

## Original Article

# Using *In silico* LD clumping and meta-analysis of genome-wide datasets as a complementary tool to investigate and validate new candidate biomarkers in Alzheimer's disease

Christopher Medway<sup>1</sup>, Hui Shi<sup>1</sup>, James Bullock<sup>1</sup>, Holly Black<sup>1</sup>, Kristelle Brown<sup>1</sup>, Baharak Vafadar-isfahani<sup>2</sup>, Balwir Matharoo-ball<sup>2</sup>, Graham Ball<sup>2</sup>, Robert Rees<sup>2</sup>, Noor Kalsheker<sup>1</sup>, Kevin Morgan<sup>1</sup>

<sup>1</sup>Department of Clinical Chemistry, Institute of Genetics, School of Molecular Medical Sciences. A Floor, West Block, QMC, Nottingham. NG7 2UH; <sup>2</sup>The John Van Geest Cancer Research Centre, School of Science and Technology, Nottingham Trent University, Nottingham, UK.

Received November 25, 2009, accepted March 12, 2010, available online: March 18, 2010

**Abstract:** Despite the recent wealth of genome-wide association studies, insufficient power may explain why much of the heritable contribution to common diseases remains hidden. As different SNP panels are genotyped by commercial chips, increasing study power through meta-analysis is made problematic. To address these power issues we suggest an approach which permits meta-analysis of candidate SNPs from multiple GWAS. By identifying correlated SNPs from different platforms ( $r^2=1$ ), using PLINK's 'clumping' method, we generated combined p-values (using Fisher's combined and random effects meta-analysis) for each clump. P-values were corrected for the number of clumps (representing the number of independent tests). We also explored to what extent commercial platforms tag HapMap SNPs within these candidate genes. To illustrate this approach, and to serve as 'proof-of-principle', we used 3 late-onset Alzheimer's disease GWAS datasets to explore SNP-disease associations in 4 new candidate genes encoding cerebro-spinal fluid biomarkers for Alzheimer's disease; Fibrinogen  $\gamma$ -chain (FGG), SPARC-like1 (SPARCL1), Contactin-1 (CNTN1) and Contactin-2 (CNTN2). Genes encoding current Alzheimer's biomarkers; APP (A $\beta$ ), MAPT (Tau) and APOE were also included. This method identified two SNP 'clumps'; one clump in APOE (rs4420638) and one downstream of CNTN2 (which harboured rs7523477 and rs4951168) which were significant following random effects meta-analysis ( $P < 0.05$ ). The latter was linked to three conserved SNPs in the 3'-UTR of CNTN2. We cannot rule out that this result is a false positive due to the large number of statistical tests applied; nevertheless this approach is easily applied and might well have utility in future '-omics' studies.

**Keywords:** PLINK, Clumping, Alzheimer's disease, genome-wide association study (GWAS), CNTN1, CNTN2, SPARCL1, FGG, APOE, meta-analysis

## 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

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  $\gamma$ -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 A $\beta$  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 – A $\beta$ , 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

**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

Study	Sample Size (Case/Control)	Ancestry	Mean Age	Genotyping Chip	Number of SNPs (after QC)
Reiman <i>et al.</i> (2007) [9]	1411 (861/550)	US (N.European)	>65	Affymetrix 500K	312,316
Li <i>et al.</i> (2008) [14]	1489 (753/736)	Canada and UK (N.European)	>65	Affymetrix 500K	469,438
Carrasquillo <i>et al.</i> (2009) [4]	1998 (799/1199)	US (N.European)	>60	Illumina Human- Hap300	313,330

**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)

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

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 Carrasquillo *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/haploview/>) 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 Reiman *et al.* and Li *et al.* 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 dataset1.assoc,dataset2.assoc,dataset3.assoc*; (4) *clump-p1 1*; (5) *clump-p2 1*; (6) *clump-r2 0.99*.

