A framework for systematic promoter motif discovery and expression profiling from high dimensional brain transcriptome data

Jeremy A. Lieberman

Submitted in partial fulfillment of the
requirements for the degree
of Master of Arts
in the Graduate School of Arts and Sciences

Program in Biotechnology
Department of Biological Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

A framework for systematic promoter motif discovery and expression profiling from high dimensional brain transcriptome data

Jeremy A. Lieberman

Understanding the regulatory logic of genes across discrete brain substructures can elucidate the basis for neural network connectivity and the cause of disease. Promoter motifs, in particular, that govern high or low expression gene networks present an important fulcrum for phenotypic behavior. Using the Allen Institute Brain Atlas we took various clustering approaches to find closely regulated genes, and generated substructure specific expression profiles to run through FIRE, a motif discovery algorithm and iPAGE, a functional ontology algorithm. Notably, we found a single large cluster of genes that had tightly coordinated behavior across hundreds of brain substructures, as well as a unique upstream promoter signature, yet highly diverse ontological characteristics. We also present a BRain EXpression Profile ASSembly script (BEXPASS) whose output is customized for FIRE and iPAGE input. Lastly we look at language processing and speech control areas of the brain and put forward recommendations for promoters that can serve as part of DNA constructs for optogenetic research an emerging neuroscientific research method that uses bacterial light-gated ion channel protein, channelrhodopsin (ChR1 or ChR2), as an activity control tool to activate neural pathway signaling.

ACKNOWLEDGEMENTS

I would like to thank the members of Professor Saeed Tavazoie's lab for supporting my research. In particular, Saeed Tavazoie and Panos Oikonomou for their mentorship and guidance. Also to Peter Freddolino for his expertise of the R language.

Table of Contents

Introduction……..…..1

Materials and Methods…..…….5

Data……..…….……...12

Discussion……...…….14

References…...…….27

I. Introduction

Understanding genetic network architecture of the human brain and the cis- regulatory elements that control transcript abundance can elucidate the basis for regional function and substructures involved with disease. The Allen Institute's Brain Atlas project takes a step towards generating the required biological data by taking histological samples from an exhaustive geography of three dimensional coordinates in the developing mouse and human brains as well as the adult mouse and human brains(1),(2),(3). We present BEXPASS, a script that generates continuous distribution absolute and fold induction expression profiles for any selected brain substructure. With these expression profiles, and combined with only a few software packages developed in the Tavazoie lab and other academic labs, many high level conclusions for guiding neuroscientific research can be made.

As part of the Adult Human Brain Atlas Project Hawrylycz et al. generated six data sets from six male and female post-mortem brains ages 18-68. Conditions for exclusion included: brain injuries/cancer, drug/alcohol history, chronic renal failure, and history of infectious diseases. The data sets are freely available online at the Brain Atlas website. Statistical tests were applied to the data to confirm uniformity across brains and establish basic hierarchies. Analysis showed very strong correlation (Pearson coefficient = 0.98) across the six brains sampled. A paired t-test between left and right hemispheres yielded no significance confounding any meta-transcriptomic basis for left and right brain functionality. Hawrylycz et al. clustered genes across all coordinates into similarly expressing modules / clusters using weighted gene co-expression analysis (WGCNA). Those 13 modules

organized into groups of genes that are highly enriched for major cell types (neurons, oligodendrocytes, astrocytes, microglia) with each cell type having ~400 “hub” genes. Organization by cell type corresponded to previous studies about the anatomical distribution of those cell types (4). Hawrylycz et al. concluded that in a transcriptome context, brain histological diversity comes from the combinatorial mosaic of neural cell types being expressed in different quantities across regions (i.e. neocortex is enriched for neural cells).

Figure 1. WGCNA Modules and hub genes of Brain Atlas data. The top band shows 13 color-coded modules of the whole genome (genes in grey not belonging to a module). The second band represents hub genes for major brain cell types that are found enriched in various modules (turquoise, neurons; yellow, oligodendrocytes; purple, astrocytes; white, microglia).

To illustrate the degree of local variance Hawrylycz et al. examined the transcriptional signatures of the hippocampus subregions through analysis of variance (ANOVA) and found that “showed distinctive expression patterns sufficiently robust to cluster together like-samples while distinguishing subdivisions from one another.” This would further confirm the fact that local subregions were transcriptionally, and thus histologically, unique.

The determination of upstream transcription factor binding sites, or lack thereof, and thus the transcription factors that bind to it are a first level in understanding regulatory networks. F.I.R.E. (FIRE) is a regulatory motif discovery algorithm that quantifies the dependency between the presence / absence of a given motif in an DNA 5’ upstream / RNA 3’ UTR promoter region for a given expression profile using mutual information (5):

$$I(\text{motif};\text{expression}) = \sum_{i=1}^{2}\sum_{j=1}^{N_e} P(i,j)\log\frac{P(i,j)}{P(i)P(j)}$$

An information score (I) is generated using this formula for all the possible 8,192 DNA and 16,384 RNA 7mers in upstream and UTR regions, respectively. The sequences with the highest information scores are deemed "seeds" while remaining 7mers are discarded. At the beginning of an optimization procedure all seed motifs have 1 nucleotide added at the beginning and end flank so that all seeds now are 9mers. FIRE then iteratively changes single nucleotides from the original seeds as well as introducing degeneracy as an attempt to improve the information score of the motif with each change. Each information score improvement reflects the biophysical affinity or control that a transcription factor has with a given sequence through information theory differing from other biophysical models that attempt to directly measure Gibbs free energy (6). As FIRE changes the nucleotide at each position or introduces degeneracy the seed information score changes and the process continues: for example, A can be changed to C, G, T, [AC] , [ACT] , [ACG] , [ACT], etc. This continues until all possible changes are exhausted and a maximally informative score is reached (7). Overlapping and redundant motifs are avoided by measuring the information score of the newest iteration of a motif being optimized over the information score of all previous iterations of the seed against a tradeoff parameter. This tradeoff parameter serves as a variable to control the level of redundancy across all the motifs; the higher the tradeoff the more unique each motif.

A two-step randomization process discards 7mers that do not meet a threshold for statistical significance. First, expression profiles are randomly shuffled and the information values calculated between the unchanged motif profile and the shuffled expression profile. This is repeated N_r times with only

motifs being deemed statistically significant ($p < (1/Nr)$) if and only if it is greater than all N_r random information values. The second test involves removing one-third of genes removed and the motif information score being recalculated against the remaining two-thirds. This process if repeated 10,000 times with only motifs that maintain robustness in 6/10 of those jack knife tests being retained.

The combined purpose of FIRE, the Brain Atlas Project data sets, BEXPASS, and other research tools is to elucidate promoter motifs, their control over downstream reading frames, and the networks and signaling cascades they affect. Beyond established high throughput methods for drug-molecule interaction, optogenetics is an emerging field in neuroscience that utilizes the bacterial light-gated ion channel protein, channelrhodopsin, as an activity control tool to activate neural pathway signaling (8). Opsins are a family of proteins conserved across a large number of species comprising of seven-transmembrane domain receptors and a chromophore molecule capable of absorbing light of a certain wavelength. The approximately 350 amino acid long opsin N-terminus protrudes into the extracellular space while the C-terminus into the cytoplasm with the chromophore covalently linked within one of the seven helical domains. Chromophores found in opsins are vitamin-A based retinaldehydes with subtypes varying across species (9). Exposing a bacterial (*Chlamydomonas reinhardtii*) cell expressing channelrhodopsin (ChR1 or ChR2) to light can cause the transmembrane domain of the opsin to open its ion channel and depolarize the plasma membrane by allowing anions to flow in. This changes the electrochemical gradient which in neural in neural cells creates an action potential along the axon which is the basis for neural circuit communication. The evolutionary purpose of this is for *Chlamydomonas*

reinhardtii to use its phototaxic ability to position itself in relation to the sun for photosynthetic growth. Until now two methods for optogenetics have been used in model organisms like algae, drosophila, and mice. The first method involves transgenically inserting a channelrhodopsin gene into cells through a viral vector with using the Cre-Lox recombinant mechanism. Until recently, engineering the requisite lox sequences into loci of choice and achieving sufficient opsin expression was difficult. Currently, numerous mouse strains expressing a variety of transgenic opsin proteins are commercially available (10). The second method for optogenetics uses the endogenously expressed channelrhodopsin proteins in *chlamydomonas reinhardtii* and other bacterial species. The shortcoming in this method is that one is limited to studying bacterial behavior in relation to phototaxis.

