

# Identity-by-Descent Mapping to Detect Rare Variants Conferring Susceptibility to Multiple Sclerosis

Rui Lin<sup>1</sup>, Jac Charlesworth<sup>1</sup>, Jim Stankovich<sup>1,2</sup>, Victoria M. Perreau<sup>3</sup>, Matthew A. Brown<sup>4</sup>, ANZgene Consortium<sup>¶</sup>, Bruce V. Taylor<sup>1\*</sup>

**1** Menzies Research Institute Tasmania, University of Tasmania, Hobart, Australia, **2** Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, **3** Centre for Neurosciences, Department of Anatomy and Neuroscience, University of Melbourne, Melbourne, Australia, **4** University of Queensland Diamantina Institute, Brisbane, Australia

## Abstract

Genome-wide association studies (GWAS) have identified around 60 common variants associated with multiple sclerosis (MS), but these loci only explain a fraction of the heritability of MS. Some missing heritability may be caused by rare variants that have been suggested to play an important role in the aetiology of complex diseases such as MS. However current genetic and statistical methods for detecting rare variants are expensive and time consuming. ‘Population-based linkage analysis’ (PBLA) or so called identity-by-descent (IBD) mapping is a novel way to detect rare variants in extant GWAS datasets. We employed BEAGLE fastIBD to search for rare MS variants utilising IBD mapping in a large GWAS dataset of 3,543 cases and 5,898 controls. We identified a genome-wide significant linkage signal on chromosome 19 (LOD=4.65;  $p = 1.9 \times 10^{-6}$ ). Network analysis of cases and controls sharing haplotypes on chromosome 19 further strengthened the association as there are more large networks of cases sharing haplotypes than controls. This linkage region includes a cluster of zinc finger genes of unknown function. Analysis of genome wide transcriptome data suggests that genes in this zinc finger cluster may be involved in very early developmental regulation of the CNS. Our study also indicates that BEAGLE fastIBD allowed identification of rare variants in large unrelated population with moderate computational intensity. Even with the development of whole-genome sequencing, IBD mapping still may be a promising way to narrow down the region of interest for sequencing priority.

**Citation:** Lin R, Charlesworth J, Stankovich J, Perreau VM, Brown MA, et al. (2013) Identity-by-Descent Mapping to Detect Rare Variants Conferring Susceptibility to Multiple Sclerosis. PLoS ONE 8(3): e56379. doi:10.1371/journal.pone.0056379

**Editor:** Amanda Ewart Toland, Ohio State University Medical Center, United States of America

**Received:** November 7, 2012; **Accepted:** January 8, 2013; **Published:** March 5, 2013

**Copyright:** © 2013 Lin et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was supported by the National Health and Medical Research Council (NHMRC, Australia) [Project 605511]; and the Australian Research Council [LP110100473]. MAB was supported by an NHMRC Senior Principal Research Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: Bruce.Taylor@utas.edu.au

¶ Membership of the Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) is provided in the Acknowledgments.

## Introduction

Multiple sclerosis (MS) is a complex neurological disease of the central nervous system (CNS) triggered by environmental and genetic factors. There is considerable evidence for a significant genetic component to MS susceptibility, such as a higher concordance rate in monozygotic twins (24%–30%) than dizygotic twins (3%–5%) [1,2]. As for other immune diseases, genome-wide association studies (GWAS) have been highly successful for MS: uncovering around 60 common genetic variants associated with disease [3–13]. The majority of these variants lie near genes with known functions in the immune system and these variants have also been associated with other autoimmune diseases, often in the opposite direction [14]. Virtually all of the variants confer modest increases in disease risk, the outstanding exception being the strong association with the *HLA-DRB1\*15:01* allele in the major histocompatibility complex (MHC), which was first detected in the 1970's [15,16].

Despite this success, the variants identified by GWAS to date only explain 18–24% of the heritability of MS [13,17]. While much of the missing heritability is probably explained by common variants of even smaller effect sizes, some heritability may be

explained by rare variants of larger effect size. Standard analysis of GWAS data is not designed to detect associations with rare variants that many believe may play an important role in the aetiology of complex traits [18–20]. Interestingly, GWAS have had less success for putative neurodegenerative diseases, such as Parkinson's disease, than for MS. For these diseases, family-based approaches detecting rare variants have been more successful [21,22]. This raises the possibility that rare variants under negative selection pressure are relatively more important in the genetic architecture of neurodegenerative processes, whereas common variants under balancing selection are more important in the genetic architecture of immunological processes. Discovery of rare MS susceptibility variants may alter perspectives on the relative importance of immunological & neurodegenerative processes in MS onset.

Standard analyses of GWAS data are not designed to detect associations with low frequency variants ( $MAF \leq 5\%$ ), and other strategies are required. One approach is to re-sequence loci containing common susceptibility variants identified from GWAS studies. This strategy was used to detect rare variants in *IFIH1* conferring protection to type I diabetes [18]. However this strategy

**Table 1.** Sample numbers from GWAS (after cleaning).

GWAS dataset	Country of origin		No. Case	No. Control	Total	No. SNPs
	Cases	Controls				
ANZgene [3]	AUS, NZ	US, UK	1,608	3,404	5,012	300,900
WTCCC2 [13]	AUS	-	766	-	766	586,393
QIMR [32]	-	AUS	-	1,516	1,516	529,292
GeneMSA [6]	US	US	878	805	1,683	550,677
Total			3,252	5,725	8,977	274,735*

\*The number of SNPs that passed QC in all 4 GWAS datasets.

