Database

An online database for brain disease research

Brandon W Higgs*†1, Michael Elashoff†1, Sam Richman†1 and Beata Barci2

Address: 1Elashoff Consulting LLC, Germantown, MD 20876, USA and 2Stanley Medical Research Institute, Bethesda, MD, 20814-2142, USA

Email: Brandon W Higgs* - brandon@elashoffconsulting.com; Michael Elashoff - michael@elashoffconsulting.com;
Sam Richman - sam@elashoffconsulting.com; Beata Barci - barcib@stanleyresearch.org

* Corresponding author   †Equal contributors

Abstract

Background: The Stanley Medical Research Institute online genomics database (SMRIDB) is a comprehensive web-based system for understanding the genetic effects of human brain disease (i.e. bipolar, schizophrenia, and depression). This database contains fully annotated clinical metadata and gene expression patterns generated within 12 controlled studies across 6 different microarray platforms.

Description: A thorough collection of gene expression summaries are provided, inclusive of patient demographics, disease subclasses, regulated biological pathways, and functional classifications.

Conclusion: The combination of database content, structure, and query speed offers researchers an efficient tool for data mining of brain disease complete with information such as: cross-platform comparisons, biomarkers elucidation for target discovery, and lifestyle/demographic associations to brain diseases.

Background

Brain disease studies based on experiments using genome-wide measurements with microarrays are traditionally challenging as compared to other disease areas. The biological results are often hindered by statistical issues of small sample sizes, small effect sizes, and patient-to-patient variability [1-3]. Also, clinical information for patients is typically sparse, such that unknown clinical covariates can either confound or confuse many of the gene expression patterns and trends, as opposed to the primary disease. Corrections using such clinical information can greatly improve inference in determining markers for disease, as well as elucidating patterns within the disease.

Technical problems in microarray data can also affect the analyses. Meaningful results are often limited by array platform-to-platform comparisons and overall organization/presentation of large data sets/results. Studies conducted on disparate platforms are inherently more difficult to analyze than those conducted on the same platform [4]. Cross-platform comparisons present analysis challenges due to differences in scaling and sensitivity (to name a few) which introduce inconsistencies in reproducibility [5-8]. Large data sets and comprehensive results summaries present another challenge that requires good organization of both analytical and bioinformatics information (e.g. expression profiles, gene summary information, pathway diagrams, fold change value comparisons, etc.) into a user-friendly format to facilitate efficient data mining. A relational web-based tool that logically com-

Table 1: Patient demographic variables for all diseases

All Patient Variables	Values
Age >45	Yes/No
PMI > 30	Yes/No
Brain pH > 6.5	Yes/No
Left Brain	Yes/No
Sex	Male/Female
Smoking at Time of Death	Yes/No
Herpes simplex virus 1 OD	High/Low
Herpes simplex virus 2 OD	High/Low
Toxolgg OD	High/Low
EBV OD	High/Low
HHV6 OD	High/Low
CMV OD	High/Low
Hervk 18 SNP	Positive/Negative
Hervk 18 Expression	Positive/Negative

Table 3: Patient demographic variables for Schizophrenic patients

Schizophrenic Patient Variables	Values
Schizophrenia Severity	Severe/Other
Schizophrenia Heavy Alcohol Use	Yes/No
Schizophrenia Heavy Drug Use	Yes/No
Schizophrenia	Paranoid/Undiff
Schizophrenia Sudden Death	Yes/No
Schizophrenia Suicide	Yes/No
Schizophrenia Lifetime Antipsychotics >45,000	Yes/No
Schizophrenia Antichollnergic	Yes/No
Schizophrenia Antidepressants	Yes/No
Schizophrenia Stabilizer	Yes/No
Schizophrenia Lithium	Yes/No
Schizophrenia Valproate	Yes/No

bines all of these factors can enhance researchers' ability to determine the underlying genomic patterns in brain disease.

The SMRIDB is an online data warehouse and analytical system designed to aid researchers in understanding the biological associations both between and within the brain disorders of schizophrenia, bipolar, and major depression. This open source database combines genomic patterns of brain disease with patient clinical metadata into a user-friendly query interface to enable efficient data mining for purposes of biomarker discovery and elucidating biological mechanisms of brain disease. The metadata includes a full summary of clinical history for each patient with hyperlinks to disease-level information, such that demographic- and lifestyle-associated effects can be determined as they relate to brain disorders. The genomic data has been compiled from 12 separate labs (identified as studies), each data set generated from brain tissue isolated from two controlled populations of 165 patients, diagnosed with one of the three brain disorders (plus unaffected control brain tissue). This genomic data has been

Table 2: Patient demographic variables for Bipolar patients

Bipolar Patient Variables	Values
Bipolar Severity	Severe/Other
Bipolar Heavy Alcohol Use	Yes/No
Bipolar Heavy Drug Use	Yes/No
Bipolar Psychotic Feature	Yes/No
Bipolar Sudden Death	Yes/No
Bipolar Suicide Status	Yes/No
Bipolar Lifetime Antipsychotics >0	Yes/No
Bipolar Antipsychotics	Yes/No
Bipolar Antidepressants	Yes/No
Bipolar Mood Stabilizer	Yes/No
Bipolar Lithium	Yes/No
Bipolar Valproate	Yes/No

generated across 6 separate human array platforms (Affymetrix: hgu133a, hgu133plus, hgu95av2, Agilent, Codelink, and cDNA custom array) providing patterns/trends and analytical inferences that are not limited by platform dependencies.

Construction and content
Bioinformatics mappings
NCBI's Database for Annotation, Visualization and Integrated Discovery (DAVID 2.0) was used as the standard source for gene annotation information [9]. The primary fields extracted from DAVID include: LocusLink, gene symbol, and gene summary. Additional annotations include gene product mappings to the Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Ontology Consortium (GO) for pathway and GO terms/classes, respectively. For Affymetrix arrays, queries were based on the Affymetrix probe ID (AFFYID). For other arrays, the Genbank accessions (GENBANK) were used.

