Introduction

Discovering genuine statistical associations between common human diseases and genetic variation not only offers the possibility for pre-symptomatic diagnosis, but could also identify novel pathways involved in disease. Given the complex and poorly understood aetiology of many common diseases, this is a valuable way of identifying new targets for treatment. However, discovering a genuine association appears to be increasingly challenging and unearthing the heritable contribution to common diseases has proven difficult [1]. Although the genetics of some common diseases are better understood than others, which may be due to greater genetic homogeneity, methodologically there are a number of reasons why a large portion of genetic risk may remain hidden. Exploring candidate genes based on prior knowledge of disease pathophysiology may not detect novel pathways. In addition many studies have been insufficiently powered to look at common diseases which are either genetically heterogeneous or harbour variants conveying small effects. Finally research to date has largely concentrated on common single nucleotide polymorphisms (SNPs) and other forms of genetic variation (rare SNPs and structural variation) remain
Alzheimer’s disease biomarkers

relatively poorly understood. The latest SNP ‘chips’ increasingly cover structural variation e.g. copy number variation (CNV) and next generation sequencing will address the contribution of rare variation; emerging data will address both of these issues which current GWAS have not explored.

The advent of genome-wide association studies (GWAS) appears to have negated the issue of choosing a suitable candidate given that it tests SNPs throughout the genome based on linkage disequilibrium (LD) between SNPs. However, many GWAS remain underpowered to detect SNPs with modest effects (Odds Ratio (OR) > 1.5) [2]. Meta-analysis of GWAS datasets is one tool that can reconcile these power issues. While an attractive method, combining genotype level data is often not feasible because different genotyping platforms have been used (each with a unique SNP array). Furthermore genotype level data is not always publically available and summary statistics (p-values rather than allele counts) may be the only available data.

Importantly, because underpowered GWAS may well mask genuine association signals, performing meta-analysis on the ‘best SNP hits’ (SNPs below a defined p-value threshold) will be of limited utility. The ‘best practise’ will be to subject all GWAS SNPs to meta-analysis when using summary statistics. Given that this will be unmanageable when combining 500,000 plus SNPs, an alternative candidate gene meta-analysis approach is attractive. The approach we describe combines the coverage afforded by GWAS to the traditional candidate gene approach, enriching the objectives of both. This has the added advantage of reducing the number of independent tests, thereby relaxing the conservative nature of correcting for multiple testing.

Consequently, in this paper we describe an LD-aware method of candidate gene meta-analysis of summary data using GWAS datasets generated from different chip platforms. Late-onset Alzheimer’s disease (LOAD) is an ideal disease to illustrate this approach because it fulfils a number of the challenges outlined above. Of the 12 LOAD GWAS to date, only three have been sufficiently powered to detect significant SNP associations in novel candidate genes; CLU (OR = 0.86), PICALM (OR = 0.86), CR1 (OR = 1.21) and PCDH11X (OR = 1.30) [3-5]. The remaining 9 studies were underpowered to detect SNPs conveying modest effects (OR<1.5), and the only replicable associations were within the APOE locus [6-14].

We have chosen to perform meta-analysis on four genes which encode potential LOAD cerebro-spinal fluid (CSF) biomarkers; Fibrinogen γ-chain (FGG), Contactin-1 (CNTN1), Contactin-2 (CNTN2) and SPARC-like-1 (SPARCL1) [15]. Contactin-1, Contactin-2 and SPARC-like-1 are cell adhesion molecules, all of which are ubiquitously expressed in the brain and nervous system and are vital for neurodevelopment [16-23]. These have been shown to interact at the protein level with the Amyloid Precursor Protein (APP), which is an established candidate as it contains the Aβ peptide (a traditional AD biomarker) – liberation of which contributes to amyloid plaque formation and downstream pathology as described by the ‘amyloid cascade hypothesis’ [24, 25]. Interestingly, Contactin-1/2 potentially modifies APP cleavage and downstream signalling carried out by liberated domains [26-28]. As a component of fibrinogen, the FGG glycoprotein chain is involved in homeostasis and inflammation. Whilst fibrinogen is predominantly expressed in the liver, it is present in high concentrations in LOAD patients and is associated with plaques [29-31].

Given that LOAD is expected to be a heterogeneous disease, where novel pathways explain disease risk, candidate gene meta-analysis of GWAS (which are individually underpowered to detect modest genetic effects) may allow genuine associations to surface. This paper describes a readily implemented combined proteomic and candidate gene meta-analysis approach providing an additional avenue for biomarker validation.