This method pooled SNP p-values (*-clump-verbose*) from all three datasets (*-clump dataset1.assoc,dataset2.assoc,dataset3.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 ( $\text{corr-}P < 0.05$ ) 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  $\geq 70\%$  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-

**Table 3.** Number of SNPs from each study (Reiman *et al.*[9], Li *et al.*[14] and Carrasquillo *et al.*[4]) assayed in each biomarker extended linkage disequilibrium block. Number of SNPs tagged ( $r^2>0.8$ ) in each study is quantified as a percentage of HapMap CEU biomarker SNPs using the same extended linkage block co-ordinates. As there are no HapMap SNPs in APOE (CEU), this analysis was not possible.

Gene	Reiman <i>et al.</i> (2007) [9]	Li <i>et al.</i> (2008) [14]	Carrasquillo <i>et al.</i> (2009) [4]	Pooled
CNTN1	62 SNPs: 56%	96 SNPs: 63%	72 SNPs: 72%	166 SNPs: 80%
CNTN2	96 SNPs: 85%	143 SNPs: 94%	92 SNPs: 80%	226 SNPs: 96%
Fibrinogen Locus (FGA/FGB/FGG)	15 SNPs: 35%	18 SNPs: 35%	23 SNPs: 39%	40 SNPs: 54%
SPARCL1	14 SNPs: 64%	21 SNPs: 75%	17 SNPs: 82%	29 SNPs: 89%
APP	67 SNPs: 73%	96 SNPs: 81%	72 SNPs: 75%	160 SNPs: 92%
MAPT	31 SNPs: 79%	40 SNPs: 81%	14 SNPs: 85%	52 SNPs: 89%

tional regions e.g. coding sequence, promoter regions, splice sites, intronic regulatory regions etc. which potentially merit further 'in depth' investigation.

## Results

### Biomarker Gene Coverage

Despite genotyping fewer SNPs across the genome, the Illumina 300K chip Carrasquillo *et al.* 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 ( $r^2 = 1$ ), 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,  $corr-P = 9.24 \times 10^{-33}$ ). 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 ( $r^2>0.9$ ) with three conserved SNPs in the 3'-UTR on CNTN2 (**Figure 1**).

## Discussion

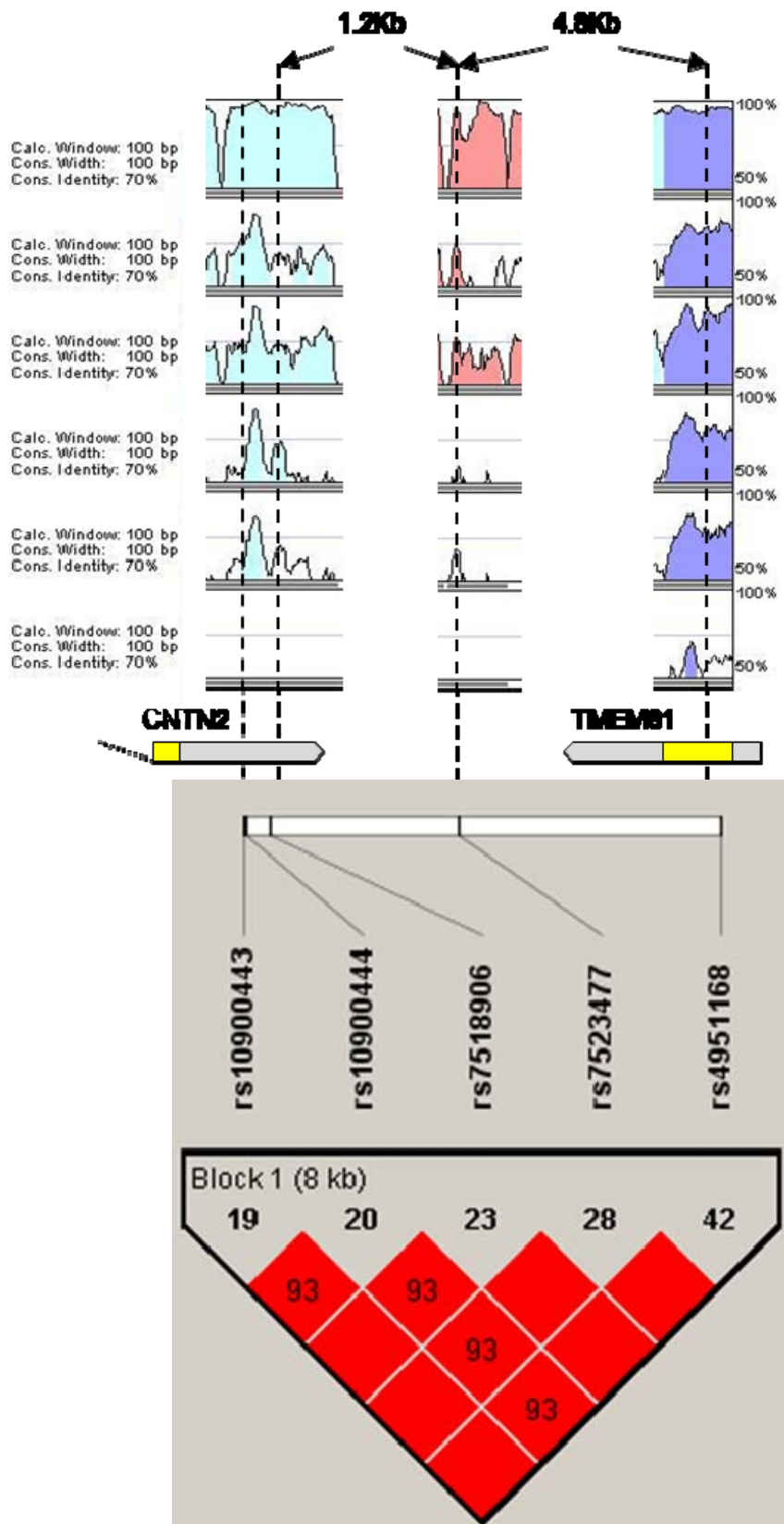
Here we have illustrated an LD-aware approach to allow meta-analysis of summary data generated from multiple GWAS datasets. This is valuable because a large portion of GWAS to date have been underpowered to detect genuine SNP associations with common diseases – which may in part explain the missing heritability. 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 ( $r^2 = 1$ ) 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

