

Finding disease candidate genes by liquid association

Ker-Chau Li^{*†}, Aarno Palotie^{‡§¶}, Shinsheng Yuan[†], Denis Bronnikov^{§#},
Daniel Chen[§], Xuelian Wei^{*}, Oi-Wa Choi[§], Janna Saarela[#] and
Leena Peltonen^{#**}

Addresses: ^{*}Department of Statistics, UCLA, 8125 Math Sciences Bldg, Los Angeles, California 90095-1554, USA. [†]Institute of Statistical Science, Academia Sinica, Academia Road, Nankang, Taipei 115, Taiwan. [‡]The Finnish Genome Center and Department of Clinical Chemistry, University of Helsinki, Haartmaninkatu, 00290 Helsinki, Finland. [§]The Broad Institute of Harvard and MIT, Cambridge Center, Cambridge, Massachusetts 02142, USA. [¶]Department of Pathology and Laboratory Medicine, Gonda Research Center, UCLA, Los Angeles, California 90095-1766, USA. [¶]Department of Human Genetics, UCLA, 695 Charles E. Young Drive South, Los Angeles, California 90095-1766, USA. [#]National Public Health Institute, Helsinki, Finland, Biomedicum Helsinki, Haartmaninkatu, 00290 Helsinki, Finland. ^{**}Department of Medical Genetics, University of Helsinki, Biomedicum Helsinki, Haartmaninkatu, 00290 Helsinki, Finland.

Correspondence: Ker-Chau Li. Email: kcli@stat.ucla.edu. Leena Peltonen. Email: leena.peltonen@ktl.fi

Published: 4 October 2007

Genome **Biology** 2007, **8**:R205 (doi:10.1186/gb-2007-8-10-r205)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2007/8/10/R205>

Received: 16 April 2007

Revised: 23 August 2007

Accepted: 4 October 2007

© 2007 Li *et al*; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

A novel approach to finding candidate genes by using gene expression data through liquid association is developed and used to identify multiple sclerosis susceptibility candidate genes.

Background

Studies aiming to identify susceptibility genes in complex diseases have proceeded along two lines. The traditional candidate gene approach is limited by our ability to come up with a comprehensive list of biologically related genes. On the other hand, the 'hypothesis free' approach relies on genome-wide scans for disease loci, typically via linkage in exceptionally large families or via association in case control studies. Multiple sclerosis (MS), which is one of the most common neurologic disorders affecting young adults, is characterized by demyelination and reactive gliosis [1]. Analogous to many complex traits, genome scans in MS have identified numerous chromosomal loci often with only a nominal evidence for linkage to MS [2-6]. With the notable exception of the human leukocyte antigen (major histocompatibility complex [MHC]) locus on 6p21, evidence for specific MS genes emerging from these studies is still scanty. Thus far, the only associated non-HLA genes replicated in multiple populations are the *PRKCA* gene [7] and the recently reported *IL2RA* and *IL7R* genes [8]. For MS, as for most complex traits, the loci derived from link-

age scans have remained quite wide because of multiple uncertainties concerning the disease model in statistical analyses. To expedite the process of gene identification in these wide DNA regions, we need novel approaches to identify potentially involved pathways and to prioritize genes on identified loci for further sequencing efforts.

Our idea is to turn to full genome functional studies for these goals. As illustrated in Figure 1, our approach takes advantage of the availability of abundant microarray data and a wealth of genomic/proteomic knowledge base from the public domain. Our intention is to integrate information from both the candidate gene and the full genome scan (thus far mostly family-based linkage) approaches. In this report we use two previously reported MS susceptibility genes, identified in the same study sample [7,9], namely *MBP* and *PRKCA*, as the lead to probe microarray gene expression data for functionally associated genes. High score genes, identified by statistical data analysis, are followed up by an extensive literature search for their biologic relevance.

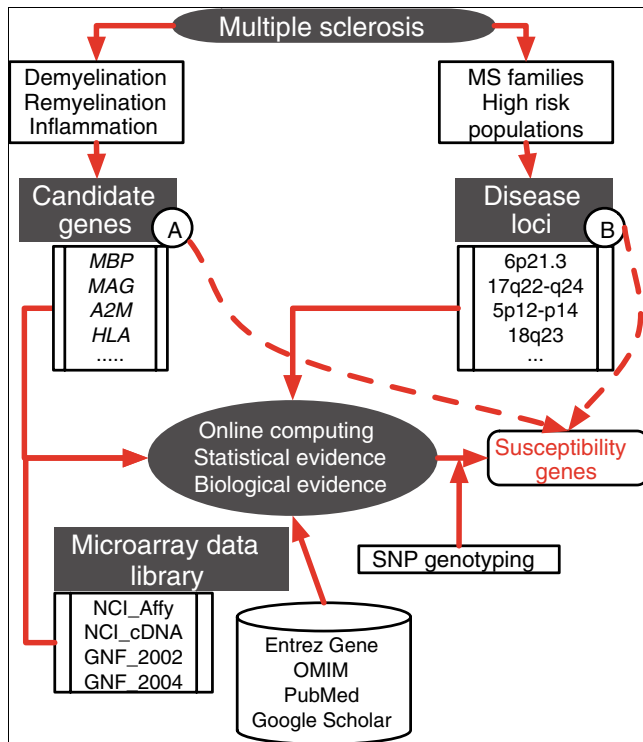


Figure 1
 Federated functional genomics approach. The two dashed lines in this diagram indicate (a) the candidate gene approach and (b) the full-genome scan approach to finding susceptibility genes. Information from both approaches is used to guide the functional genomic study on multiple sets of microarray gene expression databases. This approach is powered by online statistical computation and a biomedical literature search.

Four large expression datasets are employed in this study (see Materials and methods, below). The first two, namely NCI_cDNA and NCI_Affy, are expression profiles for US National Cancer Institute (NCI)'s 60 human cancer cell lines reported by two different research teams [10,11]. The other two databases, GCN_2002 and GCN_2004, provide expression profiles for a diverse array of human tissues [12,13]. Together, they offer a glimpse into transcript regulation under a wide spectrum of physiologic conditions.

