GEB Final Research Report
Rast and Min

Abstract

Herbivory describes animals that have adapted to subsist primarily or entirely upon plant material. While this is an ancient trait believed to have first appeared in the Mesozoic period more than 66 million years ago, herbivory has remerged from omnivorous or carnivorous lineages multiple times in an example of convergent evolution. Among mammals, herbivory is common in some clades such as ungulates, and rare in others like carnivoramorpha. We are therefore interested in instances where mammals have evolved herbivorous eating habits despite close evolutionary distance to omnivores or carnivorous ancestors. For example, the giant panda, a member of the Carnivora order, subsists almost entirely upon bamboo shoots despite maintaining the digestive system of a carnivore (Guo et al.). Likewise, the gorilla is a rare primate that subsists primarily on mostly foliage. Finally, the egyptian fruit bat (R. aegyptiacus) is a megabat that subsists on fruit. Foraging behavior and diet have been implicated largely in D2 medium spiny neurons and expressing D1 and D2 medium spiny neurons, so we have chosen to study the effect of epigenetic regulation in these cell types for dietary behavior. In this work, we compare the predicted epigenetic regulation of D1 and D2 medium spiny neurons in the striatal neural region between herbivorous and non-herbivorous animals. We identify highly differently expressed open chromatin regions and assess statistical significance using a t-test. We assess the gene ontology of genes near transcriptionally regulated areas. While herbivory is a readily apparent trait, identification of regulatory genes involved in diet may have wide-ranging implications. Furthermore, mastery of these techniques and a positive result on this trait may demonstrate feasibility in the study of more complex traits, such as neurological disorder or somatosensation.
In terms of what we learned from our data analysis, we learned that the top D2 upregulated gene is closest to the Spata13 gene which is related to submissive behavior and that the top D2 downregulated gene is closest to LRIG2. The top D1 upregulated gene is closest to DLGAP4 which is a signaling molecule whose knockout has caused intense aggressive behavior, and the top D1 downregulated gene is MAPK1. There were no significant categories in the D2 upregulated genes and in the upregulated D1 genes when the background was the entire genome, and in the downregulated D1 genes in general, but there were significant categories in the D2 upregulated genes when the background was the entire genome; one of these categories is a GO cellular component category, NMDA selective glutamate receptor complex. There were many significant categories for the upregulated D1 genes when the background was just all D1 peaks; in general, the number of categories saw a significant increase when the background was changed from all peaks to just all D2 or D1 peaks. Some categories we found that were related to the cell types of interest were GRIN2B, NLGN1, and DSCALM1.

Introduction

• Background on the trait of interest (from project trait checkpoint)

Herbivore refers to a kind of biological interaction, usually refers to an organism that mainly feeds on plants. Herbivores have adapted to plant substances as their main food ingredients in anatomy and physiology. Therefore, in the long evolutionary process, the mouthparts, digestive system and appearance are formed to be suitable for edible plants. Herbivores can be further divided into graminivore, folivore, frugivore, granivore, nectarivore, etc. In fact, many herbivores eat different parts of plants, such as roots and seeds. The eating habits of some herbivores will change with the seasons, especially in temperate regions, where they will have different foods at different times.

Phenotypic convergence between distantly related taxa often mirrors adaptation to similar selective pressures and may be driven by genetic convergence, including the same metabolic and regulatory pathways, protein-coding genes, or even identical amino acid substitutions in the same gene (Hu et al).

They co-evolved complex protective mechanisms to avoid or resist the plant's defensive traits. Although protective chemicals extracted from plants can cause serious negative effects, such as liver damage, heart failure, or even death, herbivores may still consume them because of dietary advantages, palatability. Basically, taste, olfaction, and somatosensation in the oral and nasal cavities, as well as post-oral chemical sensory mechanisms of the digestive tract, are used to detect and avoid potentially poisonous plants (Startek et al.).

Significant evidence has demonstrated the emergence of the herbivory trait as the result of convergent evolution in mammals (Ley et al, Muegge et al). While herbivory is common in the animal kingdom, gut microbiome analysis does not show significant correlation with phylogenetic tree, unlike that which would be expected if dietary shifts occurred rarely from a common ancestor. Rather, gut microbiomes can be grouped by function, not evolutionary distance. Further evidence of convergent evolution of the herbivory trait can be found from examination of the phylogenetic tree itself. Several herbivorous species such as A. melanoleuca (Giant Panda) and G. gorilla (gorilla) have carnivorous or omnivorous ancestors.

Due to the large number of plants, herbivores suffer less pressure to survive in terms of food, unlike carnivores that have greater pressure to survive, which promotes evolution. The vision of herbivores is mostly monocular. The eyes are located on both sides of the face, and they maintain independent vision. What they see is the scene on both sides, which leads to their poor positioning of external things, and the eyesight of most herbivores is not very good. Evolution makes carnivorous mammals generally smarter, but carnivores include not only carnivorous mammals, but also more amphibians and reptiles. They appeared earlier in the evolutionary process, so the brain is very basic and has not evolved. The cerebral cortex, roughly equivalent to the structure of the human brainstem, has recorded many behaviors that do not require intelligence. Judging from the structure of the brain, their intelligence is far less than that of herbivorous mammals, and the sensory system is not as developed as herbivores.

Since most herbivores do not lack food, the evolutionary pressure seems to be less. Predation by ferocious beasts does not affect the continued existence of the population, because most of

the beasts catch the old, weak, sick and disabled. Healthy herbivores can easily escape the attack of beasts because they are faster and larger in size. Natural selection has not allowed most herbivores to develop in the direction of continuous brain evolution. However, some of the herbivores have intelligence standing at the apex of wild animals in nature. Elephants and chimpanzees are such species. They appeared later in evolutionary history. The brain has been more fully evolved and has a relatively high complexity.

• Background on species of interest (from project trait checkpoint)

Giant panda - Ailuropoda melanoleuca
Giant pandas are extremely professional herbivores. Although they are mainly from carnivores, they almost exclusively feed on high-fiber bamboo. Giant pandas feed on plants, but the protein and carbohydrate content of this diet looks more like a high carnivore that gets more than 70% of their diet from other animals. Giant pandas have herbivore characteristics, including a masticatory apparatus with powerful muscles, a robust mandible, large but flat teeth and a "thumb" that enables dexterous handling of bamboo. They also lost the ability to taste umami, which is usually related to meat (Hu et al.).

Western Gorilla - Gorilla gorilla
Western gorillas are great apes with a high degree of similarity to human beings. This species of non-human primates are ground dwelling herbivores. Gorillas live in complex social structures known as troops, and their diet consists primarily of leaves, stems, and little fruit. Unlike many prey herbivores, gorillas exhibit complex characteristics found in other primates such as tool use, emotional capability, and sociability. They live and move around as a family unit for safety, protection and peace. When danger comes, male gorillas will call out as an alarm to all members, and they can also avoid some enemies by climbing trees.

Egyptian Fruit Bat - Rousettus aegyptiacus
The egyptian fruit bat is a frugivore, meaning that its diet consists nearly entirely of fruits. It is nocturnal, and forages for its diet beginning at dusk. Commonly it will eat loquats, figs, or wild dates. It is notable for the ability to travel long distances to feed from known abundant food sources, up to 20km. Additionally, the fruit bat will consume as much as 150% of its body weight in fruit. R. aegyptiacus is prey to several predators such as hawks, owls, or falcons. They are also notable for their use of echolocation to navigate and avoid predators, as one of the only megabats to do so.

Background species
Rather than performing a paired analysis of closely related herbivores and non-herbivores, our group found better results when comparing a select number of herbivores against a broad background set of non-herbivores or carnivorous species. This motivated our selection of background species that did not share close evolutionary distance, such as C. Lupis (wolf), M. mungo (banded mongoose), H. parvula (dwarf mongoose), S. suricatta (meerkat), C. ferox (fossa), and H hyaena (hyena). Rationale for this selection is that emergence of herbivore as a neural trait, as seen in Gorilla and Giant Panda (both of whom share digestive tracts capable of digestion of non-plant matter) is a rare evolutionary phenomenon. Therefore, a broad set of

background species can be used for comparison, rather than targeted paired comparison between closely related species. Empirically, we observed strong statistical correlation when using a broad set of background species rather than a paired test.