Materials and methods

Candidate Gene

In addition to the four new potential biomarker genes (CNTN1, CNTN2, FGG and SPARCL1, unpublished observations), three currently assayed LOAD biomarkers – Aβ, Tau and Apolipoprotein-E (APP, MAPT and APOE genes respectively) were also subjected to the same in silico approaches. When analysing FGG, we included the entire fibrinogen locus (FGA, FGG and FGB) as these genes are in close proximity and share...
Alzheimer’s disease biomarkers

Table 1. Summary of GWAS datasets; Reiman et al. [9], Li et al. [14] and Carrasquillo et al. [4]. Sample and SNP numbers taken from PLINK output files

<table>
<thead>
<tr>
<th>Study</th>
<th>Sample Size (Case/Control)</th>
<th>Ancestry</th>
<th>Mean Age</th>
<th>Genotyping Chip</th>
<th>Number of SNPs (after QC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li et al. (2008) [14]</td>
<td>1489 (753/736)</td>
<td>Canada and UK (N.European)</td>
<td>&gt;65</td>
<td>Affymetrix 500K</td>
<td>469,438</td>
</tr>
</tbody>
</table>

Table 2. Summary of biomarker gene size (kb) and location (chromosome and base-pair co-ordinates) and the size (Kb) and co-ordinates of the extended linkage disequilibrium block (ascertained using HapMap CEU data, release 22)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome</th>
<th>Size (kb)</th>
<th>Gene Co-ordinates</th>
<th>Extended LD Block Co-ordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNTN1</td>
<td>12</td>
<td>378</td>
<td>39372625-39750361</td>
<td>661 39323424-39984609</td>
</tr>
<tr>
<td>CNTN2</td>
<td>1</td>
<td>35</td>
<td>203278953-203313759</td>
<td>832 202974054-203705392</td>
</tr>
<tr>
<td>FGA</td>
<td>4</td>
<td>8</td>
<td>155723730-155731347</td>
<td></td>
</tr>
<tr>
<td>FGB</td>
<td>4</td>
<td>8</td>
<td>155703596-155711686</td>
<td>247 155589391-155835920</td>
</tr>
<tr>
<td>FGG</td>
<td>4</td>
<td>9</td>
<td>155744737-155753352</td>
<td></td>
</tr>
<tr>
<td>SPARCL1</td>
<td>4</td>
<td>56</td>
<td>88613514-88669530</td>
<td>104 88604470-88709064</td>
</tr>
<tr>
<td>APP</td>
<td>21</td>
<td>290</td>
<td>26174733-26465003</td>
<td>565 25967838-26533077</td>
</tr>
<tr>
<td>MAPT</td>
<td>17</td>
<td>134</td>
<td>41327624-41461544</td>
<td>136 41326624-41462544</td>
</tr>
<tr>
<td>APOE</td>
<td>19</td>
<td>4</td>
<td>50100879-50104489</td>
<td>46 50077599-50124397</td>
</tr>
</tbody>
</table>

a number of SNPs courtesy of the LD architecture within this locus.

GWAS datasets for in silico analysis

Of the GWAS to date, we have obtained subject-level genotype data for two; Reiman et al. [9] and Carrasquillo et al. [4], and summary data for one other; Li et al. [14] (Table 1). Datasets were converted to PLINK (v1.5) (http://pngu.mgh.harvard.edu/~purcell/plink/) input files (.MAP and .PED) and SNP IDs converted to dbSNP reference numbers where necessary to ensure consistency across datasets [32]. Genotyping quality control measures had already been applied prior to release, and no additional data pruning was performed. Samples that were common to both Carasquillo et al. and Reiman et al. cohorts were removed from the latter. Extended Linkage Disequilibrium (LD) Block of Biomarker Genes

In order to maximise the coverage of each bio-
marker gene, base-pair co-ordinates of flanking SNPs in LD ($r^2 > 0.8$) with biomarker gene SNPs were identified using LD plots generated in HaploView v4.1 ([http://www.broad.mit.edu/mpg/haplovieview/]) from HapMap CEU genotype data (Release 22) [33]. The co-ordinates of these extended linkage blocks formed the parameters for all gene-centric analyses (Table 2).

**Assessing SNP Coverage of Biomarker Genes**

SNP coverage offered by each genotyping platform, Affymetrix 500K [9, 14] and Illumina HumanHap 300K (Carrasquillo et al.), was quantified as a percentage of total HapMap CEU SNPs. A list of SNP IDs falling within the predetermined biomarker gene co-ordinates (plus 2Kb at either flank) was generated using the ‘--write-snplist’ command in PLINK. This was repeated for both genotyping platforms and each biomarker gene.

