RESEARCH ARTICLE

A classifier driven approach to find biomarkers for affective disorders from transcription profiles in blood

Wiktor Mazin[1,8*], Joseph A. Tamm[2], Irina A. Antonijevic[2], Aicha Abdourahman[2], Munish Das[2], Roman Artymyshyn[2], Birgitte Søgaard[3], Mary Walker[2], Danka Savic[4], Gordana Matic[5], Svetozar Damjanović[6], Ulrik Gether[7], Thomas Werge[8], Lars V. Kessing[9], Henrik Ullum[10], Eva Haastrup[10], Eric Vermetten[11], Paul Markovitz[12], Erik Mosekilde[1] and Christophe P.G. Gerald[2]

[1] Department of Physics, Technical University of Denmark, Lyngby 2800, Denmark

[2] Neuroinflammation Unit, Lundbeck Research USA, Paramus, NJ 07652, USA

[3] Clinical & Quantitative Pharmacology, H. Lundbeck A/S, Valby 2500, Denmark

[4] Vinča Institute of Nuclear Sciences, Laboratory for Theoretical and Condensed Matter Physics, University of Belgrade, Belgrade 11001, Serbia

[5] Institute for Biological Research, Siniša Stanković, Department of Biochemistry, University of Belgrade, Belgrade 11060, Serbia

[6] School of Medicine, Clinical Center of Serbia, Institute of Endocrinology, Diabetes and Metabolic Diseases, University of Belgrade, Belgrade 11000, Serbia

[7] Lundbeck Foundation Center for Biomembranes in Nanomedicine (CBN), Department of Neuroscience and Pharmacology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen N 2200, Denmark

[8] Research Institute for Biological Psychiatry, Mental Health Centre Sct. Hans, Copenhagen University Hospitals, Roskilde 4000, Denmark

[9] Psychiatric Centre Copenhagen, University Hospital of Denmark, Copenhagen 2100, Denmark

[10] Department of Clinical Immunology, the Blood Bank, University Hospital of Copenhagen, Copenhagen 2100, Denmark

[11] Leiden University Medical Center, Leiden, The Netherlands, and Arq Psychotrauma Expert Group, Diemen, The Netherlands

[12] Mood and Anxiety Research, Inc., Ventura, CA 93103, USA

Abstract: Gene expression profiles in blood are increasingly being used to identify biomarkers for different affective disorders. We have selected a set of 29 genes to generate expression profiles for healthy control subjects as well as for patients diagnosed with acute post-traumatic stress disorder (PTSD) and with borderline personality disorder (BPD). Measurements were performed by quantitative polymerase chain reaction (qPCR). Using the actual data in an anonymous form we constructed a series of artificial data sets with known gene expression profiles. These sets were used to test 14 classification algorithms and feature selection methods for their ability to identify the correct expression patterns. Application of the three most effective algorithms to the actual expression data showed that control subjects can be distinguished from BPD patients based on differential expression levels of the gene transcripts Gi2, GR and MAPK14, targets that may have links to stress related diseases. Controls can also be distinguished from acute PTSD patients by differential expression levels of the transcripts for ERK2 and RGS2 that are known to be associated with mood disorders and social anxiety. We conclude that it is possible to identify informative transcription profiles in blood samples from individuals with affective disorders.

Keywords: feature selection, mental disorders, gene expressions, gene panel

*Correspondence to: Wiktor Mazin, Research Institute for Biological Psychiatry, Mental Health Centre Sct. Hans, Copenhagen University Hospitals, Roskilde 4000, Denmark; Email: wiktor.mazin@regionh.dk

Received: October 20, 2015; **Accepted:** December 13, 2015; **Published Online:** March 18, 2016

Citation: Mazin W, Tamm J A, Antonijevic I A, *et al*. 2016, A classifier driven approach to find biomarkers for affective disorders from transcription profiles in blood. *Advances in Precision Medicine*, vol.1(1): 48–65. http://dx.doi.org/10.18063/APM.2016.01.003.

Introduction

The molecular basis of affective disorders is still poorly understood. Psychiatric disorders such as major depression or post-traumatic stress disorder are heterogeneous disorders that are considered to arise from a complex interplay of several genes and environmental factors. Due to the heterogeneous and polygenic nature of these disorders, they are also difficult to treat effectively. In the case of depression, 20% of the affected individuals show full remission with the current antidepressants. This might be related to the fact that research in the field of psychiatry historically has focused mainly on drugs targeting monoamine receptors and transporters, in various combinations, leading to only slightly different profiles[1].

On this background, there is clearly a need to better understand the biological basis of complex psychiatric disorders and to identify biomarkers that go beyond monoamine transporters and receptors. This could pave the way for the development of new drugs and, at the same time, make it possible to predict which of the current drugs to prescribe to a given patient. Transcriptional biomarkers in blood may help us achieve these goals, and during the last few years a variety of such biomarkers have been suggested as alternatives to brain markers[1,2]. This is the case, for instance, for post-traumatic stress disorder[3–6], bipolar disorder[7–10], and major depression[11–14]. Use of markers in the blood is a standardized approach and, as only a blood sample is needed, the approach is fast, inexpensive and practically non-invasive. However, data analysis to identify possible biomarkers will not be straightforward as complex psychiatric diseases are likely to involve a number of molecular changes. Hence, we believe it will require a multivariate approach to identify useful patterns in the data. Given that many different kinds of multivariate algorithms are available, a major challenge in biomarker research will be the identification of appropriate methods of analysis.

To address this question, we measured gene expression of a focused panel of targets in whole blood samples from control subjects and patients who were diagnosed with either borderline personality disorder or post-traumatic stress disorder. Next, based on the real data, we constructed simulated data sets with known differences between the simulated controls and simulated patients. Testing of 14 different classification algorithms and feature selection methods with the simulated data provided an understanding of how well each algorithm performed. Finally, using the best algorithms from the simulation we applied these algorithms to the real data sets to examine their utility in practice.

Materials and Methods

The Selected Genes

We searched the literature for gene expression changes linked to affective and anxiety disorders in some way (such as altered expression in disease state or alteration with drug treatment). Due to limited availability of human studies in blood samples, we considered results from both human and animal experiments, from both blood and brain tissue, and from both RNA and protein expression measurements. We screened our initial list to determine which targets could be reliably detected in blood with qPCR (quantitative Polymerase Chain Reaction) and arrived at the genes listed in Table 1. Supplementary figure S1 provides links to references supporting the rationale for the inclusion of the selected targets in the panel. The selected genes are involved in specific biological functions, such as cellular growth and proliferation as well as cell death/apoptosis. Some of the genes are associated with inflammatory and immunological pathways such as immune response or the development and function of the hematological system. We found no correlation above 0.3 between any of the gene expressions and BMI or age in healthy controls[15].

Ethics Statement

All of the clinical protocols used to enroll healthy control subjects or patients diagnosed with an affective disorder (Table 2) were approved by a local ethics

Table 1. The selected 29 genes and associated functions

Gene	Accession Number	Function
ADA (adenosine deaminase)	NM_000022	metabolism or immune response
ARRB1 (beta-arrestin 1)	L04685	GPCR signaling or immune response
ARRB2 (beta-arrestin 2)	BC007427	GPCR signaling or immune response
CD8 alpha (CD8 antigen alpha polypeptide)	M12824	immune response
CD8 beta (T-cell surface glycoprotein CD8 beta chain)	M37601	immune response
CREB1 (cAMP responsive element binding protein 1)	NM_134442	cell growth or proliferation
CREB2 (cAMP responsive element binding protein 2)	M86842	cell growth or proliferation
DPP4 (dipeptidyl peptidase 4)	M74777	metabolism
ERK1 (extracellular signal-related kinase 1)	M84490	cell growth or proliferation
ERK2 (extracellular signal-related kinase 2)	M84489	cell growth or proliferation
Gi2 (G protein, alpha-inhibiting activity polypeptide 2)	X04828	GPCR signaling or cell growth or proliferation
Gs (G protein, alpha-stimulating activity polypeptide 1)	AF493897	GPCR signaling or cell growth or proliferation
GR (glucocorticoid receptor)	X03225	glucocorticoid signaling or stress response
INDO (indoleamine pyrrole 2,3-dioxygenase)	NM_002164	inflammation
IL-1β (interleukin-1 beta)	NM_000576	inflammation
IL-6 (interleukin-6)	M14584	inflammation
IL-8 (interleukin-8)	M28130	inflammation
MAPK14 (mitogen-activated protein kinase 14) (p38 MAPK)	L35253	proliferation or inflammation
MAPK8 (mitogen-activated protein kinase 8)	AY893269	stress response
MKP1 (dual specificity phosphatase 1)	X68277	proliferation
MR (mineralocorticoid receptor)	M16801	glucocorticoid receptor signal. or stress response
ODC1 (ornithine decarboxylase)	NM_002539	cell death or apoptosis
P2X7 (purinoreceptor P2X7) (P2RX7)	NM_002562	inflammation or cell death
PBR (peripheral-type benzodiazepine receptor)	BC001110	stress response or neurosteroid biosynthesis
PREP (prolyl endopeptidase)	D21102	metabolism
RGS2 (regulator of G-protein signaling 2)	NM_002923	G protein signaling
S100A10 (S100 calcium-binding protein A10) (p11)	NM_002966	monoamine signaling
SERT (serotonin transporter)	NM_001045	monoamine signaling
VMAT2 (vesicle monoamine transporter 2)	L23205	monoamine signaling

This selection of genes is based on the literature, incorporating information from both human and animal experiments and from both RNA and protein expression measurements. The list was screened to select targets that could reliably be detected in blood with qPCR. See supplementary figure S1 for more information.