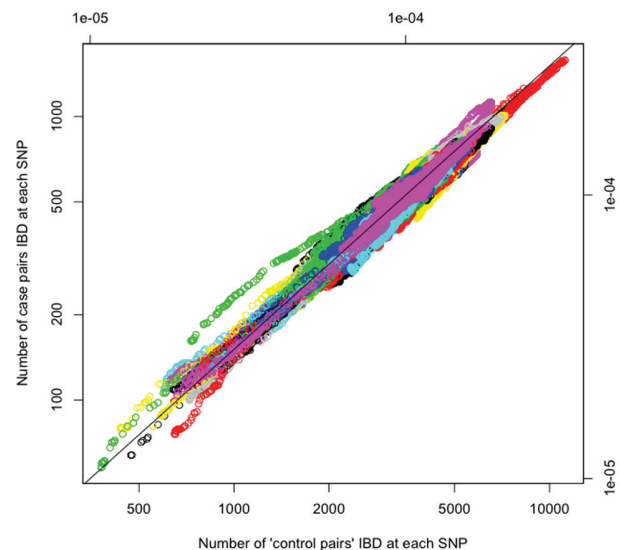
AUS = Australia, NZ = New Zealand.

doi:10.1371/journal.pone.0056379.t001

precludes the identification of new loci. Eventually it will be possible to overcome this limitation by whole genome sequencing, but it remains prohibitively expensive to perform adequately-powered studies. An alternative is to re-analyse GWAS data using identity-by-descent (IBD) mapping [23], also referred to as 'population-based linkage analysis' (PBLA) [24]. PBLA describes linkage analysis applied at the population level to detect mega base-scale regions where cases have inherited long haplotypes from distant ancestors, 10–100 generations ago. IBD mapping is performed on the unrelated individuals to determine whether these mega-base scale regions are identical and inherited from a common ancestor. If the common ancestor lived more than ten generations ago the individuals will share very short tracts of genetic material, and a shared haplotype that is very rare is also very likely to be IBD. HapMap Phase 3 identified that lower frequency variants should, on average, be younger than more common variants; and thus display a greater extent of haplotype sharing [25]. Therefore, if case pairs can be detected with long shared haplotypes (generally one to five megabases) inherited from distant common ancestors, then rare variants influencing disease risk can be localised. Even when whole genome sequencing becomes cheap enough to pursue with substantial sample sizes, IBD mapping may still help reduce the massive multiple testing problem by prioritizing regions. This is similar to the technique of prioritising association signals in regions of linkage [26].

Several methods of IBD mapping have been published: these include PLINK [24], GERMLINE [27], BEAGLE IBD [28] and BEAGLE fastIBD [29]. The models employed by PLINK and GERMLINE assume SNPs are in linkage equilibrium, and so 'pruning' of SNPs [24] is required to avoid false positives due to under-estimates of population haplotype frequencies. However pruning of SNPs in incomplete linkage disequilibrium (LD) discards potentially useful information and reduces power. BEAGLE IBD and fastIBD implement a variable length Hidden Markov Model [30] to account for LD and model haplotype frequencies more accurately. BEAGLE fastIBD runs considerably faster than BEAGLE IBD (of the order of 1000 times faster with large GWAS datasets). This is mainly because 1) it does not formally model IBD status ('IBD'/'not IBD') between pairs of individuals using a Hidden Markov Model as in BEAGLE IBD; 2) it stores haplotype frequencies in a data dictionary (as in GERMLINE) which means computational time scales with sample size  $n$  like  $n \log n$  instead of  $n^2$ .

To detect MS rare variants, we here use BEAGLE fastIBD to perform an IBD analysis on several large MS GWAS datasets comprised 3543 cases and 5898 controls. We identified a region of high significance on chromosome 19q13.43, with a genome-wide

**Figure 1.** Plots of raw data of IBD with one point for each SNP.

The green region is obvious outstanding from the black line, which indicates the proportion of case pairs in this region higher than that of control pairs. The black line represents where the proportion of case pairs equal to control pairs.

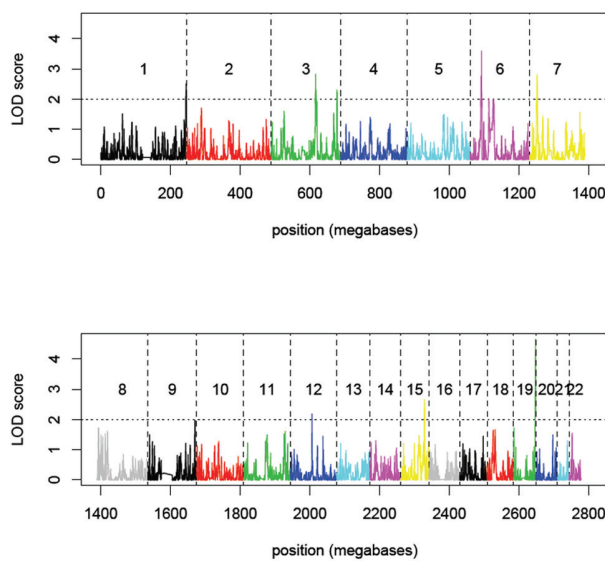
doi:10.1371/journal.pone.0056379.g001

significant localisation signal ( $p = 1.9 \times 10^{-6}$ ;  $\text{LOD} = 4.65$ ) using thresholds based on IBD segment length greater than 3 cM and the probability  $p$ -value less than  $10^{-9}$  ( $3\text{cM}_{1e-9}$ ). This locus was deemed genome-wide significant according to the recently established genome-wide significance thresholds set for IBD mapping [31]. Analysis of expression data and investigation of genes in this area support the hypothesis for regulation of gene expression in this region to impact upon development or health of CNS tissue. Our analyses also illustrate some of the practical issues to deal with in IBD analyses, and demonstrate that IBD mapping can form a potentially powerful method for detecting rare variants in unrelated individuals at the population level.

