Text mining and information extraction for the life sciences: an enhanced science approach

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 16 mei 2013
om 10.30 uur precies

door

Wilco Wilhelmus Maria Fleuren

geboren op 8 april 1978
te Nijmegen
This work was performed within the NBIC BioRange project “A Systems Bioinformatics Approach For Evaluating And Translating Drug-Target Effects In Disease Related Pathways” (BR4.2) and continued within the Netherlands eScience project “Creation of food specific ontologies for food focused text mining and knowledge management studies” (027.011.306).

Title: Text mining and information extraction for the life sciences: an enhanced science approach.

Copyright © 2013 Wilco Fleuren, Nijmegen, The Netherlands

ISBN/EAN: 978-90-9027438-6

Cover design: Walter van Rooij, Buro Brandstof

Printed by: CPI Wöhrmann Print Service, Zutphen
"Vorsprung durch Technik", Audi AG
# TABLE OF CONTENTS

| Chapter 1 | General introduction | 6 |
| Chapter 2 | CoPub update: CoPub 5.0 a text mining system to answer biological questions | 22 |
| Chapter 3 | Identification of new biomarker candidates for glucocorticoid induced insulin resistance using literature mining | 34 |
| Chapter 4 | Prednisolone-induced changes in gene-expression profiles in healthy volunteers | 56 |
| Chapter 5 | Prednisolone induces the Wnt signaling pathway in 3T3-L1 adipocytes | 80 |
| Chapter 6 | Prednisolone-induced differential gene expression in mouse liver carrying wild type or a dimerization-defective glucocorticoid receptor | 106 |
| Chapter 7 | Identification of T lymphocyte signal transduction pathways | 134 |
| Chapter 8 | Functional annotation of microorganisms using text mining | 166 |
| Chapter 9 | General discussion | 190 |

Summary | 200 |
Sanenvatting | 206 |
Dankwoord | 213 |
Curriculum vitae | 217 |
Bibiliography | 219 |
CHAPTER 1
General introduction
General introduction

Products for human consumption and intake need to be safe and should not be harmful. Hence the pharmaceutical and food industry spent millions of dollars on research and development (R&D) to develop better and safer products. The safety of drugs and food products is tested in experiments before they are marketed. These experiments result in enormous amounts of biological data that need to be analyzed in order to draw conclusions, to initiate follow-up studies, and to generate new hypothesis. This so-called data driven research is becoming more and more important in these industries and requires specialists which are able to extract useful information out of these biological data. Results from experiments are published in scientific papers and reports for which the abstracts often are deposited in the publicly available Medline database. Medline contains over 22 million of journal citations and abstracts from biomedical literature and is growing exponentially each year. Extraction of useful knowledge from these abstracts in order to annotate results from experiments requires methods which can do this extraction automatically.

Text mining and Information extraction

Text mining (TM) and automatic information extraction (IE) has evolved over the years into a sophisticated and specialized discipline in the bio medical sciences. The process of TM and IE involves high quality information extraction from literature and subsequently linking this information to experimental data. Currently, the most used approaches to retrieve knowledge from text are by co-occurrences and by natural language processing (NLP) [1].

Co-occurrence-based methods

Co-occurrence based methods use the simple concept that two keywords that co-occur in the same body of text often are functionally related. For example the co-occurrence of retinol-binding protein 4 (RBP4) and insulin resistance suggests a functional relationship between the gene and the disease [2, 3]. However, also many keywords co-occur in text without being functionally related. Therefore these methods often use a statistically scoring based on frequency of occurrence of both keywords, to add a degree of significance to the relationship. Co-occurrence based methods do not provide information about the type of relation between keywords but are therefore able to detect any type of relationship.
NLP-based methods

NLP based methods use prior knowledge to retrieve information out of text. This prior knowledge can be for instance how language is structured or specific knowledge on how biological information is mentioned in the literature. NLP is defined as "the processing of natural language, i.e. human languages by computers". In contrast to co-occurrence based systems, NLP systems are phrase based and include information about the relationship between two keywords, e.g. gene A inhibits gene B or gene C is involved in disease Y. NLP based systems are very suitable to search for specific relationships. However these systems are limited to predefined relationships and are often very computationally intensive.

Drug discovery

The first part of this thesis describes the application of bioinformatics and especially IE and TM to study drug (side) effects. The development of a new drug is very complex, time-consuming and expensive. On average the development of a drug takes about twelve years and costs are in the order of 1 billion US dollars. Already in the United States only, approximately just one out of six drugs that enter the clinical testing pipeline will eventually obtain approval to be marketed [4]. Despite the increasing investment in research and development (R&D), the attrition rate of drug candidates increases. Therefore early detection of failure of drug candidates is essential to reduce the drug development costs. Traditionally, the drug development process can be divided into distinct phases starting with the target discovery phase and eventually leading to the successful market approval of a drug candidate (Figure 1).

Target discovery

In the first phase possible new therapeutic targets will be identified. Information about

<table>
<thead>
<tr>
<th>Research</th>
<th>Development</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target discovery</td>
<td>Lead discovery</td>
</tr>
<tr>
<td>Lead optimization</td>
<td>Pre-clinical development</td>
</tr>
<tr>
<td>Pre-clinical development</td>
<td>Clinical trials</td>
</tr>
<tr>
<td>Clinical trials</td>
<td>Registration</td>
</tr>
<tr>
<td>Registration</td>
<td>Marketing &amp; sales</td>
</tr>
</tbody>
</table>

Figure 1 A schematic representation of the drug development pipeline. The process ranges from the identification of potential drug targets in the Target discovery stage to the successful introduction of the drug on the market in the Marketing & sales stage.

...the function and the role of these targets in the development of the disease will be gathered and subsequently used to get an overview of the pathways and genes/proteins that are involved. This step is very important since it gives insight into the biological context of the target and its relation to the disease. Typically in this phase information...
about the target will be retrieved from earlier studies and from the scientific literature. Promising targets will be validated in both *in vitro* and *in vivo* experiments in which the therapeutic effect and the relation of the target to the disease will be further investigated.

**Lead discovery and lead optimization**

Once possible targets have been detected and verified, screening of large compound libraries starts to identify so called lead compounds, which are able to modulate the biological activity of targets. Subsequently in the lead optimization phase these identified compounds will be chemically modified to improve their potency, selectivity and efficacy. The efficacy of compounds will be determined in genome-wide association studies using microarray technology, in which changes in gene expression are measured as a result of compound treatment. These expression profiles can be compared with expression profiles from other compounds and can also be used to get an insight into the mechanisms and pathways affected.

In the target discovery, lead discovery and lead optimization phases is it important to gather as much as possible information about the target, its biological surrounding and the compound to predict (side) effects. TM and IE are important techniques to retrieve this information from the scientific literature.

**Pre-clinical development**

Compounds which showed a promising effect on the biological target will move to the *pre-clinical development* phase were they will be further tested for efficacy and safety. Compounds are tested on toxicity in animal models. Furthermore in pharmacokinetics experiments the compounds are tested for their absorption, distribution, metabolism, and excretion (ADME) properties in the body. The pre-clinical development phase is important because the ultimate safety profile of compounds will determined. Based on this safety profile decisions will be made to move the compounds to the next phase in which they will be tested in humans.

Approximately two third of the compounds fail in the pre-clinical development phase because of poor toxicity and/or pharmacokinetics properties.

**Clinical trials**

In the last phase compounds also called drug candidates will be tested in humans in so called clinical trials. Clinical trials are normally taking place in four phases. In phase I in a small group of healthy volunteers (20 - 80) the toxicity and possible side effects of the drug candidate will be tested together with assessment of the correct dosage. These tests will continue in phase II, but in larger groups (100-300) of both healthy volunteers and patients. The effectiveness of the drug candidate will be further tested. In phase III the drug candidate will be administered to large groups of people (1000-3000) to continue to investigate the effectiveness and to compare the effects with commonly used treatments.
The results from these phases will be documented and reviewed by agencies such as the European Medicines Agency (EMA) and the US Food and Drug Administration (FDA) to get market approval for the drug candidate. Phase IV trials are post-market studies in which long-term treatment is monitored to assess long-term effects of the drug. Clinical trials are time-consuming taking several years and account for 50-70% of the total drug development costs.

**Studying drug effects on different levels**

With the rising drug development costs and the increasing pressure by governments to introduce safer and more valuable drugs on the market, the pharmaceutical industry faces a challenge to reduce these costs in order to keep delivering affordable drugs on the market. Therefore early detection of failure of drug candidates is essential. To predict the effects of drug candidates it is important to understand which drug targets they are able to modulate, and what the biological surrounding of these targets are. Many drug targets influence multiple phenotypic traits, genes, proteins and pathways.

The techniques used in this thesis to study drug effects on genes, pathways and proteins

![Figure 2 Techniques used in the studies described in this thesis to study drug effects.](image-url)
Genome-wide expression profiling
Genome-wide association studies enable to study gene expression profiles of thousands of genes simultaneously. In this way drug effects can be studied by looking at changes in gene expression as a result of treatment with a drug or compound. Gene expression is measured using microarray technology. A microarray consists of spots on which pieces of DNA called probes are attached to the surface of the array. Extracted RNA from the biological sample will be labelled with a fluorescent tag and hybridized to the probes on the microarray. Subsequently, the hybridized chip is scanned to measure fluorescent intensities. Typically, drug effects on gene expression are studied by comparing the gene expression from a drug treated sample with gene expression from a control sample.

