Final Report

Abstract

Social group size varies greatly across different species, and it has been proposed that ecological (Chapman, 2012) and cognitive factors contribute to determining the social group size of a particular species (Park, 2017; Dunbar, 1998). In primates, the social group size positively correlates with relative brain size, and the effect is present in different genera of primates, suggesting convergent evolution of large/small social group size (Dunbar, 1998). The social group size varies within the three lemuroidea and three hominoids chosen in this study. Social behaviors are related to the dopaminergic system in the striatum. Socially isolated rats exhibit hypersensitivity to dopamine release in the ventral striatum, and socially housed dominant monkeys have higher density of D2 receptors in the striatum (Báez-Mendoza & Schultz, 2013). Therefore, we will study the differences in social group size in relation to levels of predicted open chromatin in D1 and D2 medium spiny neurons in the striatum. We will also investigate the results further in the context of schizophrenia.

We identified four candidates that are likely most correlated to differential expression leading to different social group sizes. Even though they are functionally related to social behaviors and are expressed mostly in D1 and D2 MSNs, they are not differentially expressed in schizophrenic vs non-schizophrenic individuals and cannot be used to build strong models for predicting the schizophrenic phenotype. If we expand the candidacy to genes associated with 200 top differential regions, even though we cannot identify a relevant neural process that can be regulated by these genes, we can build models with much higher performances in predicting the schizophrenic phenotype.

The findings of the research can reveal connections between epigenetics and social group size as well as schizophrenia. In addition, it can demonstrate evolution of social behaviors at the epigenetic level. Epigenetic evidence will aid and strengthen current analytics of social behavior among species.

### Introduction

#### Trait

The trait chosen to conduct our analysis on is social group size of different species of primates. The social group of a species is the number of conspecifics an individual tends to interact and function with. For most species, interactions between others are temporary and change based on the given circumstances, but for primates, these social groups tend to stay consistent throughout their life span (Dunbar et al., 2018).

Social groups benefit the individual in numerous ways. Being in a group decreases an individual’s risk of predation and increases the chance of finding food and resources (Chapman, 2012). However, because there are more mouths to feed, there is more competition over the resources in a group, resulting in an expansion of foraging area (Chapman, 2012). The ecological constraint model suggests that the distribution and density of resources limit a species’ group size (Chapman, 2012).

In addition to ecological factors, cognitive factors can play a role as well. Human interactions are examples demonstrating these factors. Making decisions in a group can help a person benefit from other people’s ideas (Park, 2017). Consequently, human social groups can be determined by the willingness of the individuals in a group to agree with ideas not originally thought by themselves. The credibility of these opinions is heavily influenced based on prior experiences with others in the group (Park, 2017).

The relationship between cognitive ability (inferred by brain size) and social group size is well summarized by the social brain hypothesis. It proposes that the need to navigate complex social systems in primates drives evolution towards larger brain size (Dunbar, 1998). Analysis conducted by Dunbar and Shultz shows that relative brain size positively correlates with social group size in primates (2007). In addition, this relationship is seen in different genera of primates: simians, prosimians, and apes, suggesting that larger brains evolved multiple times and that social group complexity drove convergent evolution of larger brain size (Dunbar, 1998). However, the correlation between brain size and social group size is not seen in carnivores, ungulates, bats, and birds. Instead, pair-bonding (or monogamy) is associated with larger brain size in these taxa. (Shultz and Dunbar, 2007). Therefore, to ensure that there are neurological bases in differences in social group size between species, we restricted the study to only primates.

The neural basis of social group size can be further demonstrated by several MRI and fMRI studies in humans. The studies have discovered a strong correlation between social network size and activation of regions in the amygdala as well as amygdala volume (Liu et al., 2018). Because the amygdala projects to the striatum (Cho et al., 2013), differences in social group size could also be reflected in neurological differences in the striatum. This indeed is the case. In rhesus macaques, firing rates of the neurons in the caudate are strongly regulated by social reward (Klein & Platt, 2013). A human fMRI study also shows that mutual cooperation is associated with consistent activation of nucleus accumbens and caudate nucleus, two areas in the striatum (Rilling et al., 2002).

For the reasons above, we expect to see correlation between genetic and/or epigenetic differences in striatal cells and social group size across primate species. We referred to panTHERIA for annotation of social group size to each species (Jones et al., 2009).

#### Species

| Species Name | Group | Common Name | Social Group Size |
| --- | --- | --- | --- |
| Eulemur flavifrons | lemur | blue-eyed black lemur | 1 |
| Propithecus coquereli | lemur | Coquerel’s Sifaka | 9.15 |
| Microcebus murinus | lemur | grey mouse lemur | 5 |
| Nomascus Leucogenys | hominoid | northern white-cheeked gibbon | 8.55 |
| Homo Sapiens | hominoid | human | 5.75 |
| Piliocolobus tephrosceles | hominoid | Ugandan red colobus | 4 |

**Table 1**. Species chosen to conduct analysis on.

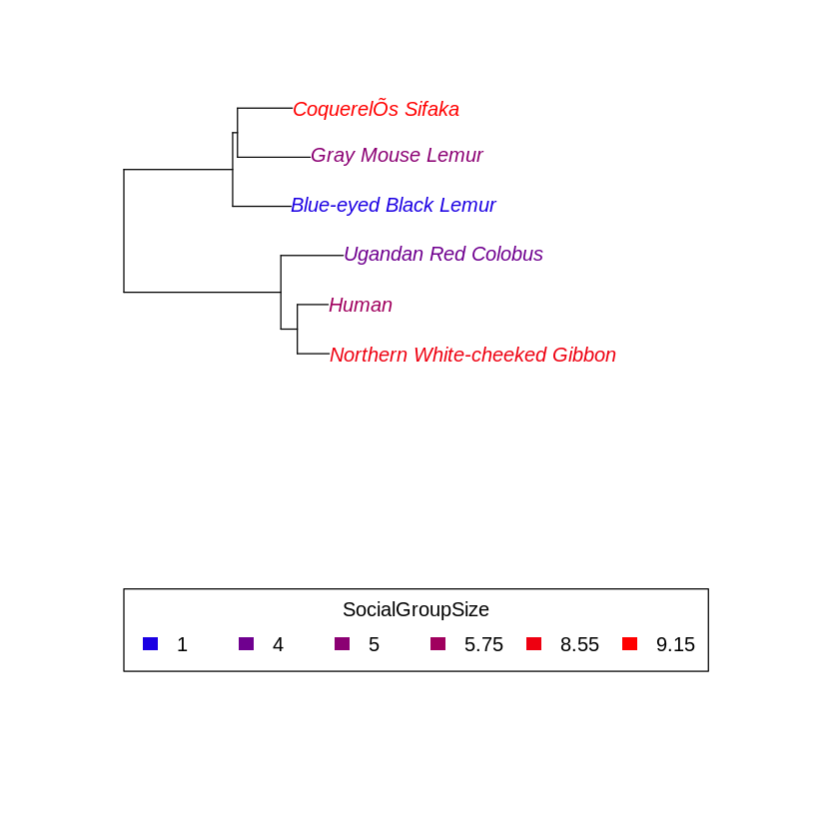
a)

|  | Eulemur flavifrons | Propithecus coquereli | Microcebus murinus |
| --- | --- | --- | --- |
| Eulemur flavifrons | -- | 4 | 5 |
| Propithecus coquereli | 4 | -- | 3 |
| Microcebus murinus | 5 | 3 | -- |

b)

|  | Nomascus Leucogenys | Homo sapien | Piliocolobus tephrosceles |
| --- | --- | --- | --- |
| Nomascus Leucogenys | -- | 3 | 4 |
| Homo sapien | 3 | -- | 7 |
| Piliocolobus tephrosceles | 4 | 7 | -- |

**Table 2**. The number of evolutionary events separating species within the same group. The numbers are derived from The Mammalian Tree on Zoonomia (“The Mammalian Tree”). **a)** The number of evolutionary events separating different species of lemurs. **b)** The number of evolutionary events separating different species of hominoids.



**Figure 1.** Evolutionary relationships between all chosen species. The tree shown here is built based on the Zoonomia tree. Social group size is shown on a color scale. There are differences within groups, suggesting convergence of social group size.

In Dunbar, Carron, and Shultz’s analysis, social group size refers to the average number of individuals observed in a group (2018). They base their analysis largely on published data describing features of various primate species, which is also how we conducted the analysis. The trait annotation (i.e. social group size data for each species) is from panTHERIA (Jones et al., 2009), and other publications are referred to for descriptions of key features related to the trait annotation.

For this study, two distinct groups of primates were chosen, lemuroidea and hominoids (Table 1). The social structure varies between the genera, with most lemurs living in multimale/multifemale groups and hominoids having smaller groups, such as families. However, there are still within-genera differences between species of the same group (Table 1 and Figure 1). Some species have larger groups and others have only sizes of one. In addition, we made sure that there are more than 3 evolutionary events (Table 2) separating species within the groups to ensure enough genetic differences to be detected. At the same time, we restricted our studies to only two genera in primates, thereby limiting the amount of genetic differences that are not specific to the trait of interest but simply brought by millions of years of evolution.

All chosen lemur species display multimale/multifemale social organization.

Eulemur Flavifrons’ groups are often separated into smaller subgroups, and Propithecus Coquereli have groups with high fluidity: males often visit and join other groups (Campbell et al., 2007). Microcebus murinus forage alone at night, but groups of females sleep together during non-mating season while males sleep alone (Campbell et al., 2007). During mating season, a single male sleeps with several females (Campbell et al., 2007).

Nomascus Leucogenys live in small groups of families, and most surveys report monogamous mating (Harding, 2012). Piliocolobus have a low-energy diet, and their feeding time correlates with group size, suggesting ecological constraints (intraspecific food competition) on their group size (Korstjens and Dunbar, 2007). Colobus group sizes can range from 4 to 20, but PIliocolobus are reported to pack in larger groups in places where they are hunted by chimpanzees, perhaps for group protection (Korstjens and Dunbar, 2007). Similar to other primates, humans live in small and highly structured social groups (Dunbar, 2020). In addition, human social groups can be divided into layers. The outer sublayer consists of people who we contact once a month and have “special ties to”, whereas the inner sublayer is the core from whom we seek emotional support from (Liu et al., 2018).

#### Cell Types

We chose to restrict our studies on predicted open chromatin levels in D1 and D2 medium spiny neurons (MSN). D1 and D2 MSNs are GABAnergic neurons in the nucleus accumbens, which is located in the ventral striatum (Soares-Cunha et al., 2019). D1 MSNs express dopaminergic receptor D1 and project directly to substantia nigra *pars reticulata* and internal pallidum (or endopeduncular nucleus in rodents), forming the direct pathway (Gagnon et al., 2017). In contrast, D2 MSNs express dopaminergic receptor D2 and project to ventral pallidum first before the circuit reaches output nuclei in basal ganglia, forming the indirect pathway (Soares-Cunha et al., 2019). Both D1 and D2 MSNs signal reward and aversion, depending on the stimulation pattern (Soares-Cunha et al., 2019).

The relationship between social behaviors and D1 and D2 MSNs has been demonstrated across species. In mice, expression of PDE10A2, a phosphodiesterase that degrades cAMP and cGMP, is specific to MSNs in the striatum (Sano et al., 2008). After knocking out the PDE10A2 gene in mice, mutant mice show increased social interactions with other mice. They have longer duration of contact in total, longer mean duration per contact, more social investigation (nose-body and nose-head contacts) without showing other behavioral changes (Sano et al., 2008). In humans, PDE10A also has the highest level of expression in caudate, putamen, and nucleus accumbens compared to other brain regions (Allen Brain Atlas). Human subject research also shows close relationships between D1 and D2 receptor availability and social behaviors. Striatal D2 receptor binding is negatively correlated with social desirability scores (Cervenka et al., 2010). In contrast, there is a positive correlation between social desirability and D1 receptors availability (Plavén-Sigray et al., 2014). People who are believed to be more socially desirable had higher D1 receptor availability (Plavén-Sigray et al., 2014). Because social groups cultivate social behaviors and are maintained through social behaviors, we believe that the given relationship between D1 and D2 MSNs and social behaviors can be extended to social group size.

