Resequencing and fine-mapping of the chromosome 12q13-14 locus associated with multiple sclerosis refines the number of implicated genes

Adrian Cortes\textsuperscript{1,‡}, Judith Field\textsuperscript{2,‡}, Evgeny A. Glazov\textsuperscript{1}, Johanna Hadler\textsuperscript{1}, ANZGene Consortium, Jim Stankovich\textsuperscript{3}, Matthew A. Brown\textsuperscript{1,*}

\textsuperscript{1}University of Queensland Diamantina Institute, Princess Alexandra Hospital, Brisbane, Queensland 4102, Australia

\textsuperscript{2}Multiple Sclerosis Division, Florey Institute of Neuroscience and Mental Health, University of Melbourne, Melbourne, Victoria 3053, Australia

\textsuperscript{3}Menzies Research Institute, University of Tasmania, Hobart, Tasmania 7000, Australia

*Corresponding Author: Professor Matthew Brown, University of Queensland Diamantina Institute, Princess Alexandra Hospital, Ipswich Road, Woolloongabba, Qld, 4102, Australia. Email: matt.brown@uq.edu.au, Fax: +61 7 3240 5946

‡The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors.
ABSTRACT

Multiple sclerosis (MS) is a common chronic inflammatory disease of the central nervous system. Susceptibility to the disease is affected by both environmental and genetic factors. Genetic factors include haplotypes in the MHC and over 50 non-MHC loci reported by genome-wide association studies. Amongst these, we previously reported polymorphisms in chromosome 12q13-14 with a protective effect in individuals of European descent. This locus spans 288kb and contains 17 genes; including several candidate genes which have potentially significant pathogenic and therapeutic implications. In this study we aimed to fine-map this locus. We have implemented a two phase study: a variant discovery phase where we have used next-generation sequencing and two target-enrichment strategies (long-range PCR and Nimblegen’s solution phase hybridization capture) in pools of 25 samples; and a genotyping phase where we genotyped 712 variants in 3,577 healthy controls and 3,269 MS patients. This study confirmed the association (rs2069502, P = 9.9 x 10^{-11}, OR = 0.787) and narrowed down the locus of association to a 86.5kb region. Although the study was unable to pinpoint the key-associated variant we have identified a 42 (genotyped and imputed) SNP haplotype block likely to harbour the causal variant. No evidence of association at previously reported low-frequency variants in CYP27B1 was observed. As part of the study we compared variant discovery performance using two target-enrichment strategies. We concluded that our pools enriched with Nimblegen’s solution phase hybridization capture had better sensitivity to detect true variants than the pools enriched with long-range PCR, whilst specificity was better in the longe-range PCR enriched pools compared to solution phase hybridization capture enriched pools; this result has important implications for the design of future fine-mapping studies.
INTRODUCTION

Multiple sclerosis (MS) is the most common chronic inflammatory disease of the central nervous system (CNS) in young adults, where the myelin sheath that protects and covers the neurons in the CNS are damaged, leading to demyelination and scarring (1). Susceptibility to MS is influenced by both genetic (2) and environmental factors (3). The predominant genetic risk factor in Northern Europeans is the HLA-DRB1*1501-DQB1*602 (HLA-DR15) haplotype in the major histocompatibility complex (MHC) (4, 5). Amongst the known environmental risk factors the geographic latitude gradient has consistent support (6), where prevalence and incidence of the disease increases with distance from the equator (7). The biological mechanisms responsible for this association are thought to involve vitamin D metabolism (8).

