Journal of Human Genetics (2010), 1–3

www.nature.com/jhg

SHORT COMMUNICATION

Genome-wide pathway analysis implicates intracellular transmembrane protein transport in Alzheimer disease

Mun-Gwan Hong[1], Andrey Alexeyenko[1], Jean-Charles Lambert[2,3,4], Philippe Amouyel[2,3,4] and Jonathan A Prince[1]

We developed and implemented software for the analysis of genome-wide association studies in the context of biological pathway enrichment and have here applied our algorithm to the study of Alzheimer disease (AD). Using genome-wide association data in a large French population, we observed a highly significant enrichment of genes involved in intracellular protein transmembrane transport, including several mitochondrial proteins and nucleoporins. An intriguing aspect of these findings is the implication that *TOMM40*, the channel-forming subunit of the translocase of the mitochondrial outer membrane complex, and a gene generally considered to be indiscernible from *APOE* because of linkage disequilibrium, may itself contribute to Alzheimer pathology. Results provide an indication that protein trafficking, in particular across the nuclear and mitochondrial membranes, may contribute to risk for AD.
Journal of Human Genetics advance online publication, 29 July 2010; doi:10.1038/jhg.2010.92

Keywords: Alzheimer; genome-wide; mitochondria; pathway; TOMM40

Genome-wide association studies are now abundant with hundreds of newly identified single loci being shown with a high degree of probability to influence a variety of traits and diseases.[1] However, for almost all tested traits only 1–2 genes are typically identified that survive correction for multiple testing in a genome-wide context, leaving the question open as to whether additional risk genes exist. An emerging approach to understanding these studies in a larger biological context is to explore the upper distribution of the most significant genes for an enrichment of certain classes of function. Because of the depth of annotation of the genome, the preferred way to do this is by means of the gene ontology (GO).[2] This is a relatively new approach, stemming from the application of GO-based analyses to gene expression data,[3] but despite promise only a handful of replicated cases of pathway enrichment have emerged.[4,5] One of the critical issues in enabling this strategy is to convert with high fidelity the single-nucleotide polymorphism (SNP) lists from genome-wide platforms to the list of the genes they represent. Toward this end, we developed a software program implemented in Perl, using as input genome-wide SNP results (primarily from PLINK[6]), that considers linkage disequilibrium (LD) across regions of significance that corrects for the inflation of significance due to gene length.[5] In brief, our software automates the process of converting genome-wide SNP lists to gene lists, beginning with the retrieval of LD structures in analogous populations with denser genotyping data (that is, HapMap). When a group of markers are in high LD in HapMap (we use an $r^2 > 0.8$ threshold), they are tied to a 'proxy cluster' treating it as a single signal. Subsequently, each marker in the original SNP list with statistically significant evidence of association with a phenotype is evaluated to see (a) if it belongs to any proxy cluster and (b) if the marker itself or any marker in the cluster is located in a genic region. Any marker or cluster that overlaps a region extending across a gene is assigned as a signal indicating the possible association of that gene. To correct the multiple-testing problem that emerges due to multiple signals across a gene, the *P*-value for each gene is adjusted by multiplication of the lowest *P*-value of the assigned signals by the number of signals. An illustration of the algorithm can be found in our earlier paper.[5] Here, we have applied this program to a genome-wide association study in a French Alzheimer disease (AD) case–control sample.

The genome-wide association study included 2032 AD cases and 5328 controls of French ancestry and was conducted on the Illumina 610 platform (Illumina, San Diego, CA, USA).[7] Appropriate institutional review board permission was obtained for this study(see Lambert *et al.*[7] for details). A total of 511 978 SNPs that passed quality control (genotypes were excluded that had call rates $<98\%$, a minor allele frequency of 1% or less, or significant deviation from Hardy–Weinberg equilibrium at $P<10^{-6}$) were parsed and converted into a list of 16 503 genes using our algorithm. We note that the maximum significance ($P=2.3\times10^{-130}$) obtained overlapped with the *TOMM40* gene, near *APOE*. Also notable is that within this set there are no genes, save around the *APOE* locus that show genome-wide significance. The resultant list of genes, the marker with highest significance

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [2]Inserm U744, Lille, France; [3]Institut Pasteur de Lille, Lille, France and [4]Université de Lille Nord de France, Lille, France
Correspondence: Dr JA Prince, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12A., Stockholm 171 77, Sweden.
E-mail: Jonathan.Prince@ki.se
Received 9 April 2010; revised 31 May 2010; accepted 30 June 2010

that is assigned to that particular gene, the number of genetic markers used for gene-based correction and a list of genes indiscernible due to LD is provided as Supplementary Table 1. For enrichment analysis we used our software together with the public domain tools provided by both the DAVID bioinformatics platform[8] and Genecodis.[9] After adjustment for gene length, there were 1351 genes that were assigned a *P*-value of 0.05 or less and these were tested for enrichment against the study base set of 16 503 genes. Importantly, testing the top genes against a default full genome base set gives an anticipated highly significant (and incorrect) enrichment of multiple high level GO categories, emphasizing the importance of using the gene lists that are actually represented on, for example, the Illumina 610 platform.

