Exploring the biological basis of affective disorders

Mazin, Wiktor

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Wiktor Mazin

Exploring the biological basis of affective disorders

PhD Thesis
July 2008

Supervisor:

Erik Mosekilde, Deptartment of Physics,
Technical University of Denmark.

Co-supervisor:

Irina Antonijevic, Director, Translational Research,
Lundbeck Research USA.

Department of Physics
Technical University of Denmark

Summary

The molecular biology of affective disorders is still poorly understood. Affective disorders like major depression or post-traumatic stress disorder are known to be heterogeneous disorders and are believed to arise from a complex interplay of multiple genes and environmental triggering factors. Due to the heterogeneous and polygenic nature of these disorders, they are also difficult to treat effectively. In the case of depression, only as few as 20% of the affected individuals show full remission with the current antidepressants. This might be related to the fact that "*biomedical research in the field of psychiatry has remained focused on treatment targets that were identified*" more or less by chance "*half a century ago. Almost all available drugs target primarily monoamine transporter and receptors, in various combinations, leading to slightly different profiles. However, these differences rarely have a clinically relevant impact in terms of efficacy or safety. Moreover, the targets have not to this day proven to be at the core of the pathophysiology of the major psychiatric disorders*" (3).

Thus, there is clearly a need to better understand the biological basis of the complex psychiatric disorders and to identify biomarkers that are related not only to monoamine transporters and receptors. "*Biological markers were and still are focused on the brain, where the pathophysiology of affective disorders is thought only to occur. Although the brain certainly is a critical site to study the biology of mental disorders, there is increasing evidence for peripheral changes associated with"* affective disorders*. "Recently, multiple forms of blood markers as alternative to brain markers have received significant attention*" (3), for instance in post-traumatic stress disorder, bipolar disorder, and major depression.

A few years ago, Lundbeck Research USA initiated an exploratory study into whole blood biomarkers for affective disorders. What made this study of particular interest, apart from focusing on peripheral signatures, was that the same ~30 carefully selected genes were measured at the mRNA level in whole blood both in depressed patients, post-traumatic stress disorder patients, and borderline personality disorder patients as well as in various control groups in the US, Denmark, UK and Serbia.

In close cooperation with first Lundbeck Research USA and later also Lundbeck A/S in Denmark, I investigated the usability of these gene expressions as whole blood biomarkers in both borderline personality disorder and post-traumatic stress disorder through comparisons with healthy subjects. Among the major questions were:

- Would the psychopathology of the examined disorders be reflected in the selected whole blood gene expression profiles?

Would there be any consistent expression differences between the various groups?
- Which exploratory methods could be used to analyze such data, and what could be learnt from applying these methods? Would the different exploratory methods basically tell the same story or would they highlight different aspects of the disorders?

In short, the overall aim of the present thesis was thus to better characterize the molecular biology of these affective disorders by the use of various exploratory analyses of gene expression profiles in whole blood. Exploratory analyses comprised bioinformatics, statistical and classification methods, and was used to generate hypotheses about the studied disorders.

A main task was obviously to compare the expression profiles of controls to those of patients. Apart from the expression data, clinical data was also available for two control groups and from some of the borderline personality disorder patients. This enabled us to look into another more subtle task of trying to identify disease subtypes (phenotypes) and healthy subjects at risk for developing depression (intermediate phenotypes).

Some of the main findings of this thesis include support for the possibility of using gene expressions in whole blood as biomarkers for affective disorders. It is shown that the expression profiles of various control groups are more similar to each other, although not identical, than to the expression profiles of different patient groups.

A simulation study identifies the most promising classifiers and variable selection methods for separating the various control and patient groups. With these classifiers, predictions about a subject's status (control vs patient) can be made solely on the basis of the gene expression profiles. In addition, gene expressions are listed that separate control and patient groups. The genes are linked to biological functions, networks and pathways.

A range of promising statistical methods to analyse the expression data are identified as well. Each method offer new interpretations of the data like establishing hypotheses about gene expression – clinical variable relationships (generating hypotheses concerning both intermediate phenotypes and disease phenotypes), identifying possible gene expression disease subtypes or revealing the stability of the gene expressions measured in the UK control group at three different time points. Being an explorative study, validation is needed to confirm or rule out these findings.

Bioinformatics is used to predict new possible biomarkers based on the selected genes. I have also attempted to predict altered gene expressions in a patient group – bipolar disorder patients that so far has not been analyzed.

In perspectives for further studies, I propose an experiment to confirm or rule out temporal gene expression oscillations as large oscillations for a gene expression might mean that the gene expression is less suitable as a biomarker or at least more complicated to use. I list requirements for constructing a Bayesian gene regulatory network. With a Bayesian network, it might be possible to predict gene regulatory behaviour in whole blood in various affective disorders.
Also, suggestions are made for other classifier approaches and other ways of searching for blood biomarkers in affective disorders. Finally, I propose clustering simulations to identify the most promising clustering methods for disease subtyping.

Dansk resumé

Vi har i dag kun begrænset viden om de psykiske lidelsers molekylærbiologi. Lidelser som depression eller posttraumatisk stresssyndrom betragtes som inhomogene lidelser og menes opstået som følge af et kompleks sammenspil af flere gener og udløsende faktorer i omgivelserne. Den inhomogene og polygene natur af disse lidelser gør det svært at behandle dem effektivt. I tilfældet af depression, er det kun omkring 20% af de berørte personer, som bliver helt raske med de nuværende antidepressiver. Dette kan skyldes, at den *"biomedicinske forskning indenfor psykiatrien stadig fokuserer på behandlings-targets, som blev identificeret mere eller mindre tilfældigt for et halvt århundrede siden. Næsten alle tilgængelige medikamenter går primært efter monoamin transportere og receptorer, i forskellige kombinationer, ledende til minimalt forskellige profiler. Desværre har disse forskelle sjældent en klinisk relevant sikkerheds- eller virkningseffekt. Til dags dato har disse targets desuden ikke vist sig at være kernen i de psykiske lidelsers sygdomsfysiologi"* (oversat efter reference nr 3).

Der er derfor klart et behov for bedre at forstå den biologiske basis af de komplekse psykiske lidelser og for at identificere biomarkører, som ikke kun er relaterede til monoamin transportere og receptorer. *"Biologiske markører var og er stadig fokuseret på hjernen, hvor psykiske lidelsers sygdomsfysiologi kun menes at finde sted. Selv om hjernen helt sikkert er et centralt område, når det drejer sig om at studere psykiske sygdommes biologi, så er der tiltagende beviser for perifere ændringer ved psykiske lidelser. For nyligt har flere typer af blodmarkører som et alternativ til hjernemarkører fået betydelig opmærksomhed"* (oversat efter reference nr 3), for eksempel, i forbindelse med posttraumatisk stresssyndrom og med både bipolær og unipolær depression.

For nogle få år siden påbegyndte Lundbeck Research USA et eksplorativt studium af blodmarkører (whole blood) for psykiske lidelser. Det, der gjorde dette studium særligt interessant, bortset fra at det fokuserede på perifere blodmarkører, var, at det var de samme ~30 særligt udvalgte gener, som blev målt på mRNA niveau i blod både i deprimerede patienter, patienter med posttraumatisk stresssyndrom og borderline-personlighedsforstyrrede (grænsepsykotiske) patienter såvel som i forskellige kontrolgrupper fra USA, Danmark, England og Serbien.

I tæt samarbejde med først Lundbeck Research USA og siden også Lundbeck A/S i Danmark har jeg undersøgt anvendeligheden af disse genekspressioner som blodmarkører i både borderline-personlighedsforstyrrelse og posttraumatisk stresssyndrom ved sammenligninger med raske personer. Nogle af de store spørgsmål var:

- Ville de undersøgte lidelsers psykopatologi være reflekteret i de udvalgte genekspressionsprofiler i blod?
  Ville der være nogle konsistente ekspressionsforskelle mellem de forskellige grupper?
- Hvilke eksplorative metoder kunne anvendes til at analysere sådanne data, og hvad kunne der læres af at anvende disse metoder? Ville forskellige eksplorative metoder basalt set fortælle den samme historie eller ville de belyse forskellige aspekter af disse lidelser?

Kort fortalt er det overordnet mål med denne afhandling således at karakterisere forskellige psykiske lidelsers molekylærbiologi i større detaljer ved anvendelsen af forskellige eksplorative analyser af genekspressionsprofiler i blod. De anvendte eksplorative analyser omfattede bioinformatik, statistiske metoder og klassifikationsmetoder, og de blev benyttet til at opstille hypoteser om de undersøgte lidelser.

Det var oplagt at sammenligne kontrolpersoners ekspressionsprofiler med patienters ekspressionsprofiler. Bortset fra ekspressionsdata, var kliniske data også tilgængelige for to kontrolgrupper og fra nogle af de borderline-personlighedsforstyrrede patienter. Dette gjorde det muligt for os at se nærmere på en mere raffineret opgave, nemlig at forsøge at identificere sygdomsundertyper (fænotyper) og raske personer i fare for at udvikle en depression (mellemliggende fænotyper).

Nogle af denne afhandlings hovedresultater understøtter muligheden for at anvende genekspressioner i blod som biomarkører for psykiske lidelser. Det bliver vist, at forskellige kontrolgruppers ekspressionsprofiler, selvom de ikke er identiske, ligner hinanden mere end de ligner forskellige patienters ekpressionsprofiler.

Et simulationsstudie identificerer klassifikationsalgoritmer og variabel selektionsmetoder, som er mest lovende til at adskille de forskellige kontrol- og patientgrupper. Med disse klassifikationsalgoritmer kan der foretages forudsigelser om et individs status (såsom rask eller syg) alene på baggrund af genekspressionsprofilerne. Derudover angives de genekspressioner, som adskiller kontrol- og patientgrupper. Gener bliver koblet til biologiske funktioner, netværk og pathways.

Ligeledes identificeres en række lovende statistiske metoder til at analysere ekspressionsdata. Hver metode byder på nye tolkninger af data såsom at opstille hypoteser om genekspression-kliniske variable forhold (opstille hypoteser vedrørende både mellemliggende fænotyper og sygdomsundertyper), identificere mulige sygdomsundertyper kun ved hjælp af genekspressioner eller afsløre stabiliteten af genekspressioner målt i den

engelske kontrolgruppe på tre forskellige tidspunkter. Siden det er et eksplorativt studie, er en efterfølgende validering nødvendig for at bekræfte eller afvise disse indledende resultater.

Bioinformatik bliver anvendt til at forudsige mulige biomarkører på baggrund af de udvalgte gener. Jeg har også forsøgt at forudsige ændrede genekspressioner i en patientgruppe – maniodepressive (bipolær depression) patienter som endnu ikke er blevet analyseret.

I en diskussion af de videre perspektiver foreslår jeg et eksperiment til at bekræfte eller afvise tidslige genekspressionsoscillationer fordi store oscillationer for en genekspression kan betyde, at den pågældende genekspression er mindre egnet som biomarkør, eller i det mindste er kompliceret at anvende. Jeg opstiller kravene for at konstruere et Bayersk genreguleret netværk. Med et Bayersk netværk vil det være muligt at forudsige genreguleret adfærd i blod i forskellige psykiske lidelser.

Afhandlingen giver også forslag til andre klassifikationsmetoder og andre måder at søge efter blodmarkører på i psykiske lidelser. Endeligt foreslår jeg clustering simuleringer for at udvælge de mest lovende clustering metoder til at identificere sygdomsundertyper.

Acknowledgements

This thesis is an account of research undertaken between August 2005 and July 2008 at the Department of Physics, Technical University of Denmark (DTU) under the supervision of Prof. Erik Mosekilde. The work was performed in close cooperation with Lundbeck Research USA with Dr. Irina Antonijevic (MD) as co-supervisor.

There are many people I would like to thank for supporting me throughout this PhD. First and foremost, I would like to thank my supervisor Erik Mosekilde for encouraging and supporting me along the way in taking on the challenge it was to start up and continue the work with Lundbeck in a new field to me. I will never forget your wise words (translated from Danish) 'whenever a great opportunity turns up, you may not know what it can lead to, but still, you should grab it – you will never regret'. This has truly been the case for me with the close and good cooperation with Lundbeck in a new and exciting field of gene expressions and biology of affective disorders. I have never regretted this work and been highly motivated from the beginning.

In particular, I owe many thanks to my co-supervisor Dr. Irina Antonijevic and senior scientist Joseph Tamm, both from Lundbeck Research USA. Irina has been a driving force in the blood marker project with her enthusiasm, and she has come up with many ideas and suggestions for approaching the data. She has continuously passed on inspiring articles and showed great interest in the suggestions and results, I have presented during the years.

It has also been a great pleasure for me working closely with Joe for the entire duration of our co-operation. Joe's great experience in molecular biology and laboratory work as well as many suggestions and questions regarding analytical approaches truly gave me the feeling that my work and analytical approaches meant a lot to him and to Irina. Joe and I worked closely together first for a month in Valby in the summer of 2006 and later during my two weeks stay at Lundbeck Research USA early spring 2008. I think that the long and close cooperation with Joe has led to a friendship that will go beyond this thesis.

Furthermore, it has also been inspiring and a pleasure working with modelling scientist Jan Bastholm Vistisen from Lundbeck DK on the classification issues. Jan's suggestions in this field and great interest in my classification work led to 3-month simulation project described in the thesis.

I would also like to thank Frank Larsen from Lundbeck DK for presenting me with the opportunity of working in this field with Irina and Joe. If it had not

been for Frank's participation in a BioSim workshop and his contacts in Lundbeck, this PhD thesis would not have been written.

Especially in the beginning of this PhD, I received statistical assistance from the Statistical Consulting Center at the Department of Informatics and Mathematical Modeling (IMM) at DTU. In this connection I would like to thank both Rune Haubo and Merethe Hansen. Also, I benefited from multiple good comments about statistical issues with associate professor Bjarne Ersbøll at IMM. Furthermore, in the beginning of the collaboration with Lundbeck, Philip Hougaard from the Biostatistics Department in Lundbeck had many valuable suggestions as how to approach the data.

I have had the pleasure of attending several relevant courses at the Bioinformatics Group at the Center for Biological Sequence Analysis, DTU in particular the inspiring course 'DNA Microarray Analysis'. I also want to thank PhD student Qiyuan Li, associate professor Chris Workman, and professor Zoltan Szallasi for many productive discussions on bioinformatic and microarray analysis approaches.

Likewise, I am grateful for the very pleasant and inspiring stay at the University of Warwick in UK with Professor David Wild and Professor David Rand, member of the BioSim Network. The academic environment at the Systems Biology Center was very stimulating. I would like to thank them both for their interest in my Bayesian approach and many fruitful discussions.

I would like to thank both Joe and Erik many times for proofreading this manuscript. I appreciate your assistance and inspiration.

My last and warmest acknowledgements go to my family. Both my parents and parents in law have encouraged me along the entire PhD work. I am sorry for not having been better at laying the work aside when I was home, and thus not been enough present with my girlfriend and two sons who constantly reminded me that there is more to life than work and research. I hope you understand and forgive. Most importantly, I thank my partner and wonderful girlfriend, Annette, for her encouragement and loving support.

This work was supported 50% by the European Union through the Network of Excellence BioSim, Contract No. LSHB-CT-2004-005137, and 50% by DTU.

_____

Wiktor Mazin

July 30, 2008

Contents

1. Introduction

There is a growing need to understand the biological basis of affective disorders. For instance, only 20% (1) to 50% (2) of individuals with depression show full remission with the current antidepressants. However, given that affective disorders are believed to arise from interactions between environmental influences and the genetic makeup of an individual that is not an easy task.

To put the extent of people suffering from affective disorders into perspective (3) "*according to the National Institute of Mental Health (NIMH), mental disorders affect an estimated 26.2% of Americans aged 18 and older in a given year. Unlike many other chronic and disabling disorders, mental illnesses strike early in life*", with e.g. unipolar depression accounting for 28% "*of the disability from all medical causes in people aged 15-44 years*" (3) (4). "*These data are in line with the Global Burden of Disease study, reporting that mental illness, including suicide, accounts for over 15% of the burden of disease in established market economies. This is more than the disease burden caused by all cancers combined*" (3).
Despite this bleak assessment, "*biomedical research in the field of psychiatry has remained focused on treatment targets that were identified serendipitously more than half a century ago. Almost all available drugs target primarily monoamine transporter and receptors, in various combinations, leading to slightly different profiles. However, these differences rarely have a clinically relevant impact in terms of efficacy or safety. Moreover, the targets have not proven to be at the core of the pathophysiology of the major psychiatric disorders to this day, which may explain the modest efficacy of all available drugs when tested in poorly defined patient populations. This is in contrast to research into other major chronic diseases, such as cancer and heart disease, that has shed light on the biology and has resulted in the successful development of new treatment targets*" (3).

"*Biomedical research that focuses on the disease rather than on treatment may improve our understanding of core biological alterations associated with psychiatric disorders*" (3). "*A better understanding of the disease biology, and the biological differences among patients, should advance the diagnostic classification, which today is entirely descriptive. In doing this, one would also expect to identify new biomarkers that may yield more efficacious treatments, at least for subgroups of patients that share core biological disturbances*" (3) thus resulting in biological signatures of the various disorders.

"*Not surprisingly, the biological markers were and still are focused on the brain, where the pathophysiology of mental disorders is thought to occur. Although the brain certainly is a critical site to study the biology of mental*

disorders, there is increasing evidence for peripheral changes associated with mental disorders" (3) (5). "*Recently, multiple forms of blood markers as alternative to brain markers have received significant attention*" (3), for instance in post-traumatic stress disorder (6), (7) bipolar disorder (8), (9) and major depression (10), (11).

The rationale for studying affective disorders via peripheral blood can summed up as follows:

a) a disease is caused by (and/or results in) gene expression changes
b) diseases in one part of the body (brain) can result in gene expression changes elsewhere (blood)
c) blood represents a minimally invasive sampling option in humans

The overall aim of this thesis is to better understand the biology of different types of affective disorders by the use of various exploratory analyses of gene expression profiles in whole blood. Four psychiatric disorders are considered in the thesis: Borderline personality disorder, post-traumatic stress disorder (PTSD), depression and bipolar disorder. Gene expression profiles in borderline personality disorder and PTSD are analyzed together with the expression profiles from several cohorts of controls. Gene expressions are not analyzed from depressed patients or from bipolar disorder patients. However, as will be explained later, gene[1] predictions will be made for depressed and bipolar disorder patients. The results of all these various analyses will be presented.

What further makes this study very interesting is that the same 25-30 gene expressions are measured in different control and patient groups and hence, one important challenge is to see whether there are relative expression differences between the groups.

Being an exploratory study, focus has been on exploratory / hypothesis generating statistical and classification approaches that will be described in later chapters. For each applied statistical method strengths and weaknesses will be mentioned as well as the reason for using that method. However, I do not consider this as a thesis in statistics, so in general I will not go into statistical details but refer to appropriate literature for further details.

Overall, it can be said that the thesis is an intersection of the fields of biology of affective disorders and gene-environment interactions combined with statistical, machine learning and bioinformatics methods.

1.1 Lundbeck

During the entire thesis I have had a close cooperation with Lundbeck Research USA and towards the end of the thesis also with Lundbeck. Focus has

[1] The words "gene" and "gene expression" are used interchangeably.

been on solving relevant challenges as defined together with Dr Irina Antonijevic (MD), Director, Translational Research, Lundbeck Research USA, senior scientist Joseph Tamm, Lundbeck Research USA and mathematical modelling scientist at the clinical department of pharmacology Jan Vistisen, Lundbeck DK.

In order to get access to relevant data gathered by Lundbeck and Lundbeck's academic collaborators a contract has been signed. Parts of the thesis may be treated confidentially by Lundbeck and will be omitted from a public version[2]. After the contract was signed, the gene expression data was made available for analysis in portions as soon as Lundbeck had the data in-house and had obtained written consent from collaborators, if necessary. All data obtained came from controls and untreated patients that all had signed an informed consent stating the guidelines for handling the data.

Portions of this thesis have appeared in a book chapter called "Perspectives for an Integrated Biomarker Approach to Drug Discovery and Development" by Irina Antonijevic, Joseph Tamm, Wiktor Mazin, et al. (in press).

Furthermore, I have made a number of presentations at various BioSim conferences and workshops about the methods used and results obtained in an anonymous format.

1.2 Main findings

This study has shown that:

- Gene expression profiles in whole blood have the potential to be used as biomarkers for affective disorders (further validation is required).
    - The expression profiles of various control groups are more similar to each other, although not identical, than to the expression profiles of different patient groups.
    - Controls can be separated from borderline personality disorder patients based on differential expression of four genes: Gi2, GR, MAPK14 and partly MR (see section 8.1)
    - Controls can be separated from acute post-traumatic stress disorder patients by differential expression levels of ARRB2, ERK2 and RGS2.
    - Controls can not be separated from remitted post-traumatic stress disorder patients – a result that is in good agreement with the clinical diagnosis.

[2] At the end of July 2008, the Lundbeck legal department has accepted the thesis with minor corrections that I have implemented.

  - Controls may be separated from trauma patients without post-traumatic stress disorder by differential expression levels of ARRB2, CREB1, ERK2, IL-6 and partly MAPK8, Gs, MKP1, and MR (see section 8.1). The performance measure PPV (positive predictive value) in this case is not great, suggesting that the separation between these groups is not strong.
  - Controls, borderline disorder patients and acute post-traumatic stress disorder patients can all be separated from each other based on differential expression of four genes: ERK1, ERK2, GR and MKP1.

- A simulation study combined with results from actual data sets has shown that the most promising classifiers and variable selection methods for separating various control and patient groups are
  - support vector machines combined with variables selection based on random forests (both for 2-group and multiple group comparisons)
  - stepwise logistic regression (only for 2-group comparisons)
  - recursive partitioning (only for multiple group comparisons)

- The most promising statistical methods to analyse the data are
  - The univariate parametric t-test and non-parametric Wilcoxon test in the 2-group case as well as the ANOVA and Kruskal-Wallis test in the multiple group case.
  - Spearman correlations for pair wise gene expression comparisons.
  - Repeated measures ANOVA for identifying differences between multiple time points measurements.
  - For gene expression disease subtyping: hierarchical clustering and heat maps (validation is needed).
  - Canonical correlation analysis for gene expression-clinical variable relationships supplemented by the univariate tests (validation is needed).

- 20 hypotheses are constructed as gene expression predictions for depressed patients. These hypotheses are based on expression patterns from controls and identified possible intermediate phenotypes. The expression patterns observed in a small group of severely depressed patients confirms some of the hypotheses.

- Possible disease subtypes / phenotypes on the gene expression level may be identified with heat maps and canonical correlation analysis (validation is needed).

- Bioinformatics may be used to predict new possible biomarkers for depression such as Hsp90, PP2A, NFkB, Ras, MHC Class I, Mek, Akt and Ap1.

Finally, bioinformatics may be used to predict altered gene expressions in an as yet unanalyzed patient group – bipolar disorder patients - for: Gs, IL-1 beta, CREB1 and ERK1.

- Expression differences of the considered genes do exist between the genders, but these differences do not seem as significant as one might think.

- Three time point measurements (Day 0 at 8 am, Day 0 at 2 pm, and Day 1 at 8 am) indicate significant expression differences for CD8 beta, IL-8, MKP1, MR and ODC1 between the time points. Validation is needed (see section 9.1).

2. Four psychiatric disorders – their symptoms, phenotypes and genetic background

This chapter provides an introductory description of the four psychiatric disorders – depression, borderline personality disorder, post-traumatic stress disorder (PTSD) and bipolar disorder - that I consider in the present thesis. These heterogeneous disorders are discussed from both a symptom perspective based on the descriptions in 'Diagnostic and Statistical Manual of Mental Disorders, 4th text revision' (DSM-IV-TR) and a gene perspective, listing the genes that are assumed to be associated with each disorder according to the literature and the online database OMIM. All information in this chapter is collected from publicly available sources. In order to illustrate the potential of the gene expression approach used, I have included a ground-breaking example of using gene expressions in blood to differentiate between individuals who developed PTSD and those who did not following a traumatic event. This study was performed by Segman et al. from Israel and considers trauma survivors who were admitted to the emergency room immediately following a traumatic event (6).

Three of the main depression hypotheses are described: The monoamine hypothesis involving a neurotransmitter such as serotonin, the hypothalamic-pituitary-adrenal (HPA) axis hypothesis involving stress and the stress hormone cortisol, and the cytokine hypothesis involving the action of pro- and anti-inflammatory cytokines in cell signaling. The three hypotheses are considered as different aspects of a single broader hypothesis at the end of the chapter where I also describe possible shared biological mechanisms between the above disorders.

Hereafter, common aspects of the psychiatric disorders are shortly described;

- *phenotype and intermediate phenotype challenges. A disease phenotype or disease subtype may be a specific symptom cluster that consists of parts of the symptoms present in a disorder. This aspect is very important as the psychiatric disorders are heterogeneous diseases with no single well-defined disease phenotype capturing all the various symptoms present.*

  Intermediate phenotypes are here understood as appearing in normal subjects who are or may be at risk to develop a disorder. These phenotypes involve some of the symptoms observed in acutely ill patients. Examination of intermediate phenotypes may represent an important step towards an improved understanding of the biology of psychiatric disorders.
- *gene-environment interactions. The significance and complexity of such interactions are clearly witnessed by the facts that mental disorders have environmental causes and that people show significant variability in their response to those causes. Environmental risk factors such as substance abuse during pregnancy, premature parental loss, and exposure to*

family conflict and violence are mentioned.
Gene-environment factors are finally discussed in the context of constructing the ultimate goal - a 'Mendeleyev table' of (genetic and non-genetic) phenotypes to help understand the 'atomic structure' underlying these complex disorders.

In order to obtain the personality traits of each disorder according to DSM-IV-TR, I have used the renowned database eMedicine.com. This database contains one of the largest and most current clinical knowledge bases available to physicians and other healthcare professionals.
DSM-IV-TR is the latest version of the "Diagnostic and Statistical Manual of Mental Disorders (DSM)" which is an American handbook for mental health professionals that lists different categories of mental disorders and the criteria for diagnosing them, according to the publishing organization the *American Psychiatric Association* (12).

Each of the following four sections mentions genes[3] associated with different affective disorders as listed in some of the literature (see the respective section for references) and in the online public available database OMIM – "Online Mendelian Inheritance in Man". The database catalogues all known diseases with a genetic component, links them to the relevant genes in the human genome, whenever possible, and provides appropriate references. OMIM is developed for the World Wide Web by NCBI, the National Center for Biotechnology Information.

Even though expression data from depressed patients was not analyzed directly in this thesis in accordance with our agreement with Lundbeck, the section on depression will be relatively more thorough than the other disorder sections. This is partly because treatment of depression is one of Lundbeck's key focus areas, and partly because it was the original intention to compare expression levels of depressed patients with the expression levels of patients suffering from the other/related disorders. Thus, the other disorders can be said to be benchmarked against depression.

2.1 Depression

Depression is a broad term for a heterogeneous disease that comes in different forms. According to the National Institute for Mental Health - NIMH (13), the most common forms are dysthymia (a less severe type of depression, but chronic form with a typical duration of more than two years) and major depression, also known as unipolar depression. Other types include psychotic depression, postpartum depression, seasonal affective disorder (SAD) and

[3] Listed genes will come from both genetic and gene expression studies. SNPs (single nucleotide polymorphisms - a variation in a gene caused by the change of a single base in DNA) in genes are believed to result in altered gene expressions.

bipolar disorder. Bipolar disorder will be treated separately later in this chapter. The focus in this section is on 'major depression'.

"*Major depression is characterized by a combination of symptoms that interfere with a person's ability to work, sleep, study, eat, and enjoy once–pleasurable activities. Major depression is disabling and prevents a person from functioning normally. An episode of major depression may occur only once in a person's lifetime, but more often, once experienced it recurs throughout a patient's life."* (13)

The DSM-IV-TR defines a major depressive episode (14) as a syndrome in which, during the same 2-week period, are at least 5 of the following symptoms present and manifest themselves as a change from a previous state of well-functioning (moreover, the symptoms "*must include either (a) or (b)):*

(a) Depressed mood
(b) Diminished interest or pleasure
(c) Significant weight loss or gain
(d) Insomnia or hypersomnia
(e) Psychomotor agitation or retardation
(f) Fatigue or loss of energy
(g) Feelings of worthlessness
(h) Diminished ability to think or concentrate; indecisiveness
(i) *Recurrent thoughts of death, suicidal ideation, suicide attempt, or specific plan for suicide*"

Depending upon the number and severity of the symptoms, a depressive episode may be specified as mild, moderate or severe. Women are twice as likely to experience depression as men.

DSM-IV-TR further includes descriptions of symptoms that must be present in various subtypes of depression. Depression can be noted to be with or without psychotic symptoms and may have melancholic or catatonic features or be classified as an atypical depression. Here it is not important to go into the details of each subtype, but just to stress that we are dealing with a clinically very heterogeneous disease. We may expect the gene expression profiles of depressed patients to reflect this heterogeneity. If it will be possible to define these different profiles, they can be used to better classify patients and tailor the development of drugs to these subtypes.

Listing the symptoms also plays an important role for some of the first analyses done in this thesis. Here clinical information from controls was combined with their gene expression profiles. This was done to predict expression response in depressed patients. For more details, see the results chapter.

2.1.1 Depression hypotheses

According to NIMH, "*there is no single cause of depression. Rather, likely to result from a combination of genetic, biochemical, environmental, and psychological factors*" (15). "*Some types of depression tend to run in families, suggesting a genetic link. However, depression can occur in people without family histories of depression as well. Genetics research indicates that risk for depression results from the influence of multiple genes acting together with environmental or other factors*". (15)
"*In addition, trauma, loss of a loved one, a difficult relationship, or a particularly stressful situation may trigger a depressive episode. Subsequent depressive episodes may occur with or without an obvious trigger*" (15).

Given the fact that no single gene causes depression and that it's a heterogeneous disease, several depression hypotheses exist. Some of the most common are the monoamine, the HPA-axis and the cytokine hypothesis which will briefly be touched upon below. Other hypotheses, like the neurogenesis hypothesis, exist but will not be described in detail here.

The monoamine hypothesis
Serotonin is a monoamine neurotransmitter and the monoamine hypothesis claims that low levels of serotonin cause depression. In addition, the hypothesis explains why antidepressants take about 6-8 weeks to work.
In order to relieve symptoms in depressed patients the levels of serotonin have to be raised. This can be done by e.g. SSRI antidepressants. They block the reuptake of serotonin thereby causing a temporary increase in the level of serotonin. However, because a negative feedback exists that counterworks the increased neurotransmitter level, see figure 1, it takes 6-8 weeks to normalize serotonin levels and relief symptoms.

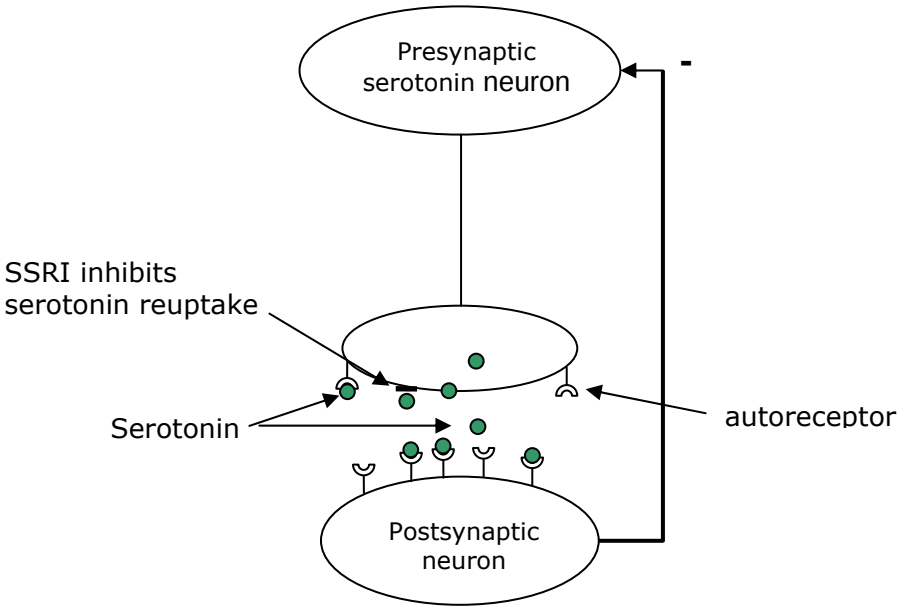


Figure 1: Illustration of the monoamine hypothesis. Increase of serotonin causes a negative feedback mechanism that reduces serotonin firing and then long-term treatment desensitizes the inhibitory serotonin presynaptic autoreceptors and transmission is enhanced.

HPA-axis (Hypothalamic-Pituitary-Adrenal) hypothesis

Stress causes the elevation of cortisol and long-term elevation of cortisol may cause depression. More precisely, long term exposure to stress causes excitatory drive of the hypothalamus. Thereby, the level of CRH (corticotropin releasing hormone) is increased and causes a sustained release of ACTH (adrenocorticotropic hormone) from the pituitary. This, in turn, increases the level of cortisol from the adrenal gland. The negative feedback loop, see figure 2, is impaired in depressed patients which causes a reduction of brain corticosteroid receptors. The end result is elevated levels of the stress hormone cortisol which, thus, may cause depression.

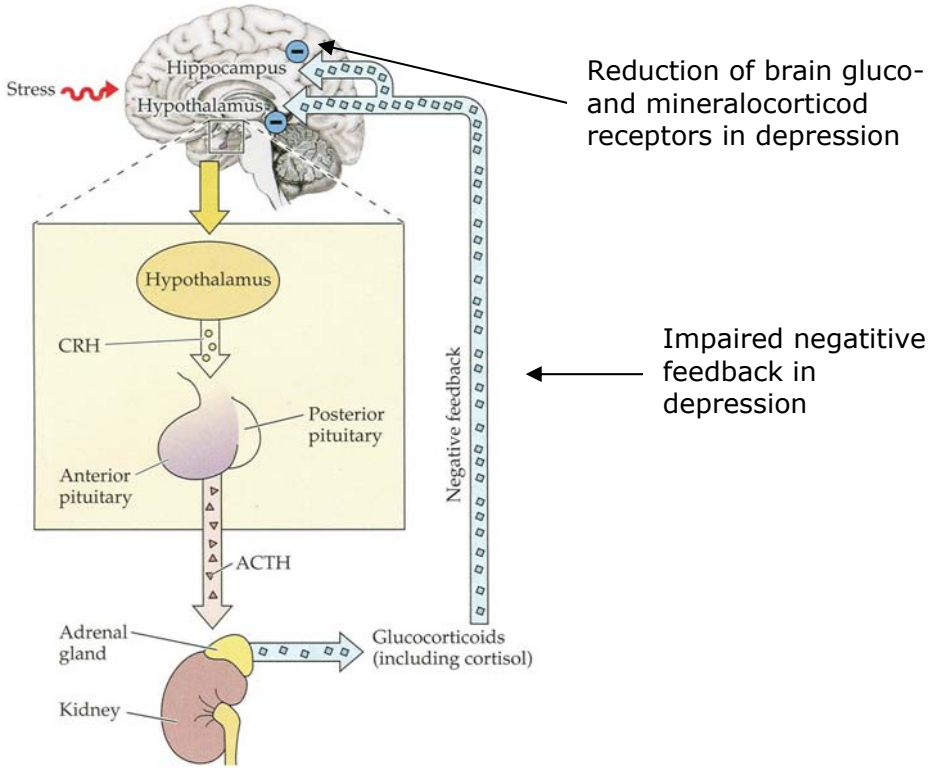


Figure 2: Illustration of the HPA-axis hypothesis. "*The hypothalamus is subject to abnormal excitatory drive from limbic system regions due to prolonged stress, resulting in sustained release of ACTH. Depression causes a reduction of brain corticosteroid receptors, resulting in subnormal negative feedback in this system and thus, elevated levels of the stress hormone cortisol*". Source: (16).

Interactions have been demonstrated (17) between the serotonergic system and the HPA axis. Cortisol may lower serotonin levels and conversely, serotonin stimulates secretion of CRH and ACTH and may modulate negative feedback of the HPA axis by glucocorticoids.

The cytokine hypothesis

The cytokine hypothesis of depression posits that depression is caused by the actions of cytokines (18). Cytokines are proteins and peptides that are used for cell signaling. They are similar in action to hormones and neurotransmitters and are sometimes loosely described as immune system hormones (19).

Some of the well-known (from the literature) pro-inflammatory cytokines are IL-1β (interleukin-1 beta), IL-6 and TNF-α (tumor necrosis factor-α), while some anti-inflammatory cytokines are IL-4, IL-10, and TNF-β (tumor necrosis factor-β).

There are many pathways known to be involved in the pathophysiology of depression that are influenced by cytokines, like the IL-6 and the IL-10 signaling pathways. Cytokines also influence neurotransmitter and HPA-axis function (20), thus creating a link to the two previous mentioned hypotheses.

2.1.2 Depression and genes

Table 1 lists genes considered to be associated with major depression together with references from the literature and OMIM. The table is not based on an exhaustive literature search but reflects the literature I have dealt with at the beginning of the thesis work. The purpose of listing the genes is to note any possible overlap with the list of genes selected by Lundbeck, see next chapter[4].

Genes associated with major depression	**Reference**
Involved in the action of monoamine neurotransmitters[5]:	
• SERT (serotonin transporter gene) encodes an integral membrane protein that transports the neurotransmitter serotonin (monoamine transporter) from synaptic spaces into presynaptic neurons and has long been associated with depression.	(10), (21), (22), (23)
• TPH1 (tryptophan hydroxylase-1)	(22)
• TPH2 (tryptophan hydroxylase-2)	(22)
• HTR1A (serotonin 5-HT-1A receptor)	(24)
• HTR2A (serotonin 5-HT-2A receptor)	(22)
• IDO (indoleamine 2,3-dioxygenase) catalyzes the rate-limiting step of tryptophan conversion.	(20)
• DRD4 (dopamine D4 receptor).	(22)
Involved in hypothalamic-pituitary-adrenal (HPA) axis regulation[5]:	
• FKBP5 (fk506-binding protein 5) plays a role in the stress hormone-regulating HPA axis.	(22)
• CREB1 (cAMP response element-binding protein 1)	(22), (25)
• CRH (corticotrophin-releasing hormone gene)	(20), (26)
• GR (glucocorticoid receptor)	(26), (27)
• MAPK14 (mitogen-activated protein kinase 14)	(20)
• MR (mineralocorticoid receptor)	(26), (27)
Involved in cytokine / immune system regulation[5]:	
• IL-1β (interleukin-1 beta) is a pro-inflammatory cytokine.	(24), (28)
• IL-6 (interleukin-6) is also a pro-inflammatory cytokine.	(24)
• IL-12 (interleukin-12)	(20)
• TNF-α (tumor necrosis factor alpha) is also a pro-inflammatory cytokine.	(20)
BCR (breakpoint cluster region)	(22)
CHRM2 (cholinergic receptor, muscarinic, 2)	(22)
DYT1 (dystonia 1, torsion, autosomal dominant)	(22)
ERK1 (extracellular signal-regulated kinase 1)	(29)
ERK2 (extracellular signal-regulated kinase 2)	(29)
MTHFR (methylenetetrahydrofolate reductase)	(22)
P2RX7 (purinergic receptor P2X, ligand-gated ion channel, 7)	(30)

Table 1: Genes associated with major depression

[4] The next chapter will also state the direction of the gene expression changes, that is, whether the expression is up- or down regulated.

[5] Certainly, some genes can play a role in e.g. both HPA-axis, cytokine and monoamine activity. Here a split is made according to descriptions of each gene's activity in OMIM and the literature and to highlight the three described depression hypotheses.

2.2 Borderline Personality Disorder (BPD)

According to NIMH, *Borderline Personality Disorder* (BPD) (31), in short borderline, "*is a serious mental illness characterized by pervasive instability in moods, interpersonal relationships, self-image, and behavior*" often in combination with pronounced "black and white" thinking. "*This instability often disrupts family and work life, long-term planning, and the individual's sense of self-identity. Originally thought to be at the "borderline" between psychosis* (a major mental disorder characterized by gross impairment of a person's perception of reality and ability to communicate and relate to others) *and neurosis* (a mental disorder characterized primarily by anxiety), *people with BPD suffer from a disorder of emotion regulation*."

The DSM-IV-TR defines BPD (32)[6] as "*a pervasive pattern of instability of interpersonal relationships, self-image and affects, as well as marked impulsivity, beginning by early adulthood and present in a variety of contexts*. *A DSM diagnosis of BPD requires any five out of nine listed criteria to be present for a significant period of time. The criteria are:*

- *(a) Frantic efforts to avoid real or imagined abandonment. [Not including suicidal or self-mutilating behavior covered in Criterion (e)]*
- *(b) A pattern of unstable and intense interpersonal relationships characterized by alternating between extremes of idealization and devaluation.*
- *(c) Identity disturbance: markedly and persistently unstable self-image or sense of self.*
- *(d) Impulsivity in at least two areas that are potentially self-damaging (e.g., promiscuous sex, eating disorders, binge eating, substance abuse, reckless driving). [Again, not including suicidal or self-mutilating behavior covered in Criterion (e)]*
- *(e) Recurrent suicidal behavior, gestures, threats, or self-mutilating behavior such as cutting, interfering with the healing of scars, or picking at oneself.*
- *(f) Affective instability due to a marked reactivity of mood (e.g., intense episodic dysphoria, irritability, or anxiety usually lasting a few hours and only rarely more than a few days).*
- *(g) Chronic feelings of emptiness, worthlessness.*
- *(h) Inappropriate anger or difficulty controlling anger (e.g., frequent displays of temper, constant anger, recurrent physical fights).*
- *(i) Transient, stress-related paranoid ideation or severe dissociative symptoms."*

Worth noticing in the above list are points (a), (f) and (i) which indicate the presence of stress-related responses, and thus, the involvement of disturbances in the HPA-axis according the HPA-axis hypothesis.

[6] eMedicine.com do not contain a specified list of BPD criteria. Wikipedia (includes references) does and is used here.

The most consistent finding in the search for causation in this disorder is a history of childhood trauma, although some researchers have suggested a genetic predisposition (32). Neurobiological research has highlighted some abnormalities in serotonin metabolism which indicates a link to the monoamine hypothesis.

The incidence has been calculated as 1-2% of the population by NIMH (31), with women three times more likely to suffer the disorder. The expression data that was analyzed in this thesis from BPD patients came mainly from women.

2.2.1 Borderline Personality Disorder and genes

In general, the literature on genes associated with BPD is very sparse, and OMIM has no record dealing with borderline. Hence, I will not show the few results in a table format. The biological underpinning of BPD is complex and poorly understood. Previous studies have emphasized the aminergic neurotransmission with serotonin and dopamine (33), (34) and HPA-axis hyperactivity in relation to the pathophysiology of BPD (35). BPD does not consist of impairment in a single neurotransmitter system. Besides aminergic neurotransmission, glucocorticoid and NMDA neurotransmission may play a part of the pathophysiology of BPD (35).

Finally, genes involved in the production of MAOA (monoamine oxidase-A) may also be involved in the development of BPD. The MAOA gene encodes for the enzyme MAOA, a potent metabolizer of serotonin and dopamine (34).

To sum up, it seems like some of the same mechanisms with neurotransmitter and HPA-axis activity are present in BPD as is the case in major depression. In young women, depression is often comorbid with BPD and here cytokines play a role (36).

2.3 Post-Traumatic Stress Disorder (PTSD)

Post-traumatic stress disorder (PTSD) "*was first brought to public attention in relation to war veterans, but it can result from a variety of traumatic incidents, such as mugging, rape, torture, being kidnapped or held captive, child abuse, car accidents, train wrecks, plane crashes, bombings, or natural disasters such as floods or earthquakes*" (37). The last couple of years PTSD has received more attention in Denmark due to the wars in former Yugoslavia, and in Afghanistan and Iraq with both refugees and soldiers affected.

NIMH (38) characterizes PTSD as "*an anxiety disorder that some people develop after seeing or living through an event that caused or threatened serious harm or death. Symptoms usually begin within 3 months of the*

incident but occasionally emerge years afterward. Symptoms include flashbacks or bad dreams, emotional numbness, intense guilt or worry, angry outbursts, feeling "on edge," or avoiding thoughts and situations that remind of the trauma", see the DSM-IV-TR description below. Not every traumatized person develops full-blown or even minor PTSD. The course of the illness varies. Some people recover within 6 months, while others have symptoms that last much longer. In some people, the condition becomes chronic.

DSM-IV-TR (39) has six criteria that has to be met for a person to be diagnosed with PTSD. Summarized, they are (40):

- *(a) "Exposure to a traumatic event (see below)*
- *(b) Persistent reexperience*
- *(c) Persistent avoidance of stimuli associated with the trauma*
- *(d) Persistent symptoms of increased arousal (e.g. difficulty falling or staying asleep)*
- *(e) Duration of symptoms more than 1 month*
- *(f) Significant impairment in social, occupational, or other important areas of functioning*

Criterion (a) (the "stressor") consists of two parts, both of which must apply for a diagnosis of PTSD. The first (a1) requires that "the person experienced, witnessed, or was confronted with an event or events that involved actual or threatened death or serious injury, or a threat to the physical integrity of self or others." The second (a2) requires that "the person's response involved intense fear, helplessness, or horror."

Almost all of the above criteria indicate a (severe) stress response and thus, the involvement of disturbances in the HPA-axis.

As can be seen from this introductory description of PTSD, environmental factors play a large role in the development of PTSD. However, genetic components are also associated with PTSD (see next section) as it is known that PTSD runs in families (40).

"The estimated lifetime prevalence of PTSD among adult Americans is 7.8%, with women (10.4%) twice as likely as men (5%) to have PTSD at some point in their lives" (40). The expression data that was analyzed in this thesis from PTSD patients were from men only and all related to war events.

2.3.1 PTSD and genes

In general, the literature on PTSD and associated genes is not abundant[7], and OMIM has no record dealing with this disorder either.

The literature indicates that serotonin might be implicated in PTSD (41). The same is true for DRD2 (dopamine receptor D2) (42) and DAT (dopamine transporter gene) (43). A recent paper (44) also points to FKBP5 (FK binding protein 5), CRH (corticotropin-releasing hormone), NET1 (noradrenaline transporter), COMT (catechol-o-methyltransferase) and GRP (gastrin-releasing peptide receptor) . Another recent paper (45) presents a search of literature that has looked at association studies involving candidate genes in the serotonin (5-HTT), dopamine (DRD2, DAT), glucocorticoid (GR), GABA (GABRB), apolipoprotein systems (APOE2), brain-derived neurotrophic factor (BDNF) and neuropeptide Y (NPY). "*The studies have produced inconsistent results, many of which may be attributable to methodological shortcomings and insufficient statistical power. They conclude that the complex etiology of PTSD, for which experiencing a traumatic event forms a necessary condition, makes it difficult to identify specific genes that substantially contribute to the disorder*".

The same paper (45) mentions the possible involvement of the HPA-axis. Since increased HPA activity is a natural reaction to stress, researchers have extensively explored this in PTSD patients. "*While traditional models of stress would predict overactivity of the HPA axis, paradoxically there is substantial evidence for decreased HPA-axis activity in patients with PTSD. It is unknown whether these HPA anomalies are caused by, or result from, PTSD*" (46). Furthermore, several papers report links between PTSD and the cytokines IL-1beta (interleukin 1-beta) (47) and IL-6 (interleukin 6) (48), (49).

At the gene expression level, a recent study by Segman and colleagues (6), based on a microarray study, identifies several hundred promising genes associated with PTSD. These authors observed peripheral blood mononuclear cell gene expression profiles in individuals seen in the emergency department shortly after a traumatic event and followed one and four months later. They found that gene expression signatures differentiated between individuals who developed PTSD and those who did not. They found that several differentiating genes were previously described as having a role in immune activation like CD2 (cluster of differentiation 2) and IL-8 (interleukin-8), and stress response/HPA-axis activity like ADM (adrenomedullin) and FKBP5 (fk506-binding protein 5). Many more genes are listed in the article.

[7] "*One reason for this lack of attention might have to do with the fact that PTSD it is a relatively new diagnosis and that, until the 1990s, it was commonly thought to be prevalent only among specific subpopulations (e.g., Vietnam War veterans) and rare in the general population. This misconception was corrected with the publication of several epidemiologic studies of trauma exposure and PTSD. These studies consistently demonstrated that both exposure to traumatic events and PTSD are common*" (36).

Even due to inconsistent results, it may seem like some of the same mechanisms with HPA-axis activity and involvement of cytokines and neurotransmitters are present in PTSD as is the case in major depression and BPD.

2.4 Bipolar Disorder (BD)

As it is the case for depression and the other affective disorders I have described, *bipolar disorder* (BD) is not a homogenous disease and has been divided into several subcategories. They are called "*bipolar I, bipolar II and cyclothymia based on the type and severity of mood episodes experienced*" (50).

According to NIMH (51), "*bipolar disorder, also known as manic-depressive illness, is a brain disorder that causes unusual shifts in a person's mood, energy, and ability to function. Different from the normal ups and downs that everyone goes through, the symptoms of bipolar disorder are severe. They can result in damaged relationships, poor job or school performance, and even suicide"*.

"*Episodes of mania and depression typically recur across the life span. Between episodes, most people with bipolar disorder are free of symptoms, but as many as one-third of people have some residual symptoms*" (51).

According to DSM-IV-TR (52), "*manic episodes are characterized by the following symptoms:*

- *(a) At least 1 week of profound mood disturbance is present, characterized by elation, irritability, or expansiveness.*
- *(b) Three or more of the following symptoms are present:*
  - *a. Grandiosity*
  - *b. Diminished need for sleep*
  - *c. Excessive talking or pressured speech*
  - *d. Racing thoughts or flight of ideas*
  - *e. Clear evidence of distractibility*
  - *f. Increased level of goal-focused activity at home, at work, or sexually*
  - *g. Excessive pleasurable activities, often with painful consequences*
- *(c) The mood disturbance is sufficient to cause impairment at work or danger to the patient or others.*
- *(d) The mood is not the result of substance abuse or a medical condition."*

DSM-IV-TR depressive episodes have already been described in a previous section of this chapter and apply to BD as well.

DSM-IV-TR further contains specifies hypomanic (a mild to moderate level of mania) and mixed episodes (symptoms of mania and depression occur simultaneously), however, they will not be described here.

Based on the DSM symptoms described above it seems that disturbances in the HPA-axis are involved in BD, as it is also the case for depression.

Studies suggest that genetics, early environment, neurobiology, and psychological and social processes are important contributory factors (50).

"*In any given year about 5.7 million American adults or about 2.6 percent of the population age 18 and older, have bipolar disorder*" (51)

2.4.1 Bipolar disorder and genes

As with the other affective disorders, bipolar disorder is a genetically heterogeneous complex trait.

Table 2 lists some of the genes associated with bipolar disorder together with references from recent literature and OMIM.

Genes associated with bipolar disorder	**Reference**
Involved in the action of monoamine neurotransmitters[8]:	
• SERT (serotonin transporter)	(53)
• HTR4 (serotonin 5-HT-4A receptor)	(53)
• DAT1 (dopamine transporter gene)	(53)
• DRD4 (dopamine D4 receptor)	(53)
• COMT (catechol-O-methyltransferase)	(53)
• GPR50 (G protein-coupled receptor 50)	(53)
• TPH2 (tryptophan hydroxylase-2)	(54)
Involved in hypothalamic-pituitary-adrenal (HPA) axis regulation[8]:	
• CRHR2 (Corticotropin-releasing hormone receptor 2)	(55)
• GR (glucocorticoid receptor)	(56)
• SEF2-1B (transcription factor 4)	(53)
Involved in cytokine / immune system regulation[8]:	
• IL-4 (interleukin 4)	(57)
• IL-6 (interleukin 6)	(57)
• TNF-α (tumor necrosis factor alpha)	(57)
BDNF (brain-derived neurotrophic factor)	(53)
BCR (breakpoint cluster region)	(53)
CACNA1C (calcium channel, voltage-dependent, L-type, alpha 1C subunit)	(58)
CLOCK (circadian locomotor output cycles kaput)	(52)

[8] As with depression, some genes can play a role in e.g. both HPA-axis, cytokine and monoamine activity. Here a split is made according to descriptions of each gene's activity in OMIM and the literature and to highlight the three described depression hypotheses.

CUX2 (cut-like 2)	(53)
DFNB31	(59)
DGKH (diacylglycerol kinase eta)	(59)
DISC1 (distrupted in schizophrenia 1)	(54)
EGFR (epidermal growth factor receptor)	(58)
GRIN2B (NMDA glutamate receptor, subunit 2B)	(54)
MTHFR (5,10 methylenetetrahydrofolate reductase)	(54)
MTND1 (complex I, subunit ND1)	(53)
MYO5B (myosin VB)	(58)
NXN (encodes the protein nucleoredoxin)	(59)
SORCS2 (SORCS receptor 2)	(59)
TRPM2 (transient receptor potential cation channel, subfamily m, member 2)	(53)
TSPAN8 (tetraspanin 8)	(58)
VGCNL1 (voltage gated channel like 1)	(59)

Table 2: Genes associated with bipolar disorder

Additional genes can be found in two recent papers dealing with gene expression (microarray) analysis in BD (60) and molecular genetics in bipolar disorder and depression (61).

Since BD contains depressive episodes, it is not surprising to find some of the same genes listed in table 2 as in table 1 (major depression). In particular, SERT, DRD4, TPH2, IL-6, TNF-α and BCR are mentioned in both tables.

Last year, the Wellcome Trust Case Control Consortium (WTCC) performed genome-wide association studies to identify genes involved in common human diseases, among them, bipolar disorder. In the British population, they examined ~2000 bipolar patients and ~3000 controls (62). In the results chapter, bioinformatics approaches will be used to link the genes inferred from the WTCC SNP data to the genes selected by Lundbeck.

2.5 Phenotypes and intermediate phenotypes

A disease phenotype (also known as an endophenotype, see the next section, or disease subtype) may be defined as a specific symptom cluster that consists of parts of the symptoms present in each disorder or be present across current diagnostic boundaries (3).

The DSM-IV-TR descriptions for the four affective disorders above clearly show that we are dealing with heterogeneous diseases and thus, no single well-defined disease phenotype captures all the various symptoms present in a psychiatric disorder. More likely, each disorder consists of a number of or perhaps even a whole spectrum of disease phenotypes. This makes it difficult (and sometimes even impossible) not only to decide on the right treatment for the individual patient but also to replicate genetic and gene expression findings in different studies and trials. To complicate things further, an affected

individual may suffer from several psychiatric disorders (comorbidity). For instance, borderline personality disorder often occurs together with mood disorders like depression or bipolar disorder. Some features of borderline personality disorder may overlap with those of mood disorders, pointing to the relevance of disease phenotypes.

In 2011, the American Psychiatric Association is planning to publish DSM-V with the aim to develop "*an etiologically based, scientifically sound (diagnostic) classification system*" (3). In order to achieve this goal, a better understanding of the biological basis of psychiatric disorders seems to be a necessary first step. Linking disease phenotypes rather than an entire disorder to biological findings like transcription profiles should be a viable approach and help to uncover the biology of distinct phenotypes. "*This point may be of particular importance when one aims to address early onset of disorders or to initiate prophylactic treatments. As the specific symptoms an individual develops depend on the individual's genetic makeup and the environmental context, objective markers that allow recognition of phenotypes associated with increased vulnerability will help select individuals for prophylactic treatment*" (3). It is here the concept of intermediate phenotypes is useful. An intermediate phenotype as defined in this thesis appears in normal subjects who are or may be at risk to develop a disorder and consists of some of the symptoms observed in acutely ill patients. "*It may be an important step towards an improved understanding of the biology of psychiatric disorders, since there is likely to be a continuum between completely healthy individuals and those with a clinically manifest psychiatric disorder*" (3). One can further imagine that the number of intermediate phenotypes will be higher in patients than in controls at risk, and when adding up, will lead to the diagnosis.

"*The importance of intermediate phenotypes is emphasized by the discussion put forward in the DSM-V research agenda on a dimensional vs categorical classification system. The dimensional approach includes an aspect often disregarded in psychiatric biomedical research, namely the examination of control populations*" (3). In the results chapter, results will be provided in support of the relevance of intermediate phenotypes by showing association between transcription patterns and psychiatrically relevant clinical variables in a control population.

"*Moreover, intermediate phenotypes seem particularly relevant for drug development, as examination of drug effects in such 'control' subjects could provide early signs for efficacy in a patient population*" (3).

"*An extension of the above approach is to address the biology of distinct clinical features across the boundaries of current diagnoses. The analysis of such complex relationships should help to characterize multiple intermediate phenotypes, which in turn may predispose for the development of certain psychiatric diseases, e.g. when exposed to environmental stressors. Examples*

in psychiatry include (3) impaired cognitive executive function, which can occur in schizophrenia, some forms of depression and in connection with substance abuse. Another example is fatigue that can occur in different psychiatric disorders such as depression and anxiety, but also in disorders associated with a high incidence of depressive disorders, such as Parkinson's disease, multiple sclerosis and obesity" (3).

Apart from the examples above, there are several suggestions in the literature as to how to define disease phenotypes. Hasler et al. (63) propose endophenotypes for major depression at two levels: Psychopathological endophenotypes comprising e.g. impaired learning and memory and impaired diurnal variation and biological endophenotypes like REM sleep abnormalities and functional and structural brain abnormalities. For bipolar disorder some of the same phenotypes are suggested plus additional ones, see (64), (65).

2.6 Gene-environment interactions

As evident from the DSM-IV-TR descriptions and genes associated with each disorder, much research goes into establishing a connection between clinical symptoms and putatively associated genes (33), (66), (67). As explained in (66), the gene-environment interaction approach has grown out of two observations: first, that mental disorders have environmental causes; second, that people show heterogeneity in their response to those causes.

In the same paper (66), "*environment risk factors for mental disorders defined up to 2006 include (but are not limited to) maternal stress during pregnancy, maternal substance abuse during pregnancy, low birth weight, birth complications, deprivation of normal parental care during infancy, childhood physical maltreatment, childhood neglect, premature parental loss, expose to family conflict and violence, stressful life events involving loss or threat, substance abuse, toxic exposures and head injury*". These environmental risk factors may then contribute to the various symptoms associated with psychiatric disorders.

This section may be considered as an extension to the previous section on phenotypes and intermediate phenotypes here with an additional focus on the genetic aspect. In the previous section, endophenotypes were mentioned. "*They provide a means for identifying the 'downstream' traits of clinical phenotypes resulting partly from environmental factors, as well as the 'upstream' consequence of genes. To be more specific, in a gene-environment perspective, endophenotypes are also assumed to be simpler than an entire complex disorder from a genetic point of view. Instead of looking for genes coding complex disorders, endophenotypic research looks for genes for simple, ideally monogenic traits that accompany the disorder and probably contribute*

to its pathophysiology" (64). "*Decreasing the complexity of the marker should also decrease the complexity of its genetic basis. If phenotypes associated with a disorder are very specialized and represent more elementary phenomena, the number of genes required to produce variations in these traits may be fewer than those involved in producing a complex psychiatric diagnostic entity. Endophenotypes are a step towards simplifying very complex diseases*" (64).

The ultimate goal might be compared (68) to the 'Mendeleyev periodic table' known from the field of chemistry. Here is would be a 'Mendeleyev table' of endophenotypes (consisting of known genetic and non-genetic factors) relevant for psychiatric disorders that would have to be constructed to help understand the 'atomic structure' underlying these complex disorders. This 'Mendeleyev table' can then be used to understand the 'molecules' or different symptoms present in a given psychiatric disorder. Only then will it be possible to understand the 'macromolecular' structure or the range of symptoms associated with a psychiatric disorder.

Finally, it should be mentioned that "*a focus on biological markers of distinct clinical features is in line with the DSM-V research agenda, which stresses the importance of studying complex relationships between biological and clinical variables*" (3). As referred to in the previous section, the first part of the thesis work dealt with testing gene-environment (clinical features) interactions along the intermediate phenotype approach. The last part of the thesis work went into another kind of gene-environment analysis, namely the classification issue. Here it was examined whether it is possible to distinguish different patient groups on the basis of their gene expression profiles.

2.7 Shared biological mechanisms

In the sections on the described psychiatric disorders, genes are listed and are involved in several different pathways. To varying degree, it seems like biological changes take place in the immune system, the HPA axis and the monoamine function in each of the four disorders. Many more biological mechanisms are most probably perturbed, but here focus will be on these three.

In figure 3 the three biological mechanisms are shown together. As explained in (69) the HPA axis is upregulated in e.g. depression "*with a down-regulation of its negative feedback controls. CRF (corticotropin-releasing factor) is hypersecreted from the hypothalamus and induces the release of ACTH (adrenocorticotropin hormone) from the pituitary. ACTH interacts with receptors on adrenocortical cells and cortisol is released from the adrenal glands; adrenal hypertrophy can also occur. Release of cortisol into the circulation has a number of effects, including elevation of blood glucose. The*

negative feedback of cortisol to the hypothalamus, pituitary and immune system is impaired. This leads to continual activation of the HPA axis and excess cortisol release. Cortisol receptors become desensitized leading to increased activity of the pro-inflammatory immune mediators and disturbances in neurotransmitter transmission".

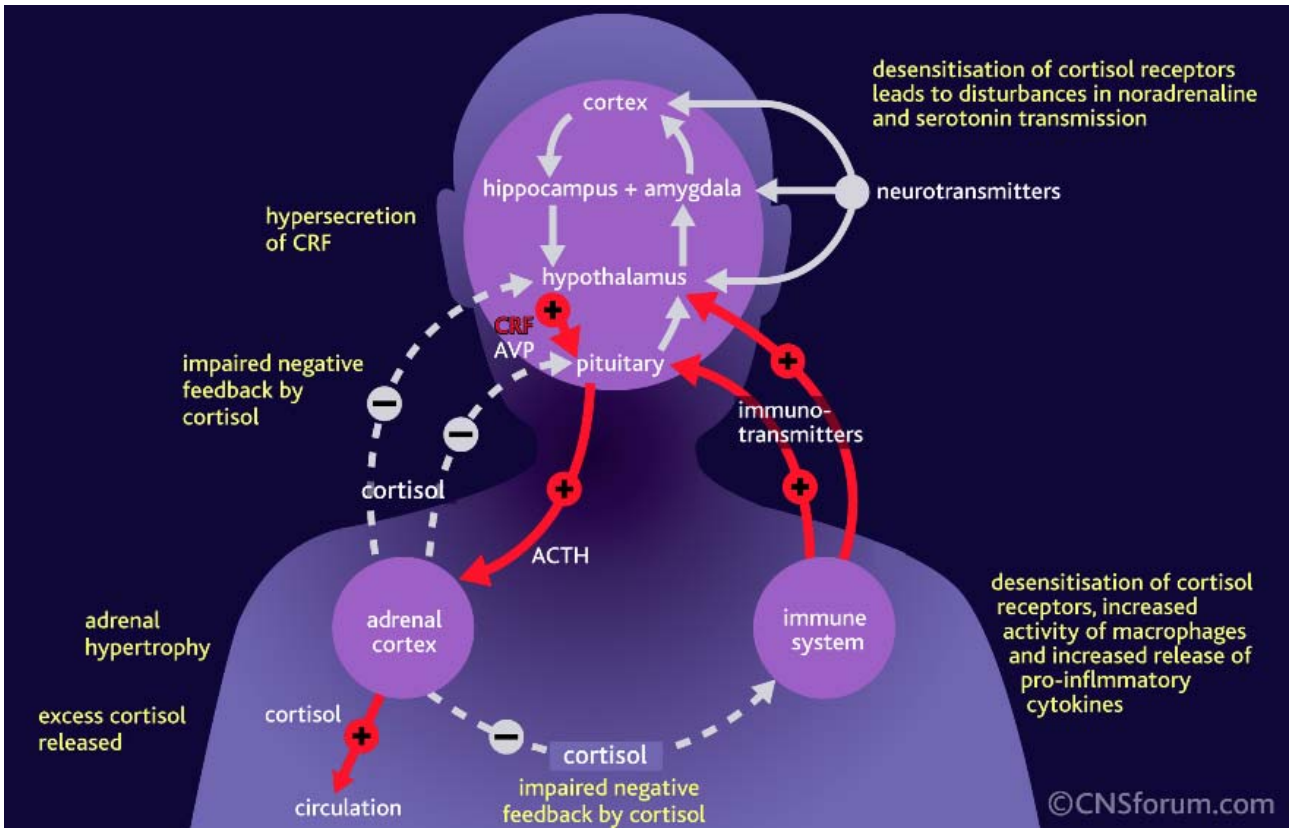


Figure 3: See text above figure. Source: Adapted after (69).