• Background on cell type or cell types of interest and potential link to the trait (from cell type checkpoint)

Overview of cell type 1
The striatum is composed of 95% spiny projection neurons (Saunders), which can be classified into several subtypes. Of these, the D1 and D2 “direct” and “indirect” pathway neurons form the majority of cell subtypes. These neurons are named for the dopamine receptors expressed on their surface, which has implications for neural function. For our first cell type analysis, we have selected a study of the D1 medium spiny neuron, a cell in the “direct” striatum pathway. The effect of the D1 dopamine receptor on the neuron is not totally characterized, however, it has been shown to enhance network connections that are consistently active by augmenting membrane potential into an elevated “up” state, facilitating the triggering of action potentials (Surmier). These cell types are largely found in the “direct” pathway which wires the nucleus accumbens to the substantia nigra.

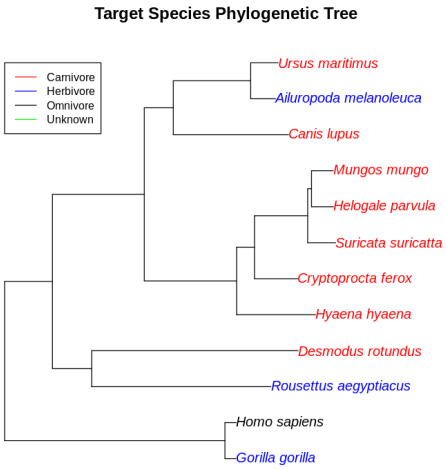
Relate cell type 1 to trait of interest
In general, the striatum has been widely implicated in dietary behaviors such as feeding, decision making, and satiation (Johnson et al.). Downregulation of striatal cell types has been shown to cause compulsive eating in murine models. More specifically, the D1 direct pathway has been widely associated with reward-seeking behavior and the activation pathway, in particular in the ventral striatum. Direct connection between striatal cell types and herbivory has been demonstrated in the implication of striatal GPR88 expression with foraging behavior (Rainwater et al.). In the work of Rainwater et al, it was found that mice with expression of GPR88 knocked out lost ability to discriminate between high caloric and low caloric rewards in a food-reinforced lever-pressing task. The inability to discern high-calorie foods was hypothesized to originate from the striatum playing a role in the action selection network. The task relates to dietary preference as herbivory or carnivory in animals with carnivorous digestive tracts such as giant panda is an action selection pathway. We hypothesize that this subclass of herbivorous animals expresses dietary preferences due to foraging action selection networks in the striatum. However, existing research does not correlate foraging behavior with cell type, as GPR88 is expressed in all medium spiny neuron cell types (Quintana et al.). We, therefore, seek to investigate with greater nuance the role of the striatum in foraging behavior.

Overview of cell type 2
The D2 MSNs are a subset of GABAergic inhibitory cells. A full ninety-five percent of the human striatum is composed of MSNs. D2 MSNs are responsible for controlling the indirect pathway within the basal ganglia. D2 MSNs that are in the indirect pathway inhibit their basal ganglia output structure. Therefore, any relevant behaviors are inhibited. The dorsal striatal D2 MSNs terminate bodily, limb, and eye movement. A variety of brain regions send excitatory input to the indirect pathway. The D2 MSNs project to the external global pallidus which, in turn, projects to

the excitatory subthalamic nucleus which then projects to the internal global pallidus and the substantia nigra. An inactivated indirect pathway means that the external global pallidus suppresses subthalamic nucleus activity. This means that downstream substantia nigra and internal global pallidus activity are decreased which, therefore, allows for higher activity in the thalamic and motor cortex neuron. In contrast, external global pallidus neurons are inhibited upon the firing of indirect pathway neurons. The subthalamic nucleus is thus disinhibited and it then starts the excitation of neurons in the substantia nigra and internal global pallidus. This results in the suppression of thalamus and motor cortex activity.

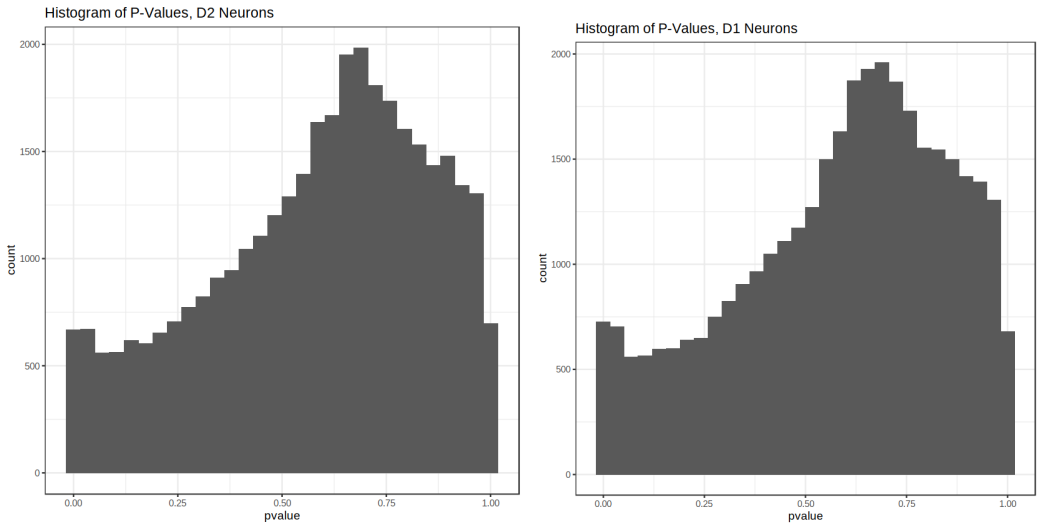
Relate cell type 2 to trait of interest
The dorsal striatal D2 MSNs are responsible for the regulation and termination of eye movement. This is related to herbivory because herbivores have evolved to have mostly monocular vision and eyes on the sides of their face. The dorsal striatal D2 MSNs were definitely involved in this process since they have evolved to control eye movement in this manner and terminate eye movement such that the eyes end up where they need to be to scan the environment. Herbivores need to view everything around them on both sides to watch out for predators. The dorsal striatal D2 MSNs are also integral in the regulation and termination of movement of other body parts. The mouthparts in particular are important for herbivores since these have evolved such that herbivores can consume plants. The dorsal striatal D2 MSNs control and terminate mouth movement to allow herbivores to properly eat plants. The ventral striatal D2 MSNs, meanwhile, are responsible for recognizing aversive stimuli. Herbivores need to recognize aversive stimuli since they need to use various mechanisms such as taste, olfaction, and somatosensation in the oral and nasal cavities and post-oral chemical sensory mechanisms in the digestive tract in order to provide a defense against plants that may very well be poisonous. The ventral striatal D2 MSNs are related to this because herbivores need to recognize poisonous plants as aversive stimuli in order to refrain from eating plants that may harm them. Another form of aversive stimuli for herbivores are predators. Herbivores also have to learn to recognize predators as aversive stimuli so that they are not killed by them.

• Visualization of species of interest in the evolutionary tree (from various checkpoints)



Results

• Calculate the p-values and magnitude of differences for each region for each cell type. Visualize them in histograms and discuss the results (from analysis checkpoint)



• Correct for multiple hypotheses and discuss the results (from analysis checkpoint).

D1

Id	meanDiff	pvalue	padj
hg38:chr20:36425128-36425628:250	0.46648127	1.611147e-06	0.05631441

hg38:chr7:126719514-126720014:250	**0.20255775**	**2.688966e-06**	**0.09398473**
hg38:chr13:76672360-76672860:250	**0.41046444**	**3.525137e-06**	**0.12320705**
hg38:chr8:85054369-85054869:250	**0.29757811**	**3.777458e-06**	**0.13202216**
hg38:chr4:146267531-146268031:250	**0.07708104**	**8.693857e-06**	**0.30384160**
hg38:chr3:192816221-192816721:250	**0.43182644**	**1.632325e-05**	**0.57046497**
hg38:chrX:19593645-19594145:250	**0.11702679**	**2.347403e-05**	**0.82034691**
hg38:chr2:211928139-211928639:250	**0.14099885**	**2.846728e-05**	**0.99481759**
hg38:chr22:21825556-21826056:250	**-0.06531153**	**5.334995e-01**	**1.00000000**
hg38:chr8:41731592-41732092:250	**0.21223869**	**1.803290e-01**	**1.00000000**