Protein profiling using anti-body arrays
Differentially expressed proteins can be measured using anti-body arrays. This enables to study drug effects on proteins. Instead of using DNA probes, protein arrays consist of antibodies that are attached to the microarray surface. These antibodies are able to detect target proteins from the biological sample, by means of specific labelled secondary antibodies. Currently, protein arrays can only hold a couple of hundreds of antibodies, this in contrast to DNA microarrays. However with the introduction of antibody arrays the gap between the effects seen at gene expression level and the effects seen at protein level can be decreased.

Text mining to interpret experiments
Co-occurrence based TM methods have been successfully used to analyze microarray data, identifying gene-disease and gene-drug relationships [5-8].
We developed CoPub a text mining tool that is able to search for occurrences of different types of keywords from a wide variety of thesauri, e.g. gene names, disease names, drug names and pathway names, in Medline abstracts. CoPub uses the principle that a co-occurrence of two keywords in the same abstract is an indication of a functional relation between both keywords. The significance of a found relation is described by the R-scaled score, which is an indication of the strength of a co-occurrence of two keywords given their individual frequencies of occurrence [9]. CoPub has originally been developed for the interpretation of microarray data [5, 9] and its technology has been expanded to other areas, such as the identification of novel hidden relations [10]. In chapter 2 of this thesis we describe CoPub 5.0, which combines all CoPub technologies with extensive search and filter algorithms that can be used to answer a variety of biological questions.

Gene literature networks
Studying genes which are related to a disease or potential drug target enables to better...
understand which mechanisms are involved. For this it is important to study genes in their functional context. Typically such a functional context can be the biological process in which the gene is involved. Examples of such a biological process can be for instance the glycolysis in which in ten steps glucose will be converted into pyruvate. The functional context of phosphofructokinase, an enzyme in the glycolysis, can be defined by the other enzymes in this process. Although the functional context for phosphofructokinase in this example is straight forward and well defined, in most cases the functional context of a gene is not that straight forward.

Another way to define a functional context for a gene is to search for co-occurrences of this gene with other genes in Medline abstracts. Hence in this way a literature network of co-occurring genes can be created. This network can be used to study drug effects by means of mapping drug induced gene expression data or by mapping drug related information from the literature onto the network. This approach not only enables to study the effects of a drug on a single gene but also on the genes with which it is functionally related in the network. In chapter 3 of this thesis this approach is used to study the effects of dexamethasone, a synthetic glucocorticoid used as an anti inflammatory agent for the treatment of inflammatory diseases, on a network of insulin resistance (IR) related genes.

**Glucocorticoids**

The studies described in chapters 3-5 of this thesis focus on the metabolic side effects as a result of treatment with synthetic glucocorticoids (GCs).

GCs are naturally occurring hormones which are released by the adrenal cortex in response to stress. They are part of the feed-back mechanism of the immune system in response to inflammation. The most abundant form in the human body is cortisol, which binds with high affinity to the glucocorticoid receptor (GR). This receptor is expressed in almost every cell in the body and unbound GR is mainly present in the cytosol of the cell. Studies indicated that GR regulates up to 10 -20% of the human genome in different cell types, thus showing the important role of this receptor. GC effects can roughly be divided into two types [11, 12]. The best studied effects are the so called genomic effects, in which the GC-GR complex moves into the nucleus of the cell, where it regulates gene expression. The GR-GC complex enhances or represses gene transcription by direct binding as a homodimer to the glucocorticoid-responsive-element (GRE), by binding to other transcription factors or by direct GRE binding and interactions with transcription factors bound to neighboring sites of a gene (Figure 3).

These genomic effects take place within a couple of hours. However GCs induce also rapid effects which are taking place within minutes. These so called non-genomic effects are far less well studied and are thought to be mediated by affecting the properties of cell membranes or through the binding of intracellular or membrane-bound glucocorticoid receptors [13, 14].
Therapeutic usage
Because of their immunosuppressive properties, GCs are the most commonly prescribed anti-inflammatory drugs for the treatment of inflammatory diseases since the mid 1950s, with the introduction of prednisolone, dexamethasone and betamethasone. Patients suffering from rheumatoid arthritis, asthma and psoriasis are often treated with GCs. In fact up to 60% of the rheumatoid arthritis patients are more or less continuously treated with GCs [15, 16]. Although the treatment of these inflammatory diseases with GCs is very successful, it is hampered by dose-dependent side effects, such as osteoporosis, skin atrophy, hypertension and severe metabolic disturbances such as insulin resistance which

Figure 3 General mechanisms of Action of glucocorticoids and the glucocorticoid receptor in the inhibition of inflammation. TNFα denotes tumor-necrosis factor α, HSP heat-shock protein, mRNA messenger RNA, and P phosphate. The three mechanisms are nongenomic activation, DNA-dependent regulation, and protein interference mechanisms (e.g., NFκB elements). Black arrows denote activation, the red line inhibition, the red dashed arrow repression, and the red X lack of product (i.e., no mRNA). Reproduced with permission from [12], Copyright Massachusetts Medical Society.
eventually leads to diabetes and obesity. This limits the long-term usage of these synthetic GCs. The general assumption is that the immunosuppressive effects of GCs are mainly driven by transrepression, in which the GC-GR complex, together with other transcription factors, prevents the transcription of pro-inflammatory genes. The unwanted side-effects are mainly driven by the act of transactivation of gene expression. A lot of effort has been put into the development of novel GC ligands with an improved efficacy/side effect ratio compared to conventional GCs. This requires knowledge about the mechanisms behind GC induced metabolic side effects. In the studies described in this thesis, a combination of microarrays, antibody arrays, bioinformatics and TM/IE are used to characterize GC induced metabolic side–effects in both in vitro and in vivo studies.

**Studying microorganisms in food production and health**

IE and TM have successfully been used in drug discovery to study drug effects and can also be used in other areas, such as in the microbiology. Microorganisms play crucial roles in processes ranging from waste-product degradation in ecosystems to industrial processes where they are used for the development of fermented foods and beverages. Furthermore, microorganisms are also the main contributor to food spoilage and can be harmful causing diseases.

**Microorganisms in food**

The biggest contribution of microorganisms to food development is in the process of fermentation, in which organic compounds are transformed under the influence of bacteria and yeasts. Louis Pasteur was one of the first scientists who discovered that yeasts were responsible for alcohol fermentation [17]. This discovery initiated the scientific and industrial interests in food microbiology and resulted in that fermentation nowadays is a highly industrialized process. Lactic acid bacteria (LAB) are very important for the fermentation of food. For instance *Lactobacillus bulgaricus subsp. delbrueckii* and *Streptococcus thermophilus* strains are used for making yoghurt [18]. *Lactobacillus lactis* contributes to the flavor development of cheese [19, 20], and *Oenococcus oeni* is important for malolactic fermentation in the process of wine making, giving wine its characteristic taste and aroma [20, 21]. Estimations indicate that about one third of the human diet consists of fermented foods and drinks [22]. In addition to these food-characteristics, some LAB strains have proven health-promoting effects on humans and are used as probiotics.
**Probiotics**

Probiotics are defined as "live microorganisms which when administered in adequate amounts confer a health benefit on the consumer". The possible contribution of probiotics to human health has been proposed in a number of ways: via competitive exclusion of pathogenic bacteria, via enhancement of epithelial barrier function by modulation of signaling pathways that lead to enhanced mucus, or by modulation of the immune system of the host [23]. Probiotic effects have been established in the prevention of number of diseases including diarrhea, necrotizing enterocolitis and pouchitis. However probiotic effects have not been consistently observed in studies to other diseases such as irritable bowel syndrome, atopic dermatitis, and inflammatory bowel disease. This requires that more scientific research is needed to improve the industrial application of probiotics towards stimulation of human health.

**Community profiling studies**

In order to use microorganisms in industrialized processes for food production and to develop cures against pathogenic microorganisms, relationships between microorganisms and a given process or treatment needs to be studied. Metagenomics is a well established technique to study the dynamics of microbial communities in so-called community profiling studies. Metagenomics is the genomic analysis of microbial DNA that is extracted directly from environmental samples, by means of pyrosequencing of 16S rRNA genes and has been proposed as the most accurate approach for the assessment of taxonomy composition [24]. One of the outcomes of these metagenomics experiments are lists of microorganisms that are affected by a given process or treatment. Annotation of these lists is essential to understand how these microorganisms are involved in the studied processes.

