AD_____

Award Number: W81XWH-08-1-0420

TITLE: Brain Region and Cell Type Transcripts for Informative Diagnostics

PRINCIPAL INVESTIGATOR: Leroy E. Hood, Ph.D., M.D.

CONTRACTING ORGANIZATION: Institute for Systems Biology
Seattle WA 98103

REPORT DATE: July 2009

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
15-07-2009	Annual Report	16 JUN 2008-15 JUN 2009

4. TITLE AND SUBTITLE

Brain Region and Cell Type Transcripts for Informative Diagnostics

5a. CONTRACT NUMBER

W81XWH-08-1-0420

5b. GRANT NUMBER

Grant Proposal No. 08024001

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Leroy E. Hood, PhD, MD
Email: LHOOD@SYSTEMSBIOLOGY.ORG

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Systems Biology
Seattle WA 98103

8. PERFORMING ORGANIZATION REPORT NUMBER

ISBAnnualTATRC09

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Army Medical Research and Material Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

TATRC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Our objectives, have and will provide a set of candidate markers that can be used not only for the diagnosis of brain trauma, such as traumatic exposure to explosions, but also diseases such as brain cancer, including glioblastoma, and neural degenerative diseases, such as Alzheimer's disease.

We were able to accomplish the first of our two main objectives – we have identified a large number of brain-region specific markers and identified those that are secreted. These markers are highly important because they offer the potential to identify brain-region specific legions non-invasively through the blood. Thus, the first phase of the project has been a success.

15. SUBJECT TERMS

brain-region specific markers

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE		14	USAMRMC
					19b. TELEPHONE NUMBER *(Include area code)* (301)-619-2254, Lisa Sawyer, Contract Specialist

Reset

Table of Contents

Annual Report for DoD Grant (Award Number: W81XWH-08-1-0420)

1. Introduction

The ability to diagnose diseases of the brain is limited by the types of diagnostic tools available: biopsy of the brain searching for pathology (a dangerous and complex procedure), imaging of the brain (a generally rather low resolution procedure) and assessment of cerebral spinal fluid (a complex and painful procedure). The ability to diagnose brain pathology is even more critical to the Department of Defense because of the many of the soldiers returning from Iraq after exposure to explosions appear initially normal but later begin to exhibit the effects of traumatic brain injury. Traumatic brain injury (TBI) is estimated to affect 1.4 million Americans per year and to cost the United States' (US) economy $60 billion. Currently, 5.3 million Americans suffer TBI-related, long-term disability. As such, TBI is a major public health concern in the US. Further, given the field of operations of today's US military, TBI has also become a major socioeconomic issue for our armed forces. We set out to identify brain region and cell-type specific transcripts through analysis of data from the Allen Brain Institute as a precursor to make possible the use of blood fingerprints (e.g. cerebellum, cerebrum, basal ganglia, brain stem, etc.) that will not only allow early assessment of brain damage after exposure to trauma, but also its localization. The completion of our These brain region and cell type specific transcripts will set the stage for future studies where we use proteomics approaches to screen for the associated proteins in the blood. The development of such multi-parameter blood protein markers lies at the very heart of the predictive medicine that will emerge over the next 10 years.

2. Body

The technical objectives of our proposal were two-fold. First, we set out to use information about gene expression in individual cells of mouse brains from the Allen Brain Atlas to identify transcripts that are specific to particular regions of the brain (e.g. cerebellum) or to specific cell types (e.g. neurons). Second, we set out to identify from the identified specific transcripts those with human orthologs and to predict computationally which are most likely to be secreted or membrane-bound. As shown below, we have succeeded in our objectives in relation to brain-region specific transcripts, but more work remains to be done in order to fully achieve the same level of success for cell-type specific transcripts.

2.1. Accessing the Allen Brain Atlas (ABA)

Our first task was to acquire data from the Allen Brain Atlas in a form that we could mine to identify those transcripts that are region and cell-type specific. The Allen Institute for Brain Science offers a public API to allow access to their database – including access to the brain image, gene expression, atlas, and neuroblast datasets (Figure 1). This API also includes a REST interface to interact with selected URLs and obtain additional required information for our analyses.

To facilitate computations, we locally mounted a replication of the ABA database consisting of XML (~200 MB) and XPR (~7.5 Gb) files – and we did so both at the Institute for Systems Biology and in the Price Lab at the University of Illinois. The XML files contain the relational database that enables us to identify the image series associated with a particular gene or probe. These files also enable us to access the public API to obtain the images and models. The XPR files are a 3-D registration approach used by the Allen Brain Institute for image analysis of gene series with 100 microns of resolution.

