	4.505E-01	4.085E-01	3.720E-01	3.398E-01	3.114E-01	2.859E-01	2.630E-01	2.424E-01	2.237E-01	2.067E-01
–epistasis = 0.9	4.265E-01	2.607E-01	1.117E-01	5.141E-02	2.449E-02	1.191E-02	5.871E-03	2.924E-03	1.467E-03	7.407E-04	3.758E-04	1.914E-04	9.779E-05	5.012E-05	2.575E-05
–fast-epistasis = 0.9	3.074E-01	1.489E-01	4.122E-02	1.242E-02	3.892E-03	1.249E-03	4.069E-04	1.341E-04	4.457E-05	1.491E-05	5.013E-06	1.693E-06	5.740E-07	1.952E-07	6.653E-08
SF = 0.6	3.155E-01	1.557E-01	4.468E-02	1.394E-02	4.522E-03	1.502E-03	5.060E-04	1.730E-04	5.940E-05	2.050E-05	7.150E-06	2.500E-06	8.750E-07	3.080E-07	1.090E-07
–epistasis = 0.6	1.598E-01	4.681E-02	4.931E-03	5.746E-04	7.007E-05	8.775E-06	1.118E-06	1.442E-07	1.877E-08	2.461E-09	3.244E-10	4.296E-11	5.710E-12	7.616E-13	1.019E-13
–fast-epistasis = 0.6	1.230E-01	2.916E-02	2.036E-03	1.580E-04	1.286E-05	1.075E-06	9.152E-08	7.890E-09	6.866E-10	6.018E-11	5.305E-12	4.699E-13	4.179E-14	3.728E-15	3.335E-16

required to fully explore the existence of any interaction.

ETHICS

Approval was obtained from the ethics committee and institutional review board for the ascertainment and collection of samples (Mayo Clinic College of Medicine, Jacksonville). Written informed consent was obtained for all individuals that participated in this study.

ACKNOWLEDGMENTS

This work was supported by the Alzheimer's Research Trust and the Big Lottery Fund.

Authors' disclosures available online (http://www.j-alz.com/disclosures/view.php?id=811).

REFERENCES

[1] Sleegers K, Jean-Charles L, Bertram L, Cruts M, Amouyel P, Van Broeckhoven C (2009) The pursuit of susceptibility genes for Alzheimer's disease: progress and prospects. *Trends Genetics* **26**, 84-93.

[2] Bertram L, McQueen M, Mullin K, Blacker D, Tanzi RE *The AlzGene Database*. Available from:http://www.alzgene.org. Accessed 22 October 2010.

[3] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. *Nature* **461**, 747-753.

[4] Bertram L, Tanzi RE (2004) The current status of Alzheimer's disease genetics: what do we tell the patients? *Pharmacol Res* **50**, 385-396.

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

[6] Phillips P (2008) Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9**, 855-867.

[7] Bateson W (1910) Mendels principles of heredity. *Mol Gen Genet* **3**, 108-109.

[8] Harold D, Abraham R, Hollingworth P, Sims R, Gerrish A, Hamshere ML, Pahwa JS, Moskvina V, Dowzell K, Williams A, Jones N, Thomas C, Stretton A, Morgan AR, Lovestone S, Powell J, Proitsi P, Lupton MK, Brayne C, Rubinsztein DC, Gill M, Lawlor B, Lynch A, Morgan K, Brown KS, Passmore PA, Craig D, McGuinness B, Todd S, Holmes C, Mann D, Smith AD, Love S, Kehoe PG, Hardy J, Mead S, Fox N, Rossor M, Collinge J, Maier W, Jessen F, Schurmann B, van den Bussche H, Heuser I, Kornhuber J, Wiltfang J, Dichgans M, Frolich L, Hampel H, Hull M, Rujescu D, Goate AM, Kauwe JSK, Cruchaga C, Nowotny P, Morris JC, Mayo K, Sleegers K, Bettens K, Engelborghs S, De Deyn PP, Van Broeckhoven C, Livingston G, Bass NJ, Gurling H, McQuillin A, Gwilliam R, Deloukas P, Al-Chalabi A, Shaw CE, Tsolaki M, Singleton AB, Guerreiro R, Muhleisen TW, Nothen MM, Moebus S, Jockel K-H, Klopp N, Wichmann HE, Carrasquillo MM, Pankratz VS, Younkin SG, Holmans PA, O'Donovan M, Owen MJ, Williams J (2009) Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet* **41**, 1088-1093.

