Prediction of cannabis use disorder severity from genetic and behavioral data

Ariel Ketcherside, Milind Rao, Shikha Prashad

Team Interactions / Meetings

We have had multiple in-person and online team meetings:

Skype	In-person
Jun. 28, 2016 – Ariel, Milind, Shikha	Sep. 1, 2016 – Ariel, Shikha
Jul. 29, 2016 – Ariel, Milind, Shikha	Nov. 21, 2016 – Ariel, Shikha
Oct. 5, 2016 – Ariel, Milind, Shikha	Dec. 6-7, 2016 – Ariel, Milind
Nov. 18, 2016 – Ariel, Milind, Shikha	Feb. 3, 2017 – Ariel, Shikha
Feb. 8, 2017 – Ariel, Milind, Shikha	Feb. 22, 2017 – Ariel, Shikha
Mar. 7, 2017 – Ariel, Milind, Shikha	Feb. 27, 2017 – Ariel, Shikha
Mar. 16, 2017 – Ariel, Milind, Shikha	Mar. 9, 2017 – Ariel, Shikha
May 12, 2017 – Ariel, Milind, Shikha	May 19, 2017 – Ariel, Shikha
Aug. 29, 2017 – Ariel, Milind, Shikha	

We have also had numerous online chat interactions.

Research Progress

Our aim is to predict severity of cannabis use disorder based on their SNPs and behavioral assessments. Specifically, we want to use behavioral measures of problematic cannabis use (Aim 1) and determine whether we can more accurately predict severity by including measures of craving and withdrawal (Aim 2).

We have created an algorithm to address Aim 1. Participants were asked questions about problems related to their cannabis use from the Marijuana Problem Scale (MPS). Each question is assigned a score of 0, 1 or 2 by the participant. The total score is a sum of the individual questions resulting in a behavioral metric b_l (range 0-38) indicating the severity of their addiction for subject l. We have 800,000 SNPs collected for each participant through the GWAS chip. By combing through the literature, we have reduced the number of interesting SNPs to around 15-200 depending on the key word sieve we use. For instance, keywords such as *cannabidiol, THC, etc.* resulted in an SNP set size of 15. More general addiction related keywords gives us a larger set. For the i^{th} SNP $(i \in \{1, 2, \ldots, N\})$, let the number of risk alleles for subject l be $r_i^l = (r_i^l \in \{0,\ 1,\ 2\})$.

We first think of behavior predicting functions of the form $\alpha_i r_i^l$, where α_i are the coefficients we have to determine. We hypothesize that a few of the N alleles are important based on the literature. Suppose the weight for a particular α_{i^*} is high relative to the other coefficients. This would imply that having risk alleles for SNP i^* greatly influences severity of cannabis use disorder. If $\alpha_i = 0$ αˆI, it would imply that SNP i^* has little effect. To determine these weights, we will solve the following Lasso problem, which is a variant of regression that minimizes the loss function (e.g., squared loss) between the behavior measure and the predicted behavior measure:

$$min_{\alpha_i} \sum_l (b_l - \sum_i \alpha_i r_i^l)^2 + \lambda \sum_i |\alpha_i|$$

This can be further expanded to determine effects of pairs of SNPs which allows us to study interaction effects. This procedure works around the multiple hypotheses testing problems that arise in this high dimensionality setting. Our approach to doing this centres on coefficients α_{ij} which measures the interaction effects between SNP *i* and *j*. We assume that only a few interactions produce an effect and hence solve a similar lasso minimization problem:

$$min_{\alpha_{ij}} \sum_{l} (b_l - \sum_{ij} \alpha_{ij} r_{ij}^{l}) \; + \; \lambda \sum_{ij} |\alpha_{ij}|$$

Preliminary results

The distributions of our variables of interest reveal a small correlation between craving and the MPS score, but craving was not correlated with other variables. There is also a skew in the sample with a greater number of males than females, as is common in substance abuse related studies. Figure 2 displays the MPS score with the leading two PCA dimensions of our SNP data. PC1 accounts for general variability in the sample; however, PC2 maps directly onto CUD severity (i.e., higher MPS score), suggesting that these SNPs can predict high MPS scores. We ran our lasso variable selection algorithm, which revealed that SNPs 20, 58, and 70 have a large coefficient index (Figure 3). We investigated how the MPS score varies for these specific SNPs and found significant correlations (all $p < 0.05$) between these SNPs and the MPS score (Figure 4). We further investigated the relationship between craving and these SNPs and found a small correlation (Figure 5). Lastly, we considered interactions between SNPs to explore whether interactions are related to behavior. We found that of the approximately 5000 interactions between the 104 SNPs that we had initially identified, there are only a few interactions that have non-zero lasso coefficients and the interaction with the highest coefficient was an interaction between two of the three SNPs of interest (rs10115383 and rs10834489; Figure 6A). We further found a significant correlation between the interaction of these SNPs and the MPS score ($p < 0.001$; Figure 6B).