Individual study-level analysis
For each of the individual studies, a series of analyses were performed. Each array (representing a single patient) was subjected to a quality control (QC) analysis for chip-level parameters (e.g. scaling factor, gene calls, control gene ratios, average correlation) with respect to the reference distribution for those parameters across the arrays. This QC analysis is represented with both graphical representations (e.g. heatmaps, scatter plots, and histograms (Figure 1)) and table summaries, allowing users to readily identify those arrays determined to be outliers in the study. A total of 41 clinical demographic variables (Tables 1, 2, 3, 4) were assessed for their effects on a gene-by-gene basis. Continuous variables and ordered categorical variables were cut at values as close as possible to the median (e.g. PMI>30 vs. PMI<30; Drug Use = 'Heavy' vs. Drug Use = 'None, Light, Moderate'). The genes determined to be most significant (p-value<0.01 and fold change >1.3) for
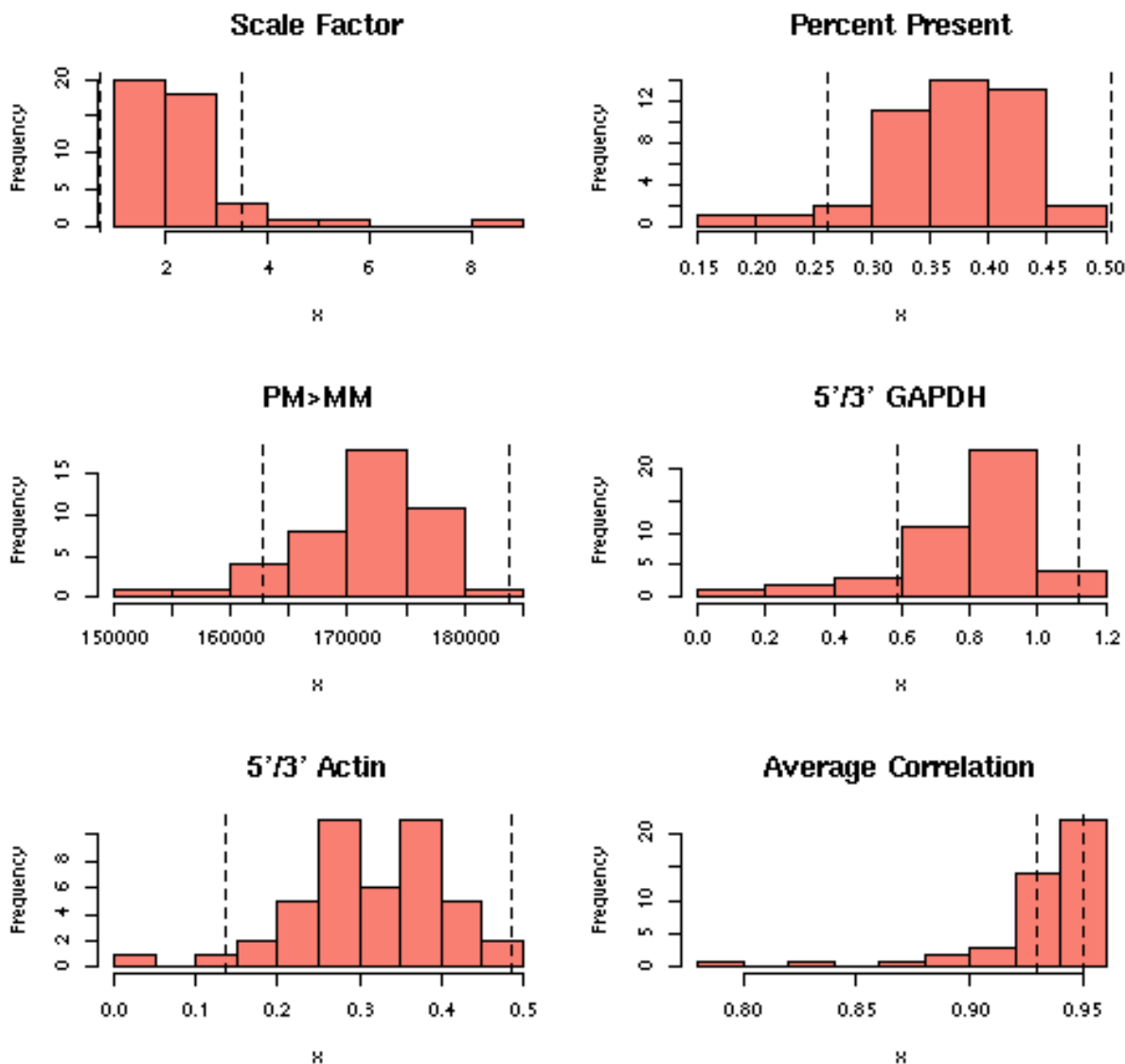
Figure 1
QC histograms. Examples of distribution thresholds used to assess outliers for an individual study.

each demographic variable is reported in a table, accompanied by a summary of the percentage of significant genes for each variable (Figure 2). Each gene found to be significant for a demographic variable links to a gene-cen-

Table 4: Patient demographic variables for Depressed patients

Depressed Patient Variables	Values
Depression Heavy Alcohol Use	Yes/No
Depression Heavy Drug Use	Yes/No
Depression Suicide	Yes/No

tric page (discussed in **Gene details page** section). Such results allow researchers to determine markers that are related to lifestyle or clinical demographical information and identify confounding variables within a disease class.

The three disease classes were analyzed to provide a list of discriminating genes (adjusted for the demographic terms that met the criteria of significance for that gene) or markers indicative of disease (Figure 3) between the control patients and each disease class (schizophrenia, bipolar, depression). In addition to table summaries (genes in table also link to their respective gene detail page), both

Demographic Analysis for Study ID 1

Gene Name	Probe ID	Fold •	P-Value	Demographic
WD repeat domain 12	218512_at	-1.300079	0.002726145	Schiz. Rate of Death
ribosomal protein L32	200674_s_at	-1.300112	0.002549543	Schiz. Rate of Death
src family associated phosphoprotein 2	204361_s_at	-1.300366	0.0026311	Hsv2 OD z-score
collagen, type I, alpha 2	202403_s_at	-1.300609	0.001424433	Schiz. Antidepressants 1
KIAA0406 gene product	212898_at	-1.301134	8.636145E-4	BP Rate of Death
manic fringe homolog (Drosophila)	204153_s_at	-1.301626	0.009516091	Schiz. Antidepressants 1
intercellular adhesion molecule 5, telencephalin	206906_at	-1.301903	0.008547785	Schiz. Anticholinergic
glyceraldehyde-3-phosphate dehydrogenase	212581_x_at	-1.301906	4.491425E-4	BP Rate of Death
GPI deacylase	220576_at	-1.303601	0.006497378	Hsv2 OD z-score
NADH dehydrogenase (ubiquinone) 1, subcomplex unknown, 2, 14.5kDa	206936_x_at	-1.304087	0.001846594	Schiz. Suicide Status
CGI-105 protein	218504_at	-1.304355	0.002839769	Schiz. Rate of Death
GCN5 general control of amino-acid synthesis 5-like 1 (yeast)	202592_at	-1.30492	0.003567047	BP Lifetime Antipsychotics
platelet-derived growth factor receptor, alpha polypeptide	203131_at	-1.305651	5.900375E-4	Schiz. Rate of Death
6-pyruvoyl-tetrahydropterin synthase/dimerization cofactor of hepatocyte nuclear factor 1 alpha (TCF1)	203557_s_at	-1.306751	0.003184587	Schiz. Rate of Death
erythrocyte membrane protein band 4.1-like 3	212681_at	-1.306797	0.00754858	Schiz. Mood Stabilizer 1/2 Valproate
phosphatidylinositol glycan, class K	209707_at	-1.307563	0.003827987	Schiz. Rate of Death
RAB11B, member RAS oncogene family	217793_at	-1.309066	0.006303166	Schiz. Antidepressants 1
ribosomal protein L22	208768_x_at	-1.309477	0.003056197	Schiz. Rate of Death

Figure 2
Demographic gene table. Table of genes determined to be significant (p < 0.01 and fold change > 1.3) with the demographic variables for an individual study.