**Microorganism focused text mining and knowledge building**

Similar to the usage of TM for the annotation of drug induced gene expression, in chapter 8 of this thesis, we describe how TM can be used to get a quick insight into the biological properties of microorganisms. Thesaurus based TM is used to annotate lists of microorganisms that are the result from community profiling studies. Furthermore TM without using thesaurus is used to study LAB to gain insight into the biological processes in which they are involved.
Thesis outline

This thesis consists of two parts. Both parts focus on the development and application of TM and IE for the annotation of biological experiments. In the first part bioinformatics and IE/TM are used to study the effects of drugs in order to unravel the underlying mechanisms and affected pathways. The majority of the studies described in this thesis focus on the mechanisms that underlie the unwanted metabolic side effects of glucocorticoid treatment.

In the second part we demonstrate that knowledge of the usage of TM in drug discovery can be used to retrieve information from literature in order to study microorganisms. A schematic summary of the work in this thesis is depicted in Figure 4.

In Chapter 2 we describe CoPub 5.0, a publicly available text mining system, which uses Medline abstracts to calculate robust statistics for keyword occurrences. In CoPub 5.0 all available CoPub technology has been integrated into a single application with a renewed user interface. CoPub 5.0 enables to answer a variety of biological questions ranging from microarray data analysis and keyword enrichment to the search for new biological relations.

Chapter 3 describes CoPubGene a method to mine Medline abstracts for functional gene-
disease relations. The method has been used to create a literature network of genes related to insulin resistance, a known side-effect of glucocorticoid treatment for which the mechanisms are not completely understood. The study gave new insights into these mechanisms and indicated a possible involvement of the sex-steroid synthesis pathway in the development of GC induced IR.

**Chapter 4** describes a genome wide expression profiling study to identify prednisolone induced gene signatures in CD4+ T lymphocytes and CD14+ monocytes derived from healthy volunteers and to link these signatures to underlying biological pathways involved in metabolic adverse effects as a result of prednisolone treatment.

**Chapter 5** describes the usage of genome wide expression data and adipokine profiling in combination with literature mining techniques to study mechanisms and signaling pathways involved in prednisolone induced metabolic disturbances in murine 3T3-L1 adipocytes. We found the wnt signaling pathway to be involved in metabolic disturbances in adipocyte cells and we suggested this pathway for follow-up experiments to study its role in GC-induced metabolic effects.

**Chapter 6** describes a genome wide gene expression profiling study to examine the role of GR dimerization in the regulation of gene expression in mouse liver. Biological pathways targeted by prednisolone in the liver of wild type (WT) mice and mice lacking the ability to form glucocorticoid receptor dimers (GRdim) have been identified and compared.

**Chapter 7** describes a study that identifies pathways and genes involved in the differentiation of Jurkat T cells towards Th1 and Th2 subtype cells, using various stimuli and pathway inhibitors. Specific gene profiles of *IL-2*, associated with a Th1 type of response, and *CCL1*, associated with a Th2 type of response, have been identified. TM analysis was used to find additional evidence for these Th1 and Th2 specific fingerprints.

In **Chapter 8** the usage of TM to add meaning to results of metagenomics experiments is demonstrated. Abstracts, in which microorganisms occur, are automatically analyzed to retrieve useful keywords that can be used to understand the biology of microorganisms and their involvement in the processes studied.

In **Chapter 9** a short discussion and some concluding remarks with regard to this thesis are described.
References

CHAPTER 2
CoPub update: CoPub 5.0 a text mining system to answer biological questions

Wilco W.M. Fleuren\textsuperscript{1,2}, Stefan Verhoeven\textsuperscript{3}, Raoul Frijters\textsuperscript{1}, Bart Heupers\textsuperscript{4}, Jan Polman\textsuperscript{3}, René van Schaik\textsuperscript{3}, Jacob de Vlieg\textsuperscript{1,3} and Wynand Alkema\textsuperscript{3}

\textsuperscript{1}Computational Drug Discovery (CDD), CMBI, NCMLS, Radboud University Nijmegen Medical Centre, \textsuperscript{2}Netherlands Bioinformatics Centre (NBIC), PO Box 9101, 6500 HB Nijmegen, \textsuperscript{3}Molecular Design and Informatics, MSD, PO Box 20, 5340 BH Oss and \textsuperscript{4}SARA Computing and Network Services, Amsterdam, The Netherlands.

\textit{Nucleic Acids Research, 2011, 39:W450-4.}
Abstract

In this article, we present CoPub 5.0, a publicly available text mining system, which uses Medline abstracts to calculate robust statistics for keyword co-occurrences. CoPub was initially developed for the analysis of microarray data, but we broadened the scope by implementing new technology and new thesauri. In CoPub 5.0, we integrated existing CoPub technology with new features, and provided a new advanced interface, which can be used to answer a variety of biological questions. CoPub 5.0 allows searching for keywords of interest and its relations to curated thesauri and provides highlighting and sorting mechanisms, using its statistics, to retrieve the most important abstracts in which the terms co-occur. It also provides a way to search for indirect relations between genes, drugs, pathways and diseases, following an ABC principle, in which A and C have no direct connection but are connected via shared B intermediates. With CoPub 5.0, it is possible to create, annotate and analyze networks using the layout and highlight options of Cytoscape web, allowing for literature based systems biology. Finally, operations of the CoPub 5.0 Web service enable to implement the CoPub technology in bioinformatics workflows. CoPub 5.0 can be accessed through the CoPub portal http://www.copub.org.
Introduction

Medline abstracts are a very useful source of biomedical information covering topics such as biology, biochemistry, molecular evolution, medicine, pharmacy and health care. This knowledge is useful to better understand the complexity of living organisms and can, for instance, be used to study groups of genes or metabolites in their biological context. In the 2008, Web Service issue of NAR, we presented CoPub as a publicly available text mining system. This system uses Medline abstracts to calculate robust statistics for keyword co-occurrences, to be used for the biological interpretation of microarray data (1,2). Since then, CoPub has been intensively used in the analysis of several microarray experiments and toxicogenomics studies [3–8]. However, literature data can be applied far beyond questions related to microarray studies. Therefore, we broadened the scope of CoPub by implementing new technology and adding new thesauri to the database. We developed a new technology called CoPub Discovery, which can be used to mine the literature for new relationships following a simple ABC-principle, in which keyword A and C have no direct relationship, but are connected via shared B-intermediates [9]. This technology can, for instance, be used to study mechanisms behind diseases, connect new genes to pathways or to find novel applications for existing drugs. To reflect all these developments, we created CoPub 5.0, which has a complete new user interface and in which we integrated all CoPub technologies. CoPub 5.0 enables the use of CoPub functionality in a very dynamic interactive manner by easily switching between multiple analysis modes and is very suitable to answer a variety of biological questions. It is also accessible using operations of the CoPub 5.0 Web Service (SOAP or JSON), which makes it possible to embed the CoPub functionality into bioinformatics workflows. CoPub 5.0 and the CoPub 5.0 Web Service can be accessed at the CoPub portal http://www.copub.org.

Methods

CoPub 5.0 has three analysis modes. A ‘term search’ mode that retrieves abstracts and

![Figure 1 Schematic representation of CoPub.](image)

The CoPub database holds co-occurrence information between categories in Medline Abstracts. The CoPub functionality can be used via three modes using the web interface or via the CoPub web services either via SOAP or JSON.
keyword relations for a single term, a ‘pair search’ mode that analyzes known or new relations between a pair of terms and a set of terms mode that deals with the relation between multiple terms (Figure 1).

‘Term search’ mode
The ‘term search’ mode provides a way to search for keywords and subsequently showing their relations with other categories in the CoPub database. This mode provides a table and cloud view which can be used to answer questions such as ‘to which diseases is this gene related?’ or ‘in which biological processes is my metabolite involved?’ For instance, the cloud view in which strongly connecting terms [i.e. high R-scaled score (1)] are displayed with a larger font, can be used to immediately show the most important relations of the term with keywords from one or more categories in the database (Figure 2A). The evidence for these relations lies in the Medline abstracts in which both terms occur. CoPub retrieves these abstracts, highlights both terms in them and ranks the abstracts which has the most term occurrences as first (Figure 2B). In the example, in Figure 2, it is shown that CXCR4 is strongly connected to its ligand CXCL12 and CXCR7, with which it forms a heterodimer (A). Also, CXCR4 is strongly connected to ‘HIV infections’ (category: disease), which is mediated by CXCR4 and to ‘stromal’ cell, to which CXCR4 is linked because of its stromal derived ligand CXCL12. In B an example is shown of the underlying abstracts for the co-occurrences.

Figure 2 An example of the term search view for the human chemokine receptor 4. In the cloud view, it is immediately clear, by the large font of the terms, that CXCR4 is strongly connected to its ligand CXCL12 and CXCR7, with which it forms a heterodimer (A). Also, CXCR4 is strongly connected to ‘HIV infections’ (category: disease), which is mediated by CXCR4 and to ‘stromal’ cell, to which CXCR4 is linked because of its stromal derived ligand CXCL12. In B an example is shown of the underlying abstracts for the co-occurrences.
‘Pair search’ mode
The ‘pair search’ mode can be used to search for specific relations between existing keywords in the CoPub database, e.g. to search for a relation between a gene and a drug. A wizard will guide the user in its search for relations between terms. CoPub will first search for co-occurrences and if no co-occurrence is found, the user can search for hidden relations using the ‘closed discovery’ mode (see ‘Hidden Relations’ section). The pair search mode can be useful to search the literature for more evidence which supports relations found in experiments, for instance, between a drug and a pathway or between a gene and a pathway or which supports hypothesis.