During the first quarter we wrote Perl-based scripts to read and parse the XML data and to retrieve the images associated by gene/ISH.

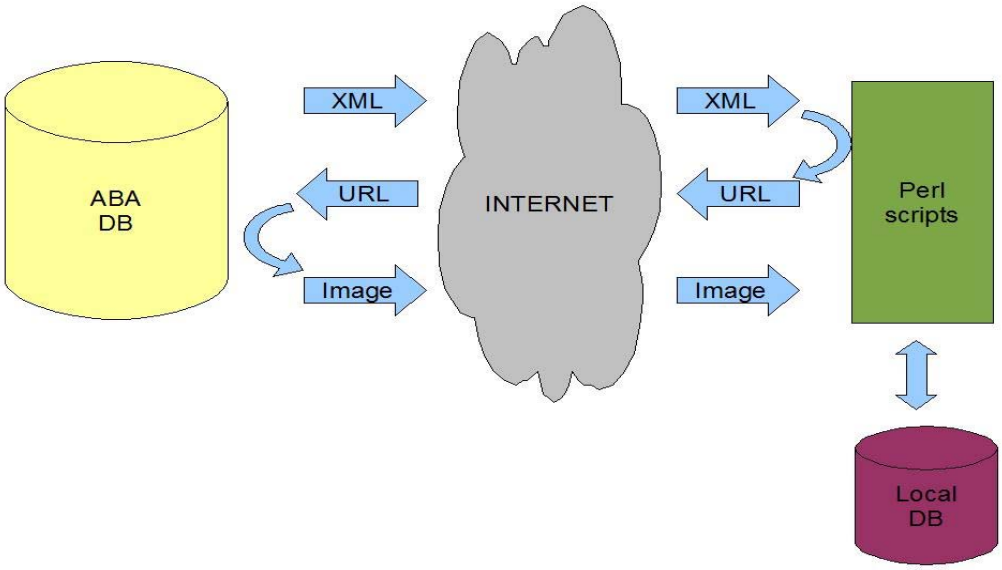

Figure 1. Basic representation of the interaction with the ABA-API.

2.2. Querying the ABA

We used AllenMiner v1.0 [1] as a means to submit queries to analyze the data represented in the XPR files. This software package allows us to extract gene expression profiles for regions of interest (ROI) according to the ABA nomenclature for brain sections or by using voxels coordinates. Also it implements an enrichment method for ROI contrasts. After we mounted and adapted the AllenMiner software to our local computational resources, we used it to scan the full database to query each annotated ROI (209 hierarchical categories) for each gene (~20,000) in each of the XPRs (~26,000). This analysis resulted in expression profiles for ~5 million gene-3D model combinations.

Many of the transcripts can be mapped into at least one ROI across the 3D model series. Some transcripts have evidence of expression in more than one 3D model (Table 1), for example when the same genes had been mapped into both the sagittal and coronal series. Duplicate data generally correlated, but in some cases we saw discrepancies between values in repeats, probably caused by image artifacts or other effects.

Table 1. Image series by gene.

Series	Frequency	Percentage %
1	1,569,613	67.92
2	660,582	28.58
3	54,529	2.36
4	6,634	0.29
5	3,421	0.15
6	8,735	0.38
7	3,416	0.15
8	352	0.02

9	177	0.01
11	368	0.02
12	605	0.03
13	416	0.02
14	612	0.03
15	179	0.01
16	625	0.03
17	417	0.02
18	209	0.01
19	191	0.01
TOTAL	**2,311,081**	**100.00**

2.3. Gene specificity

To determine the gene specificity by ROI, we tested different methods for specificity. First we calculated a conditional specificity (Q) using the Shannon entropy of the symbols [2]:

$$\begin{aligned} p_{t|g} &= \frac{w_{t|g}}{\sum_{t \in T} w_{t|g}} \\ H_g &= -\sum_{t \in T} p_{t|g} \lg(p_{t|g}) \\ Q_{g,t} &= H_g - \lg(p_{t|g}) \end{aligned}$$

where $w_{t|g}$ represents the relative expression of the gene g in the region t, H_g is the entropy for the relative expression of gene g, and $Q_{g,t}$ is the conditional specificity for gene g in the region t.

Generally, a tissue specific transcript has a Q value from 0 to 7, but in our data set the distribution is different than this norm because of the specifics of our hierarchical classification (Figure 2).

