Whole Genome Association Studies in Autistic Spectrum Disorders Revisited: A Support Vector Machine Approach

Johnston P, Hardoon D R, Ecker C, Clarke T K, Powell J & Murphy D

Background

Autistic spectrum disorders (ASDs) are moderately common, highly heritable neurodevelopmental conditions with a strong genetic basis. Several lines of evidence support genetic factors as a predominant cause of ASDs. However, investigations' using conventional genetic approaches has been slow. To date no single biological or clinical markers have yet been identified. Recent years has seen an increase use of whole genome association studies (WGAS), specifically through the establishment of collaborative efforts such as Autism Genetic Resource Exchange (AGRE) and the Autism Genome Project. Still very little light has been shed on the complex aetiology of this polygenic disorder. Support vector machines (SVM), one method of machine learning, has the ability to classify data using a mathematical function which best discriminates two groups - also highlighting the most influential discriminatory factors. However nobody has yet applied a SVM approach on WGAS.

Objectives

To analysis whole genome association data using a SVM application, classifying individuals into ASD affected or unaffected groups. This data will be used to indicate which SNPs are the most influential in terms of the classification. This is the first known study to use a SVM approach on whole genome data.

Methods

A WGAS (Affymetrix 5.0) was conducted on 2879 individuals generated at the Broad Institute and data was kindly provided to the Autism Genetics Resource Exchange. The

sample was comprised of 1385 affected and 1494 unaffected individuals, each with 390671 features (SNPs). A SVM analysis, using a linear kernel, was applied to the data using a leave-one-out procedure.

Results

SVM achieved an overall classification accuracy of 74% of the total sample. When broken down into affected and unaffected results of 54% and 94% respectively where achieved. 10 SNPs where identified as having a high weighted effect in discriminating the ASD affected from the unaffected groups. These 10 SNPs included 2 on chromosome 2 (q37.2) within the CENTG2 gene region.

Conclusion

This is the first study examining whole genome data using a SVM approach. Even though only 54% of the affected individuals where correctly classified, this analysis approach did identify SNPs within a gene found previously to be associated with ASD. This highlights the potential valuable use of SVM analysis on whole genome association data.

Acknowledgments

Data was generated at the Broad Institute and provided to AGRE by Dr Mark Daly and the Autism Consortium.