‘Set of terms’ mode
Biological research often involves a better understanding of the complexity of living organisms, for instance, to better understand the development of a disease or to gain more insight into complex signaling pathways [10,11]. This requires a systems approach in which groups of genes or metabolites are studied in relation to a disease, drugs or pathways. In CoPub 5.0, we provide such a systems approach via the ‘set of terms mode’. In this mode, a set of keywords can be uploaded either by copy-pasting them or by uploading a text file. Terms can belong to multiple categories (e.g. insulin belongs to the category human gene and to the category drug), which can be further specified to only the desired categories using the ‘Members of category’ option. An uploaded set of terms can be analyzed in a number of ways.

Set enrichment analysis.
To see with which categories the set has significant relations, an enrichment analysis can be performed. In this analysis, the relation of a given term from a category with the set is tested using the Fisher exact test against a background set. The calculated P-values are corrected using the Benjamini–Hochberg multiple testing correction method. In case the set consists of multiple categories, only one type of category can be chosen to be used as a background set. For each enriched term, the number of contributors of the set is shown and the contributors can be accessed by clicking on this number. All statistical tests are done using the R Statistics package (http://www.r-project.org).

Set annotation.
A set of terms can be annotated by searching for co-occurrences between the set and categories in the database. The cloud view immediately shows for each term in the set, the most significant associated terms (in larger font) per category. Categories from the database can be added or removed from this view. All co-occurring annotation can be downloaded from this view using the download button.
Network.

To analyze the relations between terms in the set, a literature network of the set can be created. Subsequently, the network will be visualized using the Cytoscape web plugin (http://cytoscapeweb.cytoscape.org/). Strongly connected terms have thick edges (high R-scaled score), which immediately shows important relations (Figure 3). For large networks (>500 nodes), the network can be downloaded and visualized in a standalone Cytoscape environment.

**Figure 3 Network of a group of mixed terms using the Cytoscape plugin.** In the network the gene *IL4* has strong connections to the genes *IL2* and *CCL11* and is also strongly connected to the biological process ‘isotype switching’ and ‘cytokine biosynthesis’. This is indicated by the thick edges between these nodes. Clicking on an edge will show the abstracts in which both terms occur allowing for more detailed analysis of the biological context in which the terms are related.

Add additional terms.

At any time an uploaded set can be extended with additional terms. These additional terms can be provided by the user (via ‘add additional terms’), by searching for co-occurrences between the set and categories in the database (via ‘Grow set with co-occurrences’) or by adding a specific term via the ‘term search’ mode, from which it can be added to the set using the ‘Add term to set’ button.

Hidden relations

From the ‘term search’ mode and the ‘pair search’ mode in the website, it is possible to search for hidden relations using the CoPub Discovery technology [9]. CoPub Discovery uses an ‘open discovery’ and ‘closed discovery’ process to search for new hidden relations. Both processes follow an ABC principle in which, in case of ‘open discovery’, the user provides a term A (e.g. disease) and searches the literature for hidden relations with a category (C) via intermediates (B) and in case of ‘closed discovery’, the user tests the hypothesis that, for instance, a gene (A) is related to a disease (C) and searches the literature for shared intermediates (B) which support this hypothesis. This technology
can be useful to find different roles of genes in new pathways or to get more insight into mechanisms behind diseases.

**CoPub Web Service**

The operations from the CoPub Web Service allows to embed CoPub functionality into work flows and to use it in an automatic fashion. For this, we provide to use these operations either via SOAP or via JSON. The description of these operations can be found in the help files of the CoPub 5.0 website and an example script, showing how operations can be used, is accessible via the CoPub portal http://www.copub.org.

**Discussion and conclusion**

CoPub 5.0 can be used to answer a wide variety of biological questions and bridges the gap between indexed searching of PubMed and dedicated manually curated pathway databases such as Wikipathways [12], Ingenuity Pathway Analysis (http://www.ingenuity.com) and Metacore (GeneGo) (http://www.genego.com/metacore.php). There are a number of tools that provide part of the technology offered by CoPub. For example, Chilibot [13] is an NLP based tool that retrieves abstracts for user defined pairs of terms, but has no curated dictionaries, meaning that only relations between user defined terms are found, thus limiting the possibility to discover new relations. FACTA [14] offers curated dictionaries but does not provide indirect relation searching or network possibilities. Arrowsmith [15] is a tool for the discovery of hidden relations but does not contain curated ontologies nor does it provide networking possibilities or term mode options. Furthermore, options for analyzing enrichment in terms lists are not provided by these tools, limiting their use for the analysis of approximately omics sets. The advantage of CoPub is that it integrates the approaches offered by the above methods and combines this with advanced graphical output, web service ability and multiple options for analyzing lists of terms and creating networks. The statistical framework of CoPub 5.0, together with the cloud view functionality is very suitable for the analysis of large ‘omics’ data sets. First, by running a broad scan using enrichment to get a general overview of the data and subsequently by zooming in on relevant pathways, focusing on strong connections (by means of R-scaled score) in the data. Together with the hidden relations technology, this can be used to generate new hypotheses. Future steps could include a better interface to Gene Set Enrichment Analysis (GSEA) software [16] and to incorporate Natural Language Processing (NLP) to be able to even better filter on biological relevant information.

**Acknowledgements**

The authors thank SARA Computing and Networking Services (Amsterdam, The Netherlands) for maintaining and hosting our CoPub database and web server.
Funding
Grants received from the Netherlands Bioinformatics Centre (NBIC) under the BioAssist program and from Merck Sharp & Dohme (MSD). Funding for open access charge: NBIC.

Conflict of interest statement.
None declared.
References


CHAPTER 3
Identification of new biomarker candidates for glucocorticoid induced insulin resistance using literature mining.

Wilco W.M. Fleuren\textsuperscript{1,2}, Erik J.M. Toonen\textsuperscript{3}, Stefan Verhoeven\textsuperscript{4}, Raoul Frijters\textsuperscript{1,a}, Tim Hulsen\textsuperscript{1,b}, Ton Rullmann\textsuperscript{5}, René van Schaik\textsuperscript{4}, Jacob de Vlieg\textsuperscript{1,4} and Wynand Alkema\textsuperscript{1,c}

\textsuperscript{1}Computational Drug Discovery (CDD), CMBI, NCMLS, Radboud University Nijmegen Medical Centre, P.O. Box 9101, 6500 HB Nijmegen, the Netherlands, \textsuperscript{2}Netherlands Bioinformatics Centre (NBIC), P.O. Box 9101, 6500 HB Nijmegen, The Netherlands, \textsuperscript{3}Department of Medicine; Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands, \textsuperscript{4}Netherlands eScience Center, Amsterdam, The Netherlands, \textsuperscript{5}TNO, Zeist, The Netherlands.

\textsuperscript{a}Present address: Rijk Zwaan Nederland BV, Fijnaart, The Netherlands
\textsuperscript{b}Present address: Philips Research Laboratories, Eindhoven, The Netherlands.
\textsuperscript{c}Present address: NIZO Food Research BV, Ede, The Netherlands.

\textit{BioData Min. 2013 Feb 4;6(1):2.}
Abstract

Background
Glucocorticoids are potent anti-inflammatory agents used for the treatment of diseases such as rheumatoid arthritis, asthma, inflammatory bowel disease and psoriasis. Unfortunately, usage is limited because of metabolic side-effects, e.g. insulin resistance, glucose intolerance and diabetes. To gain more insight into the mechanisms behind glucocorticoid induced insulin resistance, it is important to understand which genes play a role in the development of insulin resistance and which genes are affected by glucocorticoids.

Medline abstracts contain many studies about insulin resistance and the molecular effects of glucocorticoids and thus are a good resource to study these effects.

Results
We developed CoPubGene a method to automatically identify gene-disease associations in Medline abstracts. We used this method to create a literature network of genes related to insulin resistance and to evaluate the importance of the genes in this network for glucocorticoid induced metabolic side effects and anti-inflammatory processes.

With this approach we found several genes that already are considered markers of GC induced IR, such as phosphoenolpyruvate carboxykinase (PCK) and glucose-6-phosphatase, catalytic subunit (G6PC). In addition, we found genes involved in steroid synthesis that have not yet been recognized as mediators of GC induced IR.

Conclusions
With this approach we are able to construct a robust informative literature network of insulin resistance related genes that gave new insights to better understand the mechanisms behind GC induced IR. The method has been set up in a generic way so it can be applied to a wide variety of disease networks.

Keywords: literature mining, insulin resistance, glucocorticoids, gene networks.
Background

Glucocorticoids (GCs) are often prescribed for the treatment of inflammatory diseases such as rheumatoid arthritis, asthma, inflammatory bowel disease and psoriasis [1-3]. Despite their excellent efficacy, usage is limited because of side-effects such as insulin resistance, glucose intolerance, diabetes, central adiposity, dyslipidemia, skeletal muscle wasting and osteoporosis [4-8].