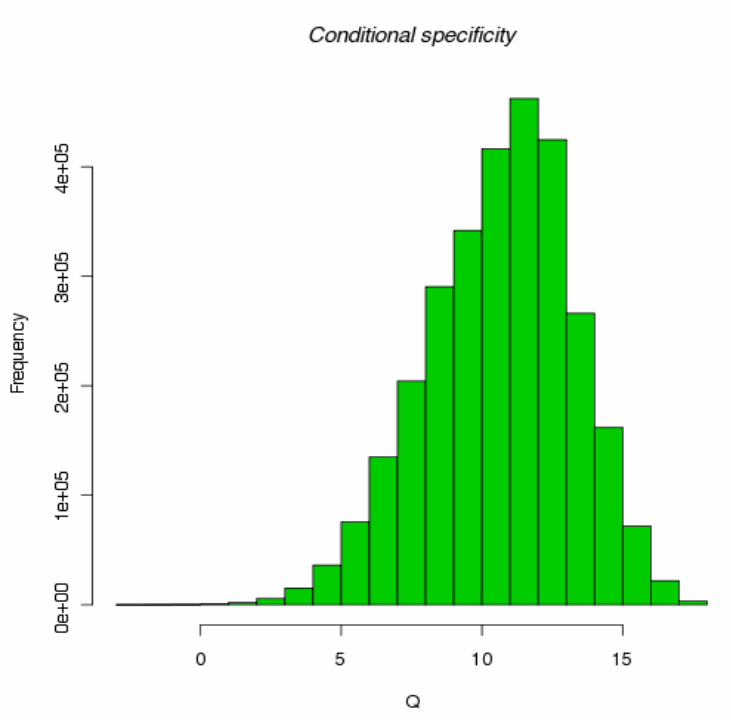

Figure 2. Q-values distribution for all gene series.

2.4. Basic brain regions

To reduce the complexity of the regions screened and calculate a specificity score Q, we selected these 17 basic non-overlapping brain regions to consider in our analyses:

1. Cerebellum (CB)
2. Cerebral cortex (CTX)
3. Hippocampal region (HIP)
4. Hippocampal formation (HPF)
5. Hypothalamus (HY)
6. Lateral septal complex (LSX)
7. Midbrain (MB)
8. Medulla (MY)
9. Olfactory areas (OLF)
10. Pons (P)
11. Pallidum (PAL)
12. Retro-hippocampal region (RHP)
13. Striatum (STR)
14. Striatum dorsal region (STRd)
15. Striatum ventral region (STRv)
16. Thalamus (TH)
17. Striatum-like amygdalar nuclei (sAMY)

We calculated the specificity of each gene in these brain regions. The expression value is normalized between duplicated images series, if require, in order to allow mixed data from coronal and sagittal samples. As an example, Rap1 interacting factor 1 homolog (yeast) of *Rif1* is a gene selected to be expressed specifically in the cerebellum region ($Q_{Rif1,CB}$ = 1.66). We validated this selection with the image series (Figure 3) and the 3D model reconstruction (Figures 4-5).