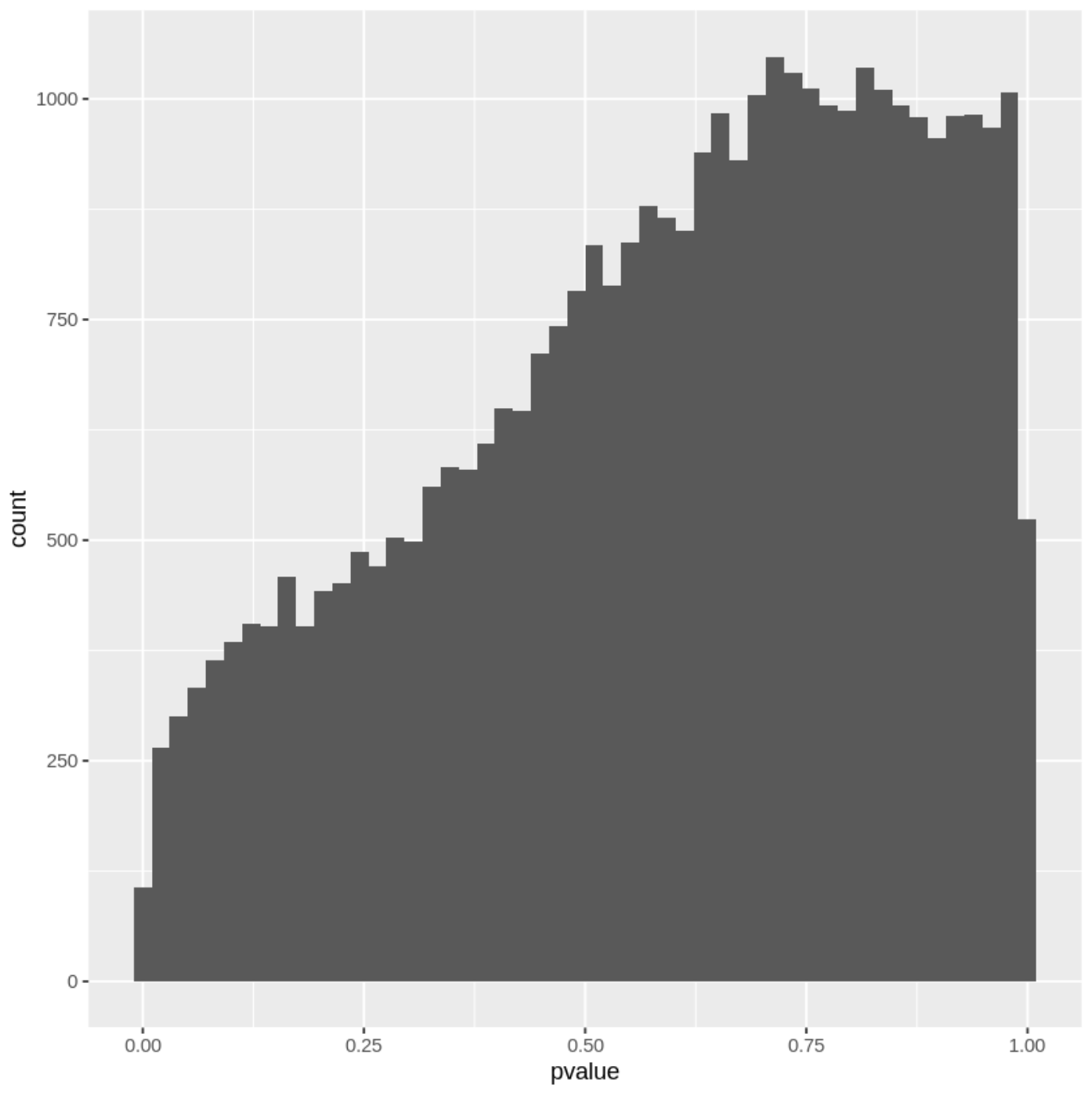
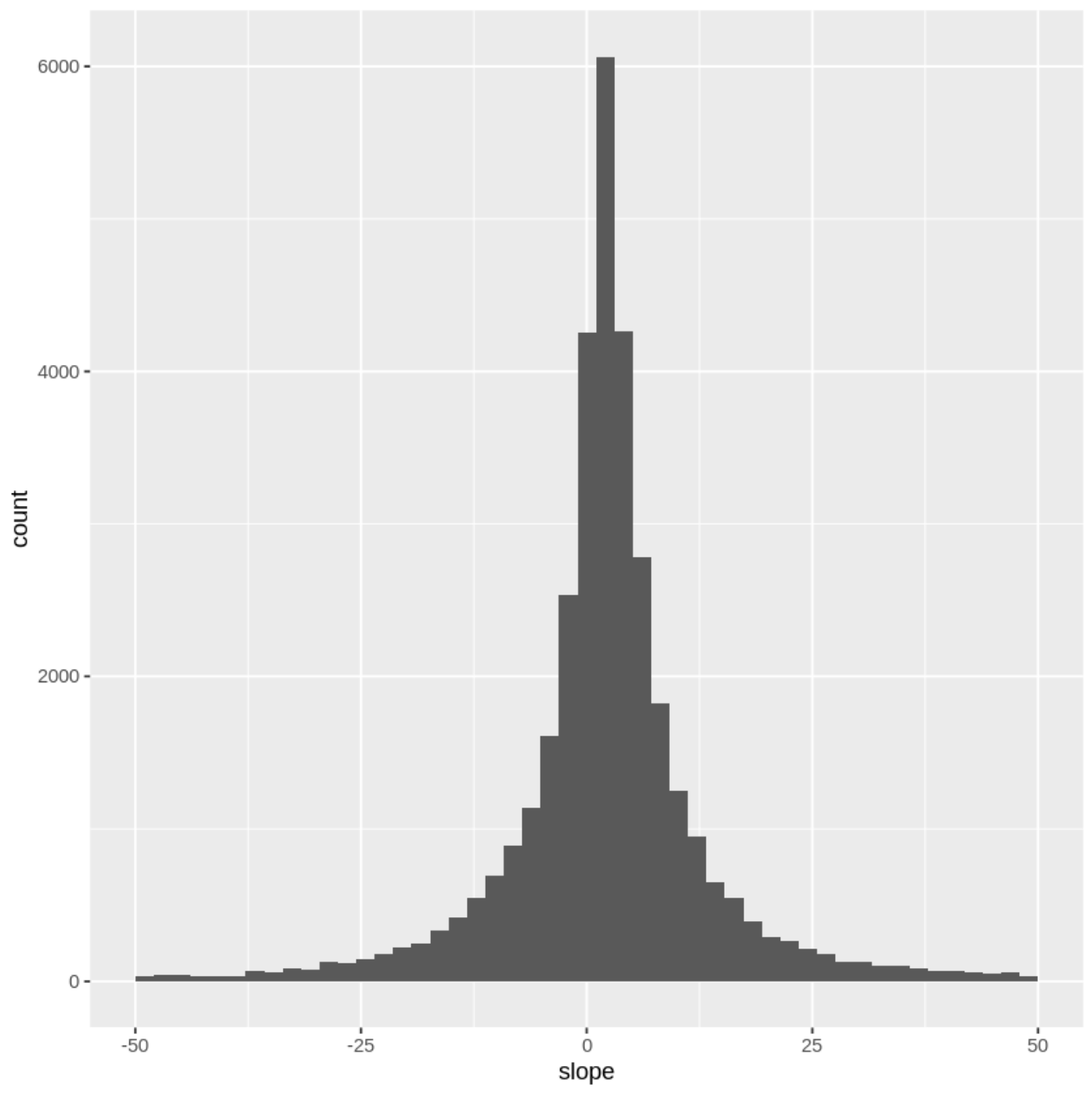
### Results

To search for differential expressions that are correlated with social group size, we used linear regression on each open chromatin peak in the ATAC-seq data, fitting predicted open chromatin levels in the six selected species against social group size. The predicted open chromatin levels at each ATAC-seq peak for each cell type are produced from a machine learning model and have been generated prior to this analysis. The slope of the fitted linear regression line represents the magnitude of differences, and the p value of the slope indicates the statistical significance. Because there is data missing at some peaks for some species, only the peaks with more than four (five or six) data points are included in the analysis to ensure that linear regression is meaningful.

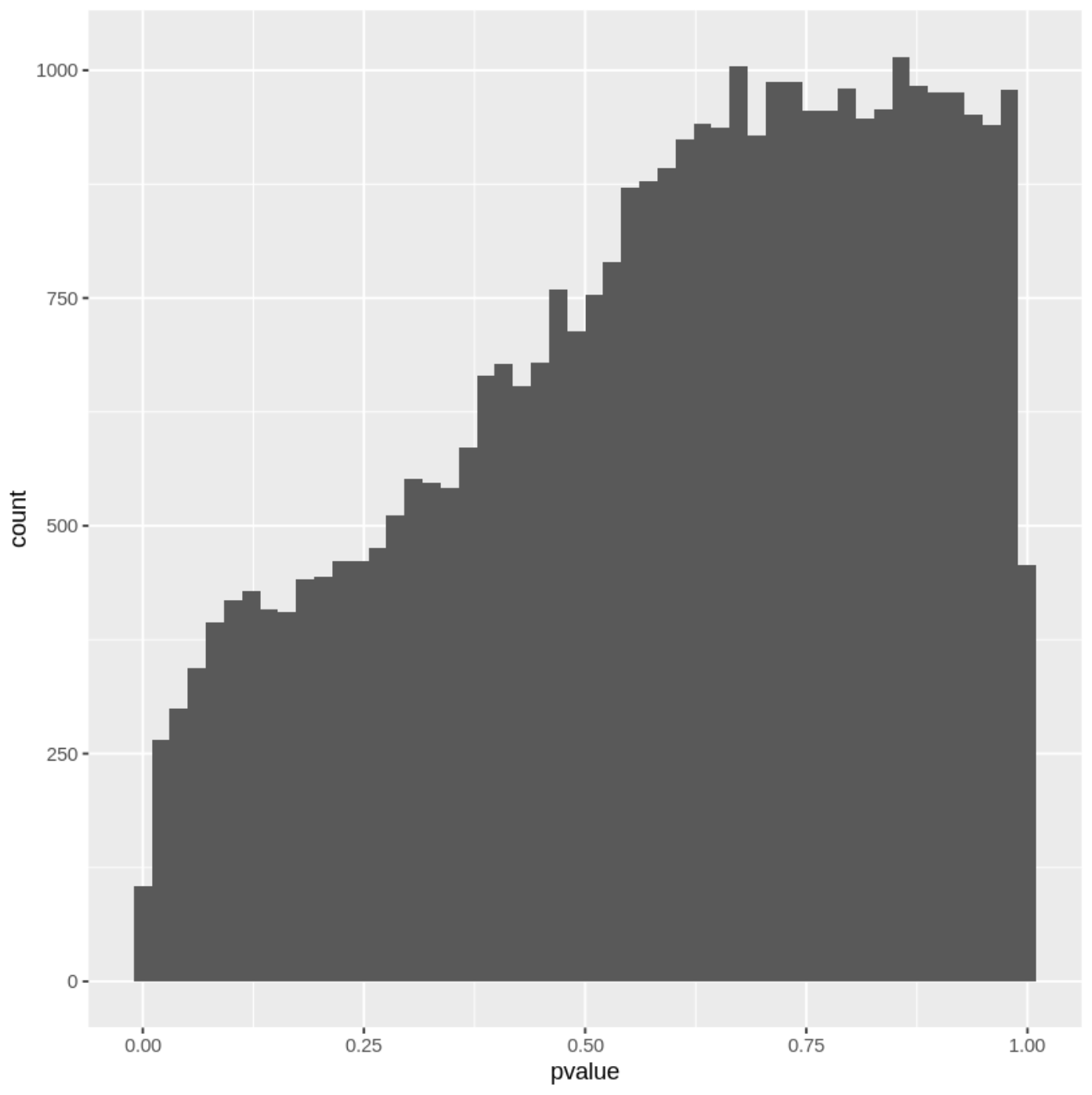
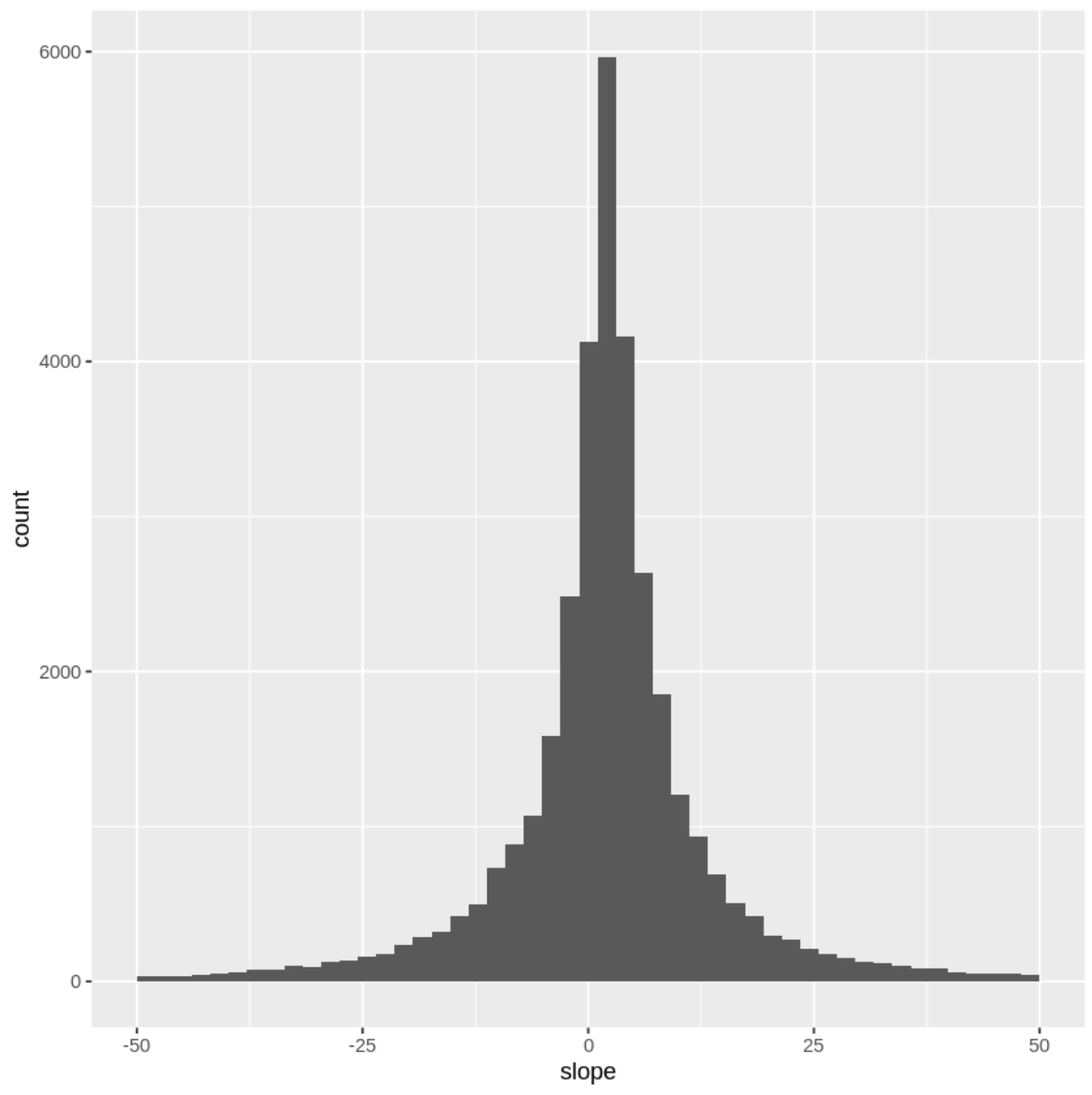
As indicated in figures 2a and 2c, the distribution of correlations is bell-shaped and symmetric with respect to 0. Therefore, there are approximately the same number of positive and negative correlations, and they tend to be weak because most correlations are around 0. There is no obvious difference between D1 and D2 MSNs.