It may be hypothesized that monoamine function, HPA axis regulation and cytokine production changes are not just present in depression (24), (70), but may contribute to the pathophysiology of other psychiatric disorders (71) like PTSD (72), (73) bipolar disorder (71) and borderline personality disorder, the latter, however, perhaps to a less degree. It may also be hypothesized that the changes in these biological systems may be reflected and measured in gene expressions in whole blood, likely showing differences depending on the disorder or endophenotype compared to healthy subjects. In this thesis, for PTSD, borderline and partly depression this combined hypothesis will be investigated and it will be seen which biological systems seem perturbed in blood in each analyzed disorder.

3. The genes selected by Lundbeck

In this chapter, I present the genes that were selected by Lundbeck for my study, in particular the 29 genes coding for the G protein coupled receptors ARRB1 and ARRB2, the cell surface proteins CD8 alpha and CD8 beta, the transcription factors CREB1 and CREB2, the kinases ERK1, ERK2, MAPK8 and MAPK14, the G proteins Gi2 and Gs, the receptors GR, MR, P2X7 and PBR, the pro-inflammatory cytokines IL-1β, IL-6 and IL-8, the regulator of G-protein signaling RGS2, the calcium binding protein S100A10, the serotonin transporter SERT, the monoamine transporter VMAT2, and the enzymes ADA, DPP4, IDO, MKP1, ODC1 and PREP. These proteins have many different functions. For example, ARRB1 uncouples G protein coupled receptors from G protein, and VMAT2 that pumps monoamine neurotransmitters from neuronal cytoplasm into synaptic vesicles.
Eleven of the genes overlap with the genes identified from the small literature search in the previous chapter.
This set of genes was chosen as the basis for studying and comparing controls and patients because of their supposed relationship to various psychiatric disorders. As listed in table 3, examples are ARRB2 that show reduced levels in leukocytes from depressed patients and GR that may be downregulated on immune cells during depression. In addition, the genes all showed good expression levels in whole blood. Whenever possible, expected up or down regulation of the gene expressions will be noted.
The purpose of the chapter is also to collect biological information concerning the set of genes in order to better understand the molecular biology and the function of subsets of the genes. Using the renowned web application Ingenuity, the selected genes are grouped according to their location in the cell, that is, in the nucleus, cytoplasm, plasma membrane or extracellular space. Also, the genes are grouped into three biological networks whose main functions deal with hematological and immunological diseases, cellular growth and proliferation as well as cell death. The relation to cell death is expressed through the neurogenesis hypothesis which posits that depression is caused by neuronal death and that antidepressant drugs cause neuronal growth. The three networks include other genes involved in the same functions, which might give ideas to new possible biomarkers. Other significant functions of the selected genes are also described, like cancer, metabolic and cardiovascular diseases, indicating the possible links between these diseases and the relevance of checking subjects for these diseases prior to enrolling them into a clinical trial for studying psychiatric diseases.
Various combinations of the above genes are involved in pathways such as glucocorticoid receptor signaling and G-protein coupled receptor signaling, the latter being a key focus area of Lundbeck.

Prior to the start of this PhD, Lundbeck had chosen a set of genes as the basis for studying and comparing controls and patients with various psychiatric disorders like major depression (MDD), borderline personality disorder (BPD) and post-traumatic stress disorder (PTSD). The reader may consult the previous chapter for descriptions of these disorders.

The set of genes is based a on literature search performed by Lundbeck. This set has been (slightly) modified throughout this thesis work due to weak expression of some genes or large expression variation of other genes. New genes were added to the list to replace those that were dropped. The most consistent sets of genes analyzed comprised of 25 or 29 gene expressions. For the Serbian group of controls and for PTSD patients with and without trauma (see next chapter on study design), six additional genes were tested. In this chapter, focus is on the 29 genes.

It should be noted, that the genes selected come from both human and animal data, from both blood and brain tissue and from both RNA and protein expression data. Furthermore, Lundbeck had refined the list of selected genes based on whether they had good expression levels in blood.

Table 3 contains the 29 genes, a short description, reasons for listing each gene according to material provided by Lundbeck and the literature, and an indication of the expected up- or down-regulation of each gene expression during various mental disorders.

Gene	**Reasons for including the gene**
ADA (adenosine deaminase)	*"Enzyme catalyzes the hydrolysis of adenosine to inosine". "Adenosine is involved in learned helplessness pathway"* (74). Lower levels in MDD subjects, possible correlation with DPP4 levels (75).
ARRB1 (beta-arrestin 1)	Uncouples GPCR (G protein coupled receptors) from G protein, involved in receptor internalization. Reduced in leukocytes from depressed patients (76) and correlated with the severity of symptoms, levels rise in rats upon antidepressant treatment (77).
ARRB2 (beta-arrestin 2)	See ARRB1 for rationale. Data do not show clear change in MDD.
CD8 alpha (CD8 antigen alpha polypeptide)	Identifies cytotoxic / suppressor T cells and is involved in T cell mediating killing. Lower CD8 expression in depressed patients.
CD8 beta (T-cell surface glycoprotein CD8 beta chain)	See CD8 alpha.
CREB1 (cAMP responsive element binding protein 1)	Transcription factor, cAMP pathways regulate T cell mediated immune responses; CREB1 stimulates neurogenesis (see text after the table) in dendrite gyrase. Associated with MDD" (78) with expected reduced levels (79), (80).
CREB2 (cAMP responsive	Transcription factor, cAMP pathways regulate T cell

element binding protein 2)	mediated immune responses.
DPP4 (dipeptidyl peptidase 4)	*"Cleaves X-proline dipeptides from the N-terminus of polypeptides"* (81); binds ADA and serves to active T cells. Protein levels are downregulated in the blood of depressed (cancer) patients; serum activity higher in men than women (81); decreased enzyme activity may mean lower immune system function in MDD (75).
ERK1 (extracellular signal-related kinase 1)	Kinase, reduced protein and mRNA in brain of depressed suicide subjects (29); In PBMC (peripheral blood mononuclear cells) balance between ERK1/2 and MAPK8/MAPK14 regulates anti- and pro-inflammatory cytokine secretion with ERK1/2 promoting neurogeneration and anti-inflammation.
ERK2 (extracellular signal-related kinase 2)	See ERK1.
Gi2 (G protein, alpha-inhibiting activity polypeptide 2)	Increased in depressed patients' platelets and normalized by antidepressant treatment (82); decreased in leukocytes of depressed patients and normalized by ECT (electroshock therapy) (83); increased in bipolar disorder, but not MDD (84).
Gs (G protein, alpha-stimulating activity polypeptide 1)	Increased in bipolar disorder, but not MDD (84); decreased in leukocytes of depressed patients and normalized by ECT (electro chock treatment) (83).
GR (glucocorticoid receptor)	Lower affinity for CORT than MR, may be downregulated on immune cells in depression (85); handling of neonatal rats increases GR/MR ratio in hippocampus and prevents depression (86); receptor level in dorsal hippocampus increased after 20 days of stress (87); interferon alpha regulates GR receptors in cell lines (88).
INDO (indoleamine pyrrole 2,3-dioxygenase)	Reduces tryptophan levels, induced by pro-inflammatory cytokines and glucocorticoids and could cause tryptophan depletion (70) (perhaps leading to less serotonin synthesis) and/or kynurenine (tryptophan metabolite) increase which leads to neurontoxicity (89). May be expected to be increased in depressed patients.
IL-1β (interleukin-1 beta)	Pro-inflammatory cytokine, upregulated in depressed patients (90); cytokines can produce symptoms of depression (70), (18).
IL-6 (interleukin-6)	Pro-inflammatory cytokine, upregulated in depressed patients (91); cytokines can produce symptoms of depression (70), (18).
IL-8 (interleukin-8)	Pro-inflammatory cytokine; upregulated in monocytes of depressed patients (92); cytokines can produce symptoms of depression (70), (18).
MAPK14 (mitogen-activated protein kinase 14) (p38 MAPK)	In PBMC, balance between ERK1/2 and MAPK8/MAPK14 regulates anti- and pro-inflammatory cytokine secretion; MAPK14 and MAPK8 expression promotes neurodegeneration and pro-inflammation. Expected increased levels in MDD (93), (94).
MAPK8 (mitogen-activated protein kinase 8)	See MAPK14.
MKP1 (dual specificity phosphatase 1)	Signalling pathways; regulates phosphorylation of ERK1/2 (and hence activity), induced by glucocorticoids, upregulated in rat hippocampus after ECT (95).

	Lower levels in MDD.
MR (mineralocorticoid receptor)	Higher affinity for CORT than GR, may be downregulated on immune cells in depression (85); handling of neonatal rats increases GR/MR ratio in hippocampus and prevents depression (86); receptor level in dorsal hippocampus increased after 10 days of stress (87).
ODC1 (ornithine decarboxylase)	First step and the rate limiting step in humans for the production of polyanimes, compounds required for cell division; expression linked to cell proliferation and immunomodulation.
P2X7 (purinoreceptor P2X7) (P2RX7)	P2X purinoreceptors are cell membrane ion channels, gated by ATP; potentially involved in neuroprotection and modulation of the inflammatory response (96) – believed to be linked to the onset of bipolar disorders (30); perhaps lower levels in MDD.
PBR (peripheral-type benzodiazepine receptor)	Widely distributed, involved in cell proliferation and immunomodulation (97).
PREP (prolyl endopeptidase)	Activity correlates with DPP4 in depressed patients; serum activity higher in men than women (81), stress induces PREP levels in the blood of responders; higher in subjects with PTSD and even higher in subjects with PTSD and depression (98).
RGS2 (regulator of G-protein signaling 2)	Acts as GTPase activation protein to terminate G protein signaling; KO mice show increased anxiety, reduced T cell proliferation, reduced IL-2 synthesis (99); RGS2 is highly expressed in lymphocytes where it can influence cytokine production (100). Low levels in MDD are expected.
S100A10 (S100 calcium-binding protein A10) (p11)	Increases localization of 5-HT-1b subtype to cell surface, expression increased in rodent brains after antidepressant treatment or ECT and decreased in animal model of depression or brains of depressed humans (101).
SERT (serotonin transporter)	Target of SSRI / SNRI / TCA antidepressant drugs; influences serotonin levels at post synaptic junctions. Lower CNS expression/binding in patients with depression.
VMAT2 (vesicle monoamine transporter 2)	Pumps monoamine neurotransmitters from neuronal cytoplasm into synaptic vesicles; VMAT2 binding higher in bipolar patients in thalamus and brainstem (102); in platelets there is a significant elevation of VMAT2 expression in MDD patients versus controls (103).

Table 3: 29 genes selected by Lundbeck

Eleven of the genes in the above list overlap with the ones identified from the limited literature search in chapter two. They are SERT, IDO, CREB1, GR, MAPK14, MR, IL-1β, IL-6, ERK1, ERK2, P2X7 (all related to depression).

The main reason for the overlap between chapter two genes not being greater with the genes listed in table 3, is that a literature search of genes involved in each disorder has not been a focus area, but was merely done to get an idea of genes involved in the psychiatric disorders described.

In the conclusion and discussion chapter, genes separating groups will be compared to expected/hypothesized expression regulations in table 3 above when possible. It must be stressed that, in general, replication of previous findings for complex polygenic diseases is always difficult. The same principal replication problems exist for gene expressions studies, especially for microarray studies, as for genetic association (SNP) studies, perhaps to an even greater extent in the latter due to the millions of SNPs present in the human genome. "*Initial positive findings are hard to replicate due to small number of samples, population stratification, phenotype definition* (see chapter 2)*, genetic heterogeneity, low relative risk, multiple testing, normalization issues, selection bias especially for the control group, and other factors*" (61).

3.1 Biological networks, functions and pathways

More biological knowledge can be obtained from the list of selected genes by looking at how different combinations of genes are involved in various biological functions and pathways. This requires updated online database tools capable of handling such bioinformatics / systems biology queries into data, here a list of genes. Several such online tools exist. The best I have worked with so far is called Ingenuity Pathways Analysis (104) and makes use of only manually curated scientific literature. It is a commercial web application that Lundbeck has access to and it will be used to gain more biological insight from the list of individual genes.

Below various outputs from Ingenuity are presented that highlight various biological network aspects of the selected genes: To begin with, the 29 gene products are grouped after their location in the cell, that is, in the nucleus, cytoplasm, plasma membrane or extracellular space (figure 4). In addition, the 29 genes are arranged into networks, the most significant[9] and relevant biological functions of combinations of genes are presented as well as the most significant[9] and relevant pathways. All this information may broaden the insight of the biological basis of the 29 selected genes and provide insight into perturbed systems, once statistical and machine learning techniques have identified sets of genes separating different control and patient groups.

Figure 4 shows the location of the 29 genes in the cell and thus relates the genes in a cellular context.

[9] The significance value associated with a functional or pathway analysis is a measure of the likelihood that the association between a set of selected genes and a given process or pathway is due to random chance. This is assessed via the right-tailed Fisher's Exact Test. The smaller the p-value the less likely that the association is random and the more significant the association.

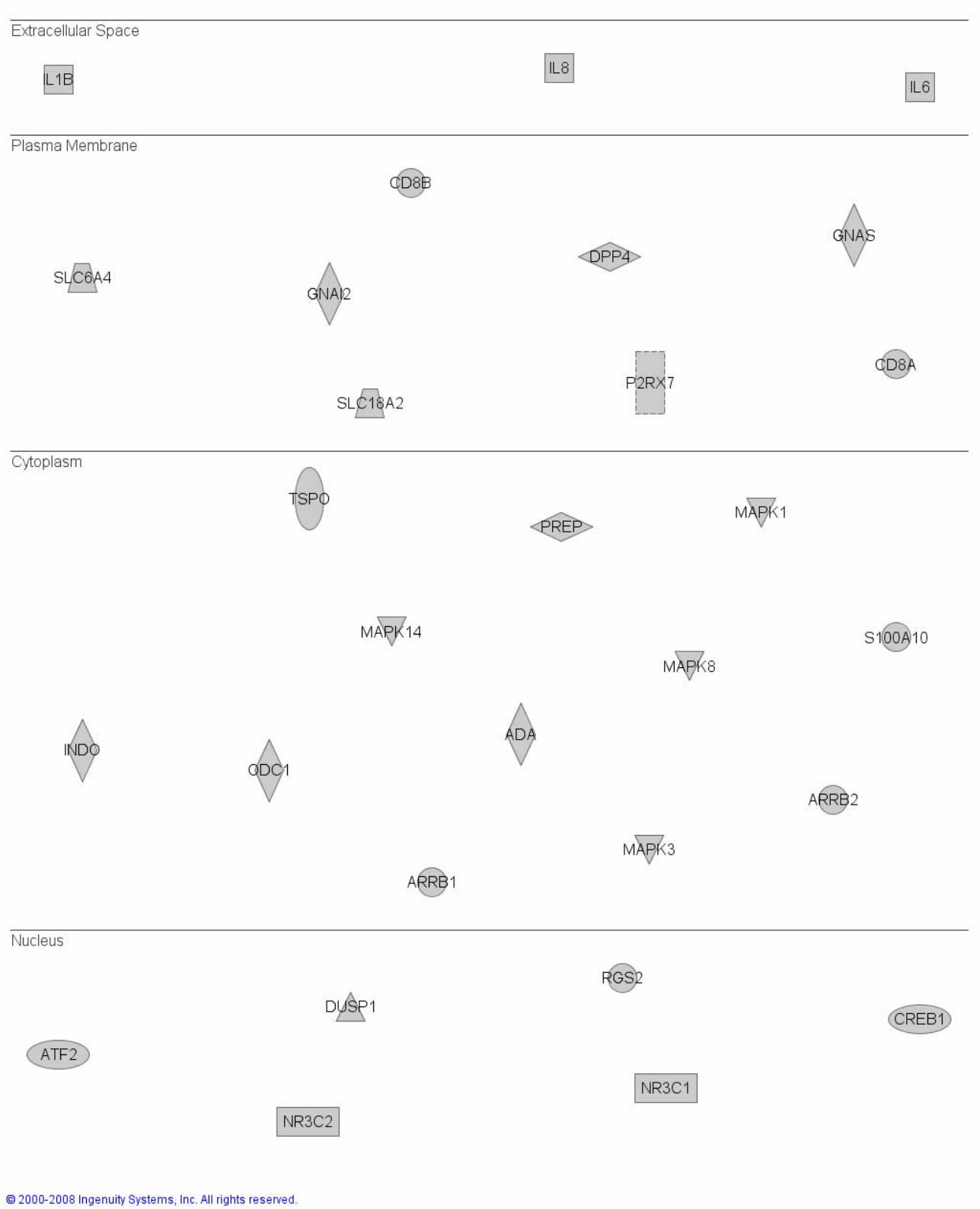


Figure 4: Location of the 29 genes in the cell. ATF2=CREB2, NR3C1=GR, NR3C2=MR, DUSP1=MKP1, TSPO=PBR, MAPK1=ERK2, MAPK3=ERK1, INDO=IDO, SLC6A4=SERT, SLC18A2=VMAT2, GNAI2=Gi2 and GNAS=Gs. Ingenuity output.

From the figure it can be seen that the gene products that function in the nucleus consists of the transcription factors CREB1/2, the HPA-axis activity regulating gluco- and mineralocorticoid receptors GR and MR as well as MKP1 and RGS2. The cytoplasm contains the kinases ERK1/2 and MAPK8/14, the arrestins ARRB1/2 as well as ADA, ODC1, PBR, PREP and S100A10. The transporters SERT and VMAT2 are located in the plasma membrane, as are the G proteins Gi2 and Gs, the transmembrane glucoproteins CD8 alpha and CD8 beta, the ion channel P2X7 as well as DPP4. The extracellular space contains all the three interleukins IL-1β, IL-6 and IL-8 related to the immune response.

Next, all the 29 genes are grouped into networks in table 4 and it can be seen all the genes group nicely into three networks:

ID	*Molecules in Network*	*Focus Molecules*	*Main Functions*
1	*Angiotensin II receptor type 1, **ARRB1**, **ARRB2**, **ATF2**, CD3, **DUSP1**, ERK1/2, G alpha, G alphai, G protein beta gamma, Ige, **IL1B**, JINK1/2, Mapk, **MAPK1**, **MAPK3**, **MAPK8**, Mek, Mek1/2, NfkB-RelA, **P2RX7**, p70 S6k, Pdgf, Pdgf Ab, PI3K, Pkg, PLA2, Pld, PP2A, Rac, Ras, Ras homolog, **S100A10**, **SLC6A4**, TCR*	*11*	*Cellular Growth and Proliferation, Connective Tissue Development and Function, Organismal Functions*
2	*Akt, Alkaline Phosphatase, AMPK, Ap1, C1q, Calcineurin protein(s), Ck2, Creb, **CREB1**, Fgf, **GNAS**, Hsp70, Hsp90, IL1, **IL6**, **IL8**, **INDO**, Insulin, Jnk, LDL, **MAPK14**, NFkB, **NR3C1**, **NR3C2**, **ODC1**, P38 MAPK, PDGF BB, Pka, Pkc(s), PLC, **RGS2**, **SLC18A2**, STAT5a/b, Tgf beta, Vegf*	*11*	*Cell Death, Hematological Disease, Immunological Disease*
3	*__ADA__, Adenylate Cyclase, Alcohol group acceptor phosphotransferase, ATAD4, beta-estradiol, Caspase, **CD8A**, **CD8B**, CDCA7, CSTB, Cyclin A, dihydrotestosterone, **DPP4**, **GNAI2**, H2-T18, HCG 1787519, Histone h3, KIAA1967, KLK15, **MAPK8**, MAPK11 PREDICTED, MHC Class I, MMD, MYC, NBPF15, NFRKB, PARP10, **PREP**, PRL2C3, RNA polymerase II, RPL21 (includes EG:79449), SSBP2, TNF, **TSPO**, ZNF267*	*8*	*Cancer, Cell Cycle, Immunological Disease*

Table 4: The 29 genes arrange nicely into three networks. For different gene names, see legend to figure 4. Ingenuity output.

Table 4 shows that e.g. network 1 contains 11 of the 29 genes (shown in bold), the top three functions/networks the 11 genes participate in and the other genes in the Ingenuity database in these networks. Looking at the top three functions/networks for the three networks some of the most relevant functions (for this thesis) have to do with hematological and immunological diseases which supports that the genes, Lundbeck has chosen to look at, are expressed in blood. Furthermore, cellular growth and proliferation as well as cell death/apoptosis is mentioned. This is interesting because one hypothesis that was not described in chapter two had to do with neurogenesis. This

hypothesis posits that depression is caused by neuronal death and that antidepressant cause neuronal growth and proliferation. It is also worth noticing cancer is being mentioned. This is interesting in the sense that cancer might cause gene expression changes in some of the selected genes, meaning it could be important to check subjects for cancer prior to their inclusion in a clinical trial.

To demonstrate the complexity of the interactions in a gene expression network, a schematic involving the genes for network 1 (from table 4) is shown in figure 5.

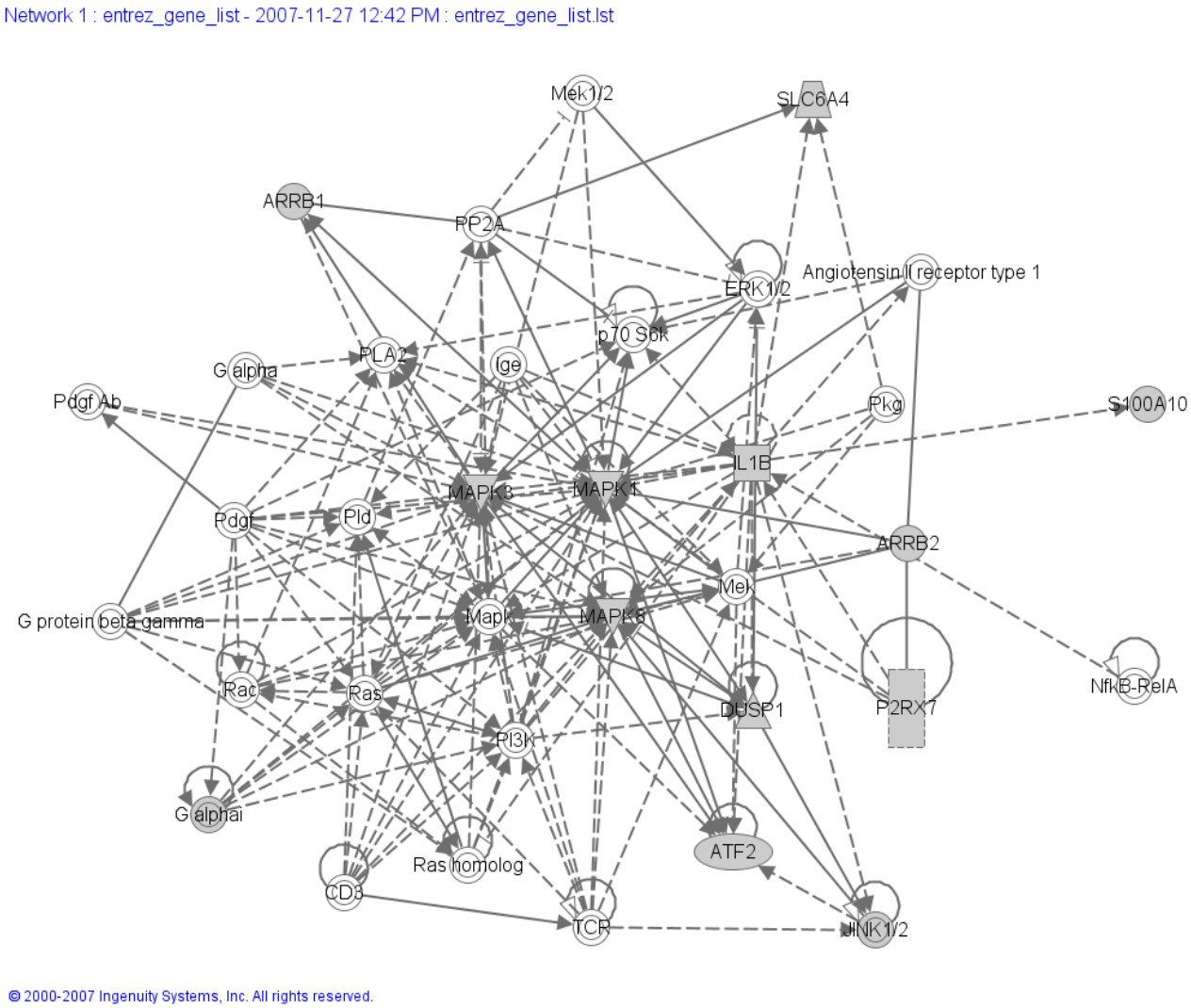


Figure 5: Network 1 from table 4 comprising 11 (shown in grey) of the 29 genes interacting with the other genes in the network. See the text for the various kinds of interactions. For different gene names, see legend to figure 4. Ingenuity output.

In figure 5, examples of interactions are: Activation/inhibition, binding, expression, phosphorylation/dephosphorylation, protein-DNA binding, protein-protein binding and transcription. Networks 1 (repeated for consensus), 2 and 3 are shown in appendix 1. Going into details about all the genes and

interactions is beyond the scope of this thesis. However, in the results chapter, bioinformatics tools will be used to predict new possible biomarkers (with a focus on protein-protein interactions) based on either a merged version of the three networks (from table 4) or with the addition of putative genes associated with bipolar disorder. Note that table 4 and e.g. figure 5 alone can give ideas regarding new biomarkers.

The top functions/networks shown in table 4 included other genes (in addition to the 29 genes) involved in the same functions. It is also possible to look into functions of combinations of the 29 genes alone. 61 combinations are significant (according to the right-tailed Fisher's Exact Test and a significance level of 1%) and listed in appendix 2. Some of the most relevant and significant combinations are listed in table 5 below (the most significant first).

Function	*Molecules*
Inflammatory Disease	*IL8, DPP4, MAPK3, MAPK8, IL6, P2RX7, CD8A, NR3C1, ODC1, GNAI2, ARRB2, MAPK14, DUSP1, ADA, IL1B, SLC6A4, S100A10*
Cell Death	*DPP4, IL8, MAPK1, MAPK3, MAPK8, P2RX7, IL6, SLC18A2, CD8A, NR3C1, ODC1, ATF2, GNAS, ARRB2, MAPK14, DUSP1, CREB1, ADA, TSPO, IL1B, S100A10*
Cellular Growth and Proliferation	*DPP4, IL8, RGS2, MAPK1, MAPK3, MAPK8, P2RX7, IL6, CD8A, NR3C1, ODC1, ATF2, GNAI2, ARRB2, MAPK14, DUSP1, CREB1, ADA, IL1B, INDO, SLC6A4, NR3C2, S100A10*
Cancer	*IL8, DPP4, MAPK1, MAPK3, MAPK8, P2RX7, IL6, CD8B, NR3C1, ODC1, ATF2, GNAI2, GNAS, ARRB2, MAPK14, ARRB1, DUSP1, CREB1, ADA, IL1B, NR3C2, S100A10*
Cardiovascular Disease	*IL8, RGS2, MAPK1, MAPK3, MAPK8, IL6, SLC18A2, NR3C1, GNAI2, MAPK14, DUSP1, CREB1, IL1B, SLC6A4, NR3C2, S100A10*
Hematological System Development and Function	*IL8, DPP4, RGS2, MAPK1, MAPK3, MAPK8, P2RX7, IL6, CD8A, NR3C1, CD8B, GNAI2, GNAS, ARRB2, MAPK14, DUSP1, CREB1, ADA, IL1B, INDO*
Immunological Disease	*IL8, DPP4, MAPK3, MAPK8, P2RX7, IL6, CD8A, NR3C1, GNAS, MAPK14, DUSP1, ADA, IL1B, SLC6A4, S100A10*
Hematological Disease	*DPP4, GNAS, IL8, MAPK14, MAPK8, ADA, IL1B, P2RX7, IL6, NR3C1*
Neurological Disease	*IL8, DPP4, MAPK8, IL6, P2RX7, SLC18A2, NR3C1, ATF2, PREP, GNAS, MAPK14, CREB1, ADA, TSPO, IL1B, SLC6A4*
Metabolic Disease	*DPP4, GNAS, DUSP1, CREB1, ADA, SLC6A4, IL1B, NR3C2, IL6, NR3C1, S100A10*
Immune Response	*IL8, DPP4, RGS2, MAPK3, MAPK8, IL6, CD8A, NR3C1, GNAI2, GNAS, ARRB2, MAPK14, DUSP1, ADA, IL1B, INDO*
Psychological Disorders	*RGS2, CREB1, IL1B, TSPO, SLC6A4*

Table 5: Functions of significant combinations of the 29 genes. For different gene names, see legend to figure 4. Ingenuity output.

Table 5 lists genes associated with the functions of inflammatory and immunological diseases as well as immune response and hematological system development and function. These functions fit very well with the HPA-axis and cytokine hypotheses of depression. SERT (SLC6A4) is involved in several of the functions among them psychological disorders and inflammatory diseases

fitting well with the monoamine hypothesis and the interaction between SERT and the interleukins/cytokines. As with the top functions/networks, cellular growth and proliferation and cell death may be good indications of a psychiatric disorder according to the neurogenesis hypothesis of depression (described above). It seems like it could be important to check subjects in clinical trials for cancer, cardiovascular, neurological and metabolic diseases as these might influence quite a lot of 29 genes. This also applies for e.g. inflammatory diseases that are not of interest like lupus or acne or immunological diseases of no interest in a trial like damage of spleen or leucopenia.

Finally, various combinations of genes are involved in different pathways. All 42 significant pathways (significance level is set to 1%) are listed in appendix 3. Below in table 6 are some of the most relevant and significant pathways ordered after significance (most significant in the top).

Pathway	*Molecules*
Glucocorticoid Receptor Signaling	*IL8, MAPK14, MAPK1, DUSP1, MAPK3, CREB1, MAPK8, IL1B, NR3C2, IL6, NR3C1*
IL-6 Signaling	*IL8, MAPK14, MAPK1, MAPK3, MAPK8, IL1B, IL6*
cAMP-mediated Signaling	*GNAI2, GNAS, RGS2, MAPK1, DUSP1, MAPK3, CREB1, ATF2*
G-Protein Coupled Receptor Signaling	*GNAI2, GNAS, RGS2, MAPK1, DUSP1, MAPK3, CREB1, ATF2*
p38 MAPK Signaling	*MAPK14, DUSP1, CREB1, IL1B, ATF2*
T Cell Receptor Signaling	*MAPK1, MAPK3, MAPK8, CD8A, CD8B*
Serotonin Receptor Signaling	*SLC6A4, SLC18A2*
Apoptosis Signaling	*MAPK1, MAPK3, MAPK8*

Table 6: Pathways of significant combinations of the 29 genes. For different gene names, see legend to figure 4. Ingenuity output.

Table 6 shows the involvement of the HPA-axis by including glucocorticoid receptor, p38 MAPK and cAMP-mediated receptor signaling. G-protein coupled receptor signaling (this list overlaps completely with the gene list of cAMP-mediated receptor signaling) is a key focus area of Lundbeck. The immune system also comes into play with IL-6 signaling and T cell receptor signaling. Monoamine activity is noted via the serotonin receptor signaling. Finally, cell death (apoptosis) signaling also seems to play a role involving three genes.

4. Study design

We are now ready to present the study design in terms of the various control and patient groups analyzed. There is a US control group of 299 individuals, a 30 person UK control group with three time point measurements on two consecutive days, a Danish control group with 89 members and a 78 person Serbian control group. Also, there are data from borderline personality disorder (BPD) patients, acute post-traumatic stress disorder (PTSD) patients, patients with trauma but without PTSD and finally expression measurements from remitted PTSD patients.
For the US control group, the Danish control group and one part of the BPD patient group, questionnaire data with clinical information is also available. Focus has been on clinical information relating to both psychological factors and covariates as both might be indicative of clinical symptoms of the studied affective disorders. Examples of psychological factors are a family history of depression, anxiety or suicide, lifetime experience of various affective disorder episodes, and sleep problems or lack of energy during the two week period prior to blood sampling. Examples of covariates are age, gender, body mass index, tobacco use, and alcohol use. The clinical variables of interest were then coded into numbers, and as an overall guideline, coding was done as intuitively as possible with, in general, a score equal to zero if the respondent had not experienced a predisposition factor. The score was then increased as the symptom level increased. Composite scores were also created like an early life stress score covering stressful life events before the age of 15 (this is an important factor predisposing to various affective disorders, see chapter 2) and vegetative symptom score, which were considered to be a better indicator of melancholic depression.
The chapter also includes a short outline of the applied 'quantitative polymerase chain reaction' (qPCR) technique for quantifying the gene expressions in blood. A short illustration shows how blood is first drawn into a tube 'freezing' the blood, how RNA is then extracted and cDNA created which is finally measured. A crucial part of quantifying gene expressions relates to normalization to account for possible variation in the amount and quality of RNA or cDNA between the biological samples obtained from the different control and patient groups across the world and measured at different time points. Here, it is described how Lundbeck chose to work with seven housekeeping genes to solve the normalization issue.
Various sources of measurement errors are described relating to the self assessment in the questionnaire responses, the interpretation of the questionnaire responses, possible error sources moving from questionnaire to excel file and also, relating to the qPCR technique and the actual clinical diagnose.
Finally, qPCR is compared to another widely applied gene expression technique – microarrays. In brief, microarrays are suitable for the measurements of

thousands of genes simultaneously, but the analytical process involves risk of over-interpreting the results due to data overfitting. On the other hand, qPCR is more sensitive and advances reproducibility of the data.

As mentioned in the introduction, blood samples have been collected for a number of control and patient groups suffering from borderline personality disorder and post-traumatic stress disorder (PTSD). Gene expressions in whole blood have then been measured for the same 25-30 genes in all samples. Depending on control and patient groups compared, the overlapping gene expressions were utilized in the analyses. Furthermore, in some cases clinical information was accessible. Table 7 sums up the data available from Lundbeck for analyses in this thesis. The results of the various analyses are presented in the results chapter.

Control/patient group	Short description	#Subjects	#Genes	Clinical information?
1) ABS controls	All US controls (both genders)	299	29	Yes, based on an US questionnaire.
- SH ABS controls	- 'Super healthy' US controls (both genders)	- 59	- 25	
2) UK controls	UK controls (males); 3 time points	30	29	No.
3) DC controls	Danish controls (both genders)	89	29	Yes, based on a Danish questionnaire.
4) Borderline patients	Two cohorts of patients (mostly females)	21	29	For one cohort only.
5) PTSD groups:	(Men only)		35	No.
- PTSD controls	- Serbian controls	78		
- Remitted PTSD patients	- Remitted Serbian PTSD patients	41		
- Acute PTSD patients	- Serbian acute PTSD patients	66		
- Trauma patients	- Serbian trauma patients without PTSD	87		

Table 7: Data available for analysis in this thesis. Control/patient group names contain the abbreviations used later in the thesis. Content is explained after the table.

Group 1) The first data available was obtained from a large US group of 299 healthy male and female subjects. 29 gene expressions (listed in table 3, chapter three) were measured in this ABS control group that consisted of four cohorts. The first of these cohorts was called the 'super healthy' controls, in short SH ABS. The SH ABS controls were selected by Lundbeck based on having a BMI (body mass index) less than 30 and having taken no drugs the last three months. By checking their questionnaire answers (more about this in the questionnaires section), it turned out that the SH ABS had in general lower stress scores, fewer symptoms and less history of family depression, anxiety or

suicide than the rest of the ABS subjects. In the SH ABS, only 25 gene expressions were measured. The reason was that Lundbeck in the beginning of the study was experimenting to identify the final set of genes to be investigated. They started with 29 genes, found that four of them showed poor expression or large variability, cutting the list down to 25 genes used for cohort 1/SH ABS. Later four additional genes were added bringing the list of genes up to the 29 genes described in the previous chapter.
All 299 controls filled out a questionnaire that is described later in this chapter. Not being a homogenous control group, the ABS group was used to make predictions about potential gene expression changes in depressed patients. There were two aspects to the predictions; the first was to detect possible gene expression trends in depressed patients and the other was to identify intermediate phenotypes among the controls.
The SH ABS group was used for various comparisons to the patients.

Group 2) The UK controls comprise 30 healthy male subjects that had their blood taken at three different time points: Day 0 at 8 am, Day 0 at 2 pm and Day 1 at 8 am. The purpose was to investigate whether any of the 29 gene expressions differed significantly between the three time points. The UK controls have been part of various comparisons to patients.

Group 3) The DC controls consists of 89 Danish healthy male and female subjects that had filled out an extensive questionnaire. This questionnaire is not the same as the one used for the US ABS controls – questionnaires will be dealt with later in this chapter. 29 gene expressions were measured in each subject. The expression values of the Danish controls have been compared to the US SH ABS controls and also been part of various other comparisons to patients.

Group 4) The expression profiles of 21 borderline patients arrived in two cohorts with the second cohort ready for analysis eight months after the first. The first cohort consisted of female patients only and clinical information was also available. The second cohort consisted of mostly females with addition of two males and, apart from gender, age and BMI, no clinical information were available.
29 gene expressions were measured in whole blood in all these patients. The borderline patients have been used to compare to control subjects and other patient groups.

Group 5) All PTSD subjects come from Serbia and are male. They are divided into 78 controls, 66 acute PTSD patients, 87 patients with trauma without PTSD and 41 remitted PTSD ‘controls’. The last three groups have experienced war events. 35 gene expressions were measured in the PTSD groups, that is, six genes were added to the analysis compared to the 29 gene list. With the PTSD groups, it has been possible to do interesting comparisons, such as comparing the expression profiles of controls with remitted patients and

compare acute patients with patients with trauma but without PTSD. The PTSD groups have also been part of other comparisons.

Publicly available data has also been analyzed with bioinformatics tools. The Wellcome Trust Case Control Consortium has compared the genetic profiles (SNP data) in blood of 2000 UK bipolar patients with 3000 UK controls. The interactions between the genes thus associated with bipolar disorder and the 29 genes are investigated, see the results chapter.

4.1 Questionnaires

The US ABS control group of 299 healthy subjects filled out a questionnaire with approximately 50 questions, the majority of them with multiple answer possibilities in the form of check boxes (see appendix 4). Below the questions are listed that were chosen by Lundbeck for coding in order to compare responses to gene expressions. Coding issues and possible sources of error are described as well.

The questions chosen for further processing related to:

1. Age
2. Gender
3. Weight and height (BMI)
4. Frequency of
    a. Tobacco use
    b. Alcohol use
5. Lifetime and three months drug use
6. Lifetime and current medical history
7. Experienced and frequency of experience during the last two weeks of
    a. Feeling low
    b. Lack of energy
    c. Less interesting in daily activities
    d. Difficulties concentrating
    e. Sleep problems
    f. Anxiety
    g. Not being able to cope with daily problems, having considered suicide
8. Experienced changes in and level of change during the last two weeks in
    a. Appetite
    b. Weight
    c. Sexual interest
9. Lifetime experience of various affective disorder episodes including alcohol and substance abuse

10. Lifetime treatment of various affective disorder episodes including alcohol and substance abuse
11. Family history of
    a. Depression
    b. Anxiety
    c. Alcohol abuse
    d. Other substance abuse,
    e. Schizophrenia/psychosis
    f. Suicide
12. Early life stressful events (before age 15)
13. Recent stressful events (in the last 12 months)

These questions can be divided into clinical variables, that is psychological predisposition factors, and covariates. The following items on the list are considered to be covariates (item 1-5) – age, gender, BMI, tobacco use, alcohol use, caffeine intake and lifetime and three months drug use. It is hypothesized that some gene expressions might be correlated with a covariate or show differences between e.g. smoking and non-smoking respondents. Item 6 - lifetime and current medical use – is also a covariate included to analyze how various diseases and inflammations influence the expression levels of the selected genes.

Some of the clinical variables in the list above (item 7 and 8) are directly related to the DSM-IV-TR depression symptoms described in Chapter 2. Furthermore, as noted in that chapter, depression in some cases runs in families. This possibility is covered in the controls by item 11 in the list. Item 12 and 13 deal with stressful life events as it is known that stressful situations might trigger a depressive episode. Also, since subsequent depressive episodes may occur with or without an obvious trigger, checking for previous episodes is covered in item 9 and 10. By having questions in the ABS questionnaire matching clinical symptoms and possible causes, observation of intermediate phenotypes becomes feasible as the clinical variables reflect an enhanced risk of developing an affective disorder. Checking for the influence of covariates could, in addition, eliminate some false positive findings.

Coding
After selecting the clinical variables and covariates of interest, they were coded into numbers that is, scored. In some cases, Lundbeck calculated composite scores as well. Examples of these two kinds of coding are given below. A complete list of applied variable coding for the ABS controls can be found in appendix 5. As an overall guideline, coding was done as intuitively as possible with, in general, a score equal to zero if the respondent had not experienced a predisposition factor, that is, a particular symptom was not present at all. The score was then increased as the symptom level increased – examples are given below.

Coding a discrete covariate like tobacco use was done by setting the tobacco score equal to zero if the respondent smoked less than a cigarette per week, otherwise the score was set equal to one:

Tobacco use	
None ever	
None, past 12 months	0
Less than 1 per week	
1 to 10 per day	
10 to 20 per day	1
Greater than 20 per day	

Tobacco use was also binned/divided into three levels, low (less than 1 per week), medium (1-10 per day) and heavy users (more than 10 per day).

Coding a discrete clinical variable like experienced anxiety was done by giving each answer possibility a score:

Anxiety level (past two weeks)	
Never	0
Sometimes	1
Most days	2
Every day	3

In the case of scoring the various family histories, a separation was done between first and second rank relative affected by an affective disorder. This should account for the fact that a respondent with a first rank relative (mother, father, child, and sibling) is more prone to develop depression (could have a genetic predisposition) than a respondent with a second rank relative (uncle, aunt, grandparent, and grandchild). There was no consideration for the number of relatives affected. Scoring a family history of e.g. alcohol abuse was then done by assigning the score zero if no relatives had a history of alcohol abuse, a score of 1 if any secondary relative had that family history, and a score of 2 if any primary relative had a family history of alcohol abuse.

The family history of depression was combined with the family histories of anxiety and suicide, since these disorders range closely and predispose to depression. The scoring of a family history of depression, anxiety and suicide was then done exactly as the example with a family history of alcohol abuse above. In the beginning, each family history was coded separately but it was found to be more advantageous to combine them.

The composite scores, defined by Lundbeck, were early life stress score (105), recent stress score (105), seven symptom score, symptom score sum and vegetative symptom score (see appendix 5). For instance, early life stress

score was defined as the sum of boxes checked for stressful events before the age of 15 (item 12 in the questionnaire list above). The top item (death of both parents) had a value of twenty and the bottom item in the list (major change in living conditions) had a value of eleven. The symptom score sum was the sum of scores for ten symptoms (feeling low, lack of energy, less interest in daily activities, difficulties concentrating, sleep problems, difficulty coping, experienced anxiety, appetite change, weight change and sexual interest change). These scores were considered as semi-continuous scores and sometimes binned into two or three bins to investigate whether binning would yield different results than correlations. This turned out to be the case (more about this in the results chapter). In the two-bin symptom score sum case, a respondent was assigned the score zero if he/she had no symptoms, otherwise the score one. In the three-bin early or recent stress score, the scores were divided into bin one if the score equaled zero, in bin two if the score was between one and thirty, otherwise in bin three. For the symptom score sum, bin one was for the respondents with a symptom score of zero, bin two if the score was between one and ten, and from eleven and above bin three applied. The composite vegetative symptom score, which were considered to be a better indicator of melancholic depression, was coding into four levels ranging from zero (no problems) to three (most problems).

The Danish DC questionnaire

The questionnaire used for the Danish control group was also a self-rated questionnaire, but it was somewhat more extensive than the above questionnaire. It consisted of around 80 main questions with, in general, more sub questions/answer possibilities per question than the US questionnaire. Since it was 30 pages long (compared to the US 11 page questionnaire), it is not included in the appendix.

Focus started and remained on questions relating directly to the US questionnaire for comparability reasons. Some of the overlapping questions related to age, gender, BMI, lifetime and three months drug use, tobacco and alcohol use, family history of affective disorders, lifetime and current medical history, recent changes in appetite, weight, sexual interest and sleep, and recent stressful life events. As far as possible, these questions were then coded the same way as the US questions, but this coding is not described further here.

Also, I never obtained the borderline disorder questionnaire, but Lundbeck coded the relevant questions just like with the US ABS questionnaire, see appendix 5.

Sources of measurement error related to the questionnaires
There are several different sources of measurement errors relating to the questionnaires, see table 8 for examples.

Source of measurement error	Description
Questionnaire responses - self assesment	Two types of errors pertain to self assessment of questionnaires. - First, the two questionnaires contain many questions, with the DC questionnaire being the most extensive. It could be assumed that in some cases respondents could get tired from answering the questions and would mark some answers without paying careful attention to the actual topic. - Second, respondents are asked to make self assessments. One could question the validity of such an approach. For instance, do respondents recall events/issues correctly (selective memory)? Do they know which disorder a family member actually suffered from, if any? These two issues could be addressed if the responses could be double checked with e.g. other family members or a family medical history, if possible. Some of self rating questions might be influenced by the respondent having a good or bad day at the time of filling out the questionnaire or influenced by a respondent's normal (and perhaps unrealistic) way of perceiving his/her abilities.
Interpreting questionnaire responses - different rating - selection of clinical variables and covariates - composite scores	- If a different rating was applied to the categorical answer possibilities of a question, it would probably not yield different results as long as the rating was sensible, but this has not been tested directly[10]. Somewhat similar to the above topic, is the issue of whether a continuous clinical variable should stay continuous or be binned into a categorical clinical variable. Both options have been tried and the results are not the same (more about this in the results chapter). - Selecting the clinical variables and covariates of interest involves subjectivity. Selecting other variables than the chosen might yield different results and interpretations. - Related to the above topic, alternative composite scores, which would also make sense, could have been calculated and would most probably yield different results and interpretations compared to the applied composite scores.
From questionnaire to excel-file - typing errors - coding errors - calculation errors	- The questionnaires are all filled out handwritten. Typing the answers into an Excel file might result in a probably small fraction of typing errors. - When the answers in Excel had to be coded, a small

[10] Two and three bin scores applied a different binning to some variables, but in general the results did not differ between the two kinds of binning.

	percentage of the coding might be mistyped. This also applies to the composite scores. - Also, the composite scores might be miscalculated due to human errors.

Table 8: Sources of measurement error and descriptions related to the questionnaires.

In general, care was taken to avoid as many of the sources of measurement error as possible, but still some errors relating e.g. to self assessment or typing/coding errors might be present in the data.

4.2 qPCR and normalization

As mentioned in the introduction, (whole) blood measurements have been used in a number of recent studies of various psychiatric disorders (6), (7), (8), (9), (10), (11) based on the assumption that peripheral changes are part of the biology of mental disorders (5).

An important preprocessing step in the study design relates to the biochemical reaction technique (qPCR method) of quantifying gene expressions from whole blood samples. All work related to qPCR and normalization was solely done by Lundbeck, so below only a brief introduction is given to qPCR, and the applied normalization method. qPCR is compared to microarrays to highlight advantages and disadvantages of these widely applied gene expression measurement techniques. Finally, some sources of qPCR measurement error are mentioned. For more information on qPCR, references are given to the literature and websites.

qPCR is short for 'quantitative polymerase chain reaction' or more accurately in this thesis 'quantitative real-time polymerase chain reaction'. It is a standard laboratory technique based on the polymerase chain reaction, which is used to amplify and simultaneously quantify a targeted DNA molecule depending on the experimental design. It enables both detection and quantification of the target. Expression levels can be reported as absolute number of copies or relative to specific reference genes. Lundbeck chose the relative method of reporting gene expression levels (see below).

"*The procedure follows the general principle of polymerase chain reaction; its key feature is that the amplified DNA is quantified as it accumulates in the reaction in real time after each amplification cycle*." (107) The amount of DNA is measured after each cycle by use of fluorescent dyes. As is the case in the Lundbeck study, real-time PCR is combined with reverse transcription to quantify messenger RNA (mRNA), enabling Lundbeck to quantify relative gene expressions in whole blood, see illustration in figure 6. Here it should be remembered that unlike the genome (DNA) which is context-independent, the transcriptome (RNA) is context-dependent, that is, the mRNA level varies with

physiology and pathology (106). Further information on qPCR can be found in e.g. (107), (106), (108) and (109).

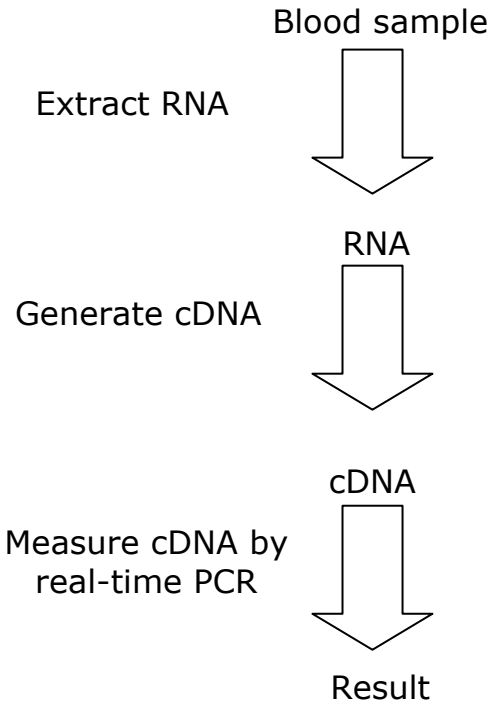


Figure 6: Real-time PCR process. First blood is drawn into a PAXgene tube having the purpose of 'freezing' the transcription profile and enabling long-term storage. Then RNA is extracted from whole blood sample, then cDNA is generated which is finally measured by quantitative real-time PCR.

A crucial part of quantifying gene expressions relates to normalization to account for possible variation in the amount and quality of RNA or cDNA between different samples and thus control for variables that could mask any underlying biological changes. This is certainly relevant in the Lundbeck study with blood samples from various control and patient groups from different countries and processed at different dates. One way to effectively compare gene expression patterns between different samples is to use normalization genes, also known as housekeeping genes (HKG). "*The term housekeeping gene refers to genes that encode for proteins whose activities are essential for the general maintenance of cell function*" (110). "*In the context of qPCR relative quantification, it is largely assumed that expression of a HKG is invariant across control and disease groups; however, the expression of some HKGs is regulated in a number of cell types and tissues*" (110) making the choice of HKGs challenging.
It is generally advised not to rely on just a single HKG (106), (109), and Lundbeck has spent quite some time and effort to identify a proper set of HKGs; Lundbeck started out by trying a method that involved comparing the expression of a gene with the expression of two HKGs in each sample. It turned out that these two HKGs could not be used to normalize every data set since different groups of subjects required different HKGs which might be due

to drug effects, disease effects and perhaps also ethnicity differences. If the same HKGs are not used in each group, expression values can not be compared. Next, Lundbeck tried using seven HKGs based on the rationale that even though all seven may not be ideal for any particular experiment, by using a large number of HKGs to normalize they would reduce the variation introduced by the noisy genes and end up with a good method. The seven HKG approach is the one chosen by Lundbeck, and was applied to most of the data available (however, not for cohort 2, 3 and 4 of the ABS control data that were normalized with two HKGs) for analysis thus allowing the comparison of every experiment to all others.

qPCR vs. microarrays

To better understand the advantages and limitations of qPCR, it can be compared to the microarray technique that is used to measure thousands of gene expressions simultaneously. Table 9 compares some of the pros and cons of the two techniques.

qPCR (real-time)		Microarrays	
Advantages	Disadvantages	Advantages	Disadvantages
• Extremely sensitive • Large dynamic range • Relatively inexpensive per sample • Focus on small number of gene expressions advances reproducibility of data (small chance of over-interpretations) • Fast	• Suitable only for relatively few selected genes at the time • Careful controls necessary to interpret data and avoid contamination.	• Suitable for the measurements of thousands of genes simultaneously. • Good at identifying new possible biomarkers.	• Less sensitive than qPCR. • Relatively expensive per sample. • May be more expensive in start-up cost than qPCR equipment • Not appropriate for a few genes • Large numbers of genes add to the complexity of the analytical process and involve risk of over-interpretation of the results.

Table 9: Comparing some pros and cons of two techniques to measure gene expressions, qPCR and microarrays (111), (112), (113).

In many studies over the last years, microarrays are used first to identify novel putative biomarkers and then qPCR is used to validate the microarray findings.

Sources of measurement error related to qPCR and normalization

Some of sources of qPCR measurement error are mentioned in table 9 and relate to RNA quality (possible degradation of RNA) and normalization (choice

and number of HKGs). Other qPCR problems relate to the PCR reaction itself (like template concentrations and inhibitors present in sample, especially in mammalian blood (106)) and reverse transcription (like RNA extraction, choice of reverse transcriptase and amount of RNA transcribed) (112).

As mentioned previously, Lundbeck has put a lot of effort in optimizing the qPCR process, making sure samples from different groups can be compared, and in that process minimized the various sources of measurement error.

Other sources of measurement error in the study design

Another potential and important source of measurement error relates to the clinical diagnose. As described in chapter two, many disease phenotypes exist within a disorder and often comorbidity with related disorders is present. Hence, we can not be absolutely sure that a patient diagnosed with a certain disorder actually has this disorder. This source of measurement error may be reduced if a patient can be assessed and obtains the same diagnose by at least two independent psychiatrists.

5. Statistical methods

As explained in the introductory chapters our study involves whole blood gene expression measurements and includes selected housekeeping genes for normalization. Because of the explorative character of the study, the US Lundbeck group and I were not sure which statistical methods would be most useful to analyze the data. Analysis of the measured qPCR expression data ranges between traditional statistics and microarray analysis with respect to the ratio of samples per variable. Chapter 5 describes various statistical methods found suitable for our exploratory analysis of the data together with the assumptions one must make before applying each of the methods (classification issues will be considered separately in the next chapter). Also described, is the reason for choosing a particular method as each method offers new interpretations of the data. This is followed in each case with an example from the available data demonstrating the usability of that method. The reasons and the applied methods are;

- *A basic and fundamental assumption for parametric statistical tests is that data is normality distributed. Normality of a gene expression may be assessed by a graphical analysis called a normal QQ plot or by various normality tests. Five different normality tests are applied.*
- *In order to identify genes separating control and patient groups based on various clinical variables, univariate tests are applied. Both parametric (t-test and ANOVA test) and nonparametric (Wilcoxon rank-sum test and Kruskal-Wallis test) tests are applied, the latter to account for non-normal expression data.*
- *In order to investigate whether any of the gene expressions differ significantly between the three time point measurements in the UK control group, repeated measures ANOVA is applied.*
- *Correlations between gene expressions, and between continuous clinical variables and gene expressions, were looked into by Spearman's nonparametric rank correlations. This approach is particularly useful to handle non-normal data and outliers.*
- *An explorative approach to identify possible clinical variable - gene expression relationships is canonical correlation analysis. The method facilitates the study of linear interrelationships among sets of multiple dependent variables (the gene expressions) and multiple independent variables (the clinical variables).*
- *Recursive partitioning is used to identify possible disease subtypes with distinct gene expression profiles via a classification tree.*
- *Cluster analysis can group objects (genes or subjects) into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters. By clustering gene expression*

profiles, disease subtypes may emerge. Heat maps implement two-way clustering combining clustering of genes with clustering of subjects.

- *Finally, in order to investigate the clinical variables that explain most of the variance in a gene, stepwise regression is applied.*

Quantitative polymerase chain reaction (qPCR) gene expression data analysis ranges between traditional statistics and microarray analysis with respect to the ratio of samples per variable. Traditional statistics often operates with many more observations/samples than the number of variables while the opposite applies to microarray gene expression analyses with more variables than the number of observations. For some groups of the available qPCR data, there are more observations than variables and for other groups, like for the borderline patients, there are more variables (gene expressions) than the number of subjects. For most groups, the ratio of the number of subjects to the number of genes is around only 1-3, making it necessary to consider both microarray analysis as well as traditional statistical methods.

Statistical and microarray oriented methods applied to gene expression data are discussed in several books (114), (115), (116), (117), (118), (119), articles (120), (121), (122), (110), (123), statistical packages (see references below) and the online encyclopedia Wikipedia.org. Quite a number of methods have been considered and tested primarily in the open-source statistical environment R / Bioconductor, secondary in SPSS, Matlab and SAS. The Statistical Consulting Center at the Department of Informatics and Mathematical Modeling at DTU has assisted a great deal with the initial R coding and the initial analyses. A list of selected methods, based on their relevance for the study (as decided in collaboration with Lundbeck and me) and applicable to the available qPCR data in this thesis, is shown in table 10. The table also briefly explains why a particular method was chosen. After the table each method is described and an example demonstrating the method is included.

This thesis has a special focus on machine learning / classification methods for predicting disease status (e.g. control vs. patient) based on gene expression profiles of subjects. Simulation studies were performed to determine the best classification and variable selection algorithms applicable to the qPCR data. The topic is described separately in the next chapter. However, some classification issues are also included in table 10.

Purpose	Method & section	Reference
Is data normally distributed?	5.1 normal QQ plots, normality tests	(116), (117)
Identify genes separating control and patient groups	5.2 univariate tests (t-test/Wilcoxon, ANOVA/Kruskal-Wallis) with multiple test correction - Variable selecting from machine learning / classification methods (next chapter)	(116), (118), (117), (121), (110), (122)
Are gene expression levels the same across different time points?	5.3 repeated measures ANOVA	(119)
Similarity between control groups	5.2 univariate tests, 5.4 correlations and correlation tests - classification (next chapter)	(116), (118), (117), (121), (110), (120), (122)
Identify (intermediate) phenotypes - clinical variable – expression relationships (covariate analysis)	5.2 univariate tests, 5.4 correlations (for continuous clinical variables), 5.5 canonical correlation analysis, 5.6 recursive partitioning	(124), (116), (118), (117), (121), (110), (120), (122)
- expression patterns only	5.7 clustering and heat maps	(116), (115), (117), (122), (123)
Which clinical variables explain most of variance in a gene?	5.8 Stepwise regression	(114)

Table 10: Overview of main statistical methods applied in this thesis. The starting point is obviously the purpose of a particular analysis and methods are then selected that can perform the analysis of interest. Each method is described after the table.

As the table illustrates, whether the gene expressions are considered as dependent or independent variables, depend on the type of analysis. This will be considered for every method, if relevant.

5.1 Normal probability plots and normality tests

A basic and fundamental assumption for parametric statistical tests is that data is normally distributed. In the univariate case, e.g. the gene expressions are considered individually, the data for each gene expression should follow a normal distribution in order to apply, for instance, a t-test or an ANOVA test. If the variation from the normal distribution is sufficiently large, the resulting parametric statistical tests are inappropriate (125). Two options exist for dealing with non-normal data; either one can try to apply a data transformation like a logarithmic transformation (recommended by the statistical department of Lundbeck and having the effect of stabilizing the variance) or one can apply a non-parametric test, described in the univariate test section.

There are different ways of assessing departure from normality. One of these is a graphical analysis, called a normal QQ (Quantile-Quantile) plot. This is a graphical method for diagnosing differences between the probability distribution of a statistical population from which a random sample has been taken, i.e. the expression data of a gene, and the normal distribution.

For a gene expression "*sample of size n, one plots n points, with the (n+1)-quantiles of the normal distribution on the horizontal axis (for k = 1, ..., n), and the order statistics of the sample on the vertical axis. If the gene expression population distribution is the same as the normal distribution this plot approximates a straight line, especially near the center. In the case of substantial deviations from linearity, the null hypothesis of sameness is rejected*" (126). Figure 7 shows two normal QQ plots, one using the SERT gene expression data for the ABS control group and the other using the natural logarithm of the same expression data. As the figure shows, applying the logarithm to the expression data makes the data resemble a normally distribution more closely than before taking the logarithm.

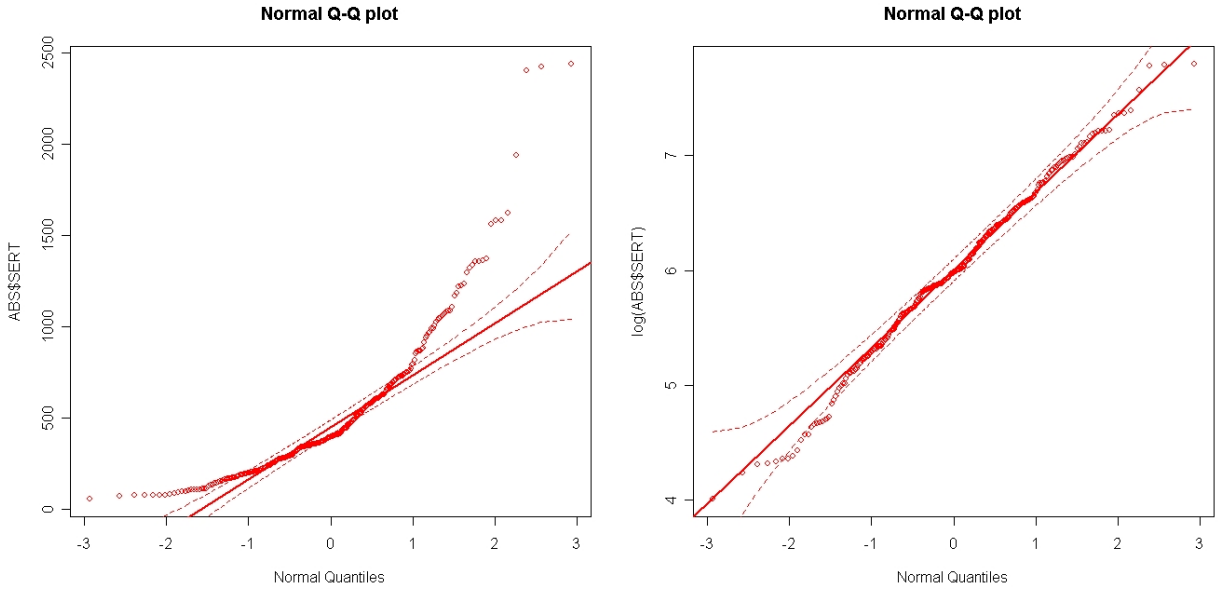


Figure 7 shows that in the case of SERT, a logarithmic transformation makes the SERT expression data follow a normal distribution. The plots were made in R.