Table 2. Basic demographic data for the control and patient groups

Group	Geographic region	N	Average age	Gender composition	MW (age)	MW (gender)
Controls	England (UK) Denmark (DC) Serbia (PTSD)	196	42	78% male	N/A	N/A
BPD	USA	21	33	90% female	0.003	P< 0.0001
Acute PTSD	Serbia	66	46	100% male	0.002	0.010
Remitted PTSD	Serbia	41	45	100% male	0.056	0.044
Trauma	Serbia	87	42	100% male	0.457	0.003

The names listed in the first column of the table are used throughout the paper to refer to the different groups. Geographic region refers to the location where the samples were collected. A Mann Whitney test (MW) was used to compare the age and gender distributions between the control group and each patient group. The p values for these comparisons are reported in the last two columns of the table.

committee as follows: UK (Brent Medical Ethics Committee), DC (Danish Ethical-Scientific Committees and the Danish Data Protection Agency), PTSD (Ethical Review Board of the University Medical School, Belgrade, Serbia), BPD (Western Institutional Review Board). The samples were collected according to all applicable laws and regulations. All individuals, at all clinical sites, read and signed an informed consent document prior to donating a blood sample.

Total RNA Isolation from Human Blood

Human blood was collected in PAXgene™ blood RNA tubes (PreAnalytiX) according to the manufacturer's instructions and stored at −80°C until processing. Whenever possible, collection was made in the morning. Prior to RNA extraction, samples were incubated overnight at room temperature and centrifuged at 3000 G for 10 minutes. The pellet was washed with water, recentrifuged, resuspended in lysis buffer (Ambion)/ 177 nM sodium acetate pH 5.5 and extracted with acid phenol/chloroform. Total RNA was purified from the aqueous phase using the RNAqueous-96 automated kit (Ambion) following the vendor's instructions. A second DNase I step was added to eliminate all contaminating genomic DNA. Desalting of the RNA was accomplished by applying the samples to a MultiScreen plate (Millipore). The desalted RNA was resuspended in water and stored at −80°C. RNA QC was performed using a Bioanalyzer.

cDNA Synthesis and Quantification

Reverse transcription of 1 µg of total RNA was accomplished using Superscript II (Invitrogen) per the vendor's protocol. The resulting cDNA was desalted using a MultiScreen plate, re-suspended in water and quantitated using Quant-it™ Oligreen ssDNA reagent (Invitrogen). Based on the QC results, the cDNA concentrations were normalized to the same concentration.

Quantitative Polymerase Chain Reaction (qPCR)

For qPCR assays, replica 96 well plates were assayed. Replica plates also contained 3 wells of water to serve as a negative control and 3 wells of reference cDNA. The utility of the reference cDNA is that it allows results from different experiments to be compared since all samples are expressed relative to the reference (see below). PCR assays were performed with hydrolysis probes on either a 7900HT Fast Real Time PCR System (Applied Biosystems) or an MX3000 instrument (Agilent) using BrilliantII FAST QPCR Master Mix (Agilent). Duplicate assay plates were run for each gene and the results were averaged.

Normalization of Gene Expression

In order to effectively compare gene expression profiles between different samples, it is essential to control for variation caused by the day to day differences in the efficiency of enzymatic reactions, instrument performance, and pipetting. The preferred way to minimize the influence of these variables is through the use of multiple normalization transcripts[16–18]. We evaluated a collection of 7 candidates (Table 3) for this purpose using GeNorm (PrimerDesign).

Table 3. Normalization genes

Gene	Gene Accession Number
B2M (beta-2-microglobulin)	NM_004048
GAPDH (glyceraldehyde-3-phosphate dehydrogenase)	NM_002046
PPIA (peptidylpropyl isomerase A)	NM_021130
RPLPO (ribosomal protein, large, P0)	NM_001002
RPL13A (ribosomal protein L13a)	NM_012423
TBP (TATA box binding protein, transcription factor IID)	NM_003194
UBC (ubiquitin C)	NM_021009

The genes were evaluated using the GeNorm software package by testing multiple groups of healthy controls or patients for expression variation. The combination of these genes achieves good normalization, as determined by a pair wise variation value (V) of 0.15 or less[19].

The goal of this process is to identify at least 3 transcripts whose expression is not influenced by variables in the study design (such as disease state, drug treatment, or demographic factors). By testing different combinations of controls and patients, we observed that the rank order of the normalization transcripts differed depending on the specific group tested (supplementary figure S2). Some transcripts were identified as stable most of the time (RPL13A, RPLPO) while others were less so (GAPDH and UBC), but there was not a consistent panel of 3 or 4 transcripts that was always the best with all combinations tested. Because our long term goal is to compare gene expression profiles across many different patient and control groups, and since such comparisons require the same normalization scheme in all cases, the best solution is to use all 7 of the normalization transcripts in combination. This is indeed a valid approach, since the degree of variation for all 7 transcripts combined is acceptable, as defined by a pairwise variation score of 0.15 or less (supplementary figure S2). An added

advantage of using a large number of genes for normalization is that this dampens the influence of any single gene whose expression may change unexpectedly due to unforeseen reasons (such as drug treatments, genetic background or ethnicity).

Transcription Data Analysis

The expression level for each unknown cDNA sample was calculated relative to the reference cDNA using the 2-delta delta C(T) method[20]. Copy numbers were determined by multiplying the relative expression values by the number of copies of each transcript contained in the reference cDNA. The copy number determinations for the reference cDNA were made separately by measuring the expression level of each transcript in the reference against a standard curve of synthetic DNA for each gene of interest.

Control Subjects and Patient Groups

Table 2 summarizes the basic demographic data for the healthy control subjects and patients examined in this investigation. BPD (borderline personality disorder) patients were diagnosed according to DSM IV guidelines and did not present other acute psychiatric symptoms. Patients diagnosed with symptoms of acute PTSD (post-traumatic stress disorder) or remitted PTSD (according to DSM IV guidelines) are veterans of military conflict. Trauma patients had been exposed to a variety of traumatic events without displaying symptoms of PTSD. At the time the blood samples were collected for the transcription profiling experiment, approximately 25% of the BPD patients were receiving treatment with antidepressant (either venlafaxine or duloxetine) and about 30% of the patients diagnosed with acute PTSD were receiving treatment with a variety of medications. We chose to utilize a large control group derived from several geographic regions for comparison to the different patient groups, rather than using much smaller, matched control groups for each comparison. The intent was to mimic some of the genetic, cultural, and dietary heterogeneity that exists in the general population and we felt that this provided the best way to evaluate our analyses.

Our decision not to utilize matched controls for each patient group resulted in data sets that are not matched by age and gender. As shown in Table 2, all of the patient groups are significantly different from the control group in terms of gender composition and two are different with respect to age distribution. Because of this imbalance, we used the control group to conduct an analysis of the impact of age and gender on the expression of each of the transcripts in our panel. No transcripts in the panel demonstrated differential expression due to age based on the observation that the Spearman correlation coefficients for age versus gene expression are less than 0.3[15]. With respect to gender, our analysis revealed that ARRB2, ERK1, IL-1β on levels between the genders is quite similar (supplementary figure S4). Regardless, proper interpretation of our results requires that we account for the possibility of gender bias as it relates to the gene expression profiles in our patient and control groups.

Classification with Variable Selection

It would obviously be of great benefit if one were able to predict the diagnosis of a patient solely based on the gene expressions in a blood sample. Such prediction is especially difficult for complex psychiatric diseases where multiple genes are assumed to be involved. While many different analysis methods exist, e.g. classification algorithms, we would like to understand which type of algorithms performs the best in relation to the present purpose. Therefore, we have set up a simulation study to test a variety of well-known classification algorithms. In this way we are able to examine how different combinations of explanatory variables are identified with various classification methods.

