Supplementary Material

Predicting the diagnosis of Autism Spectrum Disorder using Gene Pathway Analysis

Efstratios Skafidas[1], Renee Testa[2,3], Daniela Zantomio[4], Gursharan Chana[5], Ian P. Everall[5], Christos Pantelis*[,2,5]

[1] Centre for Neural Engineering, The University of Melbourne, Parkville, Victoria, Australia

[2] Melbourne Neuropsychiatry Centre, Department of Psychiatry, The University of Melbourne & Melbourne Health, Parkville, Victoria, Australia

[3] Department of Psychology, Monash University, Clayton, Vic, Australia

[4] Department of Haematology, Austin Health, Heidelberg, Vic, Australia

[5] Department of Psychiatry, The University of Melbourne, Parkville, Victoria, Australia

* corresponding author: Prof Christos Pantelis, National Neuroscience Facility (NNF), Level 3, 161 Barry Street, Carlton South, Vic 3053, Australia; email: cpant@unimelb.edu.au

Table of Contents

S1 - Gene set enrichment analysis (GSEA)

GSEA was undertaken to consider all possible genes related to pathways that might contribute to risk for autism. We were interested to examine the contribution of multiple SNPs to risk for autism, each with potentially small effect, rather than seek to identify individual or small numbers of SNPs of large effect. The latter approach, while providing some information about the genes contributing to autism, has failed to provide any ability to predict which individuals may be at risk. The approach we have taken is to identify which of the known pathways are perturbed in ASD (using KEGG canonical pathways). Here, instead of attempting to identify significance for individual SNPs or genes, we sought to identify canonical pathways that differed compared with control subjects. This has the benefit of taking into account the complex interactions of genes, and since this approach is analyzing a much smaller number of sets it considerably increases the power of our analyses.

The collections of SNPs on the Illumina platform relevant to particular pathways were compared and we determined if the SNPs related to these pathways were perturbed. Data for the AGRE cohort provided SNP information from the Illumina 550 platform. The other datasets (HAPMAP, SFARI, Wellcome Trust) provided SNP data from the Illumina 1M and 1M-Duo. The total number of SNPs consistent across the three platforms was 407,420. The number of KEGG Pathway genes examined was 5,936. For each Kegg pathway, we determined the collection of SNPs residing on genes that form part of the pathway. This was performed by firstly identifying all genes that reside on a pathway. NCBI data mapping a SNP to a gene (as described in NCBI table SNPContigLocusId -

http://www.ncbi.nlm.nih.gov/projects/SNP/snp_db_table_description.cgi?t=SNPContigLocusId) was used to identify all SNPs relevant to a pathway. This included both intronic and exonic SNPs.

The *p*-values relevant to the pathways were calculated using permutation analysis. This was necessary as we were testing a set of SNPS not a single SNP. The set-based association analysis procedure used was as described in plink (http://pmgu.mgh.harvard.edu; see also Purcell 2007, American Journal of Human Genetics, 81.). In brief, SNPs that were in linkage disequilibrium (LD) above a certain threshold were removed, so to identify independent SNPs with the highest loadings. The statistic for each set was calculated as the mean of the single SNP statistics and the dataset was permuted at least 2,000,000 times while keeping LD between SNPs constant. For the two million permutations, the listed p-value was the number of times the permuted set statistic exceeded the mean p-value for that set.

All available Kegg pathways were tested. Only those showing statistical significance were retained at $p<10^{-5}$. The significance threshold of $p<10^{-5}$ was set according to the number of pathways being examined, which was 200. Therefore significance was $<0.05/200$ (set at $< 1x\ 10^{-5}$).

Pathways were then determined. Post pathway analysis, only SNPs that were part of the significant pathways were considered. We then tested whether the collection of SNPs relevant to each pathway was more or less represented in ASD individuals versus controls. 775 SNPs were identified as being statistically significant SNPs.

S2 – Selection of informative SNPs

Identification of SNPs for the analysis was determined as follows:

The 775 SNPs identified from the pathway analysis step were then examined further to determine which SNPs were most relevant to discriminating the groups.