## Alzheimer's disease biomarkers

**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*).

Name	Gene		Reiman et al. 2007 [9]		Li et al. 2008 [14]		Carrasquillo et al. 2009 [4]		Distance between correlated SNPs (bp)	Combined p-value / Corrected p-value	Random Effects Meta-analysis	
	# of SNPs	# of SNP Clumps ( $r^2=1$ )	SNP	p-value	SNP	p-value	SNP	p-value			p-Value	Odds Ratio (95% C.I.)
<i>CNTN1</i>	235	111	-	-	-	-	-	-	-	-	-	1.23 (1.01 – 1.49)
<i>CNTN2</i>	407	178	rs4951168	0.057	rs4951168	0.339	rs7523477	0.0005	4836	3x10 <sup>-5</sup> / 0.005	<b>0.037</b>	1.18 (0.86 – 1.60)
			No proxy	-	rs10900451	0.170	rs4950982	0.001	28035	1.84x10 <sup>-4</sup> /0.033	0.300	
<i>FGA/FGB/FGG</i>	56	26	-	-	-	-	-	-	-	-	-	
<i>SPARCL1</i>	45	24	-	-	-	-	-	-	-	-	-	
<i>APP</i>	243	108	-	-	-	-	-	-	-	-	-	
<i>MAPT</i>	84	21	rs17651507	0.195	rs17651507	0.258	rs8079215	0.002	5840	4.6x10 <sup>-4</sup> /0.009	0.820	1.02 (0.83-1.27)
<i>APOE</i>	6	6	rs4420638	4.55x10 <sup>-29</sup>	rs4420638	2.26x10 <sup>-44</sup>	-	-	0	1.03x10 <sup>-72</sup> /9.24x10 <sup>-33</sup>	<b>&lt;0.0001</b>	3.36 (2.93 – 3.85)

SNP p-values from three studies; Reiman et al.[9], Li et al.[14] and Carrasquillo et al.[4] were clustered according to LD ( $r^2 = 1$ ) using PLINK. Within each cluster, one SNP from each study was subjected to a Fisher's combined Test. Only the four SNP clumps which withstood correction for multiple testing (corr- $P = 0.05/\#$  of clumps) are displayed. These four SNPs also underwent random-effects meta-analysis to address consistency in the direction of the effect. One cluster within *CNTN2* LD block (containing rs7523477 and rs4951168,) and one within *APOE* (containing rs4420638,) is supported by this analysis.



