Online Supplementary Data

Population structure confounds autism genetic classifier

T. Grant Belgard, DPhil[1]
Ivana Jankovic, BS[2]
Jennifer K. Lowe, PhD[1,3]
Daniel H. Geschwind, MD PhD[1,2,3,*]

[1]Center for Autism Research and Treatment, and Program in Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior,
[2]Department of Human Genetics,
[3]Department of Neurology,
University of California, Los Angeles, 90095.

* Corresponding author:
Daniel H. Geschwind, MD PhD
UCLA Neurogenetics Program
2309 Gonda Building
695 Charles E. Young Dr. South
Los Angeles, CA 90095-1761
Phone: 310-794-6570
Fax: 310-267-2401
dhg@mednet.ucla.edu

Daniel H. Geschwind is on the scientific advisory board of Synapdx. All other authors declare no financial conflicts of interest.

Table S1. Minor allele frequencies of SNPs in CEU versus Estonians

SNP	Weight *	Gene symbol	CEU MAF	CEU # alleles	Estonian MAF	Estonian # alleles	Estonian MAF /CEU MAF
Risk SNPs - expect Estonian MAF > CEU MAF if driven by this aspect of population structure							
rs968122	1.5555	*KCNMB4*	0.119	226	0.261	1952	**2.19**
rs9288685	0.5998	*INPP5D*	0.407	226	0.555	1950	**1.36**
rs10193128	0.5946	*INPP5D*	0.344	224	0.523	1950	**1.52**
rs7842798	0.5386	*ADCY8*	0.394	226	0.509	1946	**1.29**
rs1818106	0.5161	*PDGFD*	0.606	226	0.545	1950	0.90
rs2384061	0.4306	*ADCY3*	0.375	224	0.382	1952	**1.02**
rs2300497	0.3889	*CALM1*	0.076	224	0.116	1950	**1.53**
rs7313997	0.3567	*PTPRR*	0.049	226	0.108	1952	**2.20**
rs2239118	0.3552	*CACNA1C*	0.168	226	0.231	1952	**1.38**
Protective SNPs - expect CEU MAF > Estonian MAF if driven by this aspect of population structure							
rs1243679	-0.5674	*OR6S1*	0.066	226	0.051	1952	**0.77**
rs260808	-0.5836	*PDGFD*	0.225	226	0.091	1952	**0.40**
rs4128941	-0.6082	*AXIN2*	0.075	226	0.052	1952	**0.69**
rs769052	-0.6235	*UBE2D2*	0.097	226	0.034	1952	**0.35**
rs984371	-0.7181	*OR5L1*	0.164	226	0.208	1948	1.27
rs4308342	-0.8938	*DCK*	0.071	226	0.019	1948	**0.27**
rs905646	-0.9624	*GRM5*	0.181	226	0.113	1948	**0.62**
rs6483362	-0.9661	*GRM5*	0.128	226	0.092	1952	**0.72**
rs8053370	-1.6956	*GNAO1*	0.111	226	0.093	1950	**0.84**

Commentary on Figure 3B and Figure S3

On the face of it, this interpretation appears to conflict with Figures 3B and Figure S3 in Skafidas, et al. in which overlapping distributions of CEU controls have a low classifier score, AGRE parents have a middling classifier score, unaffected siblings of people with autism have a mid-to-high but very broad distribution of scores, and autistic individuals have a high classifier score. However, these results are consistent with our population structure interpretation. Most of the SNPs were not reported in the paper, and if population structure were a factor one would expect the SNPs with smaller effects to be more likely the result of noise. The differences in population structure would result in a higher autism classifier score for both probands and unaffected relatives in AGRE compared to the unrelated controls in CEU. Since the AGRE set was used for training, simply overfitting to the AGRE probands would result in a higher autism classifier score for AGRE probands than for unaffected relatives. Siblings without autism fall in a broad distribution of autism classifier scores that peaks between their parents and children with autism. This makes sense given the construction of the model. Individuals who are heterozygous for the minor allele were assigned 1 point while homozygotes were assigned 3 points. A given parent of a proband homozygous for a minor allele is less likely to be homozygous for the allele than an unaffected sibling of the proband for minor allele frequencies less than 0.38 (see derivation below). The minor allele frequencies of most of the reported SNPs were in this domain, suggesting even unaffected siblings should be between parents and probands. Finally, one would expect the distribution for unaffected siblings to be broader than the distribution for parents since the proportion of the autosomal genome shared by siblings is more variable than the fixed 50% shared by parents and children. We cannot discern from what subset of AGRE data this graph was generated based on the information provided by the authors.