Various statistical tests exist to assess normality (116), (127), (128). The ones applied at the beginning of the PhD are the Shapiro-Wilk test (129), the Anderson-Darling test (130), the Cramér-von-Mises criterion (131), the Lilliefors test for normality (132), and the Shapiro-Francia test for normality (133). Each test has its own way of assessing departure from normality and its own strengths and weaknesses, see the references for details about the tests. By evaluating the results from five different tests, it should be possible to

decide if the (logarithm of) expression values of a gene follow the normal distribution. In table 11, the tests are applied to the ABS SERT expression data (shown in figure 7) both the raw expression data (normalized with 2 HKGs) and applied to the natural logarithm of the SERT data as well. The higher the p-value, the more likely the data are normally distributed.

Test for normality	SERT p-value	Log(SERT) p-value
Shapiro-Wilk	1.0448e-17	0.2088
Anderson-Darling	7.2025e-30	0.1980
Cramér-von-Mises	2.5659e-06	0.2101
Lilliefors	4.3850e-17	0.0439
Shapiro-Francia	1.8886e-15	0.2901

Table 11: P-values from the five different normality tests applied to the SERT gene expression values from the ABS control group. The analyses were done in R.

With a 1% significance level to account for multiple testing (see the univariate section below), table 11 indicates that applying the logarithm to SERT improves normality of the distribution of SERT gene expression data.

In the results chapter, the five normality tests are applied to all 'raw' gene expressions values as well as to all logarithmic expression values. Some normal QQ plots are also included.

5.2 Univariate tests

The univariate tests applied in this thesis comprise the two-sample parametric t-test and the two-sample nonparametric analog, the Wilcoxon rank-sum test. Also, in the multiple sample case (more than two samples), the univariate parametric ANOVA test is used together with the non-parametric analog, the Kruskal-Wallis test. I have used the two-sample univariate tests to identify genes separating control and patient groups, and to compare control groups (see the results chapter). Both the two-sample and multiple sample univariate tests are used to establish intermediate phenotype relationships, that is, relationships between clinical variables and gene expressions.

The parametric tests are based on the assumption of normality which can be tested by the normality tests or a normal QQ plot as mentioned above. Furthermore, they are sufficiently robust to allow disregarding "*all but severe deviations from the theoretical assumptions*" (134). However, it is not always clear when the deviations are severe. To account for these situations, nonparametric tests are included. They do not assume a normal population, and are sometimes referred to as distribution-free methods. Also, since they are based on the ranks of data, they are generally more robust towards outliers than parametric tests. It should be mentioned that "*if either the parametric or nonparametric test is applicable, that is, if the data is normally*

distributed, then the former will always be more powerful than the latter" (134).

The two-sample t-test is used to test the null hypothesis that the means of two independent normally distributed populations, for instance smokers vs. non-smokers, are equal. Given two such "*data sets, each characterized by its mean, standard deviation and number of data points*", the two-sided t-test is used "*to determine whether the means are distinct, provided that the underlying distributions can be assumed to be normal*" (135), which as mentioned above, can be tested by the approaches described in method 1. If the calculated p-value is below the threshold chosen for statistical significance (see below), then the null hypothesis, stating that the two groups do not differ, is rejected in favor of an alternative hypothesis, which states that the groups do differ.
Different t-tests exist, but the applied t-test in this PhD is the Welch's t test, because it operates with unequal sample sizes and possibly unequal variances (136). Details about the t-test can be found in the references in this subsection and e.g. (137).

The nonparametric analog of the two-sample t-test, is the Wilcoxon rank-sum test also known as the Mann-Whitney U test (138), here in short, the Wilcoxon test. As with other nonparametric tests, the actual measurements are not employed, but the ranks of the measurements are used instead. As the t-test, the Wilcoxon test assesses whether two independent samples of observations come from the same distribution without, however, involving the assumption of normally distributed data. The test "*does assume that the two sample distributions have the same shape and differ only by a possible shift in location*" (139). "*The null hypothesis is that the two samples are drawn from a single population, and therefore that their probability distributions are equal*" (138). Details about the Wilcoxon test can be found in the references given above and e.g. (140).

In table 12, an example of the t-test and the Wilcoxon test applied to the dependent gene variable SERT and the independent clinical variables 'tobacco' and 'gender', is shown.

Clinical variable	**Wilcoxon test on SERT p-value**	**T-test on log(SERT) p-value**
Tobacco (non-smokers vs. smokers)	0.000352	7.62E-05
Gender (male vs. female)	0.27319	0.255765

Table 12: p-values from the Wilcoxon rank-sum test and t-test on the logarithm of SERT gene expression data from the ABS control group. For tobacco there is a significant difference (below 1%) in the SERT gene expression levels between non-smokers vs. smokers. There are no differences in the SERT levels between men and women. The analyses were done in R.

The parametric one-way ANOVA (ANalysis Of VAriance) test is used in the thesis to simultaneously compare more than two group means of a factorial

clinical variable based on independent samples from each group, for instance simultaneously comparing the gene expression levels of SERT between non-smokers, and medium and heavy smokers. Thus, the one-way ANOVA is basically an extension of the two-sample t-test applied to multiple groups (>2). "*The bigger the variation among sample group means relative to the variation of individual measurements within the groups, the greater the evidence that the hypothesis of equal means is to be rejected*" (141). The assumptions are

- normally distributed data - here meaning the logarithm of the gene expressions are used
- independent samples from each group
- variance homogeneity among groups - which is assumed. ANOVAs are "*robust, operating well even with considerable heterogeneity of variances, as long as the group sizes are equal or nearly equal*" (142).

In case of a significant result, the one-way ANOVA does not identify which group means differ. Further analysis must then be undertaken and various options exist, like applying Dunnett's test (141), or sometimes simply looking at a plot of the gene expressions in the different groups. Further details may be found in the references above and (143).

The nonparametric analogue to the one-way ANOVA test is the Kruskal-Wallis test, which is preferred when the gene expression data deviate severely from the underlying assumptions of the ANOVA. Furthermore, the Kruskal-Wallis test is only slightly influenced by differences in group variances (142).
The Kruskal-Wallis test is an extension of the Wilcoxon rank-sum test, described above, for more than two groups. Like the Wilcoxon test, it is based on ranks of the data and may be used in any situation where the parametric one-way ANOVA is applicable, however, then it will only "*be 95% as powerful as the latter*" (144). Details can be found in the references above and (145), (146).

In table 13, an example of the ANOVA and the Kruskal-Wallis test applied to the dependent gene variable ARRB1 and the independent clinical variables 'Coping' with four levels (never, sometimes, most days, or every day) and a 'Family history of depression, anxiety or suicide' with three levels (no relatives with any disease, secondary relative with any of the diseases, or primary relative with any of the diseases), is shown.

Clinical variable	Kruskal-Wallis test on ARRB1, p-value	ANOVA test on log(ARRB1), p-value
Coping (4 levels)	0.087172	0.001637
Family Dep/Anx/Sui (3 levels)	0.18548	0.121896

Table 13: p-values from the Kruskal-Wallis and the ANOVA test on the logarithm of ARRB1 gene expression data from the ABS control group. For coping there is a significant difference (ANOVA, below 1%) in the ARRB1 gene expression levels between the four levels of coping. There are no differences in the ARRB1 levels for the different levels of a family history of depression, anxiety or suicide. The analyses were done in R.

As seen in the table above for coping, the ANOVA p-value is significant while the Kruskal-Wallis p-value is not. Being an exploratory study, a result is considered of interest if either the parametric or the nonparametric analogue test result is significant.

Finally, it should be mentioned that I applied the conservative Bonferroni multiple comparison correction (147) after recommendation of Lundbeck's statistical department. The Bonferroni correction is one way of reducing the number of spurious positives, taking the number of comparisons being performed into account. In general, throughout the thesis the significance level is set to 1%, unless explicitly stated otherwise.

5.3 Repeated measures ANOVA

In the UK control group, three time measurements are made; Day 0 at 8 am, Day 0 at 2 pm and Day 1 at 8 am. In order to investigate whether any of the gene expressions differed significantly between the three time points, I applied the repeated measures ANOVA test. "*Special attention was given to these types of measurements because they cannot be considered independent. In particular, the analysis must take provisions for the correlation structure*" (148).

There are two main analytical approaches for handling repeated measures, a univariate approach and a multivariate approach (148). Here, focus is on the univariate approach as I have employed both approaches and found they yield similar results for the analyses done. The multivariate approach uses the repeated measurements as multivariate response vectors in MANOVA tests (Multivariate ANalysis Of VAriance); however, in my experience, the approach is more cumbersome to perform. Both approaches involve Mauchly's sphericity test (149): This test basically states that the variances of the differences between the repeated measurements should be about the same. If the sphericity assumption is violated, then either the Greenhouse-Geisser or the Huynh-Feldt (less conservative) corrected p-values (149) are calculated.

The univariate approach is based on the ANOVA test described in the previous section on univariate tests. Here the temporal aspect is explicitly included as a

factor in the ANOVA. This can be regarded as a two-way/two-factor ANOVA (time and subjects) (148). Just as with the ANOVA test, normality of the gene expressions is assumed (hence, I first take the logarithm of gene expressions before performing the repeated measures ANOVA) and variance homogeneity among groups are assumed. "*In addition, the univariate ANOVA approach requires that the each pair of repeated measures has the same correlation, a feature known as 'compound symmetry'*"[11] (148). The latter assumption is not valid in the UK control group as the time points are unequally spaced. However, as mentioned above this has virtually no impact on the conclusions. Further details on the repeated measures ANOVA are found in the references above.

Table 14 shows an example of the repeated measurements ANOVA in the UK control group for the CD8 beta and CREB1 three time point gene expression values. P-values for Mauchly's sphericity test are included.

p-values	CD8 beta; 3 time points	CREB1; 3 time points
Repeated measures ANOVA p-value	4.47E-05	0.2669
Mauchly's test p-value	0.8037	0.7627

Table 14: Repeated measures ANOVA results for the three time point CD8 beta and CREB1 expression measurements in the UK control group. Mauchly's sphericity test p-values are included and show that the variances of the differences between the repeated measurements are similar to each other. CD8 beta shows a significant difference between the three time points, while CREB1 does not. The analyses were done in R and SAS.

5.4 Correlations

The statistical department of Lundbeck has recommended the use of Spearman's nonparametric rank correlations due to their ability to handle non-normal data (see below) and outliers. I have applied the Spearman correlations to examine how correlated continuous clinical variables, like age and BMI, are to the gene expressions. The correlations are also used to compare control groups.

The Spearman rank correlations are particularly useful when the data is not normally distributed and, similar to Pearson correlation coefficients, they range from -1 to +1 and have no units. However, unlike the Pearson correlation coefficient, Spearman's rank correlation coefficient does not require the assumption that the relationship between the variables is linear (151).

In table 15, an example of the Spearman correlation is shown for the continuous clinical variables age and BMI vs. the gene expression values of CREB1.

[11] "*If compound symmetry is met then sphericity is also met*" (150).

Continuous clinical variable	Spearman rank correlation
Age	0.109376
BMI	-0.04609

Table 15: Spearman rank correlation coefficients for the continuous clinical variables age and BMI vs. the gene expression values of CREB1 in the ABS control group. Both correlations are weak; thus, CREB1 are practically uncorrelated with both clinical variables. The analyses were done in R.

Comparing two Spearman correlation coefficients can be done by first applying the Fisher z transformation (152), (153). Thereby, correlation coefficients between 0 and 1 are transformed to the corresponding values of Fisher's z between 0 and ∞, while correlation coefficients between 0 and -1 are transformed to z values between 0 and -∞. This "*both normalizes the underlying distributions of each of the correlation coefficients, and stabilizes the variances of these distributions*" (154). Then a test statistic is compared to the Student's t-distribution to determine if the null hypothesis of no difference between the correlation coefficients is true. Details on how to compare correlation coefficients and the Spearman correlations may be found in several of the references given in this section, e.g. (152).

In table 16, an example of comparing two Spearman correlation coefficients is shown. Here, gene expression correlations are compared between two control groups – the Danish (DC) and 'super-healthy' Americans (SH ABS).

Gene expression pair	DC Spearman correlation, N=89	SH ABS Spearman correlation, N=59	Comparing correlations, p-value
ERK2-ARRB2	0.4955	0.7700	0.0055
ERK2-DPP4	0.0829	0.1858	0.5413

Table 16: Comparing Spearman correlation coefficients between the DC and SH ABS control groups. One comparison (ERK2-ARRB2) is significant (below 1%) while the other is not. The analyses are done in R.

5.5 Canonical correlation analysis

Canonical correlation analysis (CCA) is an exploratory multivariate statistical method "*that facilitates the study of linear interrelationships among sets of multiple dependent variables* (the gene expressions) *and multiple independent variables*" (the clinical variables). CCA "*simultaneously predicts multiple dependent variables from multiple independent variables*" (155). It can use both metric and nonmetric data for either the dependent or independent variables. I have used CCA for subtyping /phenotype identification purposes in order to identify possible phenotypes among the borderline patients, by looking at the CCA relationships between the patients' clinical variables on one side and their gene expressions profiles on the other side.

CCA "*deals with the association between composites of sets of multiple dependent and independent variables. The approach develops a number of independent canonical functions that maximizes the correlation between the linear composites, also known as canonical variates, which are sets of dependent and independent variables. Each canonical function is thus based on the correlation between two canonical variates, one variate for the dependent variables and one variate for the independent variables*" (155). Each pair of canonical variates (each canonical function) seeks to maximize "*the same correlation subject to the constraint that they are to be uncorrelated with the previous pair of canonical variates*" (156). Thus, "*the first pair of canonical variates is derived so as to have the highest intercorrelation possible between the two sets of variables. The second pair of canonical variates is then derived so that is exhibits the maximum relationship between the two set of variables (variates) not accounted for by the first pair of variates. Successive pairs of canonical variates are based on residual variance, and their respective canonical correlations become smaller as each additional function is extracted*" (155).

Classical CCA is recommended to be performed with at least 10 observations per variable (155) making it practically useless for the Lundbeck data. However, in R the CCA package (124) implements a regularized version of CCA to handle such situations and intended for microarray studies or other studies where the number of variables exceeds the number of observations.

CCA has the following assumptions;

- linearity among the variables meaning that nonlinear relationships will not be captured
- normality is not required but desirable (155); the logarithm of the gene expressions are used
- homoscedasticity (homogeneity of variance) is assumed
- "*multicollinearity* (two or more variables being highly correlated) *among either variable set will confound the ability of CCA to isolate the impact of any single variable, making interpretation less reliable*" (155); this is an inherent problem with gene expression data and affects all multivariate techniques. Correlation among variables can be computed, e.g. with the Spearman correlation described previously.

Details on CCA can be found in the references above.

I will now give an example of regularized CCA performed in R. Here, I will consider the clinical variables and gene expressions of the first cohort of borderline patients, and describe a few examples of how CCA may be interpreted. The R package offers a novel graphical output to facilitate interpretation of a CCA. In this example, I did CCA with a subset of clinical variables and gene expressions. 4 gene expressions were chosen by a

univariate comparison with the SH ABS controls applying the Bonferroni correction. 6 clinical variables were chosen by first performing CCA with all clinical variables (and the 4 gene expressions) and then leaving out the ones that did not seem important (see explanation below). This procedure is recommended in (155) and (124).

The results are summarized in the figure 8 and consist of two graphs; on the left the graph of variables and on the right the graph of individuals (the 11 borderline patients). The two graphs are connected, so that the graph of variables informs about the subject groupings in the graph of individuals.

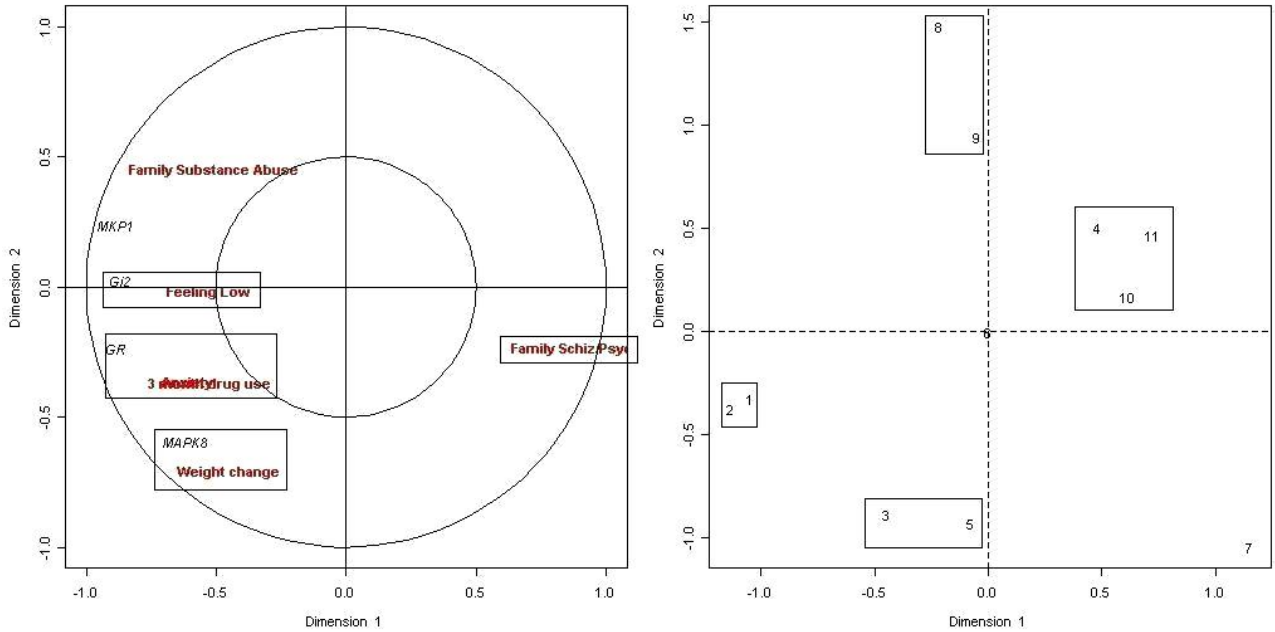


Figure 8 shows a CCA output for the borderline patients and a subset of 6 clinical variables (in red) and 4 gene expressions. There are two graphs; the graphs of variables to the left and the graph of individuals (with the 11 BPD patients) to the right with marked subgroups. The two graphs are connected; see the text after the figure. Dimension 1 and 2 represent the most and second most important canonical correlation dimension, respectively. The plots were done in R.

First, the graph of variables: Here two circles are seen - an inner circle with a radius of 0.5 and the outer circle of radius 1. Genes and variables between these two circles are the important ones according to CCA. As it can be seen there are no variables in the inner circle and this is because I have left out the unimportant variables. On this graph, "*variables* (both genes and clinical) *with a strong relation are projected in the same direction from the origin. The greater the distance from the origin, the stronger the relation*" (124). From this it can be seen that e.g. Feeling Low and Gi2 (shown in the middle left box) are strongly related and this also goes for MAPK8 and Weight Change in the third quadrant of the graph, while Family Schiz/Psyc (a family history of schizophrenia or psychosis) stands alone on the opposite side.

On the graph of individuals, different subgroups of borderline patients are shown. The two graphs are connected so that every direction, e.g. north and south or east and west in one graph corresponds to the other graph. An example: In the graph of variables, Gi2 and Feeling Low are closely related as are GR, 3 months drugs use and Anxiety in the middle left of the graph with Family Schiz/Psyc on the opposite side (as described above). In the graph of individuals, BP1 (borderline patient number 1) and BP2 are close together around the middle left (shown with a box) while e.g. BP4, BP10 and BP11 are close together at the opposite side. This can be interpreted as BP1 and BP2 have high Feeling Low, 3 months drug use, Anxiety, GR and Gi2 scores and a low Family Schiz/Psyc score at the same time. BP4, BP10 and BP11 have a high Family Schiz/Psyc score and low scores of the other variables and gene expressions. At the same time, all the borderline patients below the zero line (dimension 2), especially BP3 and BP5, have high MAPK8 expression and Weight Change scores while all the borderline patients above the zero line, especially BP8 and BP9, have low MAPK8 expression and low Weight Change scores. I have checked the gene expressions and clinical variables in the BPD data, and in general, CCA seems to reflect the patterns in data well, e.g. BP1 and BP2 always have the highest expression values for the four genes.

This example shows that CCA offers an exploratory and descriptive approach to identify possible phenotypes among the borderline patients. It can be hypothesized which borderline patients that resemble each other, measured by certain clinical variables (here Family Substance Abuse, Feeling Low, 3 months drug use, Anxiety, Weight Change and Family Schiz/Psyc) and gene expressions.

5.6 Recursive partitioning

Recursive partitioning (RP) is a non-parametric multivariate statistical method that creates a decision tree, also called a classification tree, "*that strives to correctly classify members of a population based on a dichotomous dependent variable*" (157). RP and decision trees for classification purposes are dealt with in the next chapter. Here focus is on the use of RP for identification of possible (intermediate) phenotypes with distinct gene expression profiles. I have used RP for such a purpose by identifying distinct gene expression profiles in the two control groups DC and SH ABS. These profiles could be hypothesized to belong to distinct intermediate phenotypes, see the Results chapter.

RP analyzes a set of genes jointly. "*RP first picks the gene (gene 1) most likely to separate sample labels based on the level of expression*" (233) and based on a so-called Gini splitting index (158). "*Both the gene and the threshold value are determined by the data. Then, RP may pick another gene if further improvement on classification performance can be achieved. This process can be visualized as a classification tree, in which the first branching at the top*

corresponds to gene 1, second-level branchings corresponds to gene 2, and so on. A node on the tree can be either a branching point or a terminal leaf" (233)

Some of the advantages of RP include the creating of intuitive tree models, and also "*allows varying prioritizing of misclassification costs in order to create a decision rule that has more sensitivity or specificity*" (157). A major disadvantage is that RP tends to overfit data, making it difficult to use the same decision rules on a new data set. It is possible to circumvent this problem to a certain degree, e.g., by applying a cross-validation scheme (see next chapter). Further details on RP are found in the references given above.

An example of recursive partitioning is shown in figure 9. Here RP is used to split 11 BPD patients (cohort 1 BPDs) and 296 ABS controls. All subjects are correctly placed (no misclassifications) in a decision tree created from just eight gene expressions (out of a maximum of 25). Using a maximum of six splits (genes), every control is correctly classified and the corresponding genes are easily identified.

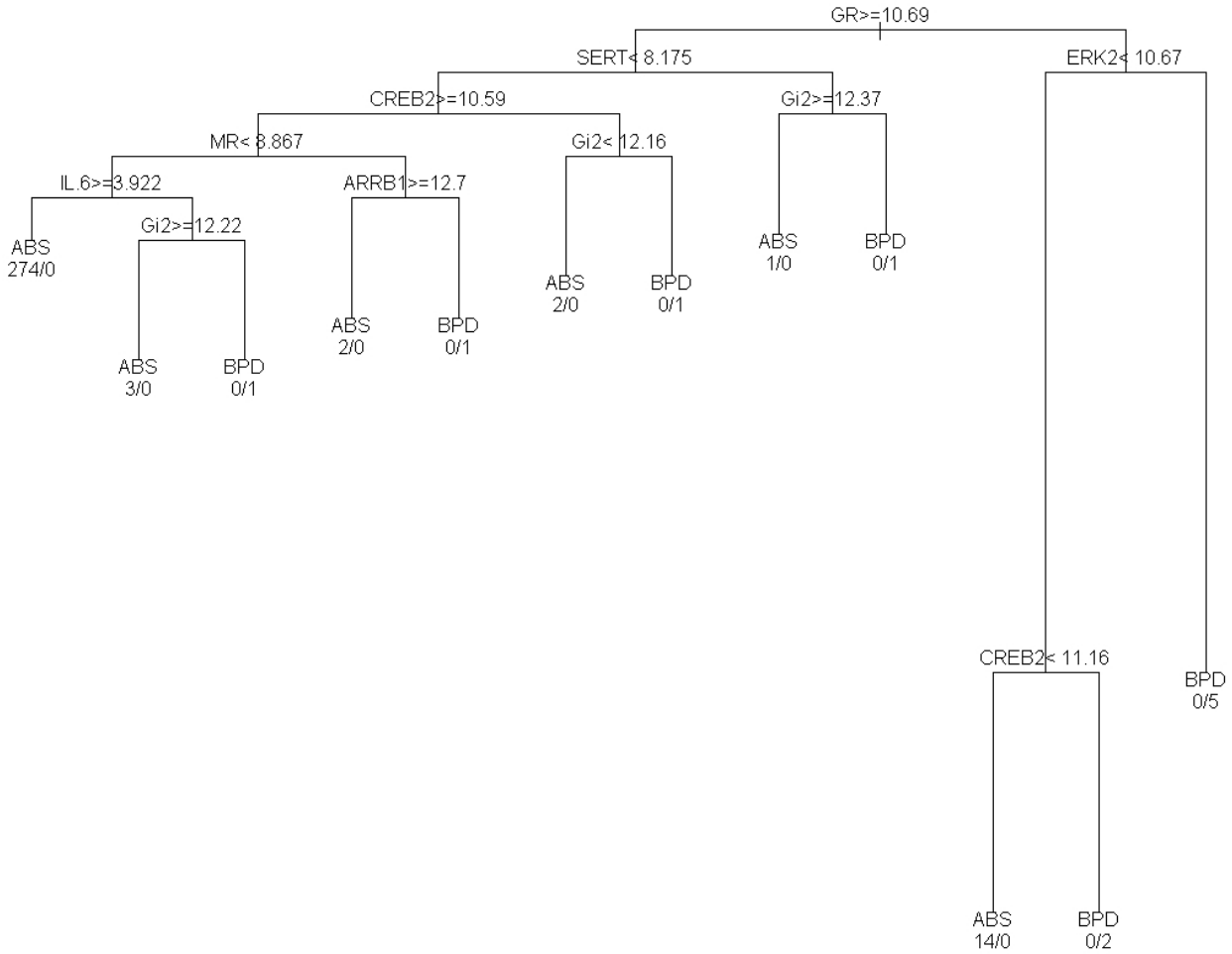


Figure 9 shows a decision tree of 11 BPD patients and 296 ABS controls. In the above tree, there are no misclassified controls. See the text after the figure for explanation of this decision tree. The 'rpart' package in R was used to carry out the recursive partitioning (158).

The right branch of the tree in Figure 9 shows that just by looking at the logarithm of the expression values of GR and ERK2, 5 of the 11 BPD patients can be correctly identified. 2 BPDs are classified by the values of GR, ERK2 and CREB2. Looking at the ABS controls, the expression levels of GR, SERT, CREB2, MR and IL-6 (left tree branch) identifies 274 controls. 14 ABS controls are identified by just GR, ERK2 and CREB2 (right branch). In this way, RP may be used to identify distinct phenotype gene expression profiles among the BPD group, and be used to identify distinct intermediate expression profiles among the ABS control group.

5.7 Clustering and heat maps

The term cluster analysis refers to a set of exploratory multivariate techniques whose primary purpose is to group objects, here genes or subjects, based on the characteristics they possess, that is pattern recognition and grouping is performed. Cluster analysis can group objects "*into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters. The attempt is to maximize the homogeneity of objects within the clusters while also maximizing the heterogeneity between the clusters*" (159). The ultimate goal is to identify naturally occurring clusters in the data, if possible. Cluster analysis is an unsupervised statistical method in the sense that it does not make use of class labels. This make clustering potentially very interesting since by clustering gene expressions, (intermediate) phenotypes can emerge without imposing any a priori hypotheses on the technique.
There are many different approaches to cluster analysis with the two main types referred to as hierarchical methods (the results of which is represented as a hierarchical tree structure) and partitioning methods (which separate the objects into a given number of clusters). Other types of clustering methods exist as well (160), (161), but here focus is on agglomerative hierarchical clustering. "*Hierarchical algorithms can be agglomerative or divisive. Agglomerative algorithms begin with each object as a separate cluster and merge them into successively larger clusters until all objects have been joined into one large cluster. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters*" (160).
A heat map is a two-way hierarchical agglomerative clustering that combines clustering of genes with clustering of subjects in one map, called a heat map because colours in the map show the relative gene expression levels, see the example at the end of the section. Hierarchical algorithms have been used extensively in the analysis of DNA microarray data (115), (162), (161). I have used agglomerative hierarchical clustering and heatmaps to identify possible phenotypes among both controls and patients.

"*An important step in any clustering is to select a distance measure, which will determine how the similarity of two objects is calculated. This will influence the*

shape of the clusters, as some objects may be close to one another according to one distance and further away according to another" (160). Two of the most commonly used for gene expression data are the Euclidean distance and Pearson correlation coefficient (162). I have used both. The Euclidian distance – unlike correlation - is sensitive to scaling and differences in average expression level, which I why I have standardized (z-scores) the data before clustering.

Another important step in clustering is the choice of the actual clustering algorithm. Several choices exist, like complete linkage clustering (maximum distance between elements of each cluster), single linkage clustering (minimum distance between elements of each cluster), average linkage clustering (mean distance between elements of each cluster) and Ward's method (seeks to join clusters whose merger leads to the smallest within-cluster variance) (160), (163). I have used average linkage and Ward's criterion to merge clusters.

Clustering is recommended to be performed in a supervised manner (159) in the sense that clustering variables, here genes, should be selected prior to clustering and, thus, not include undifferentiated variables. In the heat map example below, I use the genes, separating control and patient groups, for clustering.

Cluster analysis is not a statistical inference technique, and thus has no requirements of normality, linearity or homoscedasticity. However, substantial multicollinearity should be avoided, if possible (159).

Some of the advantages of hierarchical clustering include the simplicity by which interpretation of the results can be made from a dendrogram (hierarchical tree), the deterministic calculations, and the speed of performing the clustering. "*A disadvantage is that hierarchical clustering can be misleading because undesirable early combinations may persist throughout the analysis and lead to artificial results*" (159). Furthermore, hierarchical clustering is sensitive to the choice of similarity measure and specific clustering algorithm, meaning different choices yield different results. Further details on hierarchical clustering are found in the references given in this section.

In figure 10, I show an example of hierarchical clustering in the form of a heat map consisting of two dendrograms, one for the genes and one for subjects. The heat map is generated using 21 subjects randomly chosen from a pooled control group (DC, SH ABS and PTSD controls) and the 21 borderline disorder (BPD) patients. Only genes that separate the three groups from one another are used. In the heat map the expression profiles of two distinct BPD clusters are identified.

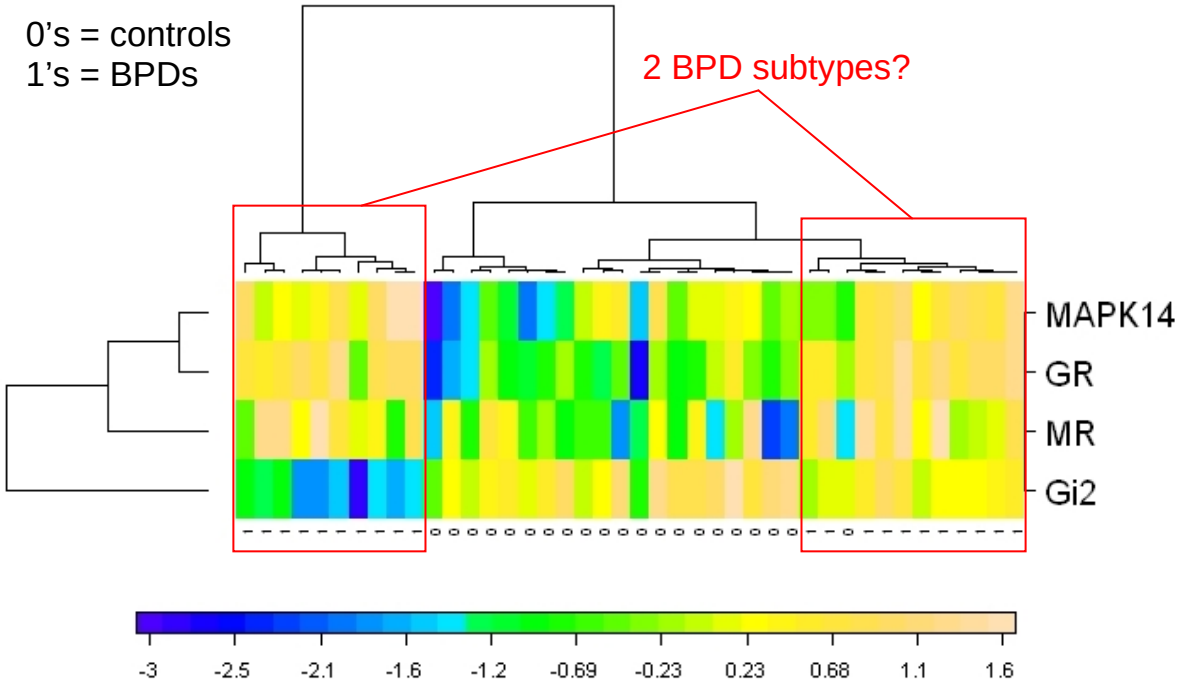


Figure 10 shows a heat map with 21 subjects from a pooled control group (DC, SH ABS and PSTD controls) and the 21 borderline disorder patients clustered at the top of the heat map. At the bottom, I have included the class labels showing two different clusters of BPD patients. These clusters may represent different phenotypes. Genes separating the control and patient groups are shown to the right. The heat map was done in R.

Finally, it should be mentioned that (unsupervised) clustering can be combined with (supervised) classification (described in the next chapter); genes may be hierarchically clustered, and then, for instance, discriminant analysis (multi-class separation) or stepwise logistic regression (two-class separation) may be applied to the clinical variables. Thus, the gene clusters may be linked to the clinical variables.

5.8 Stepwise regression

Stepwise regression is an automatic procedure in a multivariate setting suitable for establishing a linear relationship between a response (here, a gene expression) and the most important explanatory variables. Stepwise regression deals with the situation of not being able to precisely formulate a model based upon the clinical predictors, and the stepwise part narrows down the list of possible explanatory predictors. If it was possible to explain around 15% (164) of the variance in a gene expression by a linear combination of clinical variables that was to be considered a good result. By performing stepwise regression, I wanted to find out whether this also applied to the Lundbeck data and, if so, which clinical variables were the most important predictors. In this way, I have used stepwise regression as a hypothesis generating technique to identify a minimal set of clinical variables per gene expression that explain as much of the variance in a gene expression as possible.

First, a full linear model is specified with the logarithm of a gene expression equal to the sum of all depression-relevant explanatory variables. Then, stepwise regression as a combination of backwards elimination and forward selection is performed (165), (166). Backward elimination "*involves starting with all candidate variables and testing them one by one for statistical significance, and then deleting any that are not significant*" (166). Statistical significance is evaluated by the Akaike information criterion (167) which basically is "*an operational way of trading off the complexity of an estimated model against how well the model fits the data*" (234). Forward selection then involves starting with the model left from a previous backwards elimination step, trying out the left-out variables one by one and including them if they are statistically significant at the 1% significance level.
The result is a minimal linear model with the full linear model reduced to a minimum of variables significantly explaining the same amount of variation in each gene expression as the full model.
A second step is added in that a linear interaction model is made that contains the minimum set of clinical variables together with the sum of all pair-wise interactions between these variables. Stepwise regression is performed again, and the result is a minimal linear interaction model containing the minimum set of clinical variables and significant interactions explaining essentially the same amount of variation in each gene expression as the original linear interaction model. Further details on stepwise regression are found in the references given above and (168), (169). An example of the results of a stepwise regression involving the ABS control group is given in table 17.

Clinical variable	**Log(ERK2)**	**Log(ADA)**
Appetite change	X	
Feeling low : Sleep problems	i	
Sleep problems	i	
BMI		X
R^2 without interactions	14%	3%
R^2 with interactions	18%	3%

Table 17: Results of a stepwise regression on the depression-relevant clinical variables for the two gene expressions ERK2 and ADA in the ABS control group. R^2 is reported. 18% of the variance in ERK2 is explained by appetite change and a significant interaction between 'feeling low' and 'sleep problems'. 3% of the variance in ADA is explained by the single variable BMI. An 'X' denotes the clinical variable(s) remaining after a stepwise regression, and an 'i' denotes one part of an interaction term remained after a second stepwise regression. The analyses were done in R.

Stepwise regression has the advantage of producing a minimal set of explanatory predictors for a continuous response. A drawback of the stepwise regression approach is a tendency to overfit the data (165).

5.9 Other exploratory statistical methods

As mentioned in the beginning of this chapter, I have applied a number of statistical methods to the available data and described the reasons for using specific procedures. In table 18, I briefly describe a few other exploratory methods (including references) that even though their output did not seem relevant to Lundbeck, I found them of potential interest.

Other statistical methods	Description	Reference
Two-way ANOVA	Two-way ANOVA may be used to study the effects of two independent clinical variables with each variable having several levels.	(143)
MANOVA	Multivariate analysis of variance (MANOVA) is an extension of ANOVA methods to cover cases where there is more than one dependent variable (gene expression). As well as identifying whether changes in the independent clinical variables have a significant effect on the gene expressions, the technique also seeks to identify the interactions among the independent clinical variables and the association between the gene expressions, if any. In R, MANOVA is implemented in the package 'ffmanova' designed to handle collinear responses. MANOVA may be used to study certain combinations of gene expressions, e.g. biologically or pathologically relevant gene expression combinations, together with chosen clinical variables.	(170) (171)
PCA	Principal component analysis (PCA) is an unsupervised technique used to reduce multidimensional data sets to lower dimensions for analysis. PCA results in a number of principal components with the first component explaining the largest variance in the data set, the second component the second greatest variance, and so on. Each component is orthogonal to the previous one, and consists of a linear combination of all gene expressions. PCA may be used to investigate whether it is possible to separate different control and patient groups based on the principal components.	(172), (173)
SPCA	Sparse principal component analysis (SPCA) is a microarray intended variable selection method based on PCA with sparse loadings. It has the advantage of identifying sparse principal components that, unlike PCA, do not overlap, making interpretations easier. SPCA may be used in the same way as PCA described above, however, with clearer interpretations of the principal components.	(174)

PLS	Partial least squares (PLS) is a class of supervised "*methods for modelling relations between sets of observed variables by means of latent* (i.e. not observed or measured) *variables. It comprises of regression and classification tasks as well as dimension reduction techniques and modelling tools. The underlying assumption of PLS methods is that the observed data is generated by a system of process which is driven by a small number of latent variables*" (176). PLS is similar to CCA, where latent vectors with maximum correlation are extracted. PLS generates orthogonal vectors by maximizing the covariance between different sets of variables. PLS may thus be used as an alternative to CCA to identify possible (intermediate) phenotypes taking both gene expressions and clinical variables into consideration.	(175), (176)
AP clustering	Affinity propagation (AP) clustering is a novel clustering technique that works by passing messages between data points. "*At any point in time, the magnitude of each message reflects the current affinity that one data point has for choosing another data point as its 'exemplar'*" (177), i.e. a cluster center selected among actual data points. AP clustering has the advantage of only requiring a similarity measure in order to derive the 'naturally' occurring number of clusters in a data set. Thus, unlike K-means clustering, the number of clusters is not specified beforehand.	(177)

Table 18: Other potential relevant exploratory statistical methods including a short description and references.

6. Classification with variable selection

Classification with automatic variable selection offers a supervised approach to prediction of diagnosis and future events, e.g. response to treatment, as well as an algorithm-oriented approach to extracting the variables responsible for group/class separation and prediction. This chapter investigates several classification methods with related variable selection techniques. As in the last chapter, both traditional statistical approaches and microarray approaches to classification are considered.
This chapter also presents a simulation project intended to clarify which classifiers and variable selection methods are the most promising at prediction, classification and variable selection with the available Lundbeck data. The simulation study consists of two phases. In phase 1, twelve different classifiers (some with automatic built-in variable selection, some without) and two different variable selection methods are tested on a simulated data set with all variables drawn from a normal distribution. Phase 1 consists of 42 linear and nonlinear tasks (listed in appendix 6). It should be stressed that we do not know which kind of gene interactions that exist in reality, so we decided to try several simple and classical combinations as shown in appendix 6. The performance of the classifiers and variables selection methods is evaluated by the accuracy measure in a 10-fold cross-validation scheme and by the Jaccard score which indicates how well a classifier identifies the correct explanatory variables. The result of phase 1 is a list of classifiers and a variable selection method that, in general, performs badly. In phase 2, the remaining classifiers and the other variable selection method are tested on a realistic data set consisting of three control groups (the Danish, the 'super-healthy' American and the post-traumatic stress disorder (PTSD) control group) as well as the borderline personality disorder patients and acute post-traumatic stress disorder patients. 33 linear and nonlinear tasks are given to the classifiers and variable selection method (listed in appendix 7). The result of phase 2 is, in the two-group case, the recommendation of the classifiers; stepwise logistic regression, classification tree (RPART), and support vector machine combined with variable selection based on random forests. The last two classifiers are also recommended in multiple-group case.
I set up a classification and variable selection procedure suitable for real data sets with different group sizes and used to decide whether a group separation is possible or not based on permuted accuracies and univariate tests. An example is given demonstrating the two-group advised classifiers on a real data set.
Stepwise logistic regression, variable selection based on random forests and support vector machines are described in some detail in section 6.4.

In the preceding chapter on Statistical Methods, I mentioned that qPCR gene expression data analysis ranges between traditional statistics and microarray

analysis as classified by the number of samples and variables, and that for most groups in the present study, the ratio of the number of subjects to the number of gene expressions is around only 1-3, making it necessary to consider both microarray approaches and traditional statistical methods. This also applies to classification, and that is why I will consider both classical statistical approaches and microarray approaches to classification in the present chapter.

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 (178) as well as an multivariate algorithm-oriented approach to extracting the variables responsible for class separation and prediction. In the Statistical methods chapter, section 5.2, several univariate approaches, like t-tests and Wilcoxon tests, are described capable of identifying variables separating groups. These "*univariate methods 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*" (178). Multivariate variable selection approaches, on the other hand, recognize that the subset of variables with best univariate discrimination power are not the best subset of classification variables, and try to determine which combinations of variables yield high prediction accuracies. After agreement with both Danish and US Lundbeck co-operators, I looked into classifiers and automatic multivariate variable selection methods. This was done to explore the best possible classifiers and feature selection methods for the available gene expression data.

In the beginning of the classification work, I tested a few classifiers in R and identified two that seemed to perform well on the available data – 'Pelora' and 'SLR' (see below) – and have used these two classifiers for a range of classification tasks presented in the results chapter.
Pelora (179) is a microarray analysis intended algorithm based on Penalized Logistic Regression that "*combines gene selection, gene grouping and sample classification in a supervised approach*" (179). Being based on logistic regression, Pelora is intended for two class/group classification tasks, and has relaxed the assumptions of normality. Through a mean-based approach of a linear combination of predictor variables, Pelora seeks to explain the binary outcome. Penalization methods allow distinguishing irrelevant from relevant classification variables through modifying their coefficients, and thus perform intrinsic feature selection. Details on Pelora are found in the reference above as well as in (180). An example of the output of Pelora is given in figure 11 where a clear separation is seen between the two control groups, the 'super-healthy' Americans (SH ABS) and the Danes (DC).

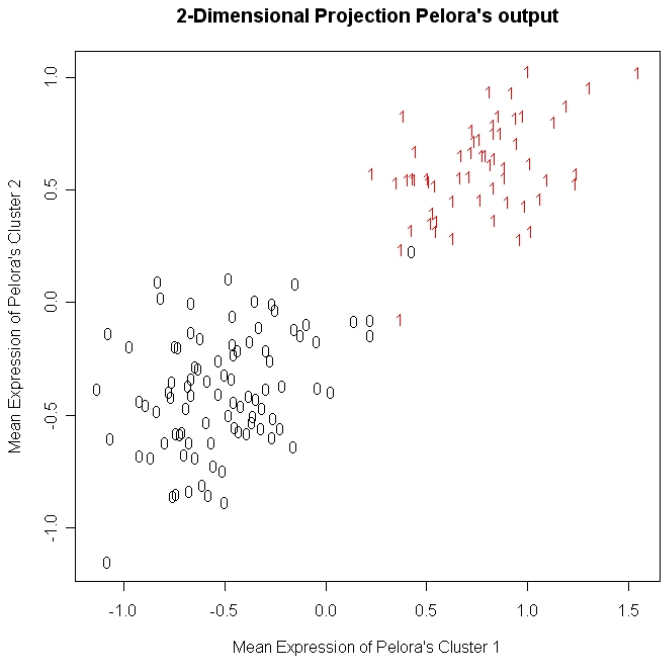


Figure 11 shows in a Pelora plot a clear separation of the two control groups; the DCs (0's) and the SH ABS (1's). On the x-axis is the mean of a scaled set of Pelora cluster 1 genes; ADA, CREB1, CREB2, MAPK14 and ODC1. The plot shows that this set of genes are sufficient for separation of the control groups. The Pelora plot was done in R.

SLR is short for Stepwise Logistic Regression, and like Pelora, SLR is also intended for response variables with a binary outcome. Apart from the binary outcome and the relaxed assumptions of normality, SLR resembles the stepwise regression method described in the previous chapter in that forward selection and backward elimination is carried out, and variable importance is evaluated with the Akaike information criterion. The reader may consult section 5.8 in the previous chapter for a description of the stepwise approach. Using SLR on the previous example yields the output gene list; ADA, CREB1, Gs and MAPK8 with an (LOOCV – Leave-One-Out Cross-Validation (181)) error rate of only 2%. A certain gene overlap between Pelora and SLR is noticed (ADA and CREB1). This example clearly demonstrates that different classification algorithms may yield good predictive results with different sets of predictor variables for the same data set.
Finally, I should mention that SLR – unlike Pelora - performs well with clinical variables as predictors. This is demonstrated in the results chapter, section 7.5.3 and 7.6.

It soon turned out, that despite the initial promising Pelora and SLR results, these classification algorithms contained some inherent limitations;

1. Per default, Pelora and SLR only consider a linear combination of the predictor variables, meaning predictor nonlinearities are not taken into account.

2. The two classifiers can only predict a threshold-based outcome (y), that is, an outcome that is above or below a threshold, meaning that an interval-based outcome can not be discovered. In the simple case, an example of an interval-based outcome is a scaled gene expression with a value of less than -1 leading to a disease state, a value between -1 and +1 being the healthy state, and a value above +1 again leading to a disease state (see more details in the Simulation study section below).

These limitations and the availability of a broad range of classifiers in R, led to a comprehensive simulation study.

6.1 Simulation study

In a close co-operation with Jan Bastholm Vistisen from Lundbeck Denmark, we decided to explore Pelora, SLR and a wide variety of linear and nonlinear classifiers and related automatic feature selection methods in more detail. I wanted to identify some classifiers and feature selection methods that would perform well with the available data, and that could take different kinds of gene interactions into account (see below). We decided to set up a simulation study to investigate which classifiers and feature selection methods that we would trust to analyze the qPCR data. Key questions were;

- Which classifiers can identify the correct explanatory genes?
- How good are the different classifiers at classifying and predicting new subjects?

Simulated data sets allow us to create data sets where we define the outcome, that is

- We know which combination of variables is used to define the outcome.
- The outcome can be defined in any desired form like be threshold- or interval-based.
- We wanted to explore various roads to a disease state, so we decided to create different variable combinations of interest; linear, ratio and product combinations of genes. According to Lundbeck Research, ratio and product gene combinations may be of biological interest (see below).

Here, it should be stressed that we did not know which kind of gene interactions that exist in reality, so we decided to try several simple and classical combinations as described above.

We decided to divide the simulation study into two phases. In phase 1, the simulated data sets had approximately the same amount of variables as in the real data; 30 variables were used. The number of samples per data set was set

to both N=100 and N=1000, which should mimic more or less the number of samples in the real data set at that time as well as the size of data set to be analyzed in the 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 were that in phase 1 all variables were drawn from a normal distribution, while in phase 2 a realistic data set was considered, see the phase 2 section later in this chapter.

In phase 1, we wanted to rule out classifiers that could not solve tasks, we deemed important, and came up with 42 linear and nonlinear tasks, see appendix 6. Here it should be noted, that had we focused on other tasks (than the ones in appendix 6 and 7), different results would probably have been obtained.

The major aspects of the phase 1 tasks were;

- As a start, the outcome was just a function of one variable above a threshold or in an interval. This was done to understand, how the classifier would perform with simple tasks.
- The outcome was then a function of different combinations of two or five variables in a linear, ratio or product manner and always either above a threshold or in an interval.

Three separate scenarios were performed:

1. Different magnitudes of two involved variables were tested ($X_1 \approx X_2$, $X_1 \approx 10*X_2$ and $X_1 \approx 100*X_2$). This was done to see whether the magnitude of involved variables played an important role.
2. Different fractions of data points classified as Y=1 (0.05, 0.20 and 0.50). This was very relevant to us, as in some cases, the number of patients was much smaller than the number of controls compared with.
3. Two populations with different mean values in gene no. 1 ranging 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))
   This was done to see how small a difference, the classifiers could detect.

Tested classifiers and automatic feature selection methods
We decided to look at broad range of classifiers. Based on classifier course material from the CBS course 'DNA Microarray Analysis' and various available classifiers in R, I came up with a list of classifiers that operated based on (more or less) different algorithms. Focus was on classifiers with either built-in variable selection or with variable selection as a pre-step to classification.

The following linear and nonlinear classifiers and variable selection methods were tested in phase 1 - in general, not in all cases, the default options in R for each classifier were used:

1. Pelora with only the first Pelora cluster used to avoid overfitting (180), (179).
2. SLR (168) and (182).
3. PLR - Penalized Logistic Regression with a stepwise variable selection (183).
4. RPART - Recursive PARTtioning (classification tree), see section 5.6 in the previous chapter and (158).
5. NB[12] – Naive Bayes is a standard classifier known to perform well (186) and (187).
6. LDA[12] – Linear Discriminant Analysis is a classical classification method (188) and (187).
7. SKNN[12] – Simple K Nearest Neighbor. K-NN is described in (115), (189) and (187).
8. Random Forest[12] (190), (191) and (192).
9. QDA[13] – Quadric Discriminat Analysis is also described in (115), and (194), (193).
10. SVM[13] – Support Vector Machines (193) and (195).
11. NNET[13] – Neural NETwork with a single-hidden-layer (193). For neural network theory, see (196).
12. LogitBoost[13] – a new boosting machine learning technique (193) and (197).

The interested reader may consult the references for details on these classifiers.

Accuracy, cross-validation and the Jaccard similarity coefficient
In order to determine the performance of a classifier for both two and multiple (phase 2) class tasks, I decided to measure the accuracy (198):

$$accuracy = \frac{\#true\ \ positives\ +\ \#true\ \ negatives}{\#true\ \ positives\ +\ \#false\ \ positives\ +\ \#true\ \ negatives\ +\ \#false\ \ negatives}$$

The accuracy was measured in each cross-validation sample and finally averaged. As recommended in (199), 10-fold stratified[14] cross-validation was

[12] Initial testing demonstrated that the variable selection method (varselrf (184),(185)) performed very well, so the classifiers NB, LDA, SKNN and random forest were started of with the varselrf-selected variables.
[13] QDA, SVM, NNET and LogitBoost were tested with two different variable selection methods: msc (based on mass spectra classification (193)) and varselrf (variable selection based on random forests).
[14] Stratified cross-validation means that each fold contains approximately the same ratio of patients to controls as is present in the entire data set.

used. This also explicitly meant that the accuracy measured would not be inflated due to overfitting.

To measure how well a classifier identified the correct variables, the Jaccard similarity coefficient (200), recommended in the 'DNA Microarray Analysis' course, was used (binary form):

$$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

"M_{11} represents the total number of attributes where the binary vectors A and B both have a value of 1.
M_{01} represents the total number of attributes where the attribute of A is 0 and the attribute of B is 1.
M_{10} represents the total number of attributes where the attribute of A is 1 and the attribute of B is 0" (200).

The Jaccard score yielded a number (0-100%) indicating how well a classifier identified variables compared to the correct explanatory variables as we had defined them. The higher the Jaccard score, the better an agreement between the two.

Phase 1 results
Now I present a few of the results from phase 1 in the form of plots with the x-axis representing the accuracy and the y-axis the Jaccard score. 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 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 12, the 16 classifier possibilities are shown for a task involving a linear combination of two variables with the outcome above a 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.

In figure 13, a linear combination of predictor variables is explored as well. However, the correlation between the variables is now 0.5, and the outcome is in an interval. Only four classifiers solve this tasks well; RF.RF, SVM.RF, QDA.RF and NNET.RF.

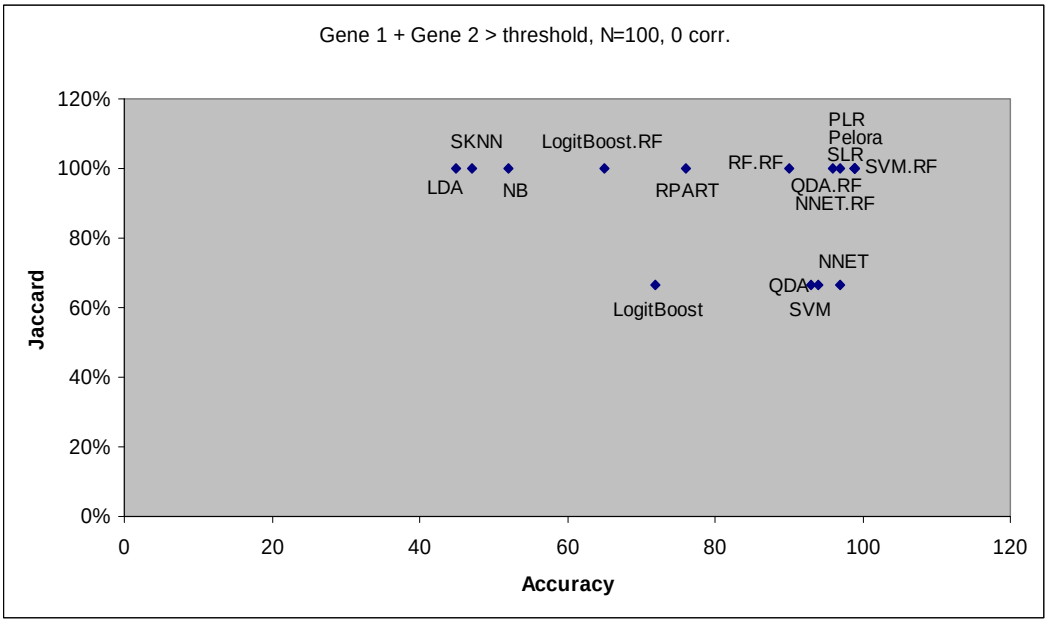


Figure 12 shows the 16 classifier possibilities solving the task shown at the top of the figure. The best performing classifiers are: PLR, Pelora, SLR, SVM.RF, QDA.RF, NNET.RF, RF.RF and RPART. The figure was made in Excel based on output from R.

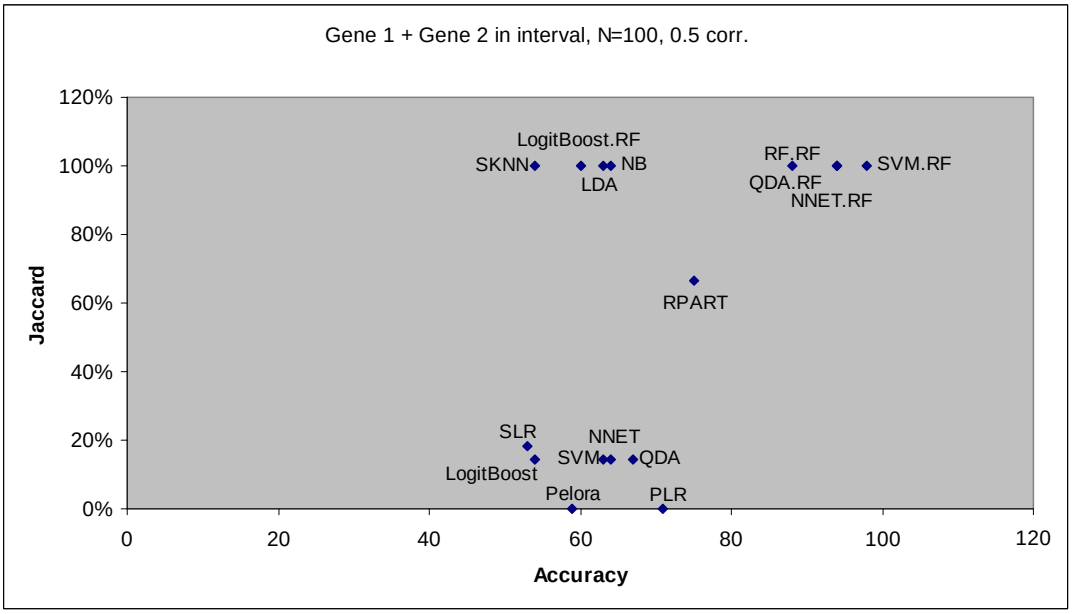


Figure 13 shows the 16 classifier possibilities solving the task shown at the top of the figure. The best performing classifiers are: RF.RF, SVM.RF, QDA.RF and NNET.RF. The figure was made in Excel based on output from R.

In figure 14, the ratio between two variables is explored for a large data set (N=1000) with 0.5 correlation between the variables, given a threshold. The best performing classifiers were; SVM.RF, RF.RF, RPART and QDA.RF. The msc variable selection method could not solve this task as all.

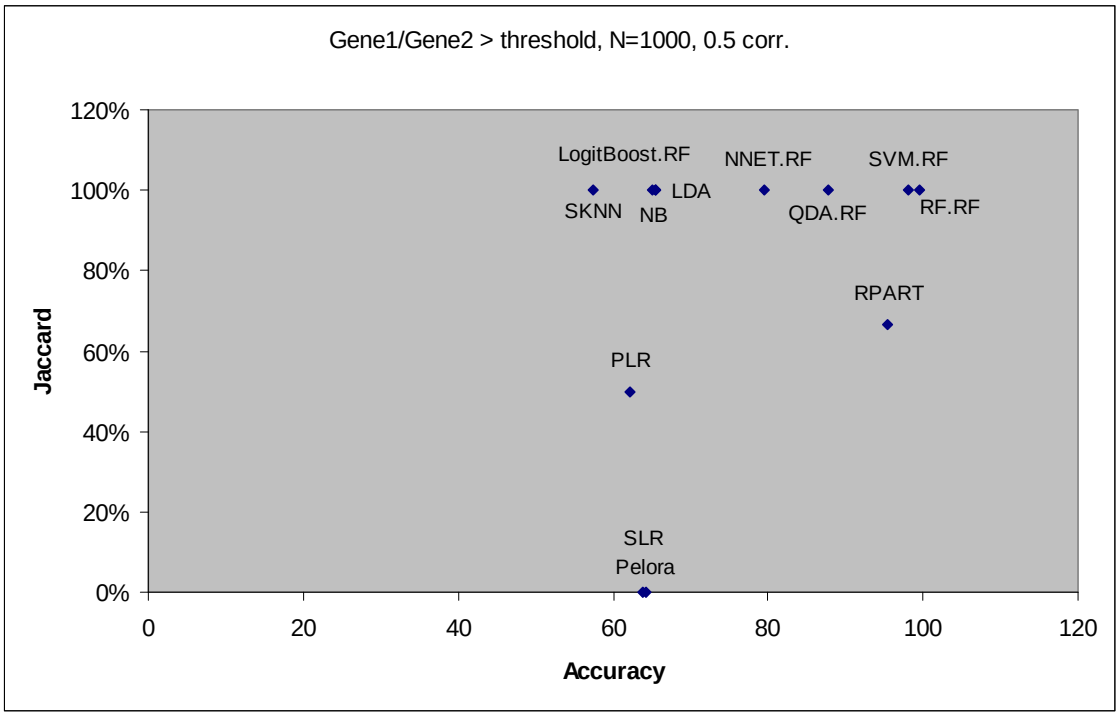


Figure 14 shows 12 classifier possibilities solving the task shown at the top of the figure. The best performing classifiers are: SVM.RF, RF.RF and QDA.RF. The msc variable selection method could not solve this task. The figure was made in Excel based on output from R.

In figure 15, the product between two variables is explored for N=100 in an interval with no correlation between the variables. The best performing classifiers were; SVM.RF, RF.RF and QDA.RF.

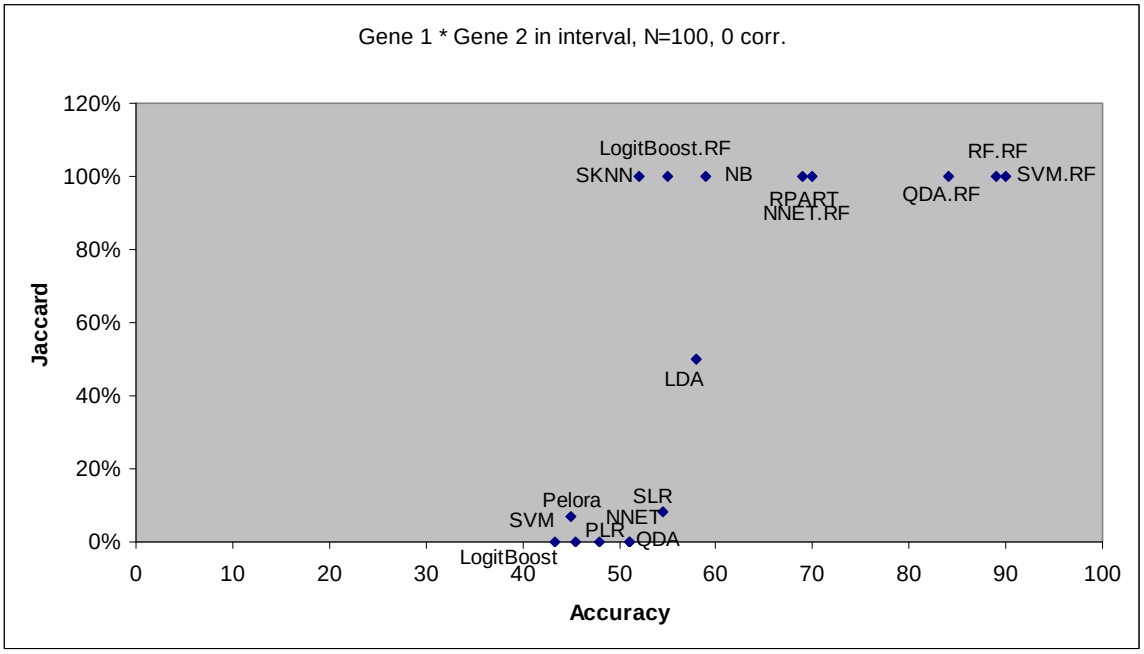


Figure 15 shows 16 classifier possibilities solving the task shown at the top of the figure. The best performing classifiers are: SVM.RF, RF.RF and QDA.RF. The figure was made in Excel based on output from R.

Finally, in figure 16, a linear combination of five variables is explored for N=100 with 0.5 correlation between the variables, given a threshold. The best performing classifiers were; Pelora and SLR.