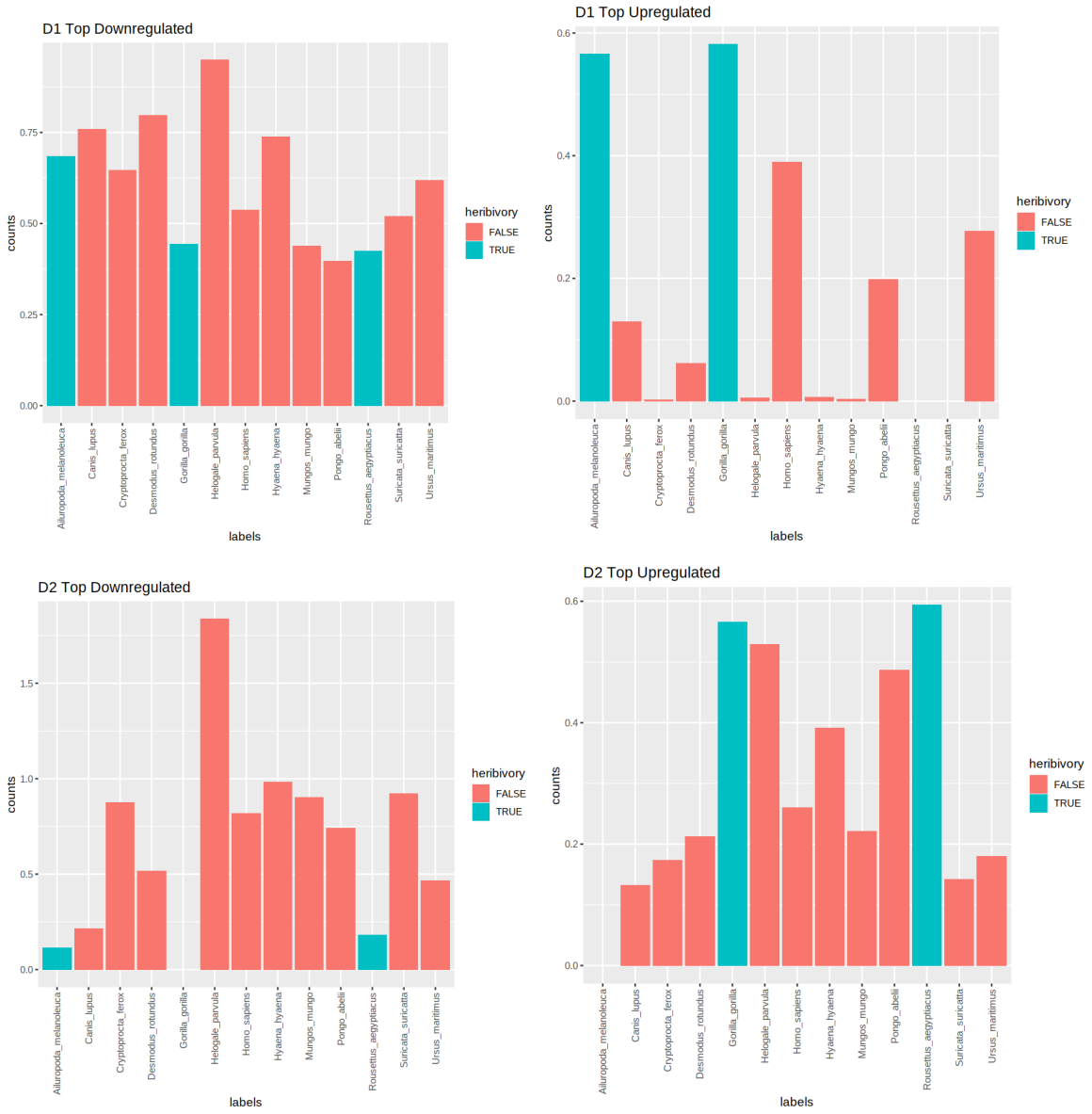
D2

Id	**meanDiff**	**pvalue**	**padj**
hg38:chr13:24052743-24053243:250	**0.33151230**	**2.213956e-06**	**0.07681986**
hg38:chr2:211928098-211928598:250	**0.45927274**	**9.281157e-06**	**0.32202829**
hg38:chr5:80764549-80765049:250	**0.19163196**	**1.430491e-05**	**0.49632329**
hg38:chr1:237181661-237182161:250	**0.16584414**	**1.560921e-05**	**0.54156160**
hg38:chr16:17579208-17579708:250	**0.38669505**	**1.672711e-05**	**0.58033048**
hg38:chr1:113243429-113243929:250	**-0.6044875 8**	**1.721167e-05**	**0.59712443**
hg38:chr3:83898738-83899238:250	**0.14859037**	**1.869567e-05**	**0.64859035**
hg38:chr7:126719515-126720015:250	**0.16183673**	**1.985113e-05**	**0.68865568**
hg38:chr8:42005238-42005738:250	**0.01321493**	**7.677331e-01**	**1.00000000**
hg38:chr8:42015658-42016158:250	**0.19821419**	**5.832651e-01**	**1.00000000**

If these plots were significantly skewed to the left, we would have discovered a large number of statistically significant regions. As we have only a slight skew left, we have a statistically underpowered study. However, even under these conditions a number of promising candidate genes have been identified. Using an adjusted significance threshold of 0.10, we can show that three regulator regions were significantly up or down regulated. Finally, given noise inherent in

the dataset a less stringent Bonferroni correction may be considered such that a wider range of peaks may be considered.

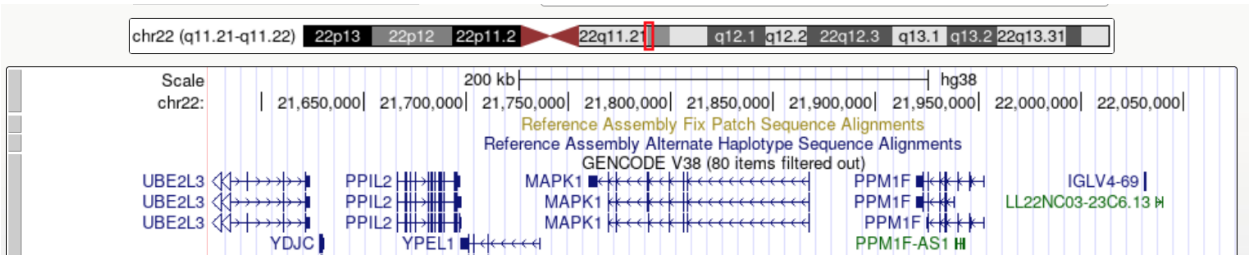
• For each cell type, find the peak that is most positively and most negatively associated with the trait based on un-adjusted p-value. Visualize the predicted open chromatin levels for that region across species in a barplot or scatterplot (from analysis checkpoint)



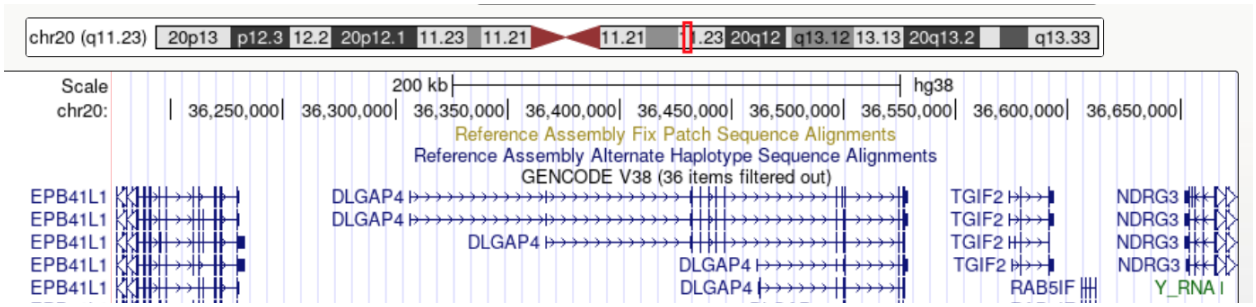
For D1 MSNs, the peak that is most negatively associated with herbivory based on the unadjusted p-value is *Helogale parvula* which is not a herbivore, and the peak that is most positively associated with herbivory is Gorilla gorilla which is a herbivore. For D2 MSNs, the peak that is most negatively associated with herbivory is also *Helogale parvula*, and the peak that is most positively associated is *Rousettus aegyptiacus* which is a herbivore.

• Find the nearest genes to those regions using the UCSC genome browser. Include screenshots to support. Discuss the function of those genes based on your literature search (from analysis checkpoint).