Disease Analysis for Study ID 1

Gene Name	Probe ID	Fold •	P-Value	disease
metallothionein 1X	204326_x_at	1.490134	0.008826667	Unaffected control/Schizo.
metallothionein 2A	212185_x_at	1.415044	0.008509821	Unaffected control/Schizo.
solute carrier family 14 (urea transporter), member 1 (Kidd blood group)	205856_at	1.405315	0.001412296	Unaffected control/Schizo.
metallothionein 1E (functional)	212859_x_at	1.394182	0.007688833	Unaffected control/Schizo.
heat shock 90kDa protein 1, alpha	211969_at	-2.197723	3.636968E-4	Unaffected control/BP
poly(rC) binding protein 2	204031_s_at	-1.891557	1.620039E-4	Unaffected control/BP
calcium/calmodulin-dependent protein kinase kinase 2, beta	213812_s_at	-1.823965	4.937033E-4	Unaffected control/BP
spinocerebellar ataxia 1 (olivopontocerebellar ataxia 1, autosomal dominant, ataxin 1)	203231_s_at	-1.819936	0.001053595	Unaffected control/BP
myelin basic protein	210136_at	-1.789055	0.002000251	Unaffected control/BP
tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, epsilon polypeptide	210317_s_at	-1.787488	1.089464E-4	Unaffected control/BP
tumor differentially expressed 1	211769_x_at	-1.783736	8.196207E-4	Unaffected control/BP
likely ortholog of mouse immediate early response, erythropoietin 4	201593_s_at	-1.756236	4.611386E-4	Unaffected control/BP
WD repeat and SOCS box-containing 1	210561_s_at	-1.743396	2.464709E-4	Unaffected control/BP
tumor differentially expressed 1	221473_x_at	-1.720738	0.001397721	Unaffected control/BP

Figure 3
Disease gene table. Table of genes determined to be significant (p < 0.01 and fold change > 1.3) with the disease for an individual study.

2D clustering heatmaps (Figure 4) and principal components scatter plots (Figure 5) are provided for a visual representation of the data. Utilizing these disease markers, the most regulated pathways and GO terms were identified for each disease comparison based on a Fisher's exact test. Each pathway and GO term (from each of the three GO functional classifications separately) is ranked by p-value for each disease comparison to indicate the most regulated pathway/GO terms (Figure 6). Additionally, each pathway and GO term in the table links to a pathway/GO detail page.

Pathway/GO details page
Within this pathway/GO detail page is a comprehensive summary of the gene expression profiles for each gene that is mapped to the associated pathway or GO term within each separate disease class. A confidence interval boxplot is provided within each disease comparison inclusive of every gene mapped to that pathway or GO term queried in the study (Figure 7), along with a link to the pathway network representation provided by KEGG. Such results allow researchers to understand the most regulated biological mechanisms and cellular sites for each disease class.

Gene details page
For every probe across the 6 array platforms, primary annotations were determined such that each probe is mapped to either a gene name or EST identifier (refer to **Bioinformatics mappings** section for mapping criteria). So each gene summary page contains probe-level information for all of the 6 array platforms and 12 studies within the database. In addition to general bioinformatics annotations (e.g. biological summary, LocusLink ID, PubMed search link, and gene symbol) and pathway/GO mappings (associations with gene that link to pathway/GO-centric pages), this page contains gene expression summaries for every probe that maps to this gene across all studies (Figure 8). A cross study 'consensus' fold change was calculated for each gene and disease/demographic comparison, based on a weighted combination of the individual fold changes and standard errors for the probes that map to each gene across the platforms/studies. Weights were determined in a probeset-specific manner to account for the differing levels of precision associated with each probeset that maps to a given gene across the platforms. Confidence interval boxplots inclusive of each probe for the gene on this page are provided for the following: normalized expression across all patients, fold changes within each disease class, percent present calls for the former two comparisons, and all 41 demographic variables for the gene (Figure 9). Additionally, there is a general search engine that supports queries of gene name, symbol, pathway, GO term, and LocusLink

ID designed for direct access to any gene detail page or pathway/GO detail page.

Cross-platform analysis
To date, making comparisons across disparate gene expression platforms has been very difficult [5-8]. Chip manufacturing differences such as probe selection, processing protocols, and spot normalization algorithms contribute to variability that can distort mRNA transcript abundance measurements and introduce inconsistencies to hinder cross-platform comparisons. Some success has been demonstrated in reducing the problem to the most consistent sequence-verified gene annotations between two platforms (e.g. UniGene cluster membership) and examining correlations, ratio values, or gene calls, although sensitivity and global statistical inference of such approaches still remains a challenge [7,10-12].

The cross-platform comparisons within the SMRIDB are based on scaled representations of individual study-level analysis across studies to extract biological patterns and relationships. These cross-platform results are provided for both the gene level (Figure 10) and pathway/GO level in a study-centric (Figure 11) and gene-centric (Figure 12) visualization. For the gene-level cross-platform analysis, the fold changes and confidence intervals are calculated as described in the **Gene details page** section. For the pathway/GO-level analysis, the p-values calculated by the Fishers's exact test from each study individually for disease-related genes were scaled across studies and provided in an interactive sortable heatmap, where each cell has a clickable link to a pathway/GO details page. Additionally, this same analysis and visual representation is provided for the demographic variables (Figure 13). Such a data representation allows researchers to quickly determine the most regulated pathways or functional classifications across all platforms or for a specific demographic variable.

Utility and discussion
The user interface was constructed to enable intuitive navigating and efficient data mining. The main site contains the primary index for the database's 4 general segmented areas: Patients, Studies, Genes, and Analysis, each of which is a gateway to unique focus areas, with mutual associations between each, such as clinical information vs. genomics results and individual study content vs. cross-platform combined analyses. The Genes tab contains an open text search engine (with partial matches) to enable queries by gene, LocusLink, or pathway for any single or combined study results.