If our data were to be randomly generated, then we expect to see a uniform distribution from 0 to 1 in the distribution of p values. The incline shown in figures 2b and 2d means that we are getting fewer significant p values (low p values) than random, implying considerable noise in the data. In addition, after correcting for multiple hypotheses, none of the adjusted p values is below 0.1 in either D1 or D2 MSNs. Therefore, strictly speaking, no peak has significant correlation to social group size in D1 and D2 MSNs. It is likely that the epigenetic differences resulting from evolution are so strong that they mask the signals of interest. It could also be the case that the study is statistically underpowered because the number of species is low.

a) b)



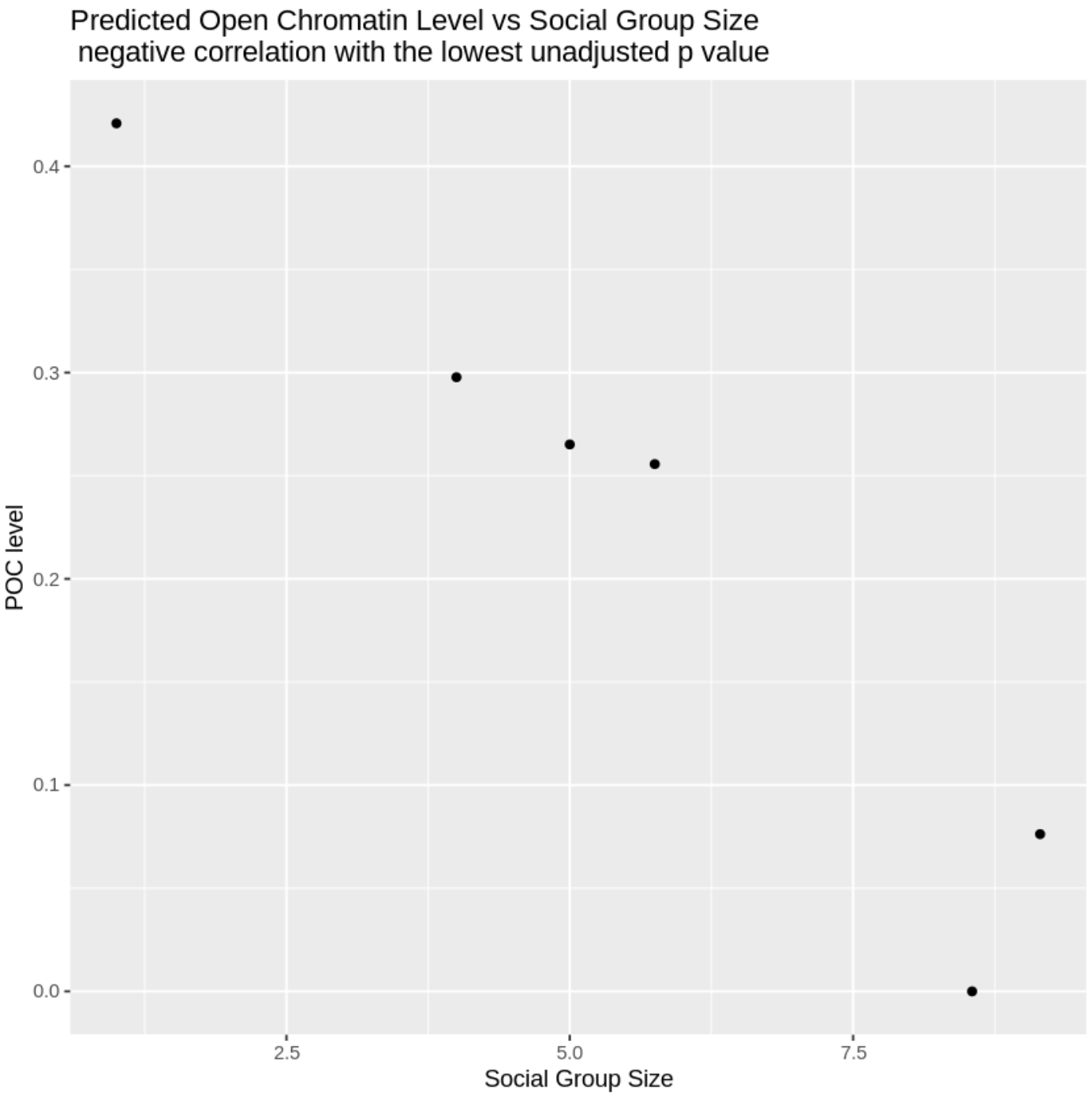
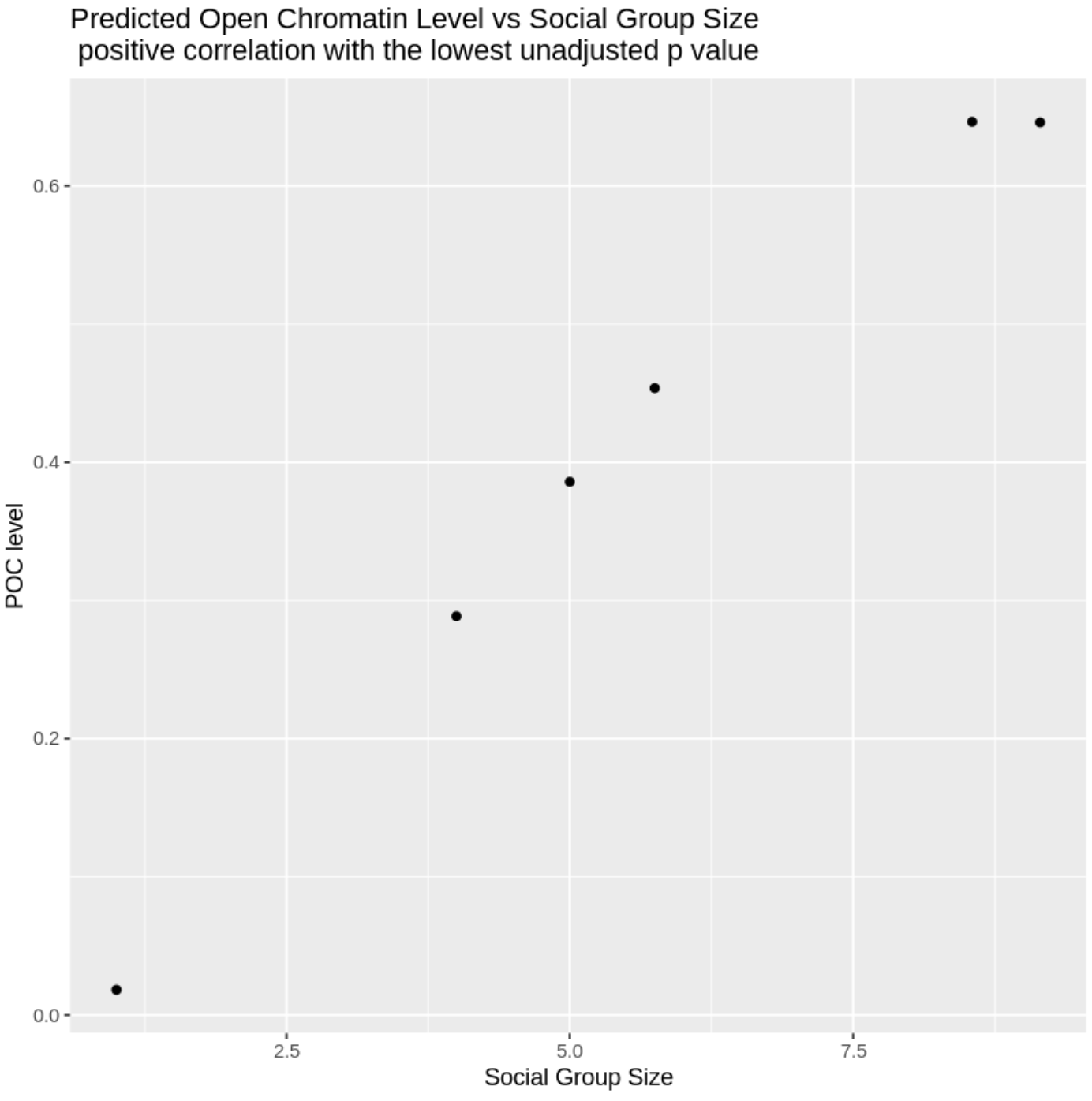
c) d)