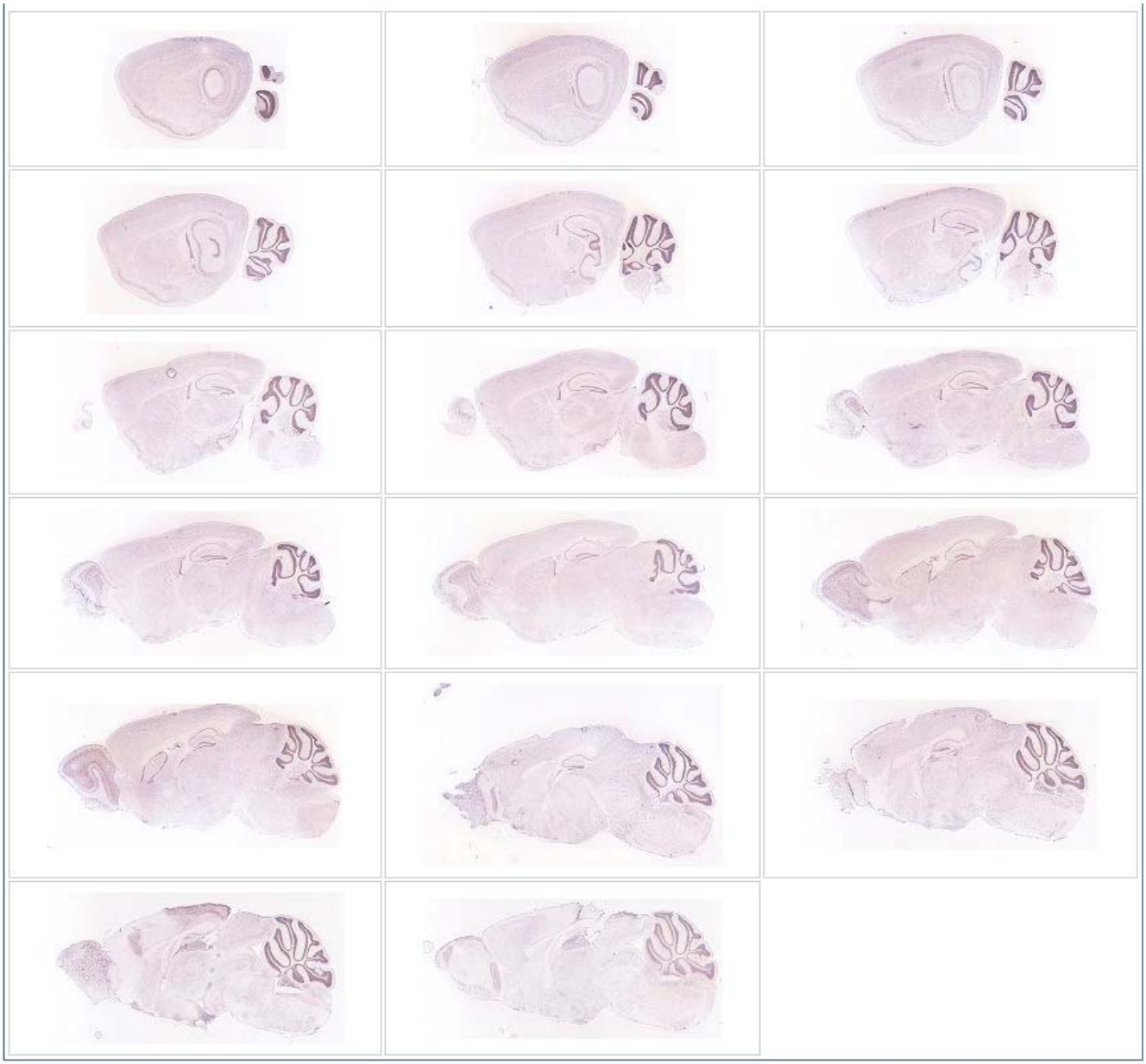

Figure 3. ISH image series for *Rif1*, sagittal.