Using HaploView, SNP IDs within these files were imputed as a tag SNP. Executing the ‘tagger’ algorithm quantified the extent to which these SNPs capture ($r^2 > 0.8$) variation from a reference dataset (HapMap CEU).

**Biomarker SNP Association with LOAD**

Using the predetermined co-ordinates, each biomarker gene underwent an allelic association test ('assoc') in PLINK. This was repeated for each GWAS dataset. No attempt was made to correct p-values for covariates (APOE, age etc) as this information was not available for all datasets.

The generated assoc files were then subjected to a clumping method ('clump-verbose') in PLINK using the following commands: (1) bfile Hapmap; (2) clump-verbose; (3) clump data-set1.assoc,data-set2.assoc,data-set3.assoc; (4) clump-p1 1; (5) clump-p2 1; (6) clump-r2 0.99.

This method pooled p-values ('clump-verbose') from all three datasets ('clump data-set1.assoc,data-set2.assoc,data-set3.assoc') based on LD (in this instance only perfect proxies ($r^2 > 0.99$) were pooled together). HapMap genotype data ('bfile Hapmap') was used to calculate $r^2$ values. No threshold for p-value was imposed ('clump-p1 1,--clump-p2 1') as all SNPs were clumped irrespective of p-value for meta-analysis.

**Fisher’s Combined and Random Effects Meta-analysis**

Within each clump, one SNP from each study was selected and a summary statistic was generated using Fisher’s combined test. For each biomarker gene, combined SNP p-values were corrected for the total number of ‘clumps’ (which represents the number of independent tests). Where possible the same SNP was selected from each platform. Failing this, the closest (based on base pair co-ordinates) perfect proxy was selected. Although two GWAS [9, 14] both use the Affymetrix 500K Chip, heavy SNP drop-out during quality control in the latter meant different proxies were selected on occasions.

It is important to note that Fisher’s combined does not account for the direction of association with disease (i.e. whether the possession of an allele is a risk or protective). Therefore, some seemingly significant SNPs identified with this analysis may have OR which ‘flip’ either side of 1. To identify allele associations, which show a consistent direction of effect, SNPs remaining significant after correction for multiple testing (corr-P < 0.05) were further analysed with an odds ratio meta-analysis (DerSimonian-Laird) using StatsDirect (v2.6.6). Unlike Fisher’s combined, this analysis takes into account the direction of effect. Consequently, it is possible to have a highly significant combined p-value which reports an insignificant odds ratio meta-analysis: an apparent contradictory ‘nonsense’ event due to allele ‘flipping’.

**Functionally Conserved SNPs**

When an association to any particular SNP is discovered it is highly unlikely that the SNP in question is the actual causal/functional variant. Vista Browser ([http://pipeline.lbl.gov/cgi-bin/gateway2]) was used to explore the conservation status of putative candidate SNPs [34]. SNPs falling in conserved regions of the genome are more likely to be disease causing variants. To be considered conserved, a region had to show ≥ 70% homology across man, mouse and rat within a 100bp window. Additionally, when candidate SNPs were poorly conserved, the conservation status of SNPs in LD ($r^2 > 0.8$) were analysed for a potential functional role. This permits the mapping of any associations identified using the meta-analysis approach to potential func-
Alzheimer’s disease biomarkers

Results

Biomarker Gene Coverage

Despite genotyping fewer SNPs across the genome, the Illumina 300K chip [4] offered greater coverage of all biomarker genes than Affymetrix 500K (Reiman et al. and Li et al.). When pooled these platforms did not provide 100% coverage of any biomarker gene (CNTN2 is best at 96% coverage), and coverage of the fibrinogen locus was particularly poor (54%, Table 3). This approach was only approximate (HapMap LD values are based on modest sample numbers and represent only a small proportion of SNPs) but gave an indication of the extent of genotyping gaps in these GWAS.

SNP Meta-analysis

Of the large number of SNPs analysed (n = 1076), of which 474 ‘SNP clumps’ were independent (r² = 1), Fisher’s combined revealed only four significant ‘SNP clumps’ after correction for multiple testing; rs7523477 (CNTN2, corr-P = 0.005), rs4950982 (CNTN2, corr-P = 0.033), rs8079215 (MAPT, corr-P = 0.009) and rs4420638 (APOE, corr-P = 9.24x10⁻³⁸). Of these, only rs7523477 (P = 0.037, OR = 1.23 (95% C.I = 1.01-1.49)) and rs4420638 (P < 0.0001, OR = 3.36 (95% C.I = 2.93 – 3.85) remained significant after random effects meta-analysis (Table 4).

