Title page

Title: DLPFC Transcriptome Defines Two Molecular Subtypes of Schizophrenia

Authors: C. Harker Rhodes, MD, PhD
Elijah F. W. Bowen, MS
Caitlyn I. Lee
Jack L. Burgess
Richard Granger, PhD

Contact Information:
C. Harker Rhodes, MD, PhD
144 Sunset Rock Rd
Lebanon, NH 03766

Tel: 603-443-3360
FAX:
E-mail: CHarkerRhodes@gmail.com

Molecular Subtypes of Schizophrenia

Abstract

The Clinical Brain Disorders Branch of the Intramural Research Program at the National Institutes of Health assembled a large collection of frozen post-mortem human brains from individuals diagnosed with schizophrenia and other psychiatric diagnoses, as well as matched control individuals. Illumina HumanHT-12 v4 expression array data was collected from dorsolateral prefrontal cortex (DLPFC) and the data deposited at dbGaP (Study ID: phs000979). We report an analysis of the data from the 189 adult schizophrenics and 206 adult controls in that cohort.

Transcripts from 633 genes are differentially expressed in the DLPFC of schizophrenics as compared to the controls at levels of statistical significance which survive Bonferroni correction. Seventeen of those genes are differentially expressed at a level of statistical significance less than 10^{-8} after Bonferroni correction.

Weighted Gene Co-expression Network Analysis (WGCNA) of the schizophrenic subjects using differentially expressed gene module eigengenes divides them into two groups, "type 1" and "type 2." There are 2,635 genes differentially expressed in the DLPFC of type 2 schizophrenics at a level of statistical significance which survives Bonferroni correction. Of them 221 have P-values less than 10^{-15} after Bonferroni correction. In the type 1 schizophrenics however, only 6 genes are differentially expressed in the DLPFC at a level of statistical significance which survives Bonferroni correction. This striking difference in their DLPFC transcriptomes emphasizes the fundamental biologic difference between these two groups of patients.

Introduction

The Clinical Brain Disorders Branch of the Intramural Research Program at the National Institutes of Health assembled a large collection of frozen post-mortem human brains from individuals diagnosed with schizophrenia, or other psychiatric diagnoses as well as matched control individuals. Illumina HumanHT-12 v4 expression array data was collected from dorsolateral prefrontal cortex (DLPFC) of those brains, and the data deposited at dbGaP (Study ID: phs000979). We report an analysis of the data from the 189 adult schizophrenics and 206 adult controls in that cohort.

Robust linear mixed effect regression including as a random effect "batch" and as fixed effects RIN, sex, age at death, and race identified 633 genes which are differentially expressed in the DLPFC of schizophrenics as compared to the controls at levels of statistical significance which survive Bonferroni correction.

Recognizing the clinical heterogeneity among schizophrenics, a successful attempt was made to divide the patients in this cohort into biologically meaningful schizophrenia subtypes based on their DLPFC transcriptomes. Two subtypes were identified. The limited clinical information available in this Medical Examiner based sample of convenience does not allow a correlation of the subtypes with clinical phenotype, but they have strikingly different sets of differentially expressed genes supporting the conclusion that there are fundamental biologic differences between these two groups of patients.

Methods

Over a period of many years the Clinical Brain Disorders Branch of the NIMH intramural program assembled a large collection of frozen human brains from Medical Examiner patients and conducted detailed post-mortem psychiatric reviews to establish their diagnoses. The human tissue collection, and processing protocols have been previously described [1]. Poly-A RNA was prepared from dorsolateral prefrontal cortex and hippocampus. Illumina HumanHT-12 v4 expression array data was generated according to the manufacturer's protocols, and that data was made publicly available at dbGaP (Study ID: phs000979).

With the appropriate IRB approval and dbGaP authorization that data was downloaded to a secure high-performance cluster (two Intel Xeon x5660 2.8GHz CPUs with 6 cores each, with hyperthreading yielding 24 virtual cores, with 72GB of DDR3-1333 RAM). Data analysis done using Rstudio and protocols for reproducible research. HTML-formatted R markdown files which contain the computer code for the entire analysis are included as supplemental material.