The intended users of the database include any genomics researchers facing the persistent challenges of sensitivity for biomarker discovery and cross-platform microarray
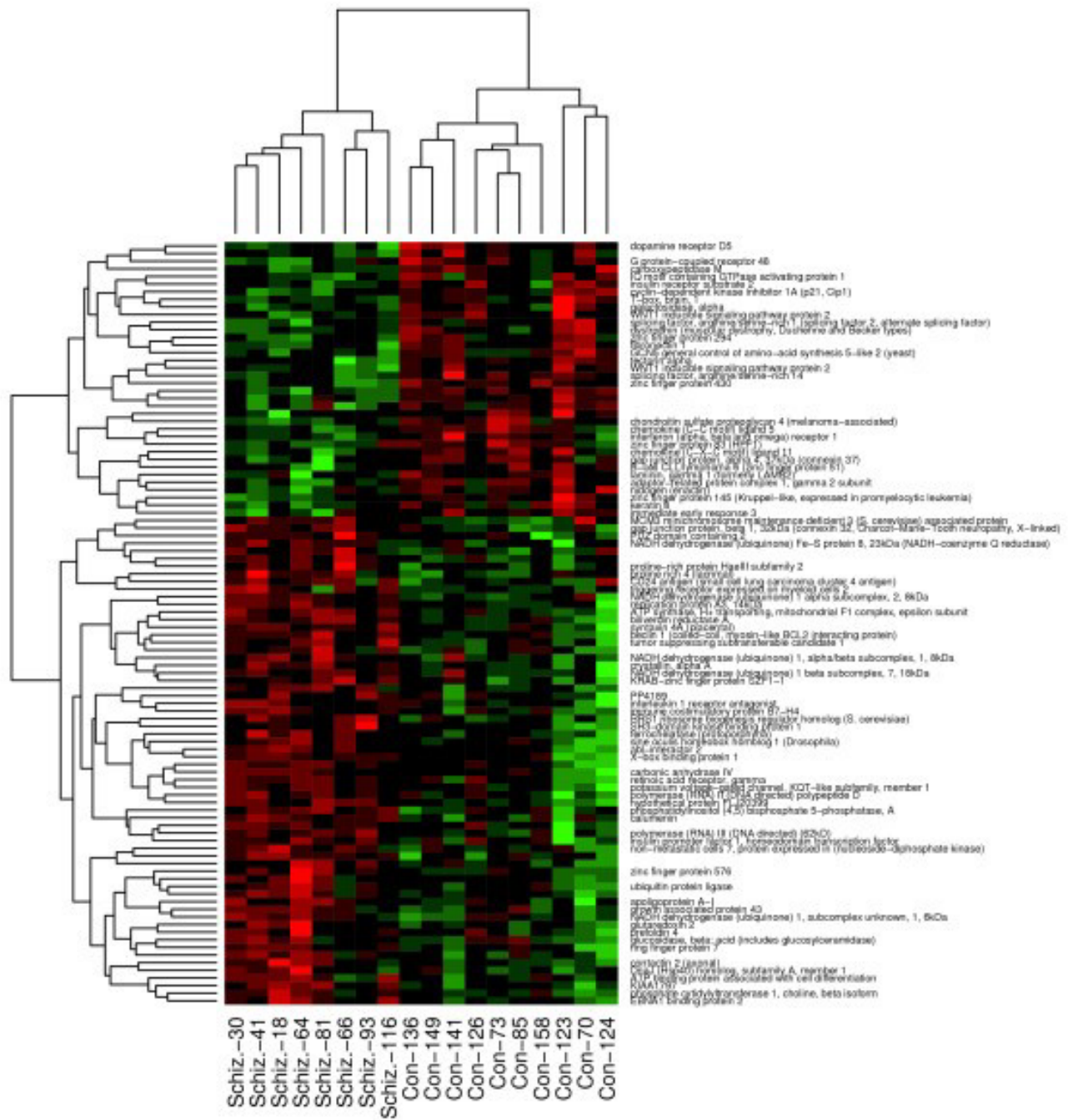
Figure 4
Study-level visuals (heatmap). Two-dimensional hierarchical clustering heatmap containing the most significant genes in schizophrenic disease for an individual study.
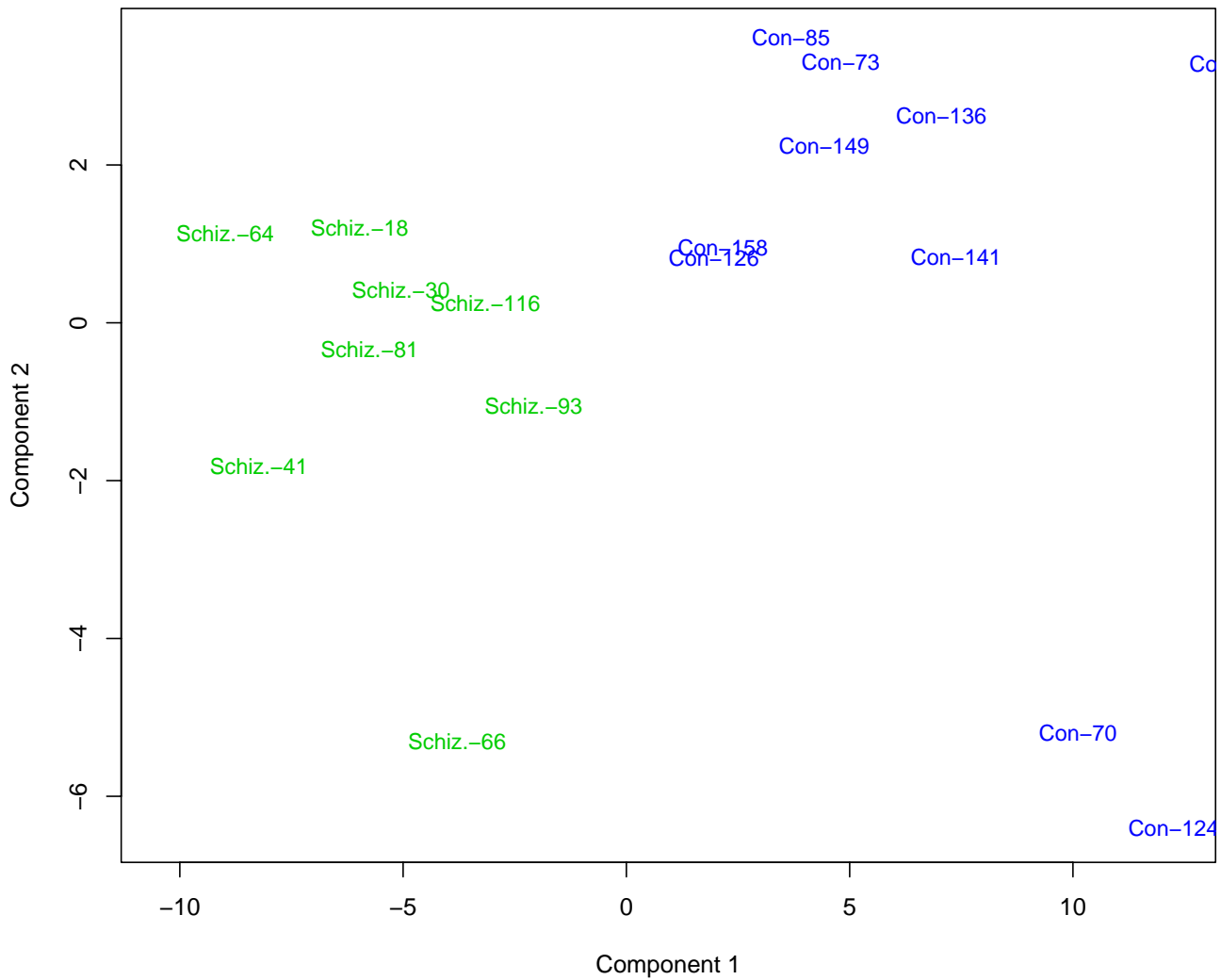
Figure 5
Study-level visuals (PCA scatter plot). Principal components plots generated with the most significant genes in schizo-phrenic disease for an individual study.