A linear classifier forms a hyperplane on the feature space of variables separating the two classes (ASD subjects versus controls). SNPs that have the greatest mean difference between populations are good candidate SNPs for group separation. In this analysis, Bonferroni correction was used, with p value set at 0.05/775, which was rounded down to 1×10^{-5}.

This procedure, however, does not ensure that the identified set of SNPs are linearly independent. In order to address such collinearity, the covariance matrix of SNPs was calculated as were all covariance matrices with one SNP removed iteratively. The covariance matrix is a real symmetric matrix, which mandates that the eigenvalues of the matrix are greater than or equal to zero. Using a property of linear algebra, namely that the trace of the matrix is equal to the sum of the eigenvalues, the contribution to the total variance of each SNP was determined by removing the SNP and calculating the difference between the trace of the covariance matrix with and without that SNP. The SNPs that contributed least to the trace of the covariance matrix were thereby removed. This process was continued until the covariance matrix was full rank. In this way the remaining 237 SNPs were not linearly dependent on each other.

It should be noted that the SNP weights were not assumed to be Gaussian. The distributions of the weights for each of the SNPs were also examined by taking

random subsamples of individuals and their genetic data, which were used to train the classifier, providing weights for each SNP with each training set. This was iteratively run 100,000 times and a histogram of the weights for each SNP was plotted. This permitted us to examine the distribution of the weights for each SNP, allowing the confidence interval for each SNP to be determined.

It should be noted that the purpose of the present study was to build a classifier for the identification of ASD subjects, as detailed above. Therefore, with regard to the SNPs that have been discarded, while these may be informative when examining the question of etiology, that question is not the subject of the present study. In future work it will be important to examine the SNPs that were discarded. However, the analysis also provided information identifying genes that conferred greater risk or resilience for ASD.

S3 - Formula for the classifier & classifier performance

(a) *Formula for the classifier*

$$Y^j = \sum_{i=1}^{N} W_i SNP_i^j + W_0$$

where Y^j, W_i, SNP_i^j correspond to the weighted output for individual j, regression coefficient weight for each SNP for each individual respectively.

That is, the sum of weight W_i x SNP_i^j {'0,1,3'} value of the relevant allele + an offset w_0 (determined by the least squares analysis).

Therefore, the weighting can be negative, so that a more deleterious effect is not necessarily assumed to be related to the minor allele. It can be either the least or the most deleterious, and the off-set can also change the contribution of those SNPs to the clinical phenotype.

As stated in the main text, an affected individual was given a value of 10 and an unaffected individual a value of -10, to provide a sufficiently large separation to maximize the distance between means. Thus, given the formula above:

Let C denote the group of controls and let A denote the group of affected individuals and W the weight vector which is defined as above W_i

The mean of each group can be shown to be equal to

$E_{j\in C}(Y^j)=\sum_{I=}^{N} W_i . E_{j\in C}(SNP_i^j) + W_0$

$E_{j\in A}(Y^j)=\sum_{I=}^{N} W_i . E_{j\in A}(SNP_i^j) + W_0$

The distance between the two means of the distribution is given by

$d = (\mu_C^T - \mu_A^T)(\mu_C - \mu_A)$

where,

$\mu_C = E_{j\in C}(SNP_i^j)$ and $\mu_A = E_{j\in A}(SNP_i^j)$.

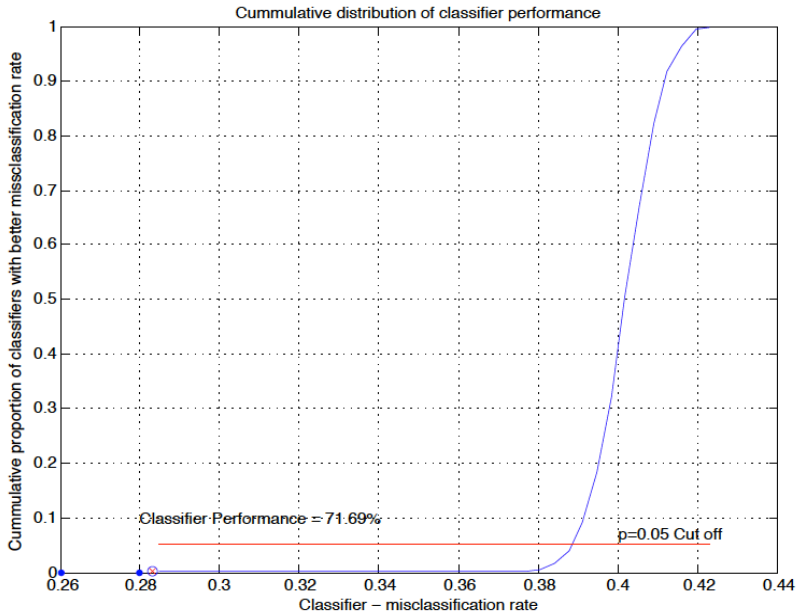
The optimal weight vector W , which maximizes the distance between the two groups, can be shown to correspond to the eigenvector with the maximum eigenvalue to the matrix