Data pre-processing

The decrypted dbGaP data includes a directory “PhenotypeFiles” containing multiple files among which is distributed the subject annotation data from this study. That information was re-formatted as a data.frame whose rows are the study subjects and whose columns contain information about the subject. (Supplemental R code #1)

The decrypted dbGaP data also includes a directory “ExpressionFiles” which contains the expression array data formatted as idat files. Using the Bioconductor package {beadarray}, that data was quantile normalized, log2-transformed, and formatted as a matrix whose rows are the study subjects and whose columns are the expression array data for each probe. At the same time the function

calculateDetection(){beadarray} which implements Illumina's method for calculating the detection scores was used to create a second matrix with those scores. (Supplemental R code #2)

The expression array dataset initially contained data from 48,107 Illumina probes. It was filtered to remove data from:

1> 2414 probes for which the log2-transformed data was "NA" or "Inf" for any of the subjects.

2> 33,158 or 73% of the probes where, based on the Illumina detection score, the level of expression was statistically significant in fewer than 841 of the 849 subjects.

3> 652 probes where the probe sequence contains a common SNP [2].

This left data from a total of 11,883 probes available for analysis. (Supplemental R code #3)

The dbGaP dataset includes expression array data from 849 individuals with a variety of psychiatric diagnoses. After restricting it to schizophrenics and controls only it contains 549 individuals and after the elimination of individuals less than 25 years old and individuals whose age is not specified, the cohort consists of 202 schizophrenics and 347 controls. (Supplemental code #4)

Identification of differentially expressed transcripts and clustering of transcripts/subjects

Illumina array probes which detect differentially expressed were identified using the robust linear mixed effect regression algorithm rlmer{robustlmm} [3] including as fixed effect covariates age, sex, race, and RIN and as a random effect covariate the expression array batch (Supplemental code #5). Ingenuity Pathway Analysis (QIAGEN; Redwood City 1700 Seaport Blvd #3, Redwood City, CA 94063) was used to identify the pathways containing the differentially expressed genes.

The differentially expressed transcripts were clustered based on transcript co-expression in the schizophrenic subjects using weighted correlation network analysis (WGCNA) [4] and three gene modules were identified (Supplemental code #6).

WGCNA was then used a second time, this time to cluster the schizophrenic patients based on the "eigengenes" for those gene modules and the relationships between the schizophrenic patients was visualized by representing the data as a graph whose nodes are patients and edges were defined as being present when the WGCNA-calculated topological overlap exceeded a threshold (Supplemental code #7).

Finally, the robust linear mixed effect regression was used a second time, this time to identify the genes differentially expressed in the DLPFC of patients with each of the schizophrenia subtypes (Supplemental code #8).

Results

Cohort demographics

This cohort is a convenience sample based on Medical Examiner cases for whom the next of kin consented to the use of post-mortem tissue for this purpose. It is, therefore, not necessarily representative of the general population and this study limitation needs to be kept in mind when interpreting these results.

As might be expected in a Medical Examiner cohort where the control subjects include accidental death and homicide victims, men are over-represented in the controls. That imbalance is much more prominent in the Caucasian than African American sub-cohorts (table 1). The cohort is, however, reasonably well balanced in terms of both ethnicity (table 2) and age (figure 1).

Table 1: Distribution of schizophrenics and controls by gender

	All subjects		Caucasians		African Americans	
	Controls	Sz	Controls	Sz	Controls	Sz
Female	61	72	16	40	43	31
Male	145	117	76	59	58	54

Table 2: Distribution of schizophrenics and controls by ethnicity

	Control	Sz
African American	101	85
Caucasian	92	99
Other	13	5

Figure 1: Distribution of schizophrenics and controls by age.

Differentially expressed genes in schizophrenics as compared to controls

The expression array data included data from 11,883 probes after censoring data from probes which did not detect mRNA in the DLPFC at a statistically significant level or which contained common polymorphisms in the probe sequence.

Robust linear mixed effects regression including RIN, gender, ethnicity, and age as fixed effects and "batch" as a random effect identified 694 array probes which detected transcripts from 633 genes which were differentially expressed in the DLPFC of the schizophrenics at a level of statistical significance which survived Bonferroni-correction. For two of those genes, SYNDIG1 (aka TMEM90B, a gene involved in the maturation of excitatory synapses) and PSMB6 (a proteasomal subunit gene), the Bonferroni-corrected P-value was less than 10^{-15}.