Derivation of probabilities that a full sibling or a parent would be homozygous given a minor allele frequency

Assuming a biallelic autosomal SNP in a homogeneous population in Hardy-Weinberg equilibrium with minor allele frequency M, the probability that a random parent of a homozygous individual would also be homozygous is M.

If we know an individual is homozygous, both of the parents must either be heterozygous or homozygous. Thus, the possible parental sets of major and minor alleles, S and s, are Ss/Ss, Ss/ss, and ss/ss. For a population in Hardy-Weinberg equilibrium, the probability of a random individual being Ss is $2M(1-M)$ and ss is M^2. If a child is known to be homozygous, the probabilities are thus as follows:

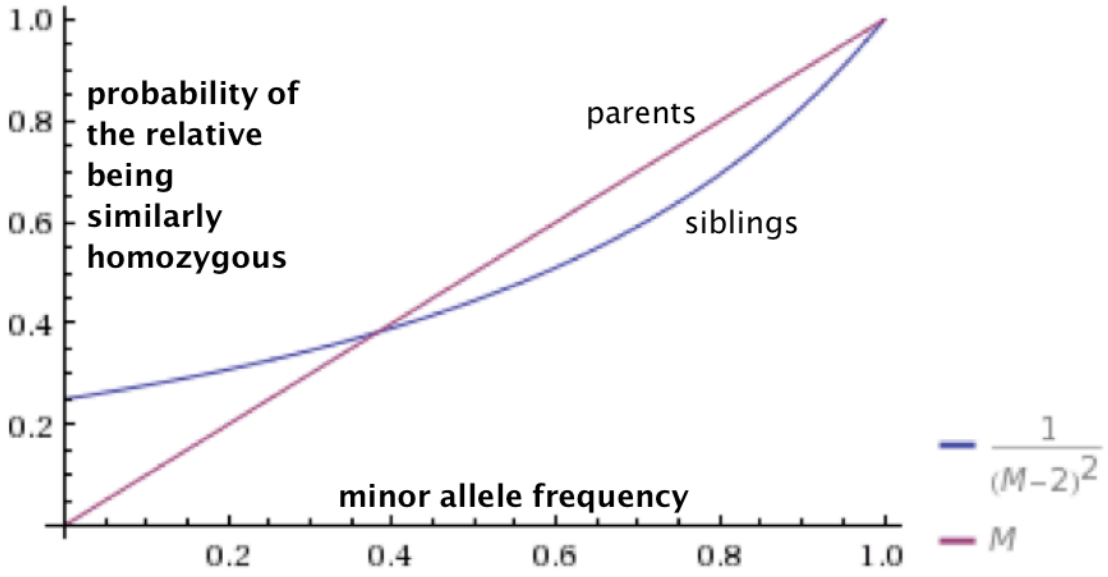
$P(Ss/Ss) = N[4M^2(1-M)^2]$

$P(Ss/ss) = P(Ss_{father}/ss_{mother}) + P(Ss_{mother}/ss_{father}) = 2N[2M^3(1-M)]$

$P(ss/ss) = N[M^4]$

where N is a normalization factor equal to $[M^2(M-2)^2]^{-1}$

Given that $P(ss_{child}|Ss_{father}/Ss_{mother}) = ¼$, $P(ss_{child}|ss_{father}/Ss_{mother}) = P(ss_{child}|Ss_{father}/ss_{mother}) = ½$ and $P(ss_{child}|ss_{father}/ss_{mother}) = 1$, we find that $P(ss_{sibling1}|ss_{sibling2}) = P(Ss/Ss)/4 + P(Ss/ss)/2 + P(ss/ss) = NM^2[(1-M)^2 + 2M(1-M) + M^2] = NM^2 = (M-2)^{-2}$



Minor allele frequency versus probability that a sibling (blue) or parent (purple) will be similarly homozygous for that SNP. $P(ss_{sibling1}|ss_{sibling2}) > P(ss_{arbitrary\_parent}|ss_{child})$ where $M < (3-\sqrt{5})/2 \cong 0.38$. Integrating over the difference of these gives from M of 0 to 1 gives 0. However, 13 of the 18 reported SNPs had a MAF below this level in both the CEU and Estonian sets (15 of the 18 in CEU alone).