Genome-wide association studies have identified over 50 non-MHC susceptibility loci (9-14). Amongst these, the Australian and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene) reported a novel association in chromosome 12q13-14 (13). This region spans 288kb (chromosome 12: 56,255,005-56,542,981; NCBI human genome build 36 coordinates) and encompasses 17 known genes. Tight linkage disequilibrium patterns in this region have hindered identification of the primary MS associated gene. Of these 17 genes, the strongest candidate MS susceptibility gene given the available genetic, immunological and epidemiological evidence is CYP27B1, a gene that encodes the enzyme 25-hydroxyvitamin D-1 alpha hydroxylase, which hydroxylates 25-hydroxyvitamin D into the bioactive form (1,25-dihydroxyvitamin D (1,25(OH)2D)). The 1,25(OH)2D form regulates calcium metabolism through the vitamin D receptor (VDR) and has important immune functions, modulating innate immunity, adaptive immunity and tolerance (15). CYP27B1 can also direct activated T cells toward a T helper type 2
anti-inflammatory phenotype (16) and induce dendritic cells with tolerogenic properties (17). Polymorphisms in the 5’-end region of *CYP27B1* are known to affect serum levels of 25(OH)D and 1,25(OH)$_2$D in type 1 diabetes patients (18, 19) and low frequency mutations affecting CYP27B1 function have been associated to MS susceptibility in multi-case families (20). Interestingly, another gene encoding the enzyme, which breaks down 1,25(OH)$_2$D, *CYP24A1*, has recently been reported to affect susceptibility to MS (14). Another plausible candidate gene in the chromosome 12q13-14 region is *CDK4* (cyclin-dependent kinase 4). *CDK4* expression can be modulated by signalling from CD40 (21), located at another MS susceptibility locus (13), and promote autoreactivity in the non-obese diabetic (NOD) mice (22). CDK4 inhibitors can ameliorate mouse collagen-induced arthritis (23) and, interestingly, has been associated with risk to rheumatoid arthritis (RA) and Celiac disease (24). CDK4 is downregulated in T cells from Japanese MS cases compared to controls (25). The chromosome 12q13-14 region has also been associated with RA, although the localisation of this association is p-telomeric of the MS one, with the peak association observed within the gene *KIF5A* (26). Transcriptome analysis on whole blood mRNA of MS patients and healthy controls has identified polymorphisms in this region associated with expression of *FAM119B* (27), a gene of unknown function. This study found a reduced association of the risk SNP with the level of expression of *CDK4* compared to *FAM119B*, the expression of which was markedly reduced with the risk SNP; no expression of *CYP27B1* was observed in whole blood cells. So far, it is not clear from the gene-mapping studies that have been done to date which of these 17 genes is causing the MS-associated signal.

To fine-map the association at chromosome 12q13-14 we have implemented an approach previously validated in human complex I deficiency (28). This approach consists on a variant discovery phase, where DNA from targeted regions is sequenced in pools of samples using next
generation sequencing technology (29). This is followed by a genotyping phase, where the comprehensive catalogue of variants in the target region is genotyped in a larger cohort of samples to fine-map the association.

Pooled sequencing has become a cost-effective approach to sequence a collection of samples on target regions and take full advantage of the high throughput of next-generation sequencing. Amongst the available target-enrichment strategies (30) we utilised two methods: long-range PCR and Nimblegen’s solution phase hybridization capture. To date no comprehensive comparison of these techniques has been reported.

**RESULTS**

**SNP discovery by pooled sequencing**

DNA from the 288kb target region from 175 MS cases and 150 controls was sequenced by pooling DNA with 25 samples per pool. Of the seven pools from cases, three were prepared using Nimblegen’s sequence capture (SeqCap), two using long-range PCR amplification and two using both techniques. Similarly, of the six pools of controls, three were prepared using Nimblegen’s sequence capture, one using long-range PCR and two using both techniques (see Methods). The four DNA pools enriched with both techniques also provided the means to compare both approaches.

High-throughput sequencing yielded large amounts of high-quality data for each pool (supplementary table 1 for pool sequencing statistics). The SNPs were identified using Syzygy (31); a variant calling algorithm that models the ploidy of the sequence data, in this case 50 chromosomes per pool. In total we identified 1,699 high-confidence variants (table 1) using both
protocols. The functional classifications of previously reported SNPs were similar to novel SNPs identified in this study, with a slightly higher proportion of novel intronic SNPs observed, and a lower proportion of novel intergenic SNPs (table 2).