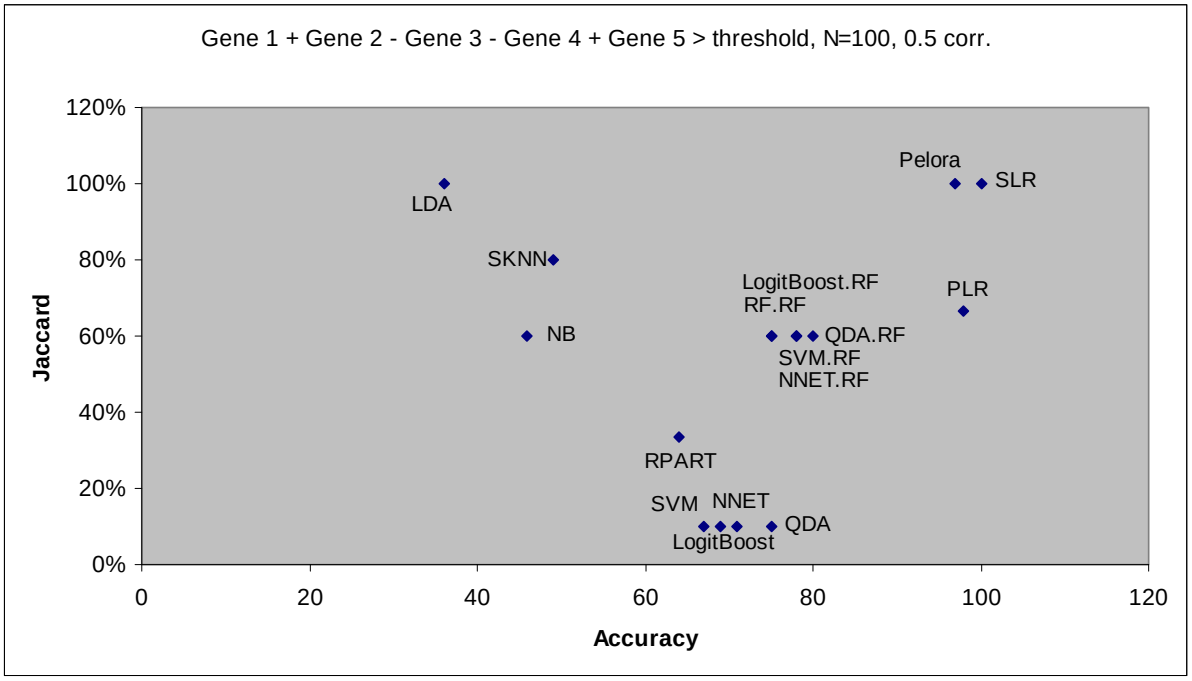


Figure 16 shows 16 classifier possibilities solving the task shown at the top of the figure. The best performing classifiers are: Pelora and SLR. The figure was made in Excel based on output from R.

By looking at the 42 plots (available on request) corresponding to the different linear and nonlinear phase 1 tasks, we concluded;

- 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 – 5 variable ratios
    - "Large products" – 5 variable products
    - 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
- LDA
- SKNN
- QDA (msc selection method)
- SVM (msc selection method)
- NNET (msc selection method)
- LogitBoost (msc and varselrf selection methods)

and would not be part of phase 2 tested classifiers. The reasons for poor performance were typically these classifiers and selection methods 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.

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

1. Pelora with only the first Pelora cluster used
2. SLR
3. PLR
4. RPART
5. QDA[15]
6. Random Forest[15]
7. SVM[15]
8. NNET[15]

The phase 2 data set consisted of the DC, SH ABS and PTSD control groups as well as the borderline personality disorder and acute PTSD patients. 25 variables/genes were included. The number of samples was 263. All data was normalized the same way; the variables used were in one case z-score standardized, as in phase 1, and in another case, real unstandardized expression values. As in phase 1, the outcome was defined as different combinations of the variables.

33 different tasks were given to the classifiers (see appendix 7) with most of them similar to the tasks in phase 1:

- 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 manner and always either above a threshold or in an interval.

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

Four separate studies were performed:

1. Completely random outcome – how would the classifiers / variable selection method perform in this situation?
2. Different fractions of data points classified as Y=1 (0.05, 0.20 and 0.50)
3. Actual data (no z-score): Different magnitudes of two involved variables were tested ($X_1 \approx X_2$, and $X_1 \approx 100\text{-}300 * X_2$)
4. 3 groups/classes: Which classifiers were able to handle multiclass classification? This issue was important to Lundbeck Research. It is generally known, that most classifiers work only on two-class data sets (201).

Phase 2 results

Below, I present five of the results from phase 2 (full documentation available on request) in the same kind of plots as in phase 1. The abbreviations in the plots are also 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 are considering are believed to be polygenic, a classifier is not of interest if it only performs well in a single variable/single gene case.

Figure 17 shows how the 8 classifiers perform with a simple linear task involving two variables and an outcome threshold. All classifiers solve this task well.

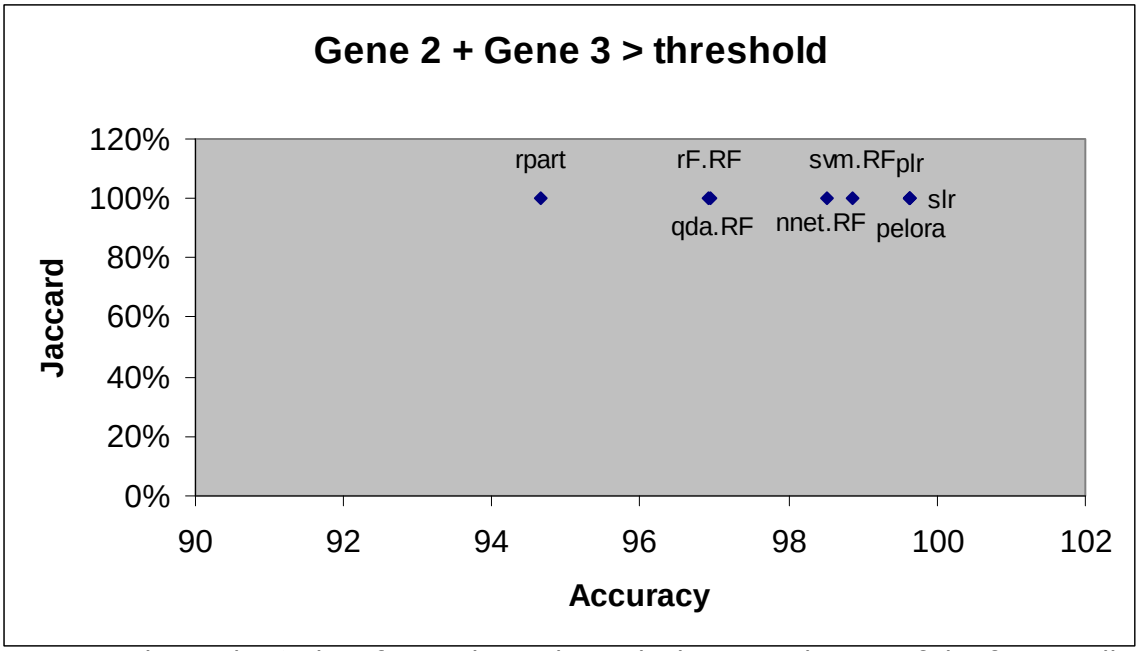


Figure 17 shows the 8 classifiers solving the task shown at the top of the figure. All classifiers perform well. NB! Notice the x-axis. The figure was made in Excel based on output from R.

However, in figure 18 the outcome is now interval dependent, otherwise with the same settings as in the previous example. The worst performing classifiers are; SLR, Pelora and PLR.

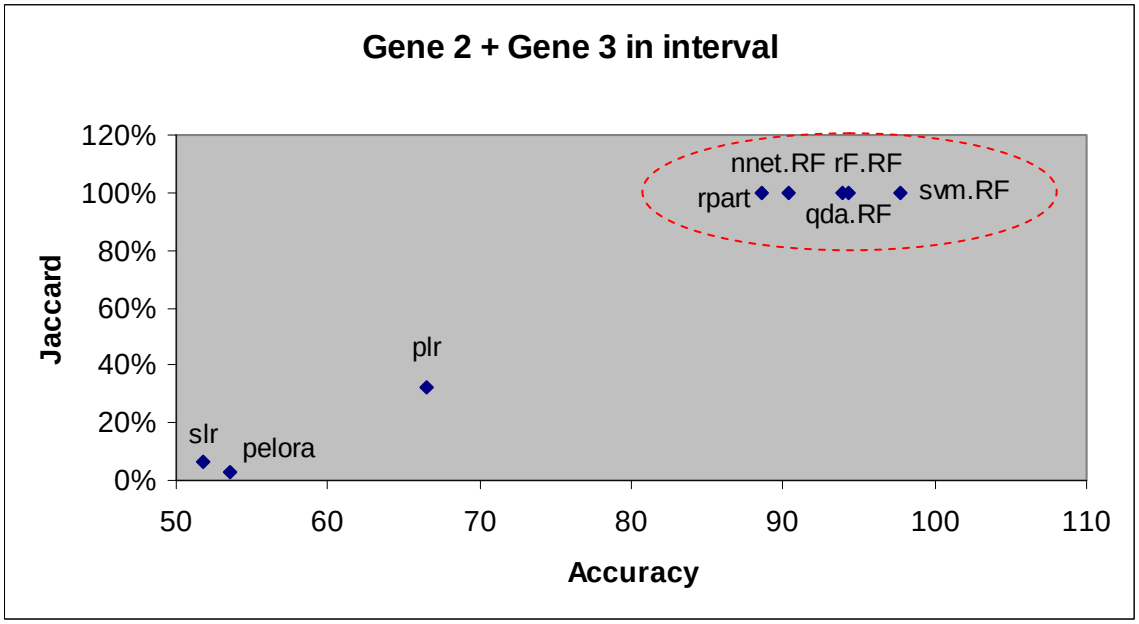


Figure 18 shows the 8 classifiers solving the task shown at the top of the figure. Three classifiers do not perform well; SLR, Pelora and PLR, while the encircled classifiers solve this task well. The figure was made in Excel based on output from R.

In figure 19, the product between two variables was explored with a simple threshold outcome. The best performing classifiers were; SVM.RF, RF.RF and PLR.

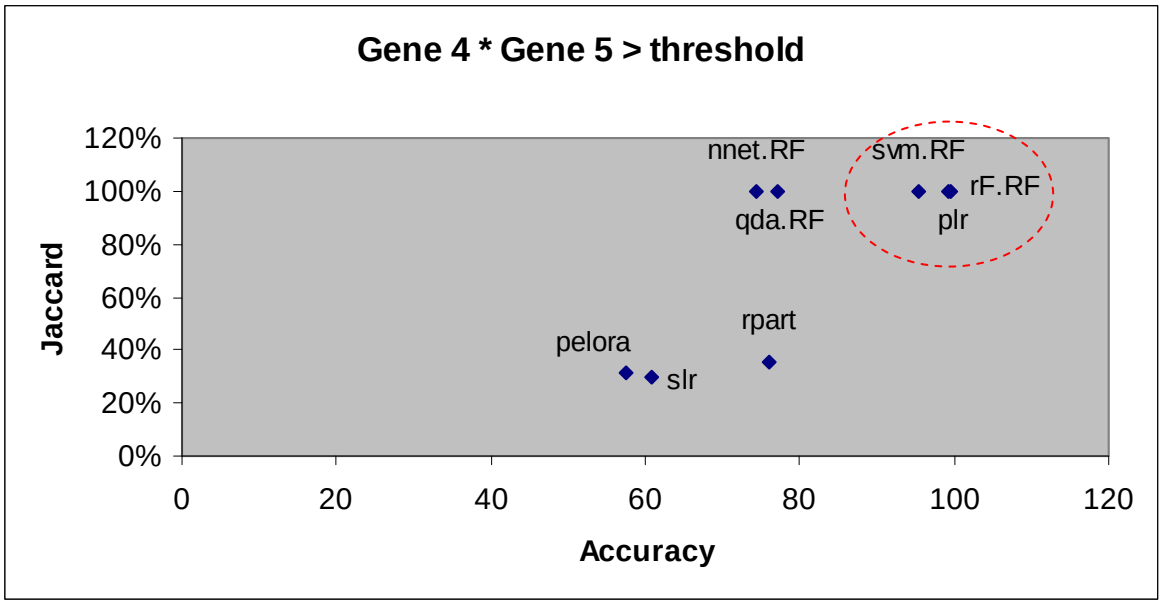


Figure 19 shows the 8 classifiers solving the task shown at the top of the figure. The encircled three classifiers perform well; SVM.RF, RF.RF and PLR. The figure was made in Excel based on output from R.

In figure 20, the ratio of two variables was explored with an interval dependent outcome. The best performing classifiers were; SVM.RF and RF.RF.

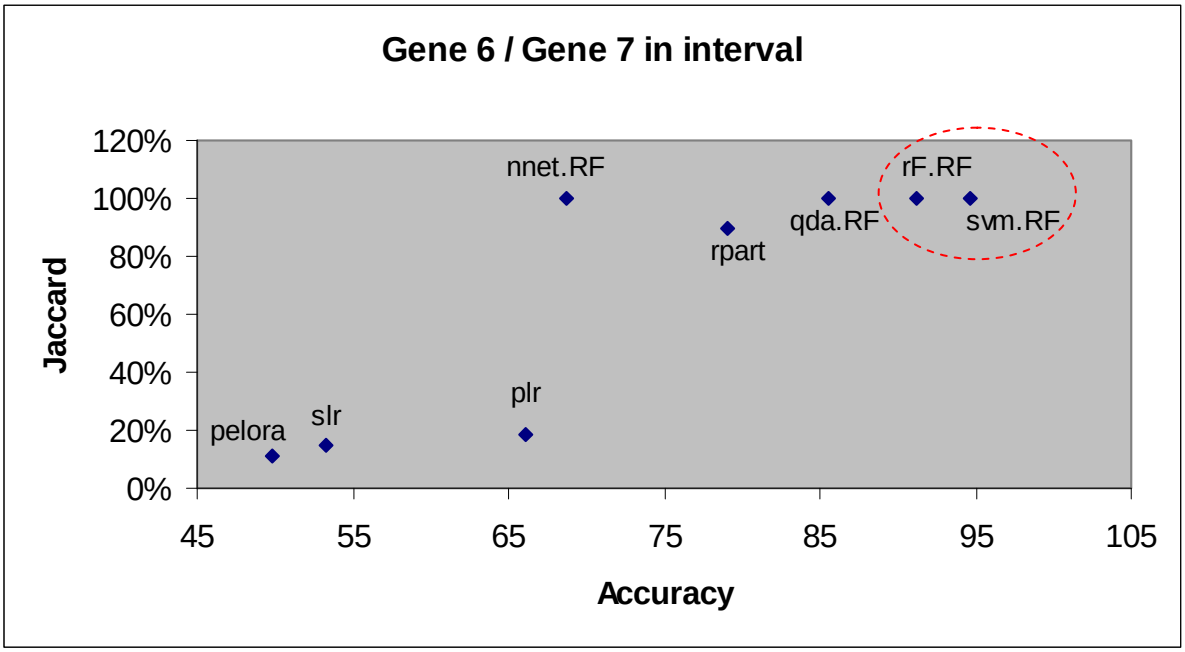


Figure 20 shows the 8 classifiers solving the task shown at the top of the figure. The encircled two classifiers perform well; SVM.RF and RF.RF. The figure was made in Excel based on output from R.

Finally, in figure 21, a linear combination of five variables was explored with a simple threshold outcome. The best performing classifiers were; SLR, Pelora and PLR.

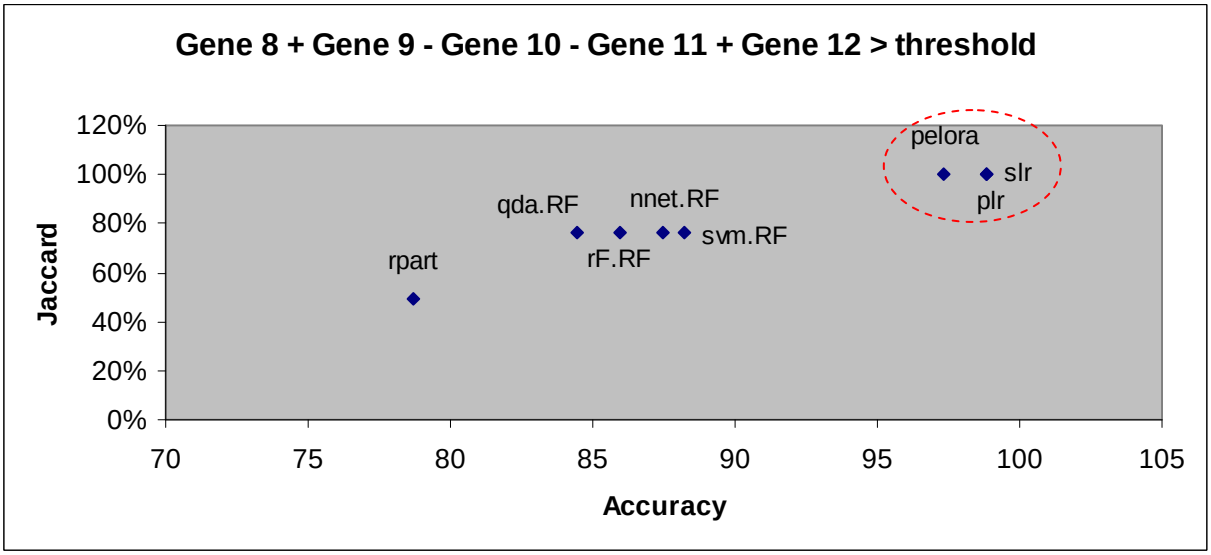


Figure 21 shows the 8 classifiers solving the task shown at the top of the figure. The encircled three classifiers perform well; SLR, PLR and Pelora. The figure was made in Excel based on output from R.

In order to compare the classifiers on a quantitatively basis, we decided to focus on classifiers that yielded an accuracy above 80% and that 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 arbitrary. In table 19, 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.

2 groups (25 tasks in total):

2 groups	%tasks solved	avg. accuracy	avg. 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

3 groups (7 tasks in total):

3 groups	%tasks solved	avg. accuracy	avg. Jaccard	Tasks solved
RPART	43%	88%	98%	r27,r28,r30
Random Forest (varselrf)*	43%	91%	98%	r28,r29,r30
QDA (varselrf)*	43%	89%	98%	r28,r29,r30
SVM (varselrf)*	43%	95%	98%	r28,r29,r30
NNET (varselrf)*	14%	97%	93%	r28

Table 19: Phase 2 tasks solved with an accuracy above 80% and a Jaccard score above 70%. The column 'Tasks solved' refer to the specific tasks solved, see appendix 7. The conclusions in the text sum up the table.

Based on the table 19 results and the 33 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 some kind of gene list consistency, that is, to have the same variables selected by two methods (although the methods share some similarity they are not alike).
- SLR solved less tasks than RPART and SVM, but more than Pelora and yielded both 100% average accuracy and average Jaccard score. Furthermore, as mentioned previously, unlike Pelora SLR is able to handle categorical clinical variables well.

- Accuracies thresholds above 80% seem to yield very high Jaccard scores > 95% and high classifier accuracies ~90-95%
- As expected standardization 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.

Based on phase 1 and phase 2 simulation studies, we recommended the use of;

Two groups:

- SVM in combination with varselrf
- RPART
- SLR (can handle linear cases with multiple variables above a threshold which neither SVM nor RPART is good at.)

Furthermore, we concluded that if groups of equal sizes could be separated with an accuracy above 80%, this would mean that we had identified the key variables/gene expressions. In the next section, I have made a classification procedure that I have used for real case classifications and variable selections.

Multiple groups (>2):

- SVM in combination with varselrf
- RPART

Also, if groups could not be separated with SVM or RPART (or additional SLR in the two-group case), we would say there were no expression differences between the groups.

These two- and multiple-group classifiers and variable selection methods are used in the Results chapter, section 7.5, to perform classification on various control and patient groups. An example of the application of the two-group classifiers is shown in the 'real case' section (section 6.3) later in this chapter.

6.2 Classification and variable selection procedure

Since the accuracy values are dependent on the group sizes, the following classification procedure is used to decide whether a group separation is possible or not, and if it is 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.

2. Calculate permuted accuracies by doing 10 permutations (due to pc temporal limitations) of the class labels leading to 10 x 10 (CV) = 100 permuted accuracies.
   I apply the permutation step in order to calculate the accuracy values expected at random in the real data set (excluding the class label) for a classifier (202).
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), otherwise use a Wilcoxon test.
4. Significant result is obtained (that is, the groups are separable) if the p-value is below the significance level 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 as a request from the US Lundbeck group.

The above five steps apply both to the two-group and multiple-group case. Furthermore, in the two-group case, to get additional useful information from the classifications, the positive and the negative predictive values are reported. "*The positive predictive value* (PPV) *is the proportion of patients with positive test results who are correctly diagnosed*" (203);

$$PPV = \frac{\#true\ \ positives}{\#true\ \ positives\ +\ \#false\ \ positives}$$

The PPV "*is the most important measure of a diagnostic method as it reflects the probability that a positive test reflects the underlying condition being tested for*" (203).

Correspondingly, "*the negative predictive value* (NPV) *is the proportion of patients* (here controls) *with negative test results who are correctly diagnosed"* (204).

$$NPV = \frac{\#true\ \ negatives}{\#true\ \ negatives\ +\ \#false\ \ negatives}$$

6.3 Real case example

To demonstrate the three classifiers SVM with varselrf, RPART and SLR on a two-group classification task, they are now used on a data set consisting of a pooled control group (DC, SH ABS, UK, PTSD controls) and the acute PTSD

patient group (25 genes). The classification and variable selection procedure is applied.

Table 20 contains the accuracy values of each classifier, both from the classification task and from permuted data sets. PPV and NPV values are reported as well.

	Accuracy (average)	PPV (average)	NPV (average)
SVM.RF	85,6%	70,0%	87,1%
Permuted.RF	78,4%		
RPART	82,2%	65,4%	88,1%
Permuted.RPART	70,3%		
SLR	88,1%	81,2%	90,1%
Permuted.SLR	78,3%		

Table 20: Accuracy, PPV and NPV values for the classification task involving a pooled control group and the acute PTSD patient group (25 genes). The table is commented in the text.

Below are the Wilcoxon two-group p-values of the 10 real case accuracies vs. the 100 permuted accuracies;

SVM.RF vs. Permuted.RF:
Wilcoxon p-value: 0.0005644

RPART vs. Permuted.RPART:
Wilcoxon p-value: 5.757e-05

SLR vs. Permuted.SLR:
Wilcoxon p-value: 1.589e-05

All these p-values are below 1%, meaning the pooled control group and the acute PTSD group is separable. In this case, just by looking at the accuracy values, it appears that the groups are separable. However, in other cases the difference between the real case and permuted accuracy values may seem small, yet be significant.

The permuted accuracies are quite high. The reason is that there are 256 controls and only 66 acute PTSD patients, meaning the group sizes are far from equal (which would have implied a permuted accuracy of approximately 50%). In this case, about 80% of the subjects (controls and patients) are controls.

Genes separating the groups:
SVM.RF[16]: ARRB2, ERK2, RGS2

[16] Due to the random aspect, different runs of RF produce (slightly) different gene lists. Focus is on the most consistent list.

RPART: ADA, ARRB1/2, CREB2, ERK1, GR, MKP1, P2X7, RGS2
SLR: ARRB1/2, CD8 beta, ERK2, IDO, IL-6, MR, PREP, RGS2

Overlapping genes: ARRB2, RGS2

The conclusion of this example is then, that the groups are separable, and that ARRB2 and RGS2 are involved in the separation. However, for use in a commercial diagnostic test, the separation of the two groups may be optimized to yield higher PPV values.

6.4 Variable selection based on random forests and SVM

While stepwise logistic regression (SLR) and recursive partitioning (RPART) have been described previously, I will now describe the particularly promising variable selection method based on random forests (varselrf) and the support vector machines (SVM) classifier in more detail.

Variable selection based on random forests
To explain variable selection based on random forests, I will start by explaining a random forest. A random forest is a classifier that consists of many classification trees (described in the Statistical methods chapter, section 5.6 and known from RPART) and outputs the class that is the most frequent of the classes output by individual trees. "*In a classification tree, each node is split using the best split among all variables. In addition to constructing each tree using a different bootstrap sample of the data, random forests change the way the classification trees are constructed. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counterintuitive strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against overfitting*" (205).

The random forests algorithm (205):

1. "*Draw ntree* (default 5000) *bootstrap samples* (with replacement) *from the original data.*
2. *For each of the bootstrap samples, grow an unpruned classification tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample mtry* (default 1) *of the predictors and choose the best split from among those variables.*
3. *Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification).*

An estimate of the error rate can be obtained, based on the training data, by the following:

1. *At each bootstrap iteration, predict the data not in the bootstrap sample (also called "out-of-bag", or OOB, data) using the tree grown with the bootstrap sample.*
2. *Aggregate the OOB predictions. (On the average, each data point would be out-of-bag around 36% of the times, so aggregate these predictions.) Calculate the error rate, and call it the OOB estimate of error rate*."

"*Variable selection based on random forests is then performed using the OOB error as a minimization criterion, by carrying out variable elimination from the random forests, by successively eliminating the least important variables (with importance as returned from random forest)*" (184). More specific (184):

1. "*With the default parameters, all forests that result from eliminating, iteratively, a fraction, of the least important variables used in the previous iteration are examined.*
2. *After fitting all forests, the OOB error rates from all the fitted random forests are examined. The solution with the smallest number of genes whose error rate is within c.sd*[17] *(default 1) standard errors of the minimum error rate of all forests are chosen*".

Support vector machine (SVM)
After varselrf has determined the variables/genes separating groups, the support vector machine classifier performs the actual classification based on these variables.

As explained in (195), considering the two-class case, the input data may be viewed as "*two sets of vectors in an n-dimensional space. An SVM will construct a separating hyperplane in that space, one which maximizes the 'margin' between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating one, which are 'pushed up against' the two data sets*", see figure 22. "*Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes. The hope is that, the larger the margin or distance between these parallel hyperplanes, the better the generalization error of the classifier will be*" (195).

[17]" *Setting c.sd = 1 is similar to the common '1 standard error rule', used in the classification tree literature; this strategy can lead to solutions with fewer genes than selecting the solution with the smallest error rate, while achieving an error rate that is not different, within sampling error, from the 'best solution'*"(184).

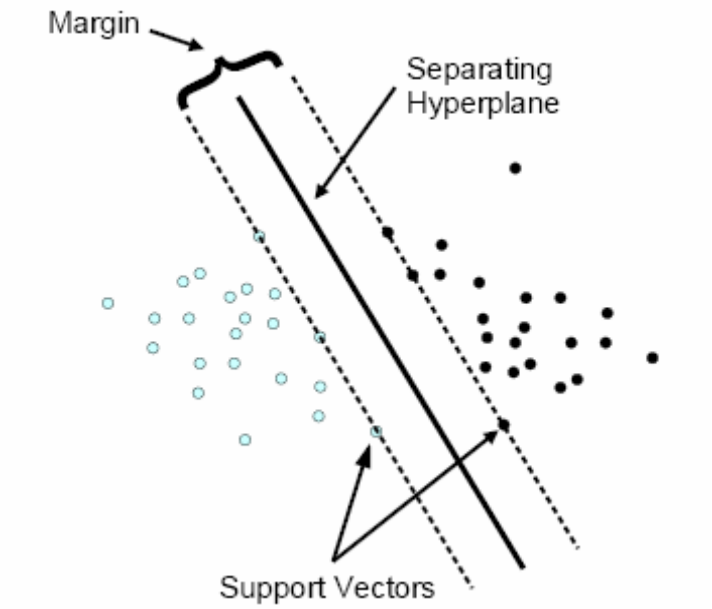


Figure 22 shows SVM classification in the linear case. A separating hyperplane (thick bold line), the two parallel hyperplanes, the margin and the support vectors are shown. The figure appears in (206).

In the actual implementation of SVM in R for classification purposes, the so-called kernel trick is applied, meaning we are performing nonlinear classification with an otherwise linear classifier. "*The kernel trick is a method for using a linear classifier algorithm to solve a nonlinear problem by mapping the original nonlinear observations into a higher-dimensional space, where the linear classifier is subsequently used; this makes a linear classification in the new space equivalent to nonlinear classification in the original space*" (207). The specific kernel applied is the radial basis function (208), which is recommended for classification (206). Also, SVMs are known to be very sensible to the proper choice of parameters, and the before mentioned reference recommends checking a range of parameter combinations. This I do by tuning the two available SVM parameters, gamma (parameter needed for all kernels except linear) and cost (cost of constraints violation) in the parameter space; gamma = 2^(-2:2)[18], cost = 2^(1:8) based on (208). This parameter space was always searched for every classification task.

[18] '2^(-2:2)' means gamma is tested in the grid interval from 2^{-2} to 2^{2}.

7. Results

The US Lundbeck group and I defined a number of questions to be considered, see table 21 in the present chapter (this table is based on the questions listed in table 10, chapter 5, and revised to include bioinformatics and classification tasks). This chapter describes answers to these questions by providing results obtained by means of bioinformatics and of various statistical and classification methods. The questions/purposes and results comprise

- *Bioinformatic predictions of new possible biomarkers; using the web application Ingenuity, new possible biomarkers like Hsp90, PP2A and NFkB (and others) are predicted.*
- *Bioinformatic predictions of altered gene expressions in an as yet unanalyzed patient group – bipolar disorder patients; Gs, IL-1 beta, CREB1 and ERK1 are predicted to show altered expression.*
- *Various gene expression and clinical data relationships are reported, e.g.*
    - *20 hypotheses are constructed as gene expression predictions for depressed patients. These hypotheses are based on expression patterns from controls and identified possible intermediate phenotypes.*
    - *Possible borderline personality disorder (BPD) subtypes are identified by recursive partitioning (RP) looking at the BPD patients and the Danish and 'super-healthy' American controls. At the same time, RP shows that the two control groups are more similar than the patient group.*
    - *Other possible BPD subtypes are identified with canonical correlation analysis incorporating both clinical variables and gene expressions.*
- *The effect of pooling both two control groups and all control groups are investigated and recommended.*
- *Repeated measures ANOVA test results indicate that five gene expressions differ significantly between three time point measurements; CD8 beta, IL-8, MKP1, MR and ODC1.*
- *Classification results indicating genes separating*
    - *controls from borderline personality disorder (BPD) patients*
    - *controls from acute post-traumatic stress disorder (PTSD) patients*
    - *controls from trauma patients (without PTSD). This separation is not convincing due to low performance measures.*
    - *controls can not be separated from remitted post-traumatic stress disorder patients – a result that is in good agreement with the clinical diagnosis.*
- *Possible disease subtypes both in BPD and PTSD are identified by the use of clustering and heat maps.*

In close cooperation with the US Lundbeck group, we defined a number of questions to be considered. I have looked into these questions by the use of various bioinformatic approaches and by the statistical and classification techniques described in earlier chapters. In order to structure all the different results I obtained by analyzing the qPCR data, I have summed up the main purposes, the applied methods, and the data used for the analyses in table 21. This table also lists the section in which the results are presented. The table is inspired by the statistical summary table 10 from the Statistical methods chapter, and revised to include bioinformatics and classification tasks. It should be noted, that sometimes a purpose in the table answers more than one question. With the many different aims of our analyses, some of them are overlapping. In the table and in the relevant sections, I make the reader aware of any such overlapping purposes.

Analysis purpose	**Method**	**Data / group**	**Section**
Bioinformatic predictions:			7.1
- Prediction of new possible biomarkers	- bioinformatics; STRING and Ingenuity	- List with 29 genes	7.1.1
- Which gene expressions are expected to be regulated in bipolar disorder patients?	- bioinformatics; NCBI's SNP database, STRING and Ingenuity	- Two articles and list with 29 genes	7.1.2
Is the qPCR data normally distributed?	normal QQ plots, normality tests	ABS[19] controls	7.2
Identify clinical variable – gene expression relationships[20]:			7.3
- Biomarkers for depression – 20 hypotheses	- univariate tests, correlations	- ABS[19] controls	7.3.1
- Gene expression subgroups identified via RP	- recursive partitioning	- ABS[19] and DC controls, BPD patients	7.3.2
- Possible BPD phenotypes through CCA	- canonical correlation analysis	- BPD patients	7.3.3
Special focus: - which clinical variables explain the most variance in a gene?	- stepwise regression	- ABS[19] controls	7.3.4
- any gender differences in the expression profiles?	- Pelora / SLR (initial classifiers), correlations, univariate tests	- ABS[19] and DC controls	7.3.5
- pooling of two control groups into one group?	- univariate tests, correlations and correlation tests, classification	- ABS[19] and DC controls	7.3.6
- should all control groups be pooled into a single large group?	- univariate tests, plots, Pelora / SLR	- all control groups	7.3.7

[19] In table 21, I do not differentiate between the ABS group and the SH ABS group, and just call it 'ABS controls'.

Are gene expression levels the same across different time points?	repeated measures ANOVA	UK controls	7.4
Variable selection (identifying genes separating various control and patient groups) and classification.	varselrf and variable selection from the classification methods RPART and SLR Classifiers: SVM (with varselrf), RPART and SLR	ABS[19], UK and DC controls, PTSD controls, remitted PTSDs, BPD patients, acute PTSDs and trauma patients	7.5
- 2-group comparisons	- SVM (with varselrf), RPART, SLR		7.5.1
- multiple group comparisons	- SVM (with varselrf) and RPART		7.5.2
Special focus: - Genes and clinical variables separating control and patient group	- SLR	- ABS[19] controls and BPD patients	7.5.3
Identify gene – gene relationships (expression patterns only)	clustering and heat maps	ABS[19] controls, BPD and PTSD acute patients	7.6

Table 21 presenting the various purposes of our analysis, the applied methods (bioinformatics, statistical, classification), the data used for analysis, and the sections containing the corresponding results.

The table contains a column for the data used to analyze a purpose. It should be noted that I have analyzed the different purposes with the control and patient data available at time of the analysis, and always in agreement with Lundbeck US. This also explicitly means that all purposes or methods have not been analyzed using all group combinations, as we have not found it necessary to investigate all permutations, even though this might be possible. In general, the results illustrate the usability of the various methods, which is a major purpose of the study.

7.1 Bioinformatic predictions

Prior to any statistical or classification analysis in R, bioinformatics can be used to answer important prediction questions. One such question concerns the identification of new possible biomarkers based on the gene list with the 29 genes, and the results from this analysis are presented in section 7.1.1. Bioinformatics may also give the answers to a question concerning which gene expressions that may be expected to be regulated in a not-yet data analyzed patient group. I present such results from a bioinformatic analysis looking into SNP (single nucleotide polymorphisms - a variation in a gene caused by the

[20] Here is an example of a (broadly defined) purpose that answers more than one important question; besides identifying clinical variable – gene expression relationships (determining intermediate phenotypes), this task also gives insight into which gene expressions that are expected to be regulated in depressed patients, see section 7.3.

change of a single base in DNA) analyses from bipolar disorder patients in section 7.1.2.

7.1.1 Prediction of new possible biomarkers

The US Lundbeck group had chosen to measure the gene expressions of approximately 30 genes, see 'The genes selected by Lundbeck' chapter. The genes were chosen based on a literature search. Bioinformatics can now be used to predict novel biomarker genes, that it could make sense to measure in the same control and patient groups. Based on the list of 29 genes, the bioinformatics web application Ingenuity lets us know that the 29 genes are gathered in three networks, see network 1, 2 and 3 in appendix 1 that also explains the gene abbreviations. These networks include various kinds of interactions between the genes. The reader may recall that in Chapter 3, I wrote, that examples of interactions were: Activation/inhibition, binding, expression, phosphorylation/dephosphorylation, protein-DNA binding, protein-protein binding and transcription.

In the CBS course 'Bioinformatics and Gene Discovery', we were taught that a protein in a protein complex may be a good biomarker if the protein complex contain other putative disease specific proteins. The more disease specific proteins in a complex, the more likely it is that another protein in the complex could be a possible biomarker. Thus, in order to predict novel biomarkers, I merged the three networks into one big network and focused on protein-protein interactions (PPI), see figure 23.

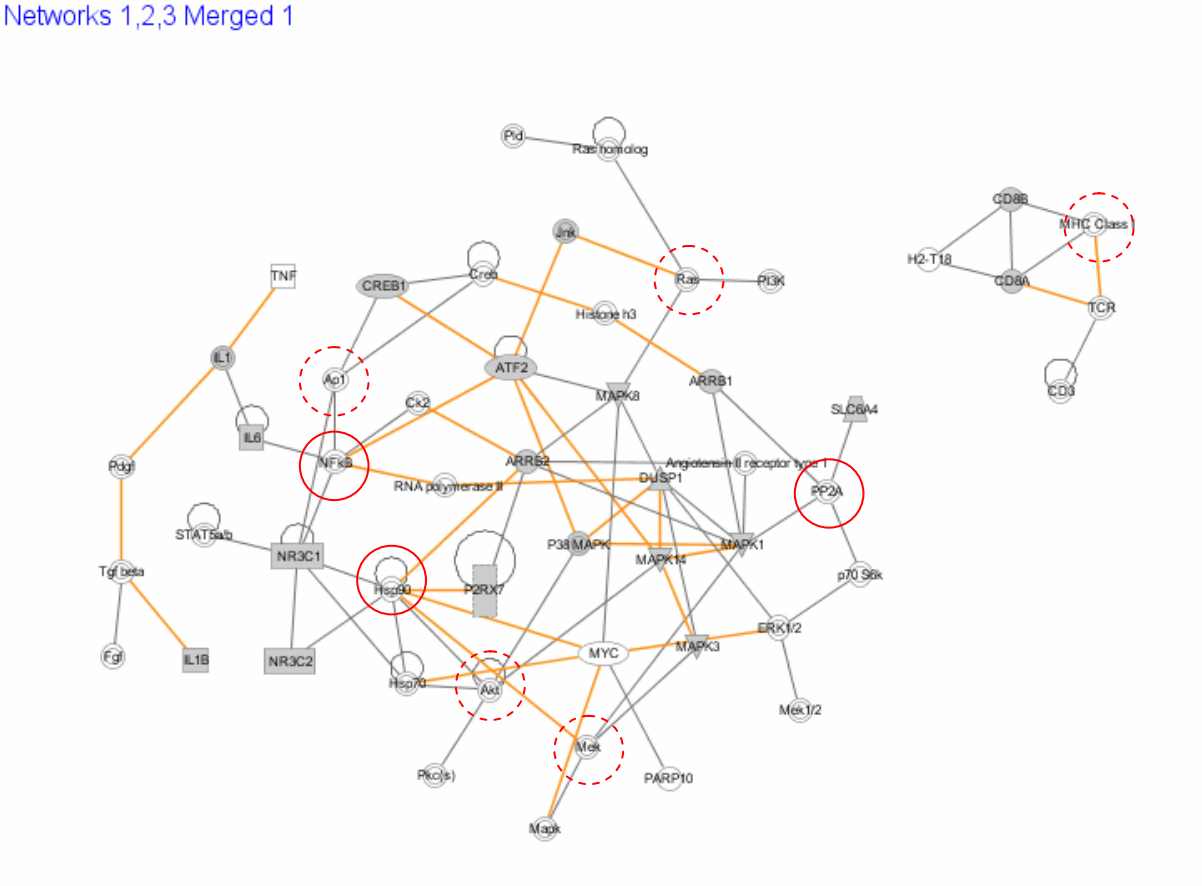


Figure 23 shows PPI interactions between 29 genes selected by Lundbeck (marked grey). The full red circled proteins are directly interacting with at least three Lundbeck selected genes, while the dotted red circled proteins directly interact with maximum two of the Lundbeck selected genes. The network was created with Ingenuity.

In figure 23, the proteins interacting with at least three of the Lundbeck selected genes/proteins are (full red circled);

- Hsp90 (heat shock protein 90kDa alpha (cytosolic), class A member 1)
- PP2A (protein phosphatase 2A activator, regulatory subunit 4)
- NFkB (nuclear factor-kappa B)

Proteins interacting with maximum two of the Lundbeck selected genes/proteins are; Ras, MHC Class I (major histocompatibility complex, class I), Mek (mitogen-activated protein kinase kinase kinase 1), Akt (protein kinase B) and Ap1 (activator protein 1).

The above 3+5 genes could be potentially new biomarkers, if they are expressed at detectable levels in whole blood.

7.1.2 Prediction of regulated gene expressions in bipolar disorder patients

Bioinformatics may also offer a qualified guess as to which gene expressions, among the 29 selected genes, that may be expected to be regulated in a patient group whose gene expressions have not been measured yet.

Last year, the Wellcome Trust Case Control Consortium (WTCC) (209) compared blood SNPs in ~2000 UK BD (bipolar disorder) patients with blood SNPs in ~3000 UK controls (210). In the article (210), WTCC identified a little more than 100 SNPs with strong or modest association to BD. Using NCBI's SNP database (211), I found that these SNPs affect 57 genes (multiple SNPs may be present in the same gene, and some SNPs are located in non-coding regions of the genome). Furthermore, Baum and colleagues looked in blood SNP in BD in both an American and a German case-control group (59). In total, their article mentioned 11 genes that might be implicated in BD. Appendix 8 lists all the 68 genes. There are no overlapping genes between the 11 and 57 genes. Also, none of 68 genes are identical with any of the 29 genes selected by Lundbeck. Furthermore, there is only a very little overlap between the 68 genes and the BD associated genes listed in chapter 2. This aspect is discussed in the next chapter.

After agreement with the US Lundbeck group, I decided to look into PPI interactions to investigate any possible protein-protein interactions with the 29 selected genes/proteins. I did this using the PPI database STRING (212) and Ingenuity (104), see figure 24 for the work flow for the whole task.

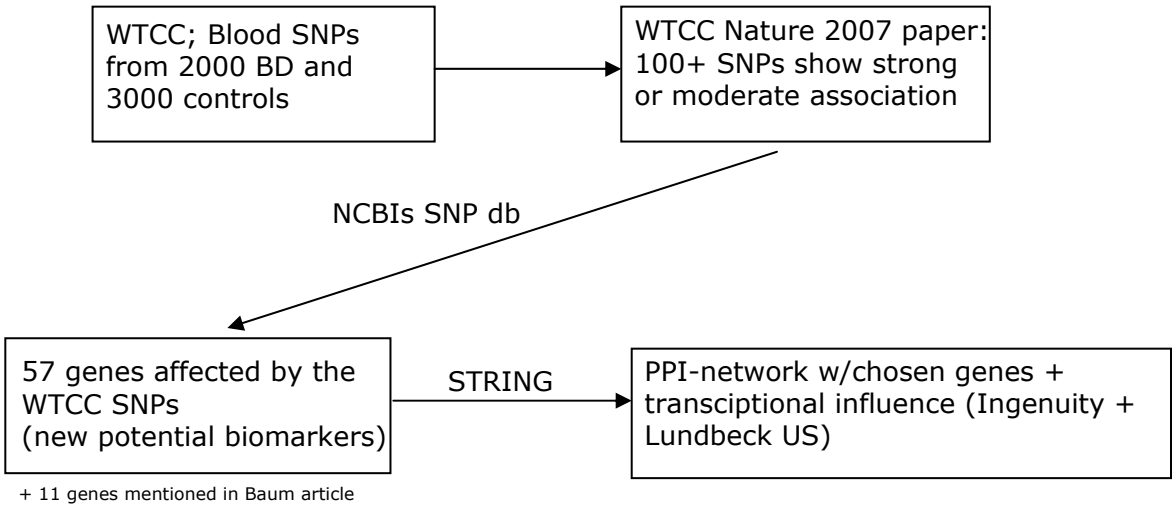


Figure 24 shows the work flow in identifying which gene expressions, among the 29 selected genes, that may be expected to be regulated in BD patients. The diagram is explained in the text.

In figure 25, I show the STRING PPI network containing the 29 selected genes/proteins and 68 BD associated proteins.

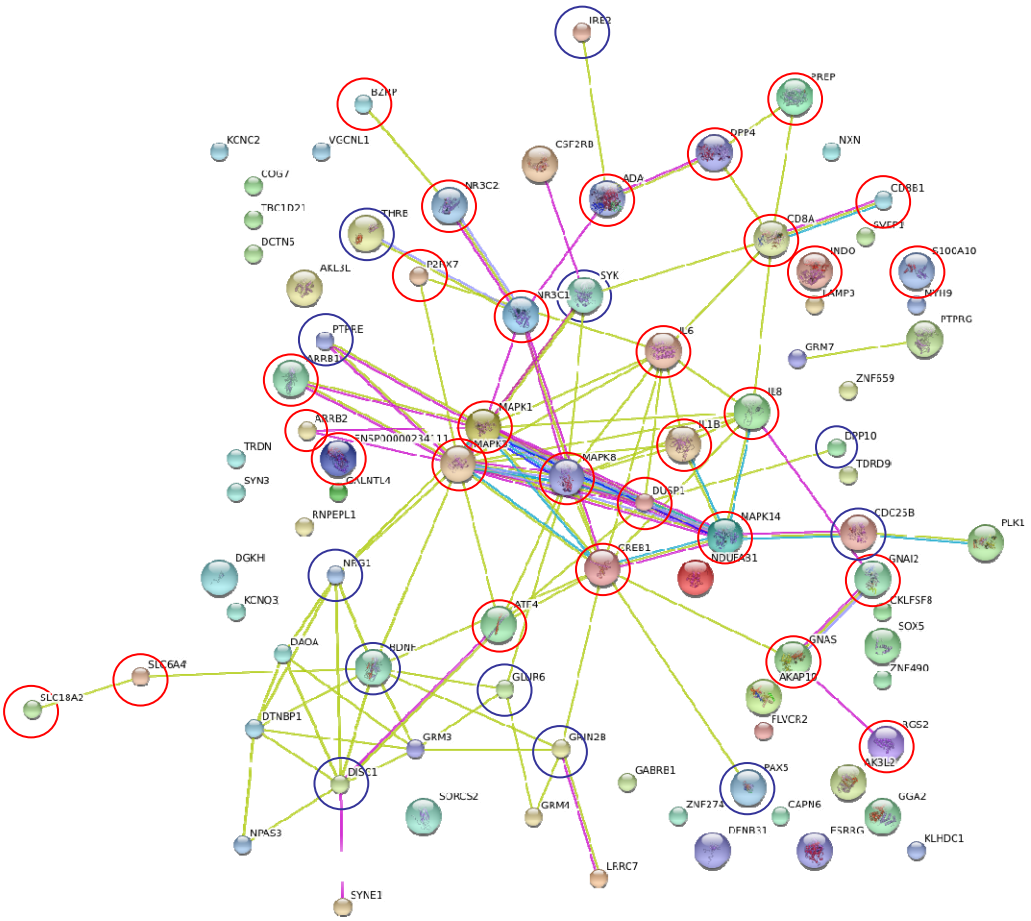


Figure 25 shows a STRING PPI network consisting of the 29 Lundbeck selected genes/proteins (red circled) and 68 BD associated proteins. Blue encircled 'BD' proteins interact with red encircled proteins. See the text for explanations of this figure.

In figure 25, the red encircled proteins are the 29 selected by Lundbeck, while the blue encircled genes are the BD associated proteins that interact with any of the 29 proteins. The 12 blue encircled interacting proteins are possible BD biomarkers[21] related to the 29 proteins:

SYK (spleen tyrosine kinase), BDNF (brain-derived neurotrophic factor), PTPRE (protein tyrosine phosphatase, receptor type, E), CDC25B (cell division cycle 25B), IRE2 (endoplasmic reticulum-to-nucleus signaling 2), DPP10 (dipeptidyl peptidase 10), THRB (thyroid hormone receptor, beta), NRG1 (neuregulin 1), DISC1 (disrupted in schizophrenia 1), PAX5 (paired box 5), GRIN2B (glutamate receptor, ionotropic, N-methyl D-aspartate 2B) and GLUR6 (Glutamate receptor 6).

[21] Before the 12 genes are considered as possible biomarkers, it is important to see if these genes are expressed at detectable levels in white blood cells. If they are not expressed in blood, they are not relevant as blood biomarkers.

In order to refine this list further, Joseph Tamm from Lundbeck US suggested focusing on interactions where gene X directly influences transcription of gene Y. This I did in Ingenuity, and shortened the list down to SYK, BDNF and THRB.

At the level of transcription, the three BD associated genes seem to affect:

- Gs
- IL-1 beta
- CREB1
- ERK1

Thus, expression differences between BDs and controls may be expected to be seen in the above four genes. At this point, it is not possible to say whether the genes are expected to be up or down regulated in BDs.

7.2 Normally distributed qPCR data?

The parametric tests applied later in this chapter assume that the qPCR gene expression data is normally distributed. In the Statistical methods chapter, section 5.1, I mention the five different normality tests; the Shapiro-Wilk test, the Anderson-Darling test, the Cramér-von-Mises criterion, the Lilliefors test for normality, and the Shapiro-Francia test for normality. I applied these five tests to the ABS control group containing 29 genes. In table 22, the number of gene expressions following a normal distribution according to the five tests is shown. I included a column called 'Ratios' dealing with various gene expression ratios that are explained in the next section (7.3).

1% significance level	Gene expressions	Ratios	Total
Anderson-Darling (normal)	0	7	7
Anderson-Darling (log10)	7	24	31
Cramer-von Mises (normal)	1	9	10
Cramer-von Mises (log10)	9	26	35
Lilliefors (normal)	3	11	14
Lilliefors (log10)	12	29	41
Shapiro-Francia (normal)	0	2	2
Shapiro-Francia (log10)	7	18	25
Shapiro-Wilk (normal)	0	2	2
Shapiro-Wilk (log10)	7	19	26

Table 22 presents the number of gene expressions and gene ratios following a normal distribution according to the five normality tests on a 1% significance level. Each test is applied to both the raw expression data and to the logarithm of the expression data. The ABS control data was investigated with 29 genes and 40 ratios. The tests were done in R.

Table 22 shows that, in general, applying a logarithmic transformation to the gene expression data and the gene ratios improves the number of gene expressions and ratios following a normal distribution. The results above also support the use of non-parametric tests as the majority of gene expressions do

not follow a normal distribution. Here is should be stressed that the standard parametric tests, I applied, are robust, that is, known to be useful when the deviations from normality are relatively small. Exactly when this is the case, is not clear from the literature, so in the next section I make use of both non-parametric tests and parametric tests on the logarithm of the expression data. We thus decided to apply the logarithm to all gene expressions.

In section 5.1, I mentioned normal QQ plots and in figure 26 and figure 27 I show a few examples of these plots for different gene expressions both applied to the raw data and to the logarithm of the expression data.

In figure 26, normal QQ plots are shown for ARRB2, both applied to the raw expression data and to the logarithm of ARRB2. ARRB2, as was the case with SERT in figure 7 in section 5.1, follows a normal distribution when the logarithm of the expression data is considered.

On the other hand, in figure 27, GR does not follow a normal distribution neither using the raw expression data nor using the logarithm of the expression data. However, as seen in figure 27, applying the logarithm diminishes the departure from normality. This is the case for most of the gene expressions, not all. Since most of the gene expressions look more or less like figure 27, no further QQ plots are shown.

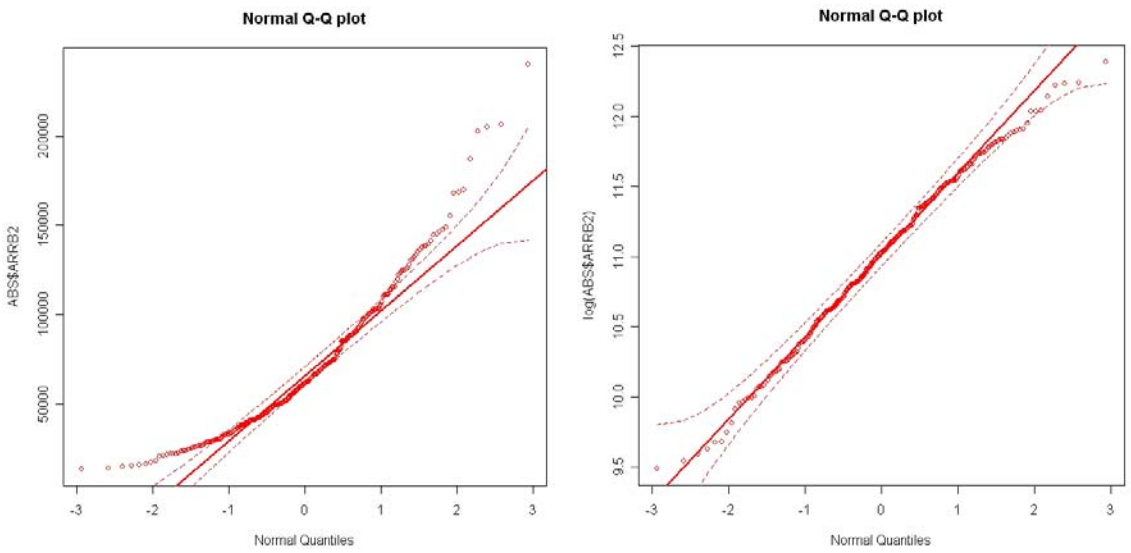


Figure 26 shows that in the case of ARRB2, a logarithmic transformation makes the ARRB2 expression data follow a normal distribution. The plots were made in R.

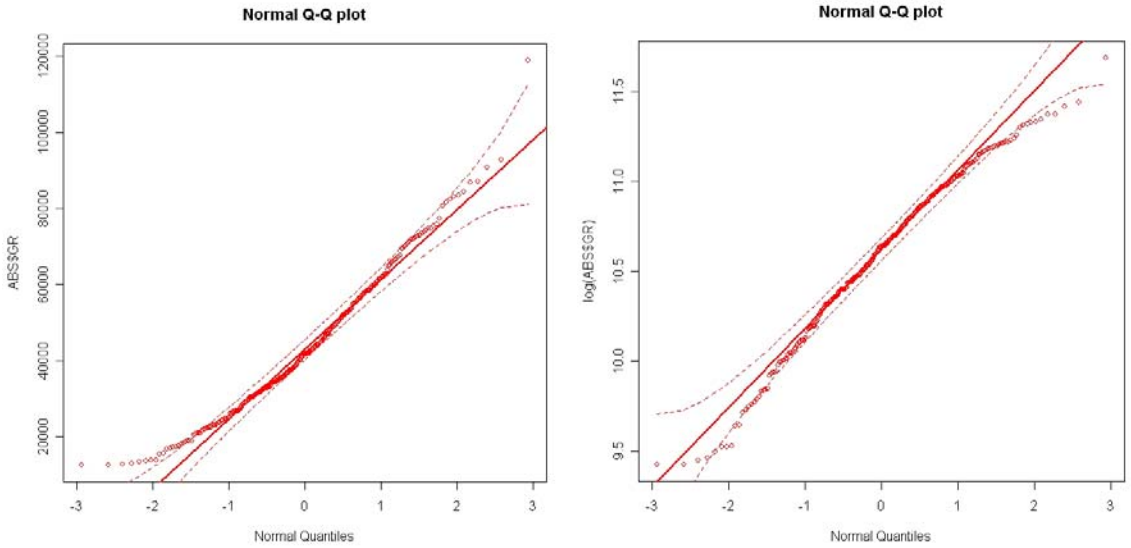


Figure 27 shows that in the case of GR, a logarithmic transformation diminishes the departure from normality. In particular, the extreme outliers in the left plot are not as evident in the right plot. The plots were made in R.

7.3 Clinical variable – gene expression relationships

As described in the Study design chapter (chapter 4), the Lundbeck study includes both questionnaire data and qPCR data for the ABS and DC control groups and for one half of a patient group (BPD). Some of the first work, the US Lundbeck group and I did, included the identification of clinical variable – gene expression relationships in the ABS control group. This was done in order to determine intermediate phenotypes, and also to predict which gene expressions that we expected to be regulated in depressed patients. Later, I looked into questionnaires and qPCR data for the borderline personality disorder (BPD) patients, and made some initial comparisons with the ABS controls. Finally, the same gene expression – questionnaire relationships found in the ABS controls were investigated in the DC controls. Below, I present the results from these three studies. The section also provides answers to four questions; 1) which clinical variables explain the most variance in a gene?, 2) are there any gender differences in the expression profiles?, 3) are the control groups similar?, and 4) should all control groups be pooled into one big control group?

7.3.1 Biomarkers for depression – 20 hypotheses

Some of the first work I did involved looking at the 299 ABS controls. We wanted to predict gene expressions that might be expected to be regulated in depressed patients solely on the basis of data from controls. This also formed the start of our work on intermediate phenotypes as explained below.

Besides directly investigating the 29 genes described in chapter 3, we decided to include gene ratios, since they could expand the window of detecting differences between groups. Ratios were also included to expand the hypothesis generating phase in that several combinations/ratios of gene expressions had a biological interest. This included 40 ratios of e.g. some anti-inflammatory cytokines to pro-inflammatory cytokines, various kinase combinations, a glucocorticoid and mineralocorticoid combination, etc., see appendix 9 for a list of 97 ratios in total. In that appendix, 57 ratios formed on the basis of Spearman correlations are included as well. We included both ratios between genes with high Spearman correlations (>0.80), which generally included correlations between gene pairs seen in the literature, and low correlations (<0.30), which in general included gene pairs not expected to be correlated in the literature. In this way, the gene ratios were partly formed on a biological basis, partly on a data driven basis which was then not necessarily biologically biased.

I applied the statistical univariate tests described in the Statistical methods chapter, section 5.2; the parametric t-test and ANOVA test on the logarithm of the gene expressions and ratios, and the non-parametric Wilcoxon and Kruskal-Wallis tests. In order to reduce the number of false positives, the significance level for individual genes was set to 1%, and for ratios of gene expressions to 0.1%. Spearman correlations were also used.

The ABS questionnaire consisted of approximately 50 questions as described in the Study design chapter. The US Lundbeck group picked questions and composite scores defining depression related intermediate phenotypes on the basis of the DSM-IV-TR clinical descriptions of depression in chapter 2. The interested reader is referred to the Study design chapter for the chosen questions and composite scores as well as the coding aspect[22].

Expression patterns for individual genes or ratios that yielded significant p values with multiple questionnaire responses were used to generate hypotheses regarding patient populations. The reason for considering multiple questionnaire responses, and not individual responses, was a wish to further reduce the number of false positives. As the results below indicate, even the strict significance levels applied to individual genes and ratios still yielded a small number of false positives. However, we did not want to set even more strict significance levels due the whole exploratory nature of the study.

[22] The coding of each clinical variable is shown in appendix 5. For tobacco use, a 3-bin coding (low, medium and heavy users) was attempted, but this led to no new results compared to the two-bin case (non-smokers vs smokers). In general, 3-bin coding of various clinical variables did not provide additional information compared to 2-bin codings.

The first results obtained were;

- 42 combinations of gene expression vs questionnaire response[23] satisfied the $p<0.01$ criteria (42/1564 comparisons).
- 65 combinations of ratio vs questionnaire response satisfied the $p<0.001$ criteria (65/4462 comparisons).
- No Spearman correlation coefficients were greater than 0.3 between the continuous clinical variables like age and BMI and any gene expression.
- Results obtained from MANOVAs, briefly described in the Statistical methods chapter, essentially represented a subset of those obtained using gene expression ratios. At this point, it was reassuring that two different statistical approaches gave similar results for each hypothesis.

Table 23 sums up the results for the 6+1 (see table text) significant individual genes whose expression correlates with multiple responses. Arrows in the table indicate whether the general trend is an increase or decrease in expression as one moves from the most normal to the most affected subjects. This was assessed by looking at various scatter plots of gene expressions, and in figure 28 a few are shown as examples (the other plots are not included).

	SERT	DPP4	ERK2	G alpha s	MKP1	PBR	MAPK14
More alcohol use						↑	
Any lifetime "treatments" (1)		↑			↑		
More tobacco use	↓						
Increased anxiety				↓			
Decreased appetite	↓		↓	↓	↓	↓	↓
Decreased concentration		↑					
Decreased energy		↑					
Increased feeling low		↑	↑		↑		
More sleep problems		↑					

Table 23 presents a summary of individual genes whose expression correlates with multiple responses. 6+1 genes show significant trends with respect to multiple questionnaire responses ($p<0.01$). MAPK14 is on the list, because both the parametric and non-parametric test results are significant. The results are commented in the working hypotheses at the end of this subsection. (1) "Treatments" in the table refers to treatment by a physician for depression, anxiety, etc. An arrow pointing upwards means that the general trend is an increase in expression as you move from the most normal to the most affected subjects. An arrow pointing downwards means that the general trend is a reduction in expression as you move from the most normal to the most affected subjects.

[23] The composite clinical semi-continuous variables were binned. Considering them as entirely continuous and correlating them with the gene expressions did not yield any interesting results.

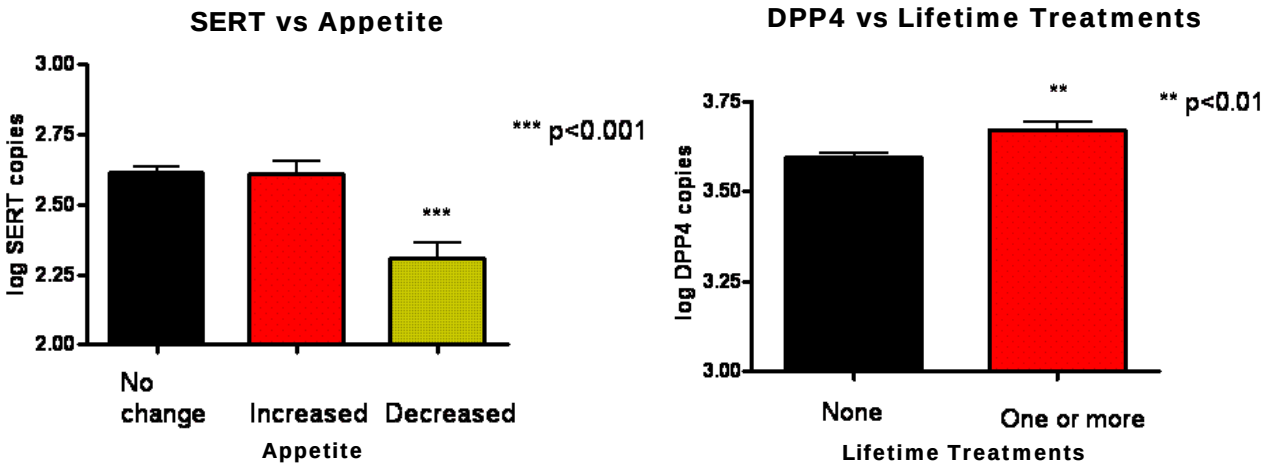


Figure 28 shows that in the case of SERT in controls, a significant decrease in gene expression is seen for a decreased appetite compared to the two other levels of appetite. For DPP4 vs lifetime treatments in controls, a significant increase is seen in subjects having at least one lifetime treatment compared to the ones not having experienced any lifetime treatment. The plots were made in GraphPad.

Table 24 sums up the results for significant gene expression ratios. 13 ratios showed significant trends with respect to multiple questionnaire responses ($p<0.001$) involving additional 10 genes compared to the significant individual genes. Arrows in this table indicate whether the general trend is an increase or reduction in ratio as one moves from the most normal to the most affected subjects. This was also assessed by looking at various plots, and in figure 29 an example with Gs/CREB2 vs anxiety is shown (other plots are not included). Also included are plots of Gs, and CREB2 vs anxiety, and here it can be seen that these plots do not contain any significant differences between the four levels of anxiety.

	ERK2/MAPK8+14	Gi2/ARRB1	Gi2/ARRB2	Gs/ARRB2	Gi2/CREB	Gs/CREB2	MKP1/ERK1	MKP1/MAPK14	ERK1/GR	ERK2/ARRB1	MAPK8/PREP	ARRB1/GR	IL-8/SERT
Gender	↓ males	↓ males		↓ males	↓ males	↓ males				↓ males			
Any illicit drug use					↑								
Lifetime "experiences" (1)		↑			↑		↑			↑	↓	↓	↑
Lifetime "treatments" (2)					↑		↑				↑		
More tobacco use		↑											
Increased anxiety						↓			↓			↓	
Decreased appetite													↑
Any drug use last 3 months *					↑		↑	↑	↓		↑		
Decreased Energy								↑					
Increased feeling low	↑		↑	↓									
More sleep problems			↑	↓									

Table 24 presents a summary of individual genes whose expression correlates with multiple responses. 13 ratios show significant trends with respect to multiple questionnaire responses ($p<0.001$). The results are commented in the working hypotheses at the end of this subsection. (1) "Experiences" refers to episodes of depression, anxiety, etc. (2) "Treatments" refers to treatment by a physician for depression, anxiety, etc. * Any drug use includes both prescription and illicit drugs. An arrow pointing upwards means the general trend is an increase in ratio as one moves from the most normal to the most affected subjects. An arrow pointed downwards indicates that the general trend is a reduction in ratio as one moves from the most normal to the most affected subjects.