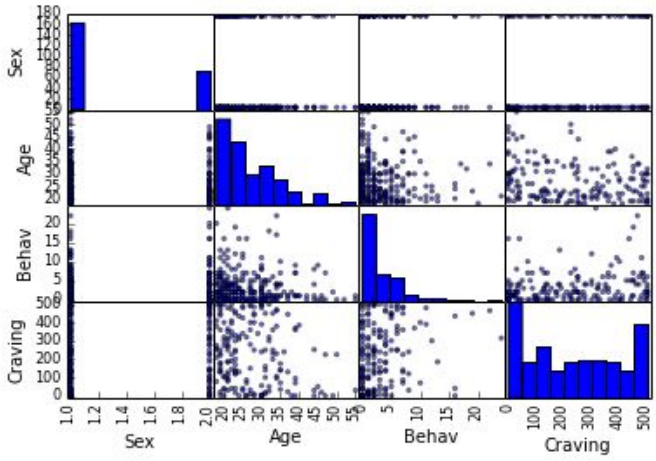

Figure 1. Distributions of variables of interest.

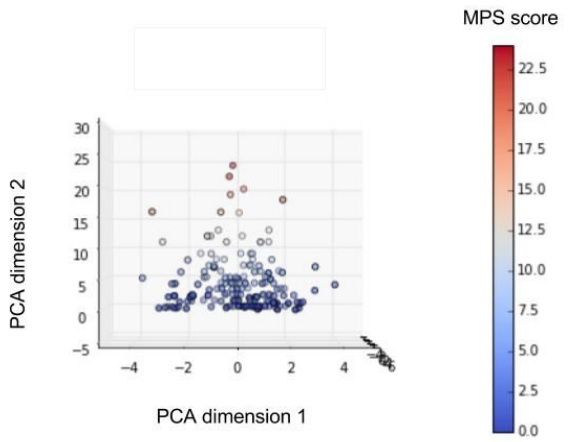

Figure 2. MPS scores with the top 2 PCA components from SNP data.

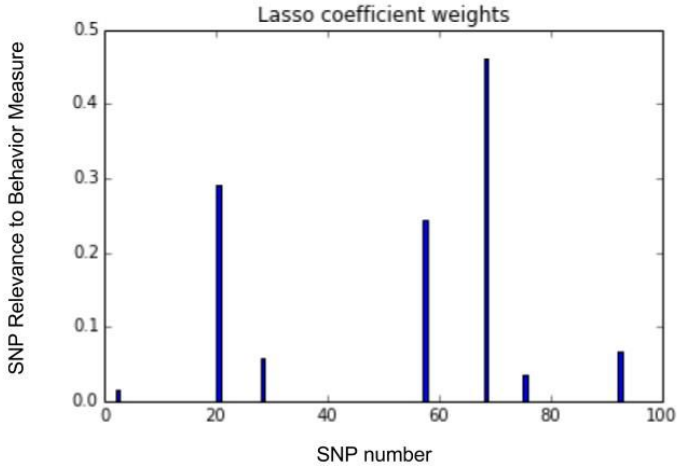

Figure 3. Lasso variable selection algorithm. This algorithm identified SNPs that were predictive of the MPS score.