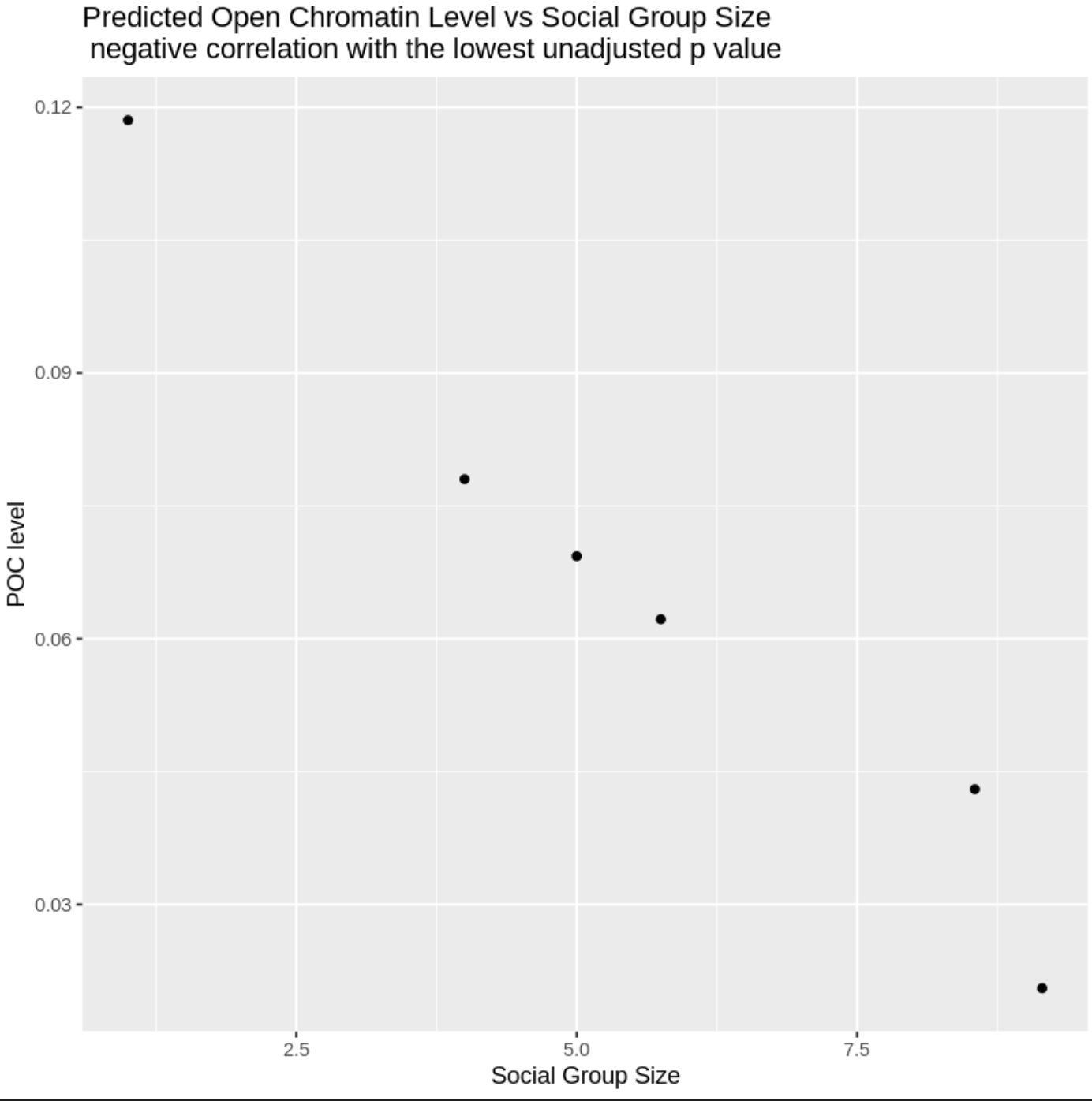
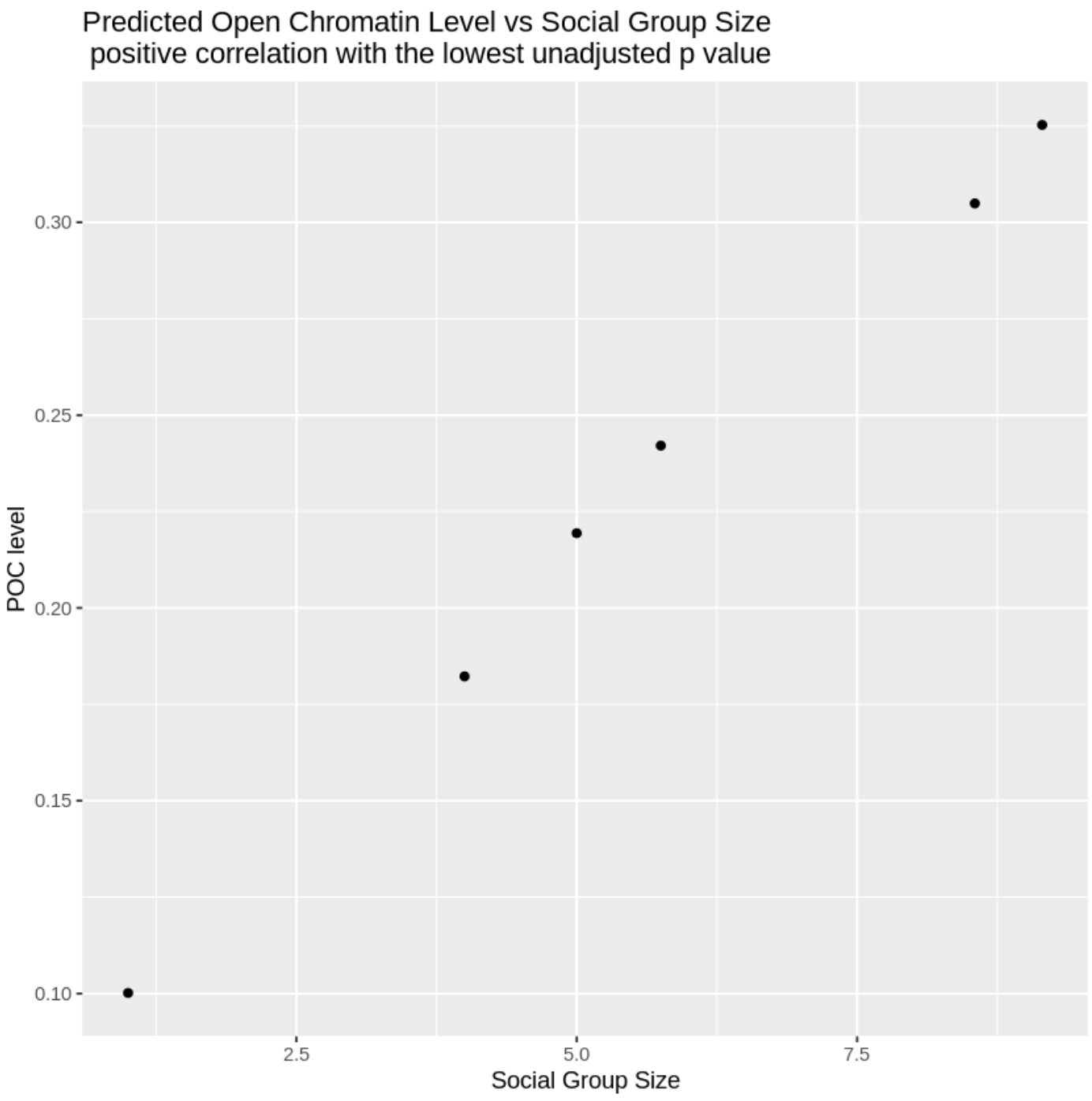
**Figure 2. a)** Distribution of the slopes by fitting predicted open chromatin levels in D1 MSN to social group size. **b)** Distribution of p values of the correlations in D1 MSN. **c)** Distribution of the slopes by fitting predicted open chromatin levels in D2 MSN to social group size. **d)** Distribution of p values of the correlations in D2 MSN.

Although the results from statistical tests are not ideal, it may still be worth investigating whether the top most positively and negatively correlated peaks are related to known biological processes that can affect social behaviors and social group size. First, we confirmed that the relationship suggested by the value of the correlation coefficient is consistent with that reflected by the raw data. The positively and the negatively correlated peaks with the lowest p value before adjustments are selected from each cell type, and the predicted open chromatin (POC) levels of these peaks are plotted against social group size. Positive correlations display an increase in POC levels (figures 3a and 3c), and negative correlations display a decrease in POC levels (figures 3b and 3d). Then, we searched for potential functions of genes that are likely to be regulated by the four peaks identified.

a) b)



c) d)

****

**Figure 3. a)** Plotting predicted open chromatin level against social group size of the most positively correlated peak in D1 MSNs. (hg38:chr3:1353150-1353650:250; p value: 8.864455e-05; adjusted p value: 1) **b)** Plotting predicted open chromatin level against social group size of the most negatively correlated peak in D1 MSNs.(hg38:chr7:104538106-104538606:250; p value: 7.502267e-04, adjusted p value: 1) **c)** Plotting predicted open chromatin level against social group size of the most positively correlated peak in D2 MSNs. (hg38:chr3:60378238-60378738:250; p value: 8.887571e-06; adjusted p value: 0.3128691) **d)** Plotting predicted open chromatin level against social group size of the most negatively correlated peak in D2 MSNs. (hg38:chr12:98,947,056-98,947,556:250; p value: 3.763250e-04; adjusted p value: 1)

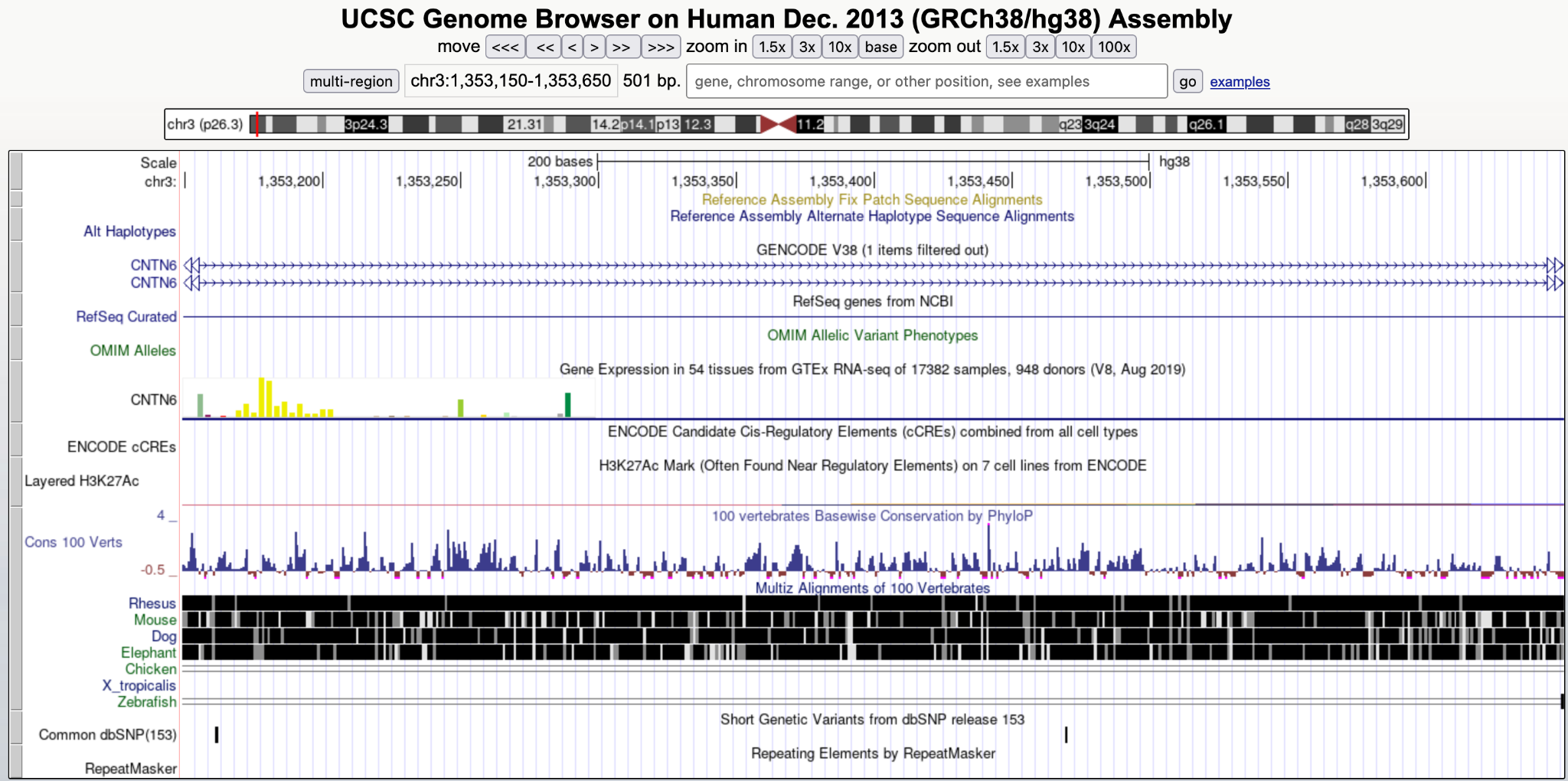
The peak most positively correlated with social group size in D1 MSNs is within an intronic region of CNTN6 (Figure 4a). Deletions and coding sequence variants in CNTN6 are enriched in individuals with autism spectrum disorders (ASD) compared to non-autistic controls (Mercati et al., 2016). CNTN6 encodes a neural cell adhesion protein that promotes neurite outgrowth and synapse formation, and it is especially important for sensory-motor pathway development in mice (Mercati et al., 2016). Because ASD is characterized by impaired social behaviors, expression level of CNTN6 can play a role in regulating social behaviors, and hence social group size.

The peak that is most negatively associated with the trait in D1 MSNs is located within an intron of LHFPL3 (Figure 4b), which can also be referred to as LHFPL4 (U.S. National Library of Medicine). LHFPL4 is found to be essential for binding to post-synaptic GABA A receptors and anchoring them to the post-synaptic membrane, forming to reliable inhibitory synapses (Davenport et al., 2017). There have been conflicting reports of the effects of LHFPL4 knockout on behaviors. Davenport et al. reported no behavioral changes in LHFPL4-/- mice (2017), whereas Wu et al. (2018) reported that complete knockout of LHFPL4 led to premature death, and incomplete knockout of it led to deficits in motor behaviors in mice. Further assessment of the behavioral consequences of LHFPL4 knockout and potential roles of LHFPL4 in producing animal behaviors, especially social behaviors, is needed. However, based on the function of LHFPL4, we can speculate its impact on social behaviors. Blocking GABA A receptor can lead to impaired sociability in rodents (Paine et al., 2020). Therefore, LHFPL4 expression level could be related to social behaviors as well.