Classification with automatic variable/feature selection offers a supervised multivariate approach to prediction of future events, e.g. response to treatment or disease course, and molecular diagnosis[21] as well as an algorithm-oriented approach to extracting the variables responsible for class separation and prediction. Common univariate methods like *t*-tests and Wilcoxon tests "*are fast and conceptually simple. However, they do not take correlations and interactions between variables into consideration, resulting in a subset of variables that may not be optimal for classification*"[21]. Multivariate variable selection approaches, on the other hand, recognize that the subset of variables with best univariate discrimination power are not necessarily the best subset of classification variables, and try to determine which combinations of variables yield high prediction accuracies. qPCR gene expression data and classification analysis range between traditional statistics and microarray analysis as measured by the number of samples and variables. For most groups in the present study, the ratio of the number of subjects to the number of measured gene

expressions is around 1–3, making it necessary to consider both microarray approaches and traditional statistical methods.

On this background, we decided to set up a simulation study to investigate which classifiers and feature selection methods would be most reliable for analysis of qPCR data, taking different kinds of gene interactions into account (see below). Key issues were identification both of classifiers able to determine the correct explanatory genes and of performance measures suitable of rating the classifiers. It is feasible from a purely technical perspective to measure relative transcription differences between control subjects and patients. However, it is difficult to place such observations into a meaningful biological context since the absolute range of mRNA expression in healthy individuals has not been defined for the transcripts in our panel. Furthermore, studies conducted in different labs are difficult to compare because different normalization schemes influence both the intensity and direction of the apparent expression changes. For these reasons, it was difficult to predict for any particular gene how the transcription profile in patients might differ from that of healthy controls. Therefore, we elected to approach the evaluation of the classifiers tested in this study in an open ended manner.

The use of simulated data sets allows us to create situations where we define the outcome, that is, we decide which combination of variables is used to define the outcome. In the simplest case, one could imagine that increased or decreased expression of one or several transcripts would be sufficient to distinguish controls from patients and achieve good classification accuracy. To mimic this, we evaluated the performance of the algorithms when the relative expression of one or more transcripts was placed above a defined threshold. Because we could not rule out the possibility that very high or very low expression of a particular transcript might result in the same biological outcome, a separate iteration involved testing the algorithms when the relative expression values fell within a defined expression interval. Smaller relative expression changes (up or down) may not be sufficient to effectively discriminate controls from patients. However, due to interactions between gene products within biological pathways, such small changes may be amplified in a way that would provide a means to discriminate the two groups. For that reason, we also tested scenarios in which either the ratio or the product of two (or more) transcripts were used for classification. As noted above, we tested situations where the ratios or products fell above a defined threshold as well as when they fell within a defined interval.

Simulation Study

The simulation study was divided into two phases. In phase 1, the simulated data sets had approximately the same number of variables as the real data; 30 variables were used. The number of samples per data set was set to both N=100 and N=1000, which represents the number of samples in the current data set as well as the size of data sets likely to be analyzed in the near future. The correlation between variables was set in some data sets to 0 and in other data sets to 0.5. The major distinction between phase 1 and phase 2 was that in phase 1 all variables were drawn from a normal distribution, while in phase 2 a realistic data set (based on actual data) was considered, see the distribution histograms in supplementary figures S5 and S6.

(1) Phase 1

In the first phase, we wanted to rule out classifiers that could not solve tasks we deemed important and, hence, defined a list of 42 linear and nonlinear tasks.

The major aspects of the phase 1 tasks were as follows. To begin with, the outcome was just a function of a single variable *x* that could fall above a threshold "a" or in a specified interval a $\leqslant x \leqslant$ b, as shown in Figure 1. This was done to understand how the classifiers would perform with simple tasks. Hereafter, the outcome was taken as a function of different combinations of two or five variables in a linear, ratio or product combination, and always either above a threshold or in a specified interval.

Three separate issues were examined. First, different magnitudes of two variables were tested (X1≈X2, X1≈10*X2 and X1≈100*X2). This was done in order to see whether the relative magnitude of the involved variables played a significant role. Next, different fractions of data points classified as Y=1 (0.05, 0.20 and 0.50) were considered. This was very relevant to us, as in some cases, the number of patients was much smaller than the number of controls. Finally, in order to determine how small a difference the classifiers could detect, two populations with different mean values in gene 1 were investigated. The mean values of gene 1 ranged from (total of five scenarios): Y=1 if Gene 1 ~N(−3,1), Y=0 if Gene 1 ~N(+3,1)) to Y=1 if Gene 1 ~N(−0.25,1), Y=0 if Gene 1 ~N(+0.25,1).

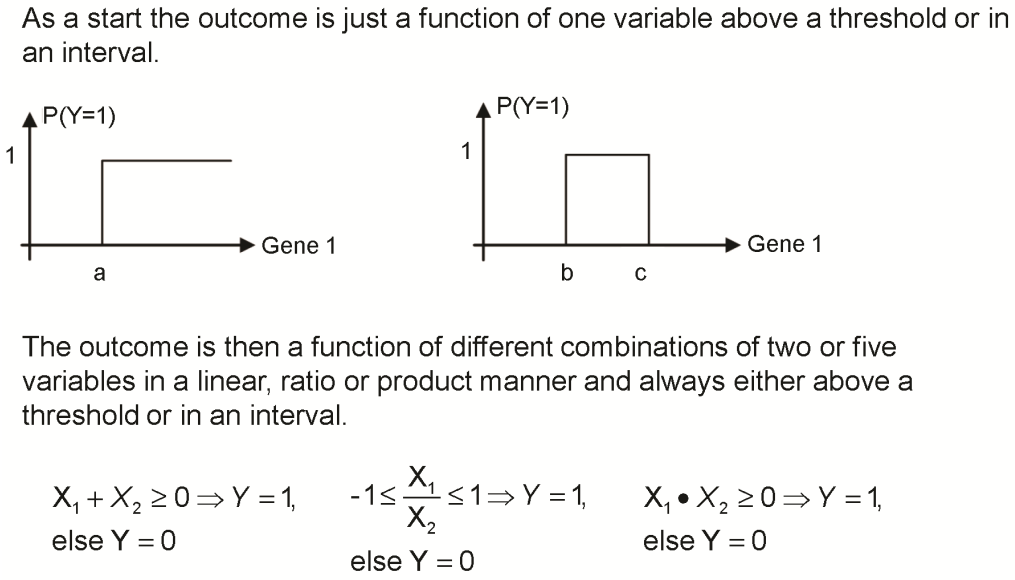

Figure 1. Simulated data sets for classifier selection.
Multiple stimulated data sets containing different known patterns of expression were constructed with N=100 or N=1000 values and different degrees of correlation between the variables. First the outcome is just a function of one variable falling above the threshold "a" or in an interval b<x<c. Hereafter, a variety of tasks are considered where the outcome is a function of two or five variables in different linear, ration and product form. A total of 14 classification algorithms were evaluate to quantify the ability to find the pattern in the various data sets.

Tested Classifiers and Automatic Feature Selection Methods

Our focus was on classifiers with either built-in variable selection or with variable selection as a pre-step to classification — all available in the statistical language R (http://www.r-project.org/). As listed in Table 4, a range of different methods were chosen for phase 1. Initial testing where the outcome was dependent on either a single variable or the sum of two variables demonstrated that the variable selection method (varselrf[40,41]) performed very well, so the classifiers NB, LDA, SKNN and random forest were started off with the varselrf-selected variables. QDA, SVM, NNET and LogitBoost were tested with two different variable selection methods: msc (based on Mass Spectra Classification[37]) and varselrf (variable selection based on random forests).

Table 4. Tested classifiers and variable selection methods available in R

	Classifier Description
1	Pelora with only the first Pelora cluster used to avoid overfitting[22,23]
2	SLR – Stepwise Logistic Regression[24,25]
3	PLR - Penalized Logistic Regression with a stepwise variable selection[26]
4	RPART - Recursive PARTioning (classification tree), see[27]
5	NB – Naive Bayes is a standard classifier known to perform well[28,29]
6	LDA – Linear Discriminant Analysis is a classical classifier[29,30]
7	SKNN – Simple K Nearest Neighbor. K-NN is described in[29,31,32]
8	Random Forest[33–35]
9	QDA – Quadric Discriminat Analysis is also described in[31,36,37]
10	SVM – Support Vector Machines[37,38]
11	NNET – Neural NETwork with a single-hidden-layer[37]
12	LogitBoost – a boosting machine learning technique[37,39]

Initial testing demonstrated that the variable selection method varselrf performed well. Hence, the classifiers NB, LDA and Random Forests were started with varselrf selected variables.

Accuracy, Cross-validation and the Jaccard Similarity Coefficient

In order to determine the performance of a classifier for both two and multiple class tasks, we decided to measure the accuracy of the predictions[42].

