Identification of Autoimmune Gene Signatures in Autism

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>doi:10.1038/tp.2011.62</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:10288460">http://nrs.harvard.edu/urn-3:HUL.InstRepos:10288460</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Identification of autoimmune gene signatures in autism

J-Y Jung1, IS Kohane1,2,3 and DP Wall1,4

The role of the immune system in neuropsychiatric diseases, including autism spectrum disorder (ASD), has long been hypothesized. This hypothesis has mainly been supported by family cohort studies and the immunological abnormalities found in ASD patients, but had limited findings in genetic association testing. Two cross-disorder genetic association tests were performed on the genome-wide data sets of ASD and six autoimmune disorders. In the polygenic score test, we examined whether ASD risk alleles with low effect sizes work collectively in specific autoimmune disorders and show significant association statistics. In the genetic variation score test, we tested whether allele-specific associations between ASD and autoimmune disorders can be found using nominally significant single-nucleotide polymorphisms. In both tests, we found that ASD is probabilistically linked to ankylosing spondylitis (AS) and multiple sclerosis (MS). Association coefficients showed that ASD and AS were positively associated, meaning that autism susceptibility alleles may have a similar collective effect in AS. The association coefficients were negative between ASD and MS. Significant associations between ASD and two autoimmune disorders were identified. This genetic association supports the idea that specific immunological abnormalities may underlie the etiology of autism, at least in a number of cases.

Translational Psychiatry (2011) 1, e63; doi:10.1038/tp.2011.62; published online 13 December 2011

Introduction

Autism spectrum disorder (ASD) is a broad spectrum of early-onset neuropsychiatric disorders characterized by severe deficits in social interaction and language, and the presence of repetitive and stereotyped behaviors and interests. Twin studies have demonstrated that ASD is largely genetic, with a 90% concordance between monozygotic twins, andheritable, with a 5–10 times higher familial risk than in the general population.2,3 Genome-wide association studies (GWASs) have discovered significant genetic markers of single-nucleotide polymorphisms (SNPs) and de novo mutations of copy number variations that may cause ASD. However, the molecular pathology of ASD is largely unclear owing to its genetic heterogeneity and the fact that only a small proportion of incidence is explained by known susceptibility loci. Given this heterogeneity, it has been suggested that cross-disease analysis between ASD and other disorders that share common phenotypic symptoms or genetically susceptible loci will shed light on our understanding of the molecular mechanisms underlying ASD.

The role of the immune system in neuropsychiatric diseases, including ASD, has long been hypothesized and is mainly supported by family cohort studies and the immunological abnormalities found in autistic patients. For example, Atladóttir et al. reported that the risk of ASD increases when a child’s mother has rheumatoid arthritis or has a family history of type 1 diabetes. Other autoimmune disorders for which epidemiological studies have shown significant association with ASD, include maternal psoriasis, maternal ulcerative colitis, and autoimmune thyroid disease (ATD). In addition, various forms of immune dysregulation, including elevated cytokine levels and increased immunoglobulin and serum protein levels, have been identified in autistic children. However, few studies have reported potential genetic components that account for associations between ASD and autoimmune disorders. In fact, only three alleles, two in the human leukocyte antigen (HLA-DR4, HLA-A2) and one in the major histocompatibility complex (MHC) of chromosome 6 (the C4B null allele) have a confirmed association with ASD. With the advent of GWAS and the availability of large amounts of genotype data from both ASD and multiple autoimmune diseases, we now have the ability to directly analyze genetic associations across these diseases.

Materials and methods

Study samples and data quality control. We obtained Illumina BeadChip 550K genotype data of 941 multiplex families (Illumina Inc., San Diego, CA, USA) with autistic children from the Autism Genetic Resource Exchange (AGRE). For the phenotype labeling, we followed the classification of the Autism Diagnostic Interview—Revised included in the AGRE phenotype database, but excluded

1Center for Biomedical Informatics, Harvard Medical School, Boston, MA, USA; 2Informatics Program, Children’s Hospital, Boston, MA, USA; 3I2b2 National Center for Biomedical Computing, Boston, MA, USA and 4Division of Genomic Medicine, Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA, USA

Correspondence: Dr DP Wall, Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, MA 02115, USA.