To quantify the sensitivity of our variant discovery framework we genotyped samples from two pools on the Immunochip genotyping chip (32) which is enriched for SNPs in the targeted region. These two pools comprised of either 25 cases or 25 controls, and were sequenced following enrichment with both protocols. The Immunochip contains 594 markers in the target region; according to the Immunochip typing, 268 out of the 594 were polymorphic in the case pool (at least one copy of the alternate allele) and 273 were polymorphic in the control pool. With this sample size the lowest reportable minor allele frequency (MAF) in each pool was 0.02. Sensitivity of the variant discovery framework was computed as the proportion of these polymorphic variants that were called with high-confidence. In our long-range PCR enriched pools the sensitivity was 0.797 for case and 0.754 for control pools. In the SeqCap enriched pools the sensitivity was 0.868 in case and 0.855 control pools. These results suggested that a better sensitivity is obtained using the SeqCap approach. Estimated MAF by read counts was also better approximated in the SeqCap libraries compared to long-range PCR libraries, with closer correlation of allele counts from the Immunochip genotype data. In the SeqCap libraries allele frequency correlation was 0.916 in cases and 0.923 in controls compared with 0.871 in cases and 0.718 in controls for long-range PCR libraries (table 3 and figures 1a-d). Including sequence variants called with lower confidence increases the sensitivity estimates but decreases correlations with MAF from Immunochip genotypes. For example, considering SeqCap in cases, the sensitivity of high quality calls was 0.868 and with all SNP calls was 0.899, whereas the MAF correlation was 0.916 for high quality variants and 0.897 for all SNP calls.
To provide a direct measurement of the false positive rate in our study we validated 150 novel SNPs using Sequenom iPLEX genotyping (33) (supplementary table 2). Of these, 80 were detected with high confidence in the long-range PCR library, 66 in the SeqCap library, and the rest had been identified with medium confidence. The positive predictive value (PPV) for previously unreported high quality SNPs was estimated to be 0.75 in the long-range PCR library and 0.60 in the SeqCap library. For previously unreported medium quality SNPs, the PPV was 0.68 in the long-range PCR library and 0.44 in the SeqCap library. This agrees with our observation that the transition-transversion ratio on the PCR library was higher than the ratio obtained in the SeqCap library variants, consistent with a lower false positive rate in the PCR library.

Fine-mapping chromosome 12q13-14 association with MS

Our next goal was to genotype the newly discovered variants and previously reported variants in the target region, in a larger cohort of samples. We successfully genotyped 712 SNPs in the target region in 3,577 healthy controls and 3,269 MS cases and reproduced the association with MS susceptibility on chromosome 12 (variant selection discussed in Methods). To increase the number of SNPs for the genetic association analysis we successfully imputed 573 SNPs in the target region from the 1,000 Genomes Project reference dataset. In total, 40 genotyped and 91 imputed SNPs were found to be associated with MS susceptibility (genome-wide significance $P < 5 \times 10^{-8}$) (partial results in table 4 and complete list in supplementary table 3). In our original report of the chromosome 12q13-14 association, the strongest association was with SNP rs703842 ($P = 4.1 \times 10^{-6}$ in the discovery GWAS, $P = 5.4 \times 10^{-11}$ overall, including the replication study (13)). In this study, we found a stronger association with the genotyped SNP rs2069502 ($P = 9.9 \times 10^{-11}$, OR = 0.787) than with the previously reported SNP rs703842 ($P = 8.5 \times 10^{-10}$, OR = 0.796; genotyped), however these two SNPs were in tight linkage disequilibrium ($r^2 = 0.88$).
This association was also stronger than the peak association observed at this locus in the IMSGC/WTCCC2 study (rs12368653, $P = 4.8 \times 10^{-7}$, OR = 0.84; genotyped) (14). Logistic regression conditioning on either rs2069502 or rs703842 also revealed no residual association in the target region ($P > 0.01$) (figure 3). GENECLUSTER (34) analysis, which can be used to assess if the pattern of association is not sufficiently characterized by the most significant SNP, suggested a one mutation model over a two mutation model (with a posterior probability of 0.50 for two mutations versus one mutation assuming both models are equally likely a priori).

GENECLUSTER identified 24 variants highly correlated ($r^2 > 0.81$) with the causal mutation which were directly genotyped (table 4). Haplotype analysis using Haploview (35) revealed five LD blocks (figure 2), one of which contained all but one SNP identified by GENECLUSTER, as highly correlated with the causal mutation. This block contains rs2069502 and rs703842 and it spans approximately 86.5kb, where the boundary was defined as the region where the p-value of association was within two log orders (base 10) from the peak of association.

From the imputed dataset, 18 SNPs were found to be highly correlated to rs2069502 ($r^2 > 0.87$, $D' > 0.95$) and highly associated to MS susceptibility ($P$-value $< 6 \times 10^{-09}$). From the 573 imputed SNPs, 362 (63.2%) were observed with high confidence in at least one pool of the discovery phase and assay design was attempted on 192 (33.5%), according to the variant selection strategy (see methods).

