Discussion guide for 27 April 2017

Remember: 5 points of your grade on Project #6 will be based on your participation in this discussion. Please come to class on Thursday prepared to discuss the paper and the critique.

Stessman et al. [2] report results of a case-control analysis of 208 genes that have been suggested as candidates for beining involved in various neurodevelopmental disorders, e.g., autism spectrum disorder (ASD), intellectual disability (ID)/developmental delay (DD), attention deficit/hyperactivity, motor and tic disorders, and language communication disorders. Previous studies had shown that there were no *de novo* variants in unaffected siblings.[1] They sequenced the genes in 15 large cohorts, with a total of 13,407 individuals exhibiting ASD, ID, or DD. They identified 61,315 variants across this set of individuals and loci, but only 2185 were restricted to a single family, i.e., "private" variants and were potentially deleterious, e.g., premature stop codons, frameshift mutations, disruptive splicing mutations. "The number of private, high-impact events identified in probands[2] was significantly greater than that in unaffected siblings in the study" (p. 516).

They used a probabilistic model (described elsewhere) to estimate that the probability of detecting 114 or more *de novo* likely gene-disruptive (LGD) events and 24 or more *de novo* severe missense (MIS30) events was abot 1.6×10^{-22}. They combined their data with pre-existing exome data to determine whether the frequency of LGD mutations or MIS30 mutations in individual genes was higher in individuals with ASD, ID, or DD than in those without these disorders. They identified 78 genes with significant associations, i.e., in which allele frequencies differed significantly between case and control populations, 32 of which had not been previously associated with these or other neurodevelopmental disorders.

Barrett et al. [1] are unconvinced by these results. Their skepticism is based on two criticisms of the analysis, but we'll consider only the first:

- Barrett et al. point out that Stessman et al. used a "two-stage" design. In the first stage, they identified candidate loci in a "discovery" sample of 5000-6000.[3] They then

[1] *De novo* variants are recognized as variants found in only a single family within a study cohort.

[2] Probands are individuals with the diagnosis of interest, in this case ASD, ID, or DD.

[3] So far as I have been can tell without digging into the references in Stessman et al., the sample size of the "discovery"sample must refer to the number of individuals included in "published sequencing studies"

sequenced these loci in a "replication" sample of 11,730 individuals. They claim that Stessman et al. calculate the false discovery rate incorrectly because they focus on only the 208 candidate loci but include both the discovery sample and the replication sample in their analysis.

The false discovery rate (FDR) both sets of authors refer to is a statistical correction applied when many significance tests are performed at once. If, for example, we run 100 t-tests on samples which were drawn from identical normal distributions and we set our significance threshhold to the conventional 5%, we expect to reject the null hypothesis of no significant difference 5 times, even though there really wasn't any difference. There are various ways of correcting for the multiple comparison problem, but all of them depend on counting the number of comparisons correctly. Barrett et al.'s claim is that Stessman et al. didn't count correctly.

Focusing on the multiple comparison critique from Barrett et al. and using your general knowledge of population genetics, explore the following questions:

- The case-control comparison involves cases and controls from 15 different cohort samples from 4 continents (North America, Europe, Asia, and Australia). How should this cohort structure affect your analysis?[4]

- The study lumps together individuals with ASD, ID, and DD as "cases" and those without any of the conditions as controls? Does this make sense? Why might you do the analysis this way rather than performing separate analyses for each of the conditions?

- What is the right way to count the number of comparisons for an FDR correction when you have a "discovery" sample and a "replication" sample? Why do Barrett et al. regard it as such a significant flaw? Are their concerns well founded?

- Set aside what you concluded in discussing the first three questions and assume that Stessman et al. are right that the 78 unique genes they identify in their Table 1 are associated with one or more of the neurodevelopmental disorders they studied. How useful would it be to know an individual's genotype at these 78 loci in diagnosing their disease? What additional information would you need to make knowing a patient's genotype useful for diagnosis if just knowing their genotype isn't enough?

mentioned in the first sentence of the **Mutation discovery** section of Stessman et al.

[4]Hint: Think about what would happen if your analysis lumped all of the cohort samples into a single population of cases and a single population of controls and suppose that ASD, ID, and DD are more common in North America.

References

[1] Jeffrey C Barrett, Joseph Buxbaum, David Cutler, Mark Daly, Bernie Devlin, Jacob Gratten, Matthew E Hurles, Jack A Kosmicki, Eric S Lander, Daniel G MacArthur, Benjamin M Neale, Kathryn Roeder, Peter M Visscher, and Naomi R Wray. New mutations, old statistical challenges. *bioRxiv*, pages 1–11, March 2017.

[2] Holly A F Stessman, Bo Xiong, Bradley P Coe, Tianyun Wang, Kendra Hoekzema, Michaela Fenckova, Malin Kvarnung, Jennifer Gerdts, Sandy Trinh, Nele Cosemans, Laura Vives, Janice Lin, Tychele N Turner, Gijs Santen, Claudia Ruivenkamp, Marjolein Kriek, Arie van Haeringen, Emmelien Aten, Kathryn Friend, Jan Liebelt, Christopher Barnett, Eric Haan, Marie Shaw, Jozef Gecz, Britt-Marie Anderlid, Ann Nordgren, Anna Lindstrand, Charles Schwartz, R Frank Kooy, Geert Vandeweyer, Celine Helsmoortel, Corrado Romano, Antonino Alberti, Mirella Vinci, Emanuela Avola, Stefania Giusto, Eric Courchesne, Tiziano Pramparo, Karen Pierce, Srinivasa Nalabolu, David G Amaral, Ingrid E Scheffer, Martin B Delatycki, Paul J Lockhart, Fereydoun Hormozdiari, Benjamin Harich, Anna Castells-Nobau, Kun Xia, Hilde Peeters, Magnus Nordenskjöld, Annette Schenck, Raphael A Bernier, and Evan E Eichler. Targeted sequencing identifies 91 neurodevelopmental- disorder risk genes with autism and developmental-disability biases. *Nature Publishing Group*, 49(4):515–526, February 2017.