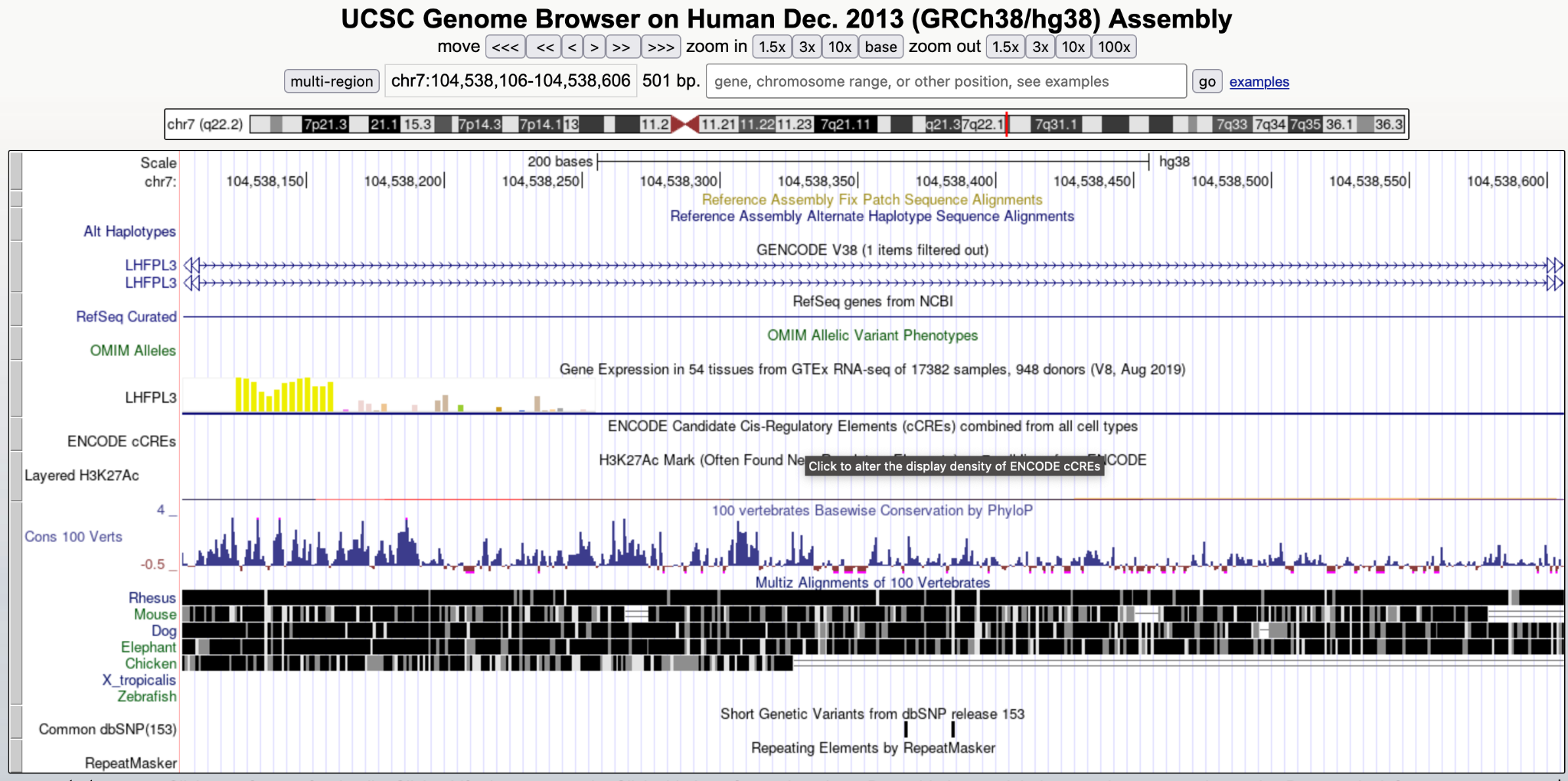
The peak that is most positively associated with the trait in D2 MSNs is in an intronic region of FHIT (Figure 4c). FHIT participates in purine metabolism and is a tumor suppressor (U.S. National Library of Medicine). It has also been reported that FHIT deletion is found in a pediatric patient diagnosed with autism spectrum disorders (Bolat & Bolat, 2021). Regression of normal language and social skills started when the patient was 2 years old. Given the association between FHIT and social skills impairment, it is possible that FHIT plays a role in social behaviors, and hence contributes to social group size.

The peak that is most negatively associated with the trait is in an intron of ANKS1B (Figure 4d). The protein ANKS1B encodes interacts with amyloid beta precursor protein and is implicated to play a role in schizophrenia, mood-disorders, obsessive-compulsive disorder, ADHD, autism spectrum disorders (ASD), and speech and motor deficits (Carbonell et al., 2019). This suggests that ANKS1B is important for a wide range of neurodevelopmental processes. Because an abnormal copy number of ANKS1B is associated with ASD, which is characterized by social skills deficits, ANKS1B plays a role in social behaviors. Therefore, it may affect social group size. However, because it affects diverse neurological processes, even if it does affect social group size, the effect may not be specific or direct.

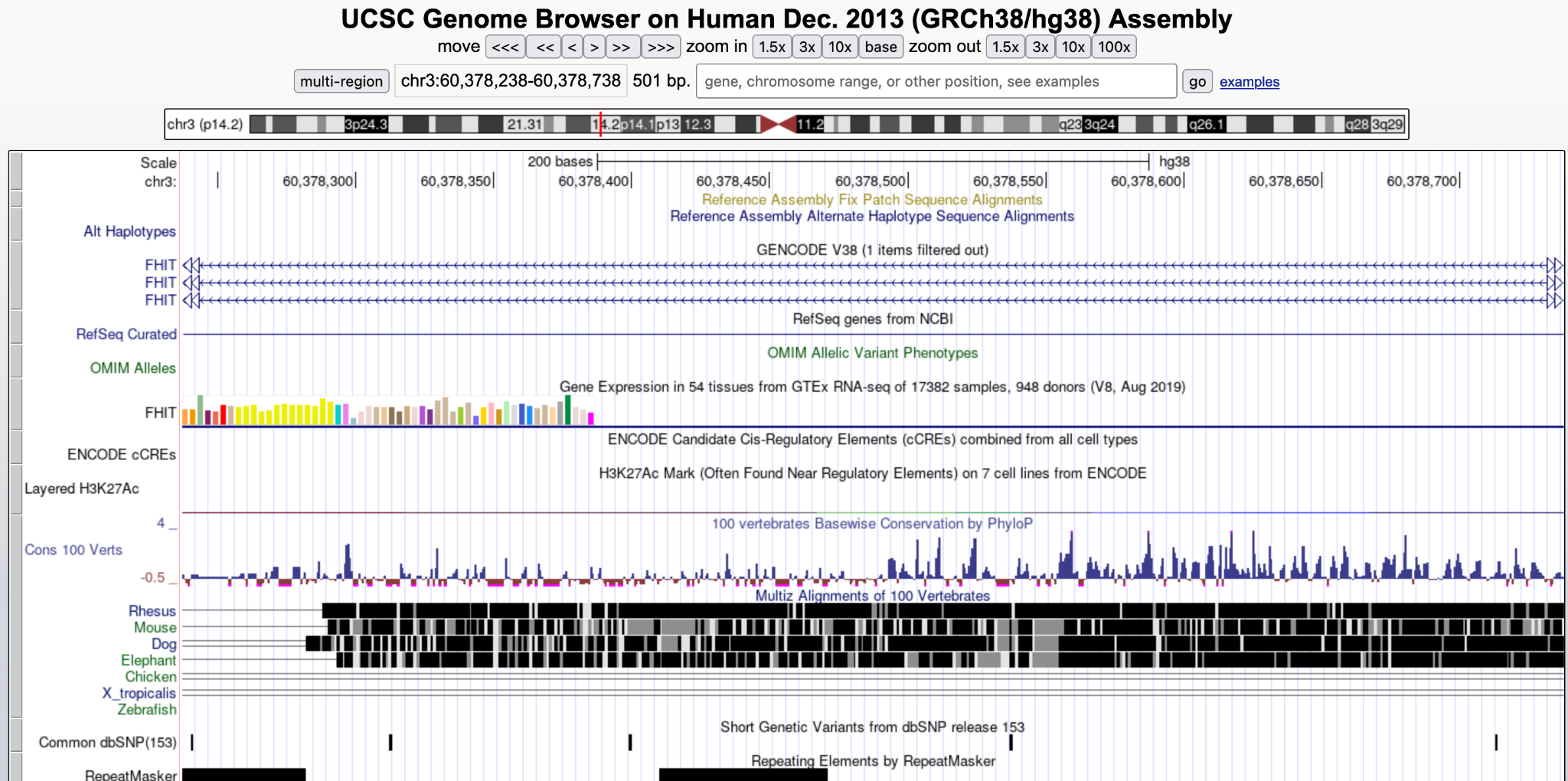
a)



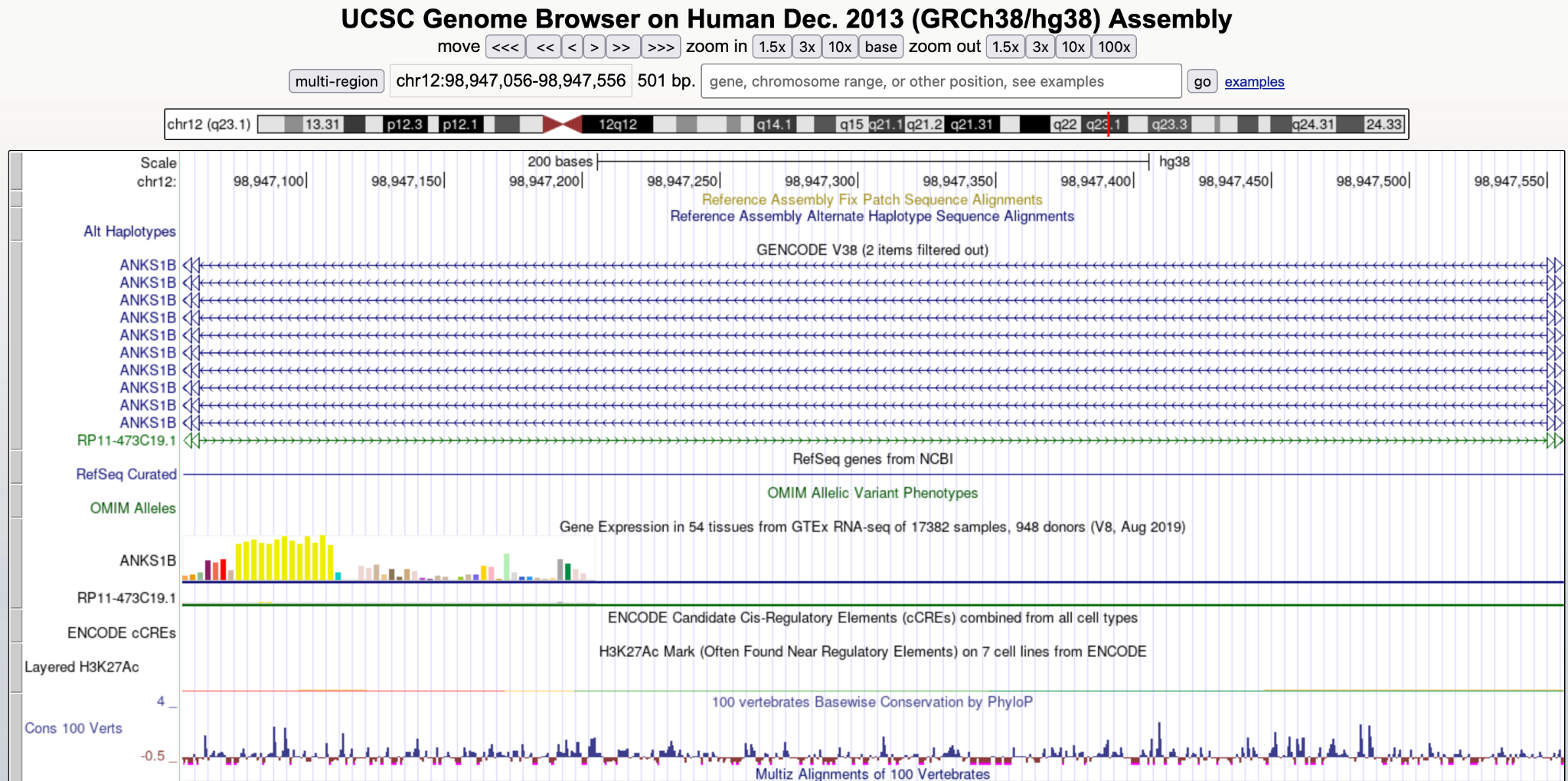
b)



c)