The accuracy was measured in each cross-validation sample and finally averaged. As recommended by Kohavi[43], 10-fold stratified cross-validation was used. This also meant that the accuracy measured would not be (too) inflated due to overfitting. To measure how well a classifier identified the correct variables, the Jaccard similarity coefficient[44] was used. The Jaccard score yields a number (0–100%) indicating how well a given classifier identifies variables compared to the correct explanatory variables as defined by us. The higher the Jaccard score is, the better

the agreement will be between the variables identified in the analysis and the predefined variables.

The phase 2 data set is based on the control group as well as the borderline personality disorder and acute PTSD patients. 25 variables/genes were included. The number of samples was 263. The mean correlation between the 25 genes was 0.26, that is, in the range of correlations considered in the phase 1 study. All data was normalized the same way; the variables used were in one case transformed to *z*-scores (standard score; $z = \frac{x - \mu}{\sigma}$, where μ is the mean and σ is the standard deviation), as in phase 1, and in another case, the actual expression values were used. These two approaches were chosen in order to see if the outcome of the classification algorithms differed. As in phase 1, the outcome was defined as different combinations of the variables. 33 different tasks were given to the classifiers (description available on request) with most of them similar to the tasks in phase 1, that is, to begin with, the outcome was just a function of one variable above a threshold or in an interval. Then, the outcome was a function of different combinations of two or five variables in a linear, ratio or product form and always either above a threshold or in a specified interval.

Three separate studies were performed: Firstly, we wanted to look at the completely random outcome to observe how the classifiers / variable selection method would perform in this situation. Secondly, by looking at different fractions of data points classified as Y=1 (0.05, 0.20 and 0.50), we could mimic the situation with unbalanced data sets. Finally, we would test the actual data (no z-score transformation) by looking at different magnitudes of two involved variables (X1≈X2, and X1≈(100-300)*X2).

Classification and Variable Selection Procedure Working with the Real Data

Since the accuracy values are dependent on the group sizes (which differ among the different patient and control groups), the following classification procedure was used to determine whether a group separation was possible, and if possible, how to report the responsible genes:

1. Calculate 10-fold stratified CV (cross-validation) accuracies in the real case scenario, i.e. with the actual control and patient data. The variable selection procedure is included in each cross-validation.

2. Calculate permuted accuracies by performing 10 permutations (the result is almost the same using 100 permutations, however, the computation times are much longer) of the class labels leading to 10 × 10 (CV) = 100 permuted / random accuracies. The permutation step is applied in order to calculate the accuracy values expected at random in the real data set (excluding the class label) for a classifier[45].

3. Compare the 10 real case accuracies with the 100 permuted accuracies using a *t*-test if the accuracy values follow a normal distribution (tested with a Shapiro-Wilk test[46]), otherwise use a Wilcoxon test[47].

4. Significant result is obtained (that is, the groups are separable) if the *p*-value is below the significance level of 1% (adjusted for multiple tests).

5. Genes corresponding to the significant result are listed. Genes are extracted from the complete data set from each classifier (selected genes may depend on classifier). Overlapping genes are reported.

To get additional useful information from the classifications, the positive and negative predictive values (PPV and NPV respectively) are reported as well together with the false positive and true positive rates (FPR and TPR respectively, the latter also known as sensitivity). They are defined as:

$$PPV = \frac{number\ of\ True\ Positives}{number\ of\ True\ Positives + number\ of\ False\ Positives}$$

$$NPV = \frac{number\ of\ True\ Negatives}{number\ of\ True\ Negatives + number\ of\ False\ Negatives}$$

$$FPR = \frac{number\ of\ False\ Positives}{number\ of\ False\ Positives + number\ of\ True\ Positives}$$

$$TPR = \frac{number\ of\ True\ Positives}{number\ of\ True\ Positives + number\ of\ False\ Negatives}$$

Results

Phase 1

An example of one of the 42 simulation results from phase 1 in the form of plots with the *x*-axis representing the accuracy and the *y*-axis the Jaccard score is shown in Figure 2. The abbreviations in the plots refer to the classifier names given in the classifier list above. A suffix of '.RF' means that the variable selection method 'varselrf' (variable selection based on random forests) was used as a pre-step to the corresponding

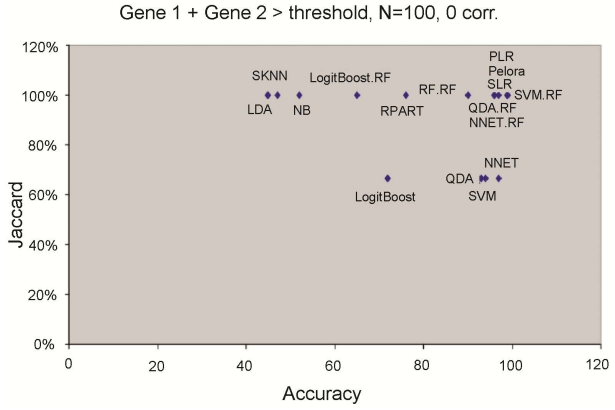

Figure 2. Classification performance when two variables sum to an outcome that exceeds a threshold.
A simulated data set (N=100) was analyzed in which the expression pattern was linked to two variables, the sum of which exceeds a given threshold. A total of 16 classifier possibilities were evaluated. The best performing classifiers are: PLR, Pelora, SLR, SVM.RF, QDA.RF, NNET.RF, RF.RF and RPART. The figure was made in Excel based on output from R. The Jaccard score is a measure of how well a classifier identifies the correct variables, two in this case, and accuracy indicates how well a classifier is at predicting the outcome, here above a given threshold.

classifier. Otherwise, the feature selection method 'msc' (variable selection based on mass spectra classification) was used as a pre-step to the quadratic discriminant analysis (QDA), Random Forest (RF), support vector machines (SVM), neural network (NNET) and boosting LogitBoost classifiers.

In Figure 2, the 16 classifier possibilities are shown for a task involving a linear combination of two variables with an outcome that exceeds a given threshold. Furthermore, there is no correlation between the variables. This plot indicates that the LDA, NB, SKNN, LogitBoost, LogitBoost.RF, QDA, NNET and SVM classifiers do not perform well for this kind of task.

By looking at 42 simulation plots (four more plots are shown in the supplementary figures S7 thru S10, the remaining plots are available on request) corresponding to the different linear and nonlinear phase 1 tasks, we concluded in a manner similar to Table 5 below;

- The variable selection method varselrf seemed very promising especially in connection with the classifiers SVM, random forest or QDA dealing with a broad range of linear and nonlinear classification problems.
- When the outcome was a linear combination of variables, SLR and Pelora did the best classification job.
- The following cases were not well handled by the above mentioned classifiers in general:
- "Large" ratios involving 5 variables
- "Large products" involving 5 variables
- Data sets of size N=100 with only 5% Y=1
- The difference between Y=0 and Y=1 variable mean values was ~0.5

Most importantly, we concluded that the following classifiers and variable selection methods in general performed the worst: NB (varself selection method), LDA (varself selection method), SKNN (varself selection method), QDA (msc selection method), SVM (msc selection method), NNET (msc selection method) and LogitBoost (msc and varselrf selection methods). Hence, these methods are not part of the classifiers tested in phase 2. The reasons for poor performance were typically that these classifiers and selection methods yielded low Jaccard score (i.e., they were not able to identify the responsible genes satisfactory) and/or bad accuracy score considered overall for the various tasks.

The worst performing classifiers from phase 1 were omitted, and thus, the following classifiers and feature selection methods would be tested in phase 2:

Table 5. Phase 2 tasks solved with accuracies above 80% and a Jaccard score above 70%

2 groups	% task solved	Average accuracy	Average Jaccard	Tasks solved
Pelora	28%	95%	97%	r1,r3,r9,r16,r17,r18,r22
SLR	36%	100%	100%	r1,r3,r9,r16,r17,r18,r20,r21,r22
PLR	28%	100%	100%	r1,r3,r5,r9,r16,r17,r18
RPART	56%	94%	97%	r1,r2,r3,r4,r6,r16,r17,r18,r20,r21,r22,r23,r24,r26
Random Forest (varselrf)	52%	95%	98%	r3,r4,r5,r6,r7,r8,r9,r20,r21,r22,r23,r24,r26
QDA (varselrf)	48%	89%	98%	r3,r4,r6,r7,r8,r9,r20,r21,r22,r23,r24,r26
SVM (varselrf)	52%	97%	98%	r3,r4,r5,r6,r7,r8,r9,r20,r21,r22,r23,r24,r26
NNET (varselrf)	16%	91%	94%	r3,r4,r9,r22

The column 'Tasks solved' refers to the specific tasks solved (available on request). r4, for instance, involves the task defined as $-0.95 \leqslant X_2+X_3 \leqslant 0.95 \Rightarrow Y = 1$, else Y = 0 and is shown in Figure 3. The conclusions in the text sum up the table.