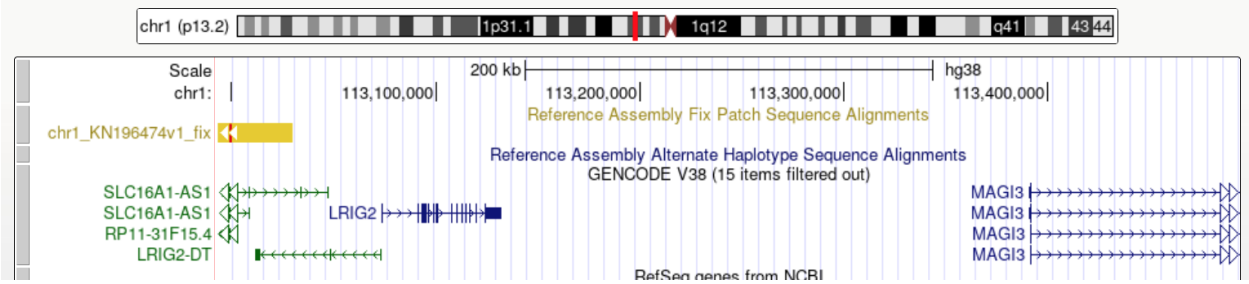
D1 Top Downregulated -- chr22:21825556-21826056



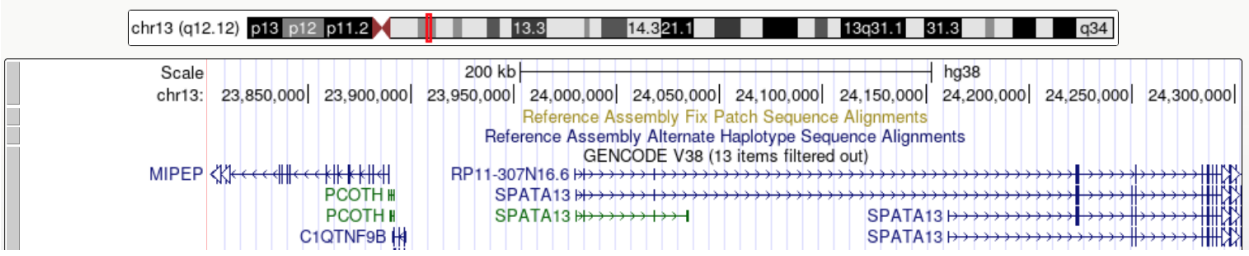
D1 Top Upregulated -- chr20:36425128-36425628



D2 Top Downregulated -- chr1:113243429-113243929



D2 Top Upregulated -- chr13:24052743-24053243



Open chromatin region	Annotation	Closest Gene	Publication
chr13:24052743-24053243	D2 Upregulated	SPATA13	Bourbia et al.

chr1:113243429-113243929	**D2 Downregulated**	**LRIG2**	Stuart et al.
chr20:36425128-36425628	**D1 Upregulated**	**DLGAP4**	Rasmussen et al. Jiang-Xie et al.
chr22:21825556-21826056	**D1 Downregulated**	**MAPK1**	Motta et al.

• Conduct the gene ontology analysis on 200 most differential regions based on the human genome coordinates using GREAT. Discuss the results. (from analysis checkpoint)

In the D2 up genes and in the D1 up genes, we saw that there were no significant categories when the background was the whole genome. On the other hand, in the D2 up genes when the background was just all D2 peaks, we observed one GO cellular component category, NMDA selective glutamate receptor complex. This is related to D2 neurons. In the D1 and D2 down genes with either of the backgrounds, we did not find any categories. In the D1 up genes, we did not find any categories with the background as the whole genome and we found more than twenty categories when the background was all peaks. This was also the case for the D2 up genes. When the background was just all D2 peaks or all D1 peaks, the number of categories increased significantly for the most part. Therefore, there were differences based on the background used. Another category related to the cell type of interest was GRIN2B which is integral in signal transmission in the brain. The protein which is encoded by the LMO3 gene is predominantly expressed in the brain. LMO1 and LMO2 on chromosome 11 are related genes and are apparently involved in T-cell leukemia in terms of chromosomal translocations. NLGN1 encodes a protein which is part of a family of proteins that are found on the neuronal cell surface. These proteins might interact with beta-neurexins via being splice site-specific ligands. Speculation abounds that CNS synapse formation and remodeling might be aided by these proteins. NLGN1 is also related to D1 neurons. As for D1, DSCALM1's protein is part of the Ig superfamily of cell adhesion molecules. These proteins assist in neuronal differentiation. The NTF2 gene's protein is a neurotrophin that allows mammalian neurons to survive and differentiate. It might help maintain the adult nervous system and might affect embryonic neuronal development.

Regulation Type	Cell Type	Background	Significant Categories
Downregulation	D1 MSN	Whole genome	None
		All peaks	None
	D2 MSN	Whole genome	None

		All peaks	None
Upregulation	D1 MSN	Whole genome	None
		All peaks	DSCALM1
			NLGN1
	D2 MSN	Whole genome	None
		All peaks	NMDA selective glutamate receptor complex
			GRIN2B
			NLGN1

MAPK1
MAPK1 is an extracellular signaling molecule. Mutations in the gene can cause Noonan syndrome, characterized by dysmorphic faces, short stature, cardiac defects, and skeletal anomalies. Noonan syndrome is not always accompanied by neurological disorder. Due to the low P-value, we suspect this is just noise.

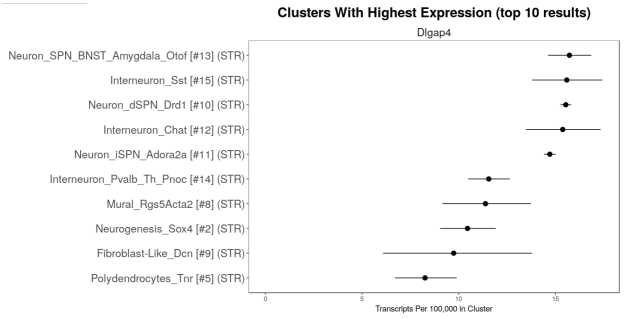
DLGAP4
DLGAP4 is a signaling molecule that can interact with potassium channels. it has been implicated in neural function. This neurotransmitter family allows for neuronal signalling at excitatory synapses. The DLGAP family is associated with with neurological disease including PTSD, ADS, and schizophrenia. In murine models, Dlgap2 knockout caused intense aggressive behavior. It therefore stands that upregulation of the gene could modulate aggressive instincts associated with dietary behavior.

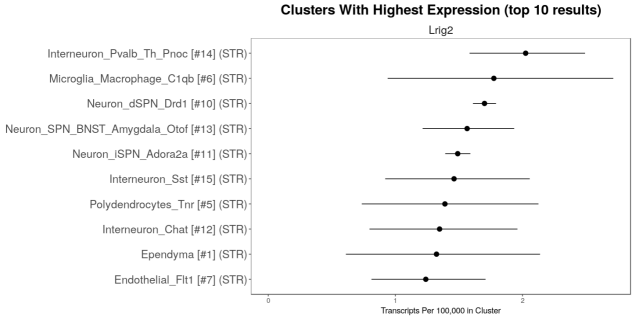
LRIG2
LRIG2 gene is closest and MAGI3 genes are closest. LRIG2 promotes epidermal growth factor signalling. However, LRIG2 has been implicated in poor survival when expressed in glioma cells. Strangely, the gene is also implicated in Urofacial Syndrome, a rare disease where patients grimace when attempting to smile. While it is clear that there is some neurological function, it is not clear what the function is.