Ingenuity pathway analysis identified proteasomal and mitochondrial pathway genes as being overrepresented in the list of differentially expressed genes. The two genes with the largest positive effect size (increased expression in schizophrenics) are MT1X and BAG3, both genes previously identified as being overexpressed in the DLPFC of schizophrenics [5] The gene with the largest negative effect size (decreased expression in schizophrenics) is NPY, a useful marker for specific subclasses of cortical GABAergic interneurons [6, 7]. The complete list of the 633 differentially expressed genes is included as on-line supplemental information.

Three gene modules were identified when weighted gene co-expression network analysis (WGCNA) was used to cluster the differentially expressed transcripts based on their co-expression in the DLPFC of the schizophrenics. The biologic significance of those gene modules is unclear, but their corresponding eigengenes proved to be very useful for the further classification of the schizophrenic patients (below).

WGCNA clustering of schizophrenics

After using WGCNA to cluster the differentially expressed genes and identify three differentially expressed “eigengenes”, WGCNA was used a second time, this time to cluster the schizophrenic patients based on those eigengenes. This analysis divides the schizophrenics into two groups, "type 1" and "type 2" (figure 2).

The differential gene expression in the DLPFC is strikingly different in these two groups of patients. There are 2,635 genes differentially expressed in the DLPFC of type 2 schizophrenics at a level of statistical significance which survives Bonferroni correction. Of them 221 have P-values less than 10^{-15} after Bonferroni correction. In the type 1 schizophrenics however, only 6 genes are differentially expressed in the DLPFC at a level of statistical significance which survives Bonferroni correction. This difference in their DLPFC transcriptomes suggests that there is a fundamental biologic difference between these two groups of patients.

The complete list of genes differentially expressed in type 1 and type 2 schizophrenics is included in the supplemental materials, but the differences between these groups of patients is perhaps best illustrated by specific examples of the differentially expressed genes (figure 3).

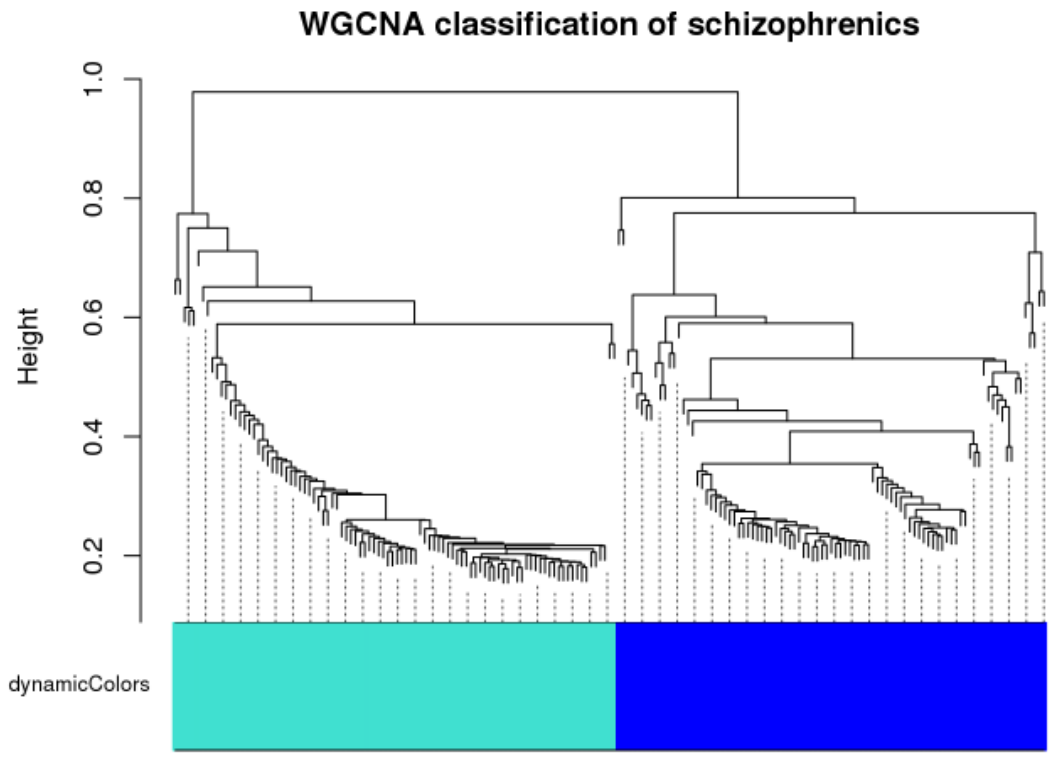

Figure 2: WGCNA clustering of the schizophrenics based on the eigengenes from the three gene modules present in the set of differentially expressed genes.