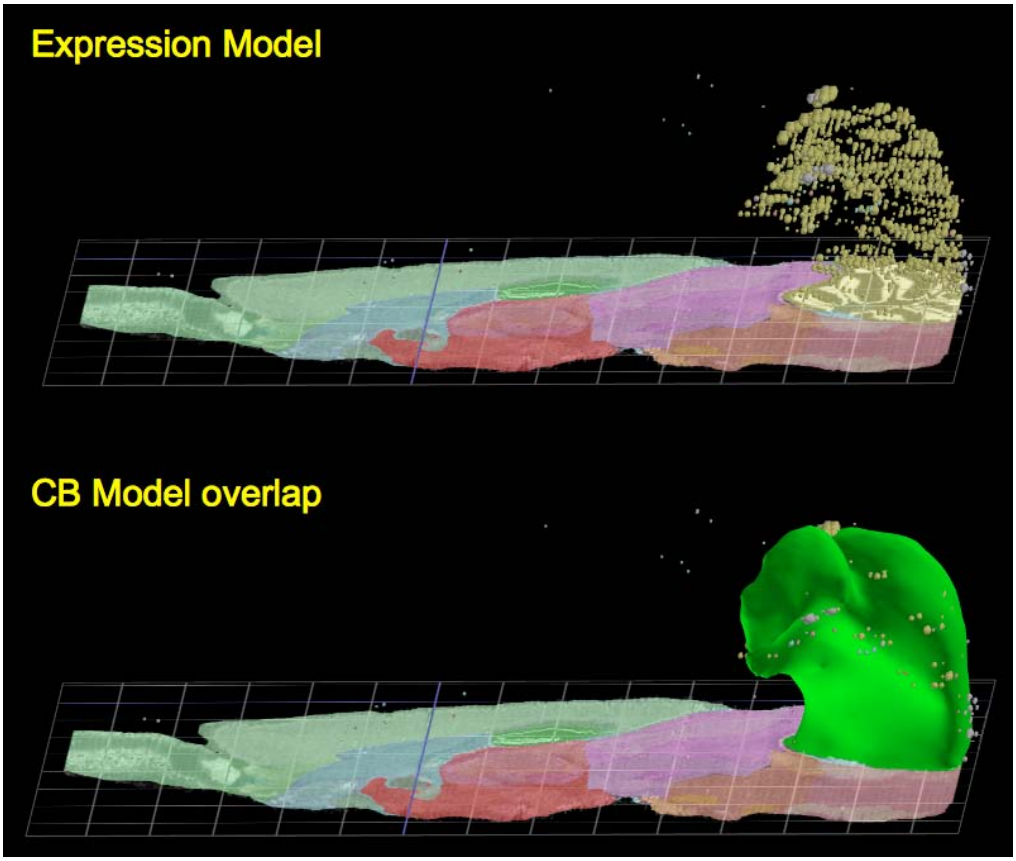

Figure 4. Expression profile for Rif1: a) gene expression 3D pattern in sagittal view, b) cerebellum 3D model overlapping the gene expression.

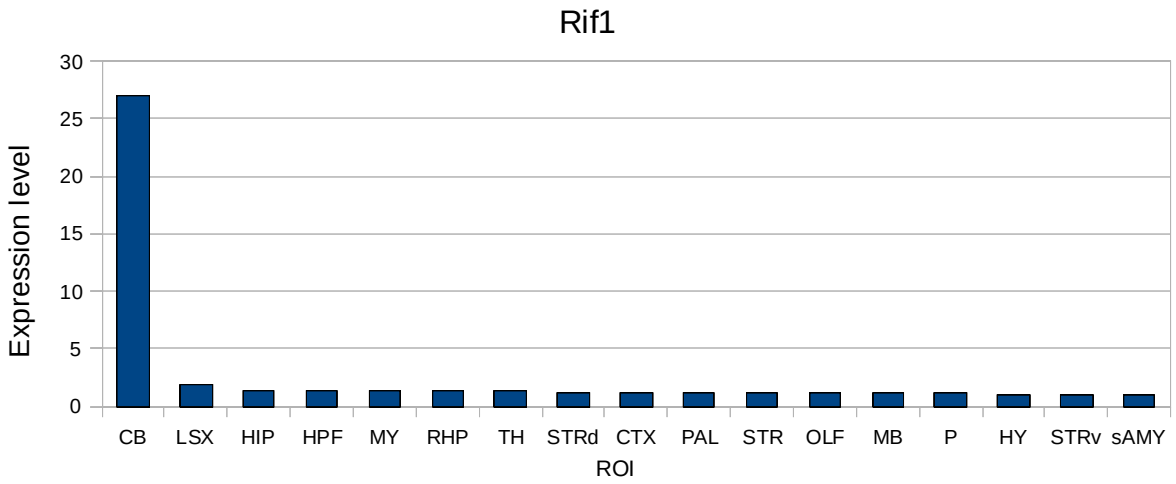

Figure 5. Expression levels of *Rif1* in the basic brain regions.

3. Key Research accomplishments

- ***Brain-region specific transcripts***

We explored beyond the basic ROIs described before. We used the hierarchical classification from the ABA to define the relationships between all ROIs. This classification defines 6 levels; we analyzed each level comparing

Table 2. ROI-specific genes.

LEVEL	ROI	Parent ROI	ROI-specific genes
1	BS	Brain	100
1	CB	Brain	131
1	CH	Brain	5902
2	CBX	CB	129
2	CNU	CH	43
2	CTX	CH	5578
2	HB	BS	27
2	IB	BS	16
2	MB	BS	4
3	CTXpl	CTX	404
3	HY	IB	1
3	MY	HB	16
3	PAL	CNU	1
3	STR	CNU	39
3	TH	IB	15
4	DORpm	TH	3
4	DORsm	TH	1
4	HPF	CTXpl	107
4	MY-mot	MY	2
4	OLF	CTXpl	172
4	PALd	PAL	1
4	STRd	STR	31
4	STRv	STR	2
5	AON	OLF	8
5	CP	STRd	31
5	GENv	DORpm	1
5	HIP	HPF	93
5	MARN	MY-mot	1
5	MOB	OLF	77
5	OT	STRv	1
5	RHP	HPF	2
5	VENT	DORsm	1
6	CA	HIP	35
6	DG	HIP	9
6	IGL	GENv	1
6	PIR	OLF	3

the gene expression in ROI versus the other ROIs in the same level from the same parent ROI. The specificity is calculated with a log-likelihood:

$$L(G \mid R) = \log_{10}(G_R) - \log_{10}(\sum G_{NR})$$

where G_R is the expression level for gene G in ROI R and G_{NR} is the expression of the same gene in the others ROIs from the same parent ROI. We selected the genes with $L > 1$, which means that the expression level is 10-fold for the selected ROI.