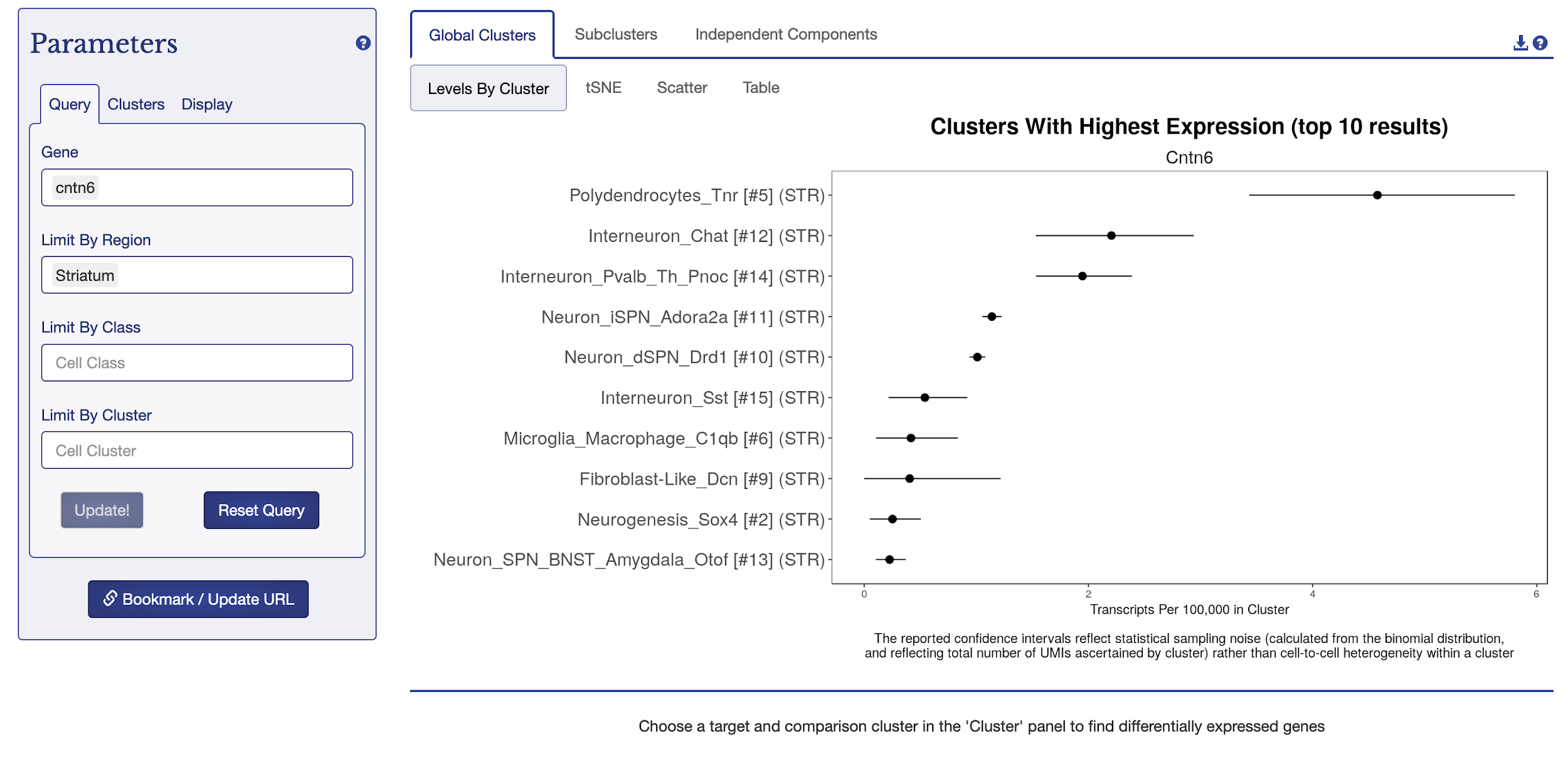
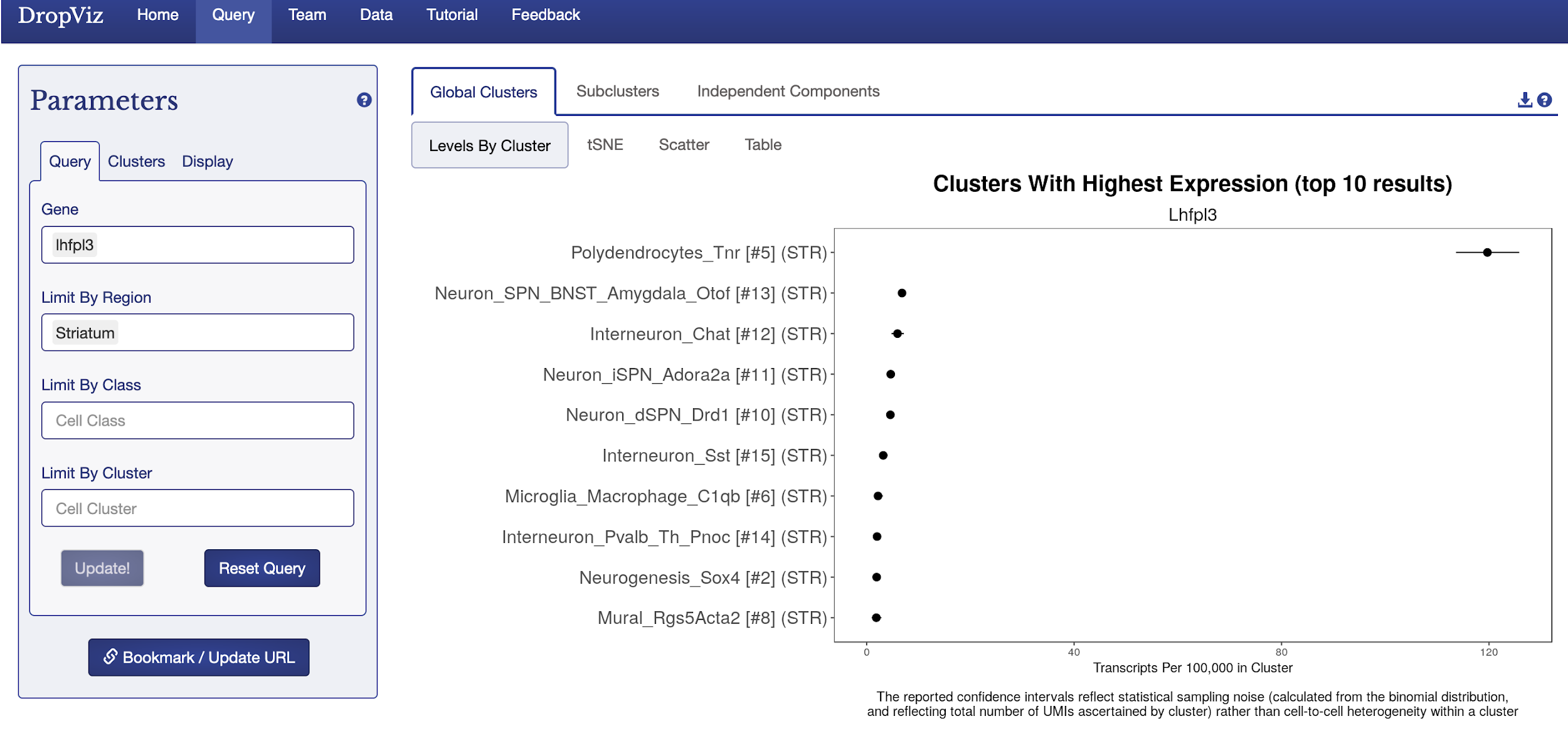
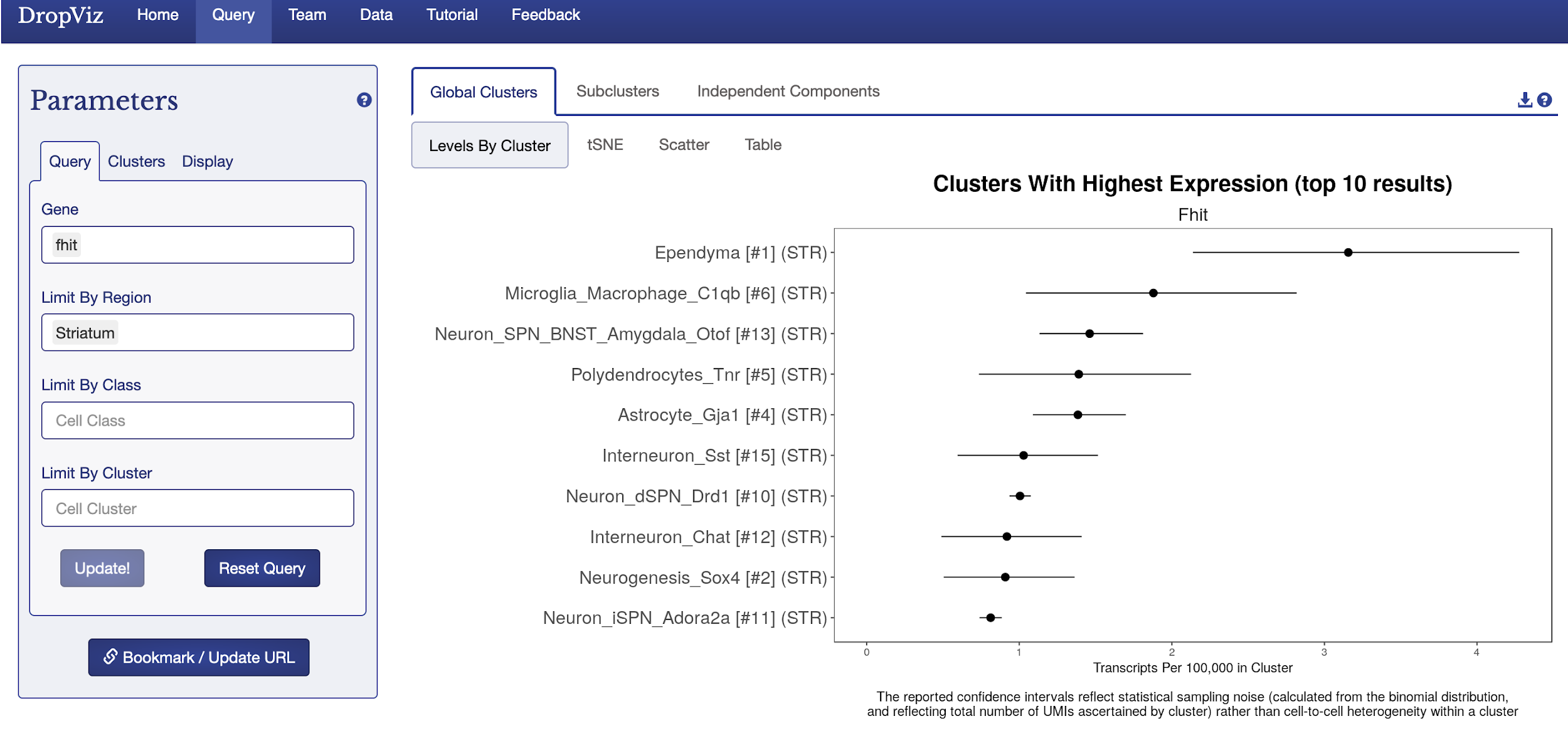


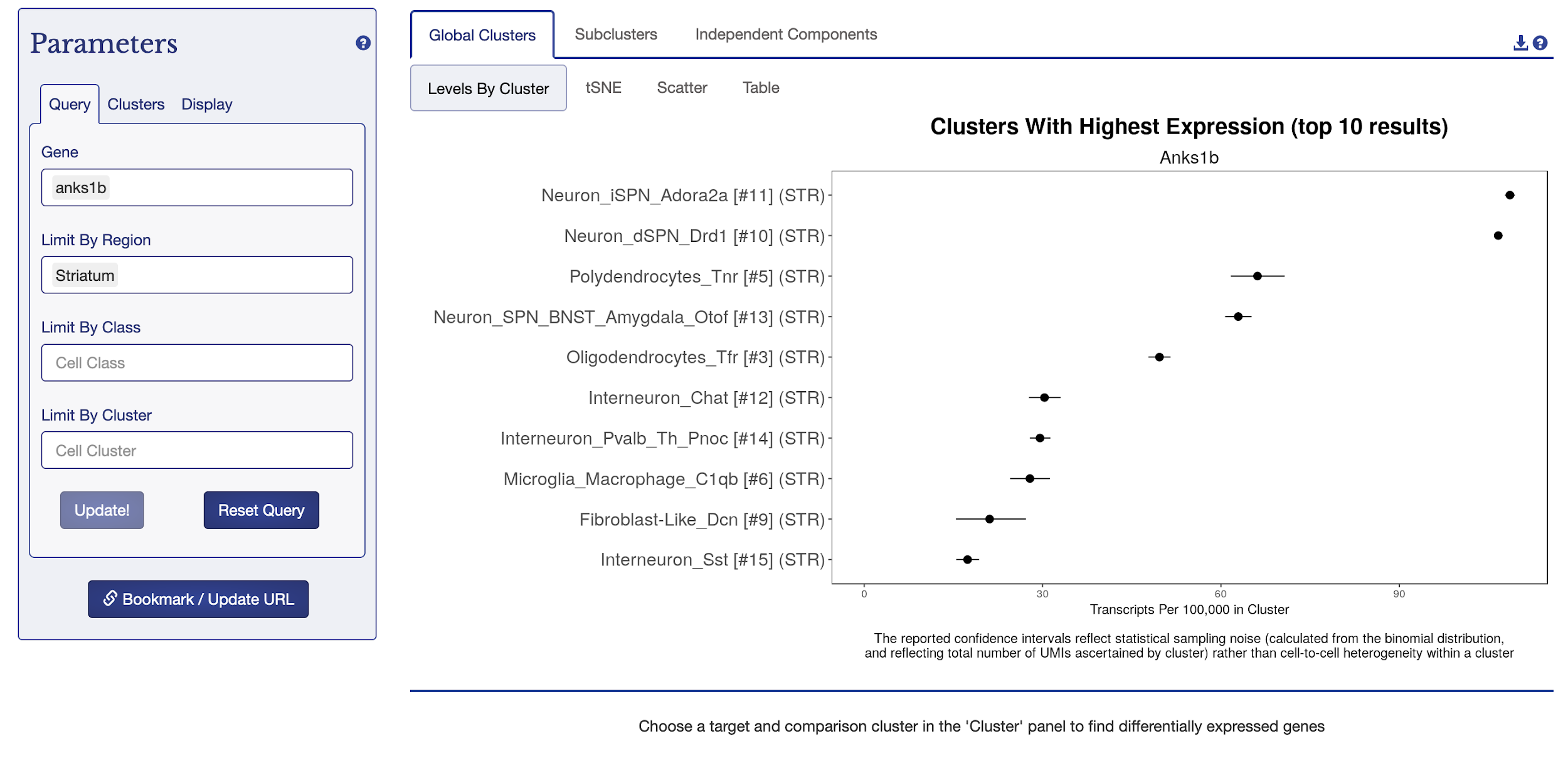
d)