## Methods

### Study subjects

All the MS cases and controls were recruited and genotyped from MS GWAS totaling 3,543 cases and 5,898 controls. Of these, 1,618 cases and 3,413 controls were from an Australian and New Zealand MS GWAS conducted by the Australian and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) [3], and those DNA samples were genotyped on the Illumina Infinium Hap370CNV array [3]. An additional 861 Australian and New Zealand MS cases were genotyped with the Illumina Human660-Quad chip as part of a GWAS performed by the International MS Genetics Consortium (IMSGC) and the Wellcome Trust Case Control Consortium-2 (WTCCC2) [13]. Controls included 1,531 unrelated Australian samples from a GWAS genotyped by Queensland Institute of Medical Research (QIMR) with the Illumina Human610-Quad chip [32], and 1064 MS cases and 954 controls genotyped with the Sentrix® HumanHap550 BeadChip from a GWAS conducted in the US (GeneMSA) [6] [accessed via dbGAP].



**Figure 2. Plot of linkage scores along the whole genome with the IBD threshold of  $3cM_{1e-9}$  (shared haplotype segment  $>3 cM$  and haplotype probability  $p < 10^{-9}$ ). Chromosome 19 has the strongest linkage signal (LOD = 4.65).**  
doi:10.1371/journal.pone.0056379.g002

### Quality control of data

Conservative quality control measures were imposed both on the individual datasets before merging, and in the combined dataset after merging: SNPs with call rates less than 0.95 or in Hardy-Weinberg disequilibrium ( $p < 10^{-7}$ ) were discarded, as were samples with call rates less than 0.98. Duplicates and close relatives were also removed. This data cleaning was performed using PLINK.

A principal components analysis (PCA) was conducted by EIGENSTRAT [33] to exclude ancestry outliers and examine population structure within the remaining samples. First, SNPs in strong LD were pruned (using the PLINK – indep command with options 50 5 1.5), and then we excluded previously identified regions of high LD [34]. Outliers in the PCA were excluded using standard settings in Eigenstrat (more than six standard deviations from the mean along the first 10 principal components). All chromosomal locations refer to Human genome version hg18.

### Running BEAGLE fastIBD and results processing

The fastIBD analyses were conducted using BEAGLE (<http://faculty.washington.edu/browning/beagle/>). In brief, genotypes for the merged, cleaned dataset were converted to BEAGLE format by using the linkage2Beagle.jar utility program. We then used the BEAGLE method for phasing the data and identifying IBD segments simultaneously, using the ‘fastibdthreshold’ option. This procedure was run 10 times for each chromosome starting with different seeds of the random number generator.

The output of these calculations was a series of “putative” IBD segments shared between pairs of individuals. Each segment comes with the following information attached: ids for the pair of individuals, first and last SNPs in the IBD segment, length of the segment in centimorgans, and probability of the two individuals both carrying the segment if it was not IBD. We filtered these segments using various maximum probabilities and minimum segment lengths, as recommended in the BEAGLE manual. Results from the 10 runs were combined by taking the union of

IBD segments detected in each run. From the final list of segments, we wrote a Perl script to count numbers of case-case pairs ( $y_i$ ), case-control pairs ( $u_i$ ) and control-control pairs ( $v_i$ ) estimated to share haplotypes IBD at each SNP.

### Analysis of IBD

We focused on the detection of loci where groups of cases have inherited rare susceptibility alleles IBD. To do this, we modelled IBD sharing  $y_i$  in case-case pairs (“case pairs”) as a function of IBD sharing in  $x_i = u_i + v_i$  in case-control pairs and control-control pairs combined (“control pairs”).

We tried various methods to model the  $y_i$  as a function of the  $x_i$ : linear regression, negative binomial regression and Poisson regression. Models were fitted using R [35] and goodness of fit was assessed by examining diagnostic plots (SR\_commands S1).

At SNPs  $i$  with more IBD sharing in cases than expected, residuals  $z_i$  from the fitted models should be large and positive. To present residuals on a scale more familiar to geneticists, we converted them to LOD scores using the formula  $LOD_i = z_i^2 / (2 * \log_e(10))$ .

At the SNPs with the highest LOD scores, we calculated the proportions of case pairs sharing IBD in various populations, and plotted networks of case and control pairs sharing IBD with each other using the R network package (<http://cran.rproject.org/web/packages/network/index.html>).

## Results

### Study samples from GWAS after cleaning

11 individuals were excluded due to call rates less than 0.98 and an additional 202 individuals were excluded because they were close relatives or duplicates. PCA was conducted on a subset of 77,856 SNPs not in LD, which were common to all sample sets. Through successive iterations 251 outliers (37 AUS cases; 9 AUS controls; 6 UK controls; 97 US cases; 102 US controls) were excluded. All datasets overlapped well after the removal of outliers (Figure S1). In summary, following cleaning there were 3,243 cases and 5,725 controls with 274,735 autosomal SNPs in the final analysis (Table 1).

### Results of IBD analysis

We detected IBD with the threshold of IBD segment greater than 3 cM and the haplotype probability p-value less than  $10^{-9}$  ( $3cM_{1e-9}$ ). A strong linkage signal was observed in the HLA region (LOD = 3.58), while the strongest signal in non-HLA region was on chromosome 19 (LOD = 4.65), which reached genome-wide significance according to the recent established genome-wide significance threshold set for IBD mapping [31].