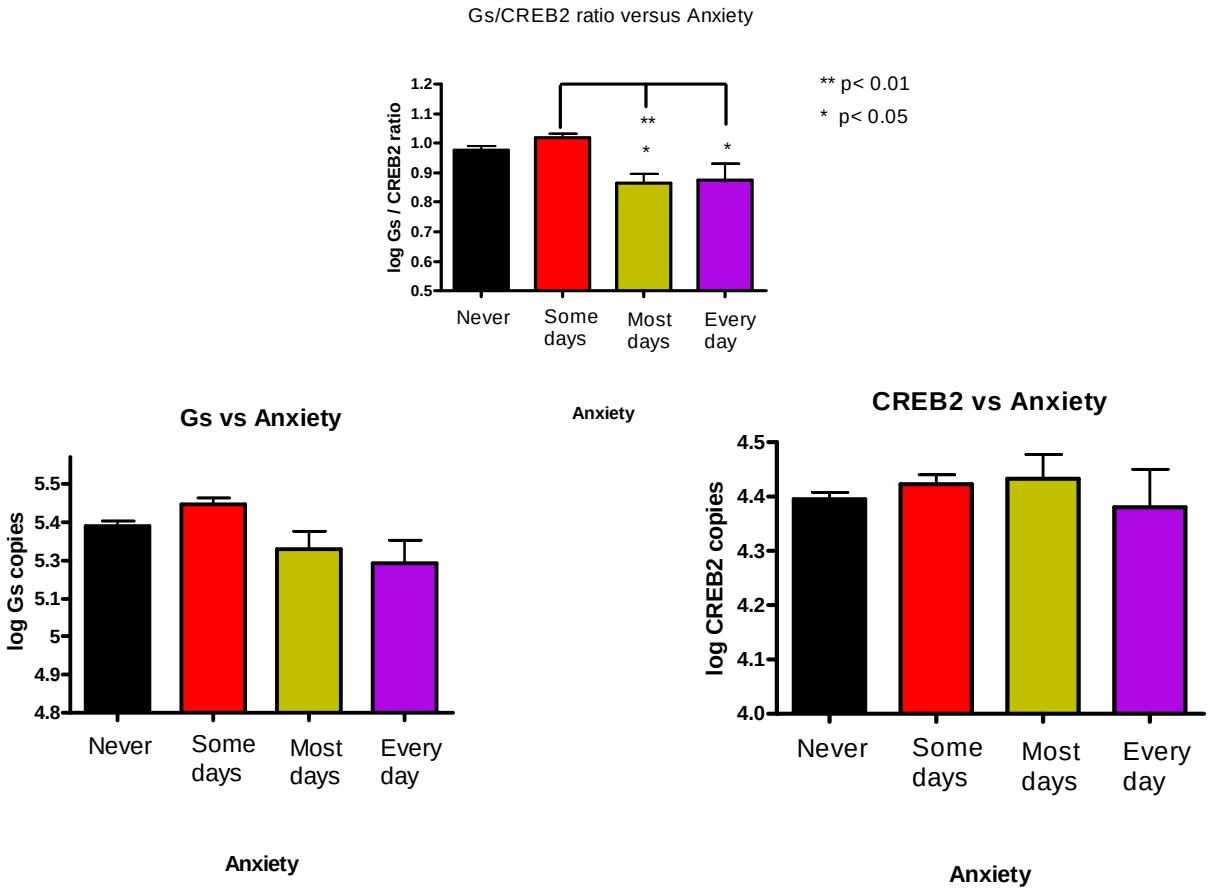


Figure 29 shows the ratio Gs/CREB2 vs the clinical variable anxiety with four levels (top plot). At the top plot, a significant difference is seen between "Some days" and "Most days". The two individual plots below show Gs vs anxiety and CREB2 vs anxiety respectively. None of these plots contain significant differences between the four levels of anxiety. The plots were made in GraphPad.

The results from table 23 and table 24 are summed up in table 25 that show gene expression patterns that may define subgroups and intermediate phenotypes. After the table, the 20 hypothesis for depression markers are listed.

Questionnaire topic	Increased expression	Decreased expression
Any illicit drug use	Gi2/CREB	
Any lifetime "experiences" (1)	Gi2/ARRB1, Gi2/CREB, MKP1/ERK1, ERK2/ ARRB1, IL-8/SERT	ARRB1/GR, MAPK8/PREP
Any lifetime "treatments" (2)	DPP4, MKP1, Gi2/CREB, MKP1/ERK1, MAPK8/PREP	
More tobacco use	Gi2/ARRB1	SERT
Increased anxiety		Gs, Gs/CREB2, ERK1/GR, ARRB1/GR
Decreased appetite	IL-8/SERT	SERT, ERK2, Gs, MKP1, PBR, MAPK14
Any drug use last 3 months	Gi2/CREB, MKP1/ERK1, MKP1/MAPK14, MAPK8/PREP	ERK1/GR
Decreased Energy	DPP4, MKP1/MAPK14	
Increased feeling low	DPP4, ERK2, MKP1, ERK2/MAPK8+14, Gi2/ARRB2	Gs/ARRB2
More sleep problems	DPP4, Gi2/ARRB2, Gs/ARRB2	
More alcohol use	PBR	
Decreased concentration	DPP4	
Males		ERK2/MAPK8+14, Gi2/ARRB1, Gs/ARRB2, Gi2/CREB, Gs/CREB2, ERK2/ARRB1

Table 25 summarizes the results from table 23 and 24. The table indicates gene expression patterns that may define intermediate phenotypes and subgroups of patients. The results are commented in the hypotheses in the text. (1) "Experiences" refers to episodes of depression, anxiety, etc. (2) "Treatments" refers to treatment by a physician for depression, anxiety, etc.

Based on table 25, the ABS control data predict the following trends in depressed patients and, at the same time, define the corresponding intermediate phenotypes:

1. SERT expression will be lower in patients - as SERT is lower in controls with decreased appetite and in controls who smoke.
2. DDP4 expression will be increased in patients - as DDP4 is increased in controls who have decreased concentration and energy, feeling increasingly low, having more sleep problems and been treated in their life for e.g. depression, anxiety, etc.
3. ERK2 expression will be altered in patients – as ERK2 is decreased in controls with decreased appetite and increased in controls who feel increasingly low.
4. G alpha s expression will be decreased in patients – as G alpha s is decreased in controls with increased anxiety and decreased appetite.
5. MKP1 expression will be altered in patients – as MKP1 is increased in controls who feel increasingly low and who have been treated in their lifetime for e.g. depression, anxiety, etc. On the other hand MKP1 is decreased in controls with decreased appetite.

6. PBR expression will be altered in patients – as PBR is increased in controls who have more than one drink per day and decreased in controls with decreased appetite.
7. MAPK14 expression will be lower in patients – as MAPK14 is decreased in controls with decreased appetite.
8. ERK2/MAPK8+14 expression will be altered in patients - as ERK2/MAPK8+14 is lower in male controls and increased in controls who feel increasingly low.
9. Gi2/ARRB1 expression will be altered in patients - as Gi2/ARRB1 is decreased in male controls, increased in controls who smoke, and increased in controls who ever experienced severe depression, severe anxiety, alcohol abuse, etc.
10. Gi2/ARRB2 expression will be increased in patients – as Gi2/ARRB2 is increased in controls with more sleep problems and increased in controls who feel increasingly low.
11. Gs/ARRB2 expression will be altered in patients – as Gs/ARRB2 is decreased in male controls, in controls who feel less low and increased in controls with more sleep problems.
12. Gi2/CREB expression will be altered in patients – as Gi2/CREB is decreased in male controls and increased in controls who had any illicit drug use, who ever experienced e.g. severe depression, who ever were treated for e.g. depression, anxiety, and who had prescription drugs and/or illicit drugs the last 3 months.
13. Gs/CREB2 expression will be decreased in patients - as Gs/CREB2 is decreased in male controls and in controls with increased anxiety.
14. MKP1/ERK1 expression will be increased in patients - as MKP1/ERK1 is increased in controls who ever experienced e.g. severe depression, severe anxiety, in controls who ever were treated for e.g. depression, anxiety, and in controls who had prescription drugs and/or illicit drugs the last 3 months.
15. MKP1/MAPK14 expression will be increased in patients - as MKP1/MAPK14 is increased in controls who had prescription drugs and/or illicit drugs the last 3 months and increased in controls with decreased energy.
16. ERK1/GR expression will be decreased in patients – as ERK1/GR is decreased in controls with increased anxiety and in controls who had prescription drugs and/or illicit drugs the last 3 months.
17. ERK2/ARRB1 expression will be altered in patients - as ERK2/ARRB1 is decreased in male controls and increased in controls who ever experienced severe depression, severe anxiety, etc.
18. MAPK8/PREP expression will be altered in patients - as MAPK8/PREP is decreased in controls who ever experienced e.g. severe depression and increased in controls who ever were treated for e.g. anxiety, and increased in controls who had prescription drugs and/or illicit drugs the last 3 months.

19. ARRB1/GR expression will be decreased in patients - as ARRB1/GR is decreased in controls who ever experienced e.g. severe depression, and in controls with increased anxiety.
20. IL-8/SERT expression will be increased in patients - as IL-8/SERT is increased in controls who ever experienced e.g. severe depression, and in controls with decreased appetite.

In the next chapter, I compare the 20 hypotheses to a small group of severely depressed patients (Lundbeck confidential data, and thus not available). At the present, it does seem like gene expression patterns can be used to segment control subjects based on various clinical variables. Expression differences were clearly identified between genders.

7.3.2 Gene expression subgroups identified via recursive partitioning

Recursive partitioning (RP) is described in the Statistical methods chapter, section 5.6. RP was mostly used as a classification technique, but in a few cases, I also applied RP for the identification of possible (intermediate) phenotypes. This was done to look for distinct gene expression profiles in the SH ABS and the DC controls. Below, I also present the result of applying RP to a matched (explained later in this subsection) data set consisting of 20 DC controls, 21 SH ABS controls and the 21 BPD patients.

In figure 30, all the DC and SH ABS controls are correctly placed (no misclassifications) in a decision tree created from just seven gene expressions. Using a maximum of five splits (genes), every control is correctly classified and the responsible genes for each subgroup are easily identified.

From figure 30, it can be seen from e.g. the right branch of the tree that just by looking at the logarithm of the expression values of MAPK14 and ODC1, 43 of the 58 SH ABS are correctly identified (0/43 in the figure means 0 subjects are misclassified and 43 subjects correctly classified) . Looking e.g. at MAPK14, CREB2 and ARRB2, 8 SH ABS are identified. In this way, RP was used to identify two distinct gene expression profiles in the total SH ABS population. Looking at the DC controls, the expression levels of MAPK14, CREB2, CREB1 and ODC1 can identify 80 out of 89 DC controls.

Neglecting the small subgroups (<5 subjects), it thus seems like there might be two SH ABS subgroups and one large DC subgroup, each subgroup identified by looking at the expression profiles of no more than 4 genes.

Pred / y	DC	SH ABS
DC	89	0
SH ABS	0	58

0 = DC
1 = SH ABS

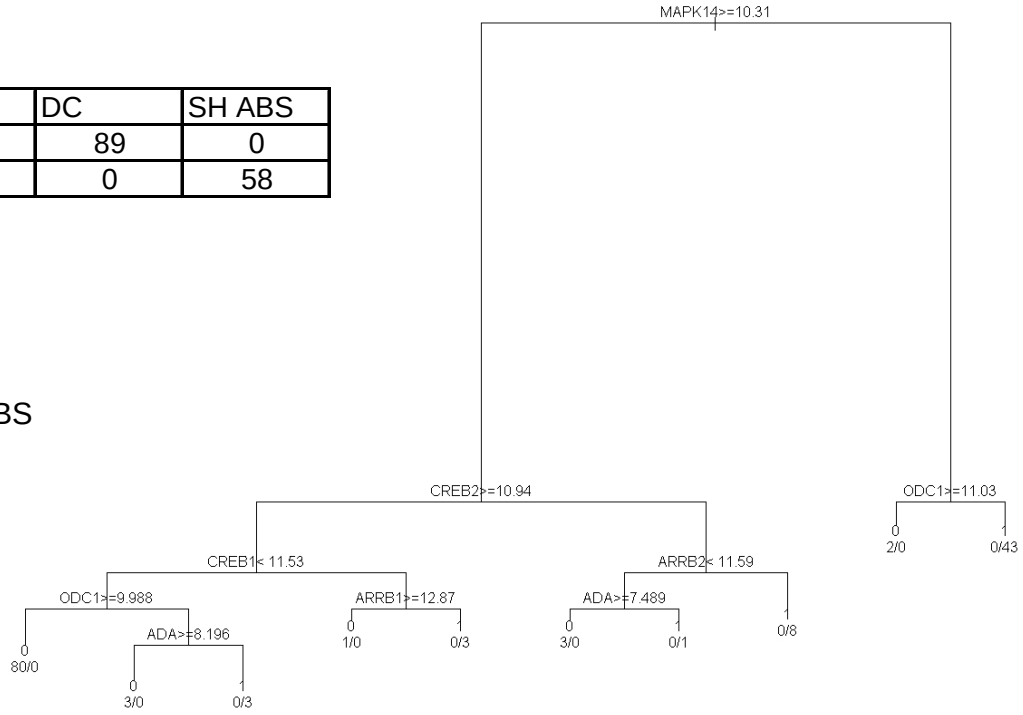


Figure 30 shows a decision tree of the DC and SH ABS controls. The table shows that there are no misclassified controls. See the text after the figure for explanation of this decision tree. The 'rpart' package in R was used to carry out recursive partitioning.

As mentioned in the statistics chapter, overfitting is an issue with RP, but still, I believe the results above indicate that RP could be used for the identification of control subgroups and thus of possibly intermediate phenotypes.

RP was also applied to three groups of almost the same size consisting of 20 DC controls, 21 SH ABS controls and the 21 BPD patients. The 21 SH ABS controls were picked by a gender and age match with the BPD patients. The 20 DC controls are a 'super healthy' DC control group matched with the SH ABS by having a BMI < 30 and not having used any drugs in the past 3 months. In figure 31, all subjects are correctly placed in the decision tree created from only four gene expressions (out of possible 25). Here, using a maximum of four genes, every subject is correctly classified.

Ignoring the small subgroups (<5 subjects) in figure 31, it can be seen that almost all BPD patients (20 out of 21) can be identified by only considering the expression value of GR (and ARRB2) in this data set (20/0/0 in the figure means 20 subjects are classified as BPD, 0 as DC, and 0 as SH ABS (the last digit)). On the other hand, 17 out of 20 DCs can be identified by GR, SERT, and CD8 beta. Almost all SH ABS (20/21) are identified by GR, SERT, CREB2 (and ADA).

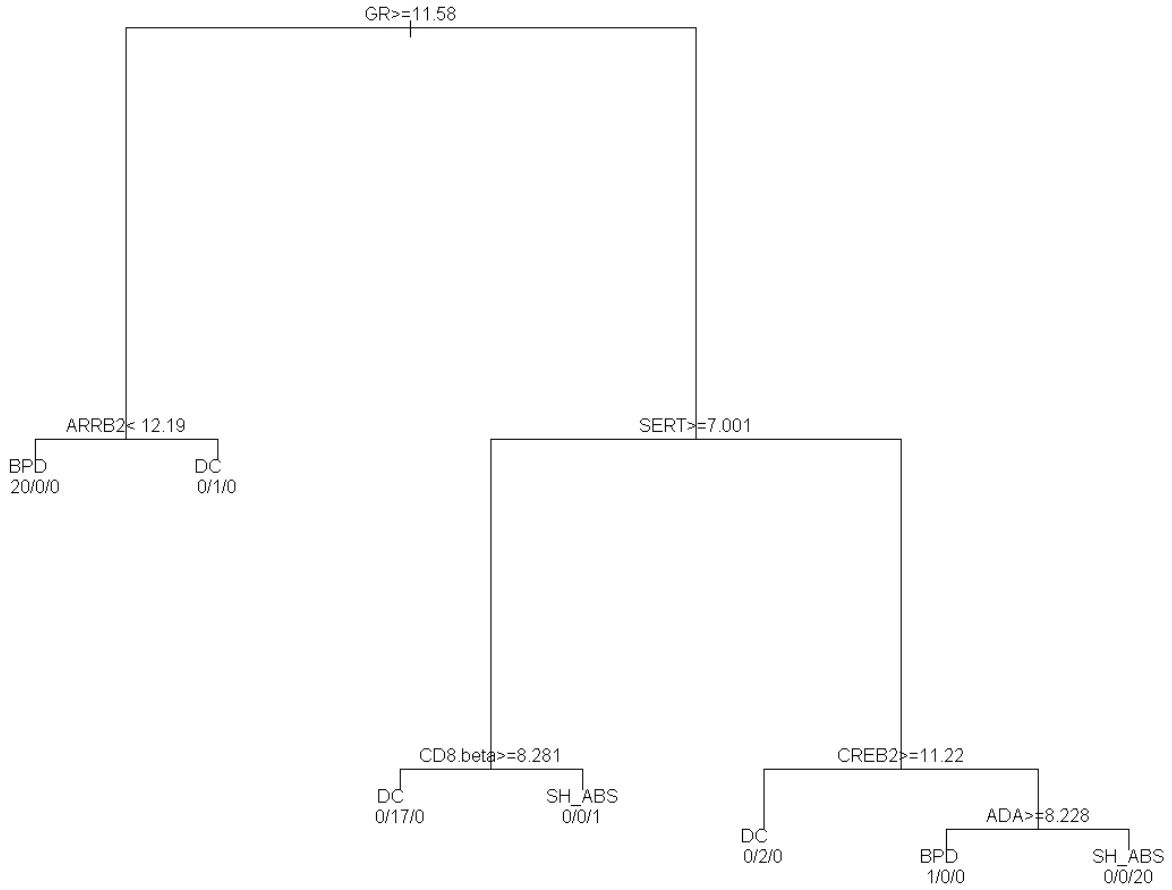


Figure 31 shows a decision tree with 20 DC controls, 21 SH ABS controls and the 21 BPD patients. There are no misclassified subjects. See the text for explanation of the decision tree. The 'rpart' package in R was used to carry out RP.

Especially in this case with only one gene expression - GR – indicating almost perfect separation between BPD patients and the controls, overfitting must be considered. However, the tree does illustrate that the two control groups are more similar than the patient group; they both share the GR threshold and diverge mainly on the value of SERT. This is an argument for pooling the controls groups (see section 7.3.6).

7.3.3 Possible BPD phenotypes through CCA

Canonical correlation analysis (CCA) was used in combination with different sets of genes to identify various possible phenotypes among the BPD patients. This was done for cohort 1, since clinical variables were available for this group. In the Statistical methods chapter, section 5.5, CCA was demonstrated on the BPD group using 4 genes.

Here I present the results of a regularized CCA analysis using 11 genes, and at the end of this subsection compare the result with the 4 gene CCA case. These 11 genes were – like the 4 genes - selected based on a comparison with the SH ABS controls. The 11 genes were differentially expressed between the two

groups assessed with univariate tests, however here without applying the Bonferroni correction. We wanted to investigate if these differentially expressed genes could give rise to hypotheses about BPD subtypes.

Like in the 4-gene CCA case, the set of 11 gene expressions and 6 clinical variables was chosen by first performing CCA with all clinical variables (and the 11 gene expressions) and then leaving out the ones with a correlation less than 0.5. The result is summarized in the figure 32 with the graph of variables to the left, and the graph of individuals to the right.

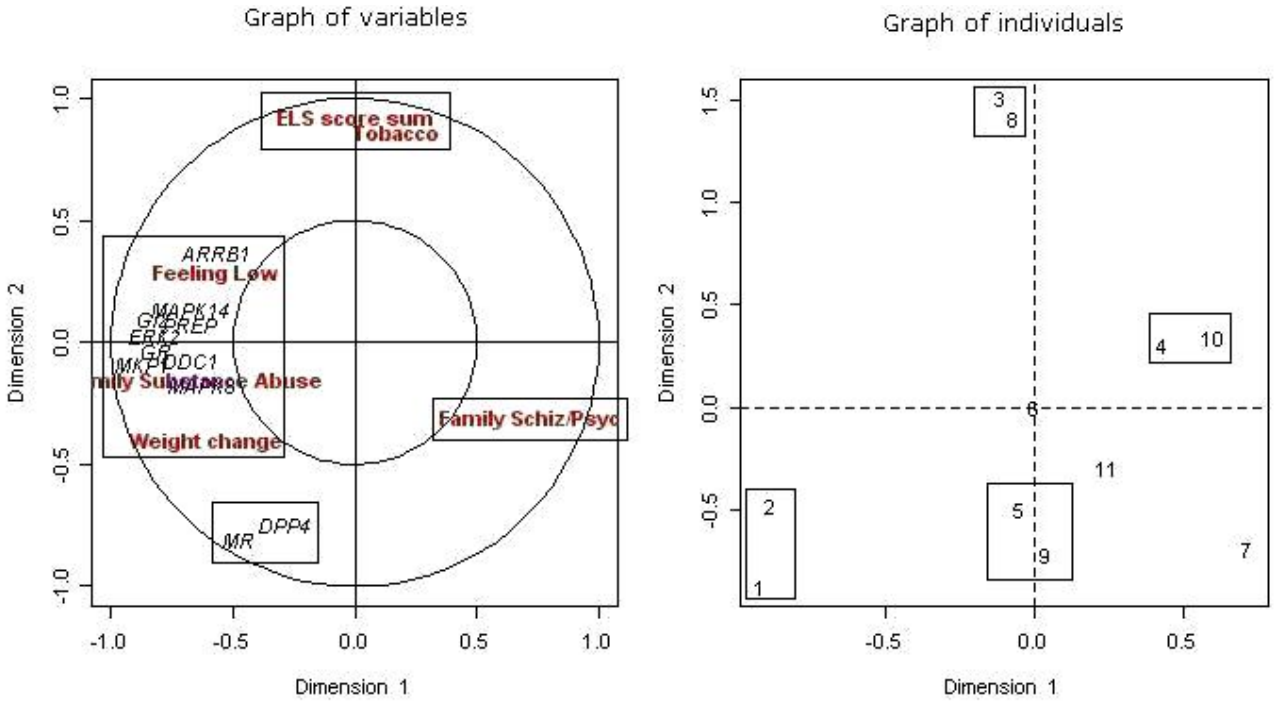


Figure 32 shows a CCA output for 11 BPD patients and a subset of 6 clinical variables and 11 gene expressions. The left graph is the graph of variables (clinical and gene expressions), and to the right, the graph of individuals (with the 11 BPD patients). The two graphs are commented in the text. Dimension 1 and 2 represent the most and second most important canonical correlation dimension, respectively. The plots were done in R.

On the graph of variables, there are no variables in the inner circle which is due to the fact that I left out the unimportant variables. I have marked several groups of clinical variables and genes with boxes. The boxes contain variables and genes with a strong relation (since they are projected in the same direction from the origin). Also, the reader may recall, that the greater the distance from the origin, the stronger the relationship. From this graph, it can be seen that e.g. ELS score sum and Tobacco (shown in the upper box) are strongly related and this also goes for DPP4 and MR in the bottom of the graph, while Family Schiz/Psyc (a family history of schizophrenia or psychosis) stands alone. Opposite is a large group of 9 genes and 3 clinical variables with

Feeling low being close to ARRB1, etc. I checked DPP4 and MR, and their correlation was 0.94, which is nicely reflected on the graph.

In the graph of variables, ELS score sum and Tobacco are closely related at the top of the graph while DDP4 and MR are closely related almost opposite (at the bottom of the graph). On the right graph (individuals), different subgroups of borderline patients are marked. In this graph, BPD3 (borderline patient number 3) and BPD8 are close together at the top (shown with a box) while e.g. BPD5 and BPD9 are close together at the bottom. This can be interpreted as BPD3 and BPD8 having high ELS score sums and high Tobacco score while BPD5 and BPD9 have low ELS score sums and low tobacco score. At the same time all the borderline patients below the zero line (dimension 2) have high DPP4 and MR expressions while all the borderline patients above the zero line have low DPP4 and MR expressions. Furthermore, BPD4 and BPD10 have a high Family Schiz/Psyc score, and low score of the group of 9 genes and 3 clinical variables. The opposite applies for BPD1 and BPD2.

Compared with the 4-gene CCA example in the statistics chapter, it can be seen that there is not a big difference between applying CCA to a 4-gene vs 11-gene analysis of the same data. Using different gene sets highlights different phenotype patterns in the data. In the 4-gene CCA case, the clinical variables are slightly different from the 11-gene CCA case; there are still 6 clinical variables, however the variables ELS and Tobacco have been replaced by Anxiety and 3 month drug use. Also, the clinical variables and genes are located similar in the two cases. On the graph of individuals, BPD1 and BPD2, and BPD10 and BPD11 are closely related in both cases. The other marked BPD subgroups are somewhat different with the biggest difference being a switch of BPD3 and BPD9 between the top and bottom marked subgroups. This is connected to the different clinical variables and genes used in the two cases.

All in all, CCA seems to generate hypotheses about subtypes of patients by linking their gene expression profiles to their clinical variables.

7.3.4 Clinical variables explaining the most variance in a gene

In the Statistical methods chapter, section 5.8, I described stepwise regression, and illustrated this statistical method with two examples in table 17. I applied stepwise regression to the ABS control groups with focus on the depression-relevant clinical variables described in the Study design chapter and in section 7.1.1. I wanted to identify a minimal set of clinical variables per gene expression that explained as much of the variance in that gene expression as possible. The cut-off was set to ~15% as explained in section 5.8. Table 26 lists gene expressions, clinical variables, and interactions between clinical variables, the latter two accounting for at least ~15% of the variance in the shown genes.

	SERT	ERK2	MKP1	ODC1	PBR	S100A10	P2X7
Gender					i1, i2 (male)		
Lifetime experiences			i1				
Lifetime treatments					i3		
Tobacco	x			i	i3		
Appetite Change	x	x	x	x	i4, i6	x	x
Coping	x						
Enjoyment				x			
Feeling low		i					i
Sleep Problems		i		x	i4, i5, i7	x	i
Weight Change	x			x	i2, i8		
Age	x				i6, i7, i8	x	
BMI				i			
ELS score	x		i1, i2				
Recent Stress score			i2				
Symptom score sum					i1, i5		
R^2 without interactions	16%	14%	14%	18%	17%	15%	17%
R^2 with interactions	16%	18%	16%	20%	36%	15%	21%

Table 26: Results of stepwise regression on the depression-relevant clinical variables for the ABS controls. R^2 is reported. An 'x' denotes the clinical variable(s) remaining after the stepwise regression, and an 'i' denotes one part of an interaction term remained after a second stepwise regression. The i-numbers indicate which variables are interacting. The logarithm is applied to all the gene expressions. The analyses were done in R.

None of the variation in the other genes (not included in table 26) was explained well by the resulting linear model (<15% of the variance). Looking at the table, e.g. 16% of the variation in SERT could be explained by the combination of tobacco use, appetite change, coping, weight change, age and ELS (early life stress) score. There were no significant interactions between any of these variables in SERT. On the other hand, looking at e.g. P2X7 17% of the variation in this gene could be explained by the combination of appetite change, feeling low and sleep problems (no interaction considered here). There was one significant interaction between feeling low and sleep problems, and by including this interaction 21% of the variation of P2X7 could be explained. The remaining results are not outlined, but these findings do indicate that a linear combination of various clinical variables with possible interactions considered, can explain some part of the variance in a gene expression.

7.3.5 Gender differences?

One of the first things I investigated was the extent of possible gene expression differences among the genders. From a clinical perspective, depression, BPD and PTSD are mental disorders known to affect more women than men (see chapter 2). In the available data, patient gender differences could not be investigated (not enough data from men among the BPD patient, and no women among the PTSD patients), so we looked at gender differences in various comparisons between and among the ABS and DC controls. Back then, I applied the initial classifiers Pelora and SLR, as well as correlation tests

and univariate tests. The actual goal was to see if gender made a significant impact on the results obtained from these classification and statistical approaches.

First, results from Pelora were explored. Since Pelora is intended for microarray analysis (see the classification chapter), we wanted to minimize any possible overfitting to the data by only considering the first Pelora cluster/ gene set. In figure 33, three Pelora plots show almost perfect separation of the DC controls vs. the SH ABS, the DC males vs. SH ABS, and the DC females vs. SH ABS, only considering the first Pelora cluster (x-axis genes).

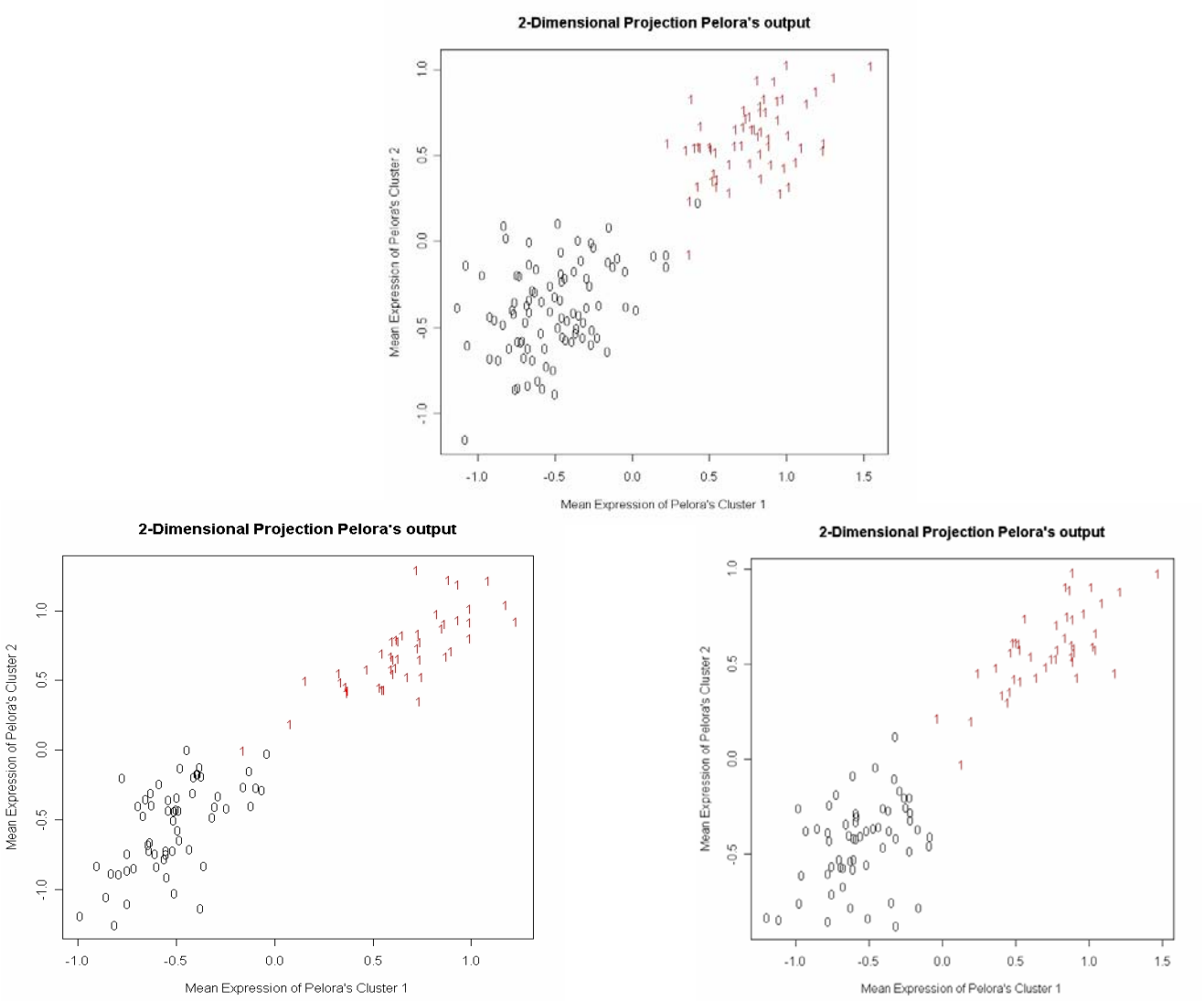


Figure 33 shows three Pelora plots, all with an almost perfect separation using only the x-axis genes. The top plot shows all DC controls vs all SH ABS controls; x-axis genes are MAPK14, ODC1, CREB1, CREB2 and ADA. The bottom left plot shows the DC males vs the SH ABS controls; x-axis genes are MAPK14, ODC1, CREB1, CREB2, ADA, MR, MAPK8 and Gs. The bottom right plot shows the DC females vs the SH ABS controls; x-axis genes are MAPK14, ODC1, CREB1 and CREB2. DC controls are the 0's and the SH ABS are the 1's. The Pelora plots were done in R.

Below is a summary of Pelora cluster/gene set 1 genes (consistent genes are marked in bold, while italics marks genes appearing in two out of three comparisons):

- DC males only vs SH ABS: **MAPK14**, **ODC1**, **CREB1**, **CREB2**, *ADA*, MR, MAPK8, Gs
- DC females only vs SH ABS: **MAPK14**, **ODC1**, **CREB1**, **CREB2**
- All DC subjects vs SH ABS: **MAPK14**, **ODC1**, **CREB1**, **CREB2**, *ADA*

These results showed that basically the same genes were identified by Pelora regardless of gender selected in the DC group.
SLR showed a similar result (data not shown).

Two negative Pelora results are shown in figure 34.

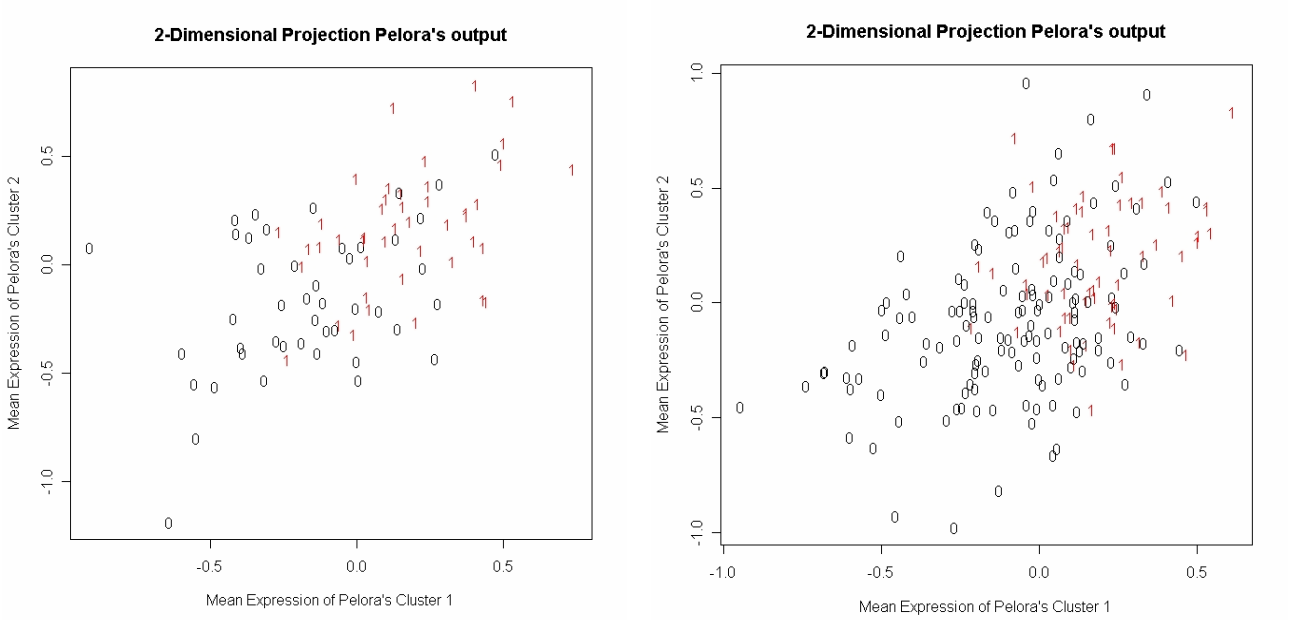


Figure 34 shows two Pelora plots. The left plot shows the DC males vs the DC females (DC controls are the 0's and the SH ABS are the 1's), and the right plot shows all male controls (0's) vs all female controls (1's) (UK controls excluded). The Pelora plots were done in R.

From figure 34, using Pelora, it was difficult to separate males from females within the DC control group (N=89), or within all controls (DC, SH ABS and PTSD controls, N=226). Likewise, SLR showed a poor performance, and picked similar genes as Pelora (results not shown).

Based on the Pelora and SLR results, gender thus did not seem to play a large role in determining gene expression patterns.

Next, we tested the differences in Spearman correlations split by gender. In total, 300 gene expression correlations were made within the DC or SH ABS

controls. The gene expression correlation pairs between the two groups were compared on the 1% significance level (as described in the statistics chapter):

- DC all subjects: 12 correlations were significantly different from the SH ABS (4%)
- DC only males : 9 were different (3%). 7/9 were seen using "all subjects"
- DC only females: 5 were different (2%). 4/5 were seen using "all subjects"

The Spearman correlations thus seemed similar regardless of gender selected in the DC group.

Finally, univariate tests were applied. The DC group was split by gender and the expression levels were compared to the SH ABS. Of 25 genes tested, 6 showed statistically different expression in one gender only (ARRB2, DPP4, IDO, IL-6, MR and RGS2). However, only one of those genes (MR) was selected by SLR or Pelora to discriminate DC controls from the SH ABS controls.

Based on the univariate comparisons, there were some expression differences detected between the genders with the DC group. However, these genes were not utilized by the multivariate techniques Pelora or SLR to distinguish the two groups. All in all, the gender differences did not seem critical.

7.3.6 Pooling of control groups into one group?

After having investigated various aspects of the ABS control group (see section 7.2, 7.3.1, and 7.3.4), we received the DC control data. Initially, we had a hypothesis stating the two control groups would be very similar in terms of gene expression profiles, and that it therefore would be natural to combine the two control groups for comparisons to patients. The two first points below made us realize that the expression profiles were not that identical:

1. Univariate tests revealed that 22 out of 25 genes were significantly different expressed between the two control groups. Only MR, P2X7, PREP did not differ significantly.

2. Pelora and SLR were able to separate the two groups based on their expression values, see figure 33 (top plot) and the previous section (7.3.5).

On the other hand,

3. Spearman correlation tests between the ABS controls (299 subjects) and the DC controls indicated that the two groups were 90% the same,

statistically speaking. This result was obtained by calculating the Spearman correlations between the 25 gene expressions within each group, and then compare gene expression pairs between the groups as described in chapter 5, section 5.4.

4. We thereafter turned to the clinical information to see if the expression differences had anything to do with clinical differences. A SH DC group was defined the same way the SH ABS was defined, that is, having no drug use the last three months and low BMI). It turned out, that the SH DC subjects had ultra clean personal and family histories (no personal lifetime psychiatric disorders of any kind, no first rank family history of schizophrenia or suicide, just one person had both a mother and father diagnosed with depression).
   The univariate tests now revealed that 15 out of 25 genes were significantly different between the groups. This could indicate that the initial differences noted between the groups (the 22 out of 25 genes) may have been influenced by the way the SH ABS subjects were selected, not because Danes and Americans were that different regarding expression.

5. The gene expression patterns in DC vs all 299 ABS subjects, with respect to specific clinical variables, was more than 90% in agreement (see appendix 10).

6. Performing Pelora and SLR with the DC and SH ABS combined vs the BPD patients yielded a very good separation of the two groups (data not shown). A similar result (data also not shown) was also obtained using Pelora and SVM with varselrf in connection with the acute PTSD patients; both these classifiers see all SH ABS as belonging to the DCs and none as acute PTSD patients.
   Therefore, the control groups, though not identical, were both segregated from the patients. These results further indicated that the two control groups were more similar than different compared to the patients.

Based on the above arguments (two arguments against and four arguments for combining the Danish and American control groups), we decided to pool the two control groups into one control group for comparison with patients.
In a broader perspective the rationale for pooling the two control groups was that, since we are dealing with controls groups from Denmark and healthy US controls, we should operate with a combined control group that spanned the expression differences in and between these groups, because in this way a "normal"/control gene expression pattern would contain more variation, and a disease expression pattern should be different from the natural gene expression variation in controls.

7.3.7 Pooling of all control groups into a single large group?

After realizing that the gender differences in the control expression profiles were not large, and we had combined the DC and SH ABS control groups into one control group, it was natural to investigate whether all control groups could be pooled into one large control group.

First, univariate tests and scatter plots were employed. A few examples are shown in figure 35 and figure 36 comprising the DC, SH ABS, SH DC, UK (time point 1) and PTSD controls. These plots show that the SH ABS appears to be the most different from the other groups, but that there, nonetheless, seems to be quite some overlap of points.

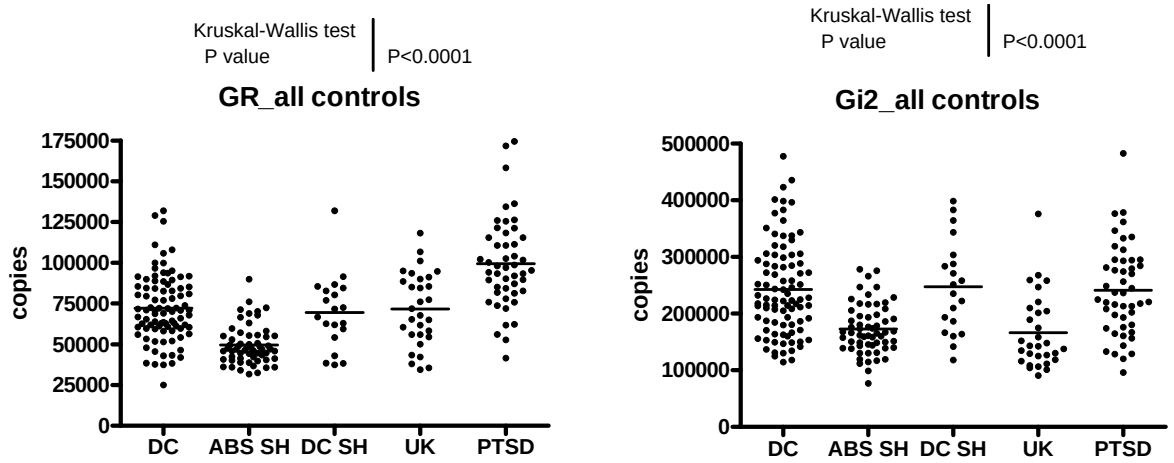


Figure 35 shows two scatter plots for the gene expression GR (left) and Gi2 (right). In the left plot, the SH ABS expression level is clearly less compared to the other groups. In the right plot, the SH ABS and UK expression levels are less compared to the other groups. The plots were done with GraphPad.

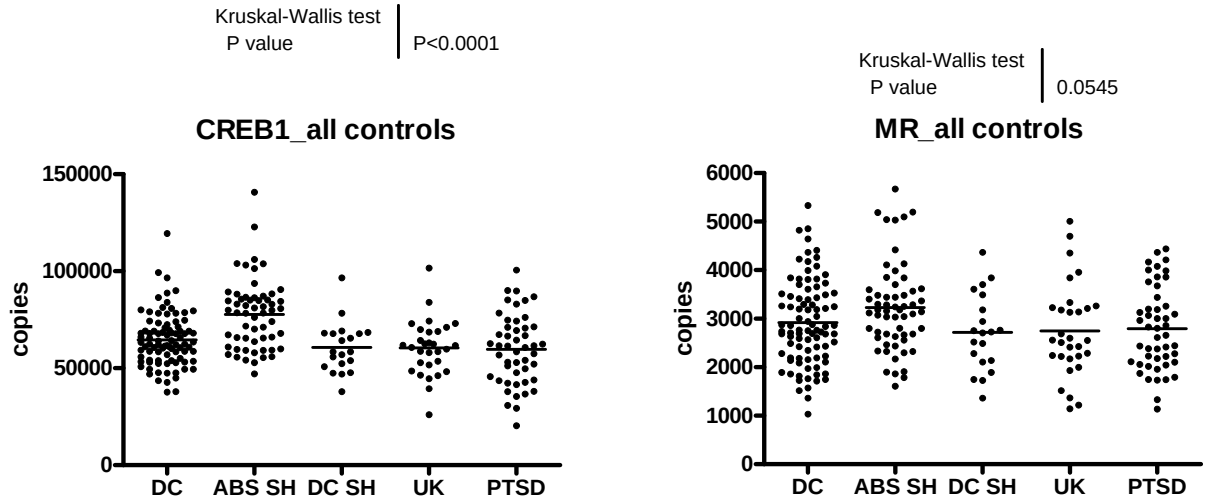


Figure 36 shows two scatter plots for the gene expression CREB1 (left) and MR (right). In the left plot, the SH ABS expression level is higher compared to the other groups. In the right plot, the expression levels of all control groups are the same. The plots were done with GraphPad.

The univariate tests and plots (all data/results not shown) showed that the expression level was:

- Clearly less in ABS SH compared to all other groups: 10/25 genes
- Low in ABS SH plus other groups: 6/25 genes
- Most/all groups were the same level: 7/25 genes
- Greater in ABS SH than all other groups: 2/25 genes

These results led us to believe the expression data for the four control groups (DC, SH ABS, UK and PTSD) overlapped somehow, and that it would be a good idea to investigate the control groups versus the patient groups in various Pelora and SLR comparisons. The results of the Pelora plots are summed up in table 27 (however, the Pelora plots are not shown). In table 27, various combinations of controls vs one patient group were used as a training data set, and then a control or patient group, that had not been part of the training data, was used as a validation data set. This was done to see how Pelora would classify an 'unknown' group in the sense that the unkown group had not been part of the classifier training process.

The results in table 27 were summed up:

- Comparing related training sets; larger data sets seemed to be better for validation performance (table 27, compare sets 1 and 2, sets 3 and 4)
- Large control groups (the bottom part of table 27) seemed to do a good job of classifying validation sets (see training set 5)
- Note 1: 100% of the controls in the training set were properly classified as controls. The patients in the training set were classified correctly 66% of the time (data not shown).
- SLR produced similar results to Pelora for each set tested.
- (It was noted how GR always appeared in the gene list.)

Set no.	Training sets	Validation sets	Score with validation set	Genes identified (1st Pelora gene set only)
1	53 SH ABS vs 16 BPD	89 DC	82% scored as controls	GR
2	16 SH ABS (matched) vs 16 BPD	20 SH DC	43% scored as controls	GR
3	89 DC vs 21 BPD	59 SH ABS	100% scored as controls	GR, Gi2
4	20 SH DC vs 21 BP	21 SH ABS	52% scored as controls	GR, ARRB2
5	59 SH ABS and 89 DC vs 21 BPD	30 UK	100% scored as controls	GR, ERK1, ERK2, ARRB2

Table 27: Summary of Pelora results for various comparisons between control and patient groups. * is commented in the text below the table, while the table is summed up in the text above the table. The analyses were done in R.

Overall, it seemed like the different control groups were more alike than they were different. Using larger control groups for Pelora and SLR seemed to produce better classification than when smaller groups were used.
Based on these results, we decided to combine multiple control groups when we had to make classification comparisons to patients, see section 7.5. Thus, like in the previous section, the pooled large control group would span the biological gene expression variation in all controls combined.

7.4 Expression levels across multiple time points

As the reader may recall, in the UK control group, three time measurements were made; Day 0 at 8 am, Day 0 at 2 pm and Day 1 at 8 am. I applied repeated measures ANOVA tests (see the statistics chapter, section 5.3) in order to investigate whether any of the gene expressions differed significantly between the three time points on the 1% significance level.

Out of 29 gene expressions, 5 gene expressions differed significantly between the time points; CD8 beta, IL-8, MKP1, MR and ODC1. In figure 37, scatter plots for MR and IL-8 are shown. For IL-8, there is a significant difference between baseline (day 0, 8 am) / 6 HR (day 0, 2 pm) and 24 HR (day 1, 8 am), while for the four other gene expressions the significant difference is between morning and afternoon measurements.

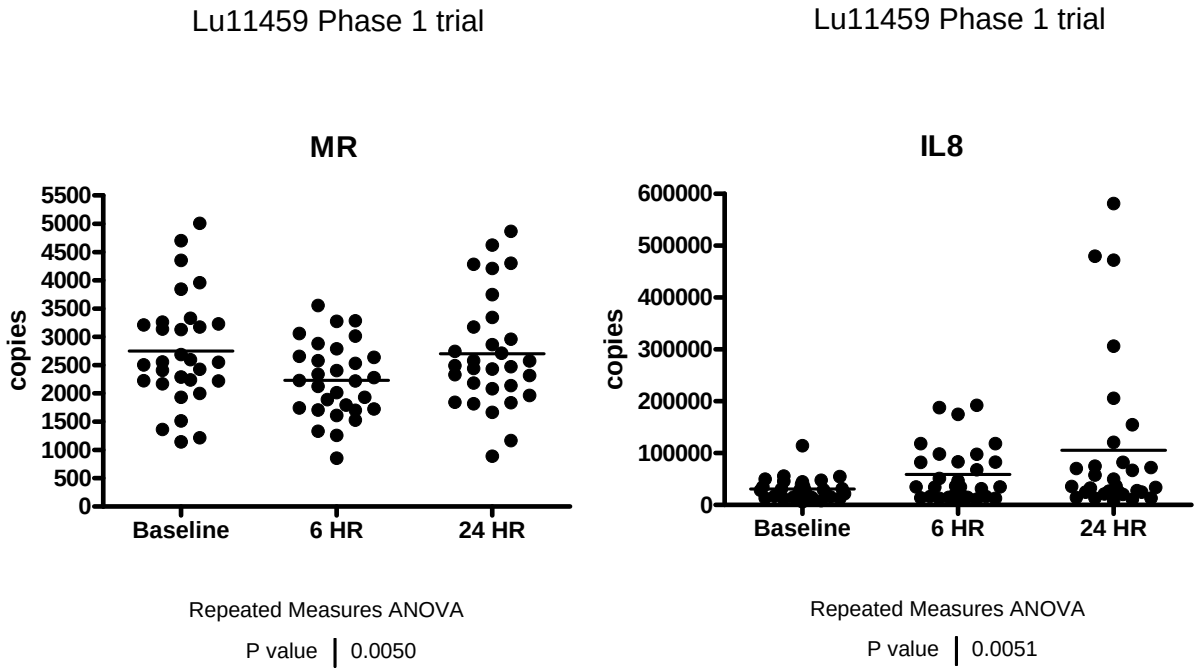


Figure 37 shows two scatter plots for the gene expressions MR (left) and IL-8 (right) at three time points. In the left plot, there is a significant difference between baseline (day 0, 8 am) / 24 HR (day 1, 8 am) and 6 HR (day 0, 2 pm). In the right plot, there is a significant difference between baseline / 6 HR and 24 HR. The plots were done with GraphPad.

At the present state of the investigation, the US Lundbeck group was not interested in including genes in the study that displayed a circadian effect in whole blood. Three time points were far from adequate to establish any possible circadian effect of the gene expressions, so I checked Pubmed for articles involving the 5 genes and circadian behavior. Disregarding other species than man and other tissue like liver, I only found MR to have a circadian pattern in man (213), (214), however not in blood.

Based on the repeated measures ANOVA results for the day and afternoon differences, and the Pubmed findings, even though circadian effects might be involved, no conclusion either way could be made with so few time points.

For IL-8, there seemed to be a difference in the expression measurements on consecutive days (both in the morning). However, inspection of the scatter plots, see the right figure in figure 37, suggested that this was caused by a few individuals (N=5), not the entire group (N=30). If the 5 subjects were removed from the analysis, the statistical difference disappeared. Also, it was not unreasonable to imagine that in any given individual, these genes could be subject to more noise, since they are proinflammatory cytokines, and expression can be altered by many factors.

Overall, the 5 genes remained in the list of genes measured in the study, and the first time point was used when UK controls were included in any analysis.

7.5 Variable selection and classification among various groups

In the classification chapter, the classifiers and variable selection techniques SVM (support vector machine) with varselrf (variable selection based on random forests), SLR (stepwise logistic regression – both for gene selection and classification) and RPART (recursive partitioning – both for gene selection and classification) were identified as the most promising methods for a multivariate approach to classification and variable selection. In this section, I present the results of applying these methods to various comparisons among the control and patient groups. Also, following the results in section 7.3.7, the various patient groups are compared to a pooled control group.

It should be noted, that at this point of the thesis work, we (the US Lundbeck group and I) did not consider univariate tests anymore for variable selection, since we were interested in accuracy, PPV (positive predictive value – relates to classification of patients) and NPV (negative predictive value – relates to classification of controls) for the applied classification techniques.

7.5.1 2-group comparisons

For the two group comparisons, it was possible to report PPV and NPV together with the accuracies. Four main 2-group comparisons were made; controls vs BPD patients, controls vs PTSD acute patients, controls vs PTSD in remission and controls vs trauma without PTSD. Below I present the results of these four comparisons in summary tables including genes selected to differentiate groups, PPV, NPV and both the actual accuracy values as well as the permuted accuracy values (in parenthesis in the table), all in percentages. I also indicate whether the accuracy values of a comparison is significant on the 1% significance level compared to permuted values as described in the 'Classification and variable selection procedure' section of the classification chapter.

In the summary tables 'All controls' were derived from 4 different subject groups (SH ABS, DC, UK and PTSD controls) and 25 gene expression values were used for comparison. 'Controls' were derived from 3 different subject groups (DC, UK and PTSD controls) and 29 gene expression values were used for comparison (the same 25 used above plus 4 additional).

First, in table 28, controls were compared with the borderline personality disorder patients. Here, 'all controls' included 254 subjects, 'controls' 196 subjects and 'BPD patients' 21 BPD patients. In general, genes selected were very similar regardless of variable selection technique. 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 as described in chapter 6. The predictive values were high except for RPART.

Comparison	Classifier	Genes selected to differentiate groups	PPV	NPV	Accuracy (permuted)
All controls vs BPD patients (25 genes)	SVM/varselrf	ERK1, Gi2, GR, MAPK14, MR	82	98	98 (92)
	SLR	Gi2, GR, MAPK14, MR	93	100	99 (92)
	RPART	Gi2, GR	68	98	96 (91)
Controls vs BPD patients (29 genes)	SVM/varselrf	Gi2, GR, MAPK14	97	98	98 (90)
	SLR	Gi2, GR, MAPK14, MR	93	99	98 (88)
	RPART	Gi2, GR	68	98	95 (88)

Table 28: Summary of controls vs BPD patients comparisons. PPV, NPV 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). The analyses were done in R.

In table 29, controls were compared to PTSD acute patients. All controls included 254 subjects, controls 196 subjects and 66 PTSD acute patients. In general, genes selected depended on the variable selection technique but not on the control group. 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 in the 29 gene comparison. The positive predictive values were considered marginal but still a lot better than the PPVs from the controls vs PTSD in remission comparison described next.

In table 30, controls were compared to PTSD in remission. All controls included 254 subjects, controls 196 subjects and 41 PTSD in remission. The two groups could not be separated by any classifier on the 1% significance level (data not shown, but the reader may notice how close the actual and permuted accuracy values are), so the gene expression profiles seemed to reflect the clinical diagnosis well. The PPV values were markedly lower than in the PTSD acute patients case (table 29).

Comparison	Classifier	Genes selected to differentiate groups	PPV	NPV	Accuracy (permuted)
All controls vs PTSD acute patients (25 genes)	SVM/varselrf	ARRB2, ERK2, RGS2	70	87	86 (78)
	SLR	ARRB1,ARRB2, CD8 beta, ERK2, IDO, IL-6, MR, PREP, RGS2	81	90	88 (78)
	RPART	ADA, ARRB1, ARRB2, CREB2, ERK1, GR, MKP1, P2X7, RGS2	65	88	82 (70)
Controls vs PTSD acute patients (29 genes)	SVM/varselrf	ARRB2, ERK2, RGS2	79	83	80 (73)
	SLR	ARRB1,ARRB2, CD8 beta, ERK2, IDO, IL-6, MR, ODC1, PREP, RGS2	77	87	84 (71)
	RPART	-	48	82	72 (64)

Table 29: Summary of controls vs PTSD acute patients comparisons. PPV, NPV 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) except for RPART in the 29 gene comparison (p-value: 0.02349). The analyses were done in R.

Comparison	Classifier	Genes selected to differentiate groups	PPV	NPV	Accuracy (permuted)
All controls vs PTSD in remission (25 genes)	SVM/varselrf	-	52	89	88 (86)
	SLR	-	46	91	86 (86)
	RPART	-	39	90	83 (81)
Controls vs PTSD in remission (29 genes)	SVM/varselrf	-	49	86	82 (82)
	SLR	-	33	86	80 (81)
	RPART	-	28	86	76 (75)

Table 30: Summary of controls vs PTSD in remission. PPV, NPV and accuracy values are in percentages. Permuted accuracy values are in parenthesis in the last column. No accuracy values are significant compared to permuted values (data not shown). The analyses were done in R.

Finally, in table 31, controls were compared to patients with trauma but without PTSD. All controls included 254 subjects, controls 196 subjects and 87 trauma patients without PTSD. All accuracy values were significant compared

to permuted values. However, the PPV values were not impressive so the two groups were anyway considered to be poorly separated regardless of classifier or control group employed. It was noted that ARRB2 and ERK2 were consistently picked just as they were in the PTSD acute patients. Furthermore, Gs and IL-6 were also consistently picked. In the 25 gene comparison, RPART again performed worse than the other classifiers.

Comparison	Classifier	Genes selected to differentiate groups	PPV	NPV	Accuracy (permuted)
All controls vs trauma without PTSD (25 genes)	SVM/varselrf	ARRB2, CREB1, DPP4, ERK1, ERK2, GR, Gs, IL-6, MAPK8, MKP1	63	84	79 (72)
	SLR	ARRB2, CD8 beta, CREB1, ERK2, IL-6, MAPK8, MAPK14, MR, RGS2, SERT	63	85	80 (73)
	RPART	ARRB2, CD8 beta, CREB2, DPP4, ERK2, Gi2, Gs, MKP1	46	80	72 (63)
Controls vs trauma without PTSD (29 genes)	SVM/varselrf	ARRB2, CREB1, DPP4, ERK1, ERK2, Gs, IL-6, IL-8, MAPK8, MKP1, MR, PBR, PREP, SERT	59	79	74 (65)
	SLR	ADA, ARRB2, CD8 beta, CREB1, ERK2, Gs, IL-6, MAPK14, MKP1, MR, RGS2, VMAT2, IL-1 beta	59	80	73 (64)
	RPART	ARRB2, ERK2, Gs, IL-6, IL-8, MKP1, PREP, SERT	63	82	76 (58)

Table 31: Summary of controls vs patients with trauma but without PTSD. PPV, NPV 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). The analyses were done in R.

7.5.2 Multiple group comparisons

Since our focus was on 2-group comparisons, I did not do many multiple (>2) group comparisons (even though there were several possible combinations). Thus, below I present two comparisons. It should be remembered that PPV and NPV were not reported for these multiple group comparisons, and that only the classifiers SVM with varselrf and RPART were applied.

In the first comparison, three groups were compared; all controls (SH ABS, DC, UK and PTSD controls) vs. BPD patients vs. PTSD acute patients. 25 genes were used in this comparison. The result is summed up in table 33. The accuracy values were high and highly significant compared to the permuted values (data not shown). It was noted that ERK1, ERK2, GR and MKP1 were picked by both variable selection methods. ERK2 was a key gene picked in the 2-group comparison with the same pooled control group vs. PTSD acute patients, while GR was a key gene picked in the similar 2-group comparison

with BPD patients. MKP1 appeared in the former 2-group comparison as well (see table 29), while ERK1 appeared in both comparisons (see table 28 and 29). It was also noted that the RPART selected genes DDP4 and MAPK8 were not part of any 2-group comparison with the large pooled control group; neither compared with the BPD patients nor compared with the PTSD acute patients.

Classifier	Genes selected to differentiate groups	Accuracy (permuted)
SVM/varselrf	ARRB2, ERK1, ERK2, GR, MAPK14, MKP1	80 (22)
RPART	DPP4, ERK1, ERK2, GR, MAPK8, MKP1, MR	78 (22)

Table 33: Summary of the three group comparison including 25 genes; all controls (SH ABS, DC, UK and PTSD controls) vs. BPD patients vs. PTSD acute patients. Permuted accuracy values are in parenthesis in the last column. All accuracy values are in percentages. Both accuracy values were significant compared to the permuted values (data not shown). The analyses were done in R.

In the second multiple group comparison, four groups were compared; PTSD controls vs. PTSD acute patients vs. PTSD in remission vs. patients with trauma without PTSD. This comparison comprised 35 genes (see table 7, chapter 4). The result is summed up in table 32. Here, it can be seen that only RPART is able to separate the four groups and that only with a relative poor accuracy. The reason for the latter is that 2-group comparisons showed no separation between controls and remitted patients. There is some overlap between the selected genes now and genes selected by RPART in various 2-group comparisons described in the previous subsection. Besides two new genes in the list (EGR2 and MMP9), the overlap is not perfect as it must be remembered that pooled control groups were utilized in the 2-group comparisons.

Classifier	Genes selected to differentiate groups	Accuracy (permuted)
SVM/varselrf	-	36 (25)
RPART	CD8 beta, CREB1, EGR2, IL-6, IL-8, MAPK14, MKP1, MMP9, MR, ODC1, P2X7, PBR, PREP, RGS2	36 (26)

Table 32: Summary of the four group comparison including 35 genes; PTSD controls vs. PTSD acute patients vs. PTSD in remission vs. patients with trauma without PTSD. Permuted accuracy values are in parenthesis in the last column. All accuracy values are in percentages. Only the RPART actual accuracy was significant compared to the permuted values (p-value: 0.001204). The analyses were done in R.

7.5.3 Genes and clinical variables separating ABS controls from BPD patients

Here follows a short section on the use of SLR (stepwise logistic regression) with clinical variables as well as gene expressions for a comparison of the same groups.

Clinical data for patients was only available for 11 BPD patients (cohort 1). As reported in the classification chapter, SLR performed well on clinical variables. For the ABS control group, consisting of 299 subjects, clinical information and gene expression data were available, and utilized in e.g. section 7.3.1 and 7.3.4. In this subsection, a way is shown to link clinical variables with gene expression profiles using a classifier approach.

The 299 ABS controls and 11 BPD patients were separated using SLR, both on the gene and the clinical level, see table 34.

ABS vs. BPD with SLR		LOOCV
Genes selected to differentiate groups	ADA, CD8 beta, DPP4, ERK2, GR, MKP1, MR, RGS2, VMAT2	97,7%
Clinical variables selected to differentiate groups	Recent stress score sum, Coping score, 7 symptom score sum, Family Alcohol Abuse, Tobacco score, Lifetime drug use, Lifetime experiences	97,4%

Table 34: Summary of genes and clinical variables selected by SLR to differentiate between the ABS control group and 11 BPD patients. The Leave-One-Out Cross-Validation accuracy is reported in the last column. The analyses are done in R.