1. Pelora with only the first Pelora cluster used
2. SLR
3. PLR
4. RPART
5. QDA
6. Random Forest
7. SVM
8. NNET

QDA, Random Forest, SVM and NNET were tested with the variable selection method varselrf.

Phase 2 Simulation Results

In Figure 3, an example of the results from phase 2 is presented. In this case a non-linear task is examined. Four more plots are shown in the supplementary figures S11 through S14; the remaining figures are available on request. The abbreviations in the plots are the same as in phase 1. Here it should be emphasized that one-gene simulations were only performed to see how the classifiers dealt with this task, and they are not included in the examples below. Since the diseases we consider are believed to be polygenetic, a classifier is not of interest if it only performs well in a single variable/single gene case. In order to compare the classifiers on a quantitative basis, we decided to focus on classifiers that yielded an accuracy above 80% and had a Jaccard similarity score above 70%. This we believed would reflect good-performing classifiers on the available qPCR data, even though these percentages were chosen more or less arbitrarily. In Table 5,

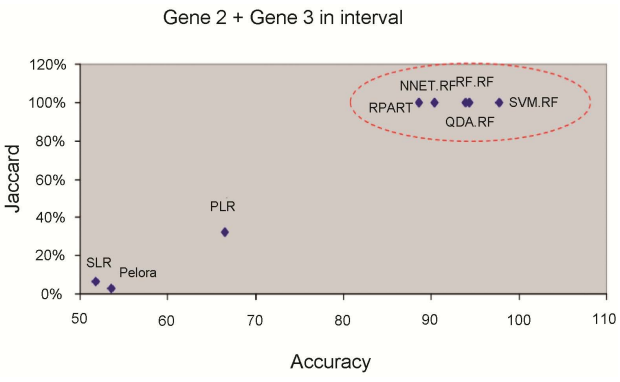

Figure 3. Classification performance when two variables sum to an outcome in a specified interval.

A simulated data set (N=263) was analyzed in which the expression pattern was linked to two variables, the sum of which falls within a specified interval. Here, a total of 8 classifiers were evaluated in phase 2. Three classifiers do not perform well; SLR, Pelora and PLR, while the encircled classifiers solve this task well. Note, however, that several of the classifiers only perform well in connection with varselft (variable selection based on random forests). Jaccard score and accuracy are as defined in Figure 2.

the eight classifiers are listed together with information on the percentage of tasks solved, the average accuracy (above 80%), the average Jaccard score (above 70%) and the specific tasks solved.

Based on Table 5 and the 33 simulation plots, we concluded;

- Above an accuracy threshold of 80% and a Jaccard score of 70%, RPART, SVM (varselrf) and Random Forest (varselrf) solved the largest fraction of given tasks.
- SVM and Random Forest solved the same tasks, however, SVM yielded a slightly higher accuracy.
- RPART was very good at dealing with one variable (threshold and interval) incl. small fraction of 1's. Furthermore, RPART identified some of the same variables as varselrf did. It was reassuring to have the same variables selected by two methods (although the methods share some similarity they are not identical).
- SLR solved less tasks than RPART and SVM, but more than Pelora and yielded both 100% average accuracy and 100% average Jaccard score. Furthermore, unlike Pelora SLR is able to handle categorical clinical variables well.
- Accuracy thresholds above 80% seem to yield very high Jaccard scores > 95% and high classifier accuracies ~90–95%.
- As expected standardization (e.g., scaling by subtraction of the mean and division with the standard deviation) of data yielded the same results as unstandardized data.
- NB! In general, all classifiers were used with default settings or default recommendations.

Optimizations of various settings would probably lead to different results, including a possible greater risk of over-fitting. Based on phase 1 and phase 2 simulation studies, we decided to focus on:

- SVM in combination with varselrf
- RPART
- SLR

SLR can handle linear cases with multiple variables above a threshold a task that neither SVM nor RPART is good at. Furthermore, we defined key variables / gene expressions as those that permit groups of equal sizes to be separated with an accuracy > 80%. If groups could not be separated with SVM, RPART or SLR, we would say there were no expression differences between the groups.

Results with Actual Data from Control Subjects and Patients

Variable Selection and Classification Among Various Groups

The most promising algorithms identified above were applied to various comparisons between actual control and patient groups. Four main two-group comparisons were made; controls vs borderline personality disorder (BPD) patients, controls vs PTSD acute patients, controls vs remitted PTSD patients and controls vs individuals who suffered trauma without developing PTSD. Below the results of these four comparisons are presented in summary tables including genes selected to differentiate groups, PPV (positive predictive value), NPV (negative predictive value) and both the actual accuracy values as well as the permuted accuracy values (in parentheses in the tables), all in percentages. It is also mentioned below for each table whether the accuracy values of a comparison is significant on the 1% significance level compared to permuted values. Finally, in supplementary figure S15 the number of true and false positives and negatives for each comparison and each classifier is shown.

The controls were derived from 3 different subject groups (DC, UK and Serbian controls) and the 29 gene expression values were used for comparison. Firstly, the controls (N=196) were compared with the BPD patients (N=21) in Table 6. In general, the genes selected were very similar regardless of variable selection technique, with Gi2, GR, and MAPK14 repeatedly identified by the algorithms. Furthermore, all accuracies were high and significant compared to the permuted values (data not shown). The reasons for the high permuted values are unbalanced data sets. The positive predictive values were high except for RPART.

Table 7 compares controls to PTSD acute patients (N=66). In general, the genes selected depended on the variable selection technique. It was noted that *ARRB2*, *ERK2*, and *RGS2* were consistently picked. Also, it was noted that RPART performed worse than the other classifiers, just as in the case with controls vs BPD patients. All accuracies were significant compared to the corresponding permuted values (data not shown) except for RPART. The positive predictive and sensitivity values were considered marginal.

In Table 8, controls were compared to remitted PTSD patients (N=41). The two groups could not be separated by any classifier on the 1% significance level (data not shown, and note how close the actual and permuted accuracy values are). The PPV and sensitivity values were markedly lower than in the PTSD acute patient case (Table 7). Finally, in Table 9, controls were compared to patients suffering trauma but without PTSD (N=87). All accuracy values were significant compared to permuted values. However, the PPV and sensitivity values were not impressive so the two groups are poorly separated regardless of classifier employed. It was noted that *ARRB2* and *ERK2* were consistently picked just as they were in the PTSD acute patients. Otherwise, *Gs, MKP1* and *IL-6* were also consistently picked.

Table 6. Summary of controls vs BPD patients

Classifier	Genes selected to differentiate groups	PPV	NPV	FPR	TPR	Accuracy (permuted)
SVM/varselrf	*Gi2, GR, MAPK14*	97	99	1	87	98 (90)
SLR	*Gi2, GR, MAPK14, MR*	93	99	1	88	98 (88)
RPART	*Gi2, GR*	68	98	3	75	95 (88)

In the table positive and negative predictive values (PPV and NPV), false positive and true positive rates (FPR and TPR) and accuracy values are in percentages. Permuted accuracy values are in parentheses in the last column and explained in the second step in the "Classification and variable selection procedure working with the real data" section. All accuracy values are significant compared to permuted values (data not shown). The analyses were done in R

Table 7. Summary of controls vs PTSD acute patients

Classifier	Genes selected to differentiate groups	PPV	NPV	FPR	TPR	Accuracy (permuted)
SVM/varselrf	*ARRB2, ERK2, RGS2*	72	82	6	37	80 (73)
SLR	*ARRB1,ARRB2, CD8 beta, ERK2, IDO, IL-6, MR, ODC1, PREP, RGS2*	77	87	8	60	84 (71)
RPART	-	48	82	20	49	72 (64)

In the table PPV, NPV, FPR, TPR and accuracy values are in percentages. Permuted accuracy values are in parentheses in the last column. All accuracy values are significant compared to permuted values (data not shown) except for RPART (p-value: 0.02349). The analyses were done in R.

Table 8. Summary of controls vs PTSD in remission

Classifier	Genes selected to differentiate groups	PPV	NPV	FPR	TPR	Accuracy (permuted)
SVM/varselrf	-	49	86	6	22	82 (82)
SLR	-	33	86	10	28	80 (81)
RPART	-	28	86	14	28	76 (75)

In the table PPV, NPV, FPR, TPR and accuracy values are in percentages. Permuted accuracy values are in parentheses in the last column. No accuracy values are significant compared to permuted values (data not shown).