Figure 1 is a scatterplot of case-pair sharing  $y_i$  versus control-pair sharing  $x_i$  as each of the 274,735 SNPs  $i$ . Using different colours to represent SNPs on different chromosomes, an outlier group of SNPs with relatively high case pair sharing on one chromosome stands out in green.

### Fitting, testing Poisson model and converting to LOD scores

From examination of diagnostic plots (Figure S2, S3, S4), we found that the Poisson model provided the best fit for these data. Figure 2 shows a plot of residuals from the Poisson model converted to LOD scores. The highest linkage signal, corresponding to the green outlier region in Figure 1, was observed on chromosome 19 with LOD = 4.65 and  $p = 1.9 \times 10^{-6}$ . As expected, a strong signal also was observed in the HLA region (LOD = 3.58; Fig. 1).





alongside the expression change with development (**Fig. 3**). From this data we observed that the vast majority of genes in this region with high foetal expression levels are marked by both H3K9me3 methylation and bound by *KAP1* and *SETDB1* at the 3' end of the gene. A small number of genes are also bound at the 3' end of the transcript by *ZN274*. We also observed a pattern in the level of H3K9me3 methylation, with two maximum levels at about position 62,850,000 and 63,400,000 and trailing off at position 63,250,000. This bimodal pattern also occurs in the *KAP1* and *SETDB1* binding data and is even more apparent when viewing a wider view of the region. This position, marked in the figure by vertical black line, also corresponds with the position of rs159870 (chr19: 63239261) and there is break in synteny with rodent genomes in this zone.

### Comparison of IBD sharing among different populations

We next examined patterns of IBD sharing at the SNP with the highest LOD score on chromosome 19 (rs159872). We compared the proportion of IBD case pairs in different populations to determine whether there are particular populations that contribute to more IBD case pairs at this locus; and found the Tasmanian population has the highest proportion of IBD case pairs. When compared with all other combined Australian populations, the Tasmanian population significantly contributed more IBD case pairs at this locus ( $p = 0.004$ ); and was significant when compared with all other combined non-Tasmanian populations ( $p = 5.44 \times 10^{-5}$ ; **Table 2**).

### Networks of cases and controls sharing haplotypes on chromosome 19

**Figure 4** shows networks of cases and controls sharing haplotypes IBD at the SNP with the highest LOD score on chromosome 19. The biggest cluster comprises 10 cases sharing a haplotype in which 4 cases were from Melbourne, 4 from New Zealand, 1 from Sydney and 1 from USA. Another big cluster includes 2 cases from Melbourne, 3 cases from Tasmania and 4 cases from New Zealand (**Fig. 4A**). More generally, there are more networks of cases sharing haplotypes than controls.

**Table 2.** Comparison of IBD case pairs among different populations (rs159872 with the highest LOD score on chr19; LOD = 4.65).

Population	No. case	No. IBD case pairs	% IBD case pairs	p-value
TAS	308	7	$14.80 \times 10^{-5}$	Ref.
Mel	841	32	$9.06 \times 10^{-5}$	0.22
Newc	111	0	0.00	1.00
Syd	541	14	$9.58 \times 10^{-5}$	0.32
Other	32	0	0.00	1.00
AUS (non-TAS)	1525	46	$3.96 \times 10^{-5}$	0.004
NZ	540	14	$9.62 \times 10^{-5}$	0.32
US	879	22	$5.70 \times 10^{-5}$	0.033
Non-TAS	2944	82	$1.89 \times 10^{-5}$	$5.44 \times 10^{-5}$

\*% IBD case pairs = IBD pairs/case  $\times$  (case-1)/2; (Fisher's Exact Test).  
doi:10.1371/journal.pone.0056379.t002