In this genome-wide data set, we observed a highly significant enrichment of genes annotated as being involved in the biological process of intracellular transmembrane protein transport (GO:0065002, $P=7.2\times10^{-6}$ based on a hypergeometric test, $P<0.001$ based on 1000 permutations). Both Genecodis and DAVID provided equivalent results (the *P*-value for this pathway with DAVID was slightly lower at 5.2×10^{-6}). There were 18 genes that contributed to this significance and we show those specific genes, as well as the best genetic marker and its associated *P*-value in Table 1. Both DAVID and Genecodis use a hypergeometric test for significance estimation, and taking the Genecodis example, significance was derived from 18 of 1331 genes in the enriched list being association with the protein transport term, versus 69 in the total of 16 283 annotated genes. We note that the genes contributing to the signal for protein transport are dispersed widely in terms of individual significance across the top 1351 genes, emphasizing the possible existence of true association signals beyond only the first few most significant genes. A common problem with analyses of this nature is the false appearance of enrichment due to chromosomal clustering of functionally related genes.[5] For this particular analysis, all genes contributing to enrichment were located in distinct genomic loci (also shown in Table 1), with the closest genes being several megabases apart. However, there were also a few cases of ontology categories that could be dismissed because of positional clustering, the most prominent being 'cytokine activity' due to an enrichment of interferon genes that are located in tight genomic proximity (not shown).

To understand in more detail the relationships among the genes contributing to the protein transport signal, we used FunCoup,[10] which enables connections to be visualized based on genomics and experimental data, such as protein–protein interaction and gene expression correlations. We were particularly interested in how the identified protein transport pathway might be related to the *APOE* locus, which contains four genes that cannot be readily discerned due to LD (*APOE*, *TOMM40*, *PVRL2* and *BCL3*). We therefore tested these 4 genes in turn for network connectivity to the 18 genes identified by enrichment analysis. To evaluate statistical significance we developed our own custom algorithm based on a previously described randomization strategy.[11] The randomized network was thus re-wired in such a way that the number of links for each node was preserved, although its network neighbors were shuffled. The real (that is, FunCoup-predicted) network was randomized 100 times. In FunCoup, each link is characterized by a confidence value termed as final Bayesian score—a sum of individual log likelihood ratios of the integrated data sets (51 sets from 7 eukaryotes) that informed on functional coupling. For the analysis, we selected network edges with final Bayesian score 4.8 (natural logarithm), that defined a network of 14 899 genes connected with 709 343 links. After every randomization, connections between a gene of interest and a gene group were counted. These values were used to calculate the mean and s.d. Together with the respective number of links in the real network, these values produced one-sided *Z*-scores that estimated significance. In this analysis, only *TOMM40* was strongly connected to the set of 18 protein transport genes, with 4 direct and 792 indirect; that is, through a third gene, links ($P<10^{-4}$ and $P<10^{-7}$, respectively). *BCL3* had a single much weaker link (based on subcellular colocalization) to *NUP88*—but this was not significant. From Figure 1 there is a clear division into two groupings, one containing members of the nucleoporin gene family and the other consisting of mitochondrial genes. These two groups are connected, albeit weakly, by interactions between *NUP98*, *TIMM44* and *TIMM17A*. In the final network

Table 1 Enriched genes in Alzheimer disease involved in intracellular transmembrane protein transport