Pathway Analysis for Study ID 1

Pathway	Absolute RR Z-Score	P-Value •	Disease
Ubiquinonebiosynthesis	110.269	0.0	Unaffected control/BP
Alzheimer's disease	7.309	0.0	Unaffected control/BP
D-Arginine andD-ornithine metabolism	5.33	0.0	Unaffected control/BP
Alzheimer's disease	4.487	7.0E-6	Unaffected control/Schizo.
Dentatorubropallidoluysian atrophy (DRPLA)	4.357	1.3E-5	Unaffected control/BP
Phosphatidylinositol signaling system	4.293	1.8E-5	Unaffected control/BP
Parkinson's disease	3.808	1.4E-4	Unaffected control/BP
RNA polymerase	2.658	0.007861	Unaffected control/BP

Figure 6
Pathway table. Table of most regulated pathways for an individual study.

Fold Change (Bipolar)
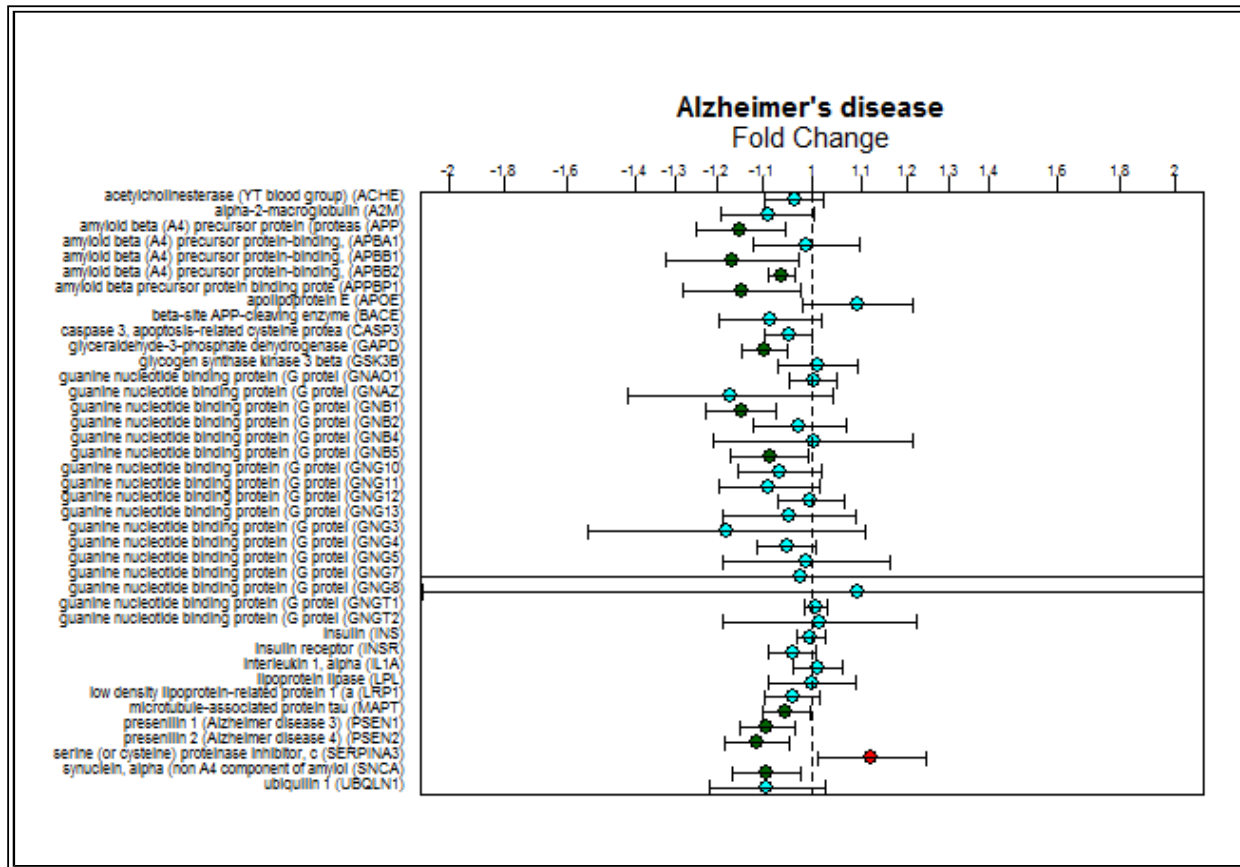
Figure 7
Fold change boxplots. Fold change (with confidence intervals) values for bipolar patients for every gene that maps to the Alzheimer's pathway.

Gene Detail for Gene ID 3858

information about the gene detail pages

Gene Symbol	RELN
Gene Name	reelin
Summary	[SUMMARY:] This gene encodes a large secreted extracellular matrix protein thought to control cell-cell interactions critical for cell positioning and neuronal migration during brain development. This protein may be involved in schizophrenia, autism, bipolar disorder, major depression and in migration defects associated with temporal lobe epilepsy. Mutations of this gene are associated with autosomal recessive lissencephaly with cerebellar hypoplasia. Two transcript variants encoding distinct isoforms have been identified for this gene. Other transcript variants have been described but their full length nature has not been determined.
Locuslink Id	5649
Pub Med Search	Search on Locuslink Id =5649

Probes

Array	Probe Id
hgu95a	37530_s_at
hgu95av2	37530_s_at
hgu133a	205923_at
hgu133plus	205923_at
codelink	NM_005045.1_PROBE1
agilent	857060
agilent	2568422
cDNA	40791