## Discussion

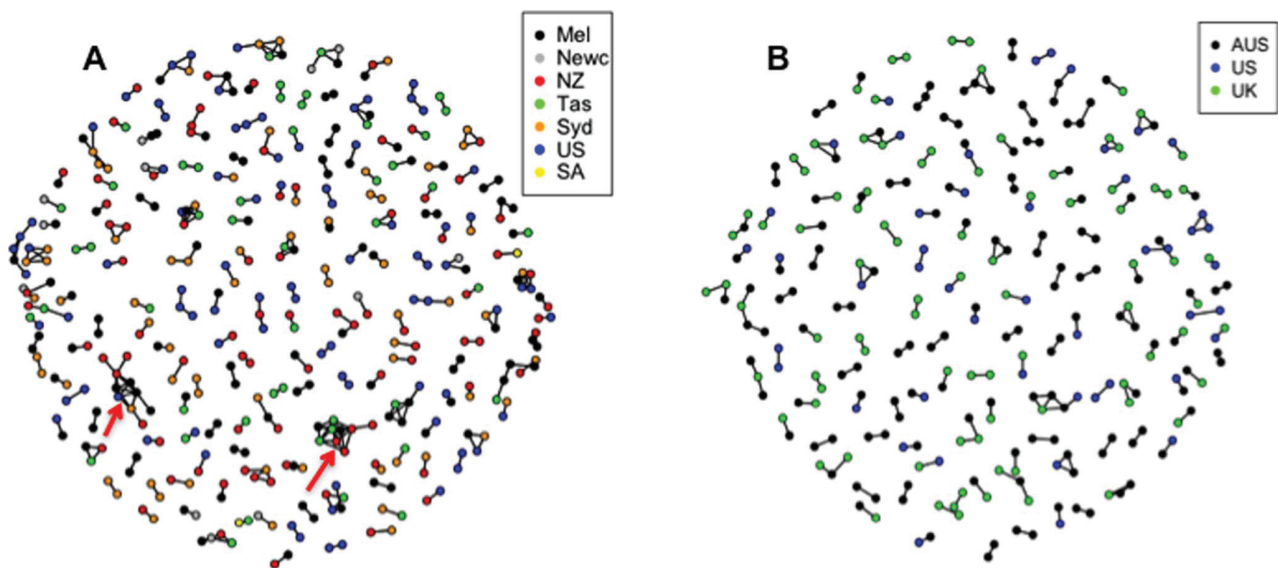
We have applied BEAGLE fastIBD for the detection of rare MS variants utilising a large-scale GWAS dataset. We identified a high linkage signal on chromosome 19 with a p-value of  $1.9 \times 10^{-6}$  (LOD = 4.65). In classical linkage analysis in small families, individuals are closely related and the segments of IBD tend to be fairly long ( $>10$  cM) which are easier to detect and less independent than IBD mapping, the generally-accepted threshold for genome-wide significance is  $p = 2.0 \times 10^{-5}$  [41]; while GWAS has more independent tests than IBD mapping, the threshold of genome-wide significance is around  $p = 5.0 \times 10^{-8}$  [42], so the threshold of p-value for IBD mapping genome-wide significant should be between  $5.0 \times 10^{-8}$  and  $2.0 \times 10^{-5}$ . Recently, researchers demonstrated that the genome-wide significance thresholds for IBD mapping depend on the IBD segment size detected or IBD generations [31]. For example, an IBD segment size of 2 cM corresponds to 25 generations and the genome-wide significance threshold is  $2.0 \times 10^{-6}$ , while the segment size of 3.2 cM corresponds to 15 generations and the genome-wide significance threshold is  $4.0 \times 10^{-6}$  [31]. The strongest non-HLA linkage signal we detected in this study used a 3 cM segment size; which corresponds to 17 generations, thus the genome-wide significance threshold is between  $4.0 \times 10^{-6}$  and  $2.0 \times 10^{-6}$ . As such, the linkage signal on chromosome 19, with a p-value of  $1.9 \times 10^{-6}$ , was determined to be genome-wide significant.

### Causal relationship between genes in linkage region and MS

Most genes in this linkage region are zinc finger (ZNF) proteins of which 32 genes have been suggested to be transcriptional regulators [43] (<http://genome.ucsc.edu/>). Seven genes (*ZN134*, *ZN135*, *ZN154*, *ZN549*, *ZN606*, *ZN671* and *ZN341*) in this region belong to the Krüppel family of ZNF genes. Only a few ZNF genes in this region have known vertebrate homologues and it includes a number of primate specific KRAB-ZNF genes [44]. In humans KRAB-ZNF genes number about 400 and make up the largest group of C2H2 transcription factors [45] which are typically expressed at low levels and involved in cell specific silencing and driving different cell lineages.

Detailed analysis of genes in this region did not reveal any direct links with MS. However examination of their expression profiles in published data revealed a shared early developmental CNS specific expression profile with 22 genes in this region being members of the expression module M20 described in [36], characterised by higher expression in all brain regions in early foetal time points followed by decreased expression prior to birth and very low expression thereafter. The M20 network of genes has a strong correlation with both neuronal differentiation and neuronal migration and a strong negative correlation with myelination [36].

Epigenetic mechanisms such as histone modification and DNA methylation are responsible for silencing many specific transcription factors including zinc finger genes, and the 3' end of many ZNF genes are specifically covered by H3K9me3 [46]. The zinc finger gene *ZN274*, located within the linkage region, is involved in gene silencing through recruitment of the histone methyltransferase complex TRIM28 (KAP1)/SETDB1 to the 3' end of specific ZNF genes [40]. Examination of H3K9me3, KAP1 and SETDB1 binding data, confirms that many of the genes in the linkage region are covered by H3K9me3 at their 3' end (**Fig. 3**). *ZN274* also interacts with *p75NTR* and is predicted to play a role in programmed cell death during development [47]. A number of the genes in this area are also highly expressed in differentiated



**Figure 4. Networks of cases and controls sharing haplotypes IBD at the SNP with the highest LOD score on chromosome 19. (A)** Networks of cases sharing haplotypes in common at the linkage region on chromosome 19. **(B)** Networks of controls sharing haplotypes in common at the linkage region on chromosome 19. Each dot represents an individual and each line connects pairs of individuals who share a haplotype. There are more big networks of cases sharing haplotypes than controls. doi:10.1371/journal.pone.0056379.g004

human neural cells compared to earlier stem cells (*ZNF549*, *ZNF324*, *ZNF548*, *ZNF264*, *ZNF671*, *ZSCAN1* and *ZSCAN18* are members of cluster A [48]). There is very little available evidence for involvement in immune cell activity for genes in this region. *ZNF304* is implicated in lymphocyte activation [49] and *ZNF274* has very high expression in activated eosinophils compared with other immune cell types [50]. Other genes in this region have relatively low expression and are not differentially regulated between immune cell types [50], as viewed in the immunological genome [51].

Together these findings suggest that many of the genes in this cluster may be involved in early differentiation of neuronal cells and potentially the silencing of genes required for myelination. Expression of ZNF genes is commonly detected in foetal brain and they are predicted to be involved in development of the nervous system, a KRAB zinc finger cluster on chromosome 8 has also been proposed to be involved in regulation of CNS development [52]. Although other clustered genes families have been shown to be co-expressed in cell types or tissues, previous studies have failed to identify coordinated expression of KRAB-ZNF gene clusters [44]. However earlier experiments did not examine the very early timepoints in CNS tissues included in the Kang dataset [36]. These expression profiles described in the M20 module are supported by two independent data set of both exon array level and RNA-seq expression data in early human CNS development available at the Allan Brain Atlas (<http://developinghumanbrain.org/>).