E-mail: dpwall@hms.harvard.edu

Keywords: autism spectrum disorder; autoimmune disorder; cross-disease analysis; genetic variation score; polygenic score

Received 5 August 2011; revised 4 November 2011; accepted 5 November 2011
individuals whose Autism Diagnostic Interview—Revised classification and Autism Diagnostic Observation Schedule classification did not agree. We examined each family’s data to find monozygotic twins, triplets, or quadruplets and elected to include only one monozygotic sibling per family, filtering 72 individuals from the data set. We also removed 92 individuals who were annotated as ‘possible non-idiopathic autism’ in the AGRE phenotype database. These cases include prematurity of less than 35 weeks gestation (45 individuals), Fragile X syndrome (12 individuals), known chromosomal abnormality (10 individuals), and other diagnosed neurogenetic disorders (10 individuals). We further applied a set of quality control filters in order to identify a stringent subset of robust SNPs. We excluded 32 individuals with genotyping rate <0.95 and 18510 SNPs with genotyping rate <0.95. We examined Mendelian error per each family trio, and removed seven individuals for Mendelian error >1 percent of all markers. SNPs with a minor allele frequency <0.05 or with a Hardy–Weinberg equilibrium exact test \( P < 0.001 \) were excluded from further analysis. After quality control, the ASD genomic data consisted of 1397 affected trios and a total of 470025 SNPs. We used the software package PLINK \(^{34}\) to conduct the transmission disequilibrium test (TDT) with all family trios passing quality control without identifying subpopulation, as the TDT is known to maintain the desired type 1 error rate in the presence of population stratification.\(^{39}\)

We obtained two groups of GWAS data from the Wellcome Trust Case Control Consortium (WTCCC) as target disease sets for measuring association with ASD. The smaller set (WTCCC4) consisted of 1500 common controls and 1000 independent cases of ankylosing spondylitis (AS), autoimmune thyroid disease, multiple sclerosis (MS), and breast cancer. It had 14436 non-synonymous SNPs plus 897 SNPs in the major histocompatibility complex region, and was genotyped using the Illumina Infinium 15K array. The larger set (WTCCC7) in the Affymetrix GeneChip 500K array (Affymetrix Inc., Santa Clara, CA, USA) consisted of 3000 shared controls and 2000 independent cases in seven diseases, including bipolar disorder, coronary artery disease, Crohn’s disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes. Shared control samples came from two sources: for the smaller set (WTCCC4), control samples were taken from the 1958 British Birth Cohort (58C) and for the larger set (WTCCC7), 1500 control samples were taken from 58C. The remaining 1500 samples were taken from blood donors recruited by the three UK Blood Services (UKBS).\(^{36,37}\) Although the ages of the former are known and past the typical age of onset for the autoimmune conditions studied, the individual age information of the UKBS group is not known. The potential variation in ages from this control group could introduce minor classification bias, as some of individuals may (or may have) develop the autoimmune disorder in the future, as discussed in the main WTCCC7 article.\(^{37}\) However, the prevalence of these autoimmune disorders is relatively rare, with a combined prevalence of rate of up to 8 percent.\(^{38}\) Therefore, although we do not have detailed age information for the UKBS samples, we assume that misclassification due to age of onset would be rare and unlikely to bias our results. We followed the quality control steps described in Burton et al.,\(^{36,37}\) and further removed SNPs with minor allele frequency <0.05 or Hardy–Weinberg equilibrium exact test \( P < 0.001 \). After quality control steps, there were 12700 SNPs remaining in the WTCCC4 set, and 469557 SNPs were retained for further analysis in the WTCCC7 set. Across all SNPs passing quality control, we used case–control association analysis with the software package PLINK.

As these data were genotyped using different platforms, we cross-referenced SNP ids and strand information as follows. First we converted all custom, non-reference SNP ids into corresponding reference SNP ids by querying the UCSC Genome Browser (http://www.genome.ucsc.edu; version May 2004/NCBI genome build 35). Then we examined all reference SNP ids to confirm that they are up-to-date by querying NCBI dbSNP (build 132) and removed six SNPs that had multiple identifiers. Next, we checked all the allelic and chromosomal position information of each data set with the HapMap genotype data (CEU founders, release 23), reconfiguring the strands when alleles did not match. Mapping SNPs into the corresponding gene regions was done by querying NCBI dbSNP (build 132). The number of intersection SNPs between ASD and WTCCC4 set was 5318, and the number of intersection SNPs between ASD and WTCCC7 set was 73331.

**Polygenic score analysis.** To identify genomic associations between diseases, we used two separate and complementary approaches, as illustrated in Figure 1. The first, the polygenic score (PS) test measured the collective effect of disease-associated SNPs from one disease on a collection of SNPs from another. The test was designed to identify associations between complex multigenic diseases that manifest through a combination of multiple variants of small individual effect. We used the PS test here to determine if collections of variants from cases with ASD appear to correlate with groups of variants from individuals with autoimmune disorders. First, we labeled a GWAS data set from one disease as the ‘source’ data set and another GWAS from a different disease as the ‘target’. From the source data set we selected groups of autosomal SNPs that corresponded to a range of nominal significance thresholds \((P_{\text{adj}})\) as source alleles, and recorded the minor allele and odds ratio information of those selected SNPs. Then, for each individual in the target data set, we calculated the PS by computing the average number of source alleles that the individual had, weighted by the logarithm-of-the-odds-ratio \((\log(\text{odds-ratio}))\) from the source data set. This polygenic score can be considered to be a measure of probabilistic similarity for those SNPs in the source disease data set and each individual in the target data set. As such, if the source and target diseases were highly similar in allelic composition, the cases would yield consistently higher scores than the controls. If there were multiple target diseases, ones that were more closely related to the source disease would yield a higher average PS, as demonstrated in International Schizophrenia Consortium.\(^{39}\) To provide this context, we ran logistic regression analysis with the PS to predict the classification of the target disease. Then, we estimated the variance explained in the target disease data by the PS using the Nagelkerke’s pseudo \(R^2\) from a model with the PS and covariates, versus that from a model without the score. We took the total number of alleles used to calculate
the PS and the numerically coded site information of individuals in the WTCCC data sets as covariates.