II. Materials/Methods

The Allen institute generated six comprehensive brain data sets extracting 393-946 anatomically discrete histological samples via manual macrodissection for larger identifiable brain structures and via laser microdissection for smaller structures. The microarray tissue samples comprised of 50 – 200mg of tissue for macrodissected structures and 0.9mm^3 of tissue for the laser microdissected regions.

The six brain sets were dissected and sampled at different times over a three year time period, 4 brains first and then the last 2 which necessitated a greater number of normalization procedures. For within brain normalization each microarray (batch) data set was fitted to flexible multivariate local regression and an applied correction to accommodate for deviations from the

batch-wise average due to non-biological biases including spatial bias on the array, GC content, and expression differences based on dissection method[i]. To allow cross brain comparison the data sets were normalized by alignment to a control sample, to a single brain mean expression value, and finally to a global mean across all six brains (11).

The microarray that was used for all samples was an Agilent 4x44K Whole Human Genome array with an additional 16 thousand customized probes (12). We imported raw microarray into and pre-processed it using RStudio, a software development suite for the R programming language, which has robust statistics libraries and toolkits. FIRE is implemented in UNIX and runs off a perl command line. Probe metadata came with annotations for Agilent probe IDs and their corresponding Entrez Gene ID. Since multiple probes can be annotated to a single Entrez Gene ID we eliminated redundancy, for the purposes of clustering, by consolidating the 58,692 Agilent probes into their respective 20,787 Entrez Gene IDs (see Figure 2 for process flow diagram).

Entrez Gene ID	Structure ID: 4143 Structure: middle temporal gyrus, Left, inferior bank of gyrus	Structure ID: 4151 Structure: inferior temporal gyrus, Left, bank of mts	Structure ID: 4270 Structure: Long Insular Gyri, Left	Structure ID: 4142 Structure: middle temporal gyrus, Left, superior bank of gyrus
41	7.825312	8.114584	8.066223	8.149979
43	6.722330	6.720349	6.429786	6.881889
47	7.265754	7.535887	7.327293	7.646113

Table 1. Representative X-Y organization of Brain Atlas Data sets. X-axis: microarray tissue samples from a 3-Dimensional coordinate. Y-axis: genes.

In this consolidation we also eliminated roughly 40% of the expression values across all brain coordinates for not beating a two background tests. These two tests included a t-test to ensure the probe's mean expression is significantly different from background and then a background subtraction signal test to establish significant difference between signal and background.

The second step involved conversion (and expansion) of the 20,787 Entrez Gene IDs to 32,665 RefSeq IDs since FIRE only accepts RefSeq IDs. This step recreated redundancy that we aimed to eliminate in Step 1; however, FIRE has a step that eliminates redundant transcripts. It was preferred to let FIRE handle the redundancy with a one-to-many conversion rather than do one-to-one conversion and have to pick a single RefSeq transcript ID that might bias FIRE's results. Conversion tables for Step 2 conversion were obtained from the Ensembl Genome Broswer (13).

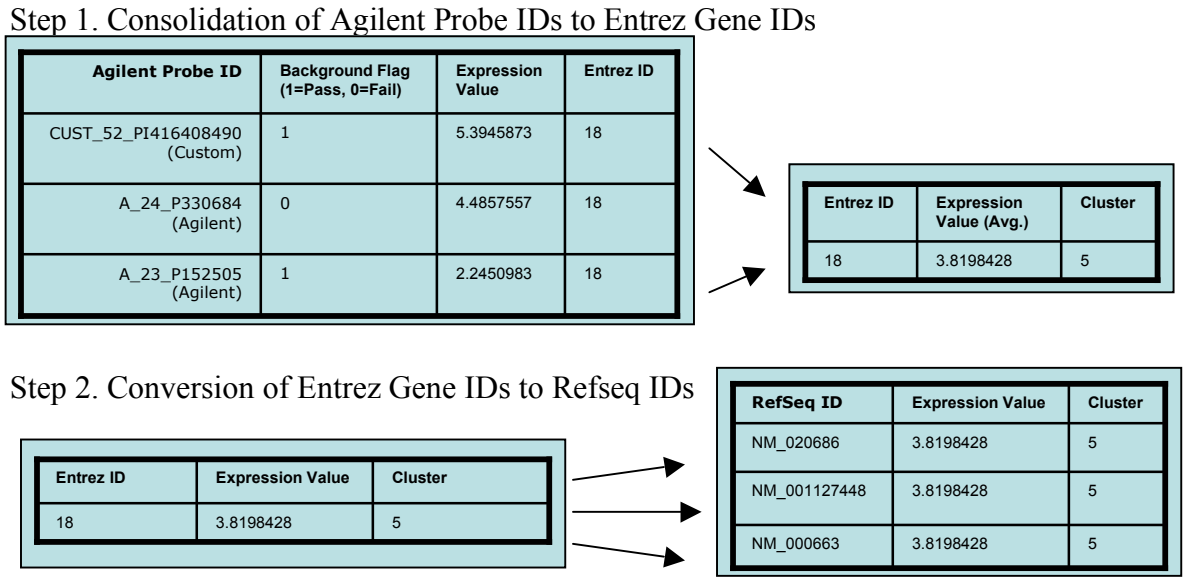


Figure 2. Probe Conversion Process Flow: Agilent to Entrez to RefSeq

For initial analysis expression values were clustered across within each brain across all coordinates using the k-means algorithm. The Hartigan and Wong method of k-means minimizes the Euclidean sum of squares distances between for each high dimensional (393-946 coordinates) gene observation.

Two methods were used to determine the number of clusters. The first is a rule of thumb in clustering, the square root of N observations divided by two: $\sqrt{(n/2)}$; the second is the elbow method which says that the optimal

number of clusters is the point at which the percentage of explained variance becomes marginal with the addition of additional cluster; explained variance is the between cluster sum of squares divided by the total sum of squares (See Table 2 for explained variances). We also clustered the data several times around the optimal number cluster number. The √(n/2) method's optimal cluster number was 102 and so in addition we clustered at multiples of 15 (± 15x): 57, 72, 87, 117, 132, 147. The elbow method's optimal cluster number was 4-6 so we clustered genes with 2-8 clusters for exhaustive variation. Following initial results from running these clusters through FIRE we subsequently clustered with 20 and 30 clusters to have coverage over the gap between 8 and 57. Each of the 98 k-means cluster combinations were ran through FIRE algorithm using default parameters and discrete distribution with each cluster serving as a bin.

	Variance Explained Based on Number of Clusters						
Brain ID	2 clusters	3 clusters	4 clusters	5 clusters	6 clusters	7 clusters	8 clusters
10021	65.3%	79.0%	84.4%	86.9%	88.4%	89.3%	89.8%
12876	65.7%	79.3%	84.2%	86.6%	87.9%	88.8%	89.3%
14380	67.8%	80.9%	86.1%	88.5%	89.9%	90.7%	91.2%
15496	66.7%	80.2%	85.6%	88.1%	89.6%	90.5%	91.0%
15697	65.8%	79.9%	85.4%	88.1%	89.6%	90.5%	91.1%
9861	64.2%	78.1%	83.3%	85.7%	87.1%	87.9%	88.5%

Table 2. Variance Explained Based on Number of Clusters. Each of the six brains cluster extremely similarly and have marginal variance explanation after 5 clusters

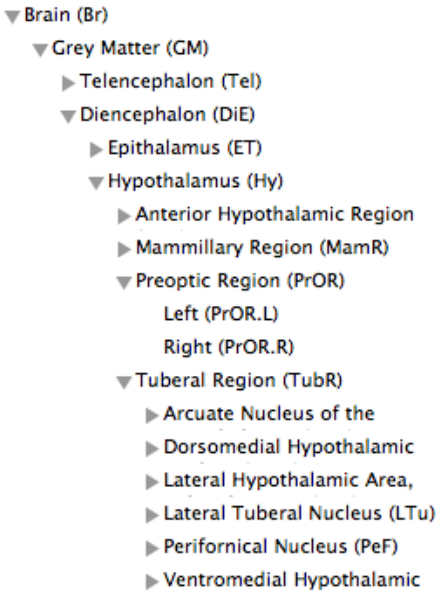


Figure 3. Ontological Representation

Each of the coordinates in the six data sets has an ID number placing it in a node on an ontological tree. This hierarchy was used to isolate coordinates of the brain to analyze which genes are highly expressed and repressed in those areas. Figure 3 shows a small example of the ontological organization to the hypothalamus used by Hawrylycz et al.