Table 9. Summary of controls vs trauma patients without PTSD

Classifier	Genes selected to differentiate groups	PPV	NPV	FPR	TPR	Accuracy (permuted)
SVM/varselrf	*ARRB2, CREB1, DPP4, ERK1, ERK2, Gs, IL-6, IL-8, MAPK8, MKP1, MR, PBR, PREP, SERT*	51	76	16	40	71 (65)
SLR	*ADA, ARRB2, CD8 beta, CREB1, ERK2, Gs, IL-6, MAPK14, MKP1, MR, RGS2, VMAT2, IL-1 beta*	59	80	17	51	73 (64)
RPART	*ARRB2, ERK2, Gs, IL-6, IL-8, MKP1, PREP, SERT*	63	83	17	61	76 (58)

In the table PPV, NPV, FPR, TPR and accuracy values are in percentages. Permuted accuracy values are in parenthesis in the last column. All accuracy values are significant compared to permuted values (data not shown).

Discussion

Classifiers and Variable Selection Methods

By using a simulation study, we were able to point to SVM (Support Vector Machines) combined with varselrf (variable selection based on random forests), RPART (Recursive Partitioning) and SLR (Stepwise Logistic Regression) as suitable classifiers and variable selection methods for analyzing our qPCR gene expression data. The same parameters were then applied to generate classification results using actual data from control subjects and patients. Whereas positive and negative predictive values (PPV and NPV) were not assessed in the simulation study design, these values were derived for group comparisons based on actual data, and could therefore be used to further rank the classifiers. RPART distinguished itself by having consistently lower PPV and accuracy values in almost all of the group comparisons. This raised concerns about the predictive ability of RPART and we decided to exclude this classifier from more detailed analyses. We therefore proposed that the most promising classifiers and variable selection methods for analyzing the qPCR data are SVM combined with varselrf, and SLR. It is possible that a better performance could be obtained by tuning the parameters of the chosen classifiers and variable selection methods, using methods described in reports from the R-project[24,27,40,48]. Another possible strategy for improving performance, at least for SLR, would be to include categorical clinical variables together with gene expression data, rather than treating them as independent variables. All strategies to further optimize must be weighed carefully against the risk of over-fitting. In addition, it should be mentioned that SVM and SLR are deterministic classifiers and that another approach relevant for psychiatry could have been to use a probabilistic classifier such as RVM (relevance vector machine).

The Simulation Study

Some consideration should be given to issues associated with the simulation study. First of all, the basis for the mathematical approach to the gene-gene interactions was an approximation, because we do not know the exact biological interactions between genes on the expression level. Therefore, we investigated a variety of simple mathematical constructs in the form of different linear and nonlinear tasks. Had we chosen to pursue alternative types of mathematical approaches, different classifiers and variable selection methods might have been selected, such as, for instance, regularized discriminant analysis, classification using generalized partial least squares, neural networks, and other types of more advanced methods.

Regarding the choice of classifier, an alternative strategy would have been to identify a single classifier capable of handling a broad range of classification tasks. In practice we were unable to identify such a classifier that would perform sufficiently well in all of our classification tasks; moreover we valued having multiple types of classifiers as a way to check for consistency among the gene signatures. It was also necessary to make some strategic choices with respect to the use of a single accuracy measure. Had we only

focused on two-group comparisons, the Matthews correlation coefficients would have been a superior choice[49]. However we were interested in multiple group (> 2 group) comparisons and, therefore, focused on a single accuracy measure with broad utility[42]. A limitation of relying on a single accuracy measure, however, is that unbalanced group sizes alone can bias the output. For this reason, we incorporated permutation tests which, albeit time consuming, were able to improve the benefit of the accuracy measure. In the case of 2-group comparisons we were also able to supplement the accuracy value with PPV, NPV, FPR and TPR as measures of predictive value and thereby generate a highly informative assessment of overall classification performance.

A final issue has to do with the stability of the classifier. Evaluation of this parameter was beyond the scope of the simulation study. However, it should be mentioned that stability can be assessed *via* modifications to the data set. One or more observations can be removed from the data set; alternatively incorrect observations can be inserted. If the classification result (including accuracy, PPV, NPV, FPR and TPR values) is obtained before and after the modification exercise, the stability parameter can be determined as a function of change in classification result relative to data permutation.

Whole Blood Biomarkers for Psychiatric Diseases

As described in the results section, several genes were differentially expressed in healthy controls and patient groups, allowing for potential insights into disease pathology, and also for the possibility of disease-associated transcriptional biomarker signatures. We found that healthy control subjects could be separated from patients with borderline personality disorder based on a common set of genes selected by multiple classifiers: *Gi2*, *GR* and *MAPK14*. As noted earlier, the expression of *Gi2* is influenced by gender (supplementary figure S3), with higher expression in female control subjects relative to male controls (supplementary figure S4). The BPD patients analyzed in this study are 90% female as compared to the control group which is 78% male, raising the possibility that *Gi2* is selected by the algorithms simply because of gender bias. If this were true, one would predict that the expression level of *Gi2* in the patients would be greater than that seen in the control group. However, comparison of the expression level of *Gi2* between the BPD patients and controls reveals that the patients actually display only 61% of the expression level as controls. This relationship is in the opposite direction that would be expected if gender were the driving force for selection of *Gi2* by the classification algorithms. For this reason, we are confident that *Gi2* has been properly identified as a transcript that is useful to discriminated controls and BPD patients.

Our analysis has also shown that healthy control subjects could be separated from patients with acute post-traumatic stress disorder using another common set of genes: *ARRB2*, *ERK2* and *RGS2*. There is the possibility that the selection of *ARRB2* by the classification algorithms is the result of gender bias. Female controls demonstrate slightly higher levels of expression of *ARRB2* than male controls (supplementary figure S4). Since our control group contains 22% females whereas the acute PTSD patient group contains no females, the slightly elevated expression that we see in the controls relative to the acute PTSD patients could be caused by gender alone. The most cautious interpretation of these results is that *ERK2* and *RGS2* expression levels are reliable discriminators of the two groups but the *ARRB2* result must be considered questionable. Conversely, healthy controls could not be significantly differentiated from remitted PTSD subjects on the basis of gene expression. This result, which may be viewed as congruent with the clinical course, suggests that gene expression values may be 'normalized' upon remission from PTSD. We cannot be sure, however, whether the outcome for remitters reflects a limitation of the 29 gene list (which was not targeted specifically to the biology of PTSD) or the analysis per se.

Finally, healthy control subjects were differentiated from subjects who experienced trauma without PTSD on the basis of seven significant and commonly selected genes: *ARRB2*, *CREB1*, *ERK2*, *IL-6*, *Gs*, *MKP1* and *MR*. Although all classifiers yielded significant results, the individual PPV and sensitivity values for the trauma group were relatively weak, indicative of an overall marginal separation. We know that the expression of both *ARRB2* and *Gs* are influenced by gender (supplementary figures S3 and S4), so as noted above care must be taken when evaluating the selection of these genes by the classification algorithms. The most conservative interpretation is to only consider *CREB1*, *ERK2*, *IL-6*, *MKP1* and *MR* as genes that are informative for separation of the controls and the trauma patients.

At first glance it may appear incongruent that the

algorithms distinguish control subjects from these patients, albeit weakly, despite the lack of a clinical diagnosis of PTSD. Rather than viewing this result as a limitation of our approach, we interpret it as support for the use of the technology. It is known that repeated trauma increases the risk for PTSD. The fact that these individuals have not been diagnosed with the disease, yet still exhibit some of the gene expression patterns displayed by patients who have been diagnosed, suggests that we may be detecting gene expression changes that are related to an increased vulnerability to develop PTSD. If confirmed, this could aid detection of individuals at risk and thus supportive treatment and efforts to avoid further trauma could be implemented to avoid development of the clinical disorder.

A brief discussion of the biology associated with informative genes is warranted. Genes of interest for BPD (*Gi2*, *MAPK14* and *GR*) can be characterized as linking peripheral stress-related hormone / neurotransmitter signals with downstream signaling cascades, to regulate the balance of pro- vs anti-inflammatory processes and cell survival vs death. For example, *Galphai2* functions in the signal transduction pathway for stress-relevant chemokines (such as *IL8*, *IL10*), neurotransmitters and hormones that activate Gi/o-coupled receptors (as opposed to Gs-, Gq- or G12/13-coupled receptors), with implications for inflammation as well as peripheral blood cell motility and survival[50]. The MAP kinase *MAPK14* (or *p38*) acts downstream of *Galphai2* to elicit inflammatory responses (including upregulation of *IL6*), phosphorylation of transcription factors and apoptosis or cell senescence (death)[51,52]. The glucocorticoid receptor *GR* is a key component of the HPA / stress axis, widely observed to be dysfunctional in mood disorders[53]. Current findings for HPA axis dysregulation in BPD are somewhat equivocal however, suggesting that additional variables such as disease segmentation or co-morbidities may need to be considered to fully understand the relationship.