Table 34 shows that whether the control and patient group is separated by the gene expressions alone or only by the selected clinical variables, an almost perfect separation of the two groups is possible. Thus, it may be hypothesized that the two sets of variables (genes and clinical variables) are closely connected, and that either of them may be used to reveal information of the other.

Finally, the genes and clinical variables may be combined so that they both are independent variables, while the class label (controls vs patients) is considered the dependent variable. SLR then selects ERK1, GR, SERT, 7 symptom score sum and recent stress score sum as the differentiating variables with a LOOCV accuracy of 98,1%. In this case, this accuracy is very close to the previous reported ones. It is noted that ERK1 and SERT are selected, none of which are listed in table 34.

7.6 Heat maps and clustering

After having identified genes separating groups in the previous section 7.5 or via univariate tests, these genes may be used in clustering as described in the statistics chapter. This was done in the very last part of the thesis work, so only a few preliminary heat maps were constructed. One of them is shown in the statistics chapter, section 5.7, in figure 10 using four genes that separated the controls from the BPD patients. These four genes were identified in the previous chapter, section 7.5.1.

In figure 38, I present another heat map generated using 30 randomly chosen subjects from a pooled control group (DC, SH ABS and PTSD controls) together with 30 randomly chosen PTSD acute patients. I did not choose more subjects for visual reasons. Only genes that separated the two groups from one another were used. These genes (ARRB2, ERK2, RGS2) were also identified in the previous subsection 7.5.1 on 2-group comparisons. In the heat map the expression profiles of two distinct PTSD acute patients clusters are identified.

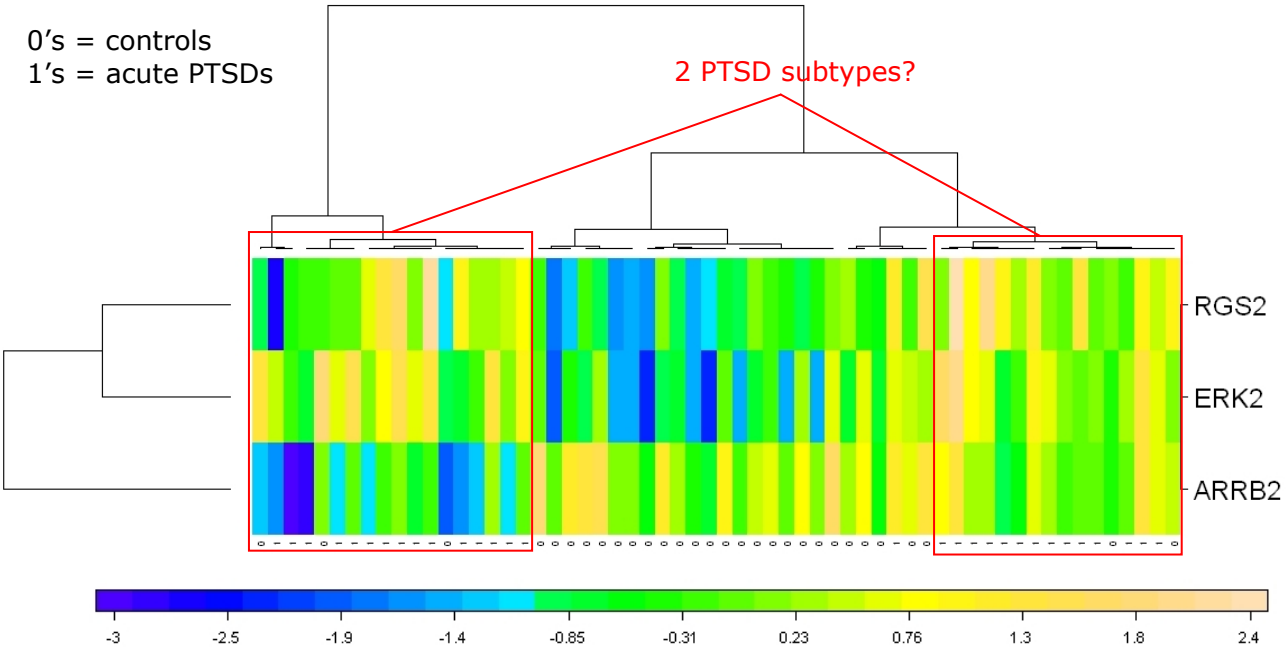


Figure 38 shows a heat map with 30 subjects randomly chosen from a pooled control group (DC, SH ABS and PSTD controls) and 30 randomly chosen PTSD acute patients clustered at the top of the heat map. At the bottom, I have included the class labels showing two different clusters of PTSD acute patients. These clusters may represent different phenotypes. Genes separating the three control and patient groups are shown to the right. The heat map was done in R.

In figure 39, a third heat map is shown, generated on the basis of data for 20 subjects from the same pooled control group as before, 20 borderline patients and 20 acute PTSD patients. Each group of 20 subjects was randomly chosen from the respective group. Only genes (ERK1, ERK2, GR, MKP1) that separated the three groups from one another were used. These genes were identified in the previous subsection 7.5.2 on multiple group comparisons. In the heat map the expression profiles of two distinct BPD and PTSD acute patients clusters are identified.

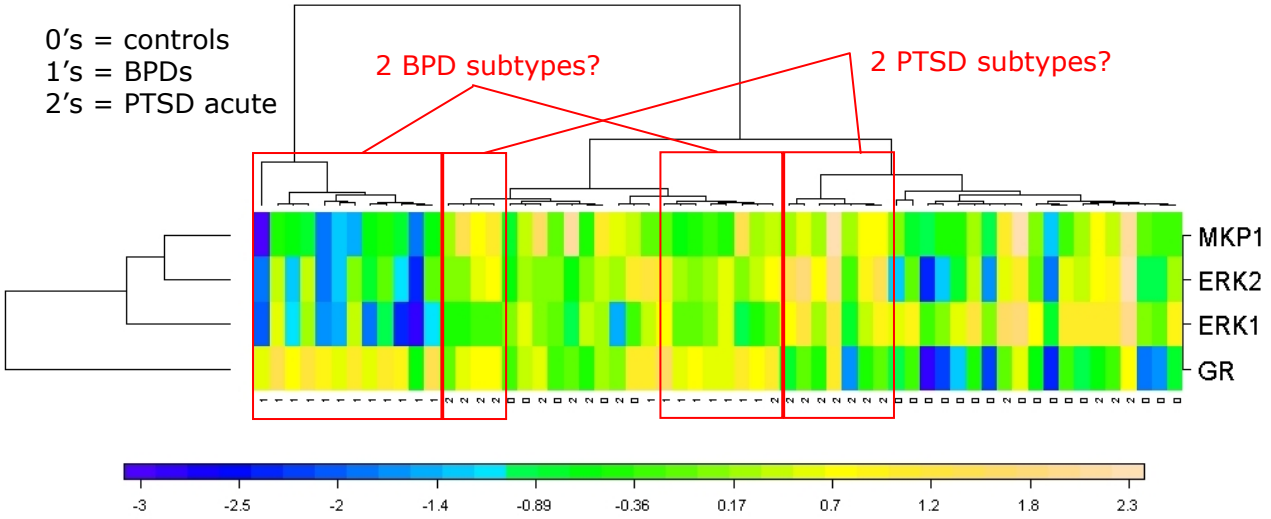


Figure 39 shows a heat map with 20 subjects from a pooled control group (DC, SH ABS and PSTD controls), 20 borderline disorder patients and 20 acute PTSD patients clustered at the top of the heat map. At the bottom, I have included the class labels showing two different clusters of BPD and PTSD patients. These clusters may represent different phenotypes. Genes separating the three control and patient groups are shown to the right. The heat map was done in R.

Due to lack of clinical data for patients[24], the principle of identifying gene expression phenotypes and linking them to clinical information is demonstrated with the SH ABS controls. Here, focus is on two distinct control subgroups within the SH ABS control group in the heat map in figure 40. 20 subjects were randomly chosen among the SH ABS controls, the BPD patients and the PTSD acute patients. Genes separating the groups are listed to the right in the figure.

Next, using the classifier and variable selection method SLR, I looked into the clinical variables selected in chapter 4 for the two subgroups of SH ABS controls[25]. SLR was able to separate the two subgroups using the clinical variables Age, BMI, Early life stress score and Anxiety score, however only with a LOOCV (Leave-One-Out Cross-Validation) accuracy of 72%. This relatively low accuracy is not surprising given that the SH ABS control group is very homogenous. Nonetheless, this result indicates that a combination of unsupervised gene clustering and supervised clinical variable classification may yield a link between distinct gene expression profiles and related distinct clinical profiles, and thus be used to identify patient subtypes or control intermediate phenotypes.

[24] Only clinical data for cohort 1, that is, 11 BPD patients is available which is not sufficient for subtyping purposes with the heat map clustering approach.
[25] SLR was used in a related task in section 7.5.3.

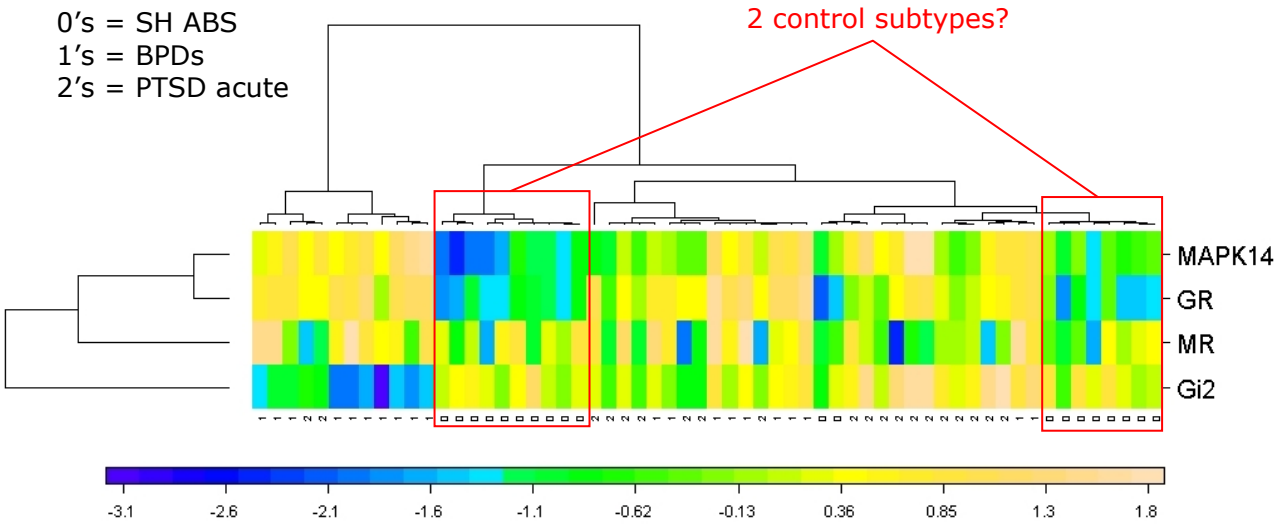


Figure 40 shows a heat map with 20 subjects from the SH ABS control group, 20 borderline disorder patients and 20 acute PTSD patients clustered at the top of the heat map. At the bottom, I have included the class labels showing two different clusters of SH ABS control subgroups. These clusters may represent different phenotypes as investigated with SLR using the SH ABS clinical information in the text. Genes separating the three control and patient groups are shown to the right. The heat map was done in R.

8. Conclusion and discussion

This chapter makes conclusions based on some of the most interesting results obtained in the previous chapters and discusses interesting aspects of the use of gene expressions in peripheral whole blood as a 'psychiatric' probe of a mental condition, see table 35. Let us first note that our results support the use of gene expressions in whole blood as biomarkers for affective disorders. Validation of the results is needed, however. In particular, I link the genes that look promising in separating various groups with biological knowledge related to molecular and cellular function, physiological system development and function, and various diseases and disorders. In this way it becomes possible to see which processes (e.g. cell death, cell signaling or hematological system development and function) are perturbed and which other diseases (like cancer and cardiovascular disease) to be aware of. Here, the separating genes are also related to the bioinformatic predictions in section 7.1.1. On this background, it is concluded that since every gene set separating the various groups is also connected to some of the most promising new biomarkers, it strengthens the arguments for including the additional biomarkers PP2A, Hsp90 and NFkB in future trials.
Based on the classifier simulation study in chapter 6 and the classification results in section 7.5, the most promising classification and variable selection methods seem to be support vector machines (SVMs) combined with variable selection based on random forests (varselrf) and stepwise logistic regression (both for 2-group comparisons). For multiple group comparisons, SVM/varselrf is recommended together with recursive partitioning. Classifier aspects are discussed in section 8.2. Both classifiers and variable selection method are used by the US Lundbeck group on other data sets.
Statistical methods are discussed in section 8.3 and based on the results. Some of the suggested methods are

- *Spearman correlations for pair wise gene expression comparisons.*
- *Repeated measures ANOVA for identifying differences between multiple time points measurements.*
- *For gene expression disease subtyping: hierarchical clustering and heat maps.*
- *Canonical correlation analysis for gene expression-clinical variable relationships supplemented by the univariate tests.*

Validation of the results obtained with the statistical methods is in general required.

Further conclusions and discussions are included concerning

- *pooling of controls (section 8.4) with expression profiles of various control groups being more similar to each other than to expression profiles of patient groups*

- *intermediate phenotype results (section 8.5) with some of the 20 depression hypotheses from section 7.3.1 being confirmed in a small group of severely depressed patients (Lundbeck confidential data)*
- *possible disease phenotypes (section 8.6) on the expression level that may be identified with heat maps and canonical correlation analysis. Clinical data from patients are needed to establish a connection between the expression subtypes and possible disease subtypes.*

Finally, bioinformatic predictions (section 8.7), gender differences (section 8.8), and time point measurements (section 8.9) are concluded on and discussed.

All in all, the study has yielded several results and pointed to different aspects that deserve to be combined, concluded upon and discussed. I have summed up the most interesting topics, arranged in order of perceived importance, in table 35 which is based upon many of the results of the previous chapters.

Topic	**Short description**	**Section**
Whole blood biomarkers for psychiatric diseases	Conclusion on and discussion of genes identified in various comparisons in section 7.5 are reported. The overlapping genes are 1) compared to the genes listed in chapter 2 2) related to the bioinformatics findings in chapter 3 dealing with the 29 Lundbeck genes 3) related to the bioinformatics predictions in section 7.1.1	8.1
Classifiers and variable selection methods	Conclusion and discussion of the applied classification algorithms and variable selection methods. Promising classifiers / variable selection methods are listed. Furthermore, the simulation study in classification chapter is discussed.	8.2
Statistical methods	Conclusion and discussion of the applied statistical methods. Promising methods are listed.	8.3
Pooling of controls	Conclusion and discussion of the results on pooling the control groups.	8.4
Intermediate phenotypes (clinical variable – gene expression relationships)	Conclusion and discussion of results concerning the 20 depression hypotheses in section 7.3.1.	8.5
Phenotypes / disease subtypes	Conclusion and discussion of results related to phenotype identification, section 7.3.3, 7.3.2 and 7.6.	8.6

Bioinformatics predictions	Conclusion and discussion of the use of bioinformatics for prediction of new possible biomarkers is discussed. Also, bioinformatics is discussed as a prediction tool in the context of a yet un-analyzed patient group (bipolar disorder patients).	8.7
Gender differences	Conclusion and discussion of the gender differences results.	8.8
Temporal measurements of gene expressions	Conclusion and discussion of the 3-time point measurement results.	8.9

Table 35 presenting the various topics that are related, concluded upon and discussed in this chapter.

8.1 Whole blood biomarkers for psychiatric diseases

Here conclusions are drawn with respect to the genes separating the various groups in section 7.5 of the previous chapter and overlapping between the classifiers[26] for consistency. These genes are compared partly to the genes reported for each affective disorder in chapter 2, partly related to the bioinformatics predictions in section 7.1.1, and at the end of this section partly related in table 36 to the bioinformatics findings in chapter 3 dealing with the 29 Lundbeck genes.

In section 7.5, comparisons between different control and patient group were made, and several genes were found to be significantly and differentially expressed between groups. For controls versus BPD patients, the overlapping genes between the classifiers, and that separated the groups were[26]: Gi2, GR, MAPK14 and partly MR (25 gene comparison). Only GR is mentioned (indirectly) in the chapter two section 'Borderline Personality Disorder and genes' in connection with the glucocorticoid neurotransmission. Besides GR, Gi2, MAPK14 and MR may be involved in the pathology of borderline personality disorder and may be novel whole blood biomarkers for BPD.
In relation to the new biomarker predictions of section 7.1.1 (here especially considering PP2A, Hsp90, NFkB and MHC Class I), GR and MR are interacting with Hsp90 on the protein level, while GR is also interacting with NFkB.

For controls versus PTSD acute patients, the separating and overlapping genes were[26]: ARRB2, ERK2 and RGS2. None of these genes are reported in the chapter two section 'PTSD and genes' and, hence, they may be involved in the pathology of post-traumatic stress disorder, and also be novel whole blood biomarkers for PTSD.
In relation to the new biomarker predictions of section 7.1.1 (again especially considering PP2A, Hsp90, NFkB and MHC Class I), ARRB2 is interacting with Hsp90 on the protein level, while ERK2 is interacting with PP2A.

[26] 2-group RPART results are disregarded as explaining in the section 8.2.

No significant genes were found to separate the controls from the remitted PTSD subjects. Thus, remitted PTSD subjects and controls have a similar gene expression profile. This result is in good agreement with the clinical diagnosis, and indicates that the gene expressions are 'normalized' upon remission.

For the final 2-group comparison, comparing the controls with the trauma patients without PTSD, the significant and overlapping genes[26] were: ARRB2, CREB1, ERK2, IL-6 and partly MAPK8 (25 gene comparison), Gs, MKP1 and MR (the latter three genes appeared in the 29 gene comparison). It should be remembered here, that the PPV values were not great, making us wonder about the validity of the results. Nonetheless, all classifiers yielded significant results. Trauma without PTSD has not been investigated in chapter two, since this group of subjects was only included for comparison reasons. OMIM has no record of trauma (without PTSD), so my informal guess would be that these eight genes may be (novel) whole blood biomarkers for trauma without PTSD, and be involved in the pathology of the disorder. Since ARRB2 and ERK2 were consistently picked as they were in PTSD acute patients, it could be hypothesized that the trauma subjects may be at risk for developing PTSD.
In relation to the new biomarker predictions of section 7.1.1 (again considering PP2A, Hsp90, NFkB and MHC Class I), ARRB2 and MR are interacting with Hsp90 on the protein level, ERK2 is interacting with PP2A, and IL-6 with NFkB.

For the multiple group comparison of controls versus BPD patients versus PTSD acute patients, the separating and overlapping genes were: ERK1, ERK2, GR and MKP1. As described above, GR has been implicated in BPD, and GR has also been implicated in PTSD (see chapter two 'PTSD and genes'). The other three genes, ERK1, ERK2 and MKP1, have not previously been associated with any of these two disorders, and may thus, besides GR, be involved in the peripheral pathology of two disorders. Furthermore, they may present new whole blood biomarkers appropriate when a comparison is made between the two disorders and controls, if relevant.
In relation to the new biomarker predictions of section 7.1.1 (again considering PP2A, Hsp90, NFkB and MHC Class I), ERK2 is interacting with PP2A on the protein level, and GR interacts with both Hsp90 and NFkB.

The last multiple group comparison (all PTSD groups versus each other) is not described here, since only one classifier yielded a significant result. This makes me doubt the value of the result, also because we have seen above that there is no difference between controls and remitted subjects.

In the various group comparisons above, every gene set separating the various groups is also connected to some of the most promising new biomarkers of section 7.1.1. This strengthens the arguments for including the additional biomarkers, PP2A, Hsp90 and NFkB, in future trials.

In order to obtain hypotheses of perturbed biological functions, I have looked closer at the separating genes mentioned above and related them to the biological functions from Ingenuity described in chapter 3 and listed fully in appendix 2. I have looked for biological functions consisting of all genes from a separating comparison. All the biological functions also contained additional genes. In table 36, I have summed up the most relevant results and arranged the biological functions after molecular and cellular functions, physiological system development and functions, and diseases and disorders (see the table).

Biological functions / Disease associated genes in	**BPD**	**BPD+1**	**acute PTSD**	**Trauma**	**Trauma+1**	**BPD and acute PTSD**
Molecular and Cellular Functions						
Cell Death				X	X	X
Cell Signalling			X			
Gene Expression				X	X	X
Molecular Transport		X	X			
Physiological System Development and Functions						
Hematological System Development and Function	X		X	X	X	X
Immune Response	X					
Diseases and Disorders						
Cancer		X		X	X	X
Cardiovascular Disease	X					X
Inflammatory Disease	X					

Table 36 lists biological functions according to appendix 2 where sets of disease associated genes participate. BPD=Gi2, GR and MAPK14; BPD+1= Gi2, GR, MAPK14 and MR; acute PTSD= ARRB2, ERK2 and RGS2; trauma=ARRB2, CREB1, ERK2 and IL-6; trauma+1=ARRB2, CREB1, ERK2, IL-6 and MAPK8; BPD and acute PTSD=ERK1, ERK2, GR and MKP1. The trauma+3 (ARRB2, CREB1, ERK2, IL-6, Gs, MKP1 and MR) is not shown for visual reasons and only appeared in 'Gene Expression'.

From table 36, it can be seen that under the category 'Molecular and Cellular Functions' different gene sets are involved in the biological functions of cell death, cellular development, cell signaling, gene expression and molecular transport.
Cell death, comprising functions associated with cellular death and survival, seems to be a relevant biological function for the trauma without PTSD associated genes (see legend to table 36) in whole blood and for the BPD and acute PTSD genes. In chapter three, the neurogenesis hypothesis was mentioned in connection with depression, so it is interesting but perhaps not that surprising to see cell death mentioned here. Cell signaling, comprising functions that are involved in intracellular signaling pathways, seems to play a role in whole blood for the genes associated with acute PTSD (again, see table 36 legend). The genes associated with trauma, and BPD and acute PTSD, are also part of the gene expression function, which confirms their basic role.

Finally, molecular transport comprising functions associated with the intra- and extracellular movement of molecules, is associated with BPD and PTSD (acute). It may not be surprising that this function may be impaired in these disorders.
Under the category 'Physiological System Development and Function', 'hematological system development and function and immune response' is listed. The former biological function includes functions associated with the normal development and function of blood, and thus makes good sense for all the disorders being measured in whole blood. The immune response biological function seems particularly activated in BPD compared to the other disorders in whole blood.
In the final category in table 36, diseases and disorders include cancer, cardiovascular and inflammatory diseases. There is an overlap between cancer genes and genes associated with all the disorders. This is not surprising (anymore) as I mentioned in chapter three that cancer might cause gene expression changes in some of the selected genes, meaning it could be important to check subjects for cancer prior to their inclusion in a clinical trial. Cardiovascular disease genes overlap with BPD associated genes and genes associated with BPD and PTSD. It could be interesting to investigate whether BPD or PTSD patients have a higher risk of cardiovascular diseases (comorbidity?) than other e.g. trauma subjects without PTSD.
Finally, there is an overlap between genes involved in inflammatory diseases and BPD. Subjects in BPD trials should therefore be screened for inflammatory diseases like damage of spleen or leucopenia as mentioned in chapter three.

Of all the group comparisons, only the gene set involved in separating controls from BPD and PTSD acute patients (ERK1, ERK2, GR and MKP1) is part of a significant biological pathway (listed in appendix 3). These four genes are part of the glucocorticoid receptor signaling pathway.

All in all, even though the amount of patient and control qPCR whole blood data analyzed in this thesis was only enough to generate hypotheses, the gene expression profiles, as described above, convincingly seem to reflect the pathology of the studied affective disorders. This seems to support the use of gene expressions in peripheral blood as biomarkers of affective disorders, showing a correlation between brain and blood (5). Several articles mentioned in chapter 1 indicated this in other blood studies for psychiatric disorders (6), (7), (8), (9), (10), (11), (also supported and reviewed in (215)) and the argument in favor of using such blood biomarkers is underpinned by the current findings, although the genes identified have to be validated in independent trials. Only then can the validity of the possible novel biomarkers be established and be exploited as a neural probe of the studied psychiatric disorders.

8.2 Classifiers and variable selection methods

The simulation study in the classification chapter identified the classifiers and variable selection methods SVM combined with varselrf (variable selection based on random forests), RPART and SLR as the best choices for analyzing the qPCR data in the two group comparison case, and in the multiple group comparison case SVM/varselrf and RPART. The simulation study did not include PPV and NPV values that were reported for the classification results in section 7.5 of the previous chapter. In almost all the 2-group comparisons of section 7.5.1, the PPV and accuracy values for the RPART results were consistently lower than the ones for two other classifiers. We (the US Lundbeck group and I) were not impressed by the bad RPART PPV results, so we decided to skip RPART in the 2-group case.
In the multiple group comparisons, PPV and NPV values were not reported. With the few multiple group comparisons made, we did not have support to exclude any of the chosen classifiers.

Thus, the most promising classifiers and variable selection methods for analyzing the Lundbeck qPCR data are SVM combined with varselrf and SLR in the 2-group comparison case, and SVM combined with varselrf and RPART in the multiple group comparison case.

For the classification results in the previous chapter the same parameters were used as in the simulation study in chapter 6. A next step can be to tune the parameters of the chosen classifiers and variable selection methods (various options exist, see (184), (208), (168) and (158)) to see if a better performance can be obtained.

Another possible way to improve the performance in the 2-group case of SLR is to include categorical clinical variables together with the gene expressions. This was demonstrated in the last part of subsection 7.5.3, although not convincingly in that case. It is not unreasonable to imagine that in other cases the combination of gene expressions and clinical variables may improve the classification performance considerably compared to using either set of variables as independent variables.

The simulation study
Some issues concerning the simulation study in the classification chapter are raised here. The basis for the mathematical approach to the gene-gene interactions was that we did not know the exact biological interactions between genes on the expression level. We investigated several classical mathematical approaches to gene interactions in the form of different linear and nonlinear tasks. Had we chosen to pursue other (mathematical) approaches, other classifiers and variable selection methods might have been recommended. Furthermore, other classifiers and variable selection techniques might have

been explored, like regularized discriminant analysis, classification using generalized partial least squares, neural networks, or other advanced methods. I mainly chose ones that I had been introduced to in the CBS course 'DNA Microarray Analysis' or read about in various articles. Also, the classifiers had to perform relatively fast in R and be able to perform some simple tasks from phase 1, all of which I tested. Having said that, had other classifiers and variable selection methods been included, I might have made different recommendations that the ones, I did.
In relation to the choice of classifier, a different strategy could have been to identify a single classifier that was able to handle a broad range of classification tasks. However, the US Lundbeck group and I decided to choose more than one classifier partly because we could not identify a single classifier that performed well in all classification tasks and partly because we pursued a strategy of having different classifiers, so that overlapping genes (between the classifiers) could be identified for consistency reasons. This was of interest for the US Lundbeck group. It should be stressed, however, that overlapping genes alone, can not be used to build a good classifier. All genes identified by a chosen classifier / variable selection method should be used.

Another issue deals with the use of the accuracy measure. Having a single accuracy measure surely was not very informative as the group sizes alone could be responsible for a large part or most of the actual accuracy value. The permutation tests, albeit a time consuming step, improved the benefit of the accuracy measure. The accuracy measure was chosen because multiple group (>2) comparisons were wanted. Had this not been the case, and had we only focused on two-group comparisons, the Matthews correlation coefficients would have been a better choice. It "*is generally regarded as a balanced measure which can be used even if the classes are of very different sizes*", and is generally regarded as being one of the best performance measures (216). Finally, it should be said that for the two-group comparisons, we supplemented the accuracy values with the PPV and NPV values. These latter values made us aware of the predictive values, which were considered informative measures for the classification performance.

Also, 10-fold stratified cross-validation was applied as recommend in (199). Other validation schemes, like LOOCV or holdout validation, might have resulted in different performance values, and hence different recommendations.

A final issue concerning the simulation study has to do with the stability of the classifiers (178). This encompasses removing one or more observations and comparing the classification result before and after this exercise. It could also include adding an incorrect observation and noticing the impact on the classification result (accuracy, PPV and NPV values) as well as on the selected variable list. I considered these general and additional tasks to be beyond the

scope and timeframe of the simulation study. Nonetheless, they are interesting aspects that might be worth looking into.

8.3 Statistical methods

The various applied statistical methods are discussed, and the promising methods for future analyses of Lundbeck qPCR data listed.

It is always a good idea to check the expression data for normality. As it was seen in section 7.2 of the previous chapter, applying the recommend (from the statistical department of Lundbeck) logarithm to the expression data made, in general, more gene expressions follow the normal distribution. Other data transformations might have been considered like the square root or the reciprocal transformation before testing for normality. Still, more than half of the gene expressions could not be considered normally distributed according to any of the 5 applied normality tests. Even though the parametric univariate tests were robust and thus allowed all but severe deviations from normality, I believed it was a good idea to include nonparametric methods. It tuned out that in most cases, the univariate results were significant using both parametric and nonparametric tests, which made us believe (more in) the results.

The basic statistical methods, univariate tests and correlations, gave a good first impression of the data and were the basis of the 20 depression hypotheses of section 7.3.1. These methods are recommended for a first impression of data and establishing basic hypotheses about the data at hand. Furthermore, they are simple methods and in my experience, simple methods are less likely to overfit the data compared to more advanced methods.

For multiple time point measurements, repeated measures ANOVA is recommended to evaluate any statistical expression difference between the time points. If Lundbeck wanted to rule out circadian variation of any of the genes, many more time measurements would be needed; the circadian aspect is discussed in the final (Perspectives) chapter.

Even tough stepwise regression in some cases could explain ~15-20% of the variance in a gene expression by a linear combination of automatically chosen clinical variables, stepwise regression was not useful for most of the gene expressions. Also, considering the relative low explanation degree listed above, stepwise regression is not recommended for this type of data.

In general, clinical data for patients were missing (except for cohort 1 of the BPD patients). This made it difficult to evaluate several of the statistical methods for phenotype identification of patients. Recursive partitioning might

be an interesting phenotyping tool, but the RP results have to be validated and compared to clinical data. Considering that RP is prone to overfitting, and that RPART did not perform well as a classifier, RP is not recommend as a first choice of applicable exploratory statistical tools.
The value of canonical correlation analysis was also difficult to evaluate due to the lack of patient clinical data. Section 7.3.3 demonstrated that CCA might be an interesting phenotyping tool. However, again, there were too few patient data to firmly establish CCA as a promising tool. With more patient clinical data and linked gene expression data, CCA could reveal an interesting link between clinical and gene expression characteristics of subgroups of patients, and is then recommended as such.
The heat map results in section 7.6 also showed some very interesting possible disease phenotypes in simple visual plots. The hierarchical clustering used for the heat maps were based on the correlation as a distance measure. The Euclidian distance could also have been used, and would presumably result in other gene expression clusters. Again, the lack of patient clinical data did not make it feasible to link the disease phenotypes with clinical characteristics using e.g. SLR as demonstrated in that section. Still, heat maps are recommended for their promising phenotyping abilities (demonstrated several times in section 7.6) and simple visual layout.

Considering the last two methods, with few clinical data, CCA may be better to identify gene-clinical relationships compared to heat maps. With more clinical data, CCA supplemented with heat map clustering and classification of the clinical variables could yield quite interesting results.

8.4 Pooling of controls

Here, results from section 7.3.6 and 7.3.7 are summed up and discussed.

In section 7.3.6, we decided to pool the DC and SH ABS control groups. This was done based on the similarity of the (SH) DC and SH ABS clinical data, the large degree of agreement between the Spearman correlations of the two control groups, and because various classifiers recognized almost perfectly controls of one control group as belonging to the other control group when compared to either the BPD or acute PTSD patient groups. There were clearly expression differences between the two control groups; first 22 out of 25 genes was significantly different expressed by univariate tests but as we defined a SH DC group matching the clinical criteria of the SH ABS group, the list of significantly different expressed genes was down to 15 genes. Pelora and SLR were also able to separate the groups based on the gene expression profiles. I think this is a good example of biological variability between control groups from different countries and different trials. Solely based on the gene expression differences, we could have chosen to pursue another strategy; the US controls should only be compared to US patients and the Danish controls

should only be compared to Danish patients, etc. However, this approach is more prone to overfitting. Also, if Lundbeck wants a more universal classifier, this classifier should have a more universal ability to recognize controls which is another argument for pooling the control groups.

In section 7.3.7, the main question was whether all control groups should be pooled into one large control group or not. Again, there were univariate gene expression differences between the four control groups; SH ABS, DC, UK and PTSD controls. The classifier SVM/varselrf was also able to separate the four groups (data not shown), and thus confirmed the expression differences between them. However, comparisons between various (pooled) control groups and patients (see section 7.3.7) using Pelora and SLR, made us conclude that the different control groups were more alike than they were different. We did those analyses before the classifier simulation study. Given the limitations of Pelora, it could have been interesting to see the results of the control versus patient comparisons using e.g. SVM/varselrf.
At the end, the four control groups were pooled into one large control group that thus spanned the biological expression variability of the various control groups.

With clinical data for the PTSD controls and UK controls, we could have studied intermediate phenotypes (see next section) more carefully after pooling all the controls into one large control group. A possibility would be to define a super healthy control group based on the clinical data from the pooled control group. Then various intermediate phenotypes for BPD and PTSD patients could be defined like in section 7.3.1 for depressed patients.
With different clinically defined intermediate phenotypes, it would be very relevant to investigate the gene expression path from absolutely healthy controls over various intermediate phenotypes to various disease phenotypes (also clinically defined). Would the expression profiles along this path be completely random or would some kind of expression continuum be revealed? Questions like this could lead to more insight into the biological basis of affective disorders.

8.5 Intermediate phenotypes

In section 7.3.1, clinical variable - gene expression relationships were investigated in the ABS control group with respect to intermediate phenotypes.

Based on univariate tests of depression related clinical variables involving both individual genes and gene ratios, table 25 was constructed indicating gene expression patterns that could define intermediate phenotypes. The table was expanded in more detail by the 20 subsequent depression hypotheses. The first seven hypotheses involved individual genes, and could thus be compared to the expected gene regulation noted in table 3, chapter 3. For SERT

(hypothesis 1) and G alpha s (hypothesis 4), the down regulated predictions in table 25 are also expected in depressed patients from the literature (table 3). The DDP4 (hypothesis 2, up regulation expected) and MAPK14 (hypothesis 7, down regulation expected) predictions do not correspond with the table 3 literature predictions. The predicted expression directions of ERK2 (hypothesis 3) and MKP1 (hypothesis 5) are ambiguous (may present different intermediate phenotypes). In table 3, both genes are expected to be down regulated, which is also predicted in table 25 only considering decreased appetite. For PBR (hypothesis 6) there was no expected up or down regulation in table 3. Thus, there seems to be a certain overlap between the findings of section 7.3.1 and the expected regulations from the literature in table 3. It is not surprising the overlap is not greater mainly given that we can not be really sure that the intermediate phenotypes are actually at risk for developing depression and that we are comparing with trials conducted by other research groups with probably different inclusion criteria, different normalization and measurement methods, etc.

Section 7.3.1 ended with stating that it did seem like gene expression patterns could be used to segment control subjects based on various depression relevant clinical variables. The conclusions on the 20 depression hypotheses were;

- Different gene expression patterns were intermediate phenotype dependent.
- Expression differences in controls were related to symptoms of depression (but no current diagnosis of depression).
- Simple statistical modeling could identify factors that explained a clinically relevant degree of the variation in gene expression.
- The expression pattern observed in a small group of severely depressed patients partially confirmed the results from control subjects.
  (This was based on Lundbeck confidential data that could not be part of the thesis)

Overall, it thus seemed like our approach for identifying intermediate phenotypes did produce results that, at least partially, could be confirmed in severely depressed patients. The gene expression intermediate phenotypes would have to be validated in further large clinical trials.

It could have been interesting to do the same exercise with relation to the BPD and PTSD disorders, where I did have access to the patient data. Furthermore, in relation to PTSD it could have been interesting as well to compare the gene expression profiles of intermediate phenotypes to the expression profiles of remitted PTSD patients, and, hence, look more into how gene expressions are ‘normalized’ by treatment as the reported in section 8.1.

8.6 Phenotypes

In the sections 7.3.3, 7.3.2 and 7.6 different possible disease phenotypes were identified using various statistical approaches. In 7.3.3, regularized canonical correlation analysis could identify four possible BPD patient subgroups with each subgroup of patients correlated to certain clinical variables and 11 gene expressions, see the section for details. These results were compared to the CCA example in the statistics chapter using the same data set, however only utilizing 4 genes. The 4-gene and 11-gene cases were found to be quite similar in the sense that the subjects in two subgroups (BPD1/BPD2 and BPD4/BPD10) stayed closely together in both cases. This could indicate that these two subgroups were fairly robust. One subject in each of the other two subgroups switched group membership due to a different set of CCA chosen variables in each case.
With so few clinical data for patients, the CCA output is only at best circumstantial evidence of 2-4 BPD phenotypes. Still, the approach of applying several (here two) gene sets to the same data and looking for consistent patient subgroups/phenotypes, is recommended. More clinical data for patients are needed to validate these results and possibly generate additional phenotype hypotheses.

For the data set analyzed and results generated in section 7.3.2, recursive partitioning only subdivided the BPD patients into one major group consisting of 20 BPD patients (based on the expression values of GR and ARRB2 only) and one BPD patient outlier based on the expression values of GR, SERT, CREB2 and ADA (see figure 31). These results were generated when the BPD patients were compared to 20 DC and 21 SH ABS controls.
In the statistics chapter, RP was used to divide the 299 ABS controls and 11 BPD patients. In this case, RP identified two BPD subgroups consisting of more than one member; there was a subgroup of 5 BPD patients identified based on the expression values of GR and ERK2 only, and a subgroup consisting of two members based on the same gene expression adding CREB2 (figure 9). Again, due to the sparse clinical data for patients, it is difficult to validate these subgroup findings, and more both patient clinical and patient gene expression data is needed for this purpose. The RP results show that depending on which control and patient data you compare, different gene expression subgroups may emerge. If this is not just a RP overfitting issue, it indicates that the same patient group may be subdivided in various ways. To investigate this aspect, it would be of great assistance to have access to confirmed clinical phenotypes/subtypes. Having clinically defined phenotypes could be used to tune the statistical methods and validate the results.

Both in statistics chapter (figure 10) and in section 7.6, heat maps were used to distinguish expression differences between both BPD subgroups and acute PTSD subgroups (figure 38 and figure 39). The different expression subgroups / phenotypes are clearly seen on these plots. Furthermore, in the last part of

section 7.6, I demonstrated how SLR might be used to link the expression subgroups with clinical data for controls. Had sufficient patient clinical data been available, the same exercise might be done for patients, and disease putative disease phenotypes identified. Of course, they would have to be validated in further trials, and possible compared to clinically known phenotypes, if possible.
In general, the clusters identified in such an unsupervised heat map manner may thus give clues to possible phenotypes among patients (and intermediate phenotypes in subjects at risk). These expression phenotypes would then have to be verified using a supervised approach like SLR with the clinical variables for these patients. It would be interesting to see how or if SLR would be able to separate, for instance, the two BPD or PTSD subtypes in figure 38 only considering these patients' clinical variables. A further step would be to investigate whether these clinical variables and values would make sense to a psychiatrist.

8.7 Bioinformatics predictions

In section 7.1.1, bioinformatics was used to predict new possible biomarkers by looking into protein-protein interactions for protein complexes. New potential biomarkers were identified; primarily Hsp90, PP2A and NFkB (each interacting with at least three of the Lundbeck proteins), and secondary Ras, MHC Class I, Mek, Akt and Ap1 (interacting with maximum two of the Lundbeck selected proteins).

The US Lundbeck group had investigated several possible biomarkers prior to the list of 29 genes described in chapter 3. The above bioinformatics predictions were of great interest to the US Lundbeck people. They considered in particular to measure PP2A (dephosphorylates ERK), Hsp90, NFkB and MHC Class I in the blood samples. A natural first step would be to see, if these genes are sufficiently expressed in blood to be measured reliably.

Other sets of new potential biomarkers could have been obtained using other web applications than Ingenuity (like e.g. STRING or Inweb from CBS), because different PPI web applications use, for instance, different data sources, different quality of data, and protein data from different species. This might easily have resulted in other recommendations for new biomarkers.

Also, if wanted, bioinformatics could have yielded other kinds of predictions like which genes have a transcriptional influence on other genes, which proteins activate or inhibit other proteins, etc. Such bioinformatics predictions could be done prior to any experiments in whole blood to better understand these aspects for the biology and pathology of affective disorders.

Bioinformatics was finally used to predict which of 29 genes selected by Lundbeck that would be altered in the yet un-analyzed patient group consisting of bipolar disorder (BD) patients. The predictions were based on two SNP studies (WTCC (210) and Baum et al. (59)) and NCBI's SNP database. There were virtually no overlap between the 68 genes associated with these SNPs and the BD associated genes listed in chapter 2. With millions of SNPs present in the genome, overfitting is always an inherent issue, and thus, it is not surprising that there is so little overlap between the genes identified through various SNP studies and furthermore compared to e.g. gene expression studies. Nonetheless, three possible BD biomarkers related to 29 genes were identified via the PPI database STRING and via Ingenuity focusing on transcriptional interactions (see below); SYK, BDNF and THRB.
At the level of transcription, the three BD associated proteins seemed to affect: Gs, IL-1 beta, CREB1 and ERK1, and thus expression differences between BD patients and controls might be expected to be seen in the above four gene expressions. These predictions have to be validated in future BD trials.

Just like in the first bioinformatics task, the above predictions might have been different using other web applications than Ingenuity, STRING and NCBI's SNP database for the reasons given there.

Another way to do the bioinformatics BD predictions could have been to include all the BD disease related genes from chapter 2, and thus investigate how all these genes, together with the 68 genes mentioned above, interact with the 29 genes selected by Lundbeck on the protein level.

Finally, it should be mentioned that the same kind of bioinformatics exercise that was done above with bipolar disorder could also be done with the borderline personality disorder and PTSD using the BPD and PTSD related genes mentioned in chapter 2 (plus additional SNP studies, if available and desired). In these cases, it would be possible to compare the bioinformatics predictions with the available qPCR blood data results. This way, we could learn about the value of such bioinformatics predictions, optimize the bioinformatics process and be better suited for future tasks and experiments.

8.8 Gender differences

In section 7.3.5 gender differences were explored in control groups using Pelora, SLR, univariate tests and Spearman correlations. There was a large overlap between the genes selected by either Pelora or SLR comparing all DC controls to all SH ABS controls, all DC males to all SH ABS controls and all DC females to all SH ABS controls. Furthermore, Pelora and SLR were not able to separate DC males versus DC females or control (DC, SH ABS and PTSD controls) males versus control (same combined control group) females with a satisfactory result. Spearman correlations between the former three

comparisons were also more than 96% in agreement. Univariate tests showed 6 genes were significantly different expressed between the DC group split by gender and compared to the SH ABS. However, since there was only a very small overlap between the genes (only MR) selected by the classifiers and the 6 significant genes, it was concluded that although some expression differences existed between the genders, they did not seem major.

Pelora and SLR are linear classifiers. It is possible that using SVM in combination with varselrf would yield a more separable result comparing the DC males to the DC females or the male controls (DC, SH ABS and PTSD controls) to the female controls. This way, although the genders were not convincingly linearly separable, they could perhaps be convincingly (accuracy > 90% and significant compared to a permuted sample) separated with a nonlinear classifier. This would have to be investigated, and if the case, the succeeding analyses should possibly be done three ways (males only, females only, combined). However, given that all the patient data was either mostly only from women (BPD patients) or solely from men (acute PTSD), a proper gender investigation into the biology of affective disorder could not be carried out. More gender specific clinical data could have shed light on this important aspect as it is known that more women than men are affected by BPD and PTSD (see chapter 2). Thus, the influence of gender on patient groups in whole blood gene expression measurement remains to be determined.

8.9 Temporal measurements of gene expressions

In section 7.4, expression differences between three time points – Day 0 at 8 am, Day 0 at 2 pm and Day 1 at 8 am - were investigated in the UK controls using repeated measured ANOVA. Five gene expression were found to differ significantly between the time points; CD8 beta, IL-8, MKP1, MR and ODC1. For IL-8 the significant difference was between the Day 0 and Day 1 measurements, however the difference was caused by only five subjects, not the entire group. Removing these (IL-8 is a proinflammatory cytokine prone to noisy behavior), the significant difference disappeared. For the other four genes, the difference was between the morning and afternoon measurements. According to Pubmed, only MR was found to display a circadian pattern in man, however not in blood. Lundbeck wished to exclude genes displaying circadian patterns from the gene list. However, due to the few time point measurements, circadian effects could be neither be confirmed nor ruled out for the four genes. The conclusion was that even though the five genes were differently expressed between the three time points, they remained in the list of genes measured in the study.

In my view, the time point measurements above highlight, nonetheless, the importance of the temporal aspect. Since expression difference do exist for some genes, it is another argument for pooling the control groups (across

various time measurements) in order to span the biological gene expression variability in controls.
Ideally, the significant different time points for the five genes should be used as the basis for a test data set in a classification task separating controls from the patient groups. No matter the time of measurement, a control should still be classified as a control, and not as a patient simply due to the time of day the blood was sampled. This remains to be determined. Other possibilities, if time of day does matter for some gene expression, are either to exclude the genes from the study or try to incorporate the time aspect into clinical trials, if possible.

Also, if possible, it is recommended to investigate any possible circadian effects for the five genes properly in order to be certain about whether to include or exclude the gene expressions. This aspect is included in the next chapter.

9. Perspectives

This final chapter presents different suggestions for further work. Some of the more interesting perspectives are

- *a suggestion for an experiment to confirm or rule out temporal gene expression oscillations. Large oscillations for a gene expression might mean that the gene expression is not suitable as a biomarker.*
  Several of the gene expressions in the present study can potentially display circadian behavior as witnessed by the repeated measure ANOVA results and Ingenuity.
  I propose a simple 24-hour experiment with blood sampled every hour for a small number of depressed patients and healthy subjects. Not only should the 29 gene expressions be measured but I also propose to measure a limited number of so-called core clock components, as the literature points to a circadian component of mood disorder.
- *requirements for constructing a Bayesian gene regulatory network. With Bayesian networks, it is often possible to extract more information out of the genes selected in the various classification tasks. Actually, the ultimate goal is to construct a causal Bayesian network in order to predict gene regulatory behavior in whole blood in various affective disorders. The potential of the Bayesian network approach was clearly demonstrated on the protein level by Sachs and colleagues in the US (217).*
  I spent a week at the University of Warwick, UK together with professor David Rand and David Wild investigating the possibility of constructing a Bayesian network with the available data and with the Sachs article in mind. The result became a wish to understand the mechanisms behind a successful Bayesian network model with the Lundbeck data, which should be obtained by making a simulated Bayesian network based on the available data.
- *suggestions for other classifier approaches. One suggestion involves the use of deterministic forests instead of random forests as the former produce a fixed set of genes every time the same classification tasks is executed. Another suggestion involves looking at classification probabilities, since they might be more informative when it comes to identify intermediate phenotypes. This section also contains additional classification oriented suggestions.*
- *other ways of searching for blood biomarkers in affective disorders. Here, I propose the use of either microarrays to perform a genome-wide expression analysis in a hypotheses-free manner or to have custom microarrays designed focusing on specific important genes or pathways in order to reduce the number of false positives.*
- *clustering simulations for disease subtyping. As a major future challenge involves disease subtyping, and only heat maps with hierarchical*

clustering have been investigated, I propose a simulation study to identify promising unsupervised clustering methods (as many different clustering techniques exist).

In this last chapter, I choose to discuss perspectives on five issues related to the thesis;

1. Temporal aspects of gene expression behavior. This includes a description of an experiment to confirm or rule out gene expression oscillations in blood.
   Furthermore, as an example of a systems biology approach, the dynamics of a minimal model involving the CREB proteins is included.

2. The requirements for constructing a Bayesian gene regulatory network only using the available non-temporal blood measurements.

3. Suggestions for other classifier approaches and classification tasks.

4. Other ways of searching for new blood biomarkers in affective disorders.

5. Unsupervised clustering simulations for subtyping purposes.

9.1 Temporal aspects of gene expression behavior

In section 7.4, four gene expressions were found to exhibit a significant difference between morning and afternoon measurements in whole blood; CD8 beta, MKP1, MR and ODC1. Furthermore, according to Ingenuity, four other genes of interest to the present study are involved in the circadian rhythm (not in man, but in mouse and rat); IL-6, CREB1, SERT and ERK2. These findings could indicate that some, perhaps at least eight of the 29 'Lundbeck' gene expressions, might exhibit a daily rhythm. This is difficult to confirm on the basis of the existing literature which is very restricted with respect to gene expression oscillations in whole blood in man.

On the other hand, the literature points to a circadian component of mood disorders (218), (219), (220), involving the core clock components[27]; CLOCK (circadian locomoter output cycles kaput), PER1, PER2, PER3 (Period, drosophila, homolog of, 1, 2 and 3 respectively), CRY1, CRY2 (cryptochrome 1 and 2 respectively), BMAL1 (aryl hydrocarbon receptor nuclear translocator-like), REV-ERBα (nuclear receptor subfamily 1, group D, member 1) and RORα (RAR-related orphan receptor alpha) (221). Moreover, between 2-10% of all genes are assumed to be transcribed in a circadian manner (222).

[27] "*Core clock components are defined as genes whose protein products are necessary for the generation and regulation of circadian rhythms within individual cells throughout the organism*" (221).

Taken together, this suggests that gene expression oscillations, both (part of) the 29 gene expressions and the core clock components, not only might occur in man, but be directly involved in the biological basis and pathology of affective disorders. In order to investigate whether circadian or ultradian[28] gene expression oscillations occur at all in whole blood and in affective disorders, I propose an experiment where a blood sample is obtained every hour for a 24-hour period. This sampling should be repeated two or three times to see if the results are consistent. Both the 29 gene expressions as well as the 9 core clock components could be measured, if the core clock components are expressed at detectable levels in whole blood and not show too much variability (see chapter 4). The obtained data should then be analyzed, starting by inspecting the multiple 24-hour plots for each gene expression.

The experiment could be initialized with just a few healthy controls and a few MDD patients, hospitalized. Gene expressions should be normalized the same way as done in the present Lundbeck study (with 7 HKG).

The reason for including the 9 core clock components in the above experiment is to examine how these clock components oscillate at the messenger RNA level in whole blood, if at all, in controls versus patients, and to relate their oscillations to oscillations occurring in any of the 29 gene expressions. This could shed light on the temporal dynamics (including amplitude and phase shift) of gene expression behavior in affective disorders and to the causal relations between the various gene expressions. This would add to the understanding of the molecular basis of affective disorders. Would these whole blood measurements indicate any disruptions in oscillatory behavior between a healthy and a disease state?
The experiment could also indicate the variation of each gene expression during 24 hours, which either a) in case of large expression oscillations might mean that the oscillating gene expression is not suitable as a biomarker, or b) incorporated into a classifier, might improve the diagnostic classification performance. Including additional subjects could verify the results.

Also, like in the Lundbeck study described in this thesis, it would be very informative to expand the above experiment with qPCR measurements from a few acute PSTD patients, some patients with trauma without PTSD, and a few BPD patients and remitted patients.

If any of the gene expressions do indicate circadian or ultradian patterns, this might lead to the construction of a mechanism-based model to get a basic understanding of the biological mechanisms behind the oscillations, and predict qualitative behavior in various areas of the parameter space. A first step would

[28] Recurrent cycles with a period of less than 20 hours.

be to perform further tissue specific investigations to identify important components that could be subject to mathematical modelling. Transitions with various parameter adjustments from the healthy state to the diseased state could then also be studied. A detailed description is beyond the scope of this thesis, but the interested reader is referred to the paper '*Dynamics of a minimal model of interlocked positive and negative feedback loops of transcriptional regulation by cAMP-response element binding proteins*' (223).

9.2 Bayesian gene regulatory networks

The background for considering Bayesian networks is that I would like to extract more information out of the genes selected in the various classification tasks described in section 7.5. As a first approximation of the interactions between such a set of selected genes, I looked into the Spearman correlations between them. An example is shown in table 37 comprising four genes (ERK1, ERK2, GR, MKP1) separating the controls from the BPD and acute PTSD patients.

Status01			ERK1	ERK2	GR	MKP1
Controls	ERK1	Correlation Coefficient	1.000	.687(**)	.426(**)	.355(**)
		Sig. (2-tailed)	.	.000	.000	.000
		N	196	196	196	196
	ERK2	Correlation Coefficient	.687(**)	1.000	.701(**)	.638(**)
		Sig. (2-tailed)	.000	.	.000	.000
		N	196	196	196	196
	GR	Correlation Coefficient	.426(**)	.701(**)	1.000	.718(**)
		Sig. (2-tailed)	.000	.000	.	.000
		N	196	196	196	196
	MKP1	Correlation Coefficient	.355(**)	.638(**)	.718(**)	1.000
		Sig. (2-tailed)	.000	.000	.000	.
		N	196	196	196	196
BPD	ERK1	Correlation Coefficient	1.000	.887(**)	.462(*)	.590(**)
		Sig. (2-tailed)	.	.000	.035	.005
		N	21	21	21	21
	ERK2	Correlation Coefficient	.887(**)	1.000	.401	.749(**)
		Sig. (2-tailed)	.000	.	.071	.000
		N	21	21	21	21
	GR	Correlation Coefficient	.462(*)	.401	1.000	.295
		Sig. (2-tailed)	.035	.071	.	.195
		N	21	21	21	21
	MKP1	Correlation Coefficient	.590(**)	.749(**)	.295	1.000
		Sig. (2-tailed)	.005	.000	.195	.
		N	21	21	21	21
Acute PTSD	ERK1	Correlation Coefficient	1.000	.709(**)	.208	.201
		Sig. (2-tailed)	.	.000	.156	.170
		N	48	48	48	48
	ERK2	Correlation Coefficient	.709(**)	1.000	.385(**)	.293(*)
		Sig. (2-tailed)	.000	.	.007	.043
		N	48	48	48	48
	GR	Correlation Coefficient	.208	.385(**)	1.000	.469(**)
		Sig. (2-tailed)	.156	.007	.	.001
		N	48	48	48	48
	MKP1	Correlation Coefficient	.201	.293(*)	.469(**)	1.000
		Sig. (2-tailed)	.170	.043	.001	.
		N	48	48	48	48

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Table 37: Spearman correlations between four genes (ERK1, ERK2, GR, MKP1) separating the controls from the BPD and the acute PTSD patients. Significant correlations (significantly different from zero) at the 1% level are encircled. The analyses were done in SPSS.

Table 37 indicates that, in general, the significance of the correlations depends on the group, and, thus, that different correlations exist among the genes depending on group. In theory this could add a little knowledge about which correlations that get strengthened or weakened comparing the various groups. Still, correlations are not very informative. On the other hand, Bayesian networks offer insight into a possible gene regulatory behavior.

Now follows a brief introduction to Bayesian networks (BN) including reasons for considering these to explore interactions between the differentiating (classifier chosen) genes;

- In a BN, biological knowledge from the literature (tissue) can be incorporated as prior knowledge and gene-gene interactions in blood (posterior knowledge of interest) inferred.

- With a BN, a gene regulatory network of genes involved in affective disorders can then be examined.

- Ideally, I would like to infer causal relationships, if any, between the genes. In a BN, this requires either time-series (that are not available) or intervention data (explained below).
  A BN with causal relationships may serve as a first step towards a dynamic model.

- BNs "*can represent complex stochastic nonlinear relationships among multiple interacting"* genes (217).

- BNs "*are robust in the face of both noisy data and imperfectly specified hypotheses*". "*They can handle missing values*", "*are not limited to pair-wise or linear interactions between genes*" (226).

- BNs "*permit latent variables to represent unobserved factors*" (226) and thus BNs can detect both direct and indirect causal connections.

- BNs can combine data from multiple sources, e.g. questionnaire data, gene expression measurements and publicly available databases.

- "*Variables in a* BN *can be discrete or continuous*" (226).

A more thorough introduction to Bayesian inference for gene expressions is given in e.g. (224) and for graphical models (covering BNs) in (225). Furthermore, an interesting article on the subject is found in (226).

BN example
In order to understand what it requires to construct a BN that may be used to examine the differentiating genes in the Lundbeck data, I now give an example from the literature of the use of BNs in a relevant biological context – this is from the Science paper '*Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data'* by Sachs et al. (217). BN had here "*elucidated most of the traditionally reported signaling relationships and predicted novel interpathway network causalities"* (217) using only static (non-temporal) data.
In figure 41, a "*classic signaling network and points of intervention*" (217) are shown.

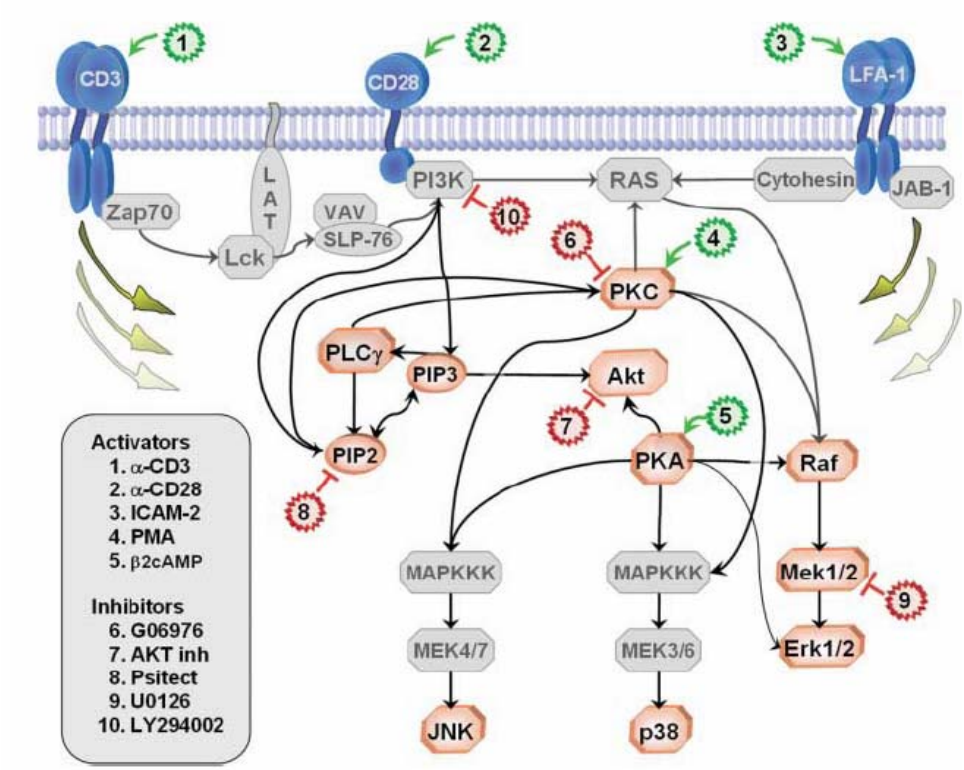
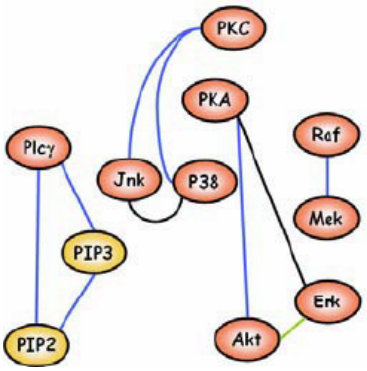


Figure 41 showing "*a classical signaling network and points of intervention*". "*Signaling nodes in color were measured directly*" by flow cytometry. "*Signaling nodes in gray were not measured, but were included to place the signaling nodes within contextual cellular pathways*". "*Arcs are used to illustrate connections between signaling molecules; in some cases, the connections may be indirect*". For more details, see (217).

A key point is the intervention data. In figure 42, observation-only (that is without any interventions) data was used to reconstruct the figure 41 signaling network.



	A. Lacking Intervention data
Expected	7/10
Reported	1/10
Reversed	N/A
Unexplained	2
Missed	11

Figure 42 shows inference results from observation data only. No arcs are recovered and the plot resembles a correlation network. This demonstrates that "*intervention data is crucial for effective interference*". For more details, see (217).

Observation data alone is clearly not enough to infer the underlying biological network. When the intervention data is included, almost the entire signaling network is recovered, see figure 43.

Fig. 3. Bayesian network inference results. (A) Network inferred from flow cytometry data represents expected outcomes. This network represents a model average from 500 high-scoring results. High-confidence arcs, appearing in at least 85% of the networks, are shown. For clarity, the names of the molecules are used to represent the measured phosphorylation sites

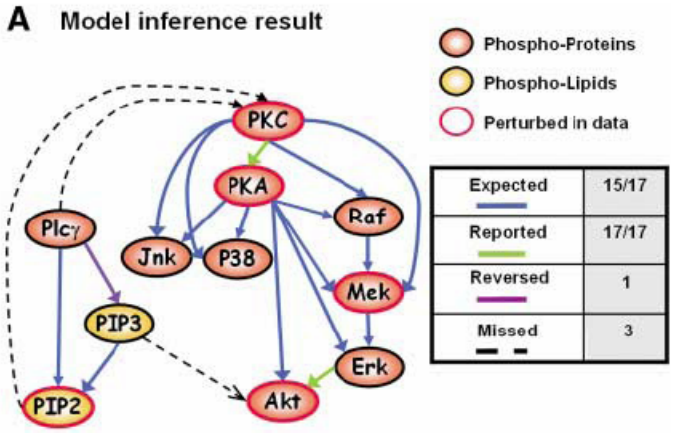


Figure 43 shows inference results using observation as well as intervention data. Almost the entire original network is recovered and novel interpathway arcs predicted, that Sachs et al. verify experimentally. For more details, see (217).

Static BNs are constrained to be acyclic, so any feedback loops can not be detected. This might be the reason for the missing arcs in figure 43.

Also, it is worth noticing that

- Causal influence could be inferred even though a protein was not perturbed (e.g. Raf -> Mek)
- Most arcs were validated by literature studies (no a priori knowledge was used – only random restarts were utilized).
- Arcs represented both direct and indirect causal connections, that is, in some cases the influence of one protein on another was mediated by a

protein that were not measured.

The main differences between the Sachs' approach and the Lundbeck data concerning possible inference of gene regulatory networks seemed to be;

1. In the Lundbeck data, we are dealing with real-world conditions / non-ideal interventions unlike in the Sachs' article, where intervention data is created with specific inhibitors and activators used to create perturbations.

2. Clinical data set sizes are used with Lundbeck unlike in the Sachs' article where thousands of data points are collected in each experiment.

3. Measurements are done on whole blood in the Lundbeck study that is, averaged over different cell types, unlike in the Sachs' article using good single-cell measurements.

On the positive side, the Sachs's article did point to the possibility of inferring gene regulatory networks based on incomplete measurements (not all variables were or could be measured) and static data only. A major question was then whether the different control and patient groups could be used as pseudo-intervention data.

I spent a week at the University of Warwick, UK together with professor David Rand and David Wild looking into the possibility of using BNs to learn more about gene regulatory behavior in whole blood with the knowledge of the Sachs's approach. Two suggestions were made to investigate whether the BN framework was likely to be successful for this task:

1. Basically, modify the Sachs' data to match the available Lundbeck data (as much as possible) and see how much network can be inferred.

2. Make a simulated BN model e.g. with 4 nodes matching the 4 genes separating the controls vs. BPD vs. acute PTSD patients, and then
    a. obtain distributions for each node sampled from the real gene expression data
    b. direct arcs (causality)
    c. simulate from the model and find out which perturbations give what information.