For each gene coordinate expression data point ($X_{i,j}$), a Z-score was calculated based on the mean and standard deviation for that unique gene across all coordinates within one brain. Initially, we looked at all substructures of the hypothalamus, amygdala, and basal ganglia. However, because of histological heterogeneity within these three rather broad brain regions of interest, we proceeded to find a "higher resolution" set of insertion candidates for substructures within them.

In our analysis we used three tools to find gene ontology enrichments within gene clusters. The first is FIRE's native gene ontology function that uses the hypergeometric distribution. The second gene ontology tool used is GOstat, which uses Fisher's Exact test (one-tailed hypergeometric) compared against a chi-squared test ($\chi 2$) and is Benjamini FDR corrected at .1 (14). The third tool used is iPAGE which, like FIRE, uses the concept of mutual information to quantify how informative an annotated pathway is to a bin of genes in a given expression profile (15).

BEXPASS is a self-contained R language script that uses the most comprehensive data set from the Brain Atlas Project and generates two whole genome

expression profiles based on a user-selected brain coordinate of choice. One of the two expression profiles is continuous ranking of absolute expression levels of each $gene_{(i)}$ in the selected brain $subregion_{(j)}$ from highest (X_{1j}) down to lowest(X_{20787j}). The second method calculates a ratio of the given gene-coordinate expression level over the gene mean across all coordinates (X_{ij}/mean(X_i)). These expression profiles are written in a format that can be immediately run through FIRE.

To assess the robustness of BEXPASS, the Allen Institute Brain Atlas data sets, and their combined functional linkage to FIRE we ran a three fold cross validation for the absolute and ratio / fold induction expression profiles for brain ID#9861. For gene indices 1-20,787 we random generated three sets of 6,929 index values ($1/3^{rd}$) without replacement. From those random values, six test sets (three absolute expression, three ratio / fold induction) were created of length 6,929. Six training sets (three absolute expression, three ratio / fold induction) were created from the remaining genes for length 13,858. The six training sets were then run through FIRE with default stringency parameters, 20 bins, and continuous distribution. The significant motifs that emerged from these six FIRE runs were then recycled and rerun through FIRE in non-discovery mode against their respective six test sets. Non-discovery mode allows a pre-selected group of motifs to be evaluated for enrichment against an expression profile and allowed us to test whether FIRE would replicate results on a smaller but highly similar expression profile. Figure 4 shows a comparison of motif signatures between training and test sets. While the enrichments and under-representations are not as significant (or deep) in the test set as they are in the training set, the general color patterns remain the same signifying FIRE's ability to reproduce results from a smaller, similar expression profile.

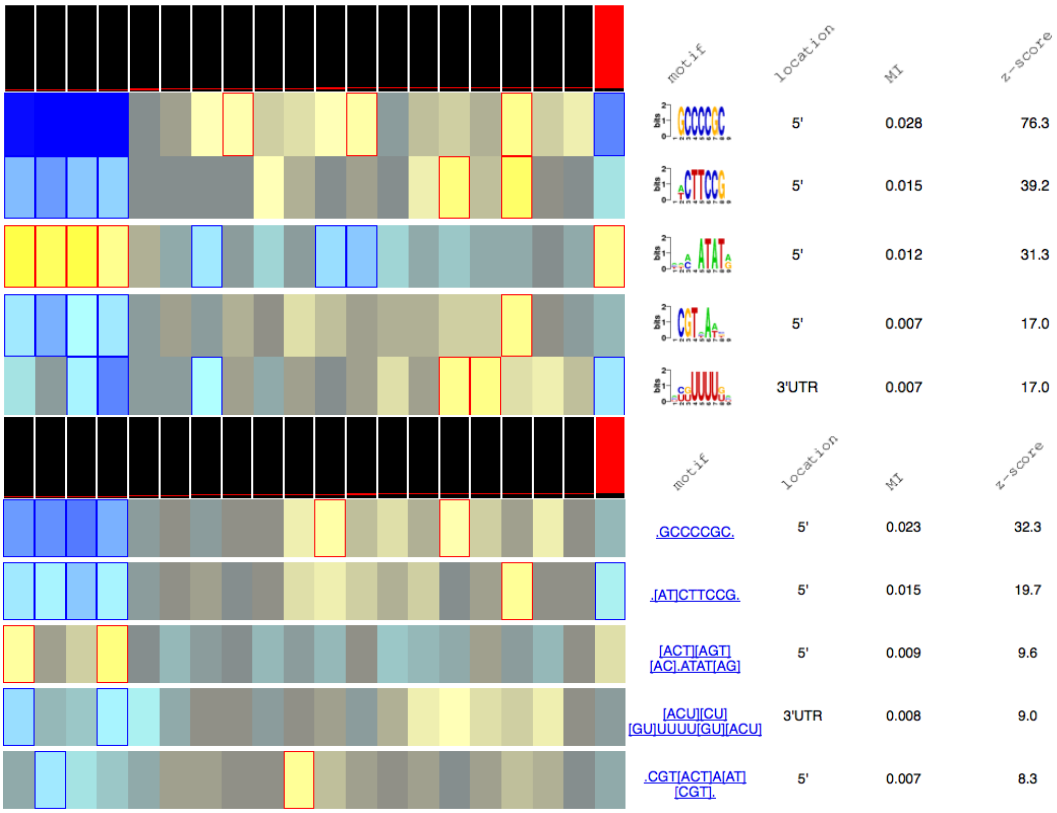


Figure 4. FIRE results for ratio / fold induction training and test set #1. A) Upper figure: training set results. B) Lower figure: test set results. Motifs 4 and 5 are in reverse order.

III. Data

The Hawrylycz et al. data sets are six tables of 58,692 probes (x-axis) and 363-946 brain coordinates (y-axis). 84% of microarray transcripts (29,412) are expressed in at least one structure[i].

Brain ID	No. of coordinates (Microarray samples)	Total Expression Readings (x58,692)	Range of Expr. (Min – Max)	Mean	Standard Deviation
10021	893	52,411,956	0 - 18.58565	4.3778	0.1505718
12876	363	21,305,196	0 - 18.52619	4.6015	0.1490856
14380	529	31,048,068	0 - 18.13379	4.6902	0.1067083
15496	470	27,585,240	0 - 18.24433	4.8131	0.1368528
15697	501	29,404,692	0 - 18.31623	4.8828	0.1231501
9861	946	55,522,632	0 - 18.38175	4.2453	0.1627419

Table 3. Summary of sample coordinate locations.

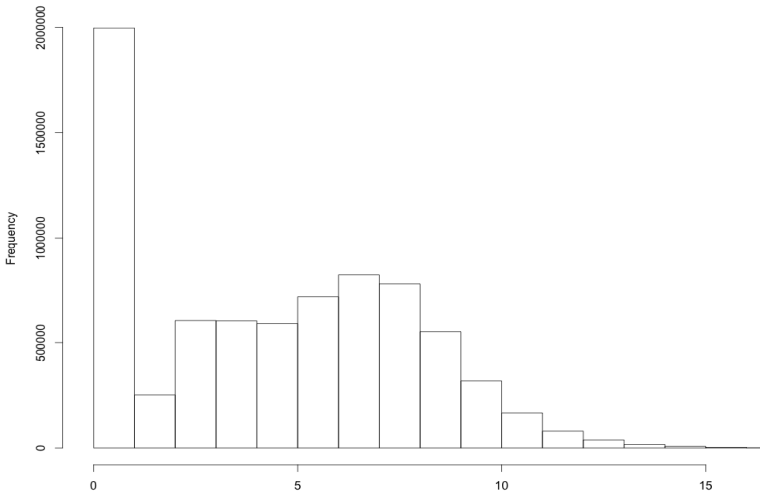


Figure 5. Distribution of expression readings in Brain ID#9861. The value of zeros is skewed due to the quantity of probe readings that do not pass background

For each brain substructure, the last node from the root in the ontological tree, anywhere from 1-11 coordinate samples were taken depending on the brain.171 brain substructures were represented with at least two samples in at least two brains. Tables 3-4 and Figure 5 show meta-statistics for the data sets. The histogram of microarray readings from brain ID#9861 is skewed towards zero (as it is in the other five brains) due to the high number of probes that did not beat background.