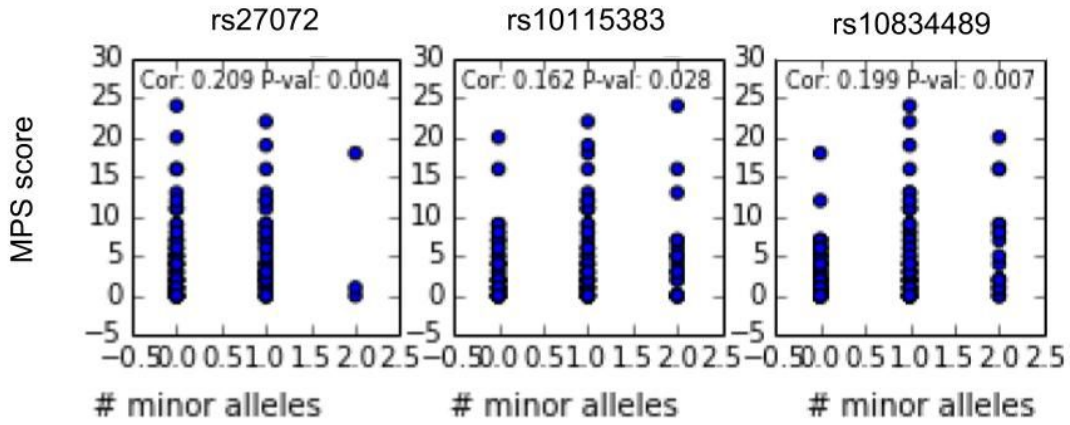

Figure 4. Correlation between identified SNPs of interest and MPS score

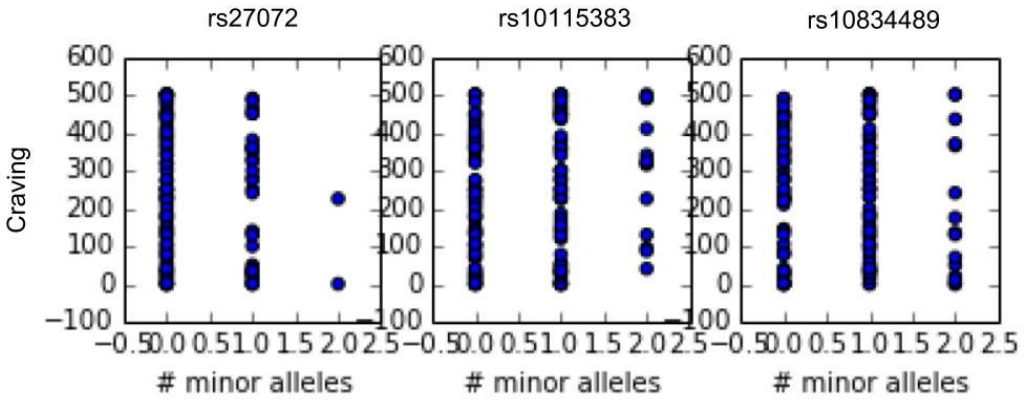

Figure 5. Correlation between identified SNPs of interest and craving.

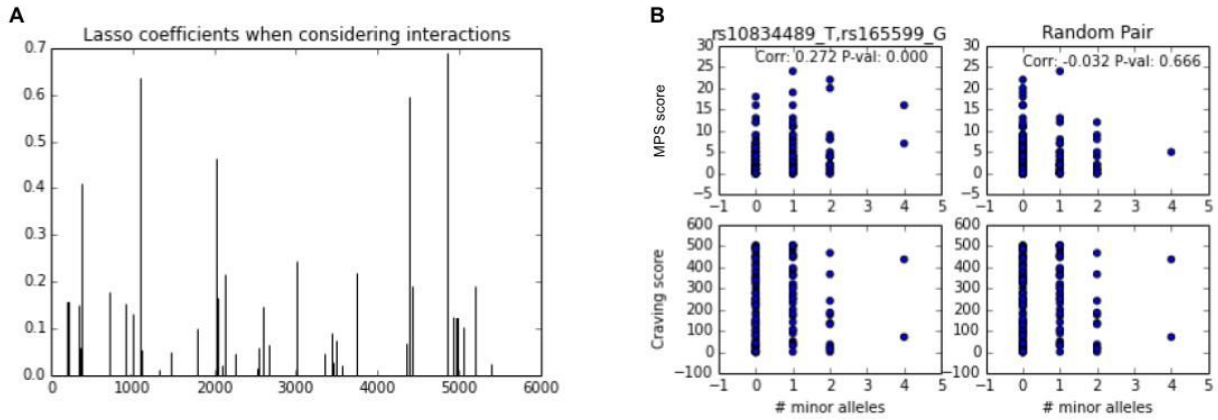

Figure 6. Interactions between SNPs. A) Interactions between SNP pairs. B) Correlation between rs10834489 and rs165599 and behavioral variables.

List of presentations, posters, conferences, publications

Presentation at NSF CSoI meeting in Purdue on December 6, 2016.
Presentation at the CSoI Virtual Brown Bag Series on March 16, 2017.

Future directions

We plan to complete the above outlined analyses to address Aim 1. We also plan to expand the algorithm to incorporate an additional measure of cannabis use (withdrawal) to determine whether it adds greater predictive power to the algorithm when combined with craving (Aim 2). Finally, we plan to present our findings at a bioinformatics / computational biology conference.

Remaining budget

Our entire $6000 budget is remaining. We plan to use a majority of it towards a presenting our findings at a conference. The following are potential conferences with dates that we are targeting:

Conference	Submission Deadline	Conference Date
73rd Annual Meeting of the Society of Biological Psychiatry	Dec. 14, 2017	May 10-12, 2018
Cognitive Neuroscience Society 2018 Annual Meeting	Oct. 2017 (not yet announced)	Mar 24-27 2018
Society for Neuroscience Annual Meeting 2018	May 2018 (not yet announced)	Nov. 2018 (not yet announced)