Gene	*Best marker*	P-*value*	*Position*	*Gene description*
NUP98	rs276885	6.68×10^{-5}	Chr11: 3.6 MB	Nucleoporin 98 kDa
NUPL2	rs858238	2.00×10^{-4}	Chr7: 23.1 MB	Nucleoporin like 2
TRAM2	rs6928665	4.00×10^{-4}	Chr6: 52.4 MB	Translocation associated membrane protein 2
NUP88	rs6502860	1.00×10^{-3}	Chr17: 5.2 MB	Nucleoporin 88 kDa
NUP160	rs7951180	1.20×10^{-3}	Chr11: 47.7 MB	Nucleoporin 160 kDa
NUP37	rs950945	2.70×10^{-3}	Chr12: 100.9 MB	Nucleoporin 37 kDa
TNKS	rs6601327	2.90×10^{-3}	Chr8: 9.4 MB	Tankyrase, TRF1-interacting ankyrin-related ADP-ribose polymerase
TRAM1	rs268652	2.90×10^{-3}	Chr8: 71.6 MB	Translocation associated membrane protein 1
NUP205	rs11984203	5.00×10^{-3}	Chr7: 134.8 MB	Nucleoporin 205 kDa
DDX19A	rs8059245	5.40×10^{-3}	Chr16: 68.9 MB	DEAD (Asp-Glu-Ala-As) box polypeptide 19A
CHCHD4	rs4685078	5.50×10^{-3}	Chr3: 14.1 MB	Coiled–coil–helix–coiled–coil–helix domain containing four
NUP133	rs927204	5.70×10^{-3}	Chr1: 227.6 MB	Nucleoporin 133 kDa
C18orf55	rs17062282	9.00×10^{-3}	Chr18: 69.9 MB	Chromosome 18 open reading frame 55
Magmas	rs611704	1.10×10^{-2}	Chr16: 4.3 MB	Mitochondrial protein, granulocyte-macrophage colony-stimulating factor signal transduction
FXC1	rs4758423	1.10×10^{-2}	Chr11: 6.4 MB	Fracture callus one homolog (rat)
TIMM44	rs12983784	1.20×10^{-2}	Chr19: 7.8 MB	Translocase of inner mitochondrial membrane 44 homolog (yeast)
TMEM48	rs1181145	1.40×10^{-2}	Chr1: 54.0 MB	Transmembrane protein 48
TIMM17A	rs2820306	1.60×10^{-2}	Chr1: 200.1 MB	Translocase of inner mitochondrial membrane 17 homolog A (yeast)

Abbreviation: LD, linkage disequilibrium.
Genes enriched in a GWAS of Alzheimer disease according to the LD-adjusted *P*-value of the most significant marker associated with each gene. The presented *P*-value is the unadjusted significance of the specified marker in the original data set.

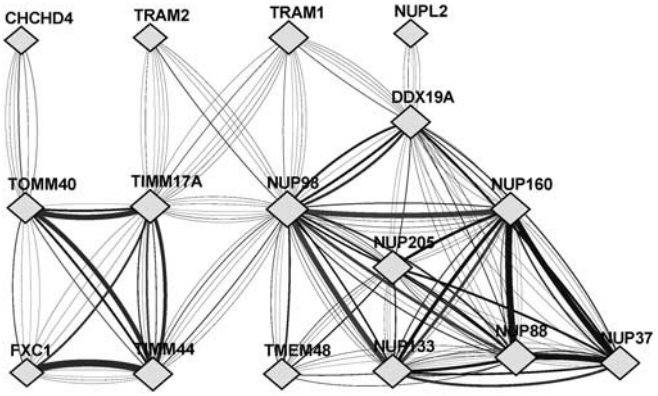

Figure 1 A network connectivity map of genes identified in the intracellular transmembrane protein transport pathway from a GWAS of Alzheimer disease. A total of 18 genes were identified, consisting of both mitochondrial and nuclear pore genes. TOMM40 is included as it connects tightly to TIMM17A and TIMM44, illustrating its importance in this pathway. No other genes from the *BCL3-PVRL2-TOMM40-APOE LD* block are significantly associated with this network. The lines (edges) between genes denote the strength and origin of connectivity and are derived primarily from mRNA coexpression, protein–protein interaction, sub-cellular colocalization and phylogenetic similarity (thicker=stronger).

(Figure 1) most of the original 18 genes are represented, with 3 (*Magmas*, *TNKS* and *C18orf55*) not having any significant connections with the remaining 15 genes. Notably, *Magmas* and *C18orf55* are mitochondrial genes, whereas *TNKS* is a nuclear pore protein.

We used gene expression data to explore the relationships of *TOMM40* and *APOE* to the known base set of genes that have been confirmed to lead to AD (*PSEN1*, *PSEN2* and *APP*). For this, a human brain sample was used with gene expression level estimates for 14 077 transcripts in 193 individuals.[12] In testing *TOMM40* and *APOE* against the base set, we observed very strong correlation of *TOMM40* to *PSEN2* ($P=1.3\times10^{-13}$, $r^2=0.24$) and a weaker association of *APOE* to *APP* ($P=4.5\times10^{-7}$, $r^2=0.12$). The other correlations were not significant at $\alpha=0.05$.

