

Submitter Name: Apurva S. Chitre
Submitted email: aschitre@health.ucsd.edu
PI Name (if different): Abraham A. Palmer
PI email (if different): aapalmer@health.ucsd.edu

Exponential increase in QTL detection with increased sample size in heterogeneous stock rats

Apurva S. Chitre¹, Oksana Poleskaya¹, Riyan Cheng¹, Katie Holl², Jianjun Gao¹, , Hannah Bimschleger¹, Angel Garcia Martinez³, Tony George⁶, Alexander F. Gileta^{1,10}, Wenyan Han³, Aidan Horvath⁴, Alesa Hughson⁴, Keita Ishiwari⁶, Christopher P. King⁵, Alexander Lamparelli⁵, Cassandra L. Versaggi⁵, Connor Martin⁶, Celine L. St. Pierre¹¹, Jordan A. Tripi⁵, Jerry B. Richards⁶, Tengfei Wang³, Hao Chen³, Shelly B. Fligel¹², Paul Meyer⁵, Terry E. Robinson⁷, Leah C. Solberg Woods⁹, Abraham A. Palmer^{1,8}

¹Department of Psychiatry, University of California San Diego; ²Human and Molecular Genetic Center, Medical College of Wisconsin; ³Department of Pharmacology, University of Tennessee Health Science Center; ⁴Department of Psychiatry, University of Michigan; ⁵Department of Psychology, University at Buffalo; ⁶Clinical and Research Institute on Addictions, University at Buffalo; ⁷Department of Psychology, University of Michigan; ⁸Institute for Genomic Medicine, University of California San Diego; ⁹Department of Internal Medicine, Wake Forest School of Medicine; ¹⁰Department of Human Genetics, University of Chicago; ¹¹Department of Genetics, Washington University in St. Louis, ¹²Molecular and Behavioral Neuroscience Institute, University of Michigan

Power analyses are often used to determine the number of animals to use in a genome wide association analysis. These analyses are typically intended to estimate the sample size needed for at least one locus to exceed a genome-wide significance threshold. A related question that is less commonly considered is the number of significant loci that will be discovered with different sample sizes. We used simulations based on a real dataset that consisted of 3,173 male and female adult N/NIH heterogeneous stock (HS) rats to explore the relationship between sample size and the number of significant loci discovered. Our simulations examined the number of loci identified in sub-samples of the full dataset. The sub-sampling analysis was conducted for four traits with low (0.15 ± 0.03), medium (0.31 ± 0.03 and 0.36 ± 0.03) and high (0.46 ± 0.03) SNP-based heritabilities. For each trait, we subsampled the data 100 times at different sample sizes (500, 1,000, 1,500, 2,000, and 2,500). We observed an exponential increase in the number of significant loci with larger sample sizes. Our results are consistent with similar observations in human GWAS and imply that future rodent GWAS should use sample sizes that are significantly larger than those needed to obtain a single significant result.