- ***Comparison with mouse genome microarrays***

Our group has experimental data for gene expression in mouse brains, using hybridizations of whole brain (Mouse Gene 1.0 ST Affymetrix arrays), as well as cortex and thalamus sections. We compared the expression levels for each gene between these microarray experiments and the ABA data (Figure 5).

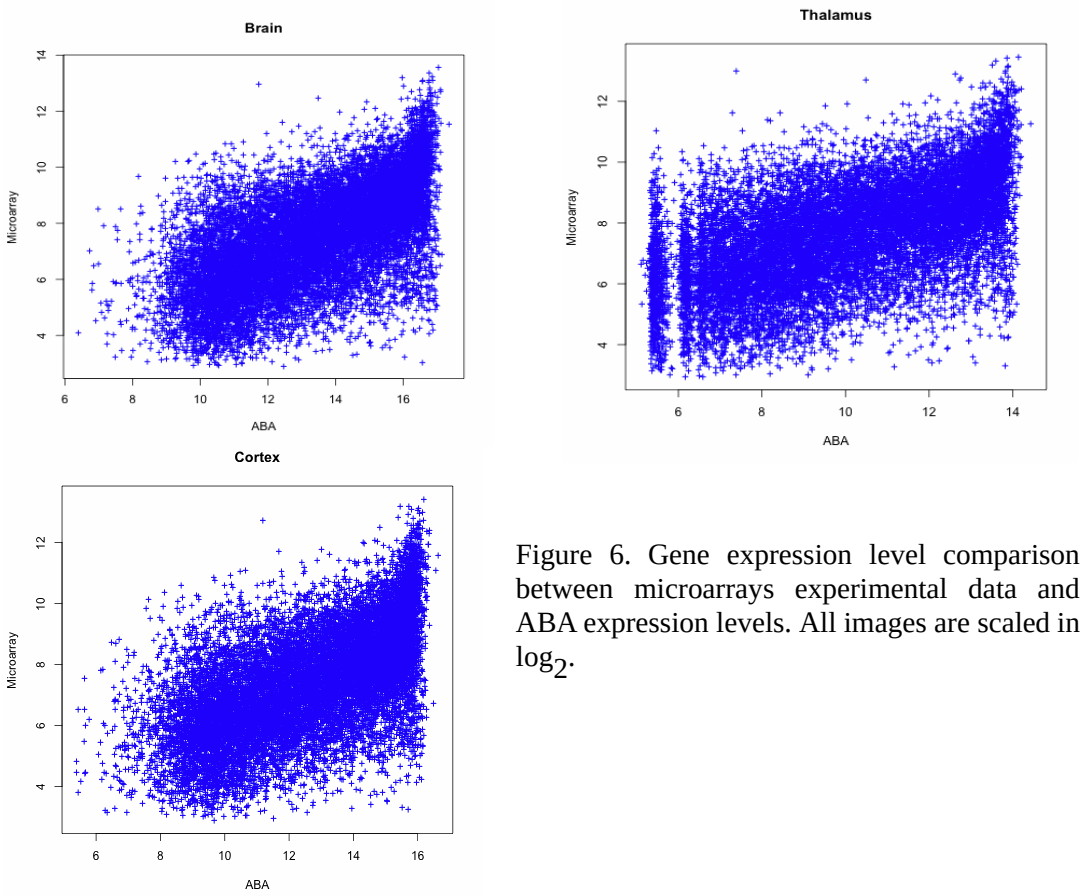

Figure 6. Gene expression level comparison between microarrays experimental data and ABA expression levels. All images are scaled in $\log_2$.

- ***Mouse secretome***

One of the primary objectives of our proposal was the definition of secreted proteins to use as candidate biomarkers that would be brain-region specific. For all the peptides present (~40,000) in the UCSC Genome Browser (knowGenesPeptide table for mm9), we computed the probability of the associated protein being secreted using the SignalP v3.0 software suite [3]. We found that ~10% of the total peptides could be secreted ($P > 0.95$). Because the gene symbol used in UCSC doesn't correspond to the gene symbol used in the ABA, we cross-referenced both databases using the gene alias. Thus, *we coupled this data with the sets of genes that are identified as being brain region specific in order to identify the candidate secreted proteins in the blood with the potential to provide brain-region specific diagnostics through non-invasive measurements.*