**Figure 4.** Showing search results from USCS genome browser. The full span of an image shows the range of a peak. In other words, the view has not been zoomed in or out. **a)** the most positively correlated peak in D1 MSNs. (hg38:chr3:1353150-1353650:250; p value: 8.864455e-05; adjusted p value: 1) **b)** the most negatively correlated peak in D1 MSNs.(hg38:chr7:104538106-104538606:250; p value: 7.502267e-04, adjusted p value: 1) **c)** the most positively correlated peak in D2 MSNs. (hg38:chr3:60378238-60378738:250; p value: 8.887571e-06; adjusted p value: 0.3128691) **d)** the most negatively correlated peak in D2 MSNs. (hg38:chr12:98,947,056-98,947,556:250; p value: 3.763250e-04; adjusted p value: 1)

To confirm that the genes that are likely regulated by the top peaks are indeed differentially expressed in D1 and D2 MSNs, we used DropViz to visualize the differential expression of the four identified genes: CNTN6**,** LHFPL3**,** FHIT, andANKS1B. Although the expression levels in DropViz are based on mouse striatum, it serves as a good enough proxy for primates. However, there are some differences in naming conventions. Neuron\_dSPN\_Drd1 is the label for D1 MSNs, and Neuron ISPN\_Adora2a is the label for D2 MSNs. The expressions of all four genes are relatively high in D1 and D2 MSNs because they appear in the top 10 most expressed cell types for all four genes (figure 5), supporting the hypothesis that the peaks identified likely have some impact on the expression of these genes. However, ANKS1B exhibits the most dramatic differential expression in D1 and D2 MSNs (figure 5d).

a)****b)****c)****

d)****

**Figure 5.** Cell type-specificity of the genes identified in Figure 4 in the mouse striatum; results generated by DropViz. Neuron\_dSPN\_Drd1 is the label for D1 MSNs, and Neuron ISPN\_Adora2a is the label for D2 MSNs. **a)** CNTN6 **b)** LHFPL3 **c)** FHIT **d)** ANKS1B

Then, we ran ontology analyses on the most differential regions. Specifically, 200 most positively correlated peaks and 200 most negatively correlated peaks (ranked by unadjusted p values) are isolated from each cell type. The ontology analyses were conducted using the whole genome as background and repeated using all peaks identified in D1 and D2 as the background in the corresponding cell types. No biological processes were found using peaks in D1 or D2 as background (\*Supplementary 1). The only biological process identified using the whole genome as background is neuron development, when the top 200 most positively correlated peaks from D1 MSNs were the foreground (Supplementary 1). Because the whole genome was used as background, the process identified could simply be related to D1 MSNs instead of the top 200 peaks most correlated with social group size. In other words, it may be irrelevant to the research question. Because dendritic and axon growth and pruning are also neuron developmental processes, the identified biological process could reflect active synaptic changes in D1 neurons.

Graduate Analysis

Throughout the entire analysis, we have been emphasizing the role epigenetics of the brain plays in social behavior and thus also in social group size. Until this point, we have made the assumption that the data is collected from healthy subjects. However, mental illnesses may affect expressions of genes in the brain, such as schizophrenia. Schizophrenia is a detrimental disease that influences the way an individual interacts with their surroundings. One of the key symptoms of the disease is social/occupational dysfunction, marked by disrupted interpersonal relations, work, or self-care (DSM-V). Given the influence the disease has on social interactions, there is reason to believe that the genes associated with peaks having significant correlations with social group size could play a role in schizophrenia as well.

To determine if the gene expression is altered for individuals with Schizophrenia, we created a binary classifier to distinguish between non-schizophrenic and schizophrenic gene expression. The RNA-seq dataset provided generously by Cathy Su and Dr. Andreas Pfenning contains 17658 genes and their corresponding expression levels in nearly half a million prefrontal cortex cells. D1 and D2 MSNs are located in the striatum instead of the prefrontal cortex, so we cannot conduct the analysis on these two cell types only. Therefore, the analysis conducted makes no distinction between different cell types, and the results are thus general to neural cells. To simplify the problem and decrease runtime, only 100,000 observations were used to train machine learning models. Each of the observations were annotated with either non-schizophrenic (encoded to 0) or schizophrenic (encoded to 1). The data is relatively balanced, with 41% schizophrenic observations and 59% non-schizophrenic observations.

The baseline model for the dataset consisting of all genes was fed through a fully connected feedforward neural network. In short, this model connects each feature (gene) to a node in a particular hidden layer. The weights are distributed based on the hidden layers to give an overall assessment of the observation. For the neural network to be binary, the model outputs either a 0 or 1 based on the gene expression for a particular individual. With five hidden layers, the model is able to perform with a validation accuracy of 93.82% (Table 3).

Now with the baseline trained, we decided to investigate the four genes associated with top correlated peaks in D1 and D2 MSNs. Since the number of genes (features) is significantly reduced, a neural network is excessive for this problem. A simple logistic regression model is sufficient as a binary classifier for the low-dimensional data. A machine learning model is built on one of the four genes or both genes identified in a particular cell type (Table 3). The accuracy score is listed in Table 3.

| Dataset | Model | Score |
| --- | --- | --- |
| All Genes | Neural Network | 0.9382 |
| FHIT Only | Logistic Regression | 0.5905 |
| ANKS1B Only | Logistic Regression | 0.5892 |
| FHIT and ANKS1B | Logistic Regression | 0.5894 |
| CNTN6 Only | Logistic Regression | 0.5892 |
| LHFPL3 Only | Logistic Regression | 0.5922 |
| CNTN6 and LHFPL3 | Logistic Regression | 0.5829 |
| D1 genes associated with top 200 positively correlated peaks | Random Forest Classifier | 0.682 |
| D1 genes associated with top 200 negatively correlated peaks | Random Forest Classifier | 0.6692 |
| D2 genes associated with top 200 positively correlated peaks | Random Forest Classifier | 0.677 |
| D2 genes associated with top 200 negatively correlated peaks | Random Forest Classifier | 0.693 |

**Table 3.** Scores of model performance of models trained on different sets of genes.

Looking at Table 3, the models using only the genes most correlated to the social group size are poor as a standalone model because an accuracy score close to 0.5 is no better than random assignment. Using both genes identified from the same cell type is not sufficient to improve the model. Even though, as indicated by the baseline model (all genes in table 3), there are epigenetic differences in neural cells between schizophrenic and non-schizophrenic individuals, the gene expression from these particular genes cannot capture the differences.

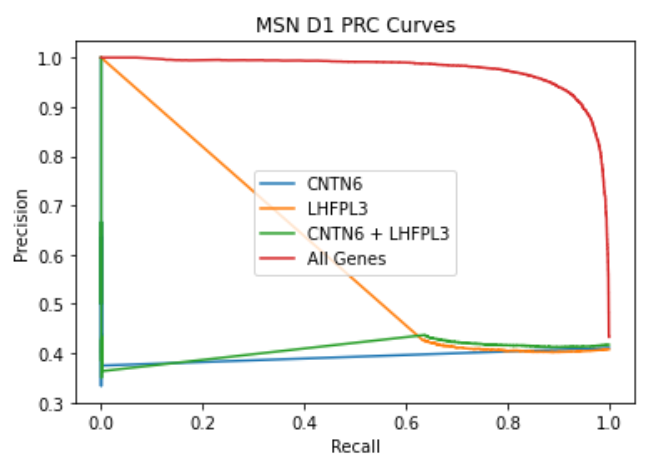
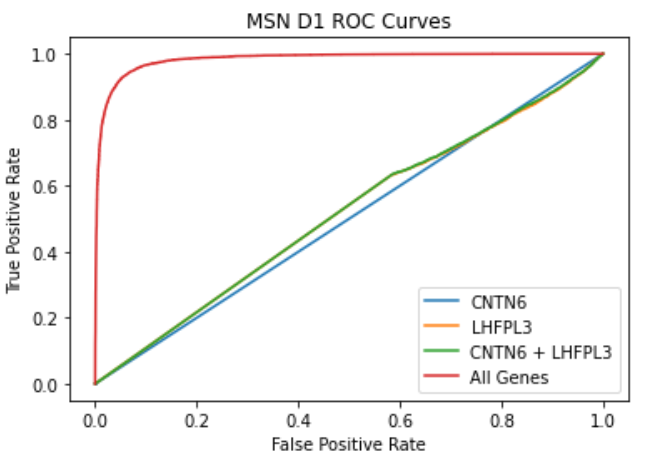
The performances of models trained on each of the four selected genes are consistent with the results from two-sample t-tests performed on normalized data. Because different numbers of cells are sequenced from different subjects and no distinction between cell types is made (as explained above), the counts for each gene from all cells of the same individual are added together and divided by the number of cells. Then, a t-test is performed on each of the four selected genes between normalized counts from schizophrenic and non-schizophrenic patients. The results are shown in Table 4. None of the p values is significant (<0.05). Therefore, these genes are not differentially expressed between schizophrenic and non-schizophrenic patients. Therefore, low performance of their respective models is to be expected. It is interesting that FHIT has the lowest p value from this analysis, and the peak it is associated with also has the lowest adjusted p value in the previous analysis. Even though neither is significant, FHIT shows a consistent trend towards significance and may be worth investigating further.

| Gene | P value from t-test |
| --- | --- |
| CNTN6 | 0.3415 |
| LHFPL3 | 0.4852 |
| FHIT | 0.1955 |
| ANKS1B | 0.2997 |

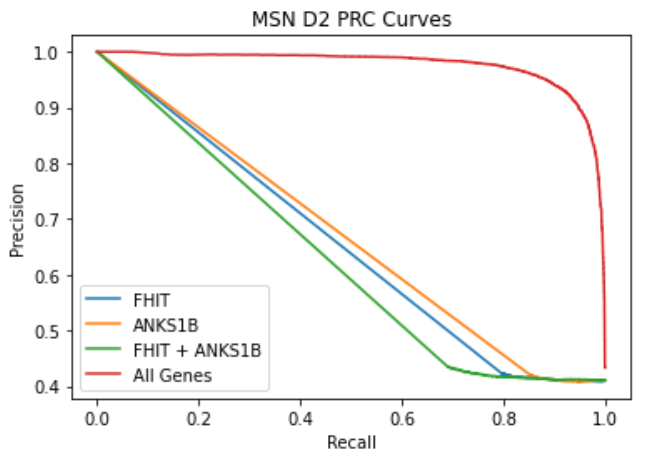
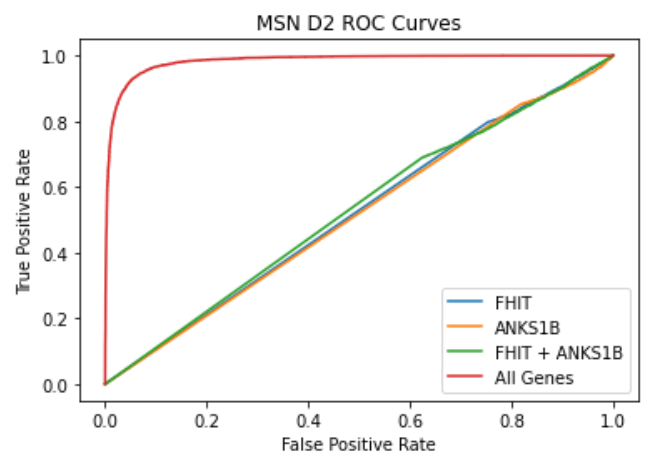
**Table 4.** P values from t-test performed on normalized data on each of the four selected genes.

Then, instead of using just the genes nearest to the top positively and negatively correlated peaks, we decided to build models using the genes associated with the top differential regions (200 positively correlated peaks and 200 negatively correlated peaks from each cell type). GREAT version 4.0.4 is used to identify the genes associated with these peaks. The parameters are: basal+extension: 5000 bp upstream, 1000 bp downstream, 1000000 bp max extension, curated regulatory domains included. The resulting 4 datasets (2 datasets for each neuron cell type) contains roughly 300 genes. The data has more features than the single featured data above and less features than the baseline. A good model to use for this data is a Random Forest Classifier. Random Forests are tree based machine learning models that split the data based on features to predict a given discrete or continuous label. The unique nature of Random Forest is that the method uses a committee of trees to predict a given label. This label follows the predictions of the majority of trees. In the case of this study, the hyperparameters of the tree depth is based on purity and the number of trees in the committee is 100.