[9] Lambert J-C, Heath S, Even G, Campion D, Sleegers K, Hiltunen M, Combarros O, Zelenika D, Bullido MJ, Tavernier B, Letenneur L, Bettens K, Berr C, Pasquier F, Fievet N, Barberger-Gateau P, Engelborghs S, De Deyn P, Mateo I, Franck A, Helisalmi S, Porcellini E, Hanon O, de Pancorbo MM, Lendon C, Dufouil C, Jaillard C, Leveillard T, Alvarez V, Bosco P, Mancuso M, Panza F, Nacmias B, Bossu P, Piccardi P, Annoni G, Seripa D, Galimberti D, Hannequin D, Licastro F, Soininen H, Ritchie K, Blanche H, Dartigues J-F, Tzourio C, Gut I, Van Broeckhoven C, Alperovitch A, Lathrop M, Amouyel P (2009) Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet* **41**, 1094-1099.

[10] Corneveaux JJ, Myers AJ, Allen AN, Pruzin JJ, Ramirez M, Engel A, Nalls MA, Chen K, Lee W, Chewning K, Villa SE, Meechoovet HB, Gerber JD, Frost D, Benson HL, O'Reilly S, Chibnik LB, Shulman JM, Singleton AB, Craig DW, Van Keuren-Jensen KR, Dunckley T, Bennett DA, De Jager PL, Heward C, Hardy J, Reiman EM, Huentelman MJ (2010) Association of CR1, CLU and PICALM with Alzheimer's disease in a cohort of clinically characterized and neuropathologically verified individuals. *Human Mol Genet* **19**, 3295-3301.

[11] Seshadri S, Fitzpatrick AL, Ikram MA, DeStefano AL, Gudnason V, Boada M, Bis JC, Smith AV, Carassquillo MM, Lambert JC, Harold D, Schrijvers EM, Ramirez-Lorca R, Debette S, Longstreth WT Jr, Janssens AC, Pankratz VS, Dartigues JF, Hollingworth P, Aspelund T, Hernandez I, Beiser A, Kuller LH, Koudstaal PJ, Dickson DW, Tzourio C, Abraham R, Antunez C, Du Y, Rotter JI, Aulchenko YS, Harris TB, Petersen RC, Berr C, Owen MJ, Lopez-Arrieta J, Varadarajan BN, Becker JT, Rivadeneira F, Nalls MA, Graff-Radford NR, Campion D, Auerbach S, Rice K, Hofman A, Jonsson PV, Schmidt H, Lathrop M, Mosley TH, Au R, Psaty BM, Uitterlinden AG, Farrer LA, Lumley T, Ruiz A, Williams J, Amouyel P, Younkin SG, Wolf PA, Launer LJ, Lopez OL, van Duijn CM, Breteler MM; CHARGE Consortium; GERAD1 Consortium; EADI1 Consortium (2010). Genome-wide Analysis of Genetic Loci Associated With Alzheimer Disease. *JAMA* **303**, 1832-1840.

[12] Carrasquillo MM, Belbin O, Hunter TA, Ma L, Bisceglio GD, Zou F, Crook JE, Pankratz VS, Dickson DW, Graff-Radford NR, Petersen RC, Morgan K, Younkin SG (2010) Replication of CLU, CR1, and PICALM Associations With Alzheimer Disease. *Arch Neurol* **67**, 961-964.

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

[14] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Human Genet* **81**, 559-575.

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

[16] Carrasquillo MM, Zou F, Pankratz VS, Wilcox SL, Ma L, Walker LP, Younkin SG, Younkin CS, Younkin LH, Bisceglio GD, Ertekin-Taner N, Crook JE, DicksonDW, Petersen RC, Graff-Radford NR (2009) Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. *Nat Genet* **41**, 192-198.

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

[18] Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Gaskell PC, Rimmler JB, Locke PA, Conneally PM, Schmader KE, Small GW, Roses AD, Haines JL, Pericak-Vance MA (1994) Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nat Genet* **7**, 180-184.

[19] Kim M-K, Moore JH, Kim J-K, Cho K-H, Cho Y-W, Kim Y-S, Lee M-C, Kim Y-O, Shin M-H, (2011) Evidence for epistatic interactions in antiepileptic drug resistance. *J Human Genet* **56**, 71-76.