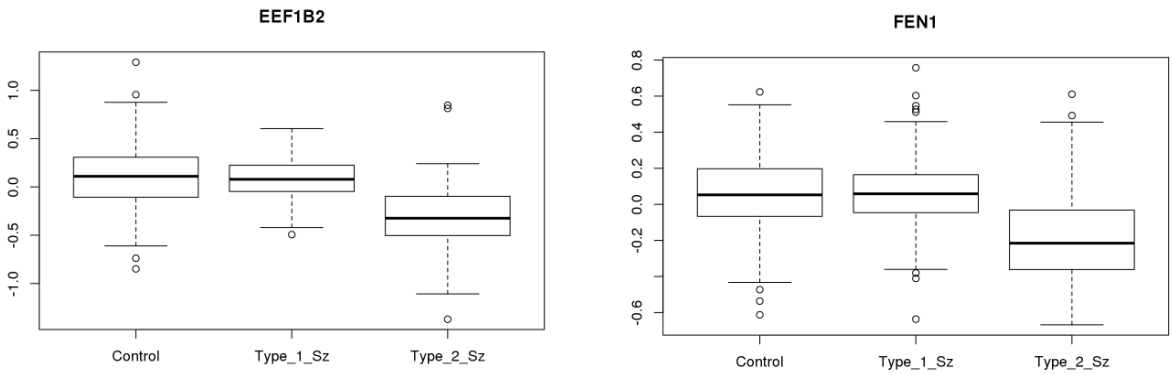

Figure 3: Two examples of the 2,635 genes differentially expressed in the DLPFC of type 2 schizophrenics at a level of statistical significance which survives Bonferroni correction.
Left: Expression of EEF1B2 (eukaryotic translation elongation factor 1 beta 2) in DLPFC.
Right: Expression of FEN1 (flap structure-specific endonuclease 1) in DLPFC.
The data is expressed as the residuals after correction for age, gender, ethnicity, RIN, and batch. In both cases the difference between the type 1 and type 2 schizophrenics is significant at a level of $P << 10^{-15}$ after Bonferroni correction for the number of probes on the array.

Biologic validation of the distinction between type 1 and type 2 schizophrenia

About half of all schizophrenics, schizoaffective patients, and bipolar patients have what has been described as a "low GABA marker" molecular phenotype based on the expression of GABA neuron markers. Specifically, this subset of schizophrenic patients have reduced expression of GAD67, parvalbumin, somatostatin, and the transcription factor LHX6 in their DLPFC [8, 9].

In this expression array dataset, the Illumina probes for somatostatin and parvalbumin do not detect transcripts at a level significantly different from zero. Neither GAD67 (GAD1) nor the LHX6 transcripts are differentially expressed at a level of statistical significance which survives Bonferroni correction for the number of probes on the Illumina array which detect transcripts expressed in the cortex. However, since we are now testing the specific hypotheses that those transcripts are down regulated in schizophrenics that Bonferroni correction is no longer appropriate. Both of those genes are down regulated in the type 2 schizophrenics, but not the type 1 schizophrenics. (The P-value without Bonferroni correction for differential expression of GAD67 in type 2 schizophrenics is 1×10^{-7}; for LHX6 the P-value is 1×10^{-6}.

Since these markers for the "low GABA marker" molecular phenotype played no role in the distinction of type 1 vs type 2 schizophrenics described here, the differential presence of this phenotype in type 1 and type 2 schizophrenics provides a biologic validation of that distinction.