Genes of interest for PTSD point similarly to G protein-coupled receptor (GPCR) signaling, specifically via expression changes for *ERK2* (a downstream MAP kinase that stimulates proliferation via transcriptional regulation) and *RGS2* (a regulator of GPCR signal duration). Other noteworthy genes include the mineralocorticoid receptor *MR*, a high affinity glucocorticoid receptor whose altered expression relative to *GR* (measured as the *GR*/*MR* ratio) is indicative of HPA / stress axis disturbance[54]; and ornithine decarboxylase (*ODC-1*), a rate-limiting enzyme for synthesis of biogenic polyamines (specifically putrescine) associated with anti-apoptotic activity. Altered inflammatory status in PTSD is further indicated by changes in the cytokine *IL6* and the cytokine-stimulated transcript for indoleamine 2,3-dioxygenase 1 (*IDO*), which diverts tryptophan into kynurenine (rather than 5-HT), with implications for cytotoxicity and altered glutamatergic signaling[55,56]. The convergence on GPCR and glucocorticoid signaling mechanisms, inflammatory processes and apoptosis is consistent with previous findings for differentially expressed blood transcripts in PTSD[3,57–59]. Genes of interest for subjects with trauma but not PTSD (*ERK2*, *MR*, and others), when grouped into pathways, are largely overlapping with genes of interest for PTSD, again highlighting a common theme of stress-induced changes in GPCR signaling, inflammatory mediators, cell proliferation and death.

Some of the recurrent transcriptional classifiers resulting from the analysis of BPD, or PTSD, or trauma without PTSD (*ERK2*, *RGS2*, *GNAI2*, *MAPK14*, *GR*) were further analyzed for functional interactions and associations using Ingenuity Pathway Analysis software (IPA). Top-ranking disease associations included inflammatory response (as expected) as well as developmental disorders, skeletal / muscular disorders, cardiovascular disease and cancer. Top-ranking canonical pathways included inflammatory processes (chemokine and interleukin signaling), HPA / stress response (CRF signaling) and once again, cancer (specific tissue factors). These disease and pathway associations, while interesting and provocative from the viewpoint of disease comorbidity and novel treatment modalities, also highlight a risk of using transcriptional signatures for classifying affective disorder, i.e., lack of disease specificity. This risk might be minimized in various ways. One approach would be to expand the candidate gene list in order to derive a more complex and distinctive transcriptional signature for the target disorder. Another strategy involves prescreening individuals for co-morbid disease such as cancer, in advance of transcriptional profiling for affective disorder. In this study, care was taken to exclude individuals with serious medical disorders to avoid complication.

This work began as a pilot study and was focused primarily on identifying discriminating algorithms with the collected data, and not designed as an extensive attempt to find all biomarkers for the affective

disorders considered. Thus, our findings regarding targets that can be used to separate BPD or acute PTSD patients from healthy controls must be considered preliminary due to the small numbers of patients involved. However, we have also applied these same algorithms to the analysis of gene expression patterns from much larger cohorts of depressed patients, and the results substantiate those presented here[60]. Comparison of gene expression patterns between 174 mild to moderately depressed patients and 198 healthy controls using SVM combined with varselrf identified 11 targets that, in combination, could separate the groups with 90% accuracy. Importantly, *Gi2*, *MAPK14*, *ARRB2* and *ERK2* were included on this list. Likewise, analysis of the expression profiles in 307 severely depressed patients versus these controls using relevance vector machine, an algorithm similar to SVM, demonstrated that *Gi2* and *MAPK14* were two of the targets providing a strong means to discriminate between the groups. Given that depression symptoms can be intertwined with both BPD and PTSD, it is reasonable that there is overlap in the gene expression profiles for these disorders. As such, the findings with the much larger depression studies provide support for the validity of the results we obtained here using much smaller BPD and PTSD patient populations. A second, related limitation is that about one third of the BPD and acute PTSD patients were receiving treatment with various medications at the time the blood sample was collected for transcription profiling. Given the small number of patients available for analysis, and the diversity of the medications in use, it was not practical to account for these variables in the analyses. Therefore, we cannot rule out that some of our observations are influenced by the variable medication regimens. Finally, potential confounders that could impact gene expression profiles, such as recent stress and early life stress[15], were not measured in the control and patients groups. Despite these possible confounds, the informative genes and associated pathways from our PTSD analysis are consistent with literature reports for these patient populations[3]. It is also noteworthy that peripheral transcripts can be correlated to some extent with gene expression patterns in the CNS[61]. The relationship between peripheral and central transcripts is born out in several other published reports of subjects with psychiatric disorder[3,4,7,8,11,12], (also supported and reviewed in[62]). Taken together, these findings point to the utility of blood transcripts as informative and useful biomarkers for affective and anxiety (or psychiatric) disorders, pending validation in independent studies.

Conclusion

In summary, we have described a process of evaluating multiple classifiers based first on simulated data and secondly on clinical data, which can support the selection of optimal classifiers for evaluation of disease-relevant transcriptional biomarkers in human blood samples. Our study demonstrates that many commonly used multivariate algorithms designed to identify patterns in complex data sets do not perform equivalently. Algorithms such as SLR and RF/SVM perform the most reliably under a broad set of conditions with simulated data sets that mimic transcription profiling data. Application of these algorithms to the analysis of actual transcription profiles from blood samples suggests that they can identify patterns that effectively discriminate control subjects from patients with affective disorders. Refinement of these approaches could lead to the identification of biomarkers that, pending further validation in studies including larger numbers of individuals, are valuable for the diagnosis of psychiatric diseases.

Author Contributions

WM analyzed the simulated gene expression data, the clinically derived gene expression data, and drafted the manuscript, AA, MD, RA, and JT executed the transcription profiling experiments to generate the expression data from the clinical blood samples, BS, IA, CG, JT, and RA designed the list of genes targeted for expression profiling , IA and CG conceived of the study and established collaborations to acquire clinical samples, BS provided blood samples for gene expression analysis from healthy control subjects, MW assisted in writing the manuscript, PM provided blood samples for gene expression analysis from patients diagnosed with borderline personality disorder, EV, DS, GM, and SD provided blood samples for gene expression analysis from patients diagnosed with acute PTSD, remitted PTSD, trauma without PTSD, and healthy control subjects, UG, TW, LK, HU, and EH provided blood samples for gene expression analysis from healthy Danish control subjects, EM acted as thesis advisor to WM and assisted in drafting of the manuscript. All authors read and approved the final manuscript.

Conflict of Interest and Funding

JT, IA, AA, MD, RA, MW, and CG were employed at Lundbeck Research USA and BS was employed by H. Lundbeck A/S during completion of this work. This work was in part financed by European Commission, via Sixth Framework Programme research project INCO-CT-2004-509213 and in part by the Ministry of Science, Serbia, Project No 41009. We acknowledge support from the European Union through the Network of Excellence BioSim (Contract No. LSHB-CT-2004-005137).

Acknowledgements

We would like to thank Jan Vistisen for his useful comments related to the construction of the simulated data sets.