In addition to the conventional similarity study, we utilized a new computational tool, termed liquid association (LA) [14-16]. The power of the LA method in identifying elements of biologic pathways has been demonstrated by its use to identify correctly genes that are involved in the urea cycle [14]. In conventional similarity analysis, we tend to rely on the correlation $\text{corr}(X,Y)$, which measures the degree of co-expression between two genes X and Y. Genes with high correlations are likely to be functionally associated. The encoded proteins may participate in the same pathway, form a common structural complex, or be regulated by the same mechanism. However, not all functionally associated genes are co-expressed; indeed, the majority of them are not. One conceivable reason for this is that gene expression can be sensitive to the often

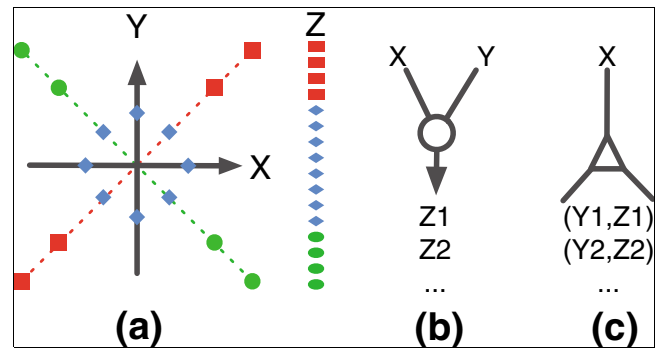


Figure 2
 Liquid association. (a) Association between genes X and Y as mediated by gene Z. When gene Z is expressed at the high level (red), a positive correlation between X and Y is observed. The association changes as the expression of Z is lowered. It eventually becomes a negative trend (green). There are two basic ways (shown in panels b and c) to apply the liquid association (LA) scoring system to guide a genome-wide search. (b) When two genes X and Y are given, compute LA score $LA(X, Y|Z)$ for every gene Z first and then output a short list of high score genes Z1, Z2, and so on. (c) When only one gene X is given, compute LA score $LA(X, Y|Z)$ for every pair of genes X,Y first and then output a short list of high score gene pairs Y1,Z1, Y2,Z2, and so on.

varying cellular state, such as presence or absence of hormones, metabolites, ion homeostasis, and so on. Two genes X and Y that are engaged in a common process under some conditions may disengage and embark on activities of their own as the cellular state changes. Consequently, two functionally related genes with a positive correlation in expression may become uncorrelated or even negatively correlated as the relevant state variable changes. If we could characterize the mediating state variable, then we might be able to detect the correlation by controlling the state variable.

Finding the mediating state variable is by no means simple. LA is a statistical device introduced for this purpose. The method is based on the assumption that the state variable is correlated with the expression of a third gene Z. If this is the case, then we may use Z to detect such a 'liquid' (as opposed to 'solid') pattern of statistical association between X and Y. Figure 2 illustrates how LA works. A liquid association score $LA(X, Y|Z)$ can be computed using a simple statistical formula given in [14]. There are two ways of applying LA. For a given pair of X and Y, one can look for genes that may mediate X,Y co-expression by computing the LA score $LA(X, Y|Z)$ for each gene Z in the genome and obtaining a genome-wide ranking. Alternatively, given one gene Z, we may ask which pairs of genes Z may mediate. With more computing effort, we can obtain $LA(X, Y|Z)$ for every pair of genes X,Y and rank their scores in order to identify the most significant pairs. We have constructed a website to facilitate online searching for genes of interest. (See Additional data file 2 [Supplementary Text 1] for an illustrative application to the Alzheimer's hallmark gene APP [amyloid-β precursor protein].)

Results

MBP-initiated genome-wide liquid association search identifies *A2M*

We started with the *MBP* gene (which encodes myelin basic protein, an integral element of the myelin sheath surrounding the neuronal extensions). This gene is critical in triggering the immune reaction in the demyelination process for experimental allergic encephalomyelitis (EAE), a rodent model of MS. Importantly, *MBP* has been implicated both in linkage and in association studies conducted in MS pedigrees of Scandinavian origin (see Haines and coworkers [4] for references).

We applied the second genome-wide LA search method (Figure 2c) to the NCI_cDNA database. By treating *MBP* as the query gene X, we evaluate the LA score for every pair of genes (Y,Z). Because there are 9,076 genes in this database, about 49 million LA scores are computed and compared with each other. The output of a short list of 25 gene pairs with the best LA scores each from the positive and the negative ends is given in Additional data file 1 (Table S1). The statistical significance of the results of this gene search procedure is discussed in Additional data file 2 (Supplementary Text 3). We find that the gene *A2M* (encoding α_2 -macroglobulin, a cytokine transporter and protease inhibitor) appears many times. We further find an interesting biologic functional association between *A2M* and *MBP* from some literature about the pathogenesis of MS. Following demyelination in human MS and rodent EAE, immunogenic *MBP* peptides are released into cerebrospinal fluid and serum (see Oksenberg and coworkers [2] for references) and *A2M* represents the major MBP-binding protein in human plasma [17]. A significant increase in α_2 -macroglobulin is found in plasma of MS patients [18]. Analogously, in rodent EAE, infusion of α_2 -macroglobulin significantly reduces disease symptoms [19].

Among the genes to which *A2M* is paired, three are found to have functional association with immunologic neurodegenerative diseases. *LYST* (lysosomal trafficking regulator, also known as *CHS1*) is the causal gene for Chediak-Higashi syndrome, an inherited immunodeficiency disease, and *CHM* (Rab escort protein 1) is responsible for an inherited human retinal blindness known as choroideremia. (For details, see Online Mendelian Inheritance in Man of the National Center for Biotechnology Information [20].) *TRIB2* (tribbles homolog 2) was identified as an autoantigen in autoimmune uveitis, a term encompassing a group of ocular inflammatory disorders with unknown causes [21]. Additionally, *MPDZ* (multiple PDZ domain protein) encodes a tight junction protein that is detected in noncompact regions of myelin, and it is thought to be required to maintain the cytoarchitecture of myelinating Schwann cells [22]. The biologic connection for other genes, many still of unknown function, is not clear. We compute the correlation between these genes and find that most of them have significant correlations. (See in Additional data file 2 [Supplementary Text 4] for more discussion.)