**Figure 1.** 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  $\geq 70\%$  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 A $\beta$  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 signifi-

cant result following meta-analysis as would be expected. This is the most replicable association seen in LOAD GWAS, and is observed due to linkage disequilibrium with the *APOE*  $\epsilon$ 4 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.

## 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 Carrasquillo 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 John Van Geest Cancer Research Centre, School of Science and Technology, Nottingham Trent University, Nottingham, UK. E-mail: kevin.morgan@nottingham.ac.uk

## References

- [1] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttacher 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 and Visscher PM. Finding the missing heritability of complex diseases. *Nature* 2009; 461: 747-753.
- [2] Bodmer W and Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 2008; 40: 695-701.
- [3] 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 JS, 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 KH, Klopp N, Wichmann HE, Carrasquillo MM, Pankratz VS, Younkin SG, Holmans PA, O'Donovan M, Owen MJ and Williams J. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nat Genet* 2009; 41: 1088-1093.
- [4] Carrasquillo MM, Zou F, Pankratz VS, Wilcox SL, Ma L, Walker LP, Younkin SG, Younkin CS, Younkin LH, Bisceglia GD, Ertekin-Taner N, Crook JE, Dickson DW, Petersen RC, Graff-Radford NR and Younkin SG. Genetic variation in PCDH11X is associated with susceptibility to late-onset Alzheimer's disease. *Nature Genetics* 2009; 41: 192-198.
- [5] Lambert JC, 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, Licastrò F, Soininen H, Ritchie K, Blanche H, Dartigues JF, Tzourio C, Gut I, Van Broeckhoven C, Alperovitch A, Lathrop M and Amouyel P. Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat Genet* 2009; 41: 1094-1099.
- [6] Abraham R, Moskvina V, Sims R, Hollingworth P, Morgan A, Georgieva L, Dowzell K, Cichon S, Hillmer AM, O'Donovan MC, Williams J, Owen MJ and Kirov G. A genome-wide association study for late-onset Alzheimer's disease using DNA pooling. *BMC Med Genomics* 2008; 1: 44.
- [7] Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, Hogan MF, Schjeide BM, Hooli B, Divito J, Ionita I, Jiang H, Laird N, Moscarillo T, Ohlsen KL, Elliott K, Wang X, Hu-Lince D, Ryder M, Murphy A, Wagner SL, Blacker D, Becker KD and Tanzi RE. Genome-wide Association Analysis Reveals Putative Alzheimer's Disease Susceptibility Loci in Addition to APOE. *Am J Hum Genet* 2008;
- [8] Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, Lince DH, Zismann VL, Beach TG, Leung D, Bryden L, Halperin RF, Marlowe L, Kaleem M, Walker DG, Ravid R, Heward CB, Rogers J, Papassotiropoulos A, Reiman EM, Hardy J, Stephan DA, Coon KD, Myers AJ, Craig DW, Webster JA, Pearson JV, Lince DH, Zismann VL, Beach TG, Leung D, Bryden L, Halperin RF, Marlowe L, Kaleem M, Walker DG, Ravid R, Heward CB, Rogers J, Papassotiropoulos A, Reiman EM, Hardy J and Stephan DA. A high-density whole-genome association study reveals that APOE is the major susceptibility gene for sporadic late-onset Alzheimer's disease.[see comment]. *Journal of Clinical Psychiatry* 2007; 68: 613-618.
- [9] Reiman EM, Webster JA, Myers AJ, Hardy J, Dunckley T, Zismann VL, Joshupura KD, Pearson JV, Hu-Lince D, Huentelman MJ, Craig DW, Coon KD, Liang WS, Herbert RH, Beach T, Rohrer KC, Zhao AS, Leung D, Bryden L, Marlowe L, Kaleem M, Mastroeni D, Grover A, Heward CB, Ravid R, Rogers J, Hutton ML, Melquist S, Petersen RC, Alexander GE, Caselli RJ, Kukull W, Papassotiropoulos A and Stephan DA. GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers. *Neu-*