Figure 8
Gene summary page (truncated). Portion of gene summary page for the gene reelin (RELN).
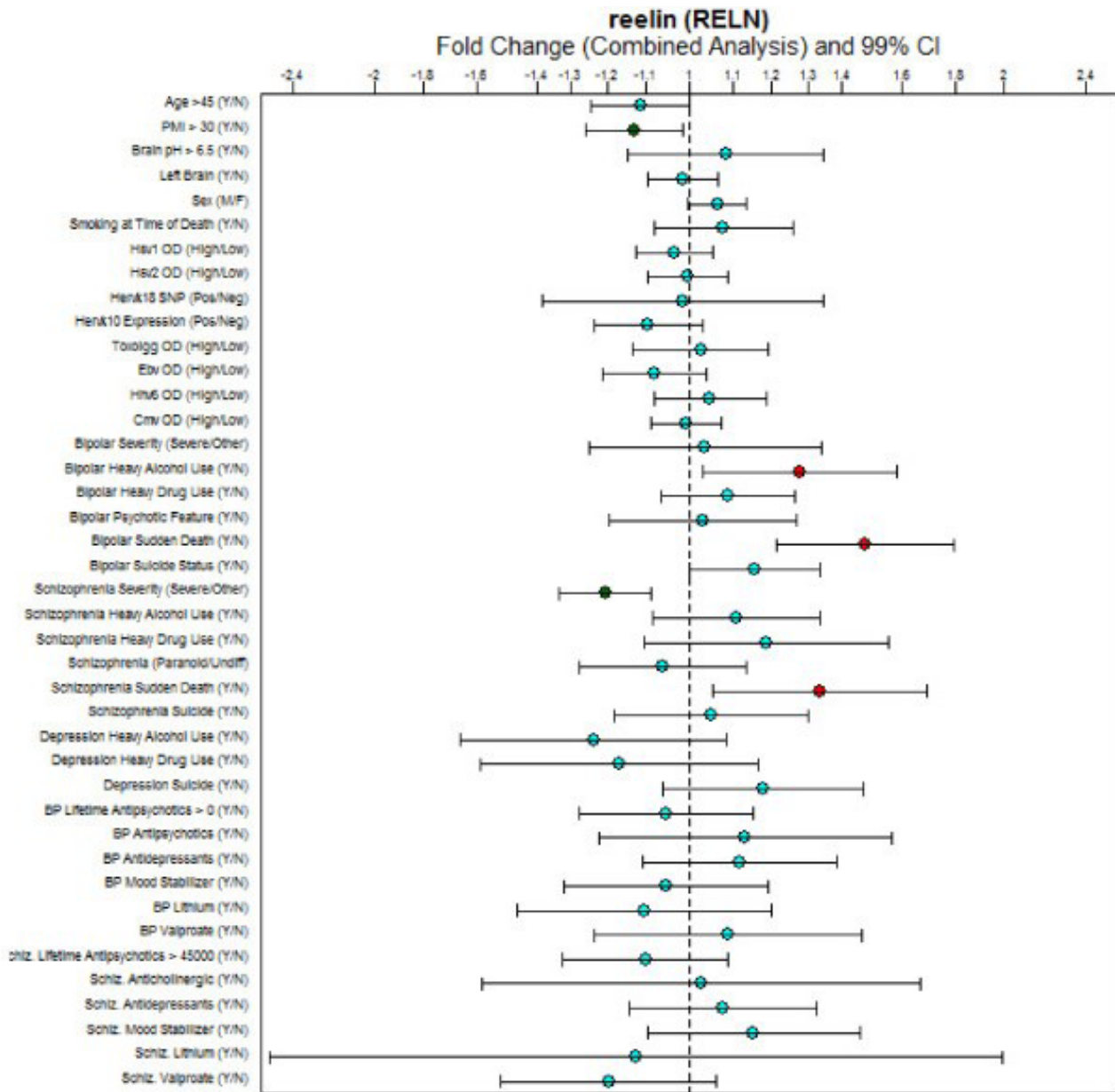
Figure 9
Fold change boxplots. Fold change (with 99% confidence intervals) for the gene reelin across all 41 demographic variables.

The Stanley Medical Research Institute
Online Genomics Database

| Patients | Studies | Genes | Analysis | About Database |

Cross Study Analysis of Disease (Genes)

Select Disease: Depression

Select Analysis Table(s): Significant probes

submit

Significant probes for Depression (up-regulation)

Total	Depression Up Regulation										
Probes	Percent of Probes with p<.05										
	0-5%	6-10%	11-15%	16-20%	21-25%	26-30%	31-35%	36-40%	41-45%	46-50%	50-100%
1-3	4050	0	0	0	0	0	67	0	0	188	317
4-5	3020	0	0	534	230	0	0	47	0	16	6
6-10	2607	102	648	499	84	72	98	14	16	13	0
11-15	357	239	23	93	21	13	5	1	1	0	0
16-20	46	44	17	10	3	1	0	0	0	0	0
21-25	13	6	5	3	1	2	0	0	0	0	0
26-30	5	4	1	1	0	1	0	0	0	0	0
31-40	2	2	0	2	0	0	1	0	0	0	0
41-50	0	0	0	0	0	1	0	0	0	0	0
>50	0	0	0	0	0	0	0	0	0	0	0