Four multiple sclerosis loci from the Finnish population and *PRKCA*

Four major loci linked to MS have been identified in Finnish families: *HLA* on 6p21, *MBP* on 18q, and loci on 17q22-24 and 5p14-p12 [23]. These loci have also been implicated in other MS study samples from more heterogeneous populations [24,25]. The large locus on 17q was further refined to a 3 megabase (Mb) region in the Finnish MS families [23]. However, little information is available in the literature concerning how various loci are related to each other biologically. Most recently, association of specific *PRKCA* alleles at 17q24 with MS both in Finnish and Canadian MS study samples has been reported [7]. Involvement of *PRKCA* in MS was also validated by an association reported in a UK population [26]. *PRKCA* encodes a regulator of immune response, making it a highly suitable candidate gene for MS. A potential functional link between the *MBP* and *PRKCA* genes was identified by Feng and coworkers [27], who showed that a golli product of the myelin basic protein gene (*MBP*) can serve as a negative regulator of signaling pathways in T lymphocytes, particularly the protein kinase C pathway.

MBP-*PRKCA*-initiated liquid association search identifies *SLC1A3*

To study the co-expression pattern between *MBP* and *PRKCA*, we took them as genes X and Y to explore the GNF_2002 database using our system. The gene with the greatest LA score was the gene *SLC1A3* (glial high affinity glutamate transporter, member 3; see Additional data file 1 [Table S2]). Interestingly, *SLC1A3* is located on 5p13.2 (36.6 to 36.7 Mb), within the previously identified MS locus on 5p [28], which is syntenic to the *EAE2* locus in mouse.

Test of the genetic relevance of *SLC1A3* to multiple sclerosis

We wished to test whether there is any genetic relevance of *SLC1A3* to MS. We selected five single nucleotide polymorphisms (SNPs) flanking the *SLC1A3* gene (Table 1) to be genotyped in our primary study set, consisting of 61 MS families from the high-risk region of Finland. The most 5' SNP, namely rs2562582, located within 2 kilobases from the initiation of the *SLC1A3* transcript, exhibited initial evidence for association with MS ($P = 0.005$) in the transmission disequilibrium test (TDT) analysis, suggesting a possible functional role for this variant in the transcriptional regulation of this gene. Moreover, as shown in Table 1, stratification of the Finnish MS families according to HLA genotype (using the SNP rs2239802, which exhibited strongest evidence for association in the Finnish families in the report by Riise Stensland and coworkers [28]), strengthened the association between the *SLC1A3* SNP and MS ($P = 0.0002$, TDT). Thus, based on LA, and supported by association analyses in an MS study sample, the presence of *SLC1A3* serves as a potential candidate to connect all four major MS loci identified in Finnish families, elucidating a potential functional relationship

Table 1**Genetic association results for SNPs located in the *SLC1A3* gene in Finnish multiple sclerosis families**

All families (n=69) and HLA stratified (n=38)

SNP ID	TDT, P value	TDT, P value	MAF (CEPH)	MAF (Finnish)
rs2562582	0.005	0.0002	0.175	0.169
rs10941306	0.344	0.5	0.317	0.378
rs1366632	0.49	0.4898	0.458	0.459
rs1544795	0.477	0.5	0.467	0.437
rs1549627	0.5	0.4765	0.44	0.45

MAF (CEPH) is the minor allele frequency, as calculated from genotyping 30 trios belonging to Centre d'Etude du Polymorphisme Humain (CEPH) panel by HapMap project. MAF (Finnish) is the minor allele frequency calculated from genotyping the Finnish families used in this study. The transmission disequilibrium tests (TDTs) were calculated using ANALYZE software package [48]. In all, 69 families are included in this analysis, of which 38 are HLA stratified (multiple sclerosis families in which the affected individual had one or two HLA single nucleotide polymorphism [SNP] 2239802 risk alleles).

between genetically identified genes and loci. We consider further evidence in the following discussion.

Further liquid association analyses

We next took *MBP* and *SLC1A3* as the query genes to conduct a genome-wide LA search in all four gene expression databases. Figure 3 and Table 2 highlight a set of genes whose biological functions are most relevant to our MS study according to the literature. The detail LA outputs are given in Additional data file 1 (Tables S3 to S6). All LA plots are easy to generate online using our website. The one for the triplet including *MBP*, *PRKCA*, and *SLC1A3* is shown in Figure 4.

For GNF_2002 data, the gene with greatest LA score is *GRM3* (glutamate receptor, metabotropic 3), followed by several genes involved in nervous diseases and neural development/functioning: *GFAP*, *CDR1*, *ROM1*, *CACNA1A*, and *GRIA3*. We also find *IL7R*, *IGHG3*, *IGLJ3*, and *HLA-A* among the highest scoring genes in this query. The identification of the *IL7R* by the LA analysis is particularly interesting because this gene was found to be associated with MS in the recent large international Whole Genome Association study [8].

MBP-*SLC1A3* initiated LA search identifies the HLA locus on 6p21

The locus of *HLA* on 6p21 is the only consensus MS locus replicated by genetic studies across different populations. Importantly, in the recent fine mapping effort with 1,068 SNPs covering the HLA locus and providing the SNP density of 1 SNP per 2 kilobases in the study sample of 4,200 individuals from Finnish and Canadian MS families [29], susceptibility to MS proved to be determined by *HLA-DRB1* alleles and their interactions. Therefore, it is especially interesting that for the GNF_2004 data, eight of the 25 genes with the best LA scores are from the HLA locus: *HLA-A* (twice), *HLA-B* (twice), *HLA-C* (twice), and *HLA-G* (twice). Other HLA genes with very high LA scores include *HLA-E*, *HLA-F*, *HLA-DRA*, and *HLA-DPB1*. We also find *B2M* (which encodes β_2 -microglobulin, the light chain of MHC class I antigen) and a

MS susceptibility gene, namely *CD45* (a T-cell receptor for galectin-1).