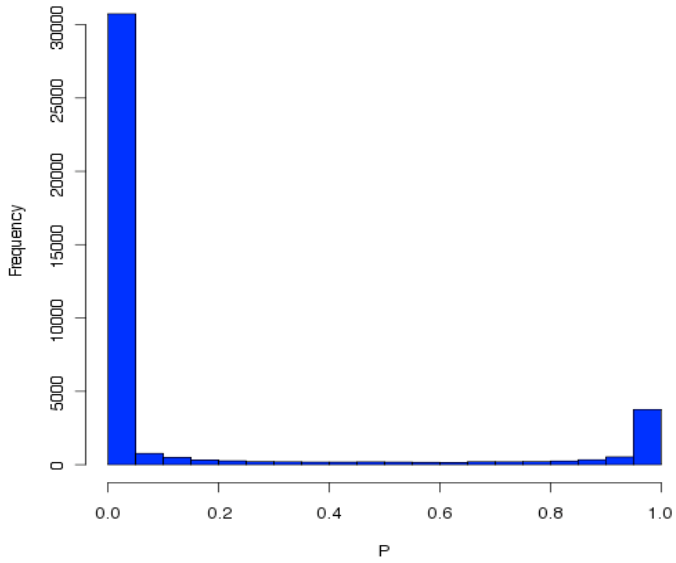

Figure 7. Distribution of P-values for SignalP predictions.

After identifying genes expected to be secreted in mouse, we applied the same methods to human genes and identified the ortholog pairs between both species, using orthology relationships reported in Mouse Genome Informatics website (http://www.informatics.jax.org/).

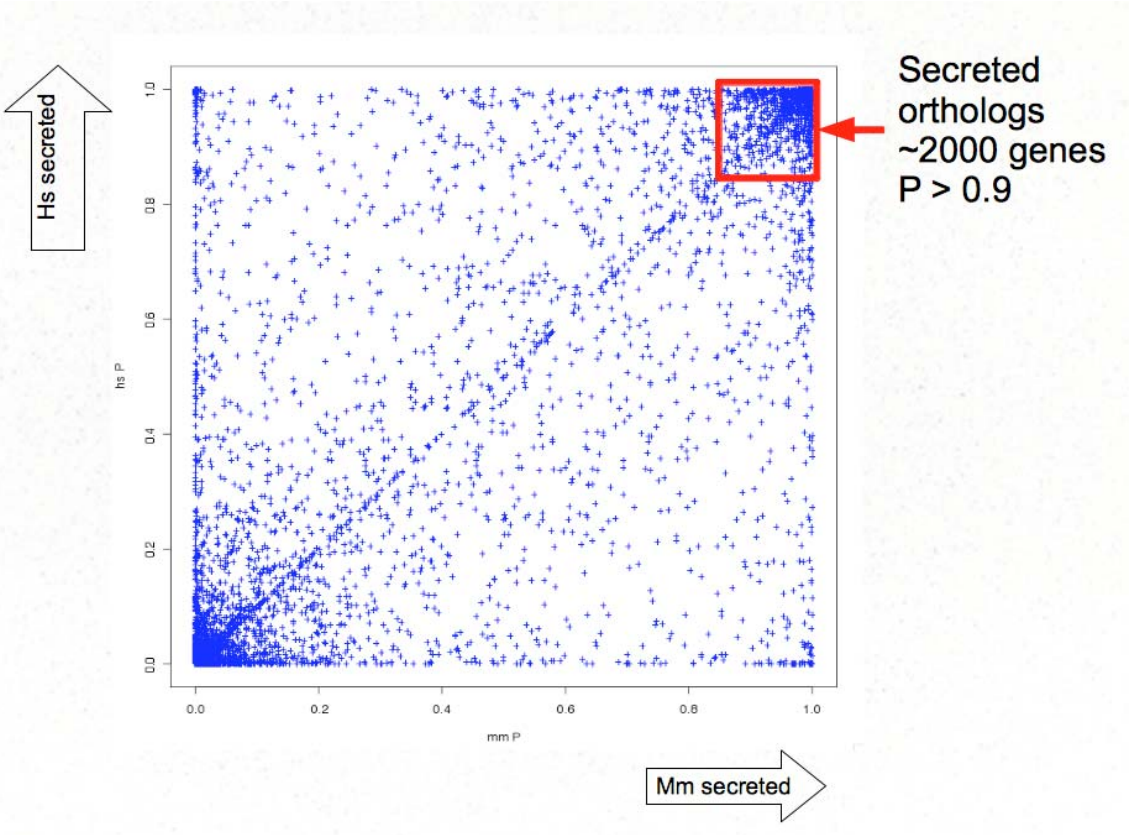

Figure 8. Signal P-values for human and mouse ortholog genes.