		Brain region	Brain 1 # of Samples Hemisphere: R(L)	Brain 2 # of Samples Hemisphere: R(L)	Sample structures	Isolation Method
Telencephalon	Cortex	Frontal Cortex	130 (63)	119 (61)	Orbital gyri; superior, middle, and inferior frontal gyri; rostral and subcallosal gyri; precentral gyrus; paracentral lobule	Macro
		Parietal Cortex	67 (32)	54 (26)	Superior and inferior parietal lobules; postcentral gyrus; paracentral lobule	Macro
		Temporal Cortex	125 (61)	74 (37)	Superior, middle, and inferior temporal gyri; fusiform gyrus; transverse gyri	Macro
		Occipital Cortex	28 (15)	43 (22)	Striate and extra-striate cortex from the cuneus and lingual gyrus; occipito-temporal gyrus; lateral occipital gyri	Macro
		Insula	10 (4)	7 (3)	Short and long insular gyri	Macro
		Cingulate Gyrus	21 (10)	27 (11)	Anterior, posterior, and retrosplenial regions of the cingulate cortex	Macro
		Parahippocampal Gyrus	13 (7)	8 (4)	Parahippocampal gyrus	Macro
		Hippocampus	60 (32)	54 (27)	CA1-CA4 pyramidal cell layers; dentate gyrus; subiculum	LMD
	Cerebral Nuclei	Striatum	34 (17)	44 (22)	Caudate nucleus; putamen; nucleus accumbens	Macro/ LMD
		Globus Pallidus	8 (4)	13 (6)	Globus pallidus	Macro/ LMD
		Basal Forebrain	7 (4)	10 (5)	Septal nuclei; cholinergic basal forebrain; bed nucleus of the stria terminalis	LMD
		Claustrum	17 (8)	11 (6)	Claustrum	LMD
		Amygdala	12 (12)	22 (9)	Lateral, basolateral, basomedial, central, and cortico-medial amygdalar nucle	LMD
Diencephalon		Dorsal Thalamus	46 (23)	39 (17)	Anterior, medial, lateral, posterior, and intralaminar nuclei of the thalamus	Macro/ LMD
		Ventral Thalamus	7 (3)	10 (5)	Reticular nucleus and zona incerta	LMD
		Subthalamus	3 (1)	3 (2)	Subthalamic nucleus	LMD
		Epithalamus	8 (3)	2 (1)	Habenular nuclei; paraventricular nucleus of the thalamus	LMD
		Hypothalamus	9 (5)	22 (11)	Anterior, lateral, posterior, and preoptic hypthalamic areas; paraventricular, supraoptic, ventromedial hypothalamic nuclei; mammillary bodies	LMD
		Mesencephelon	44 (27)	62 (34)	Cranial nerve nuclei 3 and 4; substantia nigra; red nucleus; ventral tegmental area; pretectal regions; midbrain raphe nuclei, superior and inferior colliculi	LMD
Metencephalon		Cerebellar Cortex	32 (21)	27 (18)	Cortex from the lateral hemispheres, paravermis, and vermis	Macro
		Cerebellar Nuclei	12 (5)	7 (5)	Deep cerebellar nuclei	LMD
		Basal Pons	12 (5)	12 (6)	Pontine grey	LMD
		Pontine Tegmentum	45 (22)	38 (22)	Cranial nerve nuclei 5-7; pontine reticular formation and raphe pontis; locus coeruleus; superior olivary complex	LMD
		Myelencephelon	78 (39)	85 (46)	Cranial nerve nuclei 8-12; spinal portion of the trigeminal nucleus; raphe nuclei and reticular formation of the medula; arcuate nucleus; inferior olivary complex; cuneate nucleus; gracile nucleus	LMD
		White Matter	2 (1)	1 (1)	Corpus callosum and cingulum bundle	Macro

Table 4. Summary of sample coordinate locations.

Each of the six brain data sets came with annotation files about the coordinates (seen in Table 5). Along with general information about the subregion of each microarray coordinate were Montreal Neurological Institute (MNI) coordinates, a spatial-imaging framework for cross subject brain comparison. MNI coordinates are a newer development of the Talairach Brain Atlas, which has long been the basic framework for brain spatial definition (16).

Structure ID	Slab Number	Structure Acronym	Structure Name	MNI-X	MNI-Y	MNI-Z
4077	22	PCLa-i	paracentral lobule, anterior part, Right, inferior bank of gyrus	5.9	-27.7	49.7
4323	11	Cl	Claustrum, Right	29.2	17.0	-2.9
4323	18	Cl	Claustrum, Right	28.2	-22.8	16.8
4440	18	LGd	Dorsal Lateral Geniculate Nucleus, Left	-24.6	-24.6	1.3
4266	17	CA4	CA4 field, Right	31.1	-31.3	-7.3

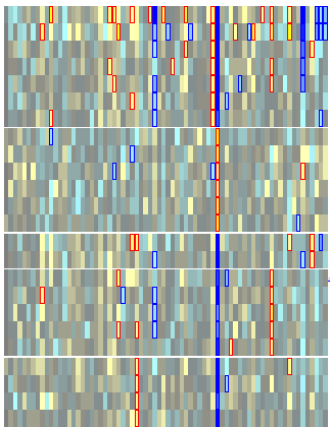
Table 5. Abridged coordinate annotation data

IV. Discussion

The growth of available biological data over the last decade has shed light on the high amount of dynamic pathways in living organisms (17). Because of the complexity of human genetic regulatory network architecture, genes can co-express and co-cluster in endless unique permutations across multiple conditions. Given these precedents it would seem counter intuitive that 4-6 clusters (based on variance explained) was determined to be the optimal cluster amount for a whole genome high coordinate expression profile. However, stratifying unique clusters from Euclidean sum of squared differential distance in >393 dimensions with a range of 0-18 simply can't truly delineate unique and dynamically integrated pathways.

The one exception to this is that across all our FIRE runs there was a group of genes, significantly composed of the olfactory receptor family, that consistently

clustered together and had a visually unique motif signature (see Figure 6A for a representative example). FIRE's native gene ontology function uses the hypergeometric distribution to identify significant gene ontologies and "olfactory receptor activity" was significant enriched in this cluster in FIRE runs going as low as three clusters. Further GO analysis was only done on the 57-147 cluster runs because this "super cluster" would be more stratified and have less noise. In this "super cluster", across 42 FIRE runs ranging from 57-147 clusters, there were 527 unique genes (356 with GO annotations, 171 without) that clustered together 42 out of 42 times. Running those 527 unique genes through GOstat with the Current Composition of Human Gene Ontology Annotation Table as the reference database yields 75 over represented gene ontologies among unique sub-groups of 20-100 genes (pval < .01, Benjamini FDR corrected at .1). Only 47 of the 527 genes in the cluster actually have annotations for olfactory receptor activity. Figures 6B and 6C show that this similarly behaving group of genes is actually quite diverse with enrichments for rhodopsin-like receptor activity and cytokine receptor activity among others. To confound the regulation of this cluster even further is that its motif regulators as found in FIRE do not explain its regulatory behavior. We only considered those motifs that have a Z-score greater than 20 are to have serious explanatory value. Across FIRE runs for the super cluster only redundant variations of the AAAATAT motif had Z-scores greater than 20. A handful of other motifs that did not appear consistently across FIRE runs had Z-scores around 10. Notably, it always showed significant under-representation of the CCGCCCC motif which is a common binding site for multiple transcriptions factors and consistently had Z-scores greater than 60.



Motif	Seed	Z-score range	Motif Name
[ACG]A.[AC]ATAT	AAAATAT	20-30	-
[AGT][AG].AATAT[ACT]			Arid5a_1
[AG][ACT]AA[AGT]TAT			Croc, Dlx3, Dlx5
AAAA[AGT]TT			SUM1, STB3, SFP1
[AGT][AG].AATAT[ACT]			Arid5a_1
[ACG]A[AG]AAT[AT]T			SFP1, SUM1
AG[ACT]CA[GT]A[AG]	ACAGAG	10-12	-
[AGT][AG]GA[GT]AGA[AGT]			GATA3
[ACT][CG]ACAG[AT]G[AGT]			DCE_S_II
[AGT]CACTC[AC]A[ACT]			-
[ACG]ACAGAG[AGT][AGT]			-
AG[ACT]CA[GT]A[AG]			-
[ACT]T[AC]TCC[ACT]	TCTCTCC	9-10	Sig1
[ACT]TCTCT[AC][CT][ACT]			-
[ACT]T[AC]T[AC]TCC[ACT]			Sig1