Additional functionally associated genes detected

The LA lists from NCI_cDNA data and NCI_Affy data also yielded several highly relevant candidate genes for MS, such as *MAG*, *IRF1*, *APOE*, *EIV2A*, and *PDGFA*. Also of interest are *SIAT8A*, *SIAT1*, *SOX4*, *SOX9*, and *EPHA2*. The protein encoded by *MAG* (myelin-associated glycoprotein) is involved in the process of myelination [30] and binds to sialic acid. *SIAT1* and *SIAT8A* are both sialyltransferases. *SOX4* and *SOX9* are involved in central nervous system development [31,32]. *SOX4* is required for the development of lymphocytes and thymocytes [33].

Results from the two NCI datasets also contain genes from the 6p21.3 locus: *TAP2*, *TRIM10*, and *HLA-DQB1*. A further investigation into the expressional association of the HLA family with *SLC1A3* using the LA method finds two highly significant genes, namely *GMFB* and *PDGFRA* (see Additional data File 1 [Tables S7 and S8]). Also, these genes are biologically relevant. *GMFB* (which encodes glia maturation factor beta) is reported to increase in astrocytes around the lesioned area after cortical cryogenic brain injury [34]. *PDGFRA* (the gene encoding platelet-derived growth factor receptor- α) is a well known marker for remyelination. The *PDGFA* supply may control oligodendrocyte progenitor cell numbers in the adult central nervous system as well as during development [35]. Interestingly, *CTNND2* (catenin delta 2; neural plakophilin-related arm-repeat protein) is the fifth most correlated gene for *SLC1A3* in the GNF_2004 data (see Additional data file 1 [Table S9]). It is also highly correlated with *MBP* (see Additional data file 1 [Table S10]).

Discussion

We here introduce a novel bio-computational approach to identifying new candidate genes for genetic and functional studies of complex human traits. The initial result from the

Table 2**Genes detected in the liquid association analysis shown in Figure 1**

Gene ID	Location	Description
<i>MBP</i>	18q23*	Myelin basic protein
<i>PRKCA</i>	17q24.2*	Protein kinase C, alpha
<i>SLC1A3</i>	5p13.2†	Solute carrier family 1 (glial high affinity glutamate transporter), member 3
<i>A2M</i>	12p13.31*	α_2 -Macroglobulin
<i>GRM3</i>	7q21.12	Glutamate receptor, metabotropic 3
<i>GFAP</i>	17q21.31	Glial fibrillary acidic protein
<i>CDR1</i>	Xq27.1	Cerebellar degeneration-related protein 1, 34kDa
<i>ROM1</i>	11q12.3	Retinal outer segment membrane protein 1
<i>CACNA1A</i>	19p13.2	Calcium channel, voltage-dependent, P/Q type, alpha 1A subunit
<i>GRIA3</i>	Xq25	Glutamate receptor, ionotropic, AMPA3
<i>SOX21</i>	13q32.1	SRY (sex determining region Y)-box 21
<i>IL7R</i>	5p13.2†	Interleukin 7 receptor
<i>IGHG3</i>	14q32.33*	Immunoglobulin heavy constant gamma 3 (G3m marker)
<i>IGLJ3</i>	22q11.1	Immunoglobulin lambda joining 3
<i>HLA-A</i>	6p21.33†	Major histocompatibility complex, class I, A
<i>HLA-B</i>	6p21.33†	Major histocompatibility complex, class I, B
<i>HLA-C</i>	6p21.33†	Major histocompatibility complex, class I, C
<i>HLA-G</i>	6p22.1†	HLA-G histocompatibility antigen, class I, G
<i>B2M</i>	15q21.1	β_2 -Microglobulin
<i>PTPRC</i>	1q31.3*	Protein tyrosine phosphatase, receptor type, C
<i>EVI2A</i>	17q11.2*	Ecotropic viral integration site 2A
<i>TAP2</i>	6p21.32*	Transporter 2, ATP-binding cassette, sub-family B (MDR/TAP)
<i>TRIM10</i>	6p21.33†	Tripartite motif-containing 10
<i>SIAT1</i>	3q27.3	Sialyltransferase 1 (beta-galactoside alpha-2,6-sialyltransferase)
<i>MAG</i>	19q13.12*	Myelin associated glycoprotein [M29273]
<i>IRF1</i>	5q23.3*	Interferon regulatory factor 1
<i>APOE</i>	19q13.31*	Apolipoprotein E
<i>PDGFA</i>	7p22*	Platelet-derived growth factor- α polypeptide [H89357]
<i>SIAT8A</i>	12p12.1	Sialyltransferase 8A
<i>SOX4</i>	6p22.3†	SRY (sex determining region Y)-box 4
<i>SOX9</i>	17q24.3†	SRY (sex determining region Y)-box 9
<i>HLA-DQB1</i>	6p21.32†	Major histocompatibility complex, class II, DQ beta 1
<i>EPHA2</i>	1p36.13	EphA2 [T74614]
<i>GMFB</i>	14q22.2	Glia maturation factor, beta
<i>PDGFRA</i>	4q12*	Platelet-derived growth factor receptor- α polypeptide [M21574]
<i>PLP1</i>	Xq22.2*	Proteolipid protein 1
<i>KLK6</i>	19q13.41	Kallikrein 6 (neurosin, zyme)
<i>PMP2</i>	8q21.13	Peripheral myelin protein 2
<i>CTNND2</i>	5p15.2*	Catenin delta 2 (neural plakophilin-related arm-repeat protein)
<i>NTRK2</i>	9q21.33	Neurotrophic tyrosine kinase, receptor type 2
<i>PKP4</i>	2q24.1	Plakophilin 4

*Genes previously reported associated with MS. †Genes located within the previously reported MS susceptibility loci.