In our variant discovery phase the SNP rs118204009, reported to be a loss of function mutation in $CYP27B1$, with reported minor allele frequency 0.67% in MS cases, and over transmitted in familial MS by the Ramagopalan et al. study (20), was not identified as a polymorphic site in the DNA pools. We genotyped this SNP in our MS and healthy control cohorts and none of the
samples were polymorphic for rs118204009. With the reported allele frequency we expected to observe 43 minor alleles in our MS cohort.

DISCUSSION

This study has confirmed the association with susceptibility to multiple sclerosis in individuals of North-Western European descent with polymorphisms in chromosome 12q13-14 previously reported by a genome-wide association study (13). Our fine-mapping study of this region suggests that the association is caused by a common polymorphism. No evidence of allelic heterogeneity or rare variant association was observed. We have shown that the region harbouring the association is under strong linkage disequilibrium and we have narrowed down the association to an 86.5kb region. We found 42 SNPs highly associated with susceptibility to MS in this region, all of which have strong pair-wise linkage disequilibrium. The study did narrow the disease-associated region down in comparison with our original study (13), and in comparison with the IMSGC/WTCCC2 study (14), where the disease-associated regions were 169.5 and 260.6 kb respectively. Moreover, in comparison with the genotyped and imputed SNPs we reported in our original study, this study increases the number of markers within the critical region that may be the true associated variant(s) from 3 to 42. We identified a stronger associated marker (rs2069502, \( P = 9.9 \times 10^{-11}, \text{OR} = 0.787 \)) than we had previously reported (rs703842, \( P = 8.5 \times 10^{-10}, \text{OR} = 0.796 \)). However, these markers both belonged to a single haplotype block of 42 SNPs, which are all in tight linkage disequilibrium with one another, and so the primary associated variant(s) at this locus could not be determined with certainty in the current study. The associations with SNPs in this block were considerably more significant than with the most significant SNP in the IMSGC/WTCCC2 study (rs12368653, \( P = 4.8 \times 10^{-7} \),
OR=0.84); though we expected to observe a stronger association with SNPs in higher LD with rs703842 than with rs12368653 as the sample overlap in the study and in the ANZgene GWAS was significant. Finally, we have confirmed that the association signal we observed is independent of the rare variant association reported in CYP27B1 in multi-case families with MS (20).

Our variant discovery analysis showed that we detected variants with high sensitivity, suggesting that the causal variant is likely to be one of these 42 variants. None of these variants is a missense mutation, and all were common (minor allele frequencies in controls between 29.2% and 34.9%). No low frequency or rare variants were shown to be associated with MS at this locus.

Thus, despite an extensive sequencing and genotyping study, we have been unable to pinpoint a directly disease linked variant at this locus. Possible explanations for this include, failure to identify and genotype the key associated variant(s), association with more than one disease-associated variant, or insufficient sample size. The variant discovery phase of our study had a high sensitivity to detect variants with population frequency similar to the frequency of the associated SNPs reported here, but it should be noted that we were unable to design genotyping assays for all variants detected in the region; hence failure to genotype key associated variants is a plausible explanation. The study was adequately powered (above 80%) to detect association with SNPs with MAF above 7% and effects size of 1.2 though underpowered to detect independent associations with lower population frequency of similar effect size to the SNPs reported by the seminal paper by Nejentsev, et al. (36) in type 1 diabetes.
Alternate approaches to dissecting this locus include expression QTL (eQTL) studies and tranethnic approaches. eQTL studies have shed some light on the likely variant, with one variant in this set having previously been reported to be associated with \textit{CYP27B1} expression in B lymphocytes (37), another correlating with serum levels of 1,25(OH)$_2$D (19), whilst a third associated with expression of \textit{FAM119B} on whole blood (27). Further eQTL studies investigating gene-expression in more relevant tissues with regard to MS pathogenesis may be valuable in further dissection of this locus.

To date it is not clear if polymorphisms in chr12q13-14 affect susceptibility to MS in other ethnic groups; if it does, a trans-ethnic gene mapping study could identify the causal variant (38-41). For example, the 42 SNP haplotype identified in this study has a weaker linkage disequilibrium pattern in the HapMap Chinese and Japanese populations (42), so it would be of interest to test these SNPs for association with MS in a large East Asian sample collection.