Thus this may be an example of a gene cluster of KRAB-ZNF genes exhibiting coordinated expression regulation, indicating the presence of a genomic regulatory block (GRB). Such regions are usually transcription factors controlled by highly conserved noncoding regions. Although the identification of GRBs remains difficult the evidence that we have collated is suggestive of two genomic regulatory blocks within the linkage region, interrupted at

the position of SNP rs159870 where there is an absence of H3K9me3 methylation and a break in synteny (reviewed in [53]).

The underlying cause for susceptibility in this region could therefore be due, not to differences in a specific gene expression or protein product, but to differences in the tight expression regulation of a GRB. As mentioned above, many of the C2H2 zinc finger genes in this region have an expression profile consistent with silencing of genes required for myelination. Further analysis needs to be undertaken to examine if these genes are co-regulated in demyelination and remyelination as well as CNS developmental states. Unfortunately, due to the species specificity of many of the KRAB-ZNF genes and the absence of rodent homologues of genes in this region, data from non-human models of demyelination and remyelination may not be useful.

Ideally, re-sequencing is the next step to refine this potential signal further. Unfortunately, resequencing of the region would be complicated since there are many gene duplications in this linkage region.

For the SNP (rs159872) with the highest LOD score on chromosome 19, we hypothesise that there are some difference between cases and controls sharing haplotypes in the linkage region among different populations. We found the Tasmanian MS population has the highest proportion of case IBD sharing, significantly higher than non-Tasmanian combined populations as well as other non-Tasmanian combined Australian populations. While Tasmania has the highest prevalence of MS in Australia, it is generally agreed that this is primarily driven by environmental effects related to, sunlight and/or vitamin D [54]. However there is also a modest founder effect in Tasmania [55], which might result in an increase in MS susceptibility driven by rare variants IBD. Interestingly, we found there are more big networks of cases sharing haplotypes than controls, and one big case network comprises 3 Tasmanian cases, 4 New Zealand cases and 2 cases from Melbourne, which may indicate the potential causal variants or gene mutations exist in those big case networks. However, this

SNP falls in a region of low/none methylation and correlates with a break in syteny, the significance of which is unclear.

### Technical considerations

Even though Beagle fastIBD is several orders of magnitude faster than Beagle IBD, IBD analysis remains moderately computationally intensive on a dataset of this size (8,977 individuals and 274,735 SNPs). For instance, on chromosome 2 with 22,607 SNPs, the computation time for each run was approximately 4.6 hours with memory requirement of 3.3 GB on 2 cores of a SGI Altix ICE 8200 HPC cluster computer node.

However, we also found IBD analysis limitations: it is only suited to discover rare variants if all variants act in the same direction in one gene. For example, the identified rare variants in *BRCA1* and *BRCA2* gene all increase risk of breast cancer [56], and the four rare variants identified in *IFIH1* gene all protect against type I diabetes [18]. If some rare variants increase risk while others in the same gene decrease risk then the signal in the region will be attenuated. In addition, we found IBD analysis is very sensitive to genotyping error, resulting in reducing signal strength. The linkage signal detected depends on a lot of markers or long haplotypes, containing up to hundreds of SNPs, a single error occurring in reading a single marker significantly reduces the signal. In our data, samples came from different GWAS using different genotyping chips in different locations, which at least in part, may decrease the potential signal strength from our study. Furthermore, resequencing would be complicated by gene duplication and repeat regions, since the linkage region detected in this study had many gene duplicates, thus replication in other independent dataset is needed.

The optimal method to detect rare disease-causing variants is whole genome sequencing of thousands of samples. When this becomes affordable, there will remain a role for IBD analysis to prioritize regions for follow-up analysis and minimize the massive multiple testing burden. Just as linkage analysis is now used to identify regions for follow-up in whole genome sequencing and exome sequencing of Mendelian disease families, and linkage analysis can be used to weight regions for GWA analysis [26].

In summary, we have applied IBD analysis to a large complex disease GWA dataset and identified a linkage signal with genome-wide significance, although it. While our most significant result is of equivocal significance, and lies in a region that is hard to validate via sequencing, we believe IBD analysis has considerable potential, particularly to help interpret whole-genome sequencing data in complex trait studies.

### Supporting Information

**Figure S1 Principal components analysis for the dataset.** Most individuals in the dataset are of predominantly northern European ancestry (right hand side), but some have southern European ancestry (left hand side) (one dot for each individual). (TIF)

**Figure S2 Fitting Poisson model for the IBD data.** All the four real lines in these four modules fit well with the default lines,

suggesting Poisson model is appropriate for this data. The residuals of the green region are higher than others. (TIF)

**Figure S3 Fitting negative binomial model for the IBD data.** All the four real lines in these four modules fit not well with the default lines, suggesting negative binomial model is not suitable for this IBD data. (TIF)

**Figure S4 Fitting linear model for the IBD data.** All the four real lines in these four modules fit not well with the default lines, suggesting linear model is not suitable for this IBD data. (TIF)

**SR\_commands S1** 1) Fitting and testing model for IBD data. 2) Plot of residuals from the Poisson model converted to LOD scores. 3) Network analysis. (PDF)

### Acknowledgments

We thank IMSGC and WTCCC2, QIMR and the ANZgene Consortium for access to the data relating to the MS associated SNPs. We thank the Tasmanian Partnership for Advanced Computing (TPAC) for access to the high performance computing facilities at the University of Tasmania.