Biological Process	P-val.	Molecular Function	P-val.	Cellular Component	P-val.
Sensory perception of chemical stimulus	<1.00e-80	Olfactory receptor activity	3.22e-79	Extracellular space	6.33e-25
Sensory perception of smell	<1.00e-80	Rhodopsin-like receptor activity	2.72e-52	Intrinsic to membrane	6.27e-20
Sensory perception	5.67e-72	G-Protein coupled receptor activity	1.36e-42	Integral to membrane	1.13e-19
Neurological system process	8.28e-52	Transmembrane signaling receptor activity	4.19e-41	Membrane Part	7.27e-16
Multicellular organismal process	6.42e-48	Receptor activity	1.88e-31	Extracellular region part	1.28e-14
System Process	2.34e-43	Signal transducer activity	1.07e-24	Membrane	1.32e-11
Defense response	8.26e-20	Molecular transducer activity	1.07e-24	Intrinsic to plasma membrane	7.87e-05
Plasma membrane	4.18e-34	Cytokine receptor activity (8)	8.03e-06	Intermediate filament	0.000306
G-protein coupled receptor signaling pathway	1.48e-36	Cytokine activity	0.00092	Intermediate filament cytoskeleton	0.000306
Cell surface receptor signaliing pathway	6.02e-30	Pancreatic ribonuclease activity	0.00258	Integral to plasma membrane	0.00068

Figures 6. A) FIRE output: representative example of olfactory cluster that is highly enriched for set of promoters and highly absent for others. The super cluster is represented in the right third of the matrix with significant under-representation for the first 7 and last 11 promoters and significant over representation for promoters 8-13. **B) Most enriched promoters.** Set of most enriched promoters for the "super cluster" across all FIRE runs. **C) Most significant Gene Ontology results for the super cluster across FIRE runs.**

WCGNA modules attained by Hawrylycz et al. mirrored anatomical distributions of neural cell types. Since those modules were significantly enriched for hub genes of each neural cell type they therefore represented an ideal expression profile for FIRE and iPAGE in order to ascertain motifs that control and characterize cell type differentiation and identify enriched pathways. Figure 8 shows strong FIRE results with numerous motifs having Z-scores greater than 20 and with most clusters

showing distinct over and under representations for specific pathways. Module 2, of which neuron hub genes make up one tenth, yielded several of the strongest transcription factor binding sites. Oligodendrocyte hub genes, which represent roughly a quarter of module 12 genes, yielded only one over-represented motif and no under represented motifs. Module 10 which is almost entirely made up of astrocyte hub genes yielded only under-representation for 1 transcription factor. Lastly, modules 8 and 9 which are entirely and half, respectively, made up of microglia hub genes yielded 2 over-represented and 5 under-represented motifs.

The "hub" genes referenced in Hawrylycz et al were originally annotated as cell specific marker genes by Oldham et al (18). Enough homology exists between *homo sapiens* and *mus musculus* that Oldham used a transcriptome database of purified mouse astrocyte, neuron, and oligodendrocyte cell colonies to identify the marker genes for each cell type. The purified cell lines came from the postnatal mouse brain at various postnatal ages from 1 day old to 30 days old in Cahoy et al (19). Cells were sorted using fluorescent-activated cell sorting (FACS) and transcriptomes were measured by Affymetrix GeneChip Arrays. While astrocytes and oligodendrocytes share functionality under the nomenclature of "glial" cells Cahoy's analysis showed that their transcriptomes are as differentially expressed from one another as they are from neurons. We took the list of genes (between 2000 and 3000 per cell type) enriched by greater than 1.5-fold and statistically different by significance analysis of microarrays (SAM) with a false discovery rate (FDR) threshold of 1% and ran them through FIRE and iPAGE with continuous distribution (fold enrichment) with default parameters. This yielded only 1-3 weak motifs per cell type. There was no overlap between positive results for FIRE and iPAGE as none of the bins that showed enriched significance for a motif displayed significance for an ontological category as well. Lowering stringency to a minR of 2 and jack knife tests to 4 yielded only more weak motifs. Six weak motifs emerged for oligodendrocytes under less stringent parameters

with four of them being located on 3' UTR; one of those motifs was a binding site for at least a dozen microRNAs. Notably, the most enriched bin of genes in neurons was significantly enriched for the biological process "chloride transport" and the KEGG pathway "Neuroactive ligand receptor interaction".

Genes greater than 20 fold enriched in the three major CNS cell types (117 genes in astrocytes, 175 genes in neurons, 83 genes in oligodendrocytes), were deemed "cell specific" genes Cahoy et al. These >20 fold enriched gene sets were rerun through FIRE and iPAGE as a single gene cluster. Results were poor with only astrocytes showing two strong motifs: the first is a 3' UTR motif that binds cyclic AMP response element CRE–BP1 and Hepatic Leukemia Factor. The second motif is undiscovered with a GAAACGC seed. iPAGE confirms the Cahoy et al.'s conclusion that oligodendrocytes and astrocytes have significantly distinct transcriptomes as >20 fold enriched cluster against background (whole RefSeq genome) showed enrichment for different GO categories (see Figure 7).

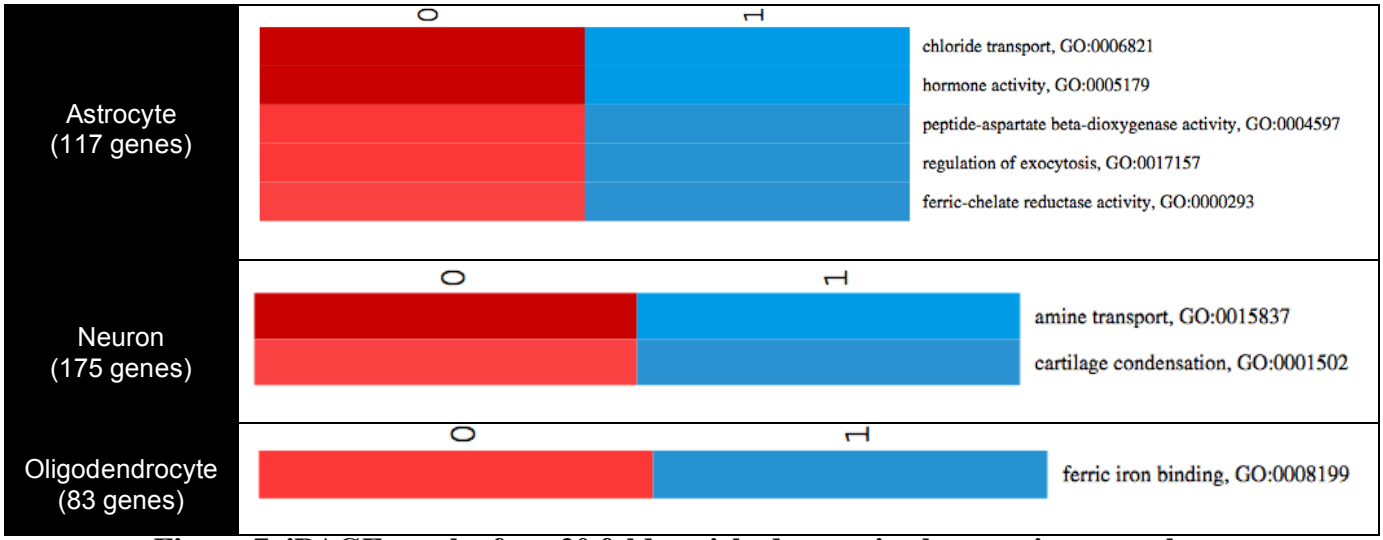


Figure 7. iPAGE results for >20 fold enriched genes in three major central nervous system cell types. Cluster 0 is >20 fold enriched genes and cluster 1 is background.