Our comparison between the two DNA enrichment approaches demonstrated better sensitivity measurements and MAF estimates with Nimblegen’s sequence capture compared to long-range PCR, but higher specificity for the long-range PCR method. With both techniques, variants that were not confidently called were predominantly singletons in the pool, even though sufficient power and coverage to detect the singletons was reached for a high proportion of the variants missed. Difficulty with singleton detection has previously been reported in similar studies in which pool sequencing was used (28, 31, 43, 44), where sensitivity has been reported to be as low as 66%. The limitations in variant discovery with pool sequencing are rapidly being overcome with indexed pool sequencing where non-uniform amplification of genomic samples do not affect the sensitivity of the variant discovery, a potential source of bias previously observed in long-range PCR pool sequencing (43). Our comparison between the sensitivity and
MAF estimates with these two techniques suggested our SeqCap pools suffered less from this imbalance than the PCR pools, as previously suggested elsewhere but not tested (44).

In conclusion, this study was able to reduce the key MS-associated region at locus chromosome 12q13-14 to 51% of the boundaries we originally reported, and to 33% of the region implicated by the much larger IMSGC/WTCCC2 MS study. We have likely identified all, or nearly all, of the potential true disease-associated variants. Nonetheless we were unable to pinpoint individual variants as being associated with disease. Our findings suggest a single key associated variant is responsible for the association at this locus. The extent of linkage disequilibrium across the locus suggests that much larger studies will be required to dissect the locus further, and therefore alternate approaches should be considered.

MATERIALS AND METHODS

Multiple sclerosis patients and healthy controls

175 cases were selected for sequencing from the 1,618 cases included in a previous genome-wide association study (13). These cases were selected to maximize haplotypic diversity. Genotypes for all 5,031 cases and controls in the GWAS were phased across a 500kb region on chromosome 12 using MACH (45); out of the 10,062 haplotypes there were 433 distinct haplotypes across the 288kb target region. There were 150 distinct haplotypes that occurred in a case and at least one other individual. One case carrying each of these haplotypes was selected for sequencing; the remaining 25 cases for sequencing were selected at random. 150 control samples were selected randomly from an Australian control sample collection. All cases and
controls provided informed written consent, and the study had been approved by the relevant ethics review boards.

The 3,269 MS cases genotyped for fine-mapping in stage 2 were all of Northern European ancestry are were part of discovery and replication arm of the ANZgene MS GWAS (13). From the MS case collection, 2,018 samples came from Australia and 1,251 from New Zealand. Average age of onset was 35.26, and female to male ratio was 2.72.

**DNA preparation and pooling**

Genomic (gDNA) DNA from Australian MS cases and controls was extracted from peripheral blood using kit based DNA extraction protocols. DNA from New Zealand MS patients was extracted from saliva self-collected into Oragene collection tubes as per the manufacturers’ instructions (DNA-Genotek). Extracted gDNA was quantified using Quant-iT™ Picogreen® (Invitrogen). DNA samples were diluted to 50 ng/µL and then pooled in equal amounts for each of the pooling strategies as described below.

**SeqCap Pools**

The sequence capture was performed using custom liquid phase sequence capture pools from Nimblegen according to the manufacturer's recommendations. Sequence capture efficiency was assessed by quantitative real-time PCR using a standard set of control primers recommended in the sequence capture protocol. Individual DNA libraries were quality-checked and quantified on Agilent 2100 Bioanalyzer using DNA1000 kit, and DNA concentration adjusted to 10 nM.

Sequencing was performed on the Illumina Genome Analyzer II using a standard 56 cycle paired-end read sequencing protocol and Illumina's sequencing reagents according to the
manufacturer's recommendations. Each library was sequenced individually on a single flow cell lane.

**PCR pools**

Genomic DNA regions corresponding to 288Kb were amplified using long range PCR with individual amplicon sizes ranging between 5 and 12Kb (primers available in the supplementary materials). For each patient, individual PCR fragments were pooled in equimolar concentration to form a single target region pool. Subsequently individual pools were combined into final pools, including 25 samples. Final pools were used as an input for the standard preparation of the Illumina's paired-end libraries. DNA quality and sequencing was performed as described above for the SeqCap pools.