MS study is encouraging. We demonstrated that using only two genes, *MBP* and *PRKCA*, as the lead to probe for functionally associated genes, the LA method was successful in identifying a number of potential MS-related genes through subtle transcription co-regulation under a wide spectrum of cellular conditions. Additionally, when *MBP* and

SLC1A3 were used as query genes in the LA analysis, the recently identified MS susceptibility gene *IL7R* was among the highest scoring, statistically significant genes.

LA allows the detection of gene co-regulation, which may only occur under specific cellular states. There is no need to specify

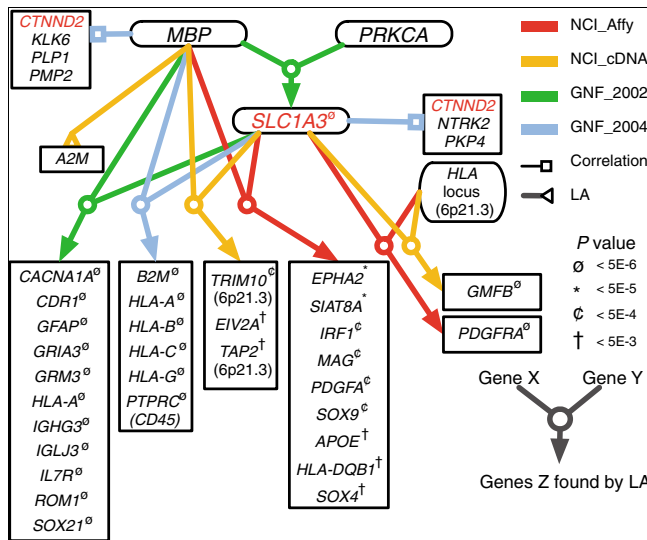


Figure 3
 SLC1A3 and related genes. Four large-scale gene expression databases are used in this study. The arrows point to the genes found using the liquid association score system, according to the search method described in Figure 2b. The color of a line/arrow shows which database is used in the analysis. P values are calculated by randomization test. For descriptions of gene symbols, see Table 2. All four major multiple sclerosis loci for the Finnish scan have representative genes in this chart: MBP from 18q23, PRKCA from 17q22-q23.2, SLC1A3 from 5p13, and the HLA locus at 6p21.3. Also shown are two separate lists of genes correlated with MBP and with SLC1A3 most strongly. CTNND2 (located at 5p15.2) is seen in both lists.

the states, and this is one of the advantages of LA [14]. Furthermore, LA can be used in conjunction with traditional correlation analysis. An online computation system to conduct both LA and correlation analysis is available at our website [36]. This platform allows users to switch conveniently from one gene expression dataset to another. Those conducting research in other diseases can easily carry out analyses similar to that presented here with a few leading genes related to the disease of interest.

Glutamate-induced excitotoxicity

Although all of the putative genes identified using LA method must be confirmed with genetic association studies in multiple populations and eventually in targeted functional studies, the putative genes identified here are highly relevant to MS. Our transcript regulatory findings portray a coherent web of molecular evidence, which supports the glutamate-induced excitotoxicity hypothesis of MS. SLC1A3 is highly expressed in various brain regions including cerebellum, frontal cortex, basal ganglia, and hippocampus. It encodes a sodium-dependent glutamate/aspartate transporter 1 (GLAST). Glutamate and aspartate are excitatory neurotransmitters that have been implicated in a number of pathologies the nervous system. Glutamate concentration in cerebrospinal fluid rises in acute MS patients [37], whereas the glutamate antagonist amantadine reduces MS relapse rate [38]. In EAE, the levels of GLAST and GLT-1 (SLC1A2) have been reported

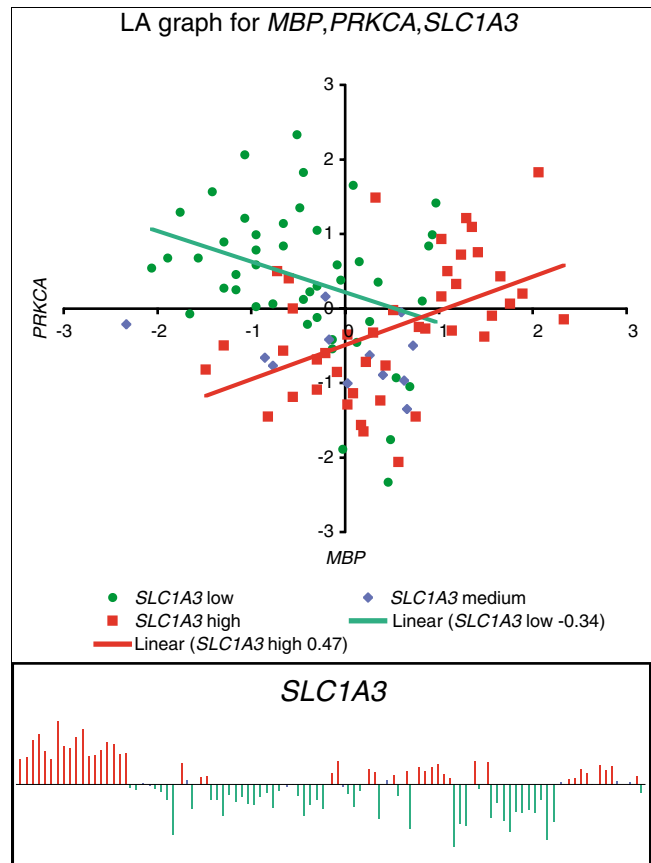


Figure 4
 Liquid association activity plot for MBP, PRKCA as mediated by SLC1A3. When SLC1A3 is upregulated (red squares), a positive association between MBP and PRKCA can be seen. The correlation vanishes when the expression of SLC1A3 is low (green dots). Liquid association measures the change in the correlation structure; the score is 0.438 for this triplet.

to be downregulated in spinal cord at the peak of disease symptoms, and no recovery was observed after remission [39]. We consider it encouraging that several lines of evidence, including both genetic association and gene expression association, are consistent with the glutamate-induced excitotoxicity hypothesis, which states that glutamate-induced excitotoxicity results in demyelination and axonal damage in MS [40].