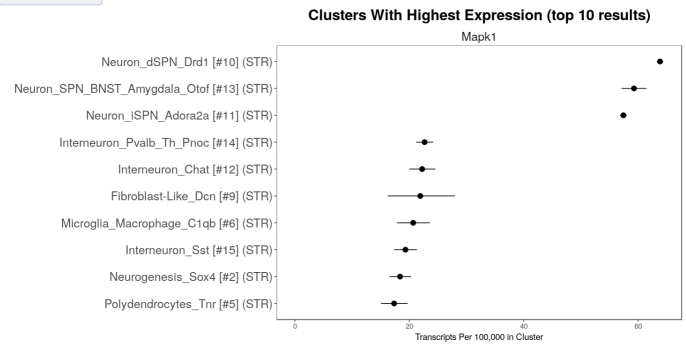
Spata13
The Spata13 gene is closest to this region. Spata13 has been shown in literature to have pronounced expression in the amygdala. Knockout of this gene in mice was shown to have neurological activity, affecting social behavior and nocturnal behavior. In particular, modulation

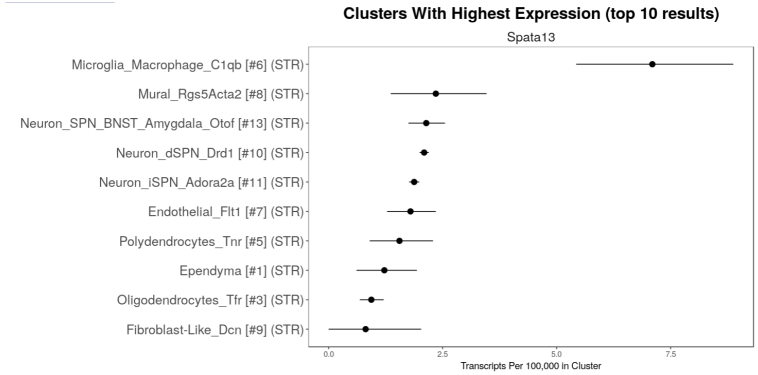
of the gene was shown to cause submissive behavior. It may be hypothesized that passive behavior seen in herbivores is related to modulation of this gene.

• NEW: Choose two genes per cell type that are near differential peaks of interest, either from the UCSC genome browser search or from the GREAT analysis. Identify the cell type-specificity of those genes in the mouse striatum. Visualize that cell type-specificity in some way.









Our group found very little differences in candidate marker gene expression between the two cell types of interest. Our cell selection (D1 and D2 medium spiny neurons, labeled in the plots above as dSPN and iSPN) contains two cell types with high degree of molecular similarity. While we hypothesized that these cell types may be related to the inhibitory and excitatory impulses found in predatory and dietary behavior, the candidate genes identified did not support this hypothesis.

Extended Results

Our group performed GWAS and eQTL analysis in order to identify whether phenotype or molecular level variation would occur with mutation in the candidate regulatory regions we identified. The following phenotypic changes were observed to occur within 100kbp of the regulatory region.

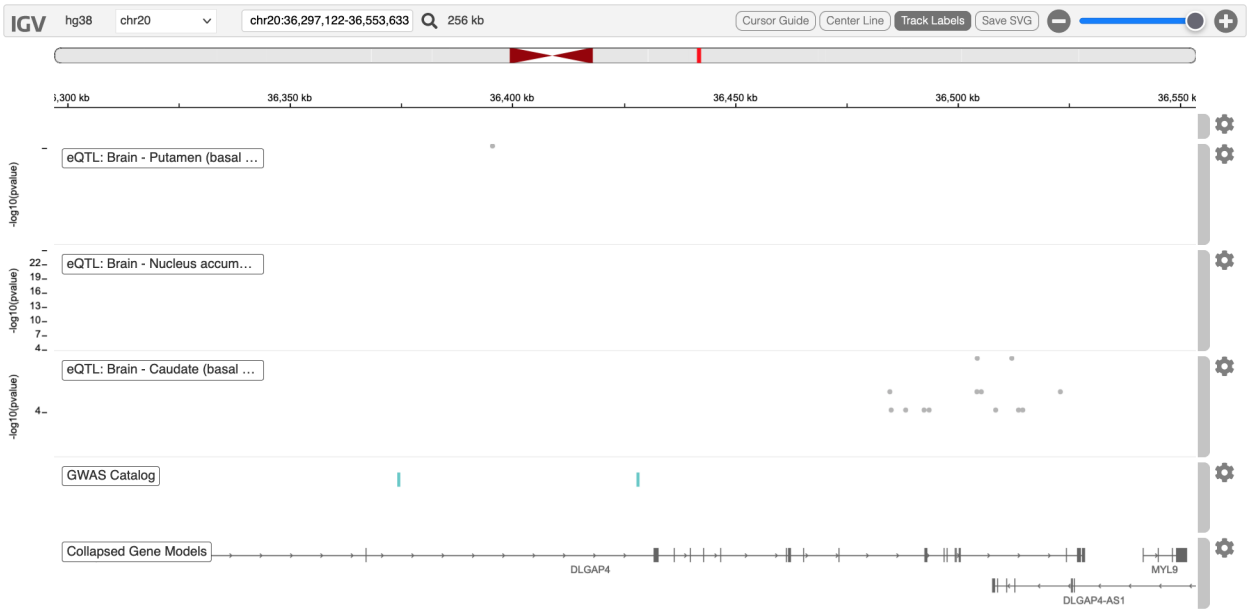
Peak	Nearest genes	Nearby GWAS	Reference	
hg38:chr20:36425128-36425628:250	**DLGAP4**			
	TGIF2			
	AAR2			
hg38:chr7:126719514-126720014:250	**GRM8**	Neuroticism Worry	Cai et al. Hill et al.	
	ZNF800			
	MIR592			
	And others			
hg38:chr13:76672360-76672860:250	KCTD12			
	ACOD1			
	CLN5	Morning person	Hu et al.	
	RP11-19D22.1			
hg38:chr22:21825556-21826056:250	**MAPK1**	Multiple sclerosis Physical activity	International Multiple Sclerosis Genetics Consortium Jung et al.	
	PPM1F			
	PPIL2			
	YPEL1			
	UBE2L3			
	YDJC			
hg38:chr13:24052743-24053243:250	**SPATA13**			
	RP11-307N16.6			
	PCOTH			
	LMO7	smooth pursuit and antisaccade eye	Lencer et al.	

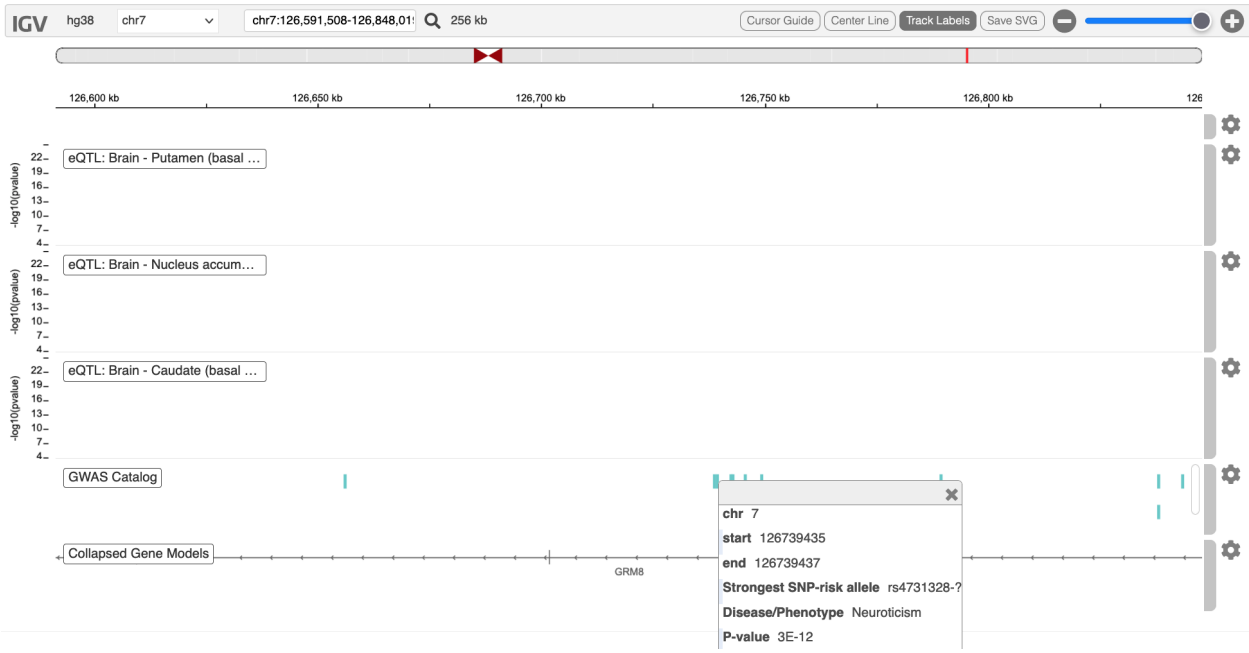
		movements		
hg38:chr1:113243429-113243929:250	LRIG2			
	MAGI3			
	SLC16A			
	PHTF1			
	MOV10			
hg38:chr2:211928098-211928598:250	**ERBB4**	BMI	Justice et al. Locke et al. Akiyama et al.	
	IKZF2			
	CPS1			