Figure 10
Summary statistic table. Gene-level summary table of significant probes across all studies for depression.
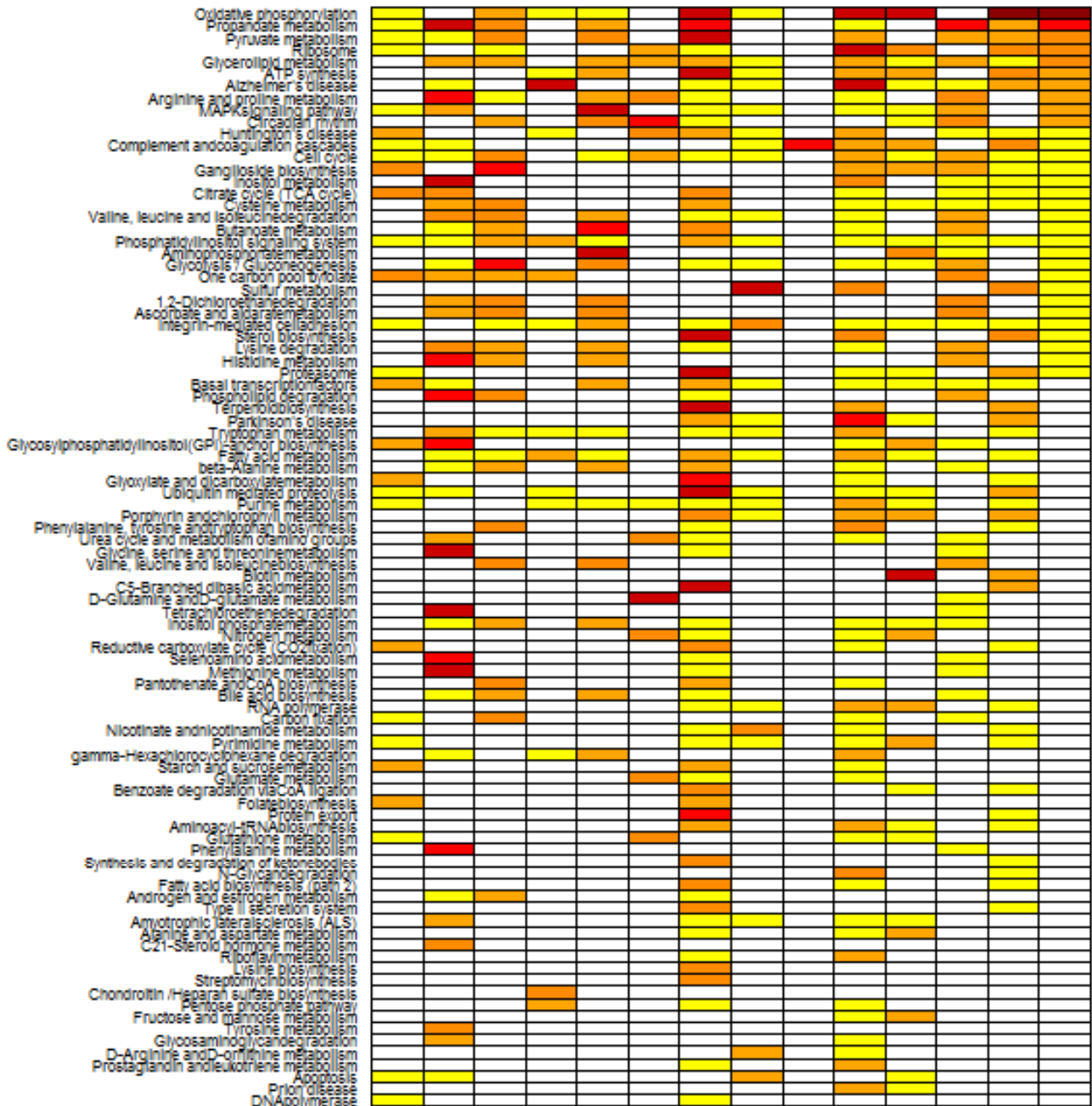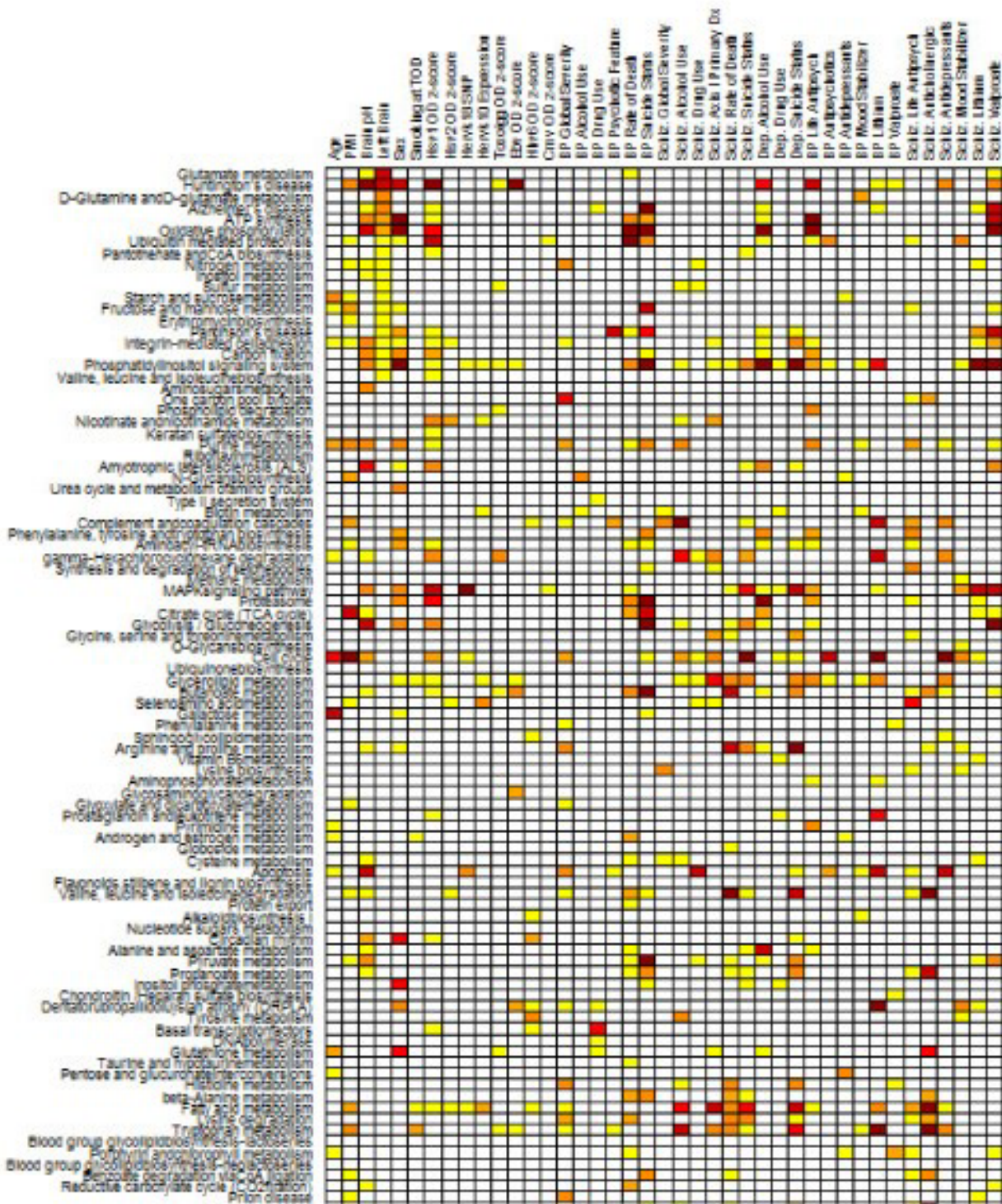
Figure 11
Pathway clickable heatmap. Study-centric clickable heatmap of top regulated pathways in schizophrenia. Each column can be sorted by a particular study or the three last summary columns. Study 12 was omitted from this visual.
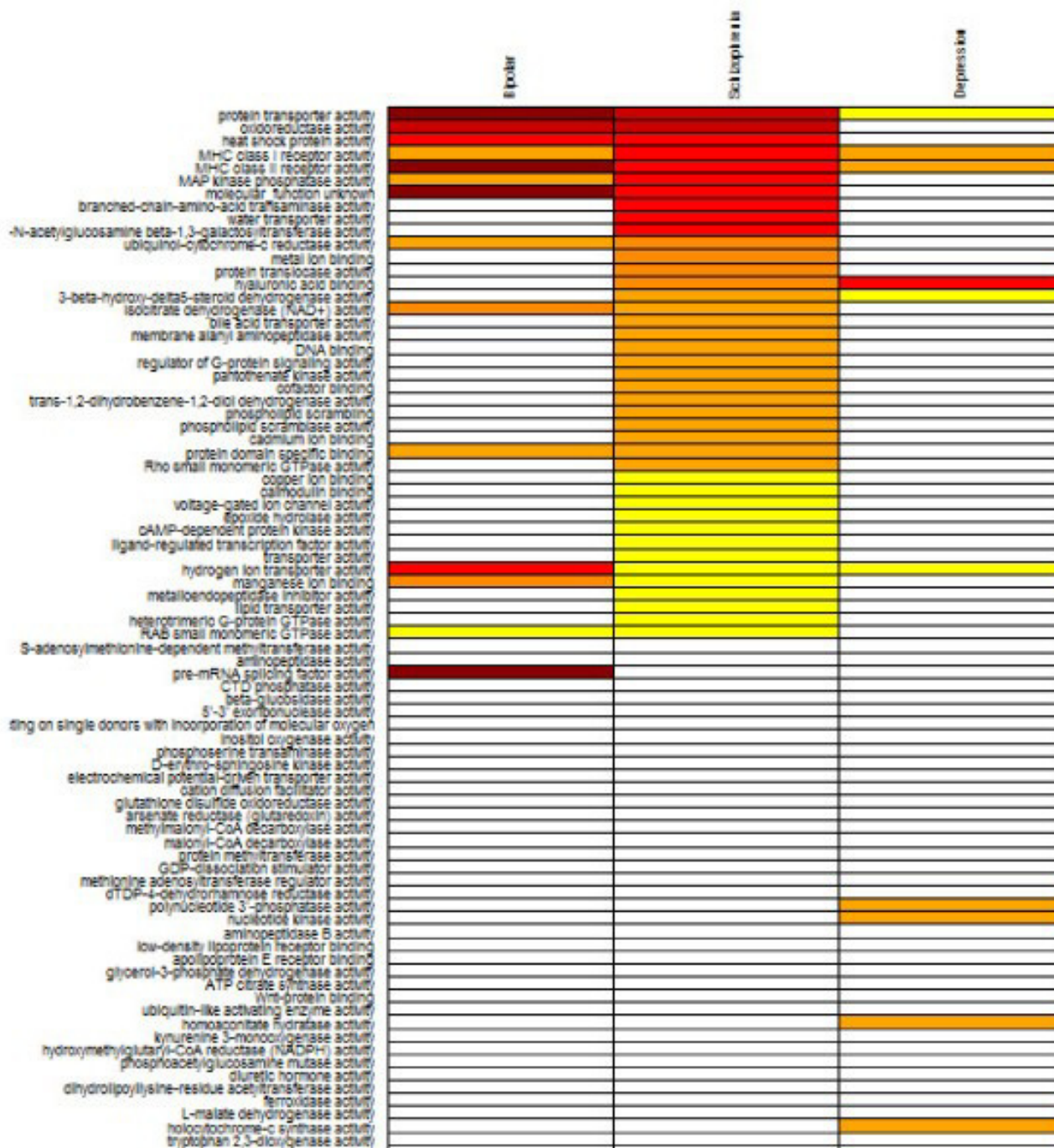
Figure 12
GO term clickable heatmap. Gene-centric clickable heatmap of top regulated GO terms (molecular function) in schizo-
phrenia. Each column can be sorted by a disease.