GCs bind to the glucocorticoid receptor (GR), which then dimerizes and translocates to the nucleus where it influences gene transcription. Positive regulation of genes (transactivation) is mainly mediated by direct binding of the GR-GC complex to glucocorticoid response elements located in the regulatory region of a target gene. The GR-GC complex may also bind to negative glucocorticoid response elements, which leads to a negative regulation of genes (transrepression). It is believed that transrepression, in which proinflammatory genes are downregulated, is mainly responsible for the efficacy of GCs as anti-inflammatory drugs [5, 7], while transactivation might be responsible for the GC-induced adverse effects [9].

An important side effect is the development of insulin resistance (IR), because it is the onset of many metabolic diseases and conditions such as obesity, diabetes mellitus and hypertension. IR is a physiological condition in which a given concentration of insulin produces a less-than-expected biological effect. These biological effects are different depending on the tissue in which they occur. For instance, under IR conditions, fat and muscle cells fail to adequately respond to circulating insulin, which results in reduced glucose uptake, and subsequently higher glucose levels in blood [10, 11]. In liver cells the IR- effects can be seen in reduced glycogen synthesis and storage, and a failure to suppress glucose production and release into the blood.

One way by which GCs induce IR is by inhibition of the recruitment of GLUT4 glucose transporter, which results in reduced insulin-stimulated glucose transport in skeletal muscle [12]. However, not all mechanisms involved in GC-induced side effects are not completely understood. To gain more insight into mechanisms behind GC induced IR, it is important to understand which genes play a role in the development of insulin resistance and which genes are affected by GCs.

It has been widely recognized that a system approach in which networks of genes in their functional context are studied, contributes to a better understanding of the mechanisms and pathways related to the disease and the drug effects [13-17]. To study a gene network related to a disease such as IR, a list of disease related genes as well as a notion of the interactions between these genes is needed.

Literature databases such as Medline contain many studies about IR and the molecular effects of synthetic glucocorticoids and thus are a good resource that can be used to create and study disease related gene networks.

The retrieval of relevant gene-disease associations out of the millions of abstracts in
Identification of new biomarker candidates for GC induced IR using literature mining

Medline is very labor intensive and thus a text mining system is needed to this in an automated fashion.

In previous work we reported about CoPub [18-20], a publicly available text mining system, which has successfully been used for the analysis of microarray data and in toxicogenomics studies [21-26]. CoPub calculates keyword co-occurrences in titles and abstracts from the entire Medline database, using thesauri for genes, diseases, drugs and pathways. We used this technology to develop CoPubGene, a rapid gene–disease network building tool. To evaluate the importance of genes in these networks we implemented a method to score the importance of genes in biological processes of interest by incorporating their functional neighborhood.

We used CoPubGene to create a network of genes related to insulin resistance and to evaluate the importance of the genes in this network for glucocorticoid induced metabolic side effects and anti-inflammatory processes.

By using this method, we identified several genes that already are considered markers of GC induced IR, such as phosphoenolpyruvate carboxykinase (PCK) and glucose-6-phosphatase, catalytic subunit (G6PC) [27, 28]. Even more importantly, we were able to identify genes involved in steroid synthesis that have not yet been recognized as mediators of GC induced IR.

Methods

CoPubGene

We constructed CoPubGene as a SOAP based web service (Table 1). This CoPub Web Service WSDL is created in Eclipse using the so-called Document Literal Wrapped style. The web service provider code is written in Perl using the SOAP::WSDL module and is available via the CoPub portal http://www.copub.org.

Retrieval of Gene-Disease associations

To create disease related gene networks, we used CoPubGene to retrieve gene-disease and gene-gene associations from Medline abstracts. Disease terms which had significant gene associations based on the R-scaled score (rs > 35) and literature count (lc > 5) in Medline abstracts, were extracted from the CoPub thesaurus.

Disease clustering

Disease clustering was done in R (http://www.r-project.org) using the pvclust R package with “complete” setting for hierarchical clustering, based on correlation distance of R-scaled scores between genes and diseases, with 100 bootstrap replications. The hierarchical cluster was visualized using Denroscope [29]. Additional gene set enrichment analysis against the GENETIC_ASSOCIATION_DB_DISEASE was done with the annotation server DAVID [30, 31].
Creation of IR gene network
CoPubGene was used to create a set of genes related to IR, by searching for associations between genes and IR in Medline abstracts using default values ($rs > 30$ and $lc > 5$). Subsequently the IR-gene network was created by connecting genes that had significant co-occurrences with each other.

Keyword enrichment analysis of IR related genes
Keyword enrichment analysis on the list of IR related genes was done against disease and drug terms from the CoPub database. Threshold values were chosen using default values.

Analysis of the IR gene network and calculation of neighbor score for genes
The IR gene network was analyzed by mapping specific occurrences of the IR related genes with ‘inflammation’ and ‘dexamethasone’ in Medline abstracts onto the network. For the evaluation of the involvement of a gene, calculation of the literature score for a given gene and a given disease term, also the effects of dexamethasone and inflammation on the connecting genes are included. The literature score for gene $g$ with term $d$ is calculated in the following way:

$$\text{Literature score}_g = \frac{g1 + Ns}{2}, \quad Ns = \frac{(rg2 * g2) + (rg3 * g3) + \ldots (rgn * gn)}{rg2 + rg3 + \ldots rgn}$$

In which $g1$ is the R-scaled score of gene $g$ with term $d$, and $Ns$ is the literature score of its neighboring genes with term d. This latter score $Ns$ is calculated using the R-scaled score of each neighboring gene of gene $g$ with term d ($g2, g3, \ldots gn$) relative to its relation (R-scaled score) with gene $g$ ($rg2, rg3, \ldots rgn$).

Results
We developed CoPubGene by creating a number of web service operations that can be used to construct networks of genes based on their co-occurrences in Medline abstracts. These web service operations can be combined to answer a variety of biological questions (Table 1). For example, the question “to what biological processes is this gene related?” can be answered by running the “get genes” and “get literature neighbours” functions. Using subsequently the “get references” function will return all the relevant pubmed entries in which the gene and keywords co-occur. By applying the “get keywords” and “get literature neighbours” functions one can retrieve all disease terms that are linked to a given drug term in the Medline abstract, or vice versa, retrieve all drug terms that are linked to a given disease term in abstracts. The networks that are created can be written to Cytoscape for downstream applications and visualizations. Also more advanced
Identification of new biomarker candidates for GC induced IR using literature mining

questions such as the construction of disease related gene networks, and subsequent calculation of keyword enrichment in this network can be addressed in an automatic way. In Table 1 the available web service operations are shown.

Table 1 List of available operations of the CoPub Web Service. Biological identifiers are used by CoPub to identify biological concepts in the system. Each biological concept has a unique identifier.

<table>
<thead>
<tr>
<th>Name</th>
<th>Operation name</th>
<th>Input</th>
<th>Output</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Get genes</td>
<td>Get_genes</td>
<td>Gene name, gene identifier</td>
<td>Biological identifier(s), with gene specific information</td>
<td>Each gene in CoPub belongs to an internal identifier (biological identifier). The Get_genes operation converts the input gene to such a biological identifier. This biological identifier serves as an input for subsequent operations.</td>
</tr>
<tr>
<td>Get Keywords</td>
<td>Get_keywords</td>
<td>Keyword</td>
<td>Biological identifier(s), with keyword specific information</td>
<td>Retrieves for a set of keywords, the Biological identifiers to which these keywords belong in CoPub. These biological identifiers serve as an input for subsequent operations.</td>
</tr>
<tr>
<td>Get references</td>
<td>Get_references</td>
<td>Biological identifier(s)</td>
<td>Literature references</td>
<td>Given a Biological identifier, retrieves all abstracts in which the term occurs.</td>
</tr>
<tr>
<td>Get literature</td>
<td>Get_literature_</td>
<td>Biological identifier(s)</td>
<td>Literature neighbors</td>
<td>Given a Biological identifier, retrieves a list of keywords which are mentioned in the literature together with the input term.</td>
</tr>
<tr>
<td></td>
<td>neighbours</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Get enriched</td>
<td>Get_enriched_</td>
<td>List of gene identifiers</td>
<td>List of enriched keywords</td>
<td>For a list of genes, this operation calculates a keyword overrepresentation.</td>
</tr>
<tr>
<td>keywords</td>
<td>keywords</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Get literature</td>
<td>Get_literature_</td>
<td>Biological identifier(s)</td>
<td>SVG / Cytoscape network</td>
<td>For a set of genes, the operation creates a network of genes.</td>
</tr>
<tr>
<td></td>
<td>network</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Get categories</td>
<td>Get_Categories</td>
<td>-</td>
<td>List of categories</td>
<td>Returns a list of categories of terms in CoPub.</td>
</tr>
<tr>
<td>Get chips</td>
<td>Get_chips</td>
<td>-</td>
<td>List of microarrays</td>
<td>Returns a list of available Affymetrix chip names in CoPub.</td>
</tr>
<tr>
<td>Version</td>
<td>Version</td>
<td>-</td>
<td>Version of code and literature</td>
<td>Returns the version of the code and literature.</td>
</tr>
<tr>
<td>Selftest</td>
<td>Selftest</td>
<td>-</td>
<td>Diagnostic information</td>
<td></td>
</tr>
</tbody>
</table>
Retrieval of gene-disease associations