International multiple sclerosis Whole Genome Association study

The recent international MS Whole Genome Association scan [8] provided additional evidence supporting an association between MS susceptibility and SLC1A3. A major component of the study is the use of Affymetrix 500K to screen common genetic variants of 931 family trios. Using the online supplementary information provided by the International MS Genetics Consortium [8] we found two SNPs, namely rs4869676 (chromosome 5: 36641766) and rs4869675 (chromosome 5: 36636676), with TDT P values of 0.0221 and 0.00399, respectively, which are in the upstream regulatory

region of the *SLC1A3* gene. In fact, within the 1 Mb region of rs486975 there are a total 206 SNPs in the Affymetrix 500K chip. No other SNPs have *P* values less than that of rs486975. The next most significant SNPs in this region are rs1343692 (chromosome 5: 35860930) and rs6897932 (chromosome 5: 35910332; the identified MS susceptibility SNP in the *IL7R* axon). The MS marker we identified, rs2562582 (chromosome 5: 36641117), less than 5 kilobases away from rs486975, was not used in the Affymetrix chip.

Although the results reported here should be considered preliminary, we propose that the genes and networks identified should be targets for additional analyses of MS in different study populations.

Use of public gene expression data

One unique feature of our approach to finding candidate genes is the use of public domain gene expression databases, of which the original experiments were not designed to study our disease of interest. For example, the two NCI-60 cell line (a panel of 60 diverse human cancer cell lines) gene expression data have primarily been used to aid anticancer drug screening, not for the study of MS. With our promising initial findings, we expect our functional genomics approach to be applicable in the initial identification of involved molecular pathways in the pathogenesis of other complex diseases. Investigators may apply our LA method or bring in other computational methods to data mine the numerous free public gene expression databases, thus reducing the time and expense associated with disease gene identification.

Materials and methods

Gene expression datasets

Four large-scale gene expression databases are employed in this study, with various numbers of conditions and genes. The first two databases give expression profiles for the 60 representative cell lines from seven cancer types that have been used in NCI's anticancer drug screen. The NCI_cDNA database uses the cDNA microarray reported by investigators from P Brown's laboratory at Stanford University [10], whereas the NCI_Affy uses Affymetrix oligonucleotide high-density HU6800 arrays [11]. The two other databases [12,13] are samples from diverse array of human tissues. GNF_2002 has a probe set for a total of 12,533 genes/clones and 101 chips (using Affymetrix U95A arrays) and GNF_2004 has a probe set for 33,689 genes and 158 chips (using Affymetrix HG-U133A and GNF1H; data downloaded from the Gene Expression Atlas [41]). The corresponding numbers for NCI_Affy are 5,611 and 60 (data downloaded from the supplementary data file of Staunton and coworkers [42]), and for NCI_cDNA they are 9,703 and 60 (data downloaded from the NCI60 Cancer Microarray Project) [43].

Liquid association

To compute the liquid association score $LA(X, Y|Z)$ for a triplet of genes, normal score transformation is first applied for each gene. After transformation, $LA(X, Y|Z)$ is given by the average of triple product between X, Y and Z: $LA(X, Y|Z) = (x_1y_1z_1 + \dots + x_my_mz_m)/m$. For a given pair (X,Y), the test of significance of an LA score is conducted by permutation, as previously described [14,16] and the *P* value is reported in each of our LA output tables. In addition, to help with the interpretation of the effect size of the LA score, two algorithms were used to find the correlation change between the state of high expression of the LA scouting gene and the state of low expression (see Additional data file 2 [Supplementary Text 5]). The LA website [36] was created to facilitate the online computation of LA. High score output genes are returned to user's browser for immediate connection to Entrez Gene. The website also generates LA graphs, performs standard correlation analysis, and provides summary information regarding gene location, functional annotation, and so on.

Multiple sclerosis association study sample

The study set used for the association analysis contained 28 multiplex MS families with multiple affected individuals, and 41 nuclear MS families (MS patient and his/her parents and, in case of a missing parent, healthy siblings were included). Twenty-two of the 28 multiplex families and all trio families originated from Southern Ostrobothnia region of Finland, which has an especially high incidence and prevalence of MS. All families were Finnish and of Caucasian descent, and they have been described in more detail by Saarela and coworkers [23]. Diagnosis of MS in affected individuals strictly followed Poser's diagnostic criteria [44]. All individuals gave informed consent and the study was approved by the Ethics Committee for Ophthalmology, Otorhinolaryngology, Neurology, and Neurosurgery in the Hospital District of Helsinki and Uusimaa. (decision 46/2002, DNRO 192/E9/02).

Genotyping

To control for sample mix-ups, all samples were genotyped for determining the sex and four microsatellite markers using the ABI 3730 (Applied Biosystems, Foster City, CA, USA). The data were compared with the known sex of the samples and checked for Mendelian errors. No Mendelian discrepancies were observed in this study set. To select the initial set of SNPs used for the association analysis, we set up the following criteria: each of the markers should be highly polymorphic in Centre d'Etude du Polymorphisme Humain (CEPH) reference families genotyped in the HapMap project and should belong to unique solid line linkage disequilibrium haplotype blocks as defined by Hapview's version 3.11 [45]. We found five highly polymorphic SNPs within and in the proximity of *SLC1A3* gene that belonged to separate haplotype blocks, according to the HapMap data. The SNPs were genotyped using multiplexed allele-specific primer extension on microarrays [46]. Primers for multiplex polymerase chain reactions were designed using in-house scripts written for the Primer3

program [47], and an in-house built software package SNP-Snapper, version 1.38beta, was utilized to call genotypes automatically.

Statistical analyses for genotyping

Allele and genotype frequencies were determined from the data, and deviation from the Hardy-Weinberg equilibrium was tested using Pearson's χ^2 test. We used the ANALYZE package to conduct TDT analyses to test for association between MS and *SLC1A3* gene [48].

