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
Jun 28, 2016 – Ariel, Milind, Shikha
Jul 29, 2016 – Ariel, Milind, Shikha
Oct 5, 2016 – Ariel, Milind, Shikha
Nov 18, 2016 – Ariel, Milind, Shikha
Feb 8, 2017 – Ariel, Milind, Shikha

In-person
Sept 1, 2016 – Ariel, Shikha
Nov 21, 2016 – Ariel, Shikha
Dec 6-7, 2016 – Ariel, Milind
Feb 3, 2017 – Ariel, Shikha
Feb 22, 2017 – Ariel, Shikha
Feb 27, 2017 – Ariel, 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. 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 $(i \in \{1,2,\dots,N\})$, let the number of risk alleles for subject l be $r_i^l = \left(r_i^l \in \{0,1,2\}\right)$.

We first think of behavior predicting functions of the form $\sum_i \alpha_i r_i^l$, 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 $\alpha_i = 0$ αˆl, 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:

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

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.

Some preliminary results are attached. In Fig. 1, we plot the behavior score with the leading two PCA dimensions of our SNP data – it appears that SNPs can predict very high behavioral scores. In Fig. 2, we ran our lasso variable selection procedure and see that SNPs 1 and 5 have a large coefficient index. We investigate this further by seeing how the behavior score varies for these two particular SNPs; it appears from Fig. 3 that the number of minor alleles has an impact on the behavior scores.

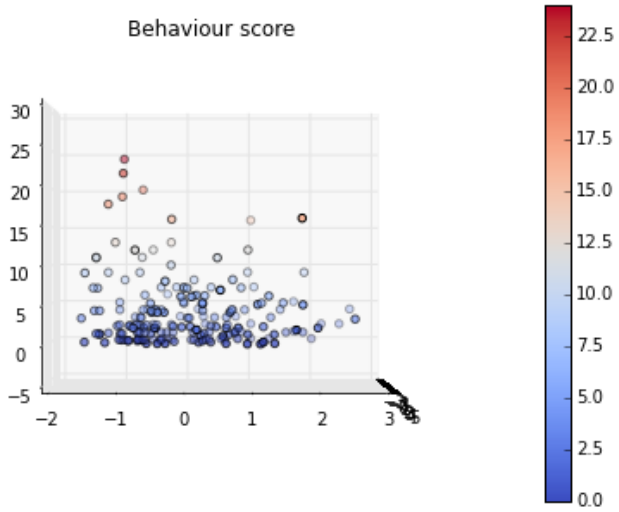

Figure 1 Behavior scores with top 2 PCA components of SNP data

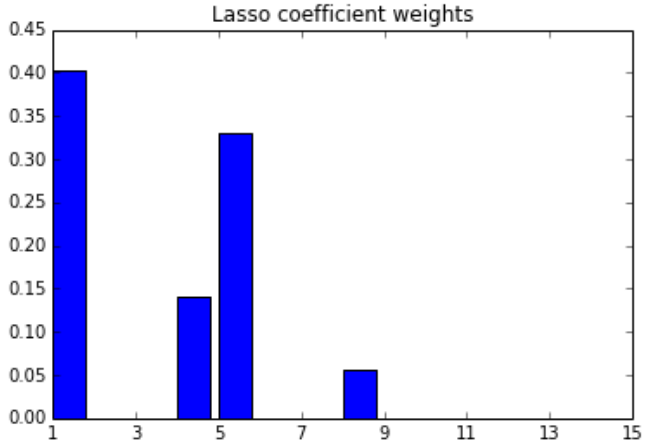

Figure 2

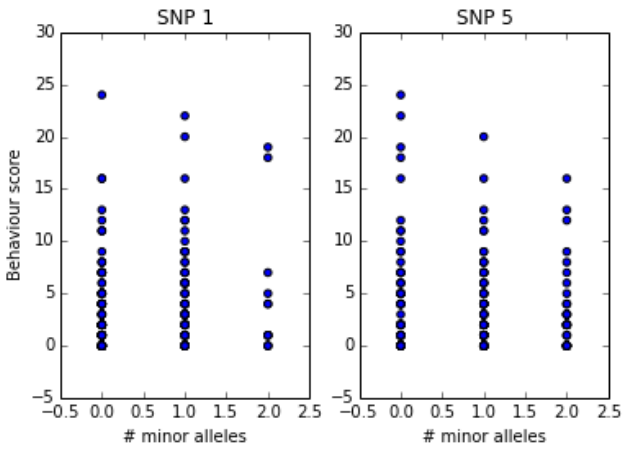

Figure 3

List of presentations, posters, conferences, publications
Presentation at NSF CSoI meeting in Purdue on December 6th.

Plans for next 6 months
We plan to complete the above outlined analyses to address Aim 1. We also plan to expand the algorithm to incorporate two additional measures of cannabis use, namely craving and withdrawal to determine if they add greater predictive power to the algorithm (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 conference in the next 6 months.