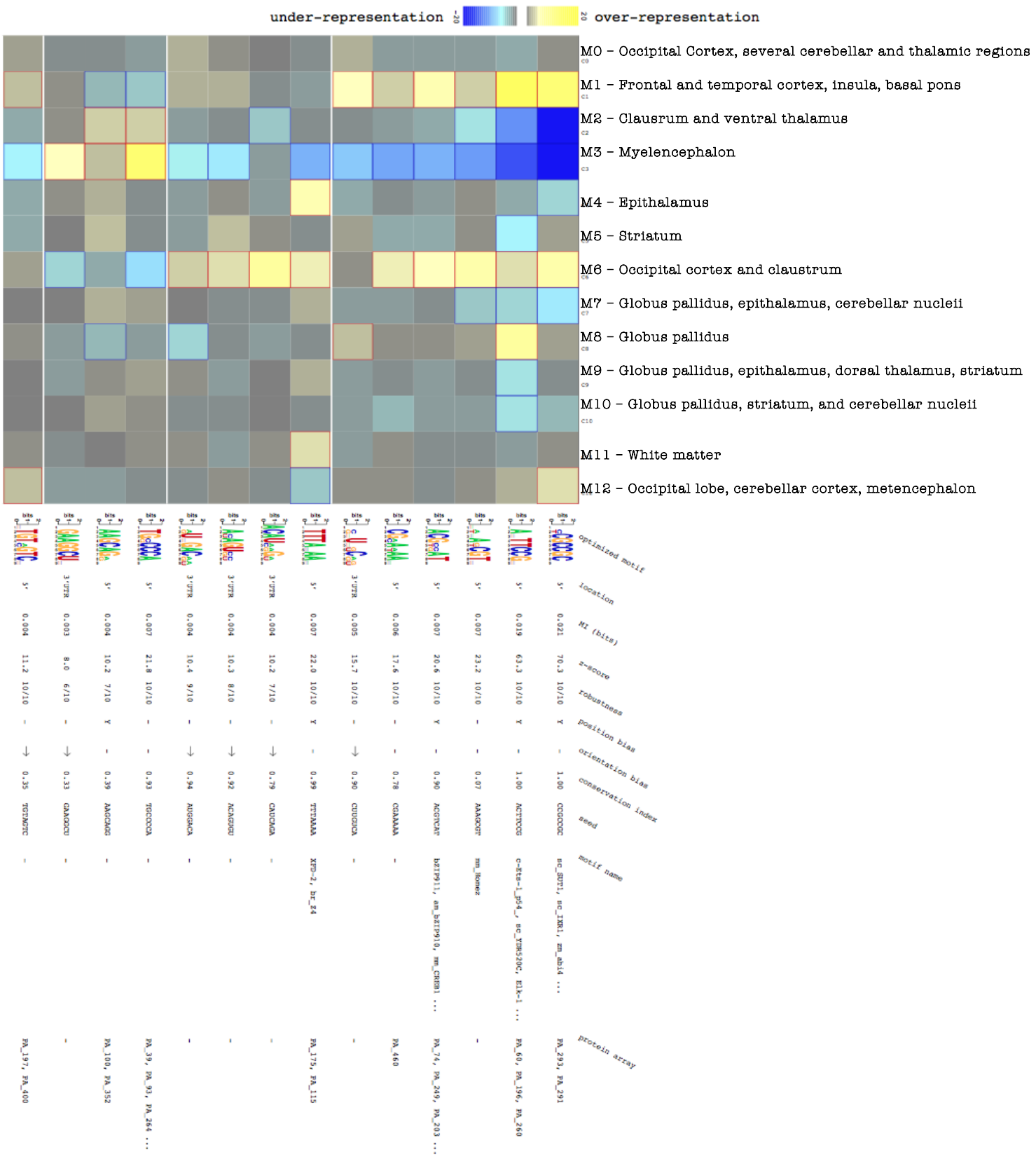


Figure 8. WCGNA Module FIRE results. Looking at this figure laterally, the X-axis represents modules from WCGNA clustering. The brain regions in which those clusters are highly expressed are listed across the top of the grid. The Y-axis represents how informative the presence/absence of each motif is within each cluster.

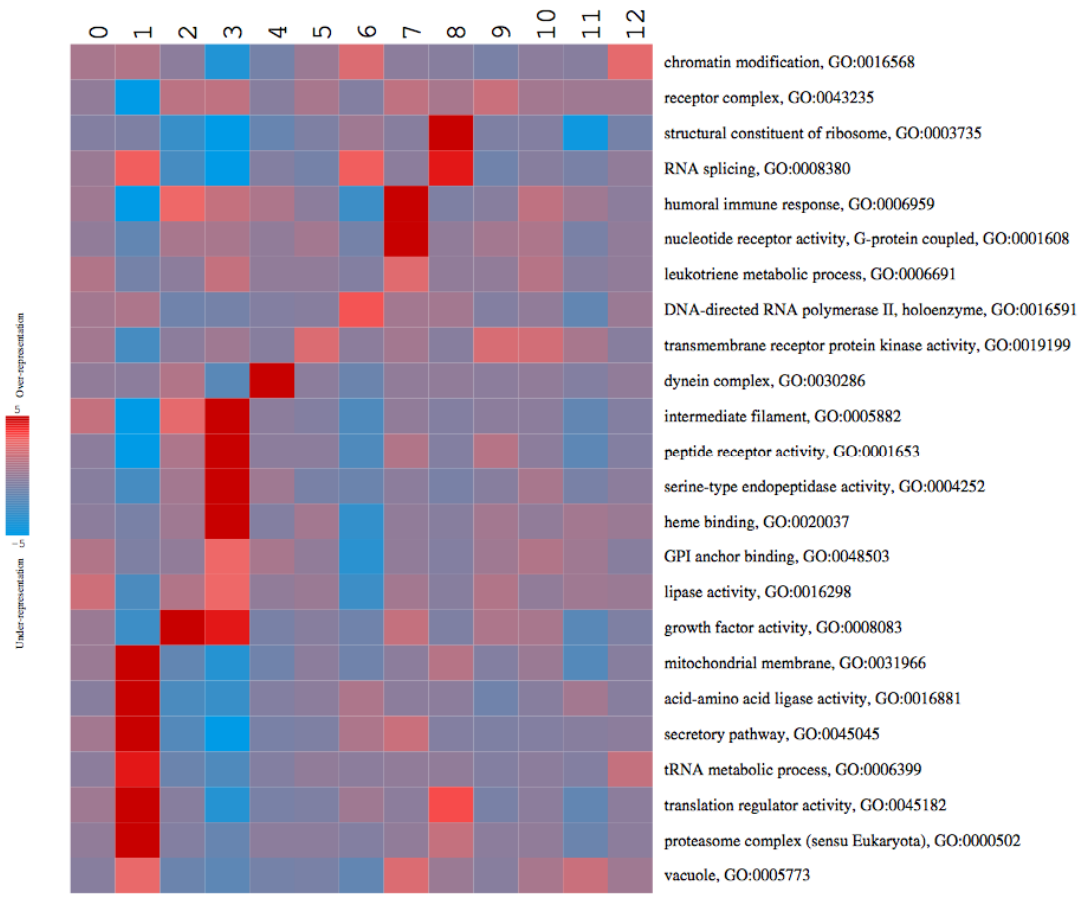


Figure 9. 13 WCGNA modules iPAGE results.

Brain Regions of Interest and their DNA constructs

The emerging recombinant technology to insert genes encoding light sensitive proteins in strategic genomic locations for optogenetic research requires comprehensive brain atlas data to identify structures where a specific promoter motif governs a network of highly expressed genes. Such promoters or genomic regions represent the basis for an artificial DNA construct.

The hypoglossal nucleus is the synapse of axons descending from the myencephalon to the hypoglossal nerve that has direct control over muscular tongue movement. Elucidation of the function of this nerve pathway has important implications for speech pathology. Using BEXPASS we assembled expression

profiles for the hypoglossal nucleus and ran it through FIRE (default stringency parameters, continuous distribution, 32 bins of ~1000 genes per bin) to yield promoters that govern high expression in the hypoglossal nucleus. Results are shown in Figure 10. On a fold induction basis there are no strong motifs for genes that are specifically enriched in this region. However, based on absolute rankings, the uncharacterized 3'UTR motif [CGU]C[AC]NUAAA is the only overrepresented in the two bins of most highly expressed genes, albeit without a particularly strong Z-score of 15.1. These two bins show ontological enrichment for, GO:0003735 structural constituent of ribosome, GO:0000786 nucleosome, GO:0005740 mitochondrial envelope, GO:0003954 NADH dehydrogenase activity, and GO:0000502 proteasome complex (sensu Eukaryota).