- ***Cell-type markers***

Moving on to our second objective, we selected cell-specific genes to be used as training sets for cell-type classification. The list of training genes was obtained from [4], we have 319 neuron-specific, 184 astrocyte-specific and 130 oligodendrocyte-specific genes. The respective image series in high resolution for each gene had been obtained using our interface for the ABA API.

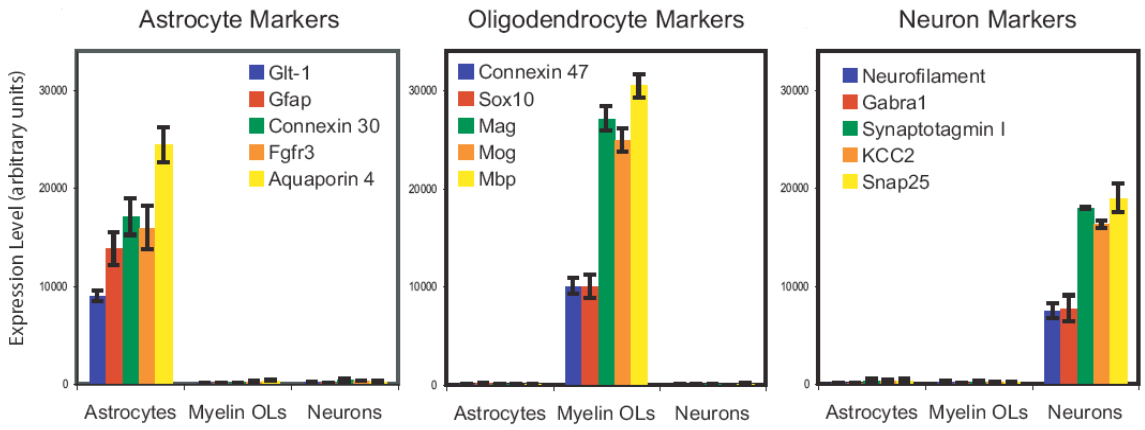

Figure 9. Expression levels of cell-type markers (adapted from [4]).

4. Reportable Outcomes

Leroy Hood has given approximately 15 keynote lectures over the past year throughout the US and in Asia and Europe—and in most of these has discussed the concept of organ-specific protein blood markers for brain and for liver. In addition, he has discussed the concept of brain-region-specific transcripts obtained from the computational analyses of data in the Allen Brain Database. Experimental work is on-going to determine how many of these candidate brain-region-specific transcripts encode proteins that are secreted into the blood. We have obtained funding that leverages this approach to identifying brain-specific blood biomarkers from the following sources: TBI DOD grant with Georgetown (put in grant number), PTSD DOD grant with Santa Barbara (put in grant number), and a strategic partnership grant from the state of Luxembourg to apply the identification of brain-specific blood markers to the study of neural degenerative diseases and brain cancer (Alzheimer's, Frontal Temporal Dementia and Glioblastoma).

5. Conclusion

In the year of funding we received from DoD, we were able to accomplish the first of our two main objectives – we have identified a large number of brain-region specific markers (See Table 2) and identified those that are secreted (see Figure 8). These markers are highly important because they offer the potential to identify brain-region specific legions non-invasively through the blood. Thus, the first phase of the project has been a success. The second goal of identifying cell-type specific markers has proved more difficult than initially anticipated. We made initial progress and have a working "training set" of markers from ~200 genes for which there is evidence for cell-type specificity. We are using these findings to develop a computational algorithm that can be used to evaluate cell-type specificity from the ABI data for the remaining almost 20,000 genes. We have thus sought an additional year of funding to pursue this second critical objective (as outlined in the continuation proposal we submitted this summer). The ultimate outcome of the research funded by the DoD is thus of high importance, promising to provide high specificity to both brain-region specific (achieved) and cell-

type specific (in progress) proteins that can be monitored non-invasive through the blood.

References

1. Davis FP and Eddy SR, "*A tool for identification of genes expressed in patterns of interest using the Allen Brain Atlas*", Bioinformatics, 2009.
2. Schug J, Schuller, WP, Kappen C., Salbaum MJ, Bucan M and Stoeckert CJ, "*Promoter features related to tissue specificity as measured by Shannon entropy*", Genome Biology, 2005.
3. Bendtsen JD, Nielsen H, von Heijne G, Brunak S, "*Improved prediction of signal peptides: SignalP 3.0*", J Mol Biol., 2004.
4. Cahoy JD, Emery B, Kaushal A, Foo LC, Zamanian JL, Christopherson KS, Xing Y, Lubischer JL, Krieg PA, Krupenko SA, Thompson WJ, Barres BA, "A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function", J Neurosci., 2008.