- ron 2007; 54: 713-720.
- [10] Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, Morgan A, Jehu L, Segurado R, Stone D, Schadt E, Karnoub M, Nowotny P, Tacey K, Catanese J, Sninsky J, Brayne C, Rubinsztein D, Gill M, Lawlor B, Lovestone S, Holmans P, O'Donovan M, Morris JC, Thal L, Goate A, Owen MJ, Williams J, Grupe A, Abraham R, Li Y, Rowland C, Hollingworth P, Morgan A, Jehu L, Segurado R, Stone D, Schadt E, Karnoub M, Nowotny P, Tacey K, Catanese J, Sninsky J, Brayne C, Rubinsztein D, Gill M, Lawlor B, Lovestone S, Holmans P, O'Donovan M, Morris JC, Thal L, Goate A, Owen MJ and Williams J. Evidence for novel susceptibility genes for late-onset Alzheimer's disease from a genome-wide association study of putative functional variants. *Human Molecular Genetics* 2007; 16: 865-873.
- [11] Poduslo SE, Huang R, Huang J and Smith S. Genome screen of late-onset Alzheimer's extended pedigrees identifies TRPC4AP by haplotype analysis. *Am J Med Genet B Neuropsychiatr Genet* 2009; 150B: 50-55.
- [12] Potkin SG, Guffanti G, Lakatos A, Turner JA, Kruggel F, Fallon JH, Saykin AJ, Orro A, Lupoli S, Salvi E, Weiner M and Macciardi F. Hippocampal atrophy as a quantitative trait in a genome-wide association study identifying novel susceptibility genes for Alzheimer's disease. *PLoS One* 2009; 4: e6501.
- [13] Lee JH, Cheng R, Graff-Radford N, Foroud T and Mayeux R. Analyses of the National Institute on Aging Late-Onset Alzheimer's Disease Family Study: implication of additional loci. *Arch Neurol* 2008; 65: 1518-1526.
- [14] Li H, Wetten S, Li L, St Jean PL, Upmanyu R, Surh L, Hosford D, Barnes MR, Briley JD, Borrie M, Coletta N, Delisle R, Dhallia D, Ehm MG, Feldman HH, Fornazzari L, Gauthier S, Goodgame N, Guzman D, Hammond S, Hollingworth P, Hsiung G-Y, Johnson J, Kelly DD, Keren R, Kertesz A, King KS, Lovestone S, Loy-English I, Matthews PM, Owen MJ, Plumpton M, Pryse-Phillips W, Prinjha RK, Richardson JC, Saunders A, Slater AJ, St George-Hyslop PH, Stinnett SW, Swartz JE, Taylor RL, Wherrett J, Williams J, Yarnall DP, Gibson RA, Irizarry MC, Middleton LT and Roses AD. Candidate single-nucleotide polymorphisms from a genomewide association study of Alzheimer disease. *Archives of Neurology* 2008; 65: 45-53.
- [15] Yin GN, Lee HW, Cho JY and Suk K. Neuronal pentraxin receptor in cerebrospinal fluid as a potential biomarker for neurodegenerative diseases. *Brain Res* 2009; 1265: 158-170.
- [16] Bizzoca A, Virgintino D, Lorusso L, Buttiglione M, Yoshida L, Polizzi A, Tattoli M, Cagiano R, Rossi F, Kozlov S, Furley A and Gennarini G. Transgenic mice expressing F3/contactin from the TAG-1 promoter exhibit developmentally regulated changes in the differentiation of cerebellar neurons. *Development* 2003; 130: 29-43.
- [17] Berglund EO, Murai KK, Fredette B, Sekerkova G, Marturano B, Weber L, Mugnaini E and Ranscht B. Ataxia and abnormal cerebellar microorganization in mice with ablated contactin gene expression. *Neuron* 1999; 24: 739-750.
- [18] Berglund EO and Ranscht B. Molecular cloning and in situ localization of the human contactin gene (CNTN1) on chromosome 12q11-q12. *Genomics* 1994; 21: 571-582.
- [19] Compton AG, Albrecht DE, Seto JT, Cooper ST, Ilkovski B, Jones KJ, Challis D, Mowat D, Ranscht B, Bahlo M, Froehner SC and North KN. Mutations in contactin-1, a neural adhesion and neuromuscular junction protein, cause a familial form of lethal congenital myopathy. *Am J Hum Genet* 2008; 83: 714-724.
- [20] Falk J, Bonnon C, Girault J-A and Faivre-Sarrailh C. F3/contactin, a neuronal cell adhesion molecule implicated in axogenesis and myelination. *Biology of the Cell* 2002; 94: 327-334.
- [21] Girard JP and Springer TA. Cloning from purified high endothelial venule cells of hevin, a close relative of the antiadhesive extracellular matrix protein SPARC. *Immunity* 1995; 2: 113-123.
- [22] Lively S and Brown IR. Localization of the extracellular matrix protein SC1 coincides with synaptogenesis during rat postnatal development. *Neurochem Res* 2008; 33: 1692-1700.
- [23] Weimer JM, Stanco A, Cheng JG, Vargo AC, Voora S and Anton ES. A BAC transgenic mouse model to analyze the function of astroglial SPARCL1 (SC1) in the central nervous system. *Glia* 2008; 56: 935-941.
- [24] Bai Y, Markham K, Chen F, Weerasekera R, Watts J, Horne P, Wakutani Y, Bagshaw R, Mathews PM, Fraser PE, Westaway D, St George-Hyslop P, Schmitt-Ulms G, Bai Y, Markham K, Chen F, Weerasekera R, Watts J, Horne P, Wakutani Y, Bagshaw R, Mathews PM, Fraser PE, Westaway D, St George-Hyslop P and Schmitt-Ulms G. The in vivo brain interactome of the amyloid precursor protein. *Molecular & Cellular Proteomics* 2008; 7: 15-34.
- [25] Hardy JA and Higgins GA. Alzheimer's disease: the amyloid cascade hypothesis. *Science* 1992; 256: 184-185.
- [26] Mattson MP. Cellular actions of beta-amyloid precursor protein and its soluble and fibrillogenic derivatives. *Physiol Rev* 1997; 77: 1081-1132.
- [27] Ma Q-H, Futagawa T, Yang W-L, Jiang X-D, Zeng L, Takeda Y, Xu R-X, Bagnard D, Schachner M, Furley AJ, Karagogeos D, Watanabe K, Dawe GS and Xiao Z-C. A TAG1-APP signalling pathway through Fe65 negatively modulates neurogenesis.[see comment]. *Nature Cell Biology* 2008; 10: 283-294.
- [28] Mattson MP and van Praag H. TAGing APP constrains neurogenesis.[comment]. *Nature Cell Biology* 2008; 10: 249-250.
- [29] van Oijen M, Witteman JC, Hofman A, Koudstaal PJ and Breteler MM. Fibrinogen is associated