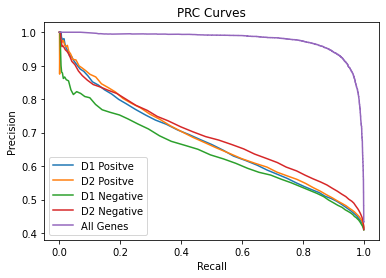
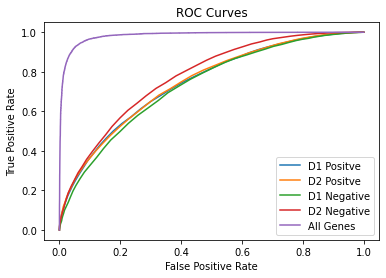
a) b)



c) d)



e) f)



**Figure 6. a)** The ROC curves for models trained on the genes associated with the most positively and negatively correlated peaks D1 MSN. **b)** The PRC curves for models trained on the genes associated with the most positively and negatively correlated peaks D1 MSN. **c)** The ROC curves for models trained on the genes associated with the most positively and negatively correlated peaks D2 MSN. **d)** The PRC curves for models trained on the genes associated with the most positively and negatively correlated peaks D2 MSN. **e)** The ROC curves for models trained on the genes associated with the top 200 positively and negatively correlated peaks in both D1 and D2 MSN. **f)** The PRC curves for models trained on the genes associated with the top 200 positively and negatively correlated peaks in both D1 and D2 MSN.

The models trained on genes associated with top differential regions performed better than those trained on single genes (Figure 6). For the D1 MSN datasets, the positive and negative correlated genes performed with a score of 0.682 and 0.6692 respectively (Table 3). For the D2 MSN datasets, the positive and negative correlated genes performed with a score of 0.677 and 0.693 respectively (Table 3). These scores are better than the previous datasets with only a single gene expression feature for the top correlation (Table 3). In figures 6e and 6f, we see the ROC and PRC curves for the 4 new models. The receiver operating characteristic curve (ROC) curve tracks the proportions of positive points correctly predicted (TP/(TP+FN)) against the proportions of negative points wrongly predicted (FP/(FP+TN) or false alarm) in the dataset. This metric serves well for datasets that have an even balance of observations for each class. A precision-recall curve (PRC) performs better at determining the accuracy of the model when there is data imbalance. As discussed earlier, this dataset is relatively balanced, but it is worth looking at PRC curves as well. Precision recall uses the metric precision and recall to calculate the curve. Precision is the number of observations that are predicted with a positive label compared to the total number of predictions made with a positive label (TP/(TP+FP)). Recall is the number of labels returned with the right label (TP/(TP+FN)). These models trained on genes associated with top differential regions perform better than random chance (straight line across the diagonal), and D2 negative performs marginally better than the others (Figure 6e and 6f). They also perform better than models trained on single genes, which have no better performances than random (Figure 6). The baseline performs near optimal in both the PRC and ROC plots. This is the case because the baseline has the most amount of information. From the models trained on genes associated with the top peaks in D1 and D2 MSN, we are able to predict well above random chance, even though these genes are not sufficient to explain all differential expressions between schizophrenic and non-schizophrenic patients. We would expect to see better results with more genes but it is worth noting that we are still only using 1.6% of the entire genetic data available. This significant increase in model prediction shows that the genes associated with social group size are also associated with schizophrenia phenotype.

Throughout the various analyses, there are no obvious differences between D1 and D2 MSNs. This is likely because D1 and D2 MSNs perform similar functions and so are epigenetically similar.

\*Supplementary 1: SupplementaryFile\_GREAT\_Results.docx

References

Park SA, Goïame S, O'Connor DA, Dreher JC (2017) Integration of individual and social information for decision-making in groups of different sizes. PLOS Biology 15(6): e2001958.<https://doi.org/10.1371/journal.pbio.2001958>

Dunbar R. I. M., Mac Carron Padraig and Shultz Susanne (2018) Primate social group sizes exhibit a regular scaling pattern with natural attractors *Biol*. Lett.142017049020170490

<http://doi.org/10.1098/rsbl.2017.0490>

Chapman, C. A. & Teichroeb, J. A. (2012) What Influences the Size of Groups in Which Primates Choose to Live? *Nature Education Knowledge* 3(10):9

Dunbar, R.I.M. (1998), The social brain hypothesis. Evol. Anthropol., 6: 178-190. [https://doi.org/10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8](https://doi.org/10.1002/(SICI)1520-6505(1998)6:5%3C178::AID-EVAN5%3E3.0.CO;2-8)

Cho, Y. T., Ernst, M., & Fudge, J. L. (2013). Cortico-amygdala-striatal circuits are organized as hierarchical subsystems through the primate amygdala. *Journal of Neuroscience*, *33*(35), 14017–14030. https://doi.org/10.1523/jneurosci.0170-13.2013

Shultz, S., & Dunbar, R. I. M. (2007). The evolution of the Social Brain: Anthropoid primates contrast with other vertebrates. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1624), 2429–2436. https://doi.org/10.1098/rspb.2007.0693

Liu, X., Liu, S., Huang, R., Chen, X., Xie, Y., Ma, R., Luo, Y., Bu, J., & Zhang, X. (2018). Neuroimaging studies reveal the subtle difference among social network size measurements and shed light on New Directions. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00461

Harding, L. E. (2012). Nomascus leucogenys (primates: Hylobatidae). *Mammalian Species*, *44*, 1–15. https://doi.org/10.1644/890.1

Korstjens, A. H., & Dunbar, R. I. (2007). Time Constraints Limit Group sizes and distribution in red and black-and-white colobus. *International Journal of Primatology*, *28*(3), 551–575.<https://doi.org/10.1007/s10764-007-9148-2>

*The Mammal Tree*. Zoonomia. (n.d.). Retrieved September 21, 2021, from<https://zoonomiaproject.org/the-mammal-tree-view/>.

Jones, K. E., Bielby, J., Cardillo, M., Fritz, S. A., O'Dell, J., Orme, C. D., Safi, K., Sechrest, W., Boakes, E. H., Carbone, C., Connolly, C., Cutts, M. J., Foster, J. K., Grenyer, R., Habib, M., Plaster, C. A., Price, S. A., Rigby, E. A., Rist, J., … Purvis, A. (2009). Pantheria: A species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology*, *90*(9), 2648–2648.<https://doi.org/10.1890/08-1494.1>

Dunbar, R. I. (2020). Structure and function in human and primate social networks: Implications for diffusion, network stability and health. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *476*(2240), 20200446.<https://doi.org/10.1098/rspa.2020.0446>

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. Neuron, 35(2), 395–405. https://doi.org/10.1016/s0896-6273(02)00755-9

Klein, J. T., & Platt, M. L. (2013). Social information signaling by neurons in primate striatum. Current Biology, 23(8), 691–696. https://doi.org/10.1016/j.cub.2013.03.022

Báez-Mendoza, R., &amp; Schultz, W. (2013). The role of the striatum in social behavior. Frontiers in Neuroscience, 7.<https://doi.org/10.3389/fnins.2013.00233>

Plavén-Sigray, P., Gustavsson, P., Farde, L., Borg, J., Stenkrona, P., Nyberg, L., Bäckman, L., & Cervenka, S. (2014). Dopamine D1 receptor availability is related to social behavior: a positron emission tomography study. *NeuroImage*, *102 Pt 2*, 590–595. <https://doi.org/10.1016/j.neuroimage.2014.08.018>

*Allen Brain Atlas*. Microarray data :: Allen Brain Atlas: Human brain. (n.d.). Retrieved October 8, 2021, from https://human.brain-map.org/microarray/search/show?exact\_match=false&search\_term=%22PDE10A%22&search\_type=gene.

Cervenka, S., Gustavsson, J. P., Halldin, C., & Farde, L. (2010). Association between striatal and extrastriatal dopamine D2-receptor binding and social desirability. *NeuroImage*, *50*(1), 323–328. https://doi.org/10.1016/j.neuroimage.2009.12.006

Gagnon, D., Petryszyn, S., Sanchez, M. G., Bories, C., Beaulieu, J. M., De Koninck, Y., Parent, A., & Parent, M. (2017). Striatal neurons expressing D1 and D2 receptors are morphologically distinct and differently affected by dopamine denervation in mice. *Scientific Reports*, *7*(1). https://doi.org/10.1038/srep41432

Sano, H., Nagai, Y., Miyakawa, T., Shigemoto, R., & Yokoi, M. (2008). Increased social interaction in mice deficient of the striatal medium spiny neuron-specific phosphodiesterase 10a2. *Journal of Neurochemistry*, *105*(2), 546–556. https://doi.org/10.1111/j.1471-4159.2007.05152.x

Soares-Cunha, C., de Vasconcelos, N. A., Coimbra, B., Domingues, A. V., Silva, J. M., Loureiro-Campos, E., Gaspar, R., Sotiropoulos, I., Sousa, N., & Rodrigues, A. J. (2019). Nucleus accumbens medium spiny neurons subtypes signal both reward and aversion. *Molecular Psychiatry*, *25*(12), 3241–3255. <https://doi.org/10.1038/s41380-019-0484-3>

Substance Abuse and Mental Health Services Administration. Impact of the DSM-IV to DSM-5 Changes on the National Survey on Drug Use and Health [Internet]. Rockville (MD): Substance Abuse and Mental Health Services Administration (US); 2016 Jun. Table 3.22, DSM-IV to DSM-5 Schizophrenia Comparison.

Bolat, G. Ü., & Bolat, H. (2021). The Role of Copy Number Variations and FHIT Gene on Phenotypic Characteristics of Cases Diagnosed with Autism Spectrum Disorder. https://doi.org/10.1159/000512171

Carbonell, A. U., Cho, C. H., Tindi, J. O., Counts, P. A., Bates, J. C., Erdjument-Bromage, H., Cvejic, S., Iaboni, A., Kvint, I., Rosensaft, J., Banne, E., Anagnostou, E., Neubert, T. A., Scherer, S. W., Molholm, S., & Jordan, B. A. (2019). Haploinsufficiency in the ANKS1B gene encoding AIDA-1 leads to a neurodevelopmental syndrome. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-11437-w

Davenport, E. C., Pendolino, V., Kontou, G., McGee, T. P., Sheehan, D. F., López-Doménech, G., Farrant, M., & Kittler, J. T. (2017). An essential role for the tetraspanin LHFPL4 in the cell-type-specific targeting and clustering of synaptic GABA A receptors. *Cell Reports*, *21*(1), 70–83. https://doi.org/10.1016/j.celrep.2017.09.025

*Dropviz*. DropViz. (n.d.). Retrieved November 30, 2021, from http://dropviz.org/.

Great input: Genomic regions enrichment of annotations tool, Bejerano Lab, Stanford University. (n.d.). Retrieved November 30, 2021, from http://great.stanford.edu/public/html/index.php.

Mercati, O., Huguet, G., Danckaert, A., André-Leroux, G., Maruani, A., Bellinzoni, M., Rolland, T., Gouder, L., Mathieu, A., Buratti, J., Amsellem, F., Benabou, M., Van-Gils, J., Beggiato, A., Konyukh, M., Bourgeois, J.-P., Gazzellone, M. J., Yuen, R. K., Walker, S., … Bourgeron, T. (2016). CNTN6 mutations are risk factors for abnormal auditory sensory perception in autism spectrum disorders. *Molecular Psychiatry*, *22*(4), 625–633. https://doi.org/10.1038/mp.2016.61

Paine, T. A., Chang, S., & Poyle, R. (2020). Contribution of GABAA receptor subunits to attention and social behavior. *Behavioural Brain Research*, *378*, 112261. https://doi.org/10.1016/j.bbr.2019.112261

U.S. National Library of Medicine. (n.d.). *FHIT fragile histidine triad diadenosine triphosphatase [Homo Sapiens (human)] - gene - NCBI*. National Center for Biotechnology Information. Retrieved November 30, 2021, from https://www.ncbi.nlm.nih.gov/gene/2272.

U.S. National Library of Medicine. (n.d.). *LHFPL3 LHFPL tetraspan subfamily member 3 [homo sapiens (human)] - gene - NCBI*. National Center for Biotechnology Information. Retrieved November 30, 2021, from https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=DetailsSearch&Term=375612.

UCSC Genome Browser Home. (n.d.). Retrieved November 30, 2021, from https://genome.ucsc.edu/.

Wu, M., Tian, H.-L., Liu, X., Lai, J. H., Du, S., & Xia, J. (2018). Impairment of inhibitory synapse formation and motor behavior in mice lacking the NL2 binding partner LHFPL4/Garlh4. *Cell Reports*, *23*(6), 1691–1705. https://doi.org/10.1016/j.celrep.2018.04.015