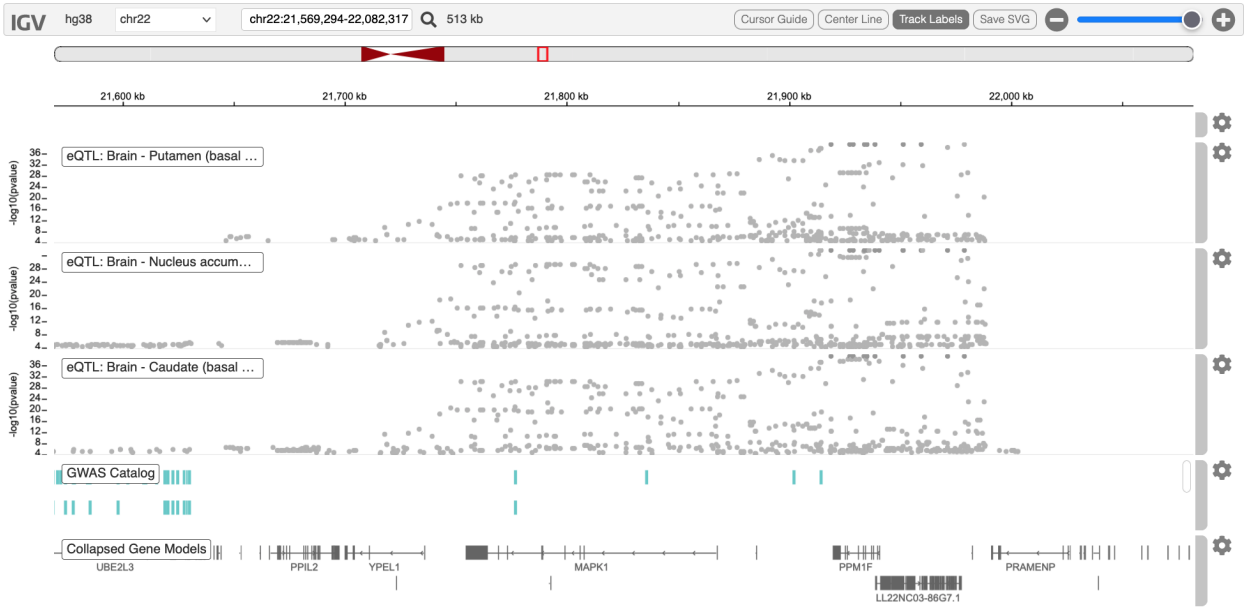
Table 4: GWAS results associated with the traits of interest

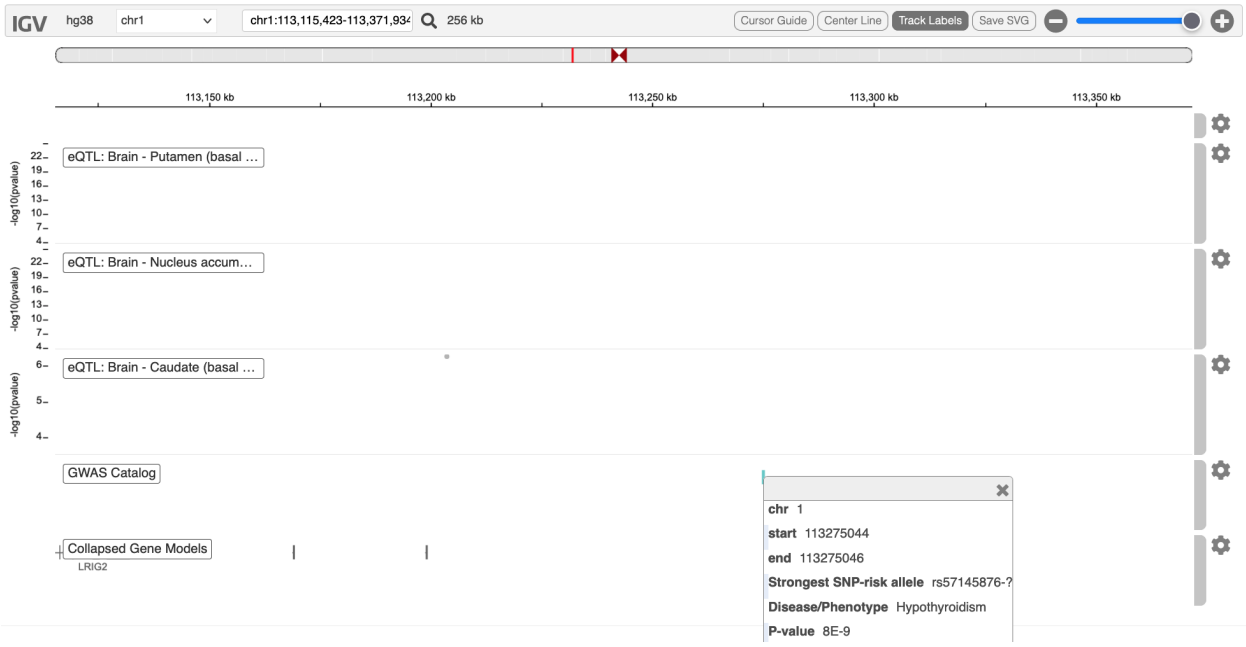
One challenge associated with this analysis is the process of mapping regulator regions to genes of interest. Regulator regions may have long range effects. The closest gene is only associated with a given regulatory region in a minority of cases. This problem is known in the field as “fine mapping,” the processing of determining which variant affects the expression levels of a trait of interest. In order to better determine how variation in open chromatin level can affect gene expression we examined known eQTLs for the regulatory regions of interest.

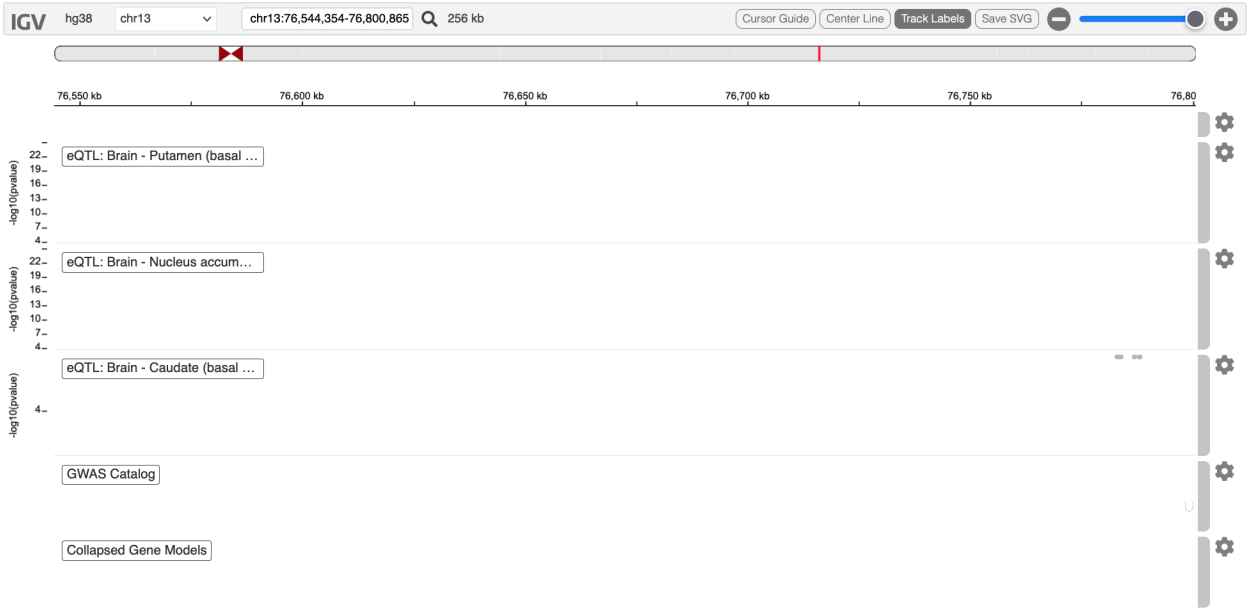
eQTLs or expression quantitative trait loci are genomic loci that influence variation in the expression level of a nearby gene. It was hoped that SNPs in identified open chromatin regions would show association with expression levels of nearby genes. However, very few SNPs were observed near the regulatory elements identified.