- with an increased risk of Alzheimer disease and vascular dementia. *Stroke* 2005; 36: 2637-2641.
- [30] Paul J, Strickland S and Melchor JP. Fibrin deposition accelerates neurovascular damage and neuroinflammation in mouse models of Alzheimer's disease. *J Exp Med* 2007; 204: 1999-2008.
- [31] Liao L, Cheng D, Wang J, Duong DM, Losik TG, Gearing M, Rees HD, Lah JJ, Levey AI and Peng J. Proteomic characterization of postmortem amyloid plaques isolated by laser capture microdissection. *Journal of Biological Chemistry* 2004; 279: 37061-37068.
- [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.
- [33] Barrett JC, Fry B, Maller J and Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 2005; 21: 263-265.
- [34] Dubchak I and Ryaboy DV. VISTA family of computational tools for comparative analysis of DNA sequences and whole genomes. *Methods Mol Biol* 2006; 338: 69-89.
- [35] Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P, Rohrer K, Zhao A, Marlowe L, Kaleem M, McCorquodale DS, 3rd, Cuellar C, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV, Heward CB, Reiman EM, Stephan D, Hardy J and Myers AJ. Genetic control of human brain transcript expression in Alzheimer disease. *Am J Hum Genet* 2009; 84: 445-458.
- [36] Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR and Cookson WO. A genome-wide association study of global gene expression. *Nat Genet* 2007; 39: 1202-1207.