Identification of pathology-related single nucleotide polymorphisms in a heterogeneous substance-abusing population

August 1, 2016

Ariel Ketcherside, MSc
Center for BrainHealth
University of Texas at Dallas
Dallas, TX 75235

Advisor: Francesca M. Filbey, PhD

Shikha Prashad, PhD
Center for BrainHealth
University of Texas at Dallas
Dallas, TX 75235

Advisor: Francesca M. Filbey, PhD

Milind Rao
Electrical Engineering
Stanford University
Stanford, CA, 94305

Advisor: Andrea Goldsmith
[advisor signature]
Andrea Goldsmith

Project Type: Pattern classification for large-scale bioinformatics data
Total Funds Requested: $6000

Problem statement

Similarities in behavior, neural activity, and genetic data have been identified between different substances of abuse (Kendler, Jacobson, Prescott, & Neale, 2003). Efforts to establish a “final common pathway” have produced a model which captures the reward and impulse control features of all substance use disorders including nicotine, marijuana, cocaine, heroin, and food (Pierce & Kumaresan, 2006; Volkow, Wang, Tomasi, & Baler, 2013; Wang, Volkow, Thanos, & Fowler, 2004). However, genetic contributions to these complex pathologies remain obfuscated, due to the volume of genetic information.

This conundrum is further complicated by the vast nature of genome wide association study (GWAS) data (Gratten, Wray, Keller, & Visscher, 2014; Visscher, Brown, McCarthy, & Yang, 2012). The large (800,000) number of single nucleotide polymorphisms (SNPs) captured per subject inflates the rate of false positive and thus, large samples sizes are needed for these high-dimensional inference tasks. Minor allele frequency further necessitates large sample sizes, as a risk allele for a disease could occur in a fraction of a percent of a population. As a result, much GWAS data is uninterpretable using naive statistical methods due to insufficient sample sizes. Thus, it is necessary to determine which alleles confer propensity toward substance use disorders in general, and toward specific substance use disorders.

Proposed activity

To that end, we will apply machine learning analysis methods to GWAS data collected across 500 individuals from three different substance-using populations, as well as healthy controls. By correlation with behavior traits like impulsivity, reward sensitivity, depression, and anxiety and subsequent pattern classification of these data, we will identify of a risk profile score characterized by the number of risk alleles carried by an individual, in conjunction with their behavioral and substance-use traits. Individuals in this study have been recruited based on self-reported frequent cannabis use, binge eating, nicotine use, or as a healthy control, and all have GWAS data that has been normalized across genome builds, ethnicity, and batch effects. We also have a variety of self-report measures assessing psychological factors (e.g. depression, anxiety, impulsivity, reward-sensitivity, IQ, memory).

In order to predict association of SNPs with psychopathology, GWAS data will undergo dimension reduction through a database search (SNPedia) for previous associations with cannabis, food, and nicotine use disorders. Only SNPs that have been previously documented will be used, for statistical power necessary to make strong structural assumptions (such as sparsity).

Aim 1. Determine alleles common in and specific to general substance use disorders (e.g. impulsivity, reward sensitivity).
Aim 2. Differentiate between substances of abuse based on a pattern classifier. Candidate SNPs and behavior measures will be used as features in a machine learning algorithm for prediction of pathology. The algorithm will be trained and tested using cross-validation approaches.

Successful completion of this project will result in identification of genetic variability that is unique to different substance use disorders, and genetic variability that is common across substance use disorders. This study is relevant to the Bioinformatics, Learning, and Statistics centers at the Center for Science of Information at Purdue University.

Goals and Outcomes

We have several expected outcomes from this research study. In the short-term, we expect that our collaboration and combined expertise will result in a novel approach to the analysis of a dataset that is currently largely uninterpretable. We will also contribute to the understanding of substance use disorders and factors that contribute to them. We hope that with our interdisciplinary backgrounds and approaches, we will be able to learn and bridge barriers between our respective fields and emerge as stronger researchers.

In the long-term, we expect to present our novel methods and findings at the 2017 Bioinformatics conference and publish a journal article. Additionally, we will write a software package to share our GWAS analysis methods with other researchers.

Proposed work statement

Ariel Ketcherside and Shikha Prashad will provide the background knowledge of GWAS and the substance using population to guide the hypotheses and desired outcome of the project. Milind Rao will provide statistical methods necessary to complete these aims. Dr. Filbey will oversee the project trajectory, and Dr. Goldsmith will be available for statistical consultation.

Team meetings will consist of video conference sessions and will occur once per month. Additionally, we will meet once in-person at the University of Texas at Dallas. All data will be shared on Google Drive or the Filbey Secure Server. Upon completion of the project, all three team members will attend the 2017 Bioinformatics (or equivalent) conference to present our work.

References

Gratten, J., Wray, N. R., Keller, M. C., & Visscher, P. M. (2014). Large-scale genomics unveils the genetic architecture of psychiatric disorders. *Nature Neuroscience*, *17*(6), 782–790. http://doi.org/10.1038/nn.3708

Kendler, K., Jacobson, K., Prescott, C., & Neale, M. (2003). Specificity of Genetic and Environmental Risk Factors for Use and AbuseDependence of Cannabis, Cocaine, Hallucinogens, Sedatives, Stimulants, and Opiates in Male Twins. *American Journal of Psychiatry*, *160*(4), 687–695.

Pierce, R. C., & Kumaresan, V. (2006). The mesolimbic dopamine system: The final common pathway for the reinforcing effect of drugs of abuse? *Neuroscience & Biobehavioral Reviews*, *30*(2), 215–238. http://doi.org/10.1016/j.neubiorev.2005.04.016

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five Years of GWAS Discovery. *American Journal of Human Genetics*, *90*(1), 7–24. http://doi.org/10.1016/j.ajhg.2011.11.029

Volkow, N. D., Wang, G.-J., Tomasi, D., & Baler, R. D. (2013). The Addictive Dimensionality of Obesity. *Biological Psychiatry*, *73*(9), 811–818. http://doi.org/10.1016/j.biopsych.2012.12.020

Wang, G.-J., Volkow, N. D., Thanos, P. K., & Fowler, J. S. (2004). Similarity Between Obesity and Drug Addiction as Assessed by Neurofunctional Imaging: A Concept Review. *Journal of Addictive Diseases*, *23*(3), 39–53. http://doi.org/10.1300/J069v23n03_04

Budget & Justification

We are requesting funding for one trip to the University of Texas at Dallas for 3 days where all the team members will meet in-person. Additionally, we will present the results of this research at the Bioinformatics (or equivalent) Conference. Our total budget is $6000 as follows:

Item	Cost per Item	Total Cost
Travel to Dallas	$350	$350
Hotel in Dallas	$250	$250
Travel to conference	$800 x 3	$2400
Registration fee for conference	$500 x 3	$1500
Hotel at conference	$500 x 3	$1500
	Total	**$6,000**