## Poster B-28

# A Computational Pipeline for the Analysis of Data from a 500,000 SNP Whole Genome Association Study of Multiple Sclerosis


**Authors:**
David Sexton (*Center for Human Genetics Research, Vanderbilt University, International Multiple Sclerosis Genetics*)
Jacob McCauley (*Center for Human Genetics Research, Vanderbilt University, International Multiple Sclerosis Genetics*)
Justin Giles (*Center for Human Genetics Research, Vanderbilt University, International Multiple Sclerosis Genetics*)
Yuki Bradford (*Center for Human Genetics Research, Vanderbilt University, International Multiple Sclerosis Genetics*)
Jonathan Haines (*Center for Human Genetics Research, Vanderbilt University, International Multiple Sclerosis Genetics*)

**Short Abstract:** The International Multiple Sclerosis Genetics Consortium (IMSGC) has undertaken a whole genome association study for regions with association to multiple sclerosis phenotypes. This study consists of genotyping 500,000 SNPs in 1000 affected family trios. New methods were developed to automatically upload, quality control, analyze, and report this data.

**Long Abstract:**

Multiple sclerosis (MS) is a complex disease affecting over 2.5 million individuals worldwide. MS is an autoimmune inflammatory disease characterized by chronic inflammation, demylenation, and gliosis. Plaques and lesions in the CNS of patients result in the inhibition of normal nerve impulse function. Epidemiological studies have detected a link in the etiology of the disease to the genetic susceptibility of affected individuals. While a region of the genome specific to MS has been identified in the Major Histocompatibility Complex (MHC) on 6p21, genomic linkage screens have failed to detect additional loci involved in this disorder. Given that it is likely that other genes are involved to a lesser extent in the disease process, these loci may not be easy to detect through linkage screens.

The International Multiple Sclerosis Genetics Consortium (IMSGC) has decided to conduct an exhaustive search of the genome to catalog the genetic variation of importance to MS. An association study has been undertaken using a collection of 500,000 single nucleotide polymorphisms (SNPs) spaced across the genome and arrayed on two Affymetrix gene chips. A retrospective cohort of 1000 parent-affected child trios were collected and probed for variation in the 500,000 SNPs. A total of 1.5 billion genotypes were collected and transferred to Vanderbilt University for storage and quality control.

Vanderbilt University Center for Human Genetics Research (CHGR) created an automated computational pipeline to upload, store, clean, and analyze the data from the association study in MS. A relational database, modeled on Duke University's Pedigene database, was created to store sample, pedigree, demographic, clinical, variation probe, and genotype data.

Software was written to automatically upload these various data types to the database as they were deposited by the collaborating institutions.

An integrated quality control application was developed to clean the data before primary and secondary analysis. Eight different quality control steps are codified in this application including:

1) A test to determine marker efficiency and accuracy for determining error rates.
2) A comparison of duplicate SNPs across the two gene chips to determine heterozygote drop-out and error rates.
3) Calculation of allele frequencies to remove markers with minor allele frequency less than 5%.
4) A test of Hardy-Weinberg equilibrium in both parents and probands to remove SNPs with a p-value less than 0.001.
5) Checks of Mendelian inheritance errors.
6) Test consistency of gender with X chromosome SNPs.
7) Test for population substructure.
8) Comparison of relationships across samples to remove sample duplication using Relpair.

Simulations were performed on a 1000 trio set to determine the relative power of the transmission disequilibrium test software, TDT and AFBAC using multiple models. These tests determined that there was no difference in power using either software package.

Primary analysis of the genotyping data encompasses a transmission disequilibrium test on the overall dataset for each SNP. SNPs are ranked by P value and those with p values greater than 1 x 10 -7 are flagged for follow up study. Linkage Disequilibrium (LD) relationships among the SNPs are calculated and compared to existing data from the HapMap project. Regions with significant LD are defined using r2 = .80. PHASE is used to determine phased haplotypes for all individuals and identify haplotypes to reduce further testing. Tests for both main and 2 way interaction effects using multifactor dimensionality reduction (MDR) software provides a non-parametric test among loci. The detection of copy number polymorphisms by use of arrayCGH and identification of SNPs in known copy number polymorphisms will help determine significant form of variation. These tests will form the initial analysis and characterization of the data and have been automated by the CHGR.

This effort has created an automated computational pipeline to analyze this unique dataset using a novel method.