Our aim was to get insight into the pathways and genes that are involved in insulin resistance, and the effect of glucocorticoids on this network. As a first step we created a list of genes associated with insulin resistance using CoPubGene. This yielded a list of 384 genes each of them connected to IR with an R scaled score (in Table 2A the top scoring genes with IR are shown, the full list is available in supplementary table 2). To evaluate the quality of this list and to investigate whether this gene list is unique for IR or whether this list contains a large number of genes that are associated with multiple diseases we constructed a gene association list for all diseases in the disease thesaurus of CoPub, using similar parameter settings as used for construction of the IR gene list. This yielded a list of disease profiles with for each disease, a number of genes connected to that disease with an R scaled score. (Table 2 shows the results for a few selected diseases, the full

Figure 1 Hierarchical cluster of disease terms from the CoPub database. The top 80 disease terms with the most gene associations are shown. Disease terms are clustered together based on having the same gene associations. Red numbers at the nodes represent approximately unbiased bootstrap values (%).
### Table 2 Part of the disease matrix, which has been used for the clustering.

A link between a gene and a disease term is given by the R-scaled score. Top scoring genes with insulin resistance (IR) are related to other metabolic disease terms (indicated in green) and not to inflammatory disease terms (indicated in red) (A). Top scoring genes with rheumatoid arthritis are related to other inflammatory disease terms (shown in green) and not to metabolic disease terms (show in red) (B). IBD: inflammatory bowel disease.

#### A

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>IR</th>
<th>Diabetes mellitus, type 2</th>
<th>Obesity</th>
<th>Rheumatoid arthritis</th>
<th>Psoriasis</th>
<th>IBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBP4</td>
<td>retinol binding protein 4, plasma</td>
<td>53</td>
<td>47</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GPR83</td>
<td>G protein coupled receptor 83</td>
<td>51</td>
<td>48</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>BSCL2</td>
<td>Bernardinelli Seip congenital lipodystrophy 2</td>
<td>51</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RETN</td>
<td>resistin</td>
<td>51</td>
<td>46</td>
<td>47</td>
<td>0</td>
<td>0</td>
<td>37</td>
</tr>
<tr>
<td>AGPAT2</td>
<td>1 acylglycerol 3 phosphate O acyltransferase</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>SERPINA12</td>
<td>serpin peptidase inhibitor, clade A, member 12</td>
<td>50</td>
<td>0</td>
<td>47</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>IRS1</td>
<td>insulin receptor substrate 1</td>
<td>50</td>
<td>43</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ADIPOR2</td>
<td>adiponectin receptor 2</td>
<td>50</td>
<td>0</td>
<td>45</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>STEAP4</td>
<td>STEAP family member 4</td>
<td>50</td>
<td>0</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ADIPOQ</td>
<td>adiponectin, C1Q and collagen domain containing</td>
<td>50</td>
<td>44</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

#### B

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>IR</th>
<th>Diabetes mellitus, type 2</th>
<th>Obesity</th>
<th>Rheumatoid arthritis</th>
<th>Psoriasis</th>
<th>IBD</th>
</tr>
</thead>
<tbody>
<tr>
<td>PADI4</td>
<td>peptidyl arginine deiminase, type IV</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>41</td>
<td>0</td>
</tr>
<tr>
<td>FCRL3</td>
<td>Fc receptor like 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>NCOA5</td>
<td>nuclear receptor coactivator 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>OLIG3</td>
<td>oligodendrocyte transcription factor 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>47</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TNFAIP3</td>
<td>tumor necrosis factor, alpha induced protein 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>46</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>PTPN22</td>
<td>protein tyrosine phosphatase, non receptor type 22</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>46</td>
<td>41</td>
<td>39</td>
</tr>
<tr>
<td>HLADRB4</td>
<td>major histocompatibility complex, class II, DR beta 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>45</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>HLADRB5</td>
<td>major histocompatibility complex, class II, DR beta 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>45</td>
<td>0</td>
<td>36</td>
</tr>
<tr>
<td>CHI3L1</td>
<td>chitinase 3 like 1</td>
<td>37</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>0</td>
<td>43</td>
</tr>
<tr>
<td>HLADRB1</td>
<td>major histocompatibility complex, class II, DR beta 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>0</td>
<td>36</td>
</tr>
</tbody>
</table>
These disease profiles were clustered using hierarchical clustering with multiscale bootstrap resampling, grouping together disease terms which have a similar profile, i.e. co-occur with the same genes (Figure 1). It appeared that a number of clusters of similar disease terms i.e. disease terms for which it is known that they have similar symptoms or have a similar mode of action, could be identified. For instance cancer related terms, such as ‘cancer of breast’, ‘cancer of prostate’ and ‘colon cancer’ are clustered together and inflammatory related disease terms such as ‘psoriasis’, ‘inflammatory bowel disease’ and ‘asthma’ are clustered together. These clusters also have high unbiased (AU) bootstrap values, indicating strong evidence for these clusters. To further confirm that the found gene-disease associations by CoPub are indeed biologically relevant, for each sub-cluster in Figure 1, we collected the union of all genes for that sub-cluster, and used these genes to perform a functional annotation analysis against the genetic association disease database using DAVID. The results of this analysis indicated that indeed similar disease terms to CoPub were found by DAVID (for the results of this analysis see supplementary tables). These analyses showed that with CoPubGene we are able to construct a relevant list of specific IR related genes that can be used for further analysis and that CoPubGene can be used to create a variety of disease related genes lists.

**Network of insulin resistance related genes.**

To create the IR gene network, we used the 384 genes from the IR gene list and connected the genes based on their co-occurrences with each other in Medline abstracts. The resulting network is shown in Figure 2A. We found that 381 genes of the IR gene list were connected to at least one other gene. We identified a number of hubs such as *peroxisome proliferator-activated receptor gamma* (PPARG), *insulin receptor substrate 1* (IRS1), *v-akt murine thymoma viral oncogene homolog 1* (AKT1), *insulin receptor* (INSR), *solute carrier family 2 (facilitated glucose transporter), member 4* (SLC2A4) and *insulin* (INS) which were connected to more than 100 other genes. The resulting network is a scale free network, as indicated by the distribution of connectivity that follows a power law distribution which is indicative for a scale free network (supplementary Figure 1) [32]. Although the above network has the characteristics of a biological network, and contains the expected genes as central hubs, without additional annotations this network representation is still largely uninformative and contains too little substructure to draw biological conclusions.

**Annotation of the network with drugs and diseases terms**

As a first step towards annotating the network and identification of sub networks with a shared biological function, we investigated which drugs and diseases in the literature are specifically linked to this network using a keyword enrichment analysis on the list of IR related genes (For details about the enrichment method see Table 1). This enrichment
yielded a number of drugs that are known drugs for the treatment of diabetes such as ‘rosiglitazone’, ‘metformin’, ‘pioglitazone’, and also ‘glucagon’ and ‘insulin’ which are frequently used for the treatment of hypoglycemia and hypoinsulinemia (Table 3A. For the full list see supplementary table 4A). Notably, among these top scoring drugs we found dexamethasone, a well known synthetic glucocorticoid. High scoring genes with dexamethasone are for instance CEBPA, SERPINA6, PCK2 and GPD1 (for a full list of genes per enriched drug term, see supplementary Table 4A.2), which also have been mentioned in the development of several metabolic diseases [33-37].

There are several top scoring over-represented terms that are related to metabolic diseases, e.g. ‘diabetes mellitus’, ‘obesity’, ‘diabetes mellitus, type 2’ and ‘hyperinsulinemia’ (Table 3B). The fact that these terms are high scoring is expected since we constructed the gene network based on the keyword insulin resistance. However we also found diseases that share a common origin with insulin resistance such as cardiovascular disease (Table 3B). The most interesting high scoring term for our particular research question was the non-metabolic term ‘inflammation’, which was represented in the network by genes such as IL6, IL18, IL1RA, SOCS1, SOCS3, CCL2 and CCR2. Several of these genes have been mentioned

**Table 3 Over-represented drug and disease terms (P-value < 0.05).** The top scoring drug terms in the IR network from the CoPub database (A). Top scoring disease terms from the CoPub database in the IR network (B).