**Variant discovery**

Base calling and sequence reads quality assessment was performed using Illumina’s Data Analysis Pipeline software v.1.6. Alignment sequence reads to the reference human genome (hg 18, UCSC assembly) was performed using the Burrows-Wheeler alignment (BWA) tool (46). Sequence alignment files conversions were performed using SAMtools (47). Reads quality score were recalibrated using Genome Analysis Toolkit (GATK) (48). Identification of SNVs was performed using SYZYGY (28) with default parameters. Variants were annotated using the ENSEMBL Perl API and custom scripts.

**Genotyping**

Sequenom iPLEX genotyping chemistry was used to genotype variations identified in our discovery set. Variations were selected randomly to capture a combination of non-synonymous
and synonymous variations within coding regions and variations identified in other non-coding genomic regions, and were based on whether they occurred within the high-confidence, medium confidence or both based on SYZYGY confidence ratings for DNA enrichment methods. PCR and extension primers were designed using the Sequenom MassArray ® designer. Assays that failed at either the design or genotyping phase were discarded, resulting in 150 assays being performed. Genotyping was performed as per the manufacturers’ instructions in 175 MS cases and 150 controls that had been used in the discovery phase.

Variants for the fine-mapping phase were selected from the variant discovery phase or from public databases and prioritized based on GoldenGate in silico assay design quality scores. From the discovery phase, variants were included if they were called with high confidence quality score in both PCR and SeqCap pools, for pools where samples were enriched with both strategies, or with high confidence in either PCR or SeqCap pools if the samples for those pools were not enriched with both strategies. From public databases, variants present in the target region in dbSNP 131 or on the Immunochip (32) and not identified in the discovery phase were included for design. Genotyping was done using a 1,536 custom GoldenGate assay. Assay design and genotyping was performed according to the manufacturer’s protocol. A total of 954 were designed in the target region and visual inspection of intensity plots identified 242 failed assays.

**Association Analysis**

Association analysis was performed using the allelic test chi-square (1df) test in PLINK v. 1.07 (49), GENECLUSTER (34) and custom scripts in R (http://www.R-project.org).
Genotype Imputation

The 6,846 cases and controls were phased and imputed with 1,000 Genomes Project data (European panel, 2011-06 haplotypes) using mach and minimac (45). Only good quality imputed SNPs were analysed (RSQ > 0.5) and association analysis on imputed data was performed with mach2data using dosage genotypes.

ACKNOWLEDGEMENTS

We thank individuals with MS in Australia and New Zealand for supporting this research and all investigators of the study who have contributed to the recruitment of MS patients. Cases and controls for this study were recruited within the Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). This work was supported by the National Health and Medical Research Council of Australia (NHMRC) [grant number 633275].

Conflict of Interest statement. None declared.
References


Legends to Figures

Figure 1: Minor allele frequencies estimated by read counts in pool sequencing libraries compared to microarray genotypes. \( \rho \) = Pearson correlation between sequence and genotyped SNPs.

Figure 2: Linkage disequilibrium (\( r^2 \)) blocks containing the 48 SNPs found to be strongly associated (\( P \)-value < 1\( \times \)10\(^{-8} \)) with multiple sclerosis susceptibility. Block 3 (boxed) contains lead SNP rs2069502 and all markers identified by GENECLUSTER analysis.

Figure 3: SNP association plot for multiple sclerosis susceptibility in target region chromosome 12q13-14. Significance of association is plotted on the \( y \)-axis (\( -\log_{10}(P\text{-value}) \)) for all SNPs genotyped (a) and after conditioning on lead SNP rs2069502 genotype (b).
### Tables

**Table 1** SNPs identified in the target region. dbSNP (%) percentage of detected SNPs present in dbSNP version 131; NS/S - non-synonymous/synonymous SNP ratio; Ti/Tv transition/transversion ratio.

<table>
<thead>
<tr>
<th></th>
<th>Long-range PCR</th>
<th>Sequence Capture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High</td>
<td>Medium</td>
</tr>
<tr>
<td>Counts</td>
<td>1,173</td>
<td>179</td>
</tr>
<tr>
<td>dbSNP(%)</td>
<td>31</td>
<td>11</td>
</tr>
<tr>
<td>NS/S</td>
<td>1.45</td>
<td>2.33</td>
</tr>
<tr>
<td>Ti/Tv</td>
<td>2.21</td>
<td>1.46</td>
</tr>
</tbody>
</table>
**Table 2** Classification of the high confidence SNPs.