In Warwick, I only managed to do some parts of suggestion 1 above and the more restrictions I put on the Sachs data (to match the Lundbeck data) the less gene regulatory network could get inferred (of course). However, two things were worth noticing: a) it was possible to infer part of a network, still and perhaps this small network could tell something about a pathway (or some

gene interactions) connected to the pathology of affective disorders. Of course it would have to be verified experimentally but the BN approach could give a hint about what to look for. b) The Sachs article dealt with protein measurements which was another obstacle in that protein expressions do not behave the same way as gene expressions. This meant that the Sachs BN model might not be suitable as a framework for deducting a regulatory network for the Lundbeck data.

This leaves suggestion 2 - making a simulated BN model which I believe would be the best thing to do. However, as this was at the end of the thesis work, I did not have time to pursue this interesting endeavor. With a simulated BN model, one would be able to predict gene regulatory behavior in whole blood in various affective disorders. The simulated results would have to be experimentally verified, and could add to the understanding of the biological basis of affective disorders.

9.3 Other classifiers and classification tasks

Having worked with the different classifiers and classification tasks in chapter 6 and observed the results in section 7.5, four perspectives come to my mind.

First of all, random forests performed very well in classification tasks in chapter 6 and showed, in general, excellent performance in section 7.5. However, random forests do not select a fixed set of genes, but the number of selected genes may vary (slightly) from each execution of the script. In order to obtain a fixed set of genes, deterministic forests might be a solution (227). They operate in a manner similar to random forests, have the same high performance, however, yield the same gene list every time for the same task. Deterministic forests are, however, not part of any R package at the present time, and would have to be either coded for optimal performance to the Lundbeck task or contact established with a possible software provider.

A second aspect concerns the inclusion of clinical variables into classification tasks. This was briefly discussed in section 8.2 concerning SLR. As the path from micro (gene expressions) to macro (clinical descriptions) is quite long and complex, it is likely that the combination of clinical variables and gene expressions might yield a better classification performance that either set of variables. It could be quite interesting to identify a range of classifiers (including Bayesian classifiers) able to handle both set of variables and perform a simulation study to identify the most promising classifiers. These results should then be compared with the classification results obtained using only gene expressions. This way we could have an idea as to how much improvement in classification performance adding clinical variables could contribute with, if any.

A third perspective worth looking into deals with a combination of the search for intermediate phenotypes and classification probabilities for classification tasks involving gene expressions. So far, we have only focused on the discrete and typically binary classification outcome – e.g. control or patient. By considering classification probabilities, it could become easier to identify intermediate phenotypes. This could easily become relevant for subjects with a classification probability of ~40-60%. These individuals should have their clinical information checked to see how they differ clinically from the rest of their group, if at all.

Finally, it should be mentioned that although treatment data was not available, the classifiers, identified in chapter 6, could be used to predict treatment response.

9.4 Searching for blood biomarkers in affective disorders

Lundbeck has collected blood samples from the different control and patient groups described in chapter 4. Half of the blood for each subject has been used to measure the gene expressions by qPCR. This leaves 2.5 ml of whole blood per subject to be used for other analyses. Below I briefly propose to use microarrays for identifying other putative blood biomarkers based on the remaining blood samples.

While the Lundbeck study focused on a predetermined set of genes, with DNA microarrays Lundbeck could perform a genome-wide gene expression analysis in a hypotheses-free manner. Microarrays were briefly described in chapter 4. Microarrays have been used in number of mental disorders, see for instance, the review articles (228), (229) and (230) that also lists various benefits and drawbacks of this technology, see table 9 (chapter 4) for an overview.

The DNA microarray technology could expand the number of possible blood biomarkers drastically compared with the original 29 gene expression biomarker list used by Lundbeck. Differentially expressed genes would need to be validated by qPCR. Due to the large ratio of measured gene expressions to the number of subjects, a large number of false positives (that partly can be diminished with multiple testing correction procedures) and overfitting are an inherent part of the analysis of microarray data. The latter is definitely also due to disease heterogeneity. Bioinformatics, e.g. with Ingenuity, would play a greater role than in the present study (chapter 3, section 7.1 and section 8.1). With many significant and differentially expressed genes, bioinformatics is crucial in the identification of the significant genes' involvement in known metabolic or signaling pathways, in elucidating which other genes in a pathway that are significant, their function, etc.
When it comes to statistical analyses and classification of microarray data, a wide range of the methods used in chapters 5 and 6 may be applied again,

since they are designed for microarray analysis purposes, e.g. random forests, Pelora, regularized CCA and heat maps besides the universal applicable univariate tests.

As mentioned above, a large number of false positives are still to be expected with a genome-wide DNA microarray. Another possibility, that could reduce the number of false positives and overfitting, is to have custom arrays designed. Such custom-made arrays could monitor expression changes in specific pathophysiologically important genes or pathways. A disadvantage of custom arrays is that one is no more performing a hypotheses-free search for biomarkers but has introduced a bias in the selection of genes. Still, the custom array approach have been used with promising results in the assessment of human stress and depression in blood leukocytes, actually using only 2.5 ml of blood (94).

9.5 Unsupervised clustering simulations

A major future challenge is disease subtyping, and in section 8.6 I have suggested various approaches for this. The single most promising subtyping approach so far seems to be hierarchical clustering and heat maps. However, no systematic investigation into unsupervised clustering methods has been performed in this thesis. Many clustering algorithms exist with some of the most popular, besides hierarchical clustering, being "*K-means, PAM* (partioning around medoids)*, SOM* (self-organizing maps)*, mixture model-based clustering and tight clustering*" (231). In order to determine which clustering method that is best suitable for the Lundbeck data, a simulation study could be undertaken that could include the clustering methods above as well as AP clustering mentioned in chapter 5.

It is beyond the scope of this thesis to go into details as how to evaluate different clustering algorithms. Some authors, see e.g. (231), evaluate clustering methods for microarray data by performing both a simulation study and looking at real data sets, like the approach taken in chapter 6.
A popular similarity measure of two clusters is the weighted Rand index (231), but other measures exist as well (232). In the later reference, several suggestions, including simulated data sets, are given as how to perform cluster validation. Here it is stressed "*that entirely objective cluster validation is possibly only on data with well-defined cluster structures*", which is why the evaluation of clustering algorithms should always include such data – simulated data as well as real data with well-defined structures, perhaps like the ones in the heat maps of figures 38 and 39.

References

(1) Depressionspiller virker kun på hver femte. www.politiken.dk 2005 Nov 23.

(2) Berton O, Nestler EJ. New approaches to antidepressant drug discovery: beyond monoamines. Nat Rev Neurosci 2006 Feb;7(2):137-51.

(3) Antonijevic I, Artymyshyn R, Forray C, Rabacchi S, Smith K, Swanson C, et al. Perspectives for an Integrated Biomarker Approach to Drug Discovery and Development. 2008.

(4) Insel TR, Scolnick EM. Cure therapeutics and strategic prevention: raising the bar for mental health research. Mol Psychiatry 2006 Jan;11(1):11-7.

(5) Sullivan PF, Fan C, Perou CM. Evaluating the comparability of gene expression in blood and brain. Am J Med Genet B Neuropsychiatr Genet 2006 Apr 5;141(3):261-8.

(6) Segman RH, Shefi N, Goltser-Dubner T, Friedman N, Kaminski N, Shalev AY. Peripheral blood mononuclear cell gene expression profiles identify emergent post-traumatic stress disorder among trauma survivors. Mol Psychiatry 2005 May;10(5):500-13, 425.

(7) Zieker J, Zieker D, Jatzko A, Dietzsch J, Nieselt K, Schmitt A, et al. Differential gene expression in peripheral blood of patients suffering from post-traumatic stress disorder. Mol Psychiatry 2007 Feb;12(2):116-8.

(8) Le-Niculescu H, Kurian SM, Yehyawi N, Dike C, Patel SD, Edenberg HJ, et al. Identifying blood biomarkers for mood disorders using convergent functional genomics. Mol Psychiatry 2008 Feb 26.

(9) Middleton FA, Pato CN, Gentile KL, McGann L, Brown AM, Trauzzi M, et al. 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. Am J Med Genet B Neuropsychiatr Genet 2005 Jul 5;136(1):12-25.

(10) Iga J, Ueno S, Yamauchi K, Motoki I, Tayoshi S, Ohta K, et al. Serotonin transporter mRNA expression in peripheral leukocytes of patients with major depression before and after treatment with paroxetine. Neurosci Lett 2005 Nov 25;389(1):12-6.

(11) Iga J, Ueno S, Yamauchi K, Numata S, Kinouchi S, Tayoshi-Shibuya S, et al. Altered HDAC5 and CREB mRNA expressions in the peripheral leukocytes of major depression. Prog Neuropsychopharmacol Biol Psychiatry 2007 Apr 13;31(3):628-32.

(12) 2008 March 3. http://en.wikipedia.org/wiki/DSM-IV

(13) 2008 March 3. http://www.nimh.nih.gov/health/publications/depression/what-is-a-depressive-disorder.shtml

(14) 2008 March 3. http://www.emedicine.com/med/topic532.htm

(15) 2008 March 3. http://www.nimh.nih.gov/health/publications/depression/causes-of-depression.shtml

(16) Rosenzweig MR, Breedlove SM, Watson NV. Biological Psychology - An Introduction to Behavioral and Cognitive Neuroscience. fourth ed. Sunderland, MA, USA: Sinauer Associates, Inc.; 2005.

(17) Tafet GE, Bernardini R. Psychoneuroendocrinological links between chronic stress and depression. Prog Neuropsychopharmacol Biol Psychiatry 2003 Sep;27(6):893-903.

(18) Dunn AJ, Swiergiel AH, de BR. Cytokines as mediators of depression: what can we learn from animal studies? Neurosci Biobehav Rev 2005;29(4-5):891-909.

(19) 2008 March 10. http://en.wikipedia.org/wiki/Cytokine

(20) Irwin MR, Miller AH. Depressive disorders and immunity: 20 years of progress and discovery. Brain Behav Immun 2007 May;21(4):374-83.

(21) Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, et al. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 2003 Jul 18;301(5631):386-9.

(22) 2008 March 3. http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=608516

(23) Kendler KS, Kuhn JW, Vittum J, Prescott CA, Riley B. The interaction of stressful life events and a serotonin transporter polymorphism in the prediction of episodes of major depression: a replication. Arch Gen Psychiatry 2005 May;62(5):529-35.

(24) Antonijevic IA. Depressive disorders -- is it time to endorse different pathophysiologies? Psychoneuroendocrinology 2006 Jan;31(1):1-15.

(25) Mizrahi C, Stojanovic A, Urbina M, Carreira I, Lima L. Differential cAMP levels and serotonin effects in blood peripheral mononuclear cells and lymphocytes from major depression patients. Int Immunopharmacol 2004 Aug;4(8):1125-33.

(26) Taylor C, Fricker AD, Devi LA, Gomes I. Mechanisms of action of antidepressants: from neurotransmitter systems to signaling pathways. Cell Signal 2005 May;17(5):549-57.

(27) Akil H. Stressed and depressed. Nat Med 2005 Feb;11(2):116-8.

(28) Kaestner F, Hettich M, Peters M, Sibrowski W, Hetzel G, Ponath G, et al. Different activation patterns of proinflammatory cytokines in melancholic and non-melancholic major depression are associated with HPA axis activity. J Affect Disord 2005 Aug;87(2-3):305-11.

(29) Dwivedi Y, Rizavi HS, Roberts RC, Conley RC, Tamminga CA, Pandey GN. Reduced activation and expression of ERK1/2 MAP kinase in the post-mortem brain of depressed suicide subjects. J Neurochem 2001 May;77(3):916-28.

(30) Lucae S, Salyakina D, Barden N, Harvey M, Gagne B, Labbe M, et al. P2RX7, a gene coding for a purinergic ligand-gated ion channel, is associated with major depressive disorder. Hum Mol Genet 2006 Aug 15;15(16):2438-45.

(31) 2008 March 4. http://www.nimh.nih.gov/health/publications/borderline-personality-disorder.shtml

(32) 2008 March 4. http://en.wikipedia.org/wiki/Borderline_personality_disorder

(33) Goodman M, New A, Siever L. Trauma, genes, and the neurobiology of personality disorders. Ann N Y Acad Sci 2004 Dec;1032:104-16.

(34) Lis E, Greenfield B, Henry M, Guile JM, Dougherty G. Neuroimaging and genetics of borderline personality disorder: a review. J Psychiatry Neurosci 2007 May;32(3):162-73.

(35) Grosjean B, Tsai GE. NMDA neurotransmission as a critical mediator of borderline personality disorder. J Psychiatry Neurosci 2007 Mar;32(2):103-15.

(36) Kahl KG, Bens S, Ziegler K, Rudolf S, Dibbelt L, Kordon A, et al. Cortisol, the cortisol-dehydroepiandrosterone ratio, and pro-inflammatory cytokines in patients with current major depressive disorder comorbid with borderline personality disorder. Biol Psychiatry 2006 Apr 1;59(7):667-71.

(37) 2008 March 5. http://www.nimh.nih.gov/health/publications/anxiety-disorders/post-traumatic-stress-disorder.shtml

(38) 2008 March 5. http://www.nimh.nih.gov/health/publications/post-traumatic-stress-disorder-research-fact-sheet.shtml

(39) 2008 March 5. http://www.emedicine.com/med/topic1900.htm

(40) 2008 March 5. http://en.wikipedia.org/wiki/PTSD

(41) Nemeroff CB, Bremner JD, Foa EB, Mayberg HS, North CS, Stein MB. Posttraumatic stress disorder: a state-of-the-science review. J Psychiatr Res 2006 Feb;40(1):1-21.

(42) Young RM, Lawford BR, Noble EP, Kann B, Wilkie A, Ritchie T, et al. Harmful drinking in military veterans with post-traumatic stress disorder: association with the D2 dopamine receptor A1 allele. Alcohol Alcohol 2002 Sep;37(5):451-6.

(43) Segman RH, Cooper-Kazaz R, Macciardi F, Goltser T, Halfon Y, Dobroborski T, et al. Association between the dopamine transporter gene and posttraumatic stress disorder. Mol Psychiatry 2002;7(8):903-7.

(44) Koenen KC. Genetics of posttraumatic stress disorder: Review and recommendations for future studies. J Trauma Stress 2007 Oct;20(5):737-50.

(45) Broekman BF, Olff M, Boer F. The genetic background to PTSD. Neurosci Biobehav Rev 2007;31(3):348-62.

(46) Radant A, Tsuang D, Peskind ER, McFall M, Raskind W. Biological markers and diagnostic accuracy in the genetics of posttraumatic stress disorder. Psychiatry Res 2001 Jul 24;102(3):203-15.

(47) Tucker P, Ruwe WD, Masters B, Parker DE, Hossain A, Trautman RP, et al. Neuroimmune and cortisol changes in selective serotonin reuptake inhibitor and placebo treatment of chronic posttraumatic stress disorder. Biol Psychiatry 2004 Jul 15;56(2):121-8.

(48) Rohleder N, Joksimovic L, Wolf JM, Kirschbaum C. Hypocortisolism and increased glucocorticoid sensitivity of pro-Inflammatory cytokine production in Bosnian war refugees with posttraumatic stress disorder. Biol Psychiatry 2004 Apr 1;55(7):745-51.

(49) Sutherland AG, Alexander DA, Hutchison JD. Disturbance of pro-inflammatory cytokines in post-traumatic psychopathology. Cytokine 2003 Dec 7;24(5):219-25.

(50) 2008 March 6. http://en.wikipedia.org/wiki/Bipolar_disorder

(51) 2008 March 6. http://www.nimh.nih.gov/health/publications/bipolar-disorder/introduction.shtml

(52) 2008 March 6. http://www.emedicine.com/med/topic229.htm

(53) 2008 March 6. http://www.ncbi.nlm.nih.gov/entrez/dispomim.cgi?id=125480

(54) Smoller JW, Gardner-Schuster E. Genetics of bipolar disorder. Curr Psychiatry Rep 2007 Dec;9(6):504-11.

(55) De L, V, Tharmalingam S, Kennedy JL. Association study between the corticotropin-releasing hormone receptor 2 gene and suicidality in bipolar disorder. Eur Psychiatry 2007 Jul;22(5):282-7.

(56) Spijker A, Hoencamp E, van Rossum E, de Rijk R, Haffmans J, Blom M. Several polymorphisms of the glucocorticoid receptor gene (NR3C1) and their associations with bipolar disorder. Journal of Affective Disorders [Vol.107 Part.Supplement 1], S74-S75. 2008.
Ref Type: Abstract

(57) Kim YK, Jung HG, Myint AM, Kim H, Park SH. Imbalance between pro-inflammatory and anti-inflammatory cytokines in bipolar disorder. J Affect Disord 2007 Dec;104(1-3):91-5.

(58) Sklar P, Smoller JW, Fan J, Ferreira MA, Perlis RH, Chambert K, et al. Whole-genome association study of bipolar disorder. Mol Psychiatry 2008 Mar 4.

(59) Baum AE, Akula N, Cabanero M, Cardona I, Corona W, Klemens B, et al. A genome-wide association study implicates diacylglycerol kinase eta (DGKH) and several other genes in the etiology of bipolar disorder. Mol Psychiatry 2008 Feb;13(2):197-207.

(60) Kato T, Kakiuchi C, Iwamoto K. Comprehensive gene expression analysis in bipolar disorder. Can J Psychiatry 2007 Dec;52(12):763-71.

(61) Kato T. Molecular genetics of bipolar disorder and depression. Psychiatry Clin Neurosci 2007 Feb;61(1):3-19.

(62) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007 Jun 7;447(7145):661-78.

(63) Hasler G, Drevets WC, Manji HK, Charney DS. Discovering endophenotypes for major depression. Neuropsychopharmacology 2004 Oct;29(10):1765-81.

(64) MacQueen GM, Hajek T, Alda M. The phenotypes of bipolar disorder: relevance for genetic investigations. Mol Psychiatry 2005 Sep;10(9):811-26.

(65) Bearden CE, Freimer NB. Endophenotypes for psychiatric disorders: ready for primetime? Trends Genet 2006 Jun;22(6):306-13.

(66) Caspi A, Moffitt TE. Gene-environment interactions in psychiatry: joining forces with neuroscience. Nat Rev Neurosci 2006 Jul;7(7):583-90.

(67) Van GS, Van BC. Genetics of personality: are we making progress? Mol Psychiatry 2003 Oct;8(10):840-52.

(68) Joober R, Sengupta S, Boksa P. Genetics of developmental psychiatric disorders: pathways to discovery. J Psychiatry Neurosci 2005 Sep;30(5):349-54.

(69) 2008 March 10. http://www.cnsforum.com/imagebank/item/HPA_DPN_DPN_3/default.aspx

(70) Schiepers OJ, Wichers MC, Maes M. Cytokines and major depression. Prog Neuropsychopharmacol Biol Psychiatry 2005 Feb;29(2):201-17.

(71) Keller PA, McCluskey A, Morgan J, O'connor SM. The role of the HPA axis in psychiatric disorders and CRF antagonists as potential treatments. Arch Pharm (Weinheim) 2006 Jul;339(7):346-55.

(72) Raison CL, Miller AH. When not enough is too much: the role of insufficient glucocorticoid signaling in the pathophysiology of stress-related disorders. Am J Psychiatry 2003 Sep;160(9):1554-65.

(73) Stam R. PTSD and stress sensitisation: a tale of brain and body Part 1: human studies. Neurosci Biobehav Rev 2007;31(4):530-57.

(74) Woodson JC, Minor TR, Job RF. Inhibition of adenosine deaminase by erythro-9-(2-hydroxy-3-nonyl)adenine (EHNA) mimics the effect of inescapable shock on escape learning in rats. Behav Neurosci 1998 Apr;112(2):399-409.

(75) Elgun S, Keskinege A, Kumbasar H. Dipeptidyl peptidase IV and adenosine deaminase activity. Decrease in depression. Psychoneuroendocrinology 1999 Nov;24(8):823-32.

(76) Matuzany-Ruban A, Avissar S, Schreiber G. Dynamics of beta-arrestin1 protein and mRNA levels elevation by antidepressants in mononuclear leukocytes of patients with depression. J Affect Disord 2005 Nov;88(3):307-12.

(77) Avissar S, Matuzany-Ruban A, Tzukert K, Schreiber G. Beta-arrestin-1 levels: reduced in leukocytes of patients with depression and elevated by antidepressants in rat brain. Am J Psychiatry 2004 Nov;161(11):2066-72.

(78) Zubenko GS, Hughes HB, III, Stiffler JS, Brechbiel A, Zubenko WN, Maher BS, et al. Sequence variations in CREB1 cosegregate with depressive disorders in women. Mol Psychiatry 2003 Jun;8(6):611-8.

(79) Carlezon WA, Jr., Duman RS, Nestler EJ. The many faces of CREB. Trends Neurosci 2005 Aug;28(8):436-45.

(80) Blendy JA. The role of CREB in depression and antidepressant treatment. Biol Psychiatry 2006 Jun 15;59(12):1144-50.

(81) Maes M, Goossens F, Lin A, De M, I, Van GA, Scharpe S. Effects of psychological stress on serum prolyl endopeptidase and dipeptidyl peptidase IV activity in humans: higher serum prolyl endopeptidase activity is related to stress-induced anxiety. Psychoneuroendocrinology 1998 Jul;23(5):485-95.

(82) Garcia-Sevilla JA, Walzer C, Busquets X, Escriba PV, Balant L, Guimon J. Density of guanine nucleotide-binding proteins in platelets of patients with major depression: increased abundance of the G alpha i2 subunit and down-regulation by antidepressant drug treatment. Biol Psychiatry 1997 Oct 15;42(8):704-12.

(83) Avissar S, Nechamkin Y, Roitman G, Schreiber G. Dynamics of ECT normalization of low G protein function and immunoreactivity in mononuclear leukocytes of patients with major depression. Am J Psychiatry 1998 May;155(5):666-71.

(84) Young LT, Li PP, Kamble A, Siu KP, Warsh JJ. Mononuclear leukocyte levels of G proteins in depressed patients with bipolar disorder or major depressive disorder. Am J Psychiatry 1994 Apr;151(4):594-6.

(85) de Kloet ER, Reul JM, Sutanto W. Corticosteroids and the brain. J Steroid Biochem Mol Biol 1990 Nov 20;37(3):387-94.

(86) Garoflos E, Panagiotaropoulos T, Pondiki S, Stamatakis A, Philippidis E, Stylianopoulou F. Cellular mechanisms underlying the effects of an early experience on cognitive abilities and affective states. Ann Gen Psychiatry 2005 Apr 6;4(1):8.

(87) Robertson DA, Beattie JE, Reid IC, Balfour DJ. Regulation of corticosteroid receptors in the rat brain: the role of serotonin and stress. Eur J Neurosci 2005 Mar;21(6):1511-20.

(88) Cai W, Khaoustov VI, Xie Q, Pan T, Le W, Yoffe B. Interferon-alpha-induced modulation of glucocorticoid and serotonin receptors as a mechanism of depression. J Hepatol 2005 Jun;42(6):880-7.

(89) Wichers MC, Koek GH, Robaeys G, Verkerk R, Scharpe S, Maes M. IDO and interferon-alpha-induced depressive symptoms: a shift in hypothesis from tryptophan depletion to neurotoxicity. Mol Psychiatry 2005 Jun;10(6):538-44.

(90) Thomas AJ, Davis S, Morris C, Jackson E, Harrison R, O'Brien JT. Increase in interleukin-1beta in late-life depression. Am J Psychiatry 2005 Jan;162(1):175-7.

(91) Alesci S, Martinez PE, Kelkar S, Ilias I, Ronsaville DS, Listwak SJ, et al. Major depression is associated with significant diurnal elevations in plasma interleukin-6 levels, a shift of its circadian rhythm, and loss of physiological complexity in its secretion: clinical implications. J Clin Endocrinol Metab 2005 May;90(5):2522-30.

(92) Suarez EC, Krishnan RR, Lewis JG. The relation of severity of depressive symptoms to monocyte-associated proinflammatory cytokines and chemokines in apparently healthy men. Psychosom Med 2003 May;65(3):362-8.

(93) Sanchez MM, Alagbe O, Felger JC, Zhang J, Graff AE, Grand AP, et al. Activated p38 MAPK is associated with decreased CSF 5-HIAA and increased maternal rejection during infancy in rhesus monkeys. Mol Psychiatry 2007 Oct;12(10):895-7.

(94) Ohmori T, Morita K, Saito T, Ohta M, Ueno S, Rokutan K. Assessment of human stress and depression by DNA microarray analysis. J Med Invest 2005 Nov;52 Suppl:266-71.

(95) Kodama M, Russell DS, Duman RS. Electroconvulsive seizures increase the expression of MAP kinase phosphatases in limbic regions of rat brain. Neuropsychopharmacology 2005 Feb;30(2):360-71.

(96) Lister MF, Sharkey J, Sawatzky DA, Hodgkiss JP, Davidson DJ, Rossi AG, et al. The role of the purinergic P2X7 receptor in inflammation. J Inflamm (Lond) 2007;4:5.

(97) Galiegue S, Tinel N, Casellas P. The peripheral benzodiazepine receptor: a promising therapeutic drug target. Curr Med Chem 2003 Aug;10(16):1563-72.

(98) Maes M, Lin AH, Bonaccorso S, Goossens F, Van GA, Pioli R, et al. Higher serum prolyl endopeptidase activity in patients with post-traumatic stress disorder. J Affect Disord 1999 Apr;53(1):27-34.

(99) Oliveira-dos-Santos AJ, Matsumoto G, Snow BE, Bai D, Houston FP, Whishaw IQ, et al. Regulation of T cell activation, anxiety, and male aggression by RGS2. Proc Natl Acad Sci U S A 2000 Oct 24;97(22):12272-7.

(100) Moratz C, Harrison K, Kehrl JH. Regulation of chemokine-induced lymphocyte migration by RGS proteins. Methods Enzymol 2004;389:15-32.

(101) Svenningsson P, Chergui K, Rachleff I, Flajolet M, Zhang X, El YM, et al. Alterations in 5-HT1B receptor function by p11 in depression-like states. Science 2006 Jan 6;311(5757):77-80.

(102) Zubieta JK, Huguelet P, Ohl LE, Koeppe RA, Kilbourn MR, Carr JM, et al. High vesicular monoamine transporter binding in asymptomatic bipolar I disorder: sex differences and cognitive correlates. Am J Psychiatry 2000 Oct;157(10):1619-28.

(103) Zucker M, Aviv A, Shelef A, Weizman A, Rehavi M. Elevated platelet vesicular monoamine transporter density in untreated patients diagnosed with major depression. Psychiatry Res 2002 Nov 15;112(3):251-6.

(104) 2008 March 17. www.ingenuity.com

(105) Miller MA, Rahe RH. Life changes scaling for the 1990s. J Psychosom Res 1997 Sep;43(3):279-92.

(106) Bustin SA, Nolan T. Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. J Biomol Tech 2004 Sep;15(3):155-66.

(107) 2008 March 25. http://en.wikipedia.org/wiki/Real-time_polymerase_chain_reaction

(108) Peirson SN, Butler JN, Foster RG. Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis. Nucleic Acids Res 2003 Jul 15;31(14):e73.

(109) Wong ML, Medrano JF. Real-time PCR for mRNA quantitation. Biotechniques 2005 Jul;39(1):75-85.

(110) Mimmack ML, Brooking J, Bahn S. Quantitative polymerase chain reaction: validation of microarray results from postmortem brain studies. Biol Psychiatry 2004 Feb 15;55(4):337-45.

(111) Papadopoulos N, Kinzler KW, Vogelstein B. The role of companion diagnostics in the development and use of mutation-targeted cancer therapies. Nat Biotechnol 2006 Aug;24(8):985-95.

(112) Wilhelm Johannsen Centre for Functional Genome Research UoC. QPCR. 2006.
Ref Type: Slide

(113) UCSF Stanford Health Care. Genetic diagnosis of inherited disorders: When, why and how to use diagnostic molecular genetic testing methods? 2002.
Ref Type: Slide

(114) Bjarne Kjær Ersbøll, Knut Conradsen. An Introduction to Statistics. 7 ed. IMM; 2005.

(115) Steen Knudsen. Guide to Analysis of DNA Microarray Data, 2nd Edition. 2 ed. Wiley-Liss; 2004.

(116) Jerrold H.Zar. Biostatistical Analysis. 4 ed. Upper Saddle River, New Jersey 07458: Prentice Hall; 1999.

(117) Joseph F.Hair, William C.Black, Barry J.Babin, Rolph E.Anderson, Ronald L.Tatham. Multivariate Data Analysis. 6 ed. Upper Saddle River, New Jersey 07458: Pearson Prentice Hall; 2006.

(118) James Lattin, Douglas Carroll, Paul Green. Analyzing Multivariate Data. 1 ed. Pacific Grove, CA 93950, USA: Duxbury Press; 2003.

(119) Glenn A.Walker. Common Statistical Methods for Clinical Research with SAS Examples. 2 ed. Cary, North Caolina 27513, USA: SAS Institute Inc.; 2004.

(120) Liang Y, Kelemen A. Associating phenotypes with molecular events: recent statistical advances and challenges underpinning microarray experiments. Funct Integr Genomics 2006 Jan;6(1):1-13.

(121) Yuan JS, Reed A, Chen F, Stewart CN, Jr. Statistical analysis of real-time PCR data. BMC Bioinformatics 2006;7:85.

(122) Speed T. Statistics and Gene Expression Analysis. 2004.
Ref Type: Unpublished Work

(123) Mataix-Cols D, Rosario-Campos MC, Leckman JF. A multidimensional model of obsessive-compulsive disorder. Am J Psychiatry 2005 Feb;162(2):228-38.

(124) Ignacio Gonzalez, Sebastien Dejean, Pascal G.P.Martin, Alain Baccini. CCA: an R package to extend canonical correlation. 2008.
Ref Type: Unpublished Work

(125) Joseph F.Hair, William C.Black, Barry J.Babin, Rolph E.Anderson, Ronald L.Tatham. Examining Your Data. Multivariate Data Analysis. 6 ed. Upper Saddle River, New Jersey 07458: Pearson Prentice Hall; 2006. p. 79.

(126) 2008 April 4. http://en.wikipedia.org/wiki/Qq_plot

(127) Joseph F.Hair, William C.Black, Barry J.Babin, Rolph E.Anderson, Ronald L.Tatham. Examining Your Data. Multivariate Data Analysis. 6 ed. Upper Saddle River, New Jersey 07458: Pearson Prentice Hall; 2006. p. 82.

(128) 2008 April 3. http://en.wikipedia.org/wiki/Normality_test

(129) 2008 April 3. http://en.wikipedia.org/wiki/Shapiro-Wilk_test

(130) 2008 April 3. http://en.wikipedia.org/wiki/Anderson-Darling_test

(131) 2008 April 3. http://en.wikipedia.org/wiki/Cram%C3%A9r-von-Mises_criterion

(132) 2008 April 3. http://en.wikipedia.org/wiki/Lilliefors_distribution

(133) 2008 April 3. http://cran.r-project.org/web/packages/nortest/index.html

(134) Jerrold H.Zar. Nonparametric statistical methods. Biostatistical Analysis. 4 ed. Upper Saddle River, New Jersey 07458: Prentice Hall; 1999. p. 145-6.

(135) 2008 April 4. http://en.wikipedia.org/wiki/T-test

(136) 2008 April 4. http://en.wikipedia.org/wiki/Welch%27s_t_test

(137) Jerrold H.Zar. Two-sample hypotheses. Biostatistical Analysis. 4 ed. Upper Saddle River, New Jersey 07458: Prentice Hall; 1999. p. 122-9.

(138) 2008 June 4. http://en.wikipedia.org/wiki/Mann-Whitney_U

(139) Glenn A.Walker. The Wilcoxon Rank-Sum Test. Common Statistical Methods for Clinical Research with SAS Examples. 2 ed. Cary, North Caolina 27513, USA: SAS Institute Inc.; 2004. p. 237-45.

(140) Jerrold H.Zar. Two-sample rank testing. Biostatistical Analysis. 4 ed. Upper Saddle River, New Jersey 07458: Prentice Hall; 1999. p. 146-55.

(141) Glenn A.Walker. One-Way ANOVA. Common Statistical Methods for Clinical Research with SAS Examples. 2 ed. Cary, North Caolina 27513, USA: SAS Institute Inc.; 2004. p. 77-90.

(142) Jerrold H.Zar. Single-factor analysis of variance. Biostatistical Analysis. 4 ed. Upper Saddle River, New Jersey 07458: Prentice Hall; 1999. p. 178-89.

(143) 2008 April 4. http://en.wikipedia.org/wiki/ANOVA

(144) Jerrold H.Zar. Nonparametric analysis of variance. Biostatistical Analysis. 4 ed. Upper Saddle River, New Jersey 07458: Prentice Hall; 1999. p. 195-200.

(145) Glenn A.Walker. The Kruskal-Wallis Test. Common Statistical Methods for Clinical Research with SAS Examples. 2 ed. Cary, North Caolina 27513, USA: SAS Institute Inc.; 2004. p. 247-56.

(146) 2008 April 7. http://en.wikipedia.org/wiki/Kruskal-Wallis_one-way_analysis_of_variance

(147) 2008 April 7. http://en.wikipedia.org/wiki/Bonferroni

(148) Glenn A.Walker. Repeated Measures ANOVA. Common Statistical Methods for Clinical Research with SAS Examples. 2 ed. Cary, North Caolina 27513, USA: SAS Institute Inc.; 2004. p. 111-56.

(149) 2008 April 8. http://en.wikipedia.org/wiki/Mauchly%27s_sphericity_test

(150) 2008 April 8. http://www.abdn.ac.uk/~psy317/personal/files/teaching/spheric.htm

(151) 2008 April 8. http://en.wikipedia.org/wiki/Spearman_correlation

(152) Jerrold H.Zar. Rank correlation. Biostatistical Analysis. 4 ed. Upper Saddle River, New Jersey 07458: Prentice Hall; 1999. p. 395-8.

(153) 2008 April 8. http://mrw.interscience.wiley.com/emrw/9780471667193/ess/article/ess5050/current/abstract

(154) Jerrold H.Zar. Comparing two correlation coefficients. Biostatistical Analysis. 4 ed. Upper Saddle River, New Jersey 07458: Prentice Hall; 1999. p. 386-8.

(155) Joseph F.Hair, William C.Black, Barry J.Babin, Rolph E.Anderson, Ronald L.Tatham. Adapted from chapter 8: Canonical Correlation Analysis. Multivariate Data Analysis. 5 ed. Upper Saddle River, New Jersey 07458: Pearson Prentice Hall; 1998.

(156) 2008 April 9. http://en.wikipedia.org/wiki/Canonical_correlation

(157) 2008 April 10. http://en.wikipedia.org/wiki/Recursive_partitioning

(158) 2008 April 10. http://cran.r-project.org/web/packages/rpart/index.html

(159) Joseph F.Hair, William C.Black, Barry J.Babin, Rolph E.Anderson, Ronald L.Tatham. Cluster Analysis. Multivariate Data Analysis. 6 ed. Upper Saddle River, New Jersey 07458: Pearson Prentice Hall; 2006. p. 555-628.

(160) 2008 April 11. http://en.wikipedia.org/wiki/Data_clustering

(161) Tibshirani R, Hastie T, Eisen M, Ross D, Botstein D, Brown P. Clustering methods for the analysis of DNA microarray data. 1999.

(162) D'haeseleer P. How does gene expression clustering work? Nat Biotechnol 2005 Dec;23(12):1499-501.

(163) James Lattin, Douglas Carroll, Paul Green. Cluster Analysis. Analyzing Multivariate Data. 1 ed. Pacific Grove, CA 93950, USA: Duxbury Press; 2003. p. 264-310.

(164) Antonijevic I, Mazin W. Stepwise regression discussion. 2007. Ref Type: Personal Communication

(165) Bjarne Kjær Ersbøll, Knut Conradsen. Choice of the "best" regression equation. An Introduction to Statistics. 7 ed. IMM; 2005.

(166) 2008 April 14. http://en.wikipedia.org/wiki/Stepwise_regression

(167) 2008 April 14. http://en.wikipedia.org/wiki/Akaike_information_criterion

(168) 2008 April 14. http://cran.r-project.org/web/packages/VR/index.html

(169) 2008 April 14. http://en.wikipedia.org/wiki/Linear_regression

(170) 2008 April 15. http://en.wikipedia.org/wiki/MANOVA

(171) 2008 April 15. http://cran.r-project.org/web/packages/ffmanova/index.html

(172) 2008 April 15. http://en.wikipedia.org/wiki/Principal_components_analysis

(173) Bjarne Kjær Ersbøll, Knut Conradsen. Principal components. An Introduction to Statistics. 7 ed. IMM; 2005.

(174) 2008 April 15. http://cran.r-project.org/web/packages/elasticnet/index.html

(175) Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform 2007 Jan;8(1):32-44.

(176) Rosipal R, Krämer N. Overview and Recent Advances in Partial Least Squares. In: Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J, editors. Subspace, Latent Structure and Feature Selection Techniques. Springer; 2006. p. 34-51.

(177) Frey BJ, Dueck D. Clustering by passing messages between data points. Science 2007 Feb 16;315(5814):972-6.

(178) Boulesteix A-L, Strobl C, Augustin T, Daumer M. Evaluating microarray-based classifiers: an overview. Cancer Informatics 2008;4:77-97.

(179) Dettling M, Bühlmann P. Finding Predictive Gene Groups from Microarray Data. Journal of Multivariate Analysis 2004;90(1):106-31.

(180) 2008 April 23. http://cran.r-project.org/web/packages/supclust/index.html

(181) 2008 April 22. http://en.wikipedia.org/wiki/LOOCV

(182) Weber G, Vinterbo S, Ohno-Machado L. Multivariate selection of genetic markers in diagnostic classification. Artif Intell Med 2004 Jun;31(2):155-67.

(183) 2008 April 23. http://cran.r-project.org/web/packages/stepPlr/

(184) 2008 April 23. http://cran.r-project.org/web/packages/varSelRF/index.html

(185) az-Uriarte R. GeneSrF and varSelRF: a web-based tool and R package for gene selection and classification using random forest. BMC Bioinformatics 2007;8:328.

(186) 2008 April 23. http://en.wikipedia.org/wiki/Naive_bayes

(187) 2008 April 23. http://cran.r-project.org/web/packages/klaR/index.html

(188) 2008 April 23. http://en.wikipedia.org/wiki/Linear_discriminant_analysis

(189) 2008 April 23. http://en.wikipedia.org/wiki/KNN

(190) 2008 April 23. http://cran.r-project.org/web/packages/randomForest/index.html

(191) 2008 April 23. http://en.wikipedia.org/wiki/Random_forest

(192) az-Uriarte R, varez de AS. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3.

(193) 2008 April 23. http://cran.r-project.org/web/packages/caMassClass/index.html

(194) 2008 April 23. http://en.wikipedia.org/wiki/Quadratic_classifier#Quadratic_discriminant_analysis

(195) 2008 April 23. http://en.wikipedia.org/wiki/Support_vector_machine

(196) 2008 April 23. http://en.wikipedia.org/wiki/Neural_network

(197) 2008 April 23. http://en.wikipedia.org/wiki/Boosting

(198) 2008 April 23. http://en.wikipedia.org/wiki/Accuracy#Accuracy_in_binary_classification

(199) Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. IJCAI 1995.

(200) 2008 April 23. http://en.wikipedia.org/wiki/Jaccard_index#Similarity_of_asymmetric_binary_attributes

(201) Cai Z, Goebel R, Salavatipour MR, Lin G. Selecting dissimilar genes for multi-class classification, an application in cancer subtyping. BMC Bioinformatics 2007;8:206.

(202) Mukherjee S, Golland P, Panchenko D. Permutation Tests for Classification. 2003 Aug 28.

(203) 2008 April 25. http://en.wikipedia.org/wiki/Positive_predictive_value

(204) 2008 April 25. http://en.wikipedia.org/wiki/Negative_predictive_value

(205) Liaw A, Wiener M. Classification and regression by randomForest. R News 2[3], 18-22. 1-12-2002.
Ref Type: Magazine Article

(206) 2008 April 28. http://cran.r-project.org/web/packages/e1071/vignettes/svmdoc.pdf

(207) 2008 April 28. http://en.wikipedia.org/wiki/Kernel_trick

(208) 2008 April 28. http://cran.r-project.org/web/packages/e1071/e1071.pdf

(209) 2008 May 6. http://www.wtccc.org.uk/

(210) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007 Jun 7;447(7145):661-78.

(211) 2008 May 6. http://www.ncbi.nlm.nih.gov/projects/SNP/

(212) 2008 May 6. http://string.embl.de/

(213) Heuser I, Deuschle M, Weber A, Kniest A, Ziegler C, Weber B, et al. The role of mineralocorticoid receptors in the circadian activity of the human hypothalamus-pituitary-adrenal system: effect of age. Neurobiol Aging 2000 Jul;21(4):585-9.

(214) Young EA, Lopez JF, Murphy-Weinberg V, Watson SJ, Akil H. The role of mineralocorticoid receptors in hypothalamic-pituitary-adrenal axis regulation in humans. J Clin Endocrinol Metab 1998 Sep;83(9):3339-45.

(215) Gladkevich A, Kauffman HF, Korf J. Lymphocytes as a neural probe: potential for studying psychiatric disorders. Prog Neuropsychopharmacol Biol Psychiatry 2004 May;28(3):559-76.

(216) 2008 June 6. http://en.wikipedia.org/wiki/Matthews_Correlation_Coefficient

(217) Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. Science 2005 Apr 22;308(5721):523-9.

(218) McClung CA. Circadian genes, rhythms and the biology of mood disorders. Pharmacol Ther 2007 May;114(2):222-32.

(219) Barnard AR, Nolan PM. When clocks go bad: neurobehavioural consequences of disrupted circadian timing. PLoS Genet 2008;4(5):e1000040.

(220) Bunney WE, Bunney BG. Molecular clock genes in man and lower animals: possible implications for circadian abnormalities in depression. Neuropsychopharmacology 2000 Apr;22(4):335-45.

(221) Ko CH, Takashi JS. Molecular components of the mammalian circadian clock. Hum Mol Genet 2006;15(2):R271-R277.

(222) Levi F, Schibler U. Circadian rhythms: mechanisms and therapeutic implications. Annu Rev Pharmacol Toxicol 2007;47:593-628.

(223) Song H, Smolen P, Av-Ron E, Baxter DA, Byrne JH. Dynamics of a minimal model of interlocked positive and negative feedback loops of transcriptional regulation by cAMP-response element binding proteins. Biophys J 2007 May 15;92(10):3407-24.

(224) Bayesian Inference for Gene Expression and Proteomics. Cambridge University Press; 2006.

(225) Bishop CM. Graphical Models. Pattern Recognition and Machine Learning. Springer; 2006. p. 359-418.

(226) Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Bayesian methods for elucidating genetic regulatory networks. IEEE Intelligent Systems 2002;17(2):37-43.

(227) Zhang H, Yu CY, Singer B. Cell and tumor classification using gene expression data: construction of forests. Proc Natl Acad Sci U S A 2003 Apr 1;100(7):4168-72.

(228) Iwamoto K, Kato T. Gene expression profiling in schizophrenia and related mental disorders. Neuroscientist 2006 Aug;12(4):349-61.

(229) Bunney WE, Bunney BG, Vawter MP, Tomita H, Li J, Evans SJ, et al. Microarray technology: a review of new strategies to discover candidate vulnerability genes in psychiatric disorders. Am J Psychiatry 2003 Apr;160(4):657-66.

(230) Iga J, Ueno S, Ohmori T. Molecular assessment of depression from mRNAs in the peripheral leukocytes. Ann Med 2008;40(5):336-42.

(231) Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC. Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics 2006 Oct 1;22(19):2405-12.

(232) Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. Bioinformatics 2005 Aug 1;21(15):3201-12.

(233) Batliwalla FM, Li W, Ritchlin CT, Xiao X, Brenner M, Laragione T, et al. Microarray analyses of peripheral blood cells identifies unique gene expression signature in psoriatic arthritis. Mol Med 2005 Jan;11(1-12):21-9.

(234) Chu LH, Chen BS. Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. BMC Syst Biol 2008;2:56.

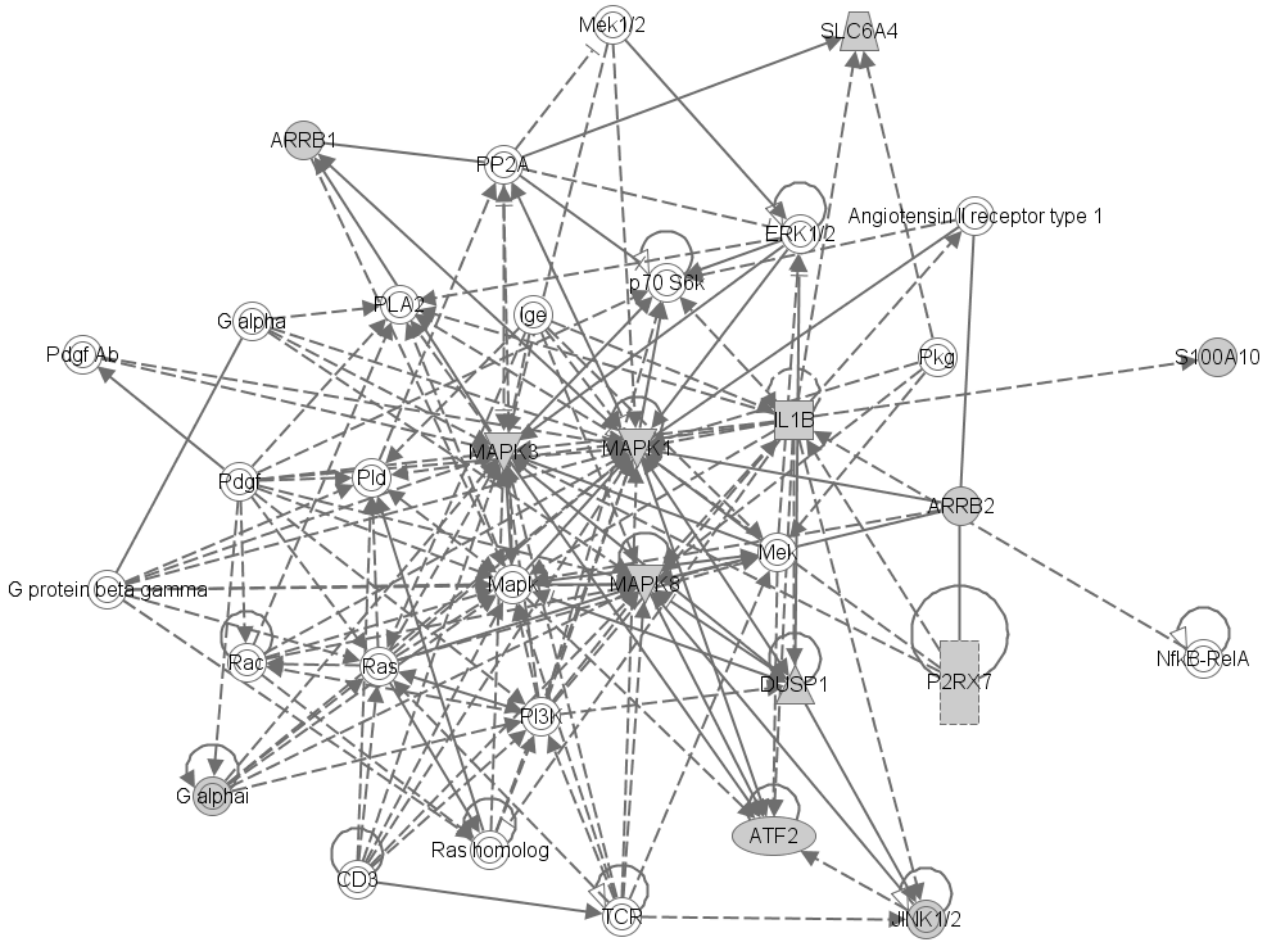
Appendices

Appendix 1: Three networks showing the 29 genes and interacting genes

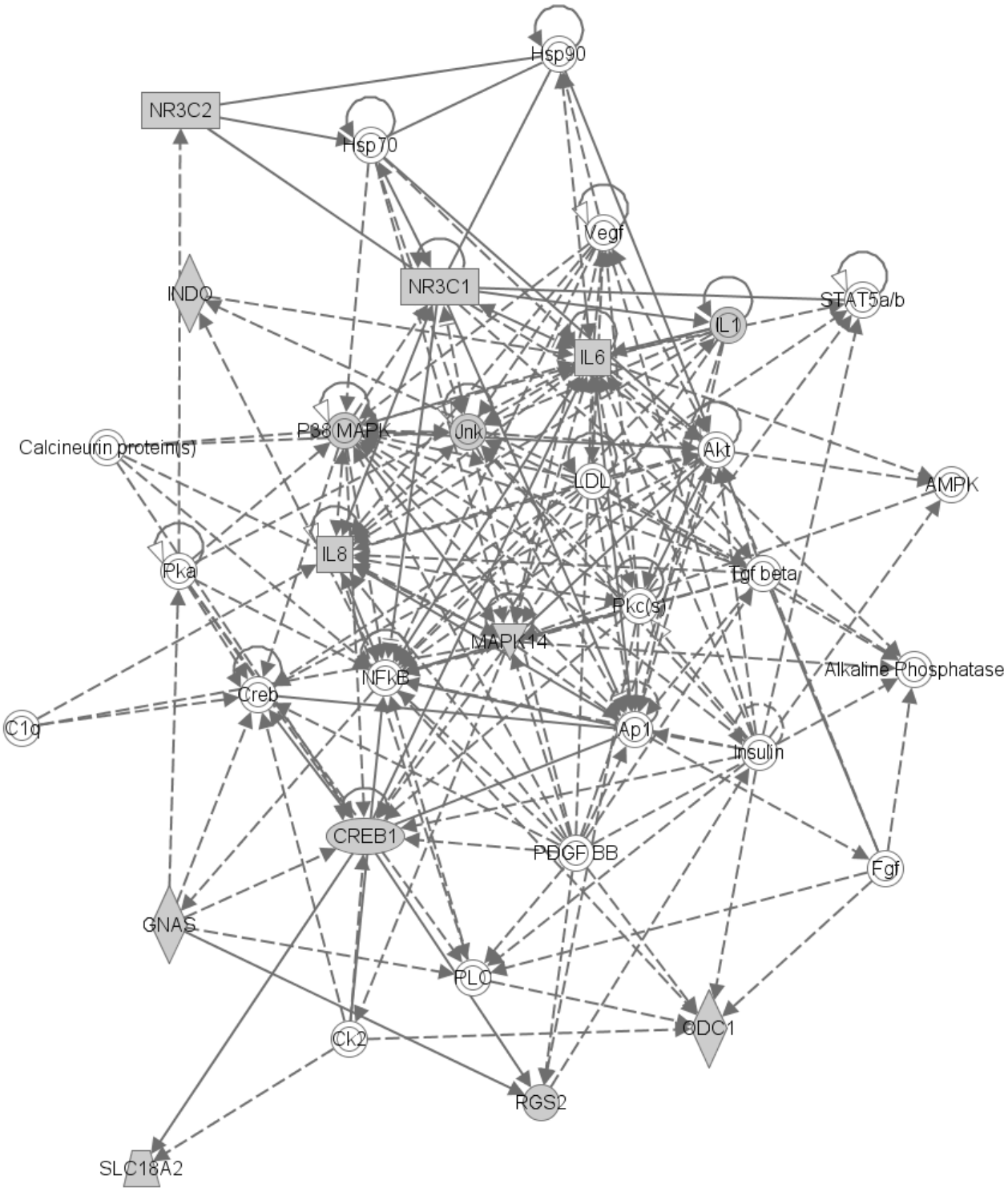
Below are shown the three networks from table 4 involving the 29 genes and interacting genes. Genes shown in grey are part of the 29 gene list.

Ingenuity requires gene names to be entered via their Entrez gene name: ATF2=CREB2, NR3C1=GR, NR3C2=MR, DUSP1=MKP1, TSPO=PBR, MAPK1=ERK2, MAPK3=ERK1, INDO=IDO, SLC6A4=SERT, SLC18A2=VMAT2, GNAI2=Gi2 and GNAS=Gs.

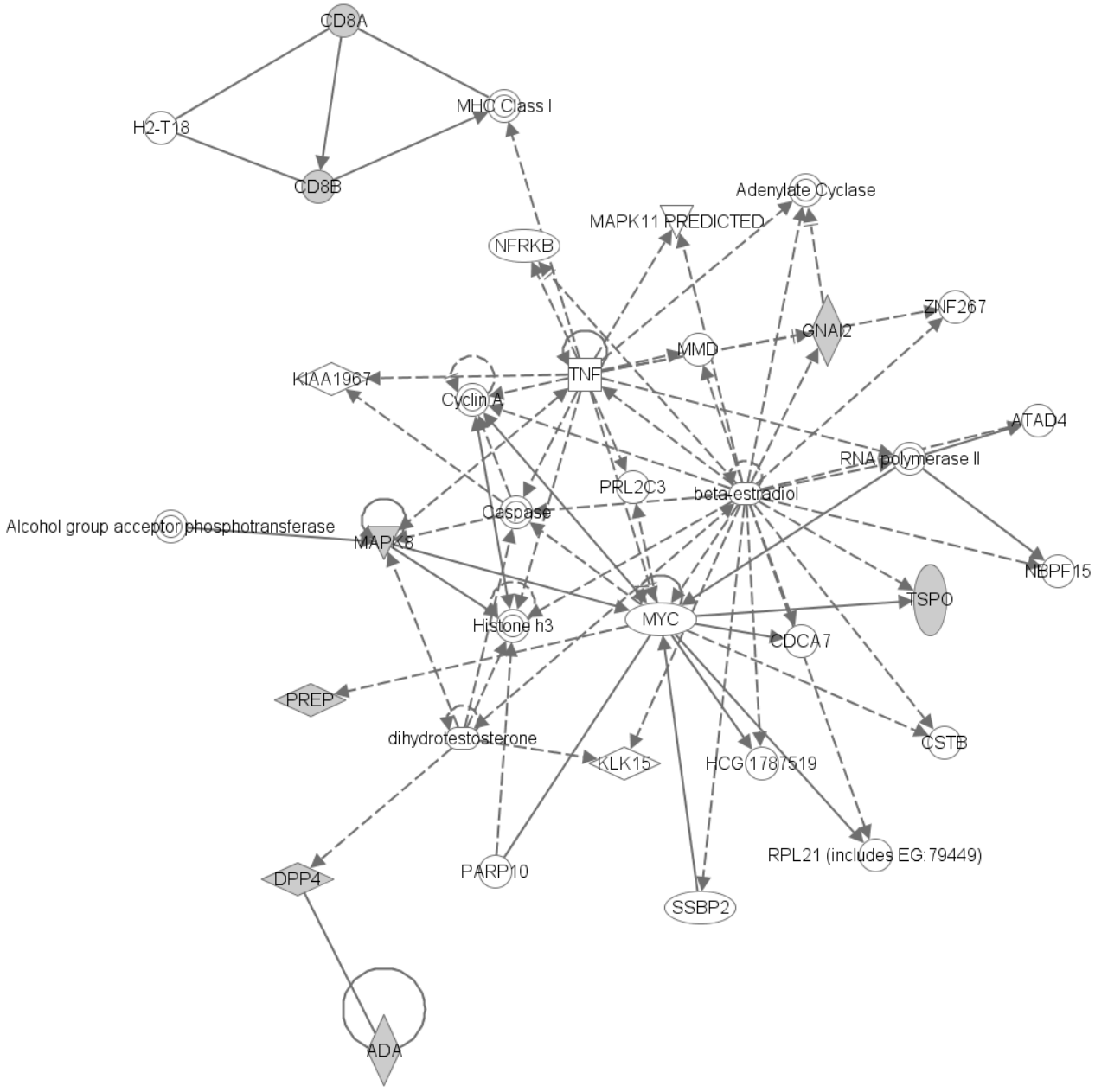
Network 1 : entrez_gene_list - 2007-11-27 12:42 PM : entrez_gene_list.lst



Network 2 : entrez_gene_list - 2007-11-27 12:42 PM : entrez_gene_list.lst



Network 3 : entrez_gene_list - 2007-11-27 12:42 PM : entrez_gene_list.lst



Appendix 2: Significant biological functions among the 29 genes

Output from Ingenuity (Ingenuity Systems®, www.ingenuity.com).

Function	p-value	Molecules
Cell Cycle	9.61E-11-2.54E-04	IL8, DPP4, MAPK1, MAPK3, MAPK8, IL6, NR3C1, ATF2, GNAS, MAPK14, ARRB1, DUSP1, CREB1, ADA, IL1B
Inflammatory Disease	2.85E-10-2.4E-04	IL8, DPP4, MAPK3, MAPK8, IL6, P2RX7, CD8A, NR3C1, ODC1, GNAI2, ARRB2, MAPK14, DUSP1, ADA, IL1B, SLC6A4, S100A10
Cell Death	4.09E-10-2.49E-04	DPP4, IL8, MAPK1, MAPK3, MAPK8, P2RX7, IL6, SLC18A2, CD8A, NR3C1, ODC1, ATF2, GNAS, ARRB2, MAPK14, DUSP1, CREB1, ADA, TSPO, IL1B, S100A10
Cellular Growth and Proliferation	6.62E-09-2.54E-04	DPP4, IL8, RGS2, MAPK1, MAPK3, MAPK8, P2RX7, IL6, CD8A, NR3C1, ODC1, ATF2, GNAI2, ARRB2, MAPK14, DUSP1, CREB1, ADA, IL1B, INDO, SLC6A4, NR3C2, S100A10
Connective Tissue Disorders	6.9E-09-2.4E-04	IL8, DPP4, MAPK3, MAPK8, IL6, P2RX7, NR3C1, ATF2, GNAI2, ARRB2, MAPK14, DUSP1, IL1B, SLC6A4, S100A10
Skeletal and Muscular Disorders	6.9E-09-2.54E-04	IL8, DPP4, MAPK1, MAPK3, MAPK8, IL6, P2RX7, NR3C1, GNAI2, GNAS, MAPK14, DUSP1, IL1B, SLC6A4, S100A10
Cancer	8.17E-09-2.54E-04	IL8, DPP4, MAPK1, MAPK3, MAPK8, P2RX7, IL6, CD8B, NR3C1, ODC1, ATF2, GNAI2, GNAS, ARRB2, MAPK14, ARRB1, DUSP1, CREB1, ADA, IL1B, NR3C2, S100A10
Cellular Development	9.04E-09-2.54E-04	DPP4, IL8, RGS2, MAPK1, MAPK3, MAPK8, IL6, P2RX7, CD8A, CD8B, NR3C1, ODC1, ATF2, ARRB2, MAPK14, DUSP1, CREB1, ADA, IL1B, S100A10
Post-Translational Modification	1.25E-08-2.12E-04	IL8, ARRB2, MAPK14, MAPK1, DUSP1, MAPK3, MAPK8, INDO, IL1B, IL6, CD8A, ODC1
Cardiovascular Disease	1.43E-08-7.69E-05	IL8, RGS2, MAPK1, MAPK3, MAPK8, IL6, SLC18A2, NR3C1, GNAI2, MAPK14, DUSP1, CREB1, IL1B, SLC6A4, NR3C2, S100A10
Hematological System Development and Function	2.43E-08-2.37E-04	IL8, DPP4, RGS2, MAPK1, MAPK3, MAPK8, P2RX7, IL6, CD8A, NR3C1, CD8B, GNAI2, GNAS, ARRB2, MAPK14, DUSP1, CREB1, ADA, IL1B, INDO
Immunological Disease	3.83E-08-8.42E-05	IL8, DPP4, MAPK3, MAPK8, P2RX7, IL6, CD8A, NR3C1, GNAS, MAPK14, DUSP1, ADA, IL1B, SLC6A4, S100A10
Gene Expression	6.32E-08-1.49E-04	IL8, MAPK1, MAPK3, MAPK8, IL6, NR3C1, ATF2, GNAS, ARRB2, ARRB1, MAPK14, DUSP1, CREB1, IL1B, NR3C2
Hematological Disease	6.39E-08-1.43E-04	DPP4, GNAS, IL8, MAPK14, MAPK8, ADA, IL1B, P2RX7, IL6, NR3C1
Developmental Disorder	6.6E-08-4.29E-06	GNAI2, DPP4, MAPK14, MAPK1, DUSP1, CREB1, MAPK8, ADA, IL1B, NR3C2, IL6, S100A10
Organ Morphology	6.6E-08-2.17E-04	GNAI2, GNAS, MAPK14, DUSP1, CREB1, MAPK8, IL1B, NR3C2, IL6, NR3C1
Behavior	1.1E-07-1.49E-04	RGS2, MAPK1, MAPK3, IL6, SLC18A2, NR3C1, GNAI2, ARRB2, ARRB1, CREB1, IL1B, SLC6A4, NR3C2
Neurological Disease	1.9E-07-2.54E-04	IL8, DPP4, MAPK8, IL6, P2RX7, SLC18A2, NR3C1, ATF2, PREP, GNAS, MAPK14, CREB1, ADA, TSPO, IL1B, SLC6A4
Lipid Metabolism	2.63E-07-7.11E-05	IL8, RGS2, MAPK1, MAPK8, P2RX7, IL6, NR3C1, GNAI2, GNAS, ARRB2, ARRB1, MAPK14, ADA, SLC6A4, IL1B, S100A10
Molecular Transport	2.63E-07-2.54E-04	IL8, RGS2, MAPK1, MAPK8, IL6, P2RX7, CD8A, SLC18A2, NR3C1, GNAI2, GNAS, ARRB2, MAPK14,

		ARRB1, ADA, SLC6A4, IL1B, NR3C2, S100A10
Small Molecule Biochemistry	2.63E-07-2.12E-04	IL8, RGS2, MAPK1, MAPK3, MAPK8, IL6, P2RX7, SLC18A2, NR3C1, GNAI2, GNAS, ARRB2, ARRB1, MAPK14, DUSP1, ADA, INDO, SLC6A4, IL1B, S100A10
DNA Replication, Recombination, and Repair	2.66E-07-2.54E-04	GNAI2, IL8, ARRB2, ARRB1, DUSP1, MAPK8, ADA, IL1B, IL6, NR3C1
Organismal Survival	5.66E-07-7.55E-07	GNAI2, GNAS, DUSP1, CREB1, MAPK8, ADA, IL1B, NR3C2, IL6, SLC18A2, CD8A, NR3C1
Metabolic Disease	6.12E-07-2.54E-04	DPP4, GNAS, DUSP1, CREB1, ADA, SLC6A4, IL1B, NR3C2, IL6, NR3C1, S100A10
Amino Acid Metabolism	6.25E-07-2.12E-04	MAPK14, MAPK1, DUSP1, MAPK3, MAPK8, INDO, IL6
Connective Tissue Development and Function	1.01E-06-1.49E-04	IL8, RGS2, ARRB2, MAPK14, MAPK1, MAPK3, CREB1, MAPK8, SLC6A4, IL1B, IL6, NR3C1
Tissue Morphology	1.31E-06-2.02E-04	IL8, MAPK3, IL6, P2RX7, NR3C1, ATF2, GNAI2, GNAS, MAPK14, CREB1, ADA, IL1B, SLC6A4, NR3C2
Reproductive System Disease	1.32E-06-1.67E-04	IL8, MAPK1, MAPK3, MAPK8, IL6, ODC1, ATF2, ARRB2, MAPK14, ARRB1, DUSP1, IL1B, S100A10
Cellular Movement	1.57E-06-2.19E-04	IL8, DPP4, MAPK1, MAPK3, MAPK8, IL6, CD8A, ODC1, GNAI2, GNAS, ARRB2, MAPK14, ARRB1, DUSP1, IL1B, S100A10
Immune Response	1.57E-06-2.23E-04	IL8, DPP4, RGS2, MAPK3, MAPK8, IL6, CD8A, NR3C1, GNAI2, GNAS, ARRB2, MAPK14, DUSP1, ADA, IL1B, INDO
Psychological Disorders	2.03E-06-1.43E-04	RGS2, CREB1, IL1B, TSPO, SLC6A4
Cell-To-Cell Signaling and Interaction	2.95E-06-2.54E-04	GNAI2, GNAS, IL8, ARRB2, MAPK8, ADA, SLC6A4, IL1B, IL6, CD8A
Immune and Lymphatic System Development and Function	2.95E-06-2.37E-04	IL8, DPP4, RGS2, MAPK1, MAPK3, MAPK8, P2RX7, IL6, CD8A, NR3C1, CD8B, MAPK14, CREB1, ADA, IL1B, INDO
Nutritional Disease	3.24E-06-4.82E-05	DPP4, MAPK8, SLC6A4, IL1B, NR3C2, IL6, NR3C1
Organismal Functions	3.92E-06-2.09E-04	ARRB2, SLC6A4, IL1B, SLC18A2, NR3C1
Endocrine System Disorders	7.15E-06-1.98E-04	DPP4, GNAS, IL8, DUSP1, CREB1, IL1B, NR3C2, IL6, NR3C1, S100A10
Endocrine System Development and Function	7.27E-06-1.98E-04	SLC6A4, IL1B, IL6, NR3C1
Nervous System Development and Function	9.34E-06-2.08E-04	MAPK1, CREB1, SLC6A4, IL1B, IL6, ATF2
Cell Signaling	1.09E-05-2.09E-04	IL8, RGS2, MAPK1, MAPK3, MAPK8, P2RX7, IL6, CD8A, SLC18A2, ATF2, GNAI2, GNAS, ARRB2, ARRB1, MAPK14, ADA, IL1B, SLC6A4
Nucleic Acid Metabolism	1.09E-05-2.09E-04	GNAI2, GNAS, RGS2, ARRB2, ARRB1, ADA, IL1B, SLC6A4, SLC18A2
Hepatic System Development and Function	1.18E-05-4E-05	MAPK1, MAPK8, IL1B, IL6
Organismal Injury and Abnormalities	1.47E-05-1.12E-04	GNAS, IL8, DUSP1, MAPK8, SLC6A4, IL1B, P2RX7, IL6
Hair and Skin Development and Function	1.81E-05-1.8E-04	IL8, MAPK8, IL6, NR3C1
Dermatological Diseases and Conditions	1.83E-05-1.83E-05	GNAI2, IL8, DUSP1, IL1B, IL6, NR3C1, ODC1, ATF2

Cellular Compromise	2.14E-05-2.48E-04	CREB1, IL1B, IL6, ATF2
Protein Trafficking	2.14E-05-2.14E-05	MAPK1, MAPK3
Cell Morphology	2.18E-05-1.62E-04	IL8, RGS2, ARRB2, MAPK14, MAPK1, MAPK3, CREB1, ADA, P2RX7, IL6, ODC1
Carbohydrate Metabolism	2.51E-05-2.54E-04	GNAS, IL8, IL1B, IL6, P2RX7
Gastrointestinal Disease	2.51E-05-2.51E-05	DUSP1, MAPK8, P2RX7, IL6
Hepatic System Disease	2.51E-05-2.54E-04	IL8, DUSP1, MAPK8, IL1B, P2RX7, IL6
Respiratory Disease	2.92E-05-4E-05	ADA, IL1B, IL6, CD8A, NR3C1
Skeletal and Muscular System Development and Function	3.01E-05-1.98E-04	IL8, MAPK1, DUSP1, IL1B, IL6
Cardiovascular System Development and Function	3.09E-05-3.09E-05	IL8, SLC6A4, IL1B
Organismal Development	4.52E-05-4.52E-05	IL8, SLC6A4, IL1B
Cellular Assembly and Organization	5.67E-05-1.62E-04	GNAI2, IL8, MAPK1, CREB1, SLC6A4, IL1B, IL6
Renal and Urological Disease	7.11E-05-2.54E-04	MAPK1, MAPK8
Embryonic Development	1E-04-1E-04	MAPK14, MAPK1, MAPK8, IL6
Cellular Function and Maintenance	1.49E-04-1.49E-04	SLC6A4, IL1B
Viral Function	1.8E-04-2.54E-04	IL8, IL1B, IL6
Vitamin and Mineral Metabolism	1.95E-04-1.95E-04	IL8, IL1B, P2RX7, IL6, CD8A
Genetic Disorder	1.98E-04-2.54E-04	GNAS, NR3C2, NR3C1

Appendix 3: Significant pathways involving the 29 genes

Below the significant (below 1%) pathways are shown for various combinations of the 29 genes. For the difference between the statistical significance (p-value) and ratio, see text after the table. Output from Ingenuity (Ingenuity Systems®, www.ingenuity.com).