<table>
<thead>
<tr>
<th>Term</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>insulin</td>
<td>358</td>
</tr>
<tr>
<td>dexamethasone</td>
<td>195</td>
</tr>
<tr>
<td>nitric oxide</td>
<td>193</td>
</tr>
<tr>
<td>estrogen</td>
<td>169</td>
</tr>
<tr>
<td>adenosine</td>
<td>151</td>
</tr>
<tr>
<td>estradiol</td>
<td>145</td>
</tr>
<tr>
<td>rosiglitazone</td>
<td>125</td>
</tr>
<tr>
<td>actinomycin</td>
<td>124</td>
</tr>
<tr>
<td>actinomycin d</td>
<td>121</td>
</tr>
<tr>
<td>glucagon</td>
<td>120</td>
</tr>
<tr>
<td>thrombin</td>
<td>108</td>
</tr>
<tr>
<td>progesterone</td>
<td>97</td>
</tr>
<tr>
<td>trypsin</td>
<td>86</td>
</tr>
<tr>
<td>nicotinamide</td>
<td>85</td>
</tr>
<tr>
<td>metformin</td>
<td>84</td>
</tr>
<tr>
<td>pioglitazone</td>
<td>82</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>insulin resistance</td>
<td>381</td>
</tr>
<tr>
<td>obesity</td>
<td>263</td>
</tr>
<tr>
<td>inflammation</td>
<td>219</td>
</tr>
<tr>
<td>diabetes mellitus</td>
<td>190</td>
</tr>
<tr>
<td>cardiovascular disease</td>
<td>181</td>
</tr>
<tr>
<td>Diabetes mellitus,type 2</td>
<td>173</td>
</tr>
<tr>
<td>Oxygen deficiency</td>
<td>164</td>
</tr>
<tr>
<td>fibrosis</td>
<td>138</td>
</tr>
<tr>
<td>hyperinsulinemia</td>
<td>137</td>
</tr>
<tr>
<td>Cancer of breast</td>
<td>131</td>
</tr>
<tr>
<td>Adiposity</td>
<td>130</td>
</tr>
<tr>
<td>cancer</td>
<td>128</td>
</tr>
<tr>
<td>starvation</td>
<td>120</td>
</tr>
</tbody>
</table>
in studies to be involved in the development of metabolic diseases. For instance, elevated levels of IL6 in subjects with obesity and diabetes showed an association between insulin resistance and IL6 [38]. Studies in mice showed that CCR2 deficiency or antagonism of this receptor resulted in attenuation of systemic insulin resistance and development of obesity, hence suggesting a modulating role of CCR2 in this [39, 40].

These results show that even with an unbiased data driven construction of a gene network, the relation between IR, dexamethasone and inflammation is discovered based on the genes that play a role in these effects. We subsequently highlighted the genes in the IR network that are related to inflammation and dexamethasone (Figure 2).

**Figure 2 Literature network of insulin resistance related genes (A).** Genes, represented by nodes are linked, based on co-occurrences in Medline abstracts. The thickness of the edge indicates the strength of the link between two genes (R-scaled score). Genes in blue have a co-occurrence with dexamethasone in Medline abstracts (R-scaled score). The strength of the link with dexamethasone is given by the color shading, ranging from no link (white) to a strong link (dark blue). The strength of the link with inflammation (R-scaled score) is given by the size of the node of the gene, ranging from no link (normal size of the node) to a strong link with inflammation (large size of the node). Sub-network for gene PPARG (B). Sub-network of Cytochrome P450s (C).

**Genes linked to inflammation and glucocorticoids in the context of insulin resistance**

From a drug development perspective it is interesting to separate the desired effect of GCs on inflammatory processes from the undesired effect on metabolic processes. To
rank each gene with respect to the relation with GC and inflammation, we calculated for each gene a literature score with dexamethasone and inflammation. Subsequently we focused on genes that score low on inflammation and high on dexamethasone (Figure 3). These genes are thought to be more exclusively related to GC induced IR. For these genes we calculated a literature neighbor score as well, by also including the relations of dexamethasone and inflammation with genes to which the gene is connected in the network. In Figure 3 it is shown that many genes which are not directly connected to inflammation (grey dots) are definitely influenced by inflammation via their connecting genes (black dots). The majority of the genes in Figure 3 are directly involved in important metabolic processes such as gluconeogenesis (PCK2, G6PC, PC and GCG), glycolysis (GCK, GCG), glucose uptake, lipid metabolism (ACACA, CHPT1, GPD1) and carbohydrate metabolism (GPD1). Other ones are directly involved in insulin signaling (GIP, IGF2, IPF1, IAPP).

Figure 3 Influence of dexamethasone and inflammation on IR genes that have a high score with dexamethasone (>25) and a low score with inflammation (<25). The direct score of these genes with dexamethasone and inflammation are shown in grey. The literature neighbor score for these genes, by also including the relations of dexamethasone and inflammation with genes to which the gene is connected in the network, are shown in black. The grey arrows indicate the migration of the gene from a direct score to a literature neighbor score.
**Sex steroid physiology in relation to insulin resistance**

Interestingly in Figure 3 we also see three cytochrome P450s, i.e. *CYP17A1*, *CYP19A1* and *CYP21A2*, which are key regulator enzymes in the steroid synthesis (Figure 4). The sub-network in Figure 2C shows the three cytochromes P450s and their direct gene neighbors. Analysis of this sub-network showed that many of the genes in the network are mentioned in studies from women suffering of the Polycystic ovary syndrome (PCOS), in which there is an imbalance of a woman's female sex hormones. PCOS is characterized by insulin resistance, possibly because of hyperandrogenism and low levels of SHBG. The latter effect has also been observed in men suffering from the metabolic syndrome [41]. Also a study by Macut et al. suggested that alterations of a cross-talk between glucocorticoid signaling and metabolic parameters, is related to PCOS pathophysiology [42].

Additional topological analysis of the sub-network using cytohubba [58] revealed that *IGF1R, HSD11B2, IGF2* and *SHBG* have a high betweenness centrality, i.e. they have many shortest paths going through them, analogous to major bridges and tunnels on a high map. Studies show that such a bottle necks play important roles in biological networks [43, 44].

*Figure 4 Steroid synthesis.* Enzymes indicated with a red box have been found in our analysis. *CYP17A1* encodes for an enzyme which has both a 17α-hydroxylase and a 17,20 lyase function. *CYP21A2* encodes for a steroid 21-hydroxylase and *CYP19A1* encodes for an aromatase. Figure derived from the image Steroidogenesis.png in Wikipedia, by David Richfield and Mikael Häggström, licensed under Creative Commons CC BY-SA 3.0 and GFDL.
Identification of new biomarker candidates for GC induced IR using literature mining

Chapter 3

CYP19A1 encodes for an aromatase which is responsible for the aromatization of androgens into estrogens, thus influencing the androgen to estrogen balance. Several studies showed that an imbalance between androgen and estrogen balance because of aromatase deficiency resulted in the development of symptoms related to the metabolic syndrome [45-48]. The fact that dexamethasone can regulate aromatase activity [49-51], suggests a role of aromatase in GC induced IR.

CYP17A1 is a key regulator of androgen synthesis and catalyzes the reactions in which pregnenolone and progesterone are converted into their 17-alpha-hydroxylated products and subsequently into Dehydroepiandrosterone (DHEA). A decline in DHEA and also its sulfated ester (DHEA-S) has been suggested to be causally linked to insulin resistance and obesity [52-55]. The possible inhibitory effects of dexamethasone on Cyp17a1 [56, 57] suggests a role in GC induced IR by this gene.

CYP21A2 is a cytochrome P450 enzyme coding for the 21-hydroxylase that is involved in the biosynthesis of the steroid hormones aldosterone and cortisol. A defect in this gene leads to Congenital adrenal hyperplasia (CAH) in which there is a disbalance in cortisol and aldosterone secretion. CAH patients are characterized by insulin resistance, lower insulin sensitivity and hyperinsulinemia [58-61]. Some studies indicate that the development of IR is because of GC treatment in this patient group [62-64]. Whether these patients develop IR because of CAH and deficiency of 21-hydroxylase, or because of the fact that they are often treated with synthetic GCs need to be elucidated.

Genes involved in osteoporosis

Another side effect of GC treatment is the development of glucocorticoid induced osteoporosis (GIOP) [65]. GIOP is characterized by reduced bone mineral density (BMD), decreased bone mass and disturbance of the bone matrix, leading to increased susceptibility to fractures. We applied CoPubGene to deduce important genes involved in GIOP by analyzing top scoring genes with OP (in total 131 genes associated with OP were found, see supplementary table 5; the network of these top scoring genes with relations to dexamethasone and inflammation is shown in supplementary figure 3. The majority of the genes are involved in bone remodeling and resorption (TNFRSF11A, TNFRSF11B, TNFSF11, SP7, CTSK), in bone mineralization (PTH, Klotho, VDR, Calca, BGLAP) or are part of the wnt signaling pathway that is involved in the regulation of bone formation (SOST, DKK1, LRP5, LRP6) [66]. Among these genes are known biomarkers of GIOP such as osteoprotegerin (encoded by TNFRSF11B) and the ligand RANK-L (encoded by TNFS11) [67]. Here we also searched for genes with a low score with inflammation. Several of these genes in the set, such as BGLAP, COL1A1 and SP7 are affected by GCs [68-72], have low associations with inflammation and therefore are interesting biomarker candidates for GiOP.
Discussion

In the work presented here we used Medline abstracts to study mechanisms and genes involved in glucocorticoid induced insulin resistance. We created CoPubGene, a number of web service operations that can be used to retrieve relevant gene-disease, gene-drug and gene-gene associations out of Medline abstracts, using the CoPub technology. The clustering of disease terms based on their associations with genes in Medline abstracts showed that CoPubGene is able to generate a list of specific IR genes that can be used for further analysis, and that this method also can be used to generate a variety of other gene disease associations. We used this clustering to evaluate the quality of disease related gene lists, generated using a text mining approach, because to our knowledge there is no real gold standard data set that covers a sufficient range of gene-disease associations that can be used. Databases such as OMIM and the KEGG disease database [73] only cover a sub range of diseases which makes these datasets difficult to use in this type of evaluation.