$M = (\mu_C - \mu_A)(\mu_C^T - \mu_A^T)$.

As the value of W is independent of W_0, W_0 can be chosen such that the two distributions of the two population means are symmetric about the origin. It is also evident that a scale factor can be chosen to place the two means at an arbitrary but symmetric location around the origin. Hence the choice of the mean value for training is arbitrary, provided that the X values have no physical significance, i.e. it does not measure a patient variable. The aim of this paper was to determine a classifier.

(b) Classifier Performance

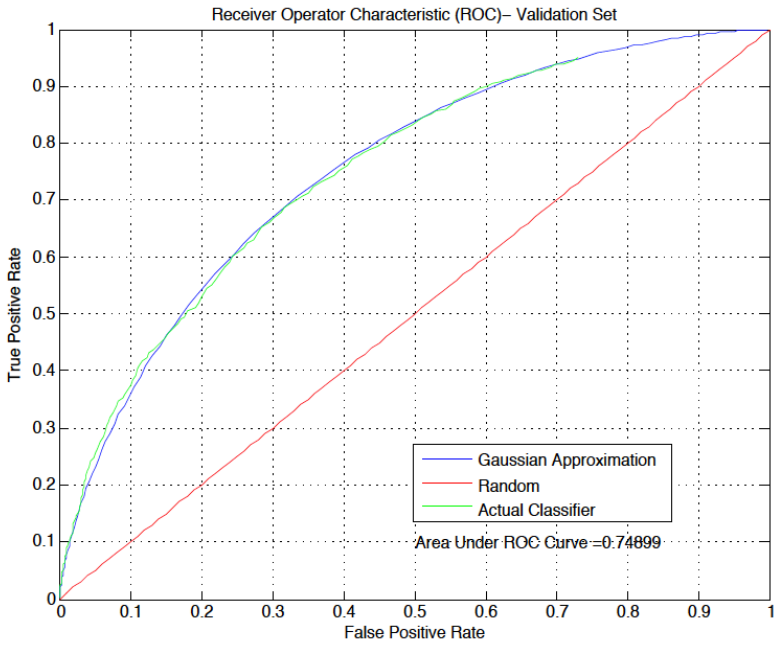
Figure S1 - Labels were randomly permuted on the training sample and the resulting classifier was used to determine clinical status in the independent validation samples. The graph below indicates the percentage of classifiers versus misclassification rate. The trained classifier on the correctly labeled data had the best performance of all other classifiers. The probability that another classifier trained on the permuted labeled data has better performance is less



than $1x10^{-3}$.

S4 – Prediction & ROC curves & area under ROC

Figure S2 - The Receiver Operator Characteristic (ROC) curve determined for the independent validation set (SFARI & WTBC), not previously seen by the classifier, showing the performance of the classifier as a function of false positive and true positive rates, as compared to random. The area under the curve was 0.749.



S5 – Distributions of Relatives and Parents in AGRE

Figure S3 - As seen in this figure, the means for the parents (mothers = 2.83, S.D. = 2.17; fathers = 2.93, S.D. = 2.34) is similar to the mean for the relatives (parents and siblings combined, mean = 2.68, S.D. = 2.27) overall. However, as seen in the main text, unaffected siblings (not meeting diagnostic criteria for ASD) fall between parents and ASD cases (mean = 4.74, S.D. = 3.80). (Mean for Controls = -0.95, S.D. = 3.01; Mean for ASD cases = 7.74, S.D. = 2.07).