Conservation of Associated Clumps

Although rs7523477 and rs4951168 (a proxy from the same clump – see Table 4) are both downstream of CNTN2, the latter falling in an exon of a different gene (TMEM81), these SNPs are in strong LD (r²>0.9) with three conserved SNPs in the 3’-UTR on CNTN2 (Figure 1).

Discussion

Here we have illustrated an LD-aware approach to allow meta-analysis of summary data generated from multiple GWAS datasets. This is valuable because a large portion of GWAS to date have been underpowered to detect genuine SNP associations with common diseases – which may in part explain the missing heritability. As datasets cannot be directly merged (due to different SNPs arrays being used) combining summary statistics will increase study power, allowing genuine associations to emerge from datasets which are individually underpowered to detect modest genetic effects. Importantly, rather than performing a genome-wide meta-analysis we have used a candidate gene approach. In this study we have selected genes encoding protein biomarkers as we believe they represent viable novel candidates worthy of exploration. By using meta-analysis, coupled with a less conservative correction for multiple testing (p-values only corrected for number of independent SNPs (r² = 1) within the candidate gene LD block), we hope to have reduced some of these power issues. The strategy we describe can be used to complement biomarker studies and other approaches for researching...
Table 4 Results of meta-analysis for LOAD biomarker genes; Fibrinogen (FGA, FGB and FGG), SPARC-like 1 (SPARCL1), Contactin-1 (CNTN1), Contactin-2 (CNTN2), Microtubule associated protein tau (MAPT), Amyloid Precursor Protein (APP) and Apolipoprotein E (APOE).

<table>
<thead>
<tr>
<th>Name</th>
<th># of SNPs</th>
<th># of SNP Clumps (r²=1)</th>
<th>SNP p-value</th>
<th>SNP p-value</th>
<th>SNP p-value</th>
<th>Distance between correlated SNPs (bp)</th>
<th>Combined p-value / Corrected p-value</th>
<th>Random Effects Meta-analysis Odds Ratio (95% C.I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNTN1</td>
<td>235</td>
<td>111</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>rs4951168</td>
<td>0.057</td>
<td>rs4951168</td>
<td>rs7523477 0.005</td>
<td>4836</td>
<td>3x10⁻⁵ / 0.005 0.037 (1.01 - 1.49) 1.18 (0.86 - 1.60)</td>
</tr>
<tr>
<td>CNTN2</td>
<td>407</td>
<td>178</td>
<td>No proxy</td>
<td>-</td>
<td>rs10900451</td>
<td>0.170</td>
<td>rs4950982 0.001</td>
<td>28035</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FGA/FGB/FGG</td>
<td>56</td>
<td>26</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SPARCL1</td>
<td>45</td>
<td>24</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>APP</td>
<td>243</td>
<td>108</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MAPT</td>
<td>84</td>
<td>21</td>
<td>rs17651507</td>
<td>0.195</td>
<td>rs17651507</td>
<td>rs8079215 0.002</td>
<td>5840</td>
<td>4.6x10⁻⁴ / 0.009 0.820 (0.83 - 1.27) 3.36 (2.93 - 3.85)</td>
</tr>
<tr>
<td>APOE</td>
<td>6</td>
<td>6</td>
<td>rs4420638</td>
<td>4.55x10⁻²⁰</td>
<td>rs4420638</td>
<td>2.26x10⁻⁴</td>
<td>0</td>
<td>1.03x10⁻⁷² / 9.24x10⁻³³ &lt;0.0001 (2.93 - 3.85)</td>
</tr>
</tbody>
</table>
SNPs within CNTN2 3'UTR are in strong LD with genotyped SNPs showing association in the meta-analysis. SNPs genotyped in GWAS datasets (rs7523477 and rs4951168) are downstream of CNTN2 – the latter falling in the only exon of TMEM81. Therefore, as rs4951168 is coding it is highly conserved. rs7523477 is intergenic and is poorly conserved in mouse and rat. Both rs7523477 and rs4951168 are in strong LD ($r^2$>0.9) with three SNPs in the 3'UTR of CNTN2 which show high conservation for a non-coding region. To be considered conserved, a region must show 70% or greater sequence similarity within a 100bp window between human, mouse and rat. Conservation output from Vista Browser (v2.0) shows conservation in 6 species compared to humans (top to bottom respectively; Rhesus Monkey, Dog, Horse, Mouse, Rat and Chicken). Where conservation ≥ 70% the plot curve is colored (Light Blue = UTR, Dark Blue = Exon and Red = Intron).
candidate genes in any complex genetic disorder. In this paper we illustrate the utility of this combined approach by using LOAD GWAS data and analysing the genes coding for four biomarkers that were detected in a proteomic screen of AD CSF (unpublished observation) as ‘proof-of-principle’.