Pathway	-log(p-value)	Ratio	Molecules
Glucocorticoid Receptor Signaling	1.11E+01	4.15E-02	IL8, MAPK14, MAPK1, DUSP1, MAPK3, CREB1, MAPK8, IL1B, NR3C2, IL6, NR3C1
IL-6 Signaling	8.67E+00	7.69E-02	IL8, MAPK14, MAPK1, MAPK3, MAPK8, IL1B, IL6
cAMP-mediated Signaling	8.46E+00	5.03E-02	GNAI2, GNAS, RGS2, MAPK1, DUSP1, MAPK3, CREB1, ATF2
G-Protein Coupled Receptor Signaling	7.69E+00	4.02E-02	GNAI2, GNAS, RGS2, MAPK1, DUSP1, MAPK3, CREB1, ATF2
FGF Signaling	7.28E+00	7.14E-02	MAPK14, MAPK1, MAPK3, CREB1, MAPK8, ATF2
PPARα /RXRα Activation	6.73E+00	3.98E-02	GNAS, MAPK14, MAPK1, MAPK3, MAPK8, IL1B, IL6
Acute Phase Response Signaling	6.73E+00	4.07E-02	MAPK14, MAPK1, MAPK3, MAPK8, IL1B, IL6, NR3C1
Neurotrophin/TRK Signaling	6.22E+00	6.85E-02	MAPK1, MAPK3, CREB1, MAPK8, ATF2
Ephrin Receptor Signaling	6.00E+00	3.02E-02	GNAI2, GNAS, MAPK1, MAPK3, CREB1, MAPK8, ATF2
Chemokine Signaling	5.92E+00	6.67E-02	GNAI2, MAPK14, MAPK1, MAPK3, MAPK8
B Cell Receptor Signaling	5.84E+00	4.05E-02	MAPK14, MAPK1, MAPK3, CREB1, MAPK8, ATF2
p38 MAPK Signaling	5.39E+00	5.26E-02	MAPK14, DUSP1, CREB1, IL1B, ATF2
T Cell Receptor Signaling	5.36E+00	4.90E-02	MAPK1, MAPK3, MAPK8, CD8A, CD8B
IL-10 Signaling	4.80E+00	5.88E-02	MAPK14, MAPK8, IL1B, IL6
ERK/MAPK Signaling	4.75E+00	2.65E-02	MAPK1, DUSP1, MAPK3, CREB1, MAPK8, ATF2
Aryl Hydrocarbon Receptor Signaling	4.55E+00	3.29E-02	MAPK1, MAPK3, MAPK8, IL1B, IL6
Xenobiotic Metabolism Signaling	4.49E+00	2.40E-02	MAPK14, MAPK1, MAPK3, MAPK8, IL1B, IL6
Fc Epsilon RI Signaling	3.96E+00	4.00E-02	MAPK14, MAPK1, MAPK3, MAPK8
Synaptic Long Term Potentiation	3.82E+00	3.60E-02	MAPK1, MAPK3, CREB1, ATF2
EGF Signaling	3.68E+00	6.38E-02	MAPK1, MAPK3, MAPK8
IL-2 Signaling	3.49E+00	5.66E-02	MAPK1, MAPK3, MAPK8
Amyloid Processing	3.46E+00	5.77E-02	MAPK14, MAPK1, MAPK3
Synaptic Long Term Depression	3.41E+00	2.45E-02	GNAI2, GNAS, MAPK1, MAPK3
Hepatic Cholestasis	3.41E+00	2.47E-02	IL8, MAPK8, IL1B, IL6
SAPK/JNK Signaling	3.27E+00	2.72E-02	MAPK1, MAPK3, MAPK8, ATF2
Parkinson's Signaling	3.13E+00	1.18E-01	MAPK14, MAPK8
PDGF Signaling	3.09E+00	4.05E-02	MAPK1, MAPK3, MAPK8
Calcium Signaling	3.02E+00	1.96E-02	MAPK1, MAPK3, CREB1, ATF2

NRF2-mediated Oxidative Stress Response	2.96E+00	2.22E-02	MAPK14, MAPK1, MAPK3, MAPK8
TGF-β Signaling	2.94E+00	3.61E-02	MAPK1, MAPK3, MAPK8
Nicotinate and Nicotinamide Metabolism	2.90E+00	2.33E-02	MAPK1, MAPK3, MAPK8
PPAR Signaling	2.83E+00	3.16E-02	MAPK1, MAPK3, IL1B
IGF-1 Signaling	2.83E+00	3.26E-02	MAPK1, MAPK3, MAPK8
Estrogen Receptor Signaling	2.61E+00	2.54E-02	MAPK1, MAPK3, NR3C1
Circadian Rhythm Signaling	2.59E+00	6.25E-02	CREB1, ATF2
Serotonin Receptor Signaling	2.50E+00	4.35E-02	SLC6A4, SLC18A2
Hepatic Fibrosis / Hepatic Stellate Cell Activation	2.39E+00	2.29E-02	IL8, IL1B, IL6
Insulin Receptor Signaling	2.38E+00	2.26E-02	MAPK1, MAPK3, MAPK8
Inositol Phosphate Metabolism	2.38E+00	1.73E-02	MAPK1, MAPK3, MAPK8
Apoptosis Signaling	2.30E+00	2.21E-02	MAPK1, MAPK3, MAPK8
Toll-like Receptor Signaling	2.18E+00	3.92E-02	MAPK14, MAPK8
PI3K/AKT Signaling	2.08E+00	1.70E-02	MAPK1, MAPK3, MAPK8

The ratio is calculated by taking the number of genes from the 29 gene set that participate in a pathway, and dividing it by the total number of Ingenuity genes in that pathway. The ratio indicates the percentage of genes in a pathway that were also found in the 29 gene list. The ratio is therefore useful for determining which pathways overlap the most with the 29 genes.

The p-value measures how likely the observed association between a specific pathway and a certain combination of the 29 genes would be if it was only due to random chance. If a p-value is very small one can be confident that the corresponding pathway is significantly associated with the 29 genes.

The ratio indicates the strength of the association, whereas the p-value measures its statistical significance.

Appendix 4: The US ABS questionnaire

Official Use ONLY

Identification number: __ __ __ __ __ __
To be used for blood sample labeling

Today's Date (mm/dd/yy) __ __ / __ __ / __ __
Time of blood collection __ __ : __ __ am / pm (circle one)

IT IS IMPORTANT THAT YOU COMPLETE THE ENTIRE QUESTIONNAIRE (ALL THE INFORMATION WILL REMAIN ANONYMOUS!) THANK YOU FOR YOUR TIME AND COOPERATION.

Age: __ __
Month / year of birth (mm/yy): __ __ / __ __

Gender:
- ☐ Male
- ☐ Female

Weight: __ __ __ lbs.
Height: __ __ ft. __ __ inches

Race or Ethnic Background:
- ☐ Native American / American Indian
- ☐ African American / Black
- ☐ White / Caucasian
- ☐ Asian (including Native Hawaiian/Other Pacific Islander)
- ☐ Hispanic
- ☐ If none of the above categories apply, please specify your race or ethnic background_____.

Marital Status: (CHECK ONLY ONE)
- ☐ Never married/single
- ☐ Never married, live with partner
- ☐ Married, live with spouse
- ☐ Separated
- ☐ Divorced
- ☐ Widowed

Employment Status: (CHECK ONLY ONE)
- ☐ Unemployed
- ☐ Full-time employee
- ☐ Part-time employee
- ☐ Homemaker, stay at home parent
- ☐ Self-employed
- ☐ Retired

Current Occupation: (CHECK THE ONE MOST APPROPRIATE)

- ☐ Manual worker
- ☐ Clerical
- ☐ Skilled /Craftsman (e.g. Carpenter, Construction Worker)
- ☐ Skilled / Office Worker
- ☐ Manager
- ☐ Professional

Tobacco use: (CHECK ALL THAT APPLY)

- ☐ Cigarettes
- ☐ Cigar
- ☐ Pipe
- ☐ Chewing tobacco

Frequency of your tobacco use: (CHECK THE ONE MOST APPROPRIATE)

- ☐ None ever
- ☐ None, past 12 months
- ☐ Less than 1 per week
- ☐ 1 to 10 per day
- ☐ 10 to 20 per day
- ☐ Greater than 20 per day

Frequency of your alcohol use: (CHECK THE ONE MOST APPROPRIATE)

- ☐ None ever
- ☐ None, past 12 months
- ☐ Less than 1 drink per week
- ☐ Less than 1 drink per day
- ☐ 1 to 5 drinks per day
- ☐ 6 to 10 drinks per day
- ☐ More than 10 drinks per day

Amount of your caffeine intake (including coffee, tea, caffeinated soda): (CHECK THE ONE MOST APPROPRIATE)

- ☐ None ever
- ☐ None, past 12 months
- ☐ Less than 1 cup per week
- ☐ Less than 1 cup per day
- ☐ 1 to 2 cups per day
- ☐ 2 to 5 cups per day
- ☐ Greater than 5 cups per day

What type of drugs have you taken at ANY time in your life? (CHECK ALL THAT APPLY)

- ☐ Tranquilizers, sleeping pills
- ☐ Antidepressants
- ☐ Anxiolytics
- ☐ Pain killers
- ☐ Anti-migraine drugs
- ☐ Anti-inflammatory drugs (e.g. aspirin, ibuprofen, acetaminophen…)
- ☐ Drugs against allergies
- ☐ Statins (cholesterol-lowering drugs)
- ☐ Drugs to treat high/low blood pressure
- ☐ Drugs to treat diabetes
- ☐ Drugs to treat thyroid disease
- ☐ Marijuana / Cannabis (grass, pot, weed, bud, Mary Jane, dope, indo, hydro)
- ☐ Amphetamine / Methamphetamine type stimulants (speed, goey, whizz, uppers, ice, glass, crystal meth Meth, poor man's cocaine)
- ☐ Cocaine (blow, nose candy, snowball, tornado, wicky stick, crack, rock)
- ☐ Ecstasy (E, Adam, XTC, eccies, the love drug, the hug drug, go, X)
- ☐ GHB (Liquid Ecstasy, Scoop, Easy Lay, Georgia Home Boy, Grievous Bodily Harm, Liquid X, and Goop)
- ☐ LSD / Hallucinogens (PCP, acid, trips, blotters, mellow, tabs)
- ☐ Inhalants (glue, solvents, aerosols)
- ☐ Methaqualone (Disco Biscuits, Lemmon 714, Lennons, Lovers, Ludes, Mandies, Mandrake, Q, Quaalude, Qualudes, Soaper, Vitamin Q)
- ☐ Heroin/morphine (smack, thunder, hell dust, big H, nose drops)
- ☐ OxyContin (Hillbilly heroin, Oxy, Oxycotton)
- ☐ Ketamine (jet, super acid, Special "K", green, K, cat Valium)
- ☐ Steroids (e.g. Prednisone, Dexamethasone, Anadrol, Oxandrin, Dianobol, Winstrol, Durabolin, Depo-Testosterone, Equipoise, other)

If you have taken drugs that are not listed above, please list them below:

_____

_____

_____

What type of drugs have you taken DURING THE LAST 3 MONTHS?
(CHECK ALL THAT APPLY)

- ☐ Tranquilizers, sleeping pills
- ☐ Antidepressants
- ☐ Anxiolytics
- ☐ Pain killers
- ☐ Anti-migraine drugs
- ☐ Anti-inflammatory drugs (e.g. aspirin, ibuprofen, acetaminophen…)
- ☐ Drugs against allergies
- ☐ Statins (cholesterol-lowering drugs)
- ☐ Drugs to treat high/low blood pressure
- ☐ Drugs to treat diabetes
- ☐ Drugs to treat thyroid disease
- ☐ Weight control pills
- ☐ Antibiotics
- ☐ Vitamins
- ☐ Marijuana/Cannabis (grass, pot, weed, bud, Mary Jane, dope, indo, hydro)
- ☐ Amphetamine / Methamphetamine type stimulants (speed, goey, whizz, uppers, ice, glass, crystal meth Meth, poor man's cocaine)
- ☐ Cocaine (blow, nose candy, snowball, tornado, wicky stick, crack, rock)
- ☐ Ecstasy (E, Adam, XTC, eccies, the love drug, the hug drug, go, X)
- ☐ GHB (Liquid Ecstasy, Scoop, Easy Lay, Georgia Home Boy, Grievous Bodily Harm, Liquid X, and Goop)
- ☐ LSD / Hallucinogens (PCP, acid, trips, blotters, mellow, tabs)
- ☐ Inhalants (glue, solvents, aerosols)
- ☐ Methaqualone (Disco Biscuits, Lemmon 714, Lennons, Lovers, Ludes, Mandies, Mandrake, Q, Quaalude, Qualudes, Soaper, Vitamin Q)
- ☐ Heroin/morphine (smack, thunder, hell dust, big H, nose drops)
- ☐ OxyContin (Hillbilly heroin, Oxy, Oxycotton)
- ☐ Ketamine (jet, super acid, Special "K", green, K, cat Valium)
- ☐ Steroids (e.g. Prednisone, Dexamethasone, Anadrol, Oxandrin, Dianobol, Winstrol, Durabolin, Depo-Testosterone, Equipoise, other)

If you have taken drugs that were not listed above, please list them below:

_____
_____
_____

Please answer the following questions regarding your general medical history
Indicate all that you EVER experienced in your life: (CHECK ALL THAT APPLY)

- ☐ Chronic Pain
- ☐ Chronic inflammation (Ulcerative colitis, Crohn's disease, Rheumatiod arthritis)
- ☐ Cardiovascular disorder
- ☐ Diabetes
- ☐ Regular Headaches, migraines
- ☐ Sexually transmitted diseases (e.g. clamydia, gonorrhea, HPV, HIV, Syphilis, Genital Herpes)
- ☐ Other chronic infections
- ☐ Allergies
- ☐ Chronic gastric problems
- ☐ Tumors/Cancer, specify

_____
_____

- ☐ Other, please specify

_____
_____
_____

Current medical problems: Indicate all that you are CURRENTLY experiencing or being treated for: (CHECK ALL THAT APPLY)

- ☐ Chronic Pain
- ☐ Chronic inflammation (Ulcerative colitis, Crohn's disease, Rheumatoid arthritis)
- ☐ Cardiovascular disorder
- ☐ Diabetes
- ☐ Gastric problems
- ☐ Headaches, migraines
- ☐ Sexually transmitted diseases (e.g. clamydia, gonorrhea, HPV, HIV, Syphilis, Genital Herpes)
- ☐ Other chronic infections
- ☐ Allergies
- ☐ Tumors/Cancer, specify
- ☐ Other, please specify

_____
_____
_____

Have you ever been hospitalized?

- ☐ No
- ☐ Yes

If yes, how often in last 12 months, please provide reason(s)

_____
_____
_____

Have you ever had surgery?

☐ No
☐ Yes

If yes, please provide reason(s) and year the surgery took place

_____ / _____ (yy)
_____ / _____ (yy)
_____ / _____ (yy)
_____ / _____ (yy)
_____ / _____ (yy)

When did you last see a doctor? (CHECK ONE)

☐ In last 4 weeks
☐ In last 6 months
☐ In last year
☐ In last 5 years
☐ In last 10 years
☐ More than 10 years ago
☐ Never

Indicate if you have experienced any of the following during the LAST 2 WEEKS and the frequency in which you experienced it:

☐ **Feeling low:**
☐ Every day ☐ Most days ☐ Sometimes ☐ Never

☐ **Lack of energy:**
☐ Every day ☐ Most days ☐ Sometimes ☐ Never

☐ **Less interest in things or unable to enjoy things you used to enjoy:**
☐ Every day ☐ Most days ☐ Sometimes ☐ Never

☐ **Difficulties concentrating:**
☐ Every day ☐ Most days ☐ Sometimes ☐ Never

☐ **Sleep problems, e.g. problems falling asleep, problems staying asleep, early morning awakening:**
☐ Every day ☐ Most days ☐ Sometimes ☐ Never

☐ **Anxiety:**
☐ Every day ☐ Most days ☐ Sometimes ☐ Never

☐ **Cannot cope with daily problems, suicide considered:**
☐ Every day ☐ Most days ☐ Sometimes ☐ Never

Indicate if you have experienced changes in any of the following and the level of change:

☐ **Appetite:**
- ☐ Increased
- ☐ Decreased
- ☐ Unchanged
- ☐ Decreased (unintentionally)

☐ **Weight:**
- ☐ Increased
- ☐ Decreased
- ☐ Unchanged
- ☐ Decreased (unintentionally)

☐ **Sexual Interest:**
- ☐ Increased
- ☐ Decreased
- ☐ Unchanged

Have you EVER experienced ANY of the following? (CHECK ALL THAT APPLY)

- ☐ Severe depression, severe mania
- ☐ Panic attacks
- ☐ Severe anxiety
- ☐ Severe obsessive or compulsive thoughts
- ☐ Alcohol abuse
- ☐ Substance abuse other than alcohol
- ☐ Psychotic episodes

Have you EVER been treated for ANY of the following? (CHECK ALL THAT APPLY)

- ☐ Depression, Mania
- ☐ Panic attacks
- ☐ Anxiety disorder
- ☐ Obsessive-compulsive disorder
- ☐ Alcohol abuse
- ☐ Substance abuse other than alcohol
- ☐ Psychotic episodes

Has anyone in your family EVER experienced or been treated for one of the following disorder(s)? (PLEASE CHECK ALL THAT APPLY, THEN SELECT WHICH MEMBER OF THE FAMILY EXPERIENCED OR WAS TREATED FOR THE DISORDER).

☐ **Depression:**

☐ Mother	☐ Daughter	☐ Sister
☐ Father	☐ Son	☐ Brother
☐ Grandmother	☐ Aunt	☐ Granddaughter
☐ Grandfather	☐ Uncle	☐ Grandson

☐ **Anxiety:**

☐ Mother	☐ Daughter	☐ Sister
☐ Father	☐ Son	☐ Brother
☐ Grandmother	☐ Aunt	☐ Granddaughter
☐ Grandfather	☐ Uncle	☐ Grandson

☐ **Alcohol abuse:**

☐ Mother	☐ Daughter	☐ Sister
☐ Father	☐ Son	☐ Brother
☐ Grandmother	☐ Aunt	☐ Granddaughter
☐ Grandfather	☐ Uncle	☐ Grandson

☐ **Other substance abuse:**

☐ Mother	☐ Daughter	☐ Sister
☐ Father	☐ Son	☐ Brother
☐ Grandmother	☐ Aunt	☐ Granddaughter
☐ Grandfather	☐ Uncle	☐ Grandson

☐ **Schizophrenia/psychosis:**

☐ Mother	☐ Daughter	☐ Sister
☐ Father	☐ Son	☐ Brother
☐ Grandmother	☐ Aunt	☐ Granddaughter
☐ Grandfather	☐ Uncle	☐ Grandson

☐ **Suicide:**

☐ Mother	☐ Daughter	☐ Sister
☐ Father	☐ Son	☐ Brother
☐ Grandmother	☐ Aunt	☐ Granddaughter
☐ Grandfather	☐ Uncle	☐ Grandson

☐ **Dementia:**

☐ Mother	☐ Daughter	☐ Sister
☐ Father	☐ Son	☐ Brother
☐ Grandmother	☐ Aunt	☐ Granddaughter
☐ Grandfather	☐ Uncle	☐ Grandson

Has anyone in your family EVER experienced or been treated for one of the following disorder(s)? (PLEASE CHECK ALL THAT APPLY, THEN SELECT WHICH MEMBER OF THE FAMILY EXPERIENCED OR WAS TREATED FOR THE DISORDER).

☐ **Amyotrophic Lateral Sclerosis:**

☐ Mother	☐ Daughter	☐ Sister
☐ Father	☐ Son	☐ Brother

☐ **Parkinson's disease:**

☐ Mother	☐ Daughter	☐ Sister
☐ Father	☐ Son	☐ Brother

☐ **Multiple Sclerosis:**

☐ Mother	☐ Daughter	☐ Sister
☐ Father	☐ Son	☐ Brother

☐ **Huntington's Chorea:**

☐ Mother	☐ Daughter	☐ Sister
☐ Father	☐ Son	☐ Brother

Indicate your exercise frequency and duration:

☐ **Daily**
☐ Less than 30 min ☐ 30 min to 1 hour ☐ Greater than 1 hour

☐ **2-3 x/week**
☐ Less than 30 min ☐ 30 min to 1 hour ☐ Greater than 1 hour

☐ **Once per week**
☐ Less than 30 min ☐ 30 min to 1 hour ☐ Greater than 1 hour

☐ **Once per month**
☐ Less than 30 min ☐ 30 min to 1 hour ☐ Greater than 1 hour

☐ **Never**

Please indicate your normal daily rhythm: (CHECK THE MOST APPROPRIATE)

☐ Rotating/split shift worker
☐ Normal bedtime after midnight
☐ Normal bedtime before midnight

Prior or current stressful live events can influence some blood parameters. Please Indicate if any of the following events happened in your life BEFORE AGE 15. (CHECK ALL THAT APPLY)

☐ Death of both parents
☐ Death of one parent
☐ Divorce of parents
☐ Death of a close family member
☐ Major personal injury or illness (e.g. cancer)
☐ Death of a close friend
☐ Gain of new family member (e.g. new sibling)
☐ Major change in health or behavior of a close family member
☐ Major change in finances
☐ Major change in living conditions (e.g. change in residence, schools)

Indicate if any of the following events happened in the LAST 12 MONTHS. (CHECK ALL THAT APPLY)

☐ Death of spouse/partner
☐ Divorce
☐ Death of a close family member
☐ Marital separation/separation from partner
☐ Fired from work
☐ Major personal injury or illness
☐ Jail term
☐ Death of a close friend
☐ Pregnancy (including miscarriage or abortion)
☐ Gain of a new family member (e.g. birth of a child)
☐ Major change in health or behavior of a close family member
☐ Major change in finances

- ☐ Retirement
- ☐ Change to a different line of work
- ☐ Marriage
- ☐ Child leaving home
- ☐ Major change in living conditions (e.g. change in residence)
- ☐ Outstanding personal achievement
- ☐ Minor violations of the law

Please provide the time of your last 3 meals:

__ __: __ __ am / pm (circle one)
__ __: __ __ am / pm (circle one)
__ __: __ __ am / pm (circle one)

FOR WOMEN ONLY

Are your menstrual cycles: (CHECK ONE)

- ☐ Regular?
- ☐ Irregular?

Date of your last menstrual period: _____/_____ (dd/mm)

Are your currently pregnant? (CHECK ONE)

- ☐ Yes
- ☐ No

Indicate the number of past pregnancies: (CHECK ONE)

- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ Greater than 5

Did you give birth in the past year? (CHECK ONE)

- ☐ Yes
- ☐ No

Are you currently breastfeeding? (CHECK ONE)

- ☐ Yes
- ☐ No

Thank you for your cooperation!

Appendix 5: Coding table with clinical variables and covariates

The table lists the coding of clinical variables and covariates from the US ABS questionnaire. The same coding was applied to the first cohort of borderline disorder patients.

Scoring questionnaire variables for ABS and borderline samples		
Variable	**Score**	**Comment**
age	as reported on questionnaire	
gender	as reported on questionnaire	
BMI	calculated from height and weight	
alcohol quantity score	< 1 drink per day = 0, > 1 drink per day = 1	
tobacco frequency score	< 1 per week = 0, > 1 per day = 1	
feeling low score	never = 0, sometimes = 1, most days = 2, every day = 3	
enjoyment score	never = 0, sometimes = 1, most days = 2, every day = 3	
sleep problems score	never = 0, sometimes = 1, most days = 2, every day = 3	
anxiety score	never = 0, sometimes = 1, most days = 2, every day = 3	
concentration score	never = 0, sometimes = 1, most days = 2, every day = 3	
sexual interest score	unchanged = 0, increased = 2, decreased = 3	
energy score	never = 0, sometimes = 1, most days = 2, every day = 3	
coping score	never = 0, sometimes = 1, most days = 2, every day = 3	
appetite change score	unchanged = 0, decreased = 0, increased = 2, decreased unintentionally = 3	people that change eating habits on purpose need to be scored differently than people who change unintentionally.
weight change score	unchanged = 0, decreased = 0, increased = 2, decreased unintentionally = 3	people who lose weight on purpose need to be scored differently than people who lose weight unintentionally.
Lifetime treatment	no personal treatments for depression or anxiety = 0, any treatments = 1	mostly depression/anxiety but also includes some alcohol or substance abuse.
Lifetime drug use	no drugs or only prescription drug use = 0, use of any drugs of abuse = 1	on the questionnaire the "drugs of abuse" start with marijuana and end with ketamine
3 month drugs	no drugs or "harmless" drugs = 0, prescription drugs only = 1, drugs of abuse = 2	"harmless" drugs are allergy meds, weight pills, vitamins, NSAID, antibiotics, prescription drugs are antidepressants, pain killers, diabetes drugs, etc

Lifetime experiences	no personal episodes of depression , anxiety, panic = 0, any episodes = 1	mostly depression/anxiety but also includes some alcohol or substance abuse.
Early life stress score	sum of boxes checked for stressful events before the age of 15. The top item (death of both parents) has a value of 20 and the bottom item in the list (major change in living conditions) has a value of 11.	also called ELS in tables, scoring adapted from the literature (105).
Recent stress score	sum of boxes checked for stressful events experienced in the past 12 months. The top item in the list (death of spouse) has a value of 20 and the bottom item in the list (minor violations of the law) has a value of 2.	also called RS in tables, scoring adapted from the literature (105).
Symptom score sum	sum of scores for 10 symptoms (feeling low, energy, interest, concentration, sleep problems, anxiety, coping, appetite change, weight change, sex interest	lowest score is 0 and highest possible is 30, sometimes we also created a 7 symptom score which does NOT include appetite, weight and sex.
Depression / Anxiety / suicide (family history)	no relatives with any disease = 0, any secondary relative with any of the diseases = 1, any primary relative with any of the diseases = 2	Secondary relative is uncle, aunt, grandparent, grandchild; primary relatives are mother, father, children, sibling; there is no consideration for the number of relatives affected.
alcohol abuse (family)	no relatives with any disease = 0, any secondary relative with any of the diseases = 1, any primary relative with any of the diseases = 2	sometimes we combine alcohol and substance abuse together but the scoring is the same method
schizophrenia / psychosis (family)	no relatives with any disease = 0, any secondary relative with any of the diseases = 1, any primary relative with any of the diseases = 2	
Substance abuse (family)	no relatives with any disease = 0, any secondary relative with any of the diseases = 1, any primary relative with any of the diseases = 2	sometimes we combine alcohol and substance abuse together but the scoring is the same method
vegetative symptom score	0 = no symptoms, 1 = some symptoms, 2 = more symptoms, 3 = most symptoms	this score combines three symptoms often associated with melancholic depression (weight loss, appetite loss, and sleep problems). The scoring is shown below.

	Scoring mechanism: 1) score the "Weight change" category as follows: unchanged = 0 increased = 1 decreased unintent = -1 decreased intent = 1 2) score the "appetite change" category as follows: unchanged = 0 increased = 1 decreased unintent = -1 decreased intent = 1 3) sum these two scores then convert as follows: all zeros stay zero all positive values become zero all negative values switch sign 4) score the "sleep problems" category as follows: never = 0 all others = 1 5) add the values for steps 3 and 4 (the range is zero to 3) (basically looking for weight loss, appetite loss, and sleep disturbances)

Appendix 6: Simulation study – phase 1 tasks

Below are tables showing the various tasks performed in phase 2 together with file names explained below.

Run	Gene file name	Response file name	Number of contributing variables	Data size	Correlation among explanatory variables	Inclusion criteria	Gene-gene interaction
1	Gene00100.csv	Response1.csv	1	100	0	Threshold	-
2	Gene001000.csv	Response2.csv	1	1000	0	Threshold	-
3	Gene00100.csv	Response3.csv	1	100	0	Interval	-
4	Gene001000.csv	Response4.csv	1	1000	0	Interval	-
5	Gene00100.csv	Response5.csv	2	100	0	Threshold	Sum
6	Gene001000.csv	Response6.csv	2	1000	0	Threshold	Sum
7	Gene00100.csv	Response7.csv	2	100	0	Interval	Sum
8	Gene001000.csv	Response8.csv	2	1000	0	Interval	Sum
9	Gene00100.csv	Response9.csv	2	100	0	Threshold	Product
10	Gene001000.csv	Response10.csv	2	1000	0	Threshold	Product
11	Gene00100.csv	Response11.csv	2	100	0	Interval	Product
12	Gene001000.csv	Response12.csv	2	1000	0	Interval	Product
13	Gene00100.csv	Response13.csv	2	100	0	Threshold	Ratio
14	Gene001000.csv	Response14.csv	2	1000	0	Threshold	Ratio
15	Gene00100.csv	Response15.csv	2	100	0	Interval	Ratio
16	Gene001000.csv	Response16.csv	2	1000	0	Interval	Ratio
17	Gene05100.csv	Response17.csv	2	100	0.5	Threshold	Sum
18	Gene051000.csv	Response18.csv	2	1000	0.5	Threshold	Sum
19	Gene05100.csv	Response19.csv	2	100	0.5	Interval	Sum
20	Gene051000.csv	Response20.csv	2	1000	0.5	Interval	Sum
21	Gene05100.csv	Response21.csv	2	100	0.5	Threshold	Product
22	Gene051000.csv	Response22.csv	2	1000	0.5	Threshold	Product
23	Gene05100.csv	Response23.csv	2	100	0.5	Interval	Product
24	Gene051000.csv	Response24csv	2	1000	0.5	Interval	Product
25	Gene05100.csv	Response25.csv	2	100	0.5	Threshold	Ratio
26	Gene051000.csv	Response26.csv	2	1000	0.5	Threshold	Ratio
27	Gene05100.csv	Response27.csv	2	100	0.5	Interval	Ratio
28	Gene051000.csv	Response28.csv	2	1000	0.5	Interval	Ratio

Separate Studies

Different magnitudes of 2 contributing variables

Run	Gene file name	Response file name	Number of contributing variables	Data size	Special feature	Inclusion Criteria	Gene-gene interaction
7	Gene00100.csv	Response7.csv	2	100	$X_1 \approx X_2$	Interval	Sum
30	Gene00100a.csv	Response30.csv	2	100	$X_2 \approx 10 \cdot X_1$	Interval	Sum
31	Gene00100b.csv	Response31.csv	2	100	$X_2 \approx 100 \cdot X_1$	Interval	Sum

Fraction of data points being classified as Y = 1

Run	Gene file name	Response file names	Number of con. variables	Data size	Special feature	Inclusion Criteria	Gene-gene interaction
32	Genefrac05100.csv	Response32.csv	1	100	$\frac{\#\{X_1 \geq 0\}}{100} = 0.05$	Threshold	-
33	Genefrac20100.csv	Response33.csv	1	100	$\frac{\#\{X_1 \geq 0\}}{100} = 0.20$	Threshold	-
34	Genefrac50100.csv	Response34.csv	1	100	$\frac{\#\{X_1 \geq 0\}}{100} = 0.50$	Threshold	-
35	Genefrac051000.csv	Response35.csv	1	1000	$\frac{\#\{X_1 \geq 0\}}{1000} = 0.05$	Threshold	-
36	Genefrac201000.csv	Response36.csv	1	1000	$\frac{\#\{X_1 \geq 0\}}{1000} = 0.20$	Threshold	-
37	Genefrac501000.csv	Response37.csv	1	1000	$\frac{\#\{X_1 \geq 0\}}{1000} = 0.50$	Threshold	-

2 populations with different mean values in gene no. 1

Run	Gene file name	Response file names	Number of con. variables	Data size	Inclusion Criteria	Gene-gene interaction
38	Mydif6.csv	Response38.csv	1	100	$Y = 1 \; if \; X_1 \sim N(-3,1)$ $Y = 0 \; if \; X_1 \sim N(+3,1)$	-
39	Mydif4.csv	Response38.csv	1	100	$Y = 1 \; if \; X_1 \sim N(-2,1)$ $Y = 0 \; if \; X_1 \sim N(+2,1)$	-
40	Mydif2.csv	Response38.csv	1	100	$Y = 1 \; if \; X_1 \sim N(-1,1)$ $Y = 0 \; if \; X_1 \sim N(+1,1)$	-
41	Mydif1.csv	Response38.csv	1	100	$Y = 1 \; if \; X_1 \sim N(-0.5,1)$ $Y = 0 \; if \; X_1 \sim N(+0.5,1)$	-
42	Mydif05.csv	Response38.csv	1	100	$Y = 1 \; if \; X_1 \sim N(-0.25,1)$ $Y = 0 \; if \; X_1 \sim N(+0.25,1)$	-

Overview of gene files

Gene file name	Distribution
Gene00100.csv	$X \sim N_{30}(0, I)$; data points=100
Gene001000.csv	$X \sim N_{30}(0, I)$;data points=1000
Gene05100.csv	$X \sim N_{30}(0, \Sigma); \sigma_{ii} = 1; \sigma_{ij} = 0.5$; data points=100
Gene051000.csv	$X \sim N_{30}(0, \Sigma); \sigma_{ii} = 1; \sigma_{ij} = 0.5$ data points=1000

Gene00100a.csv	$X_1 \sim N(1,1);\ X_2 \sim N(10,1);$ $X_i \sim N(0,1)\ \ for\ i = 3-30.$ data points=100
Gene00100b.csv	$X_1 \sim N(1,1);\ X_2 \sim N(100,1);$ $X_i \sim N(0,1)\ \ for\ i = 3-30.$ data points=100
Genefrac05100.csv	100 data points: $X_i \sim N(0,1); i = 1-30$ $\frac{\#\{X_1 \geq 0\}}{100} = 0.05$
Genefrac20100.csv	100 data points: $X_i \sim N(0,1); i = 1-30;$ $\frac{\#\{X_1 \geq 0\}}{100} = 0.20$
Genefrac50100.csv	100 data points: $X_i \sim N(0,1); i = 1-30$ $\frac{\#\{X_1 \geq 0\}}{100} = 0.50$
Genefrac051000.csv	1000 data points: $X_i \sim N(0,1); i = 1-30$ $\frac{\#\{X_1 \geq 0\}}{1000} = 0.05$
Genefrac201000.csv	1000 data points: $X_i \sim N(0,1); i = 1-30;$ $\frac{\#\{X_1 \geq 0\}}{1000} = 0.20$
Genefrac501000.csv	1000 data points: $X_i \sim N(0,1); i = 1-30$ $\frac{\#\{X_1 \geq 0\}}{1000} = 0.50$
Mydif6.csv	100 data points: $X_1^j \sim N(-3,1);\ j = 1-50$ $X_1^j \sim N(+3,1);\ j = 51-100$ $X_i \sim N(0,1); i = 2-30$
Mydif4.csv	100 data points: $X_1^j \sim N(-2,1);\ j = 1-50$ $X_1^j \sim N(+2,1);\ j = 51-100$ $X_i \sim N(0,1); i = 2-30$
Mydif2.csv	100 data points: $X_1^j \sim N(-1,1);\ j = 1-50$ $X_1^j \sim N(+1,1);\ j = 51-100$ $X_i \sim N(0,1); i = 2-30$
Mydif1.csv	100 data points:

	$X_1^j \sim N(-0.5,1);\ j = 1-50$ $X_1^j \sim N(+0.5,1);\ j = 51-100$ $X_i \sim N(0,1); i = 2-30$
Mydif05.csv	100 data points: $X_1^j \sim N(-0.25,1);\ j = 1-50$ $X_1^j \sim N(+0.25,1);\ j = 51-100$ $X_i \sim N(0,1); i = 2-30$

Overview of response files

Response file name	**Inclusion criteria**
Response1.csv Response2.csv Response32.csv Response33.csv Response34.csv Response35.csv Response36.csv Response37.csv	$X_1 \geq 0 \Rightarrow Y = 1$ $else\ Y = 0$
Response3.csv Response4.csv	$-0.67 \leq X_1 \leq 0.67 \Rightarrow Y = 1$ $else\ Y = 0$
Response5.csv Response6.csv Response17.csv Response18.csv	$X_1 + X_2 \geq 0 \Rightarrow Y = 1$ $else\ Y = 0$
Response7.csv Response8.csv Response19.csv Response20.csv	$-0.95 \leq X_1 + X_2 \leq 0.95 \Rightarrow Y = 1$ $else\ Y = 0$
Response9.csv Response10.csv Response21.csv Response22.csv	$X_1 \cdot X_2 \geq 0 \Rightarrow Y = 1$ $else\ Y = 0$
Response11.csv Response12.csv Response23.csv Response24.csv	$-0.4 \leq X_1 \cdot X_2 \leq 0.4 \Rightarrow Y = 1$ $else\ Y = 0$
Response13.csv Response14.csv Response25.csv Response26.csv	$\frac{X_1}{X_2} \geq 0 \Rightarrow Y = 1$ $else\ Y = 0$
Response15.csv Response16.csv Response27.csv Response28.csv	$-1 \leq \frac{X_1}{X_2} \leq 1 \Rightarrow Y = 1$ $else\ Y = 0$
Response30.csv	$X_1 + X_2 \geq 11 \Rightarrow Y = 1$ $else\ Y = 0$

Response31.csv	$X_1 + X_2 \geq 101 \Rightarrow Y = 1$ $else\ Y = 0$
Response38.csv	$Y_i = 1\ if\ 1 \leq i \leq 50$ $Y_i = 0\ if\ 51 \leq i \leq 100$
Response105.csv Response117.csv (N=100)	$X_1 + X_2 - X_3 - X_4 + X_5 \geq 0 \Rightarrow Y = 1$ $else\ Y = 0$
Response107.csv Response119.csv (N=100)	$-0.95 \leq X_1 + X_2 - X_3 - X_4 + X_5 \leq 0.95 \Rightarrow Y = 1$ $else\ Y = 0$
Response109.csv Response121.csv (N=100)	$X_1 \cdot X_2 \cdot X_3 \cdot X_4 \cdot X_5 \geq 0 \Rightarrow Y = 1$ $else\ Y = 0$
Response111.csv Response123.csv (N=100)	$-0.4 \leq X_1 \cdot X_2 \cdot X_3 \cdot X_4 \cdot X_5 \leq 0.4 \Rightarrow Y = 1$ $else\ Y = 0$
Response113.csv Response125.csv (N=100)	$\frac{X_1 + X_2}{X_3 - X_4 + X_5} \geq 0 \Rightarrow Y = 1$ $else\ Y = 0$
Response115.csv Response127.csv (N=100)	$-1 \leq \frac{X_1 + X_2}{X_3 - X_4 + X_5} \leq 1 \Rightarrow Y = 1$ $else\ Y = 0$

Appendix 7: Simulation study – phase 2 tasks

Below are tables showing the various tasks performed in phase 2 together with file names explained below.

Run	Gene file name	z-score standardized data	Response file name	Number of contributing variables	Data size	Inclusion criteria	Gene-gene interaction
1	allcombined_z.csv	Yes	Response1.csv	1	263	Threshold	-
2	allcombined_z.csv	Yes	Response2.csv	1	263	Interval	-
3	allcombined_z.csv	Yes	Response3.csv	2	263	Threshold	Sum
4	allcombined_z.csv	Yes	Response4.csv	2	263	Interval	Sum
5	allcombined_z.csv	Yes	Response5.csv	2	263	Threshold	Product
6	allcombined_z.csv	Yes	Response6.csv	2	263	Interval	Product
7	allcombined_z.csv	Yes	Response7.csv	2	263	Threshold	Ratio
8	allcombined_z.csv	Yes	Response8.csv	2	263	Interval	Ratio
9	allcombined_z.csv	Yes	Response9.csv	5	263	Threshold	Sum
10	allcombined_z.csv	Yes	Response10.csv	5	263	Interval	Sum
11	allcombined_z.csv	Yes	Response11.csv	5	263	Threshold	Product
12	allcombined_z.csv	Yes	Response12.csv	5	263	Interval	Product
13	allcombined_z.csv	Yes	Response13.csv	5	263	Threshold	Ratio
14	allcombined_z.csv	Yes	Response14.csv	5	263	Interval	Ratio

Separate studies

Random y

Run	Gene file name	z-score standardized data	Response file name	Number of contributing variables	Data size	Inclusion criteria	Gene-gene interaction
15	allcombined_z.csv	Yes	Response15.csv	0	263	-	-

Fraction of data points being classified as Y = 1

Run	Gene file name	z-score standardized data	Response file name	Number of contributing variables	Special feature	Data size	Inclusion criteria
16	allcombined_z.csv	Yes	Response16.csv	1	$\frac{\#\{X_1 \geq \alpha_1\}}{100} = 0.05$	263	Threshold
17	allcombined_z.csv	Yes	Response17.csv	1	$\frac{\#\{X_1 \geq \alpha_2\}}{100} = 0.20$	263	Threshold
18	allcombined_z.csv	Yes	Response18.csv	1	$\frac{\#\{X_1 \geq \alpha_3\}}{100} = 0.50$	263	Threshold

Actual data

Run	Gene file name	z-score data	Response file name	Number of contributing variables	Special feature	Data size	Inclusion criteria	Gene-gene interaction
19	allcombined.csv	No	Response19.csv	2	Different magnitudes	263	Threshold	Sum
20	allcombined.csv	No	Response20.csv	2	Equal magnitudes	263	Threshold	Sum
21	allcombined.csv	No	Response21.csv	2	Different magnitudes	263	Threshold	Ratio
22	allcombined.csv	No	Response22.csv	2	Equal magnitudes	263	Threshold	Ratio
23	allcombined.csv	No	Response23.csv	2	Different magnitudes	263	Threshold	Product
24	allcombined.csv	No	Response24.csv	2	Equal magnitudes	263	Threshold	Product
25	allcombined.csv	No	Response25.csv	2	Different magnitudes	263	Interval	Sum
26	allcombined.csv	No	Response26.csv	2	Equal magnitudes	263	Interval	Sum

3 groups/classes

Run	Gene file name	z-score standardized data	Response file name	Number of contributing variables	Data size	Inclusion criteria	Gene-gene interaction
27	allcombined.csv	Yes	Response27.csv	1	263	Interval	-
28	allcombined.csv	Yes	Response28.csv	2	263	Interval	Sum
29	allcombined.csv	Yes	Response29.csv	2	263	Interval	Ratio
30	allcombined.csv	Yes	Response30.csv	2	263	Interval	Product
31	allcombined.csv	Yes	Response31.csv	5	263	Interval	Sum
32	allcombined.csv	Yes	Response32.csv	5	263	Interval	Ratio
33	allcombined.csv	Yes	Response33.csv	5	263	Interval	Product

Overview of gene files

Gene file name	Description
allcombined_z.csv	Data set consisting of DC+SH ABS + BP + PTSD controls + PTSD acute, 7 HKG, 25 genes; z-score standardized data; N=263 (variables in columns, samples in rows)
allcombined.csv	Data set consisting of DC+SH ABS + BP + PTSD controls + PTSD acute, 7 HKG, 25 genes; N=263 (variables in columns, samples in rows)

Overview of response files

Response file name	Inclusion criteria
Response1.csv	$X_1 \geq 0 \Rightarrow Y = 1, \quad else Y = 0$
Response2.csv	$-0.67 \leq X_1 \leq 0.67 \Rightarrow Y = 1, \quad else Y = 0$

Response3.csv	$X_2 + X_3 \geq 0 \Rightarrow Y = 1, \;\; else Y = 0$
Response4.csv	$-0.95 \leq X_2 + X_3 \leq 0.95 \Rightarrow Y = 1, \;\; else Y = 0$
Response5.csv	$X_4 \cdot X_5 \geq 0 \Rightarrow Y = 1, \;\; else Y = 0$
Response6.csv	$-0.4 \leq X_4 \cdot X_5 \leq 0.4 \Rightarrow Y = 1, \;\; else Y = 0$
Response7.csv	$\frac{X_6}{X_7} \geq 0 \Rightarrow Y = 1, \;\; else Y = 0$
Response8.csv	$-1 \leq \frac{X_6}{X_7} \leq 1 \Rightarrow Y = 1, \;\; else Y = 0$
Response9.csv	$X_8 + X_9 - X_{10} - X_{11} + X_{12} \geq 0 \Rightarrow Y = 1, \;\; else Y = 0$
Response10.csv	$-0.95 \leq X_8 + X_9 - X_{10} - X_{11} + X_{12} \leq 0.95 \Rightarrow Y = 1, \;\; else Y = 0$
Response11.csv	$X_{13} \cdot X_{14} \cdot X_{15} \cdot X_{16} \cdot X_{17} \geq 0 \Rightarrow Y = 1, \;\; else Y = 0$
Response12.csv	$-0.4 \leq X_{13} \cdot X_{14} \cdot X_{15} \cdot X_{16} \cdot X_{17} \leq 0.4 \Rightarrow Y = 1, \;\; else Y = 0$
Response13.csv	$\frac{X_{18} + X_{19}}{X_{20} - X_{21} + X_{22}} \geq 0 \Rightarrow Y = 1, \;\; else Y = 0$
Response14.csv	$-1 \leq \frac{X_{18} + X_{19}}{X_{20} - X_{21} + X_{22}} \leq 1 \Rightarrow Y = 1, \;\; else Y = 0$
Response15.csv	Random y; no combination of any variables
Response16.csv	$\frac{\#\{X_1 \geq \alpha_1\}}{100} = 0.05$, $\alpha_1 = 1.7 * st.dev(X_1)$ which yields ~ 5% Y=1
Response17.csv	$\frac{\#\{X_1 \geq \alpha_2\}}{100} = 0.20$, $\alpha_2 = 0.8 * st.dev(X_1)$ which yields ~ 20% Y=1
Response18.csv	$\frac{\#\{X_1 \geq \alpha_3\}}{100} = 0.50$, $\alpha_3 = -0.15 * st.dev(X_1)$ which yields ~ 50% Y=1
Response19.csv	$X_{23} + X_{24} \geq (\mu(X_{23}) + \mu(X_{24})) \Rightarrow Y = 1, \;\; else Y = 0$. NB! $X_{23} \approx 150 * X_{24}$
Response20.csv	$X_{24} + X_{25} \geq (\mu(X_{24}) + \mu(X_{25})) \Rightarrow Y = 1, \;\; else Y = 0$. NB! $X_{24} \approx X_{25}$
Response21.csv	$\frac{X_{23}}{X_{24}} \geq \frac{\mu(X_{23})}{\mu(X_{24})} \Rightarrow Y = 1, \;\; else Y = 0$. NB! $X_{23} \approx 150 * X_{24}$
Response22.csv	$\frac{X_{24}}{X_{25}} \geq \frac{\mu(X_{24})}{\mu(X_{25})} \Rightarrow Y = 1, \;\; else Y = 0$. NB! $X_{24} \approx X_{25}$
Response23.csv	$X_{23} \cdot X_{24} \geq \mu(X_{23}) \cdot \mu(X_{24}) \Rightarrow Y = 1, \;\; else Y = 0$ NB! $X_{23} \approx 150 * X_{24}$
Response24.csv	$X_{24} \cdot X_{25} \geq \mu(X_{24}) \cdot \mu(X_{25}) \Rightarrow Y = 1, \;\; else Y = 0$.

	NB! $X_{24} \approx X_{25}$
Response25.csv	$0.5 \cdot (\mu(X_{23}) + \mu(X_{24})) \le X_{23} + X_{24} \le 1.5 \cdot (\mu(X_{23}) + \mu(X_{24})) \Rightarrow Y = 1, \quad else\, Y = 0$. NB! $X_{23} \approx 150 * X_{24}$
Response26.csv	$0.5 \cdot (\mu(X_{24}) + \mu(X_{25})) \le X_{24} + X_{25} \le 1.5 \cdot (\mu(X_{24}) + \mu(X_{25})) \Rightarrow Y = 1, \quad else\, Y = 0$ NB! $X_{24} \approx X_{25}$
Response27.csv	$X_1 \le -0.5 \Rightarrow Y = 1, \quad X_1 \ge 0.5 \Rightarrow Y = 2, \quad else\, Y = 0$
Response28.csv	$X_1 + X_2 \le -1 \Rightarrow Y = 1, \quad X_1 + X_2 \ge 1 \cdot \Rightarrow Y = 2, \quad else\, Y = 0$
Response29.csv	$\frac{X_3}{X_4} \le -0.75 \Rightarrow Y = 1, \quad \frac{X_3}{X_4} \ge 0.75 \Rightarrow Y = 2, \quad else\, Y = 0$
Response30.csv	$X_5 \cdot X_6 \le -0.25 \Rightarrow Y = 1, \quad X_5 \cdot X_6 \ge 0.25 \Rightarrow Y = 2, \quad else\, Y = 0$
Response31.csv	$X_7 + X_8 - X_9 - X_{10} + X_{11} \le -0.5 \Rightarrow Y = 1, \quad X_7 + X_8 - X_9 - X_{10} + X_{11} \ge 0.5 \Rightarrow Y = 2$ $else\, Y = 0$
Response32.csv	$\frac{X_{12} + X_{13}}{X_{14} - X_{15} + X_{16}} \le -0.75 \Rightarrow Y = 1,$ $\frac{X_{12} + X_{13}}{X_{14} - X_{15} + X_{16}} \ge 0.75 \Rightarrow Y = 2$ $else\, Y = 0$
Response33.csv	$X_{17} \cdot X_{18} \cdot X_{19} \cdot X_{20} \cdot X_{21} \le 0.05 \Rightarrow Y = 1,$ $X_{17} \cdot X_{18} \cdot X_{19} \cdot X_{20} \cdot X_{21} \ge 0.05 \Rightarrow Y = 2$ $else\, Y = 0$

Appendix 8: BD associated genes according to WTCC and Baum

In the table below, the 68 genes putatively associated with BD according to the WTCC study and according to the Baum article, are listed.

AK3L1	DTNBP1	LOC283547	SOX5
AK3L2	EARS2	LOC730018	SVEP1
AKAP10	ERN2	LOC731264	SYK
AOF1	ESRRG	LOC731914	SYN3
BDNF	FAM126A	LRRC7	SYNE1
C14orf58	GABRB1	MYH9	TBC1D21
CAPN6	GALNTL4	NDUFAB1	TDRD9
CDC25B	GGA2	NPAS3	THRB
CMTM8	GRIK2	NRG1	THSD7A
COG7	GRIN2B	NXN	TRDN
CSF2RB	GRM3	PALB2	UBPH/UBFD1
DAOA	GRM4	PAX5	VGCNL1
DCTN5	GRM7	PLK1	ZBTB44
DFNB31	KCNC2	PTPRE	ZNF274
DGKH	KCNQ3	PTPRG	ZNF490
DISC1	KLHDC1	RNPEPL1	ZNF659
DPP10	LAMP3	SORCS2	ZNF678

Appendix 9: Gene ratios

In the table below, 97 gene ratios are listed. They are formed partly on a biological basis, partly on a data driven basis; high (>0.8) and low (<0.3) Spearman correlations.

Ratio 1	IDO/SERT	Ratio 49	ERK1/ERK2
Ratio 2	ERK1/MAPK8	Ratio 50	ARRB1/MAPK8
Ratio 3	ERK1/MAPK14	Ratio 51	ARRB1/P2X7
Ratio 4	ERK2/MAPK8	Ratio 52	MAPK14/IL-1 beta
Ratio 5	ERK2/MAPK14	Ratio 53	PREP/SA100A10
Ratio 6	(ERK1+2)/MAPK8	Ratio 54	S100A10/P2X7
Ratio 7	(ERK1+2)/MAPK14	Ratio 55	ERK1/ARRB1
Ratio 8	(ERK1+2)/(MAPK8+MAPK14)	Ratio 56	CREB/ERK1
Ratio 9	ERK1/(MAPK8+MAPK14)	Ratio 57	CREB/PREP
Ratio 10	ERK2/(MAPK8+MAPK14)	Ratio 58	GR/RGS2
Ratio 11	Gi2/ARRB1	Ratio 59	ARRB1/S100A10
Ratio 12	Gi2/ARRB2	Ratio 60	CREB2/MAPK8
Ratio 13	Gs/ARRB1	Ratio 61	CREB2/S100A10
Ratio 14	Gs/ARRB2	Ratio 62	ERK1/GR
Ratio 15	(Gi2+Gs)/ARRB1	Ratio 63	GR/MAPK8
Ratio 16	(Gi2+Gs)/ARRB2	Ratio 64	MAPK8/MR
Ratio 17	(Gi2+Gs)/(ARRB1+ARRB2)	Ratio 65	ERK2/ARRB1
Ratio 18	Gi2/(ARRB1+ARRB2)	Ratio 66	CREB/MR
Ratio 19	Gs/(ARRB1+ARRB2)	Ratio 67	MAPK8/S100A10
Ratio 20	Gi2/CREB	Ratio 68	MR/PREP
Ratio 21	Gi2/CREB2	Ratio 69	ARRB2/IL-1 beta
Ratio 22	Gs/CREB	Ratio 70	ERK2/RGS2
Ratio 23	Gs/CREB2	Ratio 71	MAPK8/PREP
Ratio 24	(Gi2+Gs)/CREB	Ratio 72	MKP1/RGS2
Ratio 25	(Gi2+Gs)/CREB2	Ratio 73	RGS2/IL-1 beta
Ratio 26	(Gi2+Gs)/(CREB+CREB2)	Ratio 74	ARRB1/GR
Ratio 27	Gi2/(CREB+CREB2)	Ratio 75	IL-8/ADA
Ratio 28	Gs/(CREB+CREB2)	Ratio 76	PBR/IL-6
Ratio 29	Gi2/RGS2	Ratio 77	ODC1/PBR
Ratio 30	Gs/RGS2	Ratio 78	IL-8/IDO
Ratio 31	(Gi2+Gs)/RGS2	Ratio 79	IL-8/SERT
Ratio 32	Gi2/Gs	Ratio 80	IL-8/IL-6
Ratio 33	GR/MR	Ratio 81	IDO/IL-6
Ratio 34	MKP1/(ERK1+ERK2+MAPK8+MAPK14)	Ratio 82	IL-8/MR
Ratio 35	MKP1/(ERK1+ERK2)	Ratio 83	MPK1/ADA
Ratio 36	MKP1/(MAPK8+MAPK14)	Ratio 84	IL-8/ODC1
Ratio 37	MKP1/ERK1	Ratio 85	IL-8/CD8 alpha
Ratio 38	MKP1/ERK2	Ratio 86	PBR/SERT
Ratio 39	MKP1/MAPK8	Ratio 87	IL-8/MAPK14
Ratio 40	MKP1/MAPK14	Ratio 88	IL-8/PREP
Ratio 41	ERK2/Gi2	Ratio 89	IL-8/CD8 beta
Ratio 42	ERK1/Gi2	Ratio 90	SERT/IL-6
Ratio 43	CREB/ARRB2	Ratio 91	ADA/ODC1
Ratio 44	ERK2/GR	Ratio 92	CREB2/IL-8
Ratio 45	Gi2/GR	Ratio 93	RGS2/ADA
Ratio 46	PREP/P2X7	Ratio 94	Gs/IL-8
Ratio 47	ERK2/Gs	Ratio 95	ODC1/IL-6
Ratio 48	ARRB2/ERK1	Ratio 96	CD8 beta/PBR
		Ratio 97	ODC1/CD8 alpha

Appendix 10: Summary ABS controls and DC controls

The table below was created by Joseph Tamm based on the statistical analysis, I did in section 7.3.1, and based on visual trends noted with the graphical software Spotfire. An explanation is given after the table.

Correlation between clinical variables and gene expression in two control groups (ABS and DC)

ABS subjects
DC subjects

	CREB2	DPP4	ERK1	ERK2	GR	Gs	MAPK8	MAPK14	MKP1	MR	PBR	RGS2	S100A10	SERT	VMAT2
Family History (D/A/S)		Inc **					Inc **								
Family History (D/A/S)															
Tobacco use														Dec ***	
Tobacco use														trend down	
Lifetime experiences (D/A)	Inc ***	Inc ***		Inc **	Inc ***		Inc ***					Inc **			Inc **
Lifetime experiences (D/A)		Inc ***			trend up							trend up			
Lifetime treatments (D/A)	Inc **	Inc ***			Inc **		Inc ***		Inc **						
Lifetime treatments (D/A)					trend up	Inc **			trend up						
Appetite Change		Inc **												Dec **	
Appetite Change														trend down	
Sleep Problems		Inc **													
Sleep Problems		Inc **													
10 Symptom score (*)		Inc ***					Inc ***		trend up		Dec **				
10 Symptom score (*)			Inc **	Inc ***					Inc **		Inc **				
Vegetative symptoms		Inc **													
Vegetative symptoms															
Recent stress		Inc **													
Recent stress															

Early life stress															
Early life stress													Inc ***		
Interest in sex					Inc **						trend down				
Interest in sex								Inc **		Dec **	Inc **				

D/A/S = Depression / Anxiety / Suicide
** indicates importance on the 1% significance level, while *** indicates importance on the 0.1% significance level.
(*) The 10 symptom score is as close to the symptom score sum in the ABS controls as possible. The scores were then binned into two bins based on a threshold value of 25.

Important points to keep in mind regarding the table:
A) Sometimes there are few subjects in groups, especially in the DC data set.
B) Neither the questions nor the questionnaires were the same in the two studies.

Joseph did the following:
1) Using mine univariate analyses on ABS and DC subjects, he noted which combinations of genes and variables produced a p value < 0.01. To keep the table manageable in size (for viewing), he omitted all of the genes and variables where nothing interesting was happening.

2) Next he looked in Spotfire to see if the statistical results made sense (a believable pattern). Most of the time they did, but if not, he excluded that combination. He also noted if gene expression increased or decreased with the variable in question.

3) Whenever a gene/variable combination in ABS produced a significant p value, but DC did not, he looked at the DC scatter plot to see if there was a trend toward the ABS result. Likewise, if DC had a significant p value but ABS did not, he looked at the ABS data to see if there was a similar trend. This was the most subjective part of the process.

4) In the table, whenever ABS and DC produced the same results for a gene/variable combination, he colored the cells green. Thus, DPP4 expression increases in both control groups when people have any lifetime experiences of depression/anxiety. Whenever the control groups produced clearly opposite results, he colored the cells red (see PBR). Whenever only one control group produced a significant change (for example DPP4 and family history), he colored the cells orange.

Results:
The table reflects results for 377 combinations of gene expression / question response.
That's 13 questions (includes age and BMI which are not shown) times 29 genes = 377

23 combinations yield different results for ABS vs DC (2 are red and 21 are orange) …so 6% of the combinations produce differences. 9 combinations display changes in the same direction in both subject groups (green). 345 combinations are unchanged in both groups … 9+345=354 … so 94% of the combinations produce the same pattern.
Therefore, DC and ABS are showing very similar expression trends.