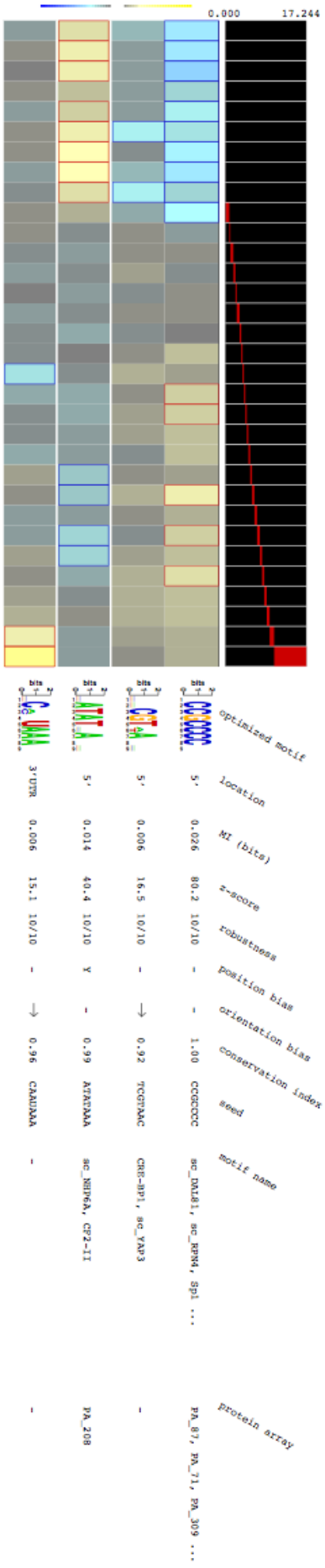
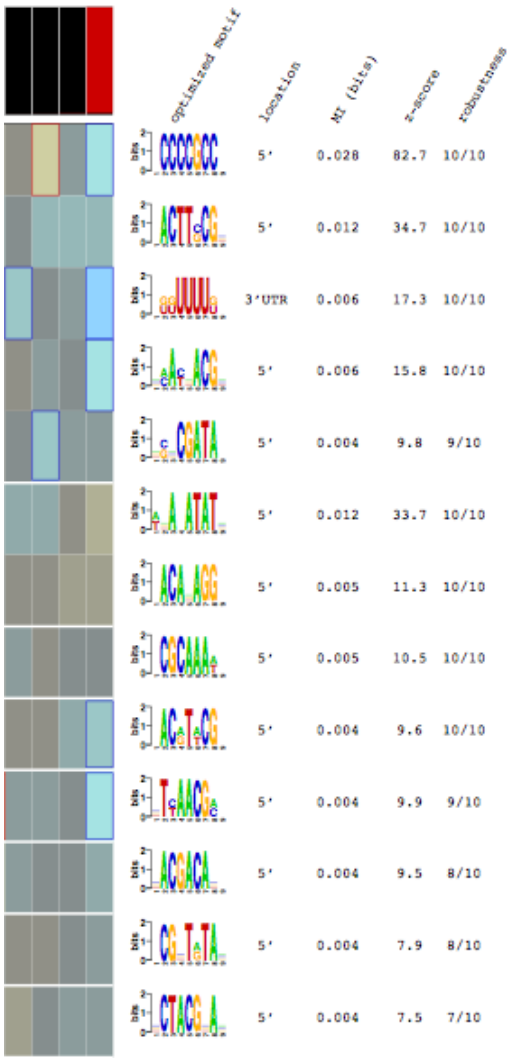


Figure 10. FIRE results for hypoglossal nucleus. A) Left, lateral view, results for expression profile on an absolute ranking basis with one good motif candidate for a DNA construct. **B)** Below, redacted results for expression profile on a ratio / fold induction basis. Only the most induced bins shown, with no good motif candidates for a DNA construct yielded.



The right cerebellum has been known to be an important language-processing center of the brain. More functionally specific, ignoring background noises and other people speaking in order to hear a specific sound of interest is the process of "suppression of interference", an area of interest in neurocognitive research (20). Filipi et al. organized a subject pool of native Italian speakers who were also conversant in English (native language=L1, second language=L2). Subjects were asked to listen to simultaneous audio tracks of sentences of different subjects in L1 and L2 and asked follow up questions about the L2 sentences in order to assess their comprehension and ability to block out L1. All of this was done while subjects while undergoing brain magnetic resonance imaging. That imaging data was mined and researchers were able correlate higher gray matter density in the right lateral paravermis of the cerebellum to better control of interference. From a ratio / fold induction expression profile there are no strong motifs that would serve as good DNA constructs for insertion into the right lateral paravermis. An absolute value expression profile yielded three overrepresented motifs. The first is CCCGCCC, a common binding motif that has showed up as the strongest motif across virtually all FIRE runs in our research. The second motif is N[ACT]ACT[AT]CCG with a strong Z-score of 37.7. These two motifs both govern the highest expression bin that is enriched for GO:0003735 structural constituent of ribosome, GO:0003954 NADH dehydrogenase activity, GO:0006334 nucleosome assembly, GO:0044455 mitochondrial membrane part, and GO:0006007 glucose catabolic process. The third motif is the uncharacterized 3' UTR motif N[CU]AAUAAA, which is very similar to the [CGU]C[AC]NUAAA motif we proposed as a construct candidate in the hypoglossal nucleus and both originated from the seed CAAUAAA.

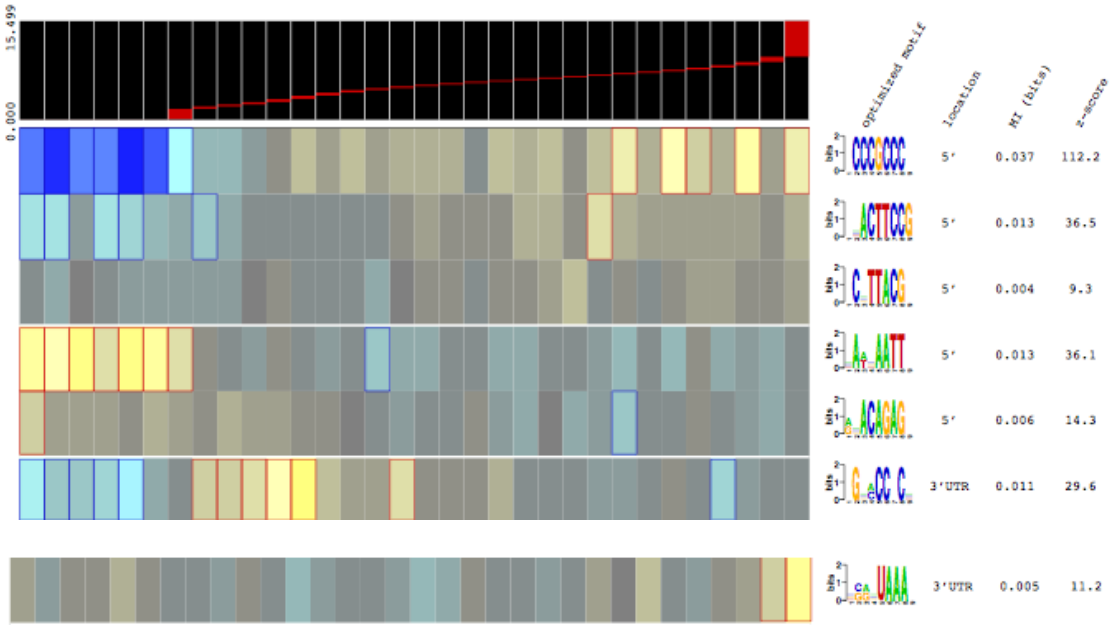


Figure 11. Abridged FIRE results for absolute value expression profile of the right paravermis.

More fundamental than the concept of suspension of interference is that the fluid interchange between and simultaneous use of L1 and L2 lies within the same brain substructures and to a lesser extent the same neural circuits (21). Crinion et al. used a similar bilingual subject pool (German-English and Japanese-English) to test whether semantic activation is independent of the language stimuli. Their method involved presenting word combinations with related meaning (trout-SALMON) or unrelated meaning (trout-HORSE). The first word (prime) and second word (target) were written in every two-by-two pairwise permutation of L1 and L2. Whole brain neuroimaging was done through positron emission tomography (PET) and functional MRI. Crinion et al. were able to showed increased activation in the left caudate when prime and target were in different languages and lowered activation levels when they were in the same language; this is evidence that "different languages are processed [to some extent] by different neural populations". Again, to find promoters governing highly expressed genes in the left body of caudate and the left head of caudate we ran

BEXPASS produced expression profiles through FIRE and iPAGE with continuous distribution, 32 bins, and default stringency parameters.

The left head of caudate the ratio / fold induction expression profile yielded a strong 5' TATA box motif for the highest expressed bin with a Z-score of 39.4; this bin was enriched for GO:0008227 amine receptor activity. The absolute value expression profile yielded two strong motifs. The first was the recurrent CCGCCCC which was overrepresented in the 5th,6th, and 7th highest expressed bin of genes; the 5th bin showed enrichment for GO:0008380 RNA splicing. The second motif was the 3' UTR characterized binding site for microRNAs, N[AU][GU]UUU[GU]U[AGU], in the 2nd-8th bins and had a Z-score of 34.9. MicroRNAs are short strands of RNA that, along with a group of proteins including RNase, form an RNA-induced silencing complex (RISC) which regulate roughly 25% of the human genome. The mechanism of action occurs when the RISC bonds with strands of complementary mRNA and silence its translation by degrading it (22). MicroRNA function has emerged as a therapeutic class of molecules primarily as a silencer of oncogenes and its role in the brain is understood primarily in areas of neurodevelopment and cellular differentiation (23). Given their role as silencers, the N[AU][GU]UUU[GU]U[AGU] motif would not be a good promoter for a DNA construct.