Analysis of the 4 potential new biomarkers revealed one SNP clump in an LD block of CNTN2 (containing rs7523477 and rs4951168) which reported a significant meta-analysis association with LOAD. rs7523477 and rs4951168 (both MAF = 15%) are located downstream of CNTN2. rs7523477 is an intergenic SNP located 1.2kb downstream in a region of poor conservation, and rs4951168 is located in an exon of TMEM81 (homo sapiens transmembrane protein 81) a further 4.8kb downstream. Whilst these SNPs are poor functional candidates, they share a number of proxies in CNTN2 3'-UTR (rs7518906, rs10900443 and rs10900444) which are conserved and are not present on either genotyping platform. However, given that we would expect 1 in 20 significant SNP-disease associations by chance, we acknowledge that this finding may be a false positive.

Whilst these may be potential candidate SNPs for regulation of CNTN2, a search of publically available genome-wide expression QTL (eQTL) data has failed to support this; rs4951168 is not significant associated with expression of CNTN2 in human brain tissue (P = 0.37 in the Myers data set [35]) and none of the SNPs linked to rs4951168 feature as a significant regulator of expression in Epstein-Barr virus-transformed lymphoblastoid cell lines (Dixon database [36]). Nevertheless, literature searches reveal Contactin-2 to be a worthy biological candidate.

Contactin-2 is a neuronal GPI-anchored cell adhesion molecule which is essential for neurodevelopment [16, 20]. The Contactin family have been shown to interact with the extracellular domain of APP. CNTN2 has been shown to regulate cleavage of APP by secretases – a process important for neurotoxic Aβ liberation and production of the APP intracellular domain (AICD), which is understood to undergo nuclear translocation and alter transcription [24, 27, 28].

One SNP clump (containing rs4420638) within the LD block of APOE reported a highly significant result following meta-analysis as would be expected. This is the most replicable association see in LOAD GWAS, and is observed due to linkage disequilibrium with the APOE ε4 SNP (rs429358). As has already been well documented no evidence was found to support a genetic role for APP or MAPT in LOAD yet currently these are the most assayed biomarkers. Lack of supportive genetic evidence does not negate the value of a potential biomarker as it is clearly evident that we have gained substantial insight from the study of these genes and their products. It will be interesting to follow emerging data as to whether APOE has any utility as a biomarker in LOAD.

Any SNPs identified using the approaches we describe could well be relevant and make a genuine contribution to the disease process. Failure to find a direct genetic association within these biomarker genes does not necessarily suggest they make no contribution to disease aetiology. The SNPs may mediate secondary effects and/or be responsive to the disease process. Disease associated functional SNPs may regulate the level of gene expression which in turn contributes to the levels present in CSF/serum. We propose that supportive proteomic/genetic data could be used to prioritise candidacy and highlight regions for further study as more data emerges.

We must however acknowledge some of the limitations of this approach. Firstly, gene variation is incompletely tested. As we have shown here, coverage of common variation offered by genotyping platforms is incomplete and this may have contributed to these genes not being identified in any GWAS to date. Furthermore, the role of rare variants (MAF < 5%) and structural variation is not addressed by GWAS. Secondly, due to differences between platform SNP panels (in some instances there were no proxies with which to perform a meta-analysis) gene coverage was further depleted. This is particularly evident in the fibrinogen locus, where poor LD architecture also renders imputation ineffective. Resequencing candidate genes will address these issues and may be vital to discovering the remaining genetic risk of common diseases. Finally, the combined dataset (n=4898) is still underpowered to detect SNPs with small effects (OR < 1.3) at 80% power. This will only be addressed as more datasets become available.
Alzheimer’s disease biomarkers

Acknowledgement

This work was supported by the Alzheimer’s Research Trust and the Big Lottery Fund. We would like to express our sincere thanks to Dr. Minerva Carassquillo and Professor Steven Younkin (Mayo Clinic, Jacksonville, Florida, USA) for kindly providing us with the data from their GWAS.

Please address correspondence to: Kevin Morgan, PhD, Department of Clinical Chemistry, Institute of Genetics, School of Molecular Medical Sciences. A Floor, West Block, QMC, Nottingham. NG7 2UH; 2The Genetics, School of Molecular Medical Sciences. A

References


Alzheimer's disease biomarkers


Int J Mol Epidemiol Genet 2010; 1(2):134-144
Alzheimer's disease biomarkers


[32] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ and Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 2007; 81: 559-575.