<table>
<thead>
<tr>
<th>Type</th>
<th>All</th>
<th>Novel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergenic</td>
<td>124 (27.6%)</td>
<td>37 (18.1%)</td>
</tr>
<tr>
<td>Intronic</td>
<td>273 (60.7%)</td>
<td>143 (70.1%)</td>
</tr>
<tr>
<td>Synonymous coding</td>
<td>18 (4.0%)</td>
<td>5 (2.5%)</td>
</tr>
<tr>
<td>Non synonymous coding</td>
<td>12 (2.7%)</td>
<td>7 (3.4%)</td>
</tr>
<tr>
<td>Splice site</td>
<td>4 (0.9%)</td>
<td>1 (0.5%)</td>
</tr>
<tr>
<td>3' prime UTR</td>
<td>15 (3.3%)</td>
<td>9 (4.4%)</td>
</tr>
<tr>
<td>5' prime UTR</td>
<td>3 (0.7%)</td>
<td>1 (0.5%)</td>
</tr>
<tr>
<td>Stop gained</td>
<td>1 (0.2%)</td>
<td>1 (0.5%)</td>
</tr>
<tr>
<td>Total</td>
<td>450</td>
<td>204</td>
</tr>
</tbody>
</table>
Table 3 Sensitivity analysis on high confidence SNPs. Numbers in parentheses represents all SNPs called (high + lower quality variants).

<table>
<thead>
<tr>
<th>Pool</th>
<th>Immuochip SNPs polymorphic</th>
<th>Library</th>
<th>SNPs Detected</th>
<th>Sensitivity</th>
<th>MAF Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td>268</td>
<td>PCR</td>
<td>192 (238)</td>
<td>0.797 (0.888)</td>
<td>0.871 (0.808)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SeqCap</td>
<td>220 (241)</td>
<td>0.868 (0.899)</td>
<td>0.916 (0.897)</td>
</tr>
<tr>
<td>Controls</td>
<td>273</td>
<td>PCR</td>
<td>199 (221)</td>
<td>0.754 (0.809)</td>
<td>0.718 (0.702)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SeqCap</td>
<td>217 (243)</td>
<td>0.855 (0.890)</td>
<td>0.923 (0.893)</td>
</tr>
</tbody>
</table>
Table 4 Association analysis and GENECLUSTER results on the 24 genotyped SNPs found to be highly correlated to the inferred causal mutation in the GENECLUSTER analysis.