References

1. Antonijevic I, Artymyshyn R, Forray C, *et al.* 2009, Perspectives for an integrated biomarker approach to drug discovery and development, in Turck C, ed. *Biomarkers for Psychiatric Disorders*. Springer.
2. Tylee D S, Kawaguchi D M and Glatt S J, 2013, On the outside, looking in: a review and evaluation of the comparability of blood and brain "-omes". *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol.162B(7): 595–603.
3. Segman R H, Shefi N, Goltser-Dubner T, *et al.* 2005, Peripheral blood mononuclear cell gene expression profiles identify emergent post-traumatic stress disorder among trauma survivors. *Molecular Psychiatry*, vol.10(5): 500–513, 425.
4. Zieker J, Zieker D, Jatzko A, *et al.* 2007, Differential gene expression in peripheral blood of patients suffering from post-traumatic stress disorder. *Molecular Psychiatry*, vol.12(2): 116–118.
5. O'Donovan A, Sun B, Cole S, *et al.* 2011, Transcriptional control of monocyte gene expression in post-traumatic stress disorder. *Disease Markers*, vol.30(2–3): 123–132.
6. Tylee D S, Chandler S D, Nievergelt C M, *et al.* 2015, Blood-based gene-expression biomarkers of post-traumatic stress disorder among deployed marines: apilot study. *Psychoneuroendocrinology*, vol.51: 472–494.
7. Le-Niculescu H, Kurian S M, Yehyawi N, *et al.* 2008, Identifying blood biomarkers for mood disorders using convergent functional genomics. *Molecular Psychiatry*, vol.14(2): 156–174.
8. Middleton F A, Pato C N, Gentile K L, *et al.* 2005, Gene expression analysis of peripheral blood leukocytes from discordant sib-pairs with schizophrenia and bipolar disorder reveals points of convergence between genetic and functional genomic approaches. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol.136(1): 12–25.
9. Munkholm K, Vinberg M, Berk M, *et al.* 2012, State-related alterations of gene expression in bipolar disorder: a systematic review. *Bipolar Disorder*, vol.14(7): 684–696.
10. Munkholm K, Peijs L, Vinberg M, *et al.* 2015, A composite peripheral blood gene expression measure as a potential diagnostic biomarker in bipolar disorder. *Translational Psychiatry*, vol.5: e614.
11. Iga J, Ueno S, Yamauchi K, *et al.* 2005, Serotonin transporter mRNA expression in peripheral leukocytes of patients with major depression before and after treatment with paroxetine. *Neurosci Letters*, vol.389(1): 12–16.
12. Iga J, Ueno S, Yamauchi K, *et al.* 2007, Altered HDAC5 and CREB mRNA expressions in the peripheral leukocytes of major depression. *Progress in Neuro-psychopharmacol & Biological Psychiatry*, vol.31(3): 628–632.
13. Hepgul N, Cattaneo A, Zunszain P A, *et al.* 2013, Depression pathogenesis and treatment: what can we learn from blood mRNA expression? *BMC Medicine*, vol.11: 28.
14. Redei E E and Mehta N S, 2015, Blood transcriptomic markers for major depression: from animal models to clinical settings. *Annals of the New York Academy of Sciences*, vol.1344: 37–49.
15. Mazin W, 2008, *Exploring the biological basis of affective disorders*, thesis, Technical University of Denmark.
16. Andersen C L, Jensen J L and Orntoft T F, 2004, Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Research*, vol.64(15): 5245–5250.
17. Jin P, Zhao Y, Ngalame Y, *et al.* 2004, Selection and validation of endogenous reference genes using a high throughput approach. *BMC Genomics*, vol.5(1): 55.
18. Huggett J, Dheda K, Bustin S, *et al.* 2005, Real-time RT-PCR normalisation; strategies and considerations. *Genes & Immunity*, vol.6(4): 279–284.
19. Vandesompele J, De P K, Pattyn F, *et al.* 2002, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology*, vol.3(7): RESEARCH0034.
20. Livak K J and Schmittgen T D, 2001, Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods*, vol.25(4): 402–408.
21. Boulesteix A-L, Strobl C, Augustin T, *et al.* 2008, Eval-

uating microarray-based classifiers: an overview. *Cancer Informatics*, vol.4: 77–97.
22. *supclust*, R project — www r-project org, <http://cran.rproject.org/web/packages/supclust/index.html>
23. Dettling M and Bühlmann P, 2004, Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis*, vol.90(1): 106–131.
24. *VR bundle - MASS*, R project — www r-project org, <http://cran.rproject.org/web/packages/VR/index.html>
25. Weber G, Vinterbo S and Ohno-Machado L 2004, Multivariate selection of genetic markers in diagnostic classification. *Artificial Intelligence in Medicine*, vol.31(2): 155– 167.
26. *Plr*, R project — www r-project org, <http://cran.r-project.org/web/packages/stepPlr/>
27. *RPART package*, R project — www r-project org, <http://cran.rproject.org/web/packages/rpart/index.html>
28. *Naive Bayes*, Wikipedia, <http://en.wikipedia.org/wiki/Naive_bayes>
29. *klaR*, R project — www r-project org, <http://cran.rproject.org/web/packages/klaR/index.html>
30. *LDA*, Wikipedia, <http://en.wikipedia.org/wiki/Linear_discriminant_analysis>
31. Knudsen S, 2004, *Guide to Analysis of DNA Microarray Data, 2nd Edition*. Wiley-Liss.
32. *KNN*, Wikipedia, <http://en.wikipedia.org/wiki/KNN>
33. *random forest*, R project — www r-project org, <http://cran.rproject.org/web/packages/randomForest/index.html>
34. *Random forest — wikipedia*, Wikipedia, <http://en.wikipedia.org/wiki/Random_forest>
35. Diaz-Uriarte R and Alvarez de A S 2006, Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, vol.7: 3.
36. *QDA*, Wikipedia, <http://en.wikipedia.org/wiki/Quadratic_classifier#Quadratic_discriminant_analysis>
37. *caMassClass*, R project — www r-project org, <http://cran.rproject.org/web/packages/caMassClass/index.html>
38. *SVM*, Wikipedia, <http://en.wikipedia.org/wiki/Support_vector_machine>
39. *Boosting*, Wikipedia, <http://en.wikipedia.org/wiki/Boosting>
40. *varselrf*, R project — www r-project org, <http://cran.rproject.org/web/packages/varSelRF/index.html>
41. Diaz-Uriarte R, 2007, GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. *BMC Bioinformatics*, vol.8: 328.
42. *Accuracy*, Wikipedia, <http://en.wikipedia.org/wiki/Accuracy#Accuracy_in_binary_classification>
43. Kohavi R, *A study of cross-validation and bootstrap for accuracy estimation and model selection*, IJCAI, <http://web.cs.iastate.edu/~jtian/cs573/Papers/Kohavi-IJCAI-95.pdf>
44. *Jaccard*, Wikipedia, <http://en.wikipedia.org/wiki/Jaccard_index#Similarity_of_asymmetric_binary_attributes>
45. Mukherjee S, Golland P and Panchenko D, 2003, *Permutation tests for classification*, MIT, <http://cbcl.mit.edu/cbcl/publications/ai-publications/2003/AIM-2003-019.pdf>
46. *Shapiro-Wilk test*, Wikipedia, <http://en.wikipedia.org/wiki/Shapiro_wilk>
47. *Mann–Whitney U*, Wikipedia, <http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U>
48. *SVM in R*, R project — www r-project org, <http://cran.rproject.org/web/packages/e1071/e1071.pdf>
49. *Matthews correlation coefficient*, Wikipedia, <http://en.wikipedia.org/wiki/Matthews_Correlation_Coefficient>
50. Han S B, Moratz C, Huang N N, *et al.* 2005, Rgs1 and Gnai2 regulate the entrance of B lymphocytes into lymph nodes and B cell motility within lymph node follicles. *Immunity*, vol.22(3): 343–354.
51. Delston R B, Matatall K A, Sun Y, *et al.* 2011, p38 phosphorylates Rb on Ser567 by a novel, cell cycle-independent mechanism that triggers Rb-Hdm2 interaction and apoptosis. *Oncogene*, vol.30(5): 588–599.
52. Zhao W, Liu M and Kirkwood K L, 2008, p38alpha stabilizes interleukin-6 mRNA via multiple AU-rich elements. *Journal of Biological Chemistry*, vol.283(4): 1778–1785.
53. Wingenfeld K and Wolf O T, 2011, HPA axis alterations in mental disorders: Impact on memory and its relevance for therapeutic interventions. *CNS Neuroscience & Therapeutics*, vol.17(6): 714–722.
54. Garoflos E, Panagiotaropoulos T, Pondiki S, *et al.* 2005, Cellular mechanisms underlying the effects of an early experience on cognitive abilities and affective states. *Annals General Psychiatry*, vol.4(1): 8.
55. Kai S, Goto S, Tahara K, *et al.* 2004, Indoleamine 2,3-dioxygenase is necessary for cytolytic activity of natural killer cells. *Scandinavian Journal of Immunology*, vol.59(2): 177–182.
56. Muller N, Myint A M and Schwarz M J, 2011, Kynurenine pathway in schizophrenia: pathophysiological and therapeutic aspects. *Current Pharmaceutical*

Design, vol.17(2): 130–136.

57. van Zuiden M, Geuze E, Willemen H L, *et al.* 2011, Pre-existing high glucocorticoid receptor number predicting development of posttraumatic stress symptoms after military deployment. *The American Journal of Psychiatry*, vol.168(1): 89–96.
58. van Zuiden M, Heijnen C J, van de S R, *et al.* 2011, Cytokine production by leukocytes of military personnel with depressive symptoms after deployment to a combat-zone: a prospective, longitudinal study. *PLoS One*, vol.6(12): e29142.
59. van Zuiden M, Geuze E, Willemen H L, *et al.* 2012, Glucocorticoid receptor pathway components predict posttraumatic stress disorder symptom development: A prospective study. *Biological Psychiatry*, vol.71(4): 309–316.
60. 2010, Posters. *Basic & Clinical Pharmacology & Toxicology*, vol.107(s1): 162–692.
61. Sullivan P F, Fan C and Perou C M, 2006, Evaluating the comparability of gene expression in blood and brain. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol.141(3): 261–268.
62. Gladkevich A, Kauffman H F and Korf J, 2004, Lymphocytes as a neural probe: potential for studying psychiatric disorders. *Progress in Neuro-psychopharmacology & Biological Psychiatry*, vol.28(3): 559–576.