Abbreviations

CEPH, Centre d'Etude du Polymorphisme Humain; EAE, experimental allergic encephalomyelitis; GLAST, glutamate/aspartate transporter 1; HLA, human leukocyte antigen; LA, liquid association; Mb, megabase; MHC, major histocompatibility complex; MS, multiple sclerosis; NCI, National Cancer Institute; SNP, single nucleotide polymorphism; TDT, transmission disequilibrium test.

Authors' contributions

SY and KCL contributed equally to statistical computing. DB, DC, JS, and OWC performed genotyping analysis. KCL, SY, and XW conducted LA analysis. KCL, AP, and LP designed the research and provided funding for research. KCL, AP, SY, DB, and LP wrote the paper.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 contains ten tables including results from both LA and correlation analyses. Additional data file 2 contains supplementary text detailing the additional data analyses mentioned in the text.

Acknowledgements

Research conducted by Li, Yuan, and Wei is supported by NSF grants DMS0201005 and DMS0406091. Li and Yuan were also supported in part by MIB, Institute of Statistical Science, Academia Sinica and grant NSC95-3114-P-002-005-Y. The research conducted by Palotie, Bronnikov, Chen, Wei, Choi, Saarela, and Peltonen is supported by NIH grant RO1 NS 43559, grants from Sigrd Juselius Foundation, Helsinki University Central Hospital Research Foundation and Center of Excellence of Disease Genetics of the Academy of Finland, and a grant from the National Multiple Sclerosis Society. We thank Sun Wei, Ching-Ti Liu, Yijing Shen, and Tun-Hsiang Yang for contributing to the LAP website development. Correspondence and requests for materials should be addressed to KL and LP. The authors are grateful to two anonymous referees for their insightful suggestions that greatly helped in improving the presentation.

References

- Trapp BD, Peterson J, Ransohoff RM, Rudick R, Mork S, Bo L: **Axonal transection in the lesions of multiple sclerosis.** *N Engl J Med* 1998, **338**:278-285.
- Oksenberg JR, Baranzini SE, Barcellos LF, Hauser SL: **Multiple sclerosis: genomic rewards.** *J Neuroimmunol* 2001, **113**:171-184.
- Ebers GC, Kukay K, Bulman DE, Sadovnick AD, Rice G, Anderson C, Armstrong H, Cousin K, Bell RB, Hader W, et al.: **A full genome**

- search in multiple sclerosis.** *Nat Genet* 1996, **13**:472-476.
- Haines JL, Ter-Minassian M, Bazyk A, Gusella JF, Kim DJ, Terwedow H, Pericak-Vance MA, Rimmler JB, Haynes CS, Roses AD, et al.: **A complete genomic screen for multiple sclerosis underscores a role for the major histocompatibility complex. The Multiple Sclerosis Genetics Group.** *Nat Genet* 1996, **13**:469-471.
- Sawcer S, Jones HB, Feakes R, Gray J, Smaldon N, Chataway J, Robertson N, Clayton D, Goodfellow PN, Compston A: **A genome screen in multiple sclerosis reveals susceptibility loci on chromosome 6p21 and 17q22.** *Nat Genet* 1996, **13**:464-468.
- Kuokkanen S, Gschwend M, Rioux JD, Daly MJ, Terwilliger JD, Tienari PJ, Wikstrom J, Palo J, Stein LD, Hudson TJ, et al.: **Genomewide scan of multiple sclerosis in Finnish multiplex families.** *Am J Hum Genet* 1997, **61**:1379-1387.
- Saarela J, Kallio SP, Chen D, Montpetit A, Jokiahio A, Choi E, Asselta R, Bronnikov D, Lincoln MR, Sadovnick AD, et al.: **PRKCA and multiple sclerosis: association in two independent populations.** *PLoS Genet* 2006, **2**:e42.
- The International Multiple Sclerosis Genetics Consortium: **Risk alleles for multiple sclerosis identified by a genomewide study.** *N Engl J Med* 2007, **357**:851-862.
- Pihlaja H, Rantamaki T, Wikstrom J, Sumelahti ML, Laaksonen M, Ilonen J, Ruutuainen J, Pirttila T, Elovaara I, Reunanen M, et al.: **Linkage disequilibrium between the MBP tetranucleotide repeat and multiple sclerosis is restricted to a geographically defined subpopulation in Finland.** *Genes Immun* 2003, **4**:138-146.
- Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, et al.: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24**:227-235.
- Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, et al.: **Chemosensitivity prediction by transcriptional profiling.** *Proc Natl Acad Sci USA* 2001, **98**:10787-10792.
- Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, et al.: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci USA* 2002, **99**:4465-4470.
- Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al.: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101**:6062-6067.
- Li KC: **Genome-wide coexpression dynamics: theory and application.** *Proc Natl Acad Sci USA* 2002, **99**:16875-16880.
- Li KC, Liu CT, Sun W, Yuan S, Yu T: **A system for enhancing genome-wide coexpression dynamics study.** *Proc Natl Acad Sci USA* 2004, **101**:15561-15566.
- Li KC, Yuan S: **A functional genomic study on NCI's anticancer drug screen.** *Pharmacogenomics J* 2004, **4**:127-135.
- Gunnarsson M, Jensen PE: **Binding of soluble myelin basic protein to various conformational forms of alpha2-macroglobulin.** *Arch Biochem Biophys* 1998, **359**:192-198.
- Jensen PE, Humle Jorgensen S, Datta P, Sorensen PS: **Significantly increased fractions of transformed to total alpha2-macroglobulin concentrations in plasma from patients with multiple sclerosis.** *Biochim Biophys Acta* 2004, **1690**:203-207.
- Hunter N, Weston KM, Bowern NA: **Suppression of experimental allergic encephalomyelitis by alpha 2-macroglobulin.** *Immunology* 1991, **73**:58-63.
- OMIM: Online Mendelian Inheritance in Man** [http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM]
- Zhang Y, Davis JL, Li W: **Identification of tribbles homolog 2 as an autoantigen in autoimmune uveitis by phage display.** *Mol Immunol* 2005, **42**:1275-1281.
- Poliak S, Matlis S, Ullmer C, Scherer SS, Peles E: **Distinct claudins and associated PDZ proteins form different autotypic tight junctions in myelinating Schwann cells.** *J Cell Biol* 2002, **159**:361-372.
- Saarela J, Schoenberg Fejzo M, Chen D, Finnilla S, Parkkonen M, Kuokkanen S, Sobel E, Tienari PJ, Sumelahti ML, Wikstrom J, et al.: **Fine mapping of a multiple sclerosis locus to 2.5 Mb on chromosome 17q22-q24.** *Hum Mol Genet* 2002, **11**:2257-2267.
- GAMES; Transatlantic Multiple Sclerosis Genetics Cooperative: **A meta-analysis of whole genome linkage screens in multiple sclerosis.** *J Neuroimmunol* 2003, **143**:39-46.
- Sawcer S, Ban M, Maranian M, Yeo TW, Compston A, Kirby A, Daly MJ, De Jager PL, Walsh E, Lander ES, et al.: **A high-density screen for linkage in multiple sclerosis.** *Am J Hum Genet* 2005,