Discussion

This reanalysis of a publicly available expression array dataset identifies 633 genes which are differentially expressed in the dorsolateral prefrontal cortex of schizophrenics as compared to controls at a level of statistical significance which survives Bonferroni correction. More importantly, it demonstrates that schizophrenics can be divided into two molecularly distinct subgroups based on their DLPFC transcriptomes. The “type 1” schizophrenics have a DLPFC transcriptome very similar to that of controls while the “type 2” schizophrenics have a strikingly different DLPFC transcriptome with 2,635 gene differentially expressed as compared to the controls. There are two important implications of these findings:

1> Studies of gene expression in schizophrenia need to take into account this molecularly defined patient heterogeneity and analyze the data from type 1 and type 2 schizophrenics separately.

2> These results suggest the testable hypothesis that, although the DLPFC is clearly abnormal in the type 2 schizophrenics, for the type 1 schizophrenics the physiologically significant pathology may be elsewhere, perhaps in the superior temporal or cingulate gyri. Tissue from both of those gyri from the individual patients included in this study is available from the Human Brain Collection Core of the NIMH intramural program.

An important strength of this study is the reliance on robust statistics. Least squares based algorithms are more efficient than the corresponding robust methods IF the data is normally distributed, but they are also exquisitely sensitive to outliers and often give misleading results when the data is from a mixed normal distribution. For a discussion of “regression diagnostics” (the statistical techniques to detect and control for these issues with least squares based algorithms) and robust statistical methods see chapter 6 of Fox and Weisberg and the online appendix “Robust Regression” to that textbook [10] or the documentation for the R package “robustlmm” [3]

The list of genes differentially expressed in the DLPFC of schizophrenics includes many potential "druggable targets". Of course, until this expression array data has been validated by qPCR or some similar method, the results from any particular array probe need to be interpreted with caution. However, the availability from the Human Brain Collection Core of the NIMH intramural program of tissue from the DLPFC of the specific individuals included in this study should facilitate those experiments.

A limitation of this work is that it is based on a single cohort and these results should be regarded as preliminary until they have been replicated in other cohorts. When doing that it will be important to keep in mind several ways in which other studies may not be strictly comparable to this one. First, because this was an array-based study, this analysis is done at the exon level. Many genes expressed in the brain have multiple transcripts which are differentially regulated; for that reason a gene-level analysis of RNAseq is not necessarily expected to yield results comparable to these. Second, this study was done using poly-A RNA. Many non-coding transcripts contain some of the same exons as the processed mRNA; for this reason neither gene- nor exon-level analysis of "Ribo-Zero RNA" is necessarily expected to give results comparable to those of this poly-A RNA based experiment.

Supplemental Information

R markdown files containing the computer code for the analyses described in this manuscript are available as on-line supplemental files.

The on-line supplemental files "Genes differentially expressed in schizophrenic DLPFC" and "Genes differentially expressed in type 2 schizophrenic DLPFC" are Excel files containing the list of those genes.

Literature cited

1. Lipska, B.K., et al., *Critical factors in gene expression in postmortem human brain: Focus on studies in schizophrenia.* Biol Psychiatry, 2006. **60**(6): p. 650-8.
2. Ramasamy, A., et al., *Resolving the polymorphism-in-probe problem is critical for correct interpretation of expression QTL studies.* Nucleic Acids Res, 2013. **41**(7): p. e88.
3. Koller, M., *{robustlmm}: An {R} Package for Robust Estimation ofLinear Mixed-Effects Models.* Journal of Statistical Software, 2016. **75**(6): p. 1-24.
4. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**: p. 559.
5. Perez-Santiago, J., et al., *A combined analysis of microarray gene expression studies of the human prefrontal cortex identifies genes implicated in schizophrenia.* J Psychiatr Res, 2012. **46**(11): p. 1464-74.
6. Karagiannis, A., et al., *Classification of NPY-expressing neocortical interneurons.* J Neurosci, 2009. **29**(11): p. 3642-59.
7. Kubota, Y., *Untangling GABAergic wiring in the cortical microcircuit.* Curr Opin Neurobiol, 2014. **26**: p. 7-14.
8. Volk, D.W., et al., *Deficits in transcriptional regulators of cortical parvalbumin neurons in schizophrenia.* Am J Psychiatry, 2012. **169**(10): p. 1082-91.
9. Volk, D.W., et al., *Cortical GABA markers identify a molecular subtype of psychotic and bipolar disorders.* Psychol Med, 2016. **46**(12): p. 2501-12.
10. Fox, J.H.S.W., *An R Companion to Applied Regression*. 2010: Sage Publications.