<table>
<thead>
<tr>
<th>SNP ID</th>
<th>Position</th>
<th>MAF Cases/Controls</th>
<th>GENECLUSTER Correlation</th>
<th>Association analysis $P$-value</th>
<th>OR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs10877011</td>
<td>56411259</td>
<td>0.31/0.35</td>
<td>0.912</td>
<td>$2.3 \times 10^{-6}$</td>
<td>0.84 (0.78-0.90)</td>
</tr>
<tr>
<td>rs2069502</td>
<td>56430932</td>
<td>0.29/0.34</td>
<td>0.965</td>
<td>$9.9 \times 10^{-11}$</td>
<td>0.79 (0.73-0.85)</td>
</tr>
<tr>
<td>rs4646536</td>
<td>56444255</td>
<td>0.29/0.33</td>
<td>0.959</td>
<td>$1.7 \times 10^{-9}$</td>
<td>0.8 (0.74-0.86)</td>
</tr>
<tr>
<td>rs703842</td>
<td>56449006</td>
<td>0.29/0.34</td>
<td>0.939</td>
<td>$8.5 \times 10^{-10}$</td>
<td>0.8 (0.74-0.86)</td>
</tr>
<tr>
<td>chr12_56449568</td>
<td>56449568</td>
<td>0.25/0.29</td>
<td>0.909</td>
<td>$2.3 \times 10^{-8}$</td>
<td>0.81 (0.75-0.87)</td>
</tr>
<tr>
<td>rs10877013</td>
<td>56451352</td>
<td>0.29/0.33</td>
<td>0.956</td>
<td>$9.2 \times 10^{-9}$</td>
<td>0.81 (0.75-0.87)</td>
</tr>
<tr>
<td>rs2291617</td>
<td>56452670</td>
<td>0.29/0.33</td>
<td>0.96</td>
<td>$2.1 \times 10^{-9}$</td>
<td>0.8 (0.74-0.86)</td>
</tr>
<tr>
<td>rs10877014</td>
<td>56453928</td>
<td>0.29/0.34</td>
<td>0.955</td>
<td>$1.7 \times 10^{-10}$</td>
<td>0.79 (0.73-0.85)</td>
</tr>
<tr>
<td>rs10877015</td>
<td>56454055</td>
<td>0.28/0.33</td>
<td>0.962</td>
<td>$1.1 \times 10^{-9}$</td>
<td>0.8 (0.74-0.86)</td>
</tr>
<tr>
<td>rs11172333</td>
<td>56456602</td>
<td>0.28/0.33</td>
<td>0.963</td>
<td>$1.4 \times 10^{-9}$</td>
<td>0.8 (0.74-0.86)</td>
</tr>
<tr>
<td>rs10877019</td>
<td>56459196</td>
<td>0.28/0.33</td>
<td>0.964</td>
<td>$1.2 \times 10^{-9}$</td>
<td>0.8 (0.74-0.86)</td>
</tr>
<tr>
<td>rs923829</td>
<td>56460573</td>
<td>0.29/0.33</td>
<td>0.958</td>
<td>$2.7 \times 10^{-9}$</td>
<td>0.8 (0.75-0.86)</td>
</tr>
<tr>
<td>rs11172335</td>
<td>56461468</td>
<td>0.28/0.33</td>
<td>0.961</td>
<td>$7.9 \times 10^{-10}$</td>
<td>0.8 (0.74-0.86)</td>
</tr>
<tr>
<td>rs10431552</td>
<td>56464415</td>
<td>0.29/0.33</td>
<td>0.96</td>
<td>$3.8 \times 10^{-10}$</td>
<td>0.79 (0.74-0.85)</td>
</tr>
<tr>
<td>rs724834</td>
<td>56466683</td>
<td>0.28/0.33</td>
<td>0.961</td>
<td>$6.1 \times 10^{-10}$</td>
<td>0.79 (0.74-0.85)</td>
</tr>
<tr>
<td>rs11172342</td>
<td>56474025</td>
<td>0.28/0.33</td>
<td>0.961</td>
<td>$4.8 \times 10^{-10}$</td>
<td>0.79 (0.74-0.85)</td>
</tr>
<tr>
<td>rs11172343</td>
<td>56474424</td>
<td>0.28/0.33</td>
<td>0.962</td>
<td>$8.5 \times 10^{-10}$</td>
<td>0.8 (0.74-0.86)</td>
</tr>
<tr>
<td>rs11172344</td>
<td>56479715</td>
<td>0.29/0.33</td>
<td>0.959</td>
<td>$1.8 \times 10^{-9}$</td>
<td>0.8 (0.74-0.86)</td>
</tr>
<tr>
<td>rs10783848</td>
<td>56482795</td>
<td>0.29/0.33</td>
<td>0.957</td>
<td>$6.2 \times 10^{-10}$</td>
<td>0.79 (0.74-0.85)</td>
</tr>
<tr>
<td>rs1875124</td>
<td>56483069</td>
<td>0.28/0.33</td>
<td>0.961</td>
<td>$5.8 \times 10^{-10}$</td>
<td>0.79 (0.74-0.85)</td>
</tr>
<tr>
<td>rs3825079</td>
<td>56489621</td>
<td>0.28/0.33</td>
<td>0.96</td>
<td>$7.6 \times 10^{-10}$</td>
<td>0.79 (0.74-0.86)</td>
</tr>
<tr>
<td>chr12_56490697</td>
<td>56490697</td>
<td>0.28/0.33</td>
<td>0.96</td>
<td>$7.6 \times 10^{-10}$</td>
<td>0.8 (0.74-0.86)</td>
</tr>
<tr>
<td>rs11172349</td>
<td>56498795</td>
<td>0.28/0.33</td>
<td>0.958</td>
<td>9.7 x10^{-10}</td>
<td>0.8 (0.74-0.86)</td>
</tr>
<tr>
<td>rs11172351</td>
<td>56499752</td>
<td>0.28/0.33</td>
<td>0.957</td>
<td>5.0 x10^{-10}</td>
<td>0.79 (0.74-0.85)</td>
</tr>
</tbody>
</table>