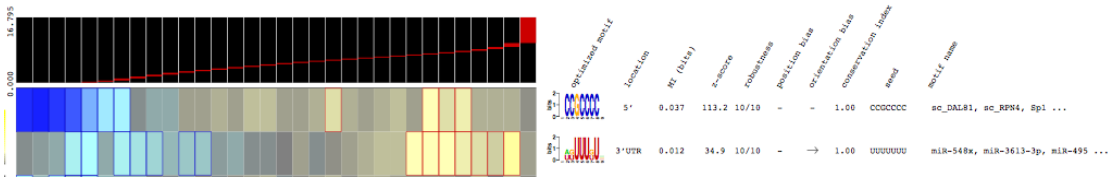


Figure 12. Absolute value expression FIRE results for the left head of caudate.

The left body of caudate ratio / fold induction expression profile yielded no strong motifs while the absolute expression profile yielded four motifs. The first motif is the recurrent 5' CCGCCCC. The second motif is the uncharacterized 5' motif [AC]N[AT]ACG[CGT]N and is highly enriched in the 4th most highly expressed bin. That bin is functionally enriched for GO:0006334 nucleosome assembly and GO:0008380 RNA splicing. The third motif is a basic leucine zipper binding site [AC]CG[AT]NATC[GT] enriched in only a single bin without any functional enrichment. The fourth motif is the uncharacterized 3' UTR motif [CGU][AGU]N[CGU]CGUU[ACU] whose bin is enriched for GO:0030286 dynein complex. See figure 13 for complete left body of caudate FIRE results.

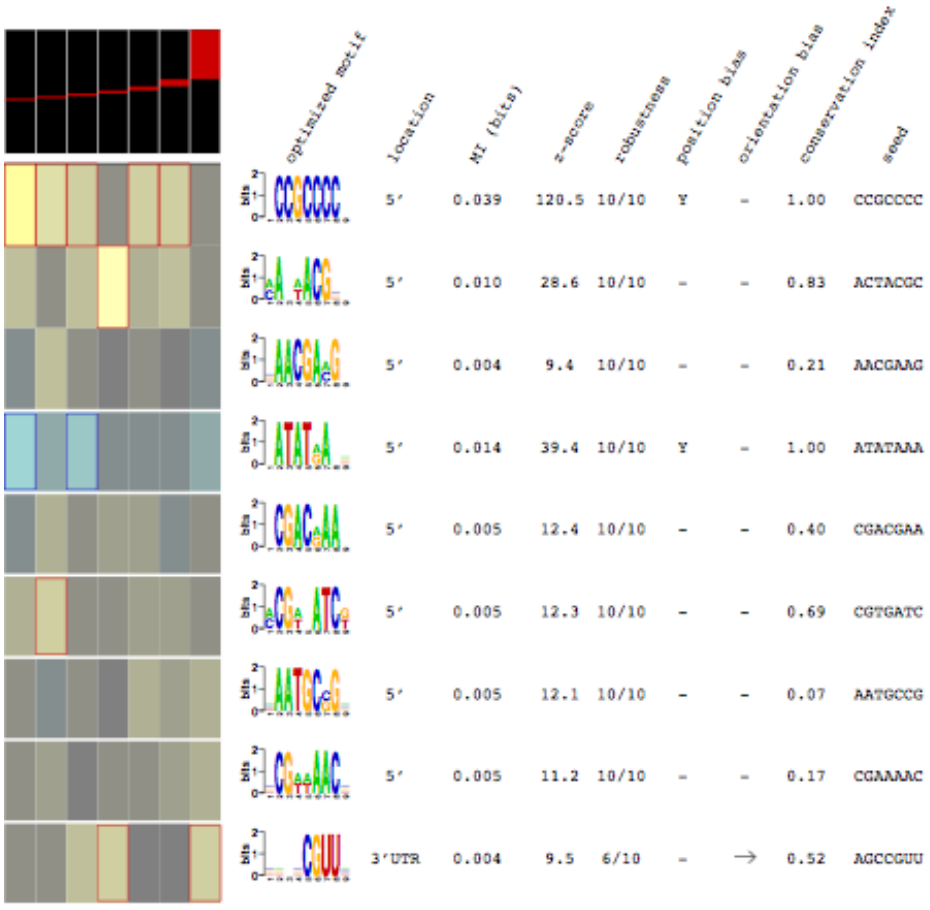


Figure 13. Abridged Absolute value expression FIRE results for Left Body of Caudate.

V. References

1. Hawrylycz et al. An anatomically comprehensive atlas of the adult human brain transcriptome. Nature 489, 391-399. Sept 2012.

2. Lein et al. Genome-wide atlas of gene expression in the adult mouse brain. Nature 445, 168-176 Jan 2007.

3. Henry AM et al. High-resolution gene expression atlases for adult and developing mouse brain and spinal cord. Mammalian Genome. 23(9-10):539-49. Oct 2012.

4. Oldham, M. C. et al. Functional organization of the transcriptome in human brain. Nature Neurosci. 11, 1271–1282. 2008.

5. Tavazoie et al. A universal framework for regulatory element discovery across all genomes and data-types. Molecular Cell. 28(2):337-50. Oct 26 2007.

6. Foat et al. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast" PNAS 102(49), 17675-17680, Dec 6 2005.

7. Supplemental materials: A universal framework for regulatory element discovery across all genomes and data-types. Molecular Cell. 28(2):337-50. Oct 26 2007.

8. Pastrana et al. Optogenetics: controlling cell function with light. Nature Methods 8, 24-25 (2011). Published online Dec 20 2010.

9. Shichida et al.Evolution of opsins and phototransduction. Phil. Trans. R. Soc. B 364, 2881–2895. 2009.

10. Madisen et al. A toolbox of Cre-dependent optogenetic transgenic mice for light-induced activation and silencing. Nat Neurosci. 15(5):793-802. Mar 25 2012. The Jackson Laboratory. http://research.jax.org/grs/optogenetics.html.

11. Allen Human Brain Atlas. Technical White Paper: Microarray Data Normalization. March 2013 v.1. http://help.brain-map.org/download/attachments/2818165/Normalization_WhitePaper.pdf?version=1&modificationDate=1361836502191

12. Allen Human Brain Atlas. Technical White Paper: Microarray Survey. Jun 2013 v.6. http://help.brain-map.org/download/attachments/2818165/Microarray_WhitePaper.pdf?version=1&modificationDate=1370991311181

13. Ensembl Genome Browser. Accessed Jun 5, 2013 http://useast.ensembl.org/index.html

14. Beißbarth et al. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics Applications Note. Vol. 20 no. 9. 2004.

15. Goodarzi et al. Revealing global regulatory perturbations across human cancers. Molecular Cell. 900-911. Dec 11 2009.

16. Lancaster JL et al, "Automated Talairach Atlas labels for functional brain mapping". Human Brain Mapping 10:120-131, 2000.

17. Morel et al. Primer on Medical Genomics Part XIV: Introduction to Systems Biology—A New Approach to Understanding Disease and Treatment. Mayo Clinic Proceedings. Mayo Clin Proceedings. 79(5):651-8. May 2004.

18. Oldham et al. Functional Organization of the Transcriptome in Human Brain. Nature Neuroscience.11(11): 1271–1282. doi:10.1038/nn.2207. Nov 2008.

19. Cahoy et al. A Transcriptome Database for Astrocytes, Neurons, and oligodendrocytes: a new resource for understanding brain development and function. The Journal of Neuroscience. 28(1):264-78. Jan 2 2008.

20. Filippi et al. The Right Posterior Paravermis and the Control of Language Interference. The Journal of Neuroscience. 31(29):10732–10740. Jul 20 2011.

21. Perani et al. The Neural Basis of First and Second Language Processing. Current Opinion in Neurobiology. Current Opinion in Neurobiology.15(2):202-6. Apr 2005.

22. Ross et al. miRNA: The New Gene Silencer. American Journal of Clinical Pathology.128(5):830-6. Nov 2007.

23. Liu et al. MicroRNA in central nervous system trauma and degenerative disorders. Physiol Genomics. 43(10): 571–580. May 2011.