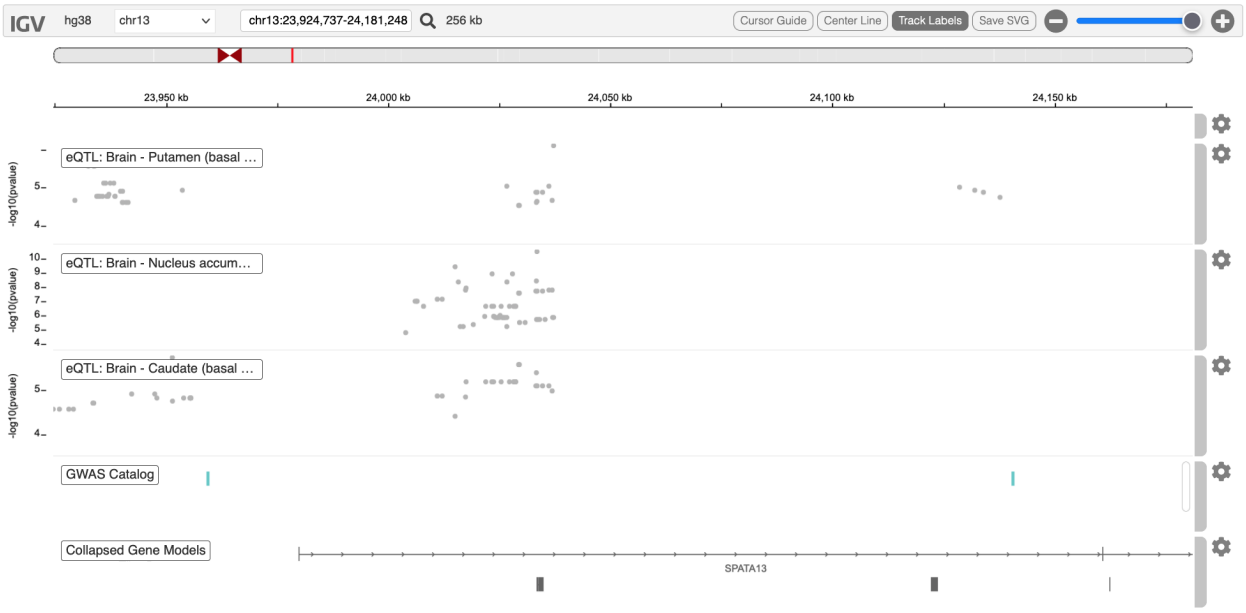

IGV
hg38
chr20
chr20:36,297,122-36,553,633
256 kb
Cursor Guide
Center Line
Track Labels
Save SVG
36,350 kb
36,400 kb
36,450 kb
36,500 kb
eQTL: Brain - Putamen (basal …
eQTL: Brain - Nucleus accum…
eQTL: Brain - Caudate (basal …
-log10(pvalue)
GWAS Catalog
Collapsed Gene Models
DLGAP4
MYL9
DLGAP4-AS1


IGV
hg38
chr7
chr7:126,591,508-126,848,01
256 kb
Cursor Guide
Center Line
Track Labels
Save SVG
126,600 kb
126,650 kb
126,700 kb
126,750 kb
126,800 kb
eQTL: Brain - Putamen (basal …
eQTL: Brain - Nucleus accum…
eQTL: Brain - Caudate (basal …
-log10(pvalue)
GWAS Catalog
Collapsed Gene Models
GRM8
chr 7
start 126739435
end 126739437
Strongest SNP-risk allele rs4731328-?
Disease/Phenotype Neuroticism
P-value 3E-12


IGV
hg38
chr22
chr22:21,569,294-22,082,317
513 kb
Cursor Guide
Center Line
Track Labels
Save SVG
21,600 kb
21,700 kb
21,800 kb
21,900 kb
22,000 kb
eQTL: Brain - Putamen (basal …
eQTL: Brain - Nucleus accum…
eQTL: Brain - Caudate (basal …
-log10(pvalue)
GWAS Catalog
Collapsed Gene Models
UBE2L3
PPIL2
YPEL1
MAPK1
PPM1F
PRAMENP
LL22NC03-86G7.1


IGV
hg38
chr1
chr1:113,115,423-113,371,934
256 kb
Cursor Guide
Center Line
Track Labels
Save SVG
113,150 kb
113,200 kb
113,250 kb
113,300 kb
113,350 kb
eQTL: Brain - Putamen (basal …
eQTL: Brain - Nucleus accum…
eQTL: Brain - Caudate (basal …
-log10(pvalue)
GWAS Catalog
Collapsed Gene Models
LRIG2
chr 1
start 113275044
end 113275046
Strongest SNP-risk allele rs57145876-?
Disease/Phenotype Hypothyroidism
P-value 8E-9


IGV
hg38
chr13
chr13:76,544,354-76,800,865
256 kb
Cursor Guide
Center Line
Track Labels
Save SVG
76,550 kb
76,600 kb
76,650 kb
76,700 kb
76,750 kb
eQTL: Brain - Putamen (basal …
eQTL: Brain - Nucleus accum…
eQTL: Brain - Caudate (basal …
-log10(pvalue)
GWAS Catalog
Collapsed Gene Models



Generation of cell-type specific predicted open chromatin map

Our group further sought to expand upon methods for open chromatin prediction. We first investigated the use of convolution neural networks for the discovery of transcription factor binding motifs. We chose to investigate the performance of support vector machine and convolutional neural network classifiers on a dataset differential open chromatin expression of SST inhibitory neurons against oligodendrocytes (Cusanovich and Hill et al.). While this cell type selection did not match the analysis above, we believe that these two cell types had the largest molecular distance, were the most easily separable cell types in our dataset, and therefore were the strongest selection for validation of the methods used in generation open chromatin predictions.

For our dataset, the performance of the SVM model outclassed the CNN. While these results did not match those reported in the literature (Lawler et al.), it may be reasoned that the smaller dataset contains more noise and is not suitable for training larger models. Indeed, when attempting to optimize our results using an architecture matching that found in (Lawler et al.) our model was unable to converge to a reasonable solution. An SVM model was then selected for extended analysis.

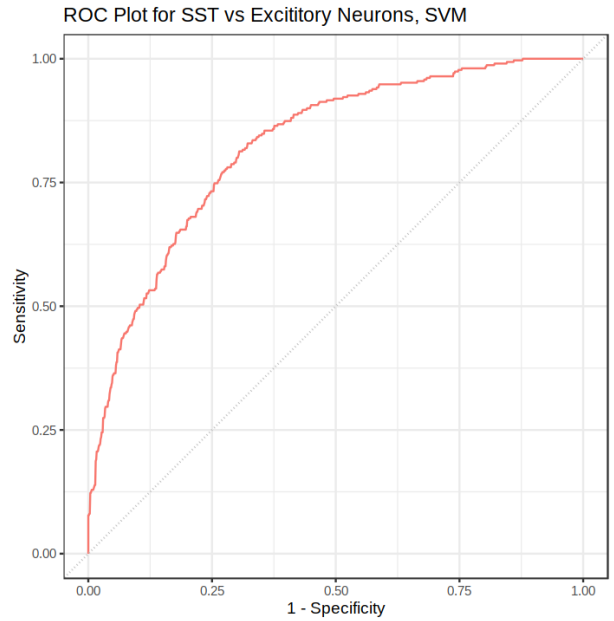


Figure X1: ROC Plot for SST Neurons vs. Oligodendrocytes, SVM. AUC = 0.8111133

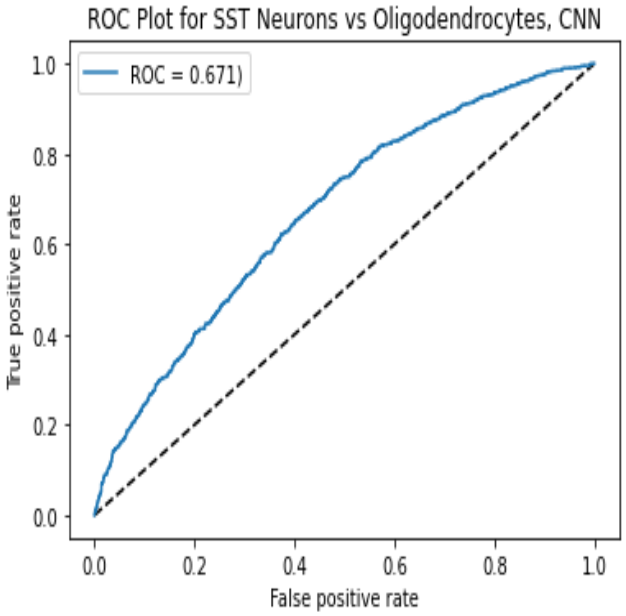


Figure X2: ROC Plot for SST Neurons vs. Oligodendrocytes, CNN. AUC = 0.671

We sought to create a map of cell-type specific chromatin accessibility in inhibitory neurons. In absence of an inhibitory vs. all dataset, we continued analysis using SST inhibitory vs. oligodendrocyte snATAC-seq data. The full human genome of chromosome 21 was binned into 10kbp regions and our SVM open chromatin prediction model was applied to each bin in order to create a map of chromatin accessibility in inhibitory neurons. Results are shown in figure X3

below. A positive score indicates differentially open chromatin in SST inhibitory neurons, where a negative score indicates differentially closed chromatin.