Figure 13
Pathway/demographic clickable heatmap. Demographic variable clickable heatmap of top regulated pathways. Each column can be sorted by a demographic variable.

comparisons. However, the content within the SMRIDB is primarily designed for biologists, clinical researchers, bio-informaticians, and scientist in the field of brain disease.

The size and scope of the SMRIDB makes it a unique contribution to genomics-based brain disease research. With combined gene expression profile summaries across 12 studies and 6 platforms, there is greater confidence in scientific findings such as biomarkers for disease, biological functional roles, and regulated pathways, as compared to results obtained from any one individual study.

Conclusion

The SMRIDB is a comprehensive data mining tool to enable researchers to elucidate the biological mechanisms of bipolar disorder, schizophrenia, and depression. A diverse patient population combine with data generated across six microarray platforms and 12 studies to provide robust results to enhance the understanding of brain disease.

Availability and requirements

The SMRIDB can be accessed at https://www.stanleygenomics.org. All users must register (name and email address) to obtain a username and password.

Authors' contributions

BWH and ME conducted the data analysis and were involved in drafting the manuscript. SR developed the web services and database backend. BB collected and catalogued the clinical information and samples. All authors read and approved the final manuscript.

References

1. Pavlidis P, Noble WS: **Analysis of strain and regional variation in gene expression in mouse brain.** *Genome Biology* 2001, **2(10):**RESEARCH0042.
2. Cho H, Lee JK: **Bayesian hierarchical error model for analysis of gene expression data.** *Bioinformatics* 2001, **20(13):**2016-25.
3. Iacobas DA, Urban M, Iacobas S, Spray DC: **Control and variability of gene expression in mouse brain and in a neuroblastoma cell line.** *Rom J Physiol* 2003, **39–40:**2002-71.
4. Jurata LW, Bukhman YV, Charles V, Capriglione F, Bullard J, Lemire AL, Mohammed A, Pham Q, Laeng P, Brockman JA, Altar CA: **Comparison of microarray-based mRMA profiling technologies for identification of psychiatric disease and drug signatures.** *J Neurosci Methods* 2004, **138(1–2):**173-88.
5. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies.** *Bioinformatics* 2002, **18:**405-412.
6. Li J, Pankratz M, Johnson JA: **Differential gene expression patterns revealed by oligonucleotide versus long cDNA arrays.** *Toxicological Sciences* 2002, **69:**383-390.
7. Mah N, Thelin A, Lu T, Nikolaus S, Kuhbacher T, Gurbuz Y, Eickhoff H, Kloppel G, Lehrach H, Mellgard B, Costello CM, Schreiber S: **A comparison of oligonucleotide and cDNA-based microarray systems.** *Physiol Genomics* 2004, **16:**361-370.
8. Tan P, Downey TJ, Spitznagel EL Jr, Xu P, Fu D, Dimitrov DS, Lempicki RA, Raaka BM, Cam MC: **Evaluation of gene expression measurements from commercial microarray platforms.** *Nucleic Acids Research* 2003, **31:**5676-5684.
9. **DAVID (Database for Annotation and Visualization and Integrated Discovery)** [http://apps1.niaid.nih.gov/david]
10. Lee JK, Bussey KJ, Gwadry FG, Reinhold W, Riddick G, Pelletier SL, Nishizuka S, Szakacs G, Annereau JP, Shankavaram U, Lababidi S, Smith LH, Gottesman MM, Weinstein JN: **Comparing cDNA and oligonucleotide array data: concordance of gene expression across platforms for the NCI-60 cancer cells.** *Genome Biology* 2003, **4:**R82.
11. Mecham B, Klus GT, Strovel J, Augustus M, Byrne D, Bozso P, Wetmore DZ, Mariani TJ, Kohane IS, Szallasi Z: **Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements.** *Nucleic Acids Research* 2004, **32:**9.
12. Nimgaonkar A, Sanoudou D, Butte AJ, Haslett JN, Kunkel LM, Beggs AH, Kohane IS: **Reproducibility of gene expression across generations of Affymetrix Microarrays.** *BMC Bioinformatics* 2003, **4:**27.