**The members of the ANZgene Consortium are:** Alan Baxter (School of Pharmacy and Molecular Sciences, James Cook University, Townsville, Australia), Allan Kermode (Department of Neurology, Sir Charles Gairdner Hospital, Nedlands, Australia), Melanie Bahlo (Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Australia), William Carroll (Department of Neurology, Sir Charles Gairdner Hospital, Nedlands, Australia), Helmut Butzkueven (Department of Medicine, Royal Melbourne Hospital, Parkville, Australia), David Booth (Westmead Millennium Institute, University of Sydney, Sydney, Australia), Graeme Stewart (Westmead Millennium Institute, University of Sydney, Sydney, Australia), James Wiley (Howard Florey Institute, University of Melbourne, Melbourne, Australia), Judith Field (Howard Florey Institute, University of Melbourne, Melbourne, Australia), Lotti Tajouri (Genomics Research Centre, Griffith University, Brisbane, Australia), Lyn Griffiths (Genomics Research Centre, Griffith University, Brisbane, Australia), Michael Barnett (Brain and Mind Research Institute, University of Sydney, Camperdown, Australia), Pablo Moscato (Hunter Medical Research Institute, Newcastle, Australia), Robert Heard (Westmead Millennium Institute, University of Sydney, Sydney, Australia), Rodney Scott (School of Biomedical Sciences, University of Newcastle, Newcastle, Australia), Shaun McColl (School of Molecular & Biomedical Science, University of Adelaide, Adelaide, Australia), Simon Foote (Australian School of Advanced Medicine, Macquarie University, Sydney, Australia), Simon Broadley (School of Medicine, Griffith University, Gold Coast Campus, Australia), Mark Slee (School of Medicine, Flinders University of South Australia, Adelaide, Australia), Steve Vucic (Western Clinical School, University of Sydney, Randwick, Australia), Trevor Kilpatrick (Centre for Neurosciences, Department of Anatomy and Neuroscience, University of Melbourne, Melbourne, Australia).

### Author Contributions

Conceived and designed the experiments: JS. Performed the experiments: MAB ANZgene Consortium. Analyzed the data: RLJS VMP. Contributed reagents/materials/analysis tools: ANZgene Consortium. Wrote the paper: RLJS VMP JC BVT.

### References

1. Sadovnick AD, Armstrong H, Rice GP, Bulman D, Hashimoto L, et al. (1993) A population-based study of multiple sclerosis in twins: update. *Ann Neurol* 33: 281–285.
2. Hansen T, Skytthe A, Stenager E, Petersen HC, Bronnum-Hansen H, et al. (2005) Concordance for multiple sclerosis in Danish twins: an update of a nationwide study. *Mult Scler* 11: 504–510.
3. The Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nat Genet* 41: 824–828.
4. Ban M, Goris A, Lorentzen AR, Baker A, Mihalova T, et al. (2009) Replication analysis identifies TYK2 as a multiple sclerosis susceptibility factor. *Eur J Hum Genet* 17: 1309–1313.