In order to better interpret the results, we included the location of known BMI-associated SNP loci, shown in red. This data was obtained from the NHGRI-EBI Catalog of human genome-wide association studies, trait EFO ID EFO_0004340. Notably, 50 out of 51 BMI-associated mutations were in regions of predicted open chromatin. Given that 94% of chromosome 23 is expected to be open, we would expect to observe this result with a p-value of 0.2967 given by chi-squared test. This is therefore not a statistically significant result.

Associated code is found in Appendix A.

Figure X3: SST inhibitory neuron map of predicted open chromatin regions on human chromosome 21. Regions associated with body mass index expression are shown in red. 50 out of 51 regions associated with BMI expression are in regions of predicted open chromatin, with a p-value of 0.2967

Appendix A

```
#Package management omitted for brevity
library("BSgenome.Hsapiens.UCSC.hg38");
genome <- BSgenome.Hsapiens.UCSC.hg38

lsgkm_predict = '/ocean/projects/ibn200014p/bnphan/src/lsgkm/bin/gkmpredict'
model_fn = file.path('lsgkm', "SSTvsOligo.model.txt")
predict_mystery_fn = 'segment.model.txt'
thecall = paste(lsgkm_predict, "test_file.fasta", model_fn, predict_mystery_fn)

model_fn = file.path('lsgkm', "SSTvsOligo.model.txt")
data <- genome$chr21
bin_size = 10000
stop = length(data) / bin_size;
score <- vector(mode="double", length=stop)
#Parallel processing crashed the kernel..
#score <- foreach(i=0:stop-1, .combine = 'c', .packages='Biostrings') %dopar% {
score2 <- foreach(i=0:stop-1, .combine = 'c', .packages='Biostrings') %do% {
        x = i*bin_size;
        y = (i+1)*bin_size - 1;
        print(i/stop);
        flush.console();
        s_set <- DNAStringSet(c(data[x:y]));

        fasta_file = tempfile(pattern = "file", tmpdir=".");
        output_file = tempfile(pattern = "file", tmpdir=".");
        writeXStringSet(s_set, fasta_file);

        thecall = paste(lsgkm_predict, fasta_file, model_fn, output_file)
        system(thecall)
        out <- read.delim(output_file, header = F)[[2]]
        write(i, file="progress/a")
        unlink(output_file);
        unlink(fasta_file);
        return(out);
}

col <-rep("black", each=length(score2))
loc<-read.csv("loc_gwas.csv", header=FALSE)
x<-unname(unlist(loc))
col[x] = 'red'
png("barplot.png", width=6000, height=350)
barplot(score2, col=col, border=NA)
```

```
dev.off()
```

References

Motta M, Pannone L, Pantaleoni F, et al (2020) Enhanced MAPK1 Function Causes a Neurodevelopmental Disorder within the RASopathy Clinical Spectrum. Am J Hum Genet 107:499–513. https://doi.org/10.1016/j.ajhg.2020.06.018

Bourbia N, Chandler P, Codner G, et al (2019) The guanine nucleotide exchange factor, Spata13, influences social behaviour and nocturnal activity. Mamm Genome 30:54–62. https://doi.org/10.1007/s00335-019-09800-9

Stuart HM, Roberts NA, Burgu B, et al (2013) LRIG2 mutations cause urofacial syndrome. Am J Hum Genet 92:259–264. https://doi.org/10.1016/j.ajhg.2012.12.002

Jiang-Xie LF, Liao HM, Chen CH, et al (2014) Autism-associated gene Dlgap2 mutant mice demonstrate exacerbated aggressive behaviors and orbitofrontal cortex deficits. Mol Autism 5:1–13. https://doi.org/10.1186/2040-2392-5-32

Saunders A, Macosko EZ, Wysoker A, et al (2018) Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell 174:1015-1030.e16. https://doi.org/10.1016/j.cell.2018.07.028

Surmeier DJ, Ding J, Day M, et al (2007) D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. Trends Neurosci 30:228–235. https://doi.org/10.1016/j.tins.2007.03.008

Rasmussen AH, Rasmussen HB, Silahtaroglu A (2017) The DLGAP family: Neuronal expression, function and role in brain disorders. Mol Brain 10:1–13. https://doi.org/10.1186/s13041-017-0324-9

Johnson PM, Kenny PJ (2010) Dopamine D2 receptors in addiction-like reward dysfunction and compulsive eating in obese rats. Nat Neurosci 13:635–641. https://doi.org/10.1038/nn.2519

Rainwater A, Sanz E, Palmiter RD, Quintana A (2017) Striatal GPR88 modulates foraging efficiency. J Neurosci 37:7939–7947. https://doi.org/10.1523/JNEUROSCI.2439-16.2017

Quintana A, Sanz E, Wang W, et al (2012) Lack of GPR88 enhances medium spiny neuron activity and alters motor- and cue-dependent behaviors. Nat Neurosci 15:1547–1555. https://doi.org/10.1038/nn.3239

Guo W, Chen Y, Wang C, et al (2020) The carnivorous digestive system and bamboo diet of giant pandas may shape their low gut bacterial diversity. Conserv Physiol 8:1–12. https://doi.org/10.1093/conphys/coz104

Hu Y, Wu Q, Ma S, Ma T, Shan L, Wang X, Nie Y, Ning Z, Yan L, Xiu Y, Wei F. Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. Proc Natl Acad Sci U S A. 2017 Jan 31;114(5):1081-1086. doi: 10.1073/pnas.1613870114. Epub 2017 Jan 17. PMID: 28096377; PMCID: PMC5293045.

Startek JB, Voets T, Talavera K. To flourish or perish: evolutionary TRiPs into the sensory biology of plant-herbivore interactions. Pflugers Arch. 2019 Feb;471(2):213-236. doi: 10.1007/s00424-018-2205-1. Epub 2018 Sep 18. PMID: 30229297.

Ley RE, Hamady M, Lozupone C, et al (2008) Evolution of mammals and their gut microbes. Science (80-) 320:1647–1651. https://doi.org/10.1126/science.1155725

Muegge BD, Kuczynski J, Knights D, et al (2011) Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science (80-) 332:970–974. https://doi.org/10.1126/science.1198719

Lawler AJ, Ramamurthy E, Brown AR, et al (2021) Machine learning sequence prioritization for cell type-specific enhancer design. Preprint

Hill WD, Weiss A, Liewald DC, et al (2020) Genetic contributions to two special factors of neuroticism are associated with affluence, higher intelligence, better health, and longer life. Mol Psychiatry 25:3034–3052. https://doi.org/10.1038/s41380-019-0387-3

Cai N, Revez JA, Adams MJ, et al (2020) Minimal phenotyping yields genome-wide association signals of low specificity for major depression. Nat Genet 52:437–447. https://doi.org/10.1038/s41588-020-0594-5

Hu Y, Shmygelska A, Tran D, et al (2016) GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. Nat Commun 7:1–9. https://doi.org/10.1038/ncomms10448

Patsopoulos NA, Baranzini SE, Santaniello A, et al (2019) Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. Science (80-) 365:. https://doi.org/10.1126/science.aav7188

Jung SY, Scott PA, Papp JC, et al (2021) Genome-wide association analysis of proinflammatory cytokines and gene-lifestyle interaction for invasive breast cancer risk: The WHI dbGaP study. Cancer Prev Res 14:41–54. https://doi.org/10.1158/1940-6207.CAPR-20-0256

Cusanovich DA, Hill AJ, Aghamirzaie D, et al (2018) A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. Cell 174:1309-1324.e18. https://doi.org/10.1016/j.cell.2018.06.052

Justice AE, Winkler TW, Feitosa MF, et al (2017) Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. Nat Commun 8:1–19. https://doi.org/10.1038/ncomms14977