**Genetic variation score analysis.** The second metric we used to compare ASD to autoimmune disorders was the genetic variance score (GVS). With GVS, a combined score of the odds-ratio and \( P \)-value was defined for each disease data set and every SNP belonging to that set. Specifically, for each disease data set \( d \) and a SNP \( s \in d \), a GVS of \([d, s]\) was defined as \( \log(\text{odds-ratio} [d, s]) \times (-\log_{10}(P\text{-value} [d, s])) \). Given that the odds-ratio was calculated with respect to the minor allele of the SNP and that the odds-ratio was greater than 1 when the minor allele was more likely to occur in the case group, the sign (log (odds-ratio)) was positive when the minor allele was the risk allele, and negative when the major allele was the risk allele. When the GVS scores of a SNP in two diseases had different signs, we assumed that the allele was protective in one disease and deleterious in the other. As an extension of this allele-specific comparison, we computed the Pearson pairwise correlation between the GVS data of two diseases. This correlation analysis enabled us to determine whether large numbers of SNPs have similar effects in two different diseases, that is, a high positive correlation coefficient indicated similar effects, whereas a strong negative coefficient indicated opposing effects of the risk alleles.

Although both of these two approaches were based on individual SNP association and framed in terms of odds-ratios and \( P \)-values, they provided two complementary assessments of the association between the genotypes of two different diseases. The PS test assigned a score per individual in the target disease set and assumed that many weakly associated SNPs with marginal odds-ratios may work collectively, such that they have stronger association test statistics than loci drawn from the null distribution. The GVS and its Pearson correlation as a similarly measure examined whether there were SNPs with strong significance in both disease sets, as GVS compared the pair-wise, allele-specific significance of each SNP.

**Results**

**PS analysis.** We examined whether groups of ASD-associated SNPs can collectively account for genomic variation in another disease, even when each individual SNP may not have a very strong effect. From this source sample, we selected sets of marginally associated alleles at five different \( P \)-value thresholds (\( P_T \), a set of SNPs with \( P < P_T \) were selected as ‘source alleles’. Then for each target disease data set, the polygenic score was calculated for each individual. We conducted logistic regression analysis with the polygenic scores in order to examine whether source alleles from ASD data could explain variances in the target disease data sets. For each ASD and autoimmune data set, the genetic variation score was calculated for each SNP. We used Pearson’s pair-wise correlation to compare the collective effect of SNPs in two diseases.

**Genetic variation score analysis.** Next, we examined whether allele-specific associations can be found among
Identification of autoimmune gene signatures in autism
J-Y Jung et al

ASD and autoimmune disease data sets via the GVS approach. As we examined the collective effect of marginally significant SNPs and found enrichment in two autoimmune diseases in the PS approach, here we focused our analysis on SNPs with at least nominally significant P values (P < 0.05). Multiple hypothesis correction was not applied to the P values when calculating GVS because we examined associations of disease pairs by correlation coefficients, which would not change after multiple hypothesis correction. Table 2 shows the Pearson correlation coefficients for all disease pairs within ASD and the WTCCC data sets, calculated with GVS. Consistent with the results from the PS analysis above, the ASD data exhibited significant positive association with AS (coefficient 0.4032) and significant negative association with MS (coefficient −0.3092). Given that the magnitude of the association between ASD and autoimmune thyroid disease was slightly positively correlated with ASD (coefficient 0.2012), whereas all other autoimmune diseases (MS, Crohn’s disease and rheumatoid arthritis) showed little association with ASD in their profiles. Table 3 summarizes the nominally significant SNPs in both ASD and the autoimmune diseases, including AS and MS. Many of these SNPs fell within already known ASD risk genes (for example, rs2034648—AGAP1) or mental disorder susceptibility genes (for example, rs11643718—SLC12A3 and rs3132468—MICB). Another group of SNPs fell within the protein coding regions of several known autoimmune risk genes (for example, rs3135559—PSORS1C1, rs3129943—C6orf10 and rs3130542—HLA-C). However, none of these have been reported previously as ASD susceptibility genes.