5. Ban M, McCauley JL, Zuvich R, Baker A, Bergamaschi L, et al. (2010) A non-synonymous SNP within membrane metalloendopeptidase-like 1 (MMEL1) is associated with multiple sclerosis. *Genes Immun* 11: 660–664.
6. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, et al. (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. *Hum Mol Genet* 18: 767–778.
7. De Jager PL, Jia X, Wang J, de Bakker PI, Ottoboni L, et al. (2009) Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet* 41: 776–782.
8. Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, et al. (2007) Risk alleles for multiple sclerosis identified by a genome-wide study. *N Engl J Med* 357: 851–862.
9. Hafler JP, Maier LM, Cooper JD, Plagnol V, Hinks A, et al. (2009) CD226 Gly307Ser association with multiple autoimmune diseases. *Genes Immun* 10: 5–10.
10. The International Multiple Sclerosis Genetics Consortium (IMSGC). (2010) Comprehensive follow-up of the first genome-wide association study of multiple sclerosis identifies KIF21B and TMEM39A as susceptibility loci. *Hum Mol Genet* 19: 953–962.
11. Jakkula E, Leppä V, Sulonen AM, Varilo T, Kallio S, et al. (2010) Genome-wide association study in a high-risk isolate for multiple sclerosis reveals associated variants in STAT3 gene. *Am J Hum Genet* 86: 285–291.
12. Sanna S, Pitzalis M, Zoledziwska M, Zara I, Sidore C, et al. (2010) Variants within the immunoregulatory CBLB gene are associated with multiple sclerosis. *Nat Genet* 42: 495–497.
13. Sawcer S, Hellenthal G, Pirinen M, Spencer CC, Patsopoulos NA, et al. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* 476: 214–219.
14. Raj T, Shulman JM, Keenan BT, Chibnik LB, Evans DA, et al. (2012) Alzheimer Disease Susceptibility Loci: Evidence for a Protein Network under Natural Selection. *Am J Hum Genet* 90: 720–726.
15. Jersild C, Dupont B, Fog T, Platz PJ, Svejgaard A (1975) Histocompatibility determinants in multiple sclerosis. *Transplant Rev* 22: 148–163.
16. Jersild C, Fog T (1972) Histocompatibility (HL-A) antigens associated with multiple sclerosis. *Acta neurologica Scandinavica Supplementum* 51: 377.
17. O’Gorman C, Lin R, Stankovich J, Broadley SA (2012) Modelling Genetic Susceptibility to Multiple Sclerosis with Family Data. *Neuroepidemiology* 40: 1–12.
18. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 324: 387–389.
19. Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, et al. (2011) Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet* 43: 519–525.
20. Ji W, Foo JN, O’Roak BJ, Zhao H, Larson MG, et al. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet* 40: 592–599.
21. Yu Z, Zhu Y, Chen-Plotkin AS, Clay-Falcone D, McCluskey L, et al. (2011) PolyQ repeat expansions in ATXN2 associated with ALS are CAA interrupted repeats. *PLoS One* 6: e17951.
22. Kang SJ, Scott WK, Li YJ, Hauser MA, van der Walt JM, et al. (2006) Family-based case-control study of MAOA and MAOB polymorphisms in Parkinson disease. *Mov Disord* 21: 2175–2180.
23. Browning SR (2008) Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics* 178: 2123–2132.
24. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81: 559–575.
25. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, et al. (2010) Integrating common and rare genetic variation in diverse human populations. *Nature* 467: 52–58.
26. Roeder K, Bacanu SA, Wasserman L, Devlin B (2006) Using Linkage Genome Scans to Improve Power of Association in Genome Scans. *Am J Hum Genet* 78: 243–252.
27. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, et al. (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res* 19: 318–326.
28. Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. *Am J Hum Genet* 86: 526–539.
29. Browning BL, Browning SR (2011) A fast, powerful method for detecting identity by descent. *Am J Hum Genet* 88: 173–182.
30. Browning SR (2006) Multilocus Association Mapping Using Variable-Length Markov Chains. *Am J Hum Genet* 78: 903–913.
31. Browning SR, Thompson EA (2012) Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics* 190: 1521–1531.
32. Medland SE, Nyholt DR, Painter JN, McEvoy BP, McRae AF, et al. (2009) Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am J Hum Genet* 85: 750–755.
33. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38: 904–909.
34. Fellay J, Shianna KV, Ge D, Colombo S, Ledergerber B, et al. (2007) A whole-genome association study of major determinants for host control of HIV-1. *Science* 317: 944–947.
35. R Development Core Team (2010) R: A language and environment for statistical computing. Vienna, Australia: R Foundation for Statistical Computing. Available: <http://www.R-project.org>.
36. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, et al. (2011) Spatio-temporal transcriptome of the human brain. *Nature* 478: 483–489.
37. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, et al. (2011) NCBI GEO: archive for functional genomics data sets – 10 years on. *Nucleic Acids Res* 39: D1005–1010.
38. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. *Genome Res* 12: 996–1006.
39. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32: D493–496.
40. Frieze S, O’Geen H, Blahnik KR, Jin VX, Farnham PJ (2010) ZNF274 recruits the histone methyltransferase SETDB1 to the 3’ ends of ZNF genes. *PLoS One* 5: e15082.
41. Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 11: 241–247.
42. Dudbridge F, Gusnanto A (2008) Estimation of significance thresholds for genome-wide association scans. *Genet Epidemiol* 32: 227–234.
43. Urrutia R (2003) KRAB-containing zinc-finger repressor proteins. *Genome Biol* 4: 231.
44. Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, et al. (2006) A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. *Genome Res* 16: 669–677.
45. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 10: 252–263.
46. O’Geen H, Squazzo SL, Iyengar S, Blahnik K, Rinn JL, et al. (2007) Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. *PLoS Genet* 3: e89.
47. Casademunt E, Carter BD, Benzel I, Frade JM, Dechant G, et al. (1999) The zinc finger protein NRIF interacts with the neurotrophin receptor p75(NTR) and participates in programmed cell death. *Embo J* 18: 6050–6061.
48. Fathi A, Hatami M, Hajihosseini V, Fattahi F, Kiani S, et al. (2011) Comprehensive gene expression analysis of human embryonic stem cells during differentiation into neural cells. *PLoS One* 6: e22856.
49. Sabater L, Ashhab Y, Caro P, Kolkowski EC, Pujol-Borrell R, et al. (2002) Identification of a KRAB-containing zinc finger protein, ZNF304, by AU-motif-directed display method and initial characterization in lymphocyte activation. *Biochem Biophys Res Commun* 293: 1066–1072.
50. Jeffrey KL, Brummer T, Rolph MS, Liu SM, Callejas NA, et al. (2006) Positive regulation of immune cell function and inflammatory responses by phosphatase PAC-1. *Nat Immunol* 7: 274–283.
51. Heng TS, Painter MW (2008) The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol* 9: 1091–1094.
52. Lorenz P, Dietmann S, Wilhelm T, Koczan D, Autran S, et al. (2010) The ancient mammalian KRAB zinc finger gene cluster on human chromosome 8q24.3 illustrates principles of C2H2 zinc finger evolution associated with unique expression profiles in human tissues. *BMC Genomics* 11: 206.
53. Navratilova P, Becker TS (2009) Genomic regulatory blocks in vertebrates and implications in human disease. *Brief Funct Genomic Proteomic* 8: 333–342.
54. Simpson S Jr., Blizzard L, Otaahal P, Van der Mei I, Taylor B (2011) Latitude is significantly associated with the prevalence of multiple sclerosis: a meta-analysis. *J Neurol Neurosurg Psychiatry* 82: 1132–1141.
55. Rubio JP, Bahllo M, Butzkueven H, van Der Mei IA, Sale MM, et al. (2002) Genetic dissection of the human leukocyte antigen region by use of haplotypes of Tasmanians with multiple sclerosis. *Am J Hum Genet* 70: 1125–1137.
56. Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, et al. (2003) Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet* 72: 1117–1130.