This study marks one of the first attempts to explore genome-wide association data in AD in the context of pathway enrichment. The enriched pathway that we have uncovered provides an intriguing indication that dysfunction of intracellular protein trafficking may be a common biological theme in AD. Although there is little support in the literature for the involvement of nucleoporin genes in AD, there is more substantial evidence for the importance of the mitochondria. In this regard, recent evidence suggests that import of β-amyloid into mitochondria may underlie β-amyloid toxicity,[13,14] in line with a larger body of evidence linking mitochondrial function to AD.[15] More importantly this process is facilitated by the translocase of the mitochondrial outer membrane complex, illustrating the potential importance of *TOMM40*, itself the highest ranked gene in this GWAS and the only gene in the BCL3-*PVRL2-TOMM40-APOE* LD block that is significantly connected to the identified pathway. It may be plausible that age-related susceptibility to β-amyloid might be mediated by a decrease in mitochondrial function that occurs with advancing age.[16,17] Import of β-amyloid into the nucleus through nucleoporins may also be an avenue worth pursuing in functional studies. Although *TOMM40* shows pathway connectivity, whereas *APOE* does not, we emphasize that we in no way make the claim that the association of the region to AD is mediated by *TOMM40*. Rather, the data indicate that *TOMM40* may also have a role in the disease, and this is echoed in the strong correlation of *TOMM40* to *PSEN2* expression. In summary, our approach rests on the idea that the genetic architecture of complex traits is not dispersed over unrelated genes in the genome, but rather the mutational events that ultimately underlie trait variance can occur in functionally related genes. While implicating intracellular protein transport in AD is a highlight of the present study, we also consider the success of identifying a significant pathway component to a complex disease an important validation of this strategy.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Swedish Medical Research Council (Grant 2007-2722).

1 McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9,** 356–369 (2008).
2 Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).
3 Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34,** 267–273 (2003).
4 Srinivasan, B. S., Doostzadeh, J., Absalan, F., Mohandessi, S., Jalili, R., Bigdeli, S. *et al.* Whole genome survey of coding SNPs reveals a reproducible pathway determinant of Parkinson disease. *Hum. Mutat.* **30,** 228–238 (2009).
5 Hong, M. G., Pawitan, Y., Magnusson, P. K. & Prince, J. A. Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum. Genet.* **126,** 289–301 (2009).
6 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
7 Lambert, J. C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M. *et al.* Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. *Nat. Genet.* **41,** 1094–1099 (2009).
8 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4,** 44–57 (2009).
9 Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C., Yang, X., Tirado, F. *et al.* GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res.* **37,** W317–322 (2009).
10 Alexeyenko, A. & Sonnhammer, E. L. Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Res.* **19,** 1107–1116 (2009).
11 Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296,** 910–913 (2002).
12 Myers, A. J., Gibbs, J. R., Webster, J. A., Rohrer, K., Zhao, A., Marlowe, L. *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* **39,** 1494–1499 (2007).
13 Anandatheerthavarada, H. K., Biswas, G., Robin, M. A. & Avadhani, N. G. Mitochondrial targeting and a novel transmembrane arrest of Alzheimer's amyloid precursor protein impairs mitochondrial function in neuronal cells. *J. Cell Biol.* **161,** 41–54 (2003).
14 Hansson Petersen, C. A., Alikhani, N., Behbahani, H., Wiehager, B., Pavlov, P. F., Alafuzoff, I. *et al.* The amyloid beta-peptide is imported into mitochondria via the TOM import machinery and localized to mitochondrial cristae. *Proc. Natl Acad. Sci. USA* **105,** 13145–13150 (2008).
15 Moreira, P. I., Duarte, A. I., Santos, M. S., Rego, A. C. & Oliveira, C. R. An integrative view of the role of oxidative stress, mitochondria and insulin in Alzheimer's disease. *J. Alzheimers Dis.* **16,** 741–761 (2009).
16 Balaban, R. S., Nemoto, S. & Finkel, T. Mitochondria, oxidants, and aging. *Cell* **120,** 483–495 (2005).
17 Hong, M. G., Myers, A. J., Magnusson, P. K. & Prince, J. A. Transcriptome-wide assessment of human brain and lymphocyte senescence. *PLoS One* **3,** e3024 (2008).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)