- 77:454-467.
26. Barton A, Woolmore JA, Ward D, Eyre S, Hinks A, Ollier WER, Strange RC, Fryer AA, John S, Hawkins CP, et al.: **Association of protein kinase C alpha (PRKCA) gene with multiple sclerosis in a UK population.** *Brain* 2004, **127**:1717-1722.
 27. Feng JM, Fernandes AO, Campagnoni CW, Hu YH, Campagnoni AT: **The golli-myelin basic protein negatively regulates signal transduction in T lymphocytes.** *J Neuroimmunol* 2004, **152**:57-66.
 28. Riise Stensland HMF, Saarela J, Bronnikov DO, Parkkonen M, Jokiahho AJ, Palotie A, Tienari PJ, Sumelahti ML, Elovaara I, Koivisto K, et al.: **Fine mapping of the multiple sclerosis susceptibility locus on 5p14-p12.** *J Neuroimmunol* 2005, **170**:122-133.
 29. Lincoln MR, Montpetit A, Cader MZ, Saarela J, Dyment DA, Tiislar M, Ferretti V, Tienari PJ, Sadovnick AD, Peltonen L, et al.: **A predominant role for the HLA class II region in the association of the MHC region with multiple sclerosis.** *Nat Genet* 2005, **37**:1108-1112.
 30. Barton DE, Arquint M, Roder J, Dunn R, Francke U: **The myelin-associated glycoprotein gene: mapping to human chromosome 19 and mouse chromosome 7 and expression in quivering mice.** *Genomics* 1987, **1**:107-112.
 31. Cheung M, Abu-Elmagd M, Clevers H, Scotting PJ: **Roles of Sox4 in central nervous system development.** *Brain Res Mol Brain Res* 2000, **79**:180-191.
 32. Cheung M, Briscoe J: **Neural crest development is regulated by the transcription factor Sox9.** *Development* 2003, **130**:5681-5693.
 33. Kuo CT, Leiden JM: **Transcriptional regulation of T lymphocyte development and function.** *Annu Rev Immunol* 1999, **17**:149-187.
 34. Hotta N, Aoyama M, Inagaki M, Ishihara M, Miura Y, Tada T, Asai K: **Expression of glia maturation factor beta after cryogenic brain injury.** *Brain Res Mol Brain Res* 2005, **133**:71-77.
 35. Woodruff RH, Fruttiger M, Richardson WD, Franklin RJM: **Platelet-derived growth factor regulates oligodendrocyte progenitor numbers in adult CNS and their response following CNS demyelination.** *Mol Cell Neurosci* 2004, **25**:252-262.
 36. **Liquid Association Website** [<http://kiefner.stat.ucla.edu/LAP2/index.php>]
 37. Stover JF, Pleines UE, Morganti-Kossmann MC, Kossmann T, Lowitzsch K, Kempinski OS: **Neurotransmitters in cerebrospinal fluid reflect pathological activity.** *Eur J Clin Invest* 1997, **27**:1038-1043.
 38. Plaut GS: **Effectiveness of amantadine in reducing relapses in multiple sclerosis.** *J R Soc Med* 1987, **80**:91-93.
 39. Ohgoh M, Hanada T, Smith T, Hashimoto T, Ueno M, Yamanishi Y, Watanabe M, Nishizawa Y: **Altered expression of glutamate transporters in experimental autoimmune encephalomyelitis.** *J Neuroimmunol* 2002, **125**:170-178.
 40. Takahashi JL, Giuliani F, Power C, Imai Y, Yong VW: **Interleukin-1beta promotes oligodendrocyte death through glutamate excitotoxicity.** *Ann Neurol* 2003, **53**:588-595.
 41. **Gene Expression Atlas** [<http://expression.gnf.org>]
 42. **Supplemental data for Staunton et al.** [<http://www.genome.wi.mit.edu/MPR/NCI60/NCI60.html>]
 43. **NCI60 Cancer Microarray Project** [<http://genome-www.stanford.edu/nci60/>]
 44. Poser CM, Paty DW, Scheinberg L, McDonald WI, Davis FA, Ebers GC, Johnson KP, Sibley WA, Silberberg DH, Tourtellotte WW: **New diagnostic criteria for multiple sclerosis: guidelines for research protocols.** *Ann Neurol* 1983, **13**:227-231.
 45. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21**:263-265.
 46. Silander K, Komulainen K, Ellonen P, Jussila M, Alanne M, Levander M, Tainola P, Kuulasmaa K, Salomaa V, Perola M, et al.: **Evaluating whole genome amplification via multiply-primed rolling circle amplification for SNP genotyping of samples with low DNA yield.** *Twin Res Hum Genet* 2005, **8**:368-375.
 47. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132**:365-386.
 48. Terwilliger JD, Ott J: **A haplotype-based 'haplotype relative risk' approach to detecting allelic associations.** *Hum Hered* 1992, **42**:337-346.