### Table 1 Logistic regression analysis results of ASD and WTCCC disease data

<table>
<thead>
<tr>
<th>Source: ASD</th>
<th>Target sample pseudo-$R^2 \times 100$ ($P$ value)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>WTCCC 4</strong></td>
<td></td>
</tr>
<tr>
<td>AS</td>
<td>0.21 (1.44e–07)</td>
</tr>
<tr>
<td>ATD</td>
<td>0.01 (1.99e–04)</td>
</tr>
<tr>
<td>BC</td>
<td>0.00 (2.29e–11)</td>
</tr>
<tr>
<td>MS</td>
<td>0.07 (2.79e–01)</td>
</tr>
<tr>
<td><strong>WTCCC 7</strong></td>
<td></td>
</tr>
<tr>
<td>BD</td>
<td>0.33 (1.54e–02)</td>
</tr>
<tr>
<td>CAD</td>
<td>0.01 (9.27e–01)</td>
</tr>
<tr>
<td>CD</td>
<td>0.11 (2.64e–01)</td>
</tr>
<tr>
<td>HT</td>
<td>0.03 (7.10e–01)</td>
</tr>
<tr>
<td>RA</td>
<td>0.05 (5.69e–01)</td>
</tr>
<tr>
<td>T1D</td>
<td>0.11 (2.34e–01)</td>
</tr>
<tr>
<td>T2D</td>
<td>0.04 (6.34e–01)</td>
</tr>
</tbody>
</table>

Abbreviations: AS, ankylosing spondylitis; ASD, autism spectrum disorder; ATD, autoimmune thyroid disease; BC, breast cancer; BD, bipolar disorder; CAD, coronary artery disease; CD, Crohn’s disease; HT, hypertension; MS, multiple sclerosis; $P_t$, P-value threshold; RA, rheumatoid arthritis; T1D, type 1 diabetes; T2D, type 2 Diabetes; WTCCC, Wellcome trust case control consortium.

### Table 2 Pairwise GVS correlation coefficients between ASD and WTCCC data sets

<table>
<thead>
<tr>
<th></th>
<th>AS</th>
<th>ATD</th>
<th>BC</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GVS$^1$ (#SNPs)</strong></td>
<td>0.4032 (28)</td>
<td>0.2012 (34)</td>
<td>0.0254 (16)</td>
<td>−0.3092 (36)</td>
</tr>
<tr>
<td><strong>GVS$^1$ (#SNPs)</strong></td>
<td>−0.0895 (276)</td>
<td>−0.0868 (267)</td>
<td>−0.0125 (288)</td>
<td>−0.0360 (250)</td>
</tr>
<tr>
<td><strong>GVS$^1$ (#SNPs)</strong></td>
<td>−0.0588 (258)</td>
<td>0.0127 (240)</td>
<td>−0.0546 (261)</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: AS, ankylosing spondylitis; ASD, autism spectrum disorder; ATD, autoimmune thyroid disease; BC, breast cancer; BD, bipolar disorder; CAD, coronary artery disease; CD, Crohn’s disease; GVS, genetic variation score; HT, hypertension; MS, multiple sclerosis; RA, rheumatoid arthritis; SNPs, single-nucleotide polymorphisms; T1D, type 1 diabetes; T2D, type 2 diabetes; WTCCC, Wellcome trust case control consortium. Number of the common SNPs in both diseases with $P<0.05$ is shown in parenthesis.

**Discussion**

Given that the magnitude of the association between ASD and the two autoimmune diseases, AS and MS, was either greater than or on par with the strength of association between what are considered now to be genetically similar autoimmune diseases, coupled with the lack of any other association of the same significance between ASD and the remaining autoimmune disorders examined, our results clearly demonstrated that there are true genomic links between ASD and the two autoimmune diseases, links that likely can inform our understanding of the genetics and treatments of ASD. However, further study and verification is required to characterize and explain these particular genomic associations. An interesting, albeit anecdotal, similarity between ASD and AS is that they both have an appreciable male bias. AS has a male-to-female ratio of approximately 2.5:1, while the male bias of ASD’s 4:1 while other autoimmune diseases, such as MS, also have a higher male bias of approximately 2.5:1, which is of the same magnitude as the male bias of ASD’s 4:1 while other autoimmune diseases. The male bias of ASD is higher than in other psychiatric disorders, such as BD (4:1), while other autoimmune diseases exhibit ratios of approximately 2.5:1.
The authors declare no conflict of interest.

Acknowledgements. We thank Drs Vincent Fusararo, Jike Cui and Todd F DeLuca for their helpful discussions while conducting the analyses. We also thank the anonymous reviewers for their insightful reviews that strengthened our manuscript. This work was supported by the National Institutes of Health under Grant No. 1R01MH090611-01A1 awarded to DPW.