Next, we studied the IR genes in their functional context, by including genes with which they co-occur in Medline abstracts. In this gene network we focused on genes that are strongly linked to dexamethasone and less strongly to inflammation. These genes are thought to be more exclusively related to GC induced IR and therefore might be interesting markers for this effect. However, all of them are to a certain extent related to inflammation, either directly or indirectly by their neighbors, which suggests that these genes cannot be used as an exclusive marker for GC induced IR. This might have consequences for the search of dissociating compounds, i.e. compounds which only have the immune suppressive properties and not the unwanted side effects. Instead the search should focus on compounds that show a reduced effect on the expression of these IR genes.

The majority of the IR genes that have a low literature neighbor score for inflammation (< 25) and a high score for dexamethasone (literature neighbor score > 25) code for enzymes and hormones directly involved in important metabolic processes, such as glycolysis, gluconeogenesis, glucose uptake and lipid metabolism. All these processes are tightly regulated by insulin. This suggests that at a first instance, the search for mechanisms of GC induced IR should be focused on these processes. Additionally, we also identified a sub network of genes involved in sex steroid synthesis that to our knowledge, not have been recognized yet as mediators of GC induced side effects. Key enzymes involved in steroid synthesis, i.e. CYP17A1, CYP21A2 and CYP19A1 keep the balance between several steroids, and an impairment of this balance could possibly result in metabolic disturbances such as IR. Additional topological analyses could further prioritize this sub-network for follow-up studies to determine the influence of GCs on sex steroid synthesis and the relation to IR. In such a study one could look at the influence of GCs on the balance between the steroids in combination with their influence
Identification of new biomarker candidates for GC induced IR using literature mining

on insulin stimulated glucose uptake in glucose sensitive tissues such as adipose and muscle tissue.

Conclusions

Using CoPubGene we are able to construct an informative literature network of IR related genes by only using information from Medline abstracts. Our approach revealed genes, that on a first glance were not considered to be involved in GC induced IR and thus gave new insights that might lead to a better understanding of the mechanisms behind GC induced IR.

Financial Contribution
Author WF was supported by the Biorange project (BR4.2) “A Systems Bioinformatics Approach For Evaluating And Translating Drug-Target Effects In Disease Related Pathways” of NBIC.

Conflict of interest
Authors have no conflict of interest.

Authors’ contributions
WF performed all data analysis, research design and wrote the paper. ET supervised with biological interpretation and helped with writing the paper. SV, TH developed the web service operations. TR, RVS, RF and JVD helped with research design. WA supervised the work and helped with research design, analyzing the data and writing the manuscript.

Supplementary information can be retrieved on request by the author of this thesis.
References


CHAPTER 4
Prednisolone-induced changes in
gene-expression profiles in healthy
volunteers

Erik J. M. Toonen\textsuperscript{1}, Wilco W. M. Fleuren\textsuperscript{2,3}, Ulla Nässander\textsuperscript{1},
Marie-José C. van Lierop\textsuperscript{1}, Susanne Bauerschmidt\textsuperscript{1}, Wim
H. A. Dokter\textsuperscript{1} and Wynand Alkema\textsuperscript{1,3}

\textsuperscript{1}MSD Research Laboratories, Oss, The Netherlands, \textsuperscript{2}Computational Drug Discovery
(CDD), CMBI, NCMLS, Radboud University Nijmegen Medical Centre and \textsuperscript{3}Netherlands
Bioinformatics Centre (NBIC), Nijmegen, The Netherlands.

Abstract

Background
Prednisolone and other glucocorticoids (GCs) are potent anti-inflammatory and immunosuppressive drugs. However, prolonged use at a medium or high dose is hampered by side effects of which the metabolic side effects are most evident. Relatively little is known about their effect on gene-expression in vivo, the effect on cell subpopulations and the relation to the efficacy and side effects of GCs.

Aim
To identify and compare prednisolone-induced gene signatures in CD4+ T lymphocytes and CD14+ monocytes derived from healthy volunteers and to link these signatures to underlying biological pathways involved in metabolic adverse effects.

Materials & methods
Whole-genome expression profiling was performed on CD4+ T lymphocytes and CD14+ monocytes derived from healthy volunteers treated with prednisolone. Text-mining analyses was used to link genes to pathways involved in metabolic adverse events.

Results
Induction of gene-expression was much stronger in CD4+ T lymphocytes than in CD14+ monocytes with respect to fold changes, but the number of truly cell-specific genes where a strong prednisolone effect in one cell type was accompanied by a total lack of prednisolone effect in the other cell type, was relatively low. Subsequently, a large set of genes was identified with a strong link to metabolic processes, for some of which the association with GCs is novel.

Conclusion
The identified gene signatures provide new starting points for further study into GC-induced transcriptional regulation in vivo and the mechanisms underlying GC-mediated metabolic side effects.

Keywords: gene signature, glucocorticoids, metabolic side effects, prednisolone, whole-genome expression profiling.
Introduction

Glucocorticoids (GCs) display strong anti-inflammatory effects and are therefore among the most prescribed agents for the treatment of numerous inflammatory conditions [1]. Exogenous GCs are used in daily clinical practice to treat inflammatory, autoimmune and allergic disorders, to attenuate organ rejection after transplantation, to treat brain edema, shock and various blood cancers [2]. Most of the effects of GCs are mediated through binding to the intracellular GC receptor (GR). This receptor can be found in almost all tissues of the human body. Upon entering the cell, GCs bind to the GR, which then dimerizes and translocates to the nucleus where it influences gene transcription. Positive regulation of genes (transactivation) is mainly mediated by direct binding of the GR homodimer to glucocorticoid response elements located in the regulatory region of a target gene. Ligand-activated GR homodimers may also bind to negative glucocorticoid response elements, which leads to a negative regulation of genes (transrepression). Transrepression is also achieved by interaction of GR monomers or homodimers with other transcription factors via protein–protein interactions [1,3]. It is believed that the transrepression pathway, in which proinflammatory genes are downregulated, is mainly responsible for the efficacy of GCs as anti-inflammatory drugs [1,3]. However, in spite of excellent efficacy, the clinical use of GCs is hampered by a wide range of dose-dependent side effects. Persistent exposure to elevated levels of circulating GCs is related to insulin resistance, glucose intolerance, diabetes, adiposity, dyslipidemia, skeletal muscle wasting and osteoporosis [1,2,4]. The mechanisms of action behind these adverse events are largely unknown. It is thought that the transactivation pathway might be responsible for these GC-induced adverse effects[5].

A number of synthetic analogs of the natural human GC cortisol have been developed, and one of the most widely used GCs in the daily clinic is prednisolone [6]. For example, prednisolone is used for the treatment of rheumatoid arthritis, inflammatory bowel disease, psoriasis and asthma [7–9]. As seen for all GCs, treatment with prednisolone also induces, besides the desired effects, a large number of (metabolic) adverse events. Therefore, there is a high medical need for improved anti-inflammatory drugs that are as effective as classical GCs but have a better safety profile. Since the precise mechanisms involved in prednisolone-induced side effects remain unclear, it is of importance to elucidate genes or gene pathways that contribute to these side effects.

In recent years, genome-wide gene-expression analysis using microarrays has become a key component to unraveling the underlying transcriptional regulation of various complex mechanisms and diseases [10,11]. Indeed, a large number of experiments detailing the effects of GC-induced transcriptional regulation have been described [12–14]. However, most of these experiments have been performed in cell lines or cultured primary cells or whole blood. Relatively little is known regarding the effects of GCs in an in vivo setting and the effect on the individual cell types that are present in blood. To
address these issues and to monitor the acute effects of GCs in an *in vivo* setting in which GC-induced transcriptional regulation is measured in the context of all other relevant signals, we performed gene-expression profiling on CD4+ and CD14+ cells isolated from healthy volunteers that were treated with prednisolone. These cell types were chosen since T cells and monocytic cells are the two most important immune cells involved in inflammation and are also accessible.

The healthy volunteers that participated in the trial demonstrated metabolic adverse events after prednisolone treatment such as a decrease in the C-peptide:glucose ratio and impaired glucose sensitivity [15]. The aim of the present study was to identify and compare acute prednisolone-induced gene signatures in CD4+ T lymphocytes and CD14+ monocytes. In addition, we aimed to link these signatures to underlying biological pathways involved in previously described metabolic adverse effects [15].

The results demonstrate that prednisolone induces significant changes in the transcriptional activity of T cells and monocytes, with a large overlap in the gene sets that are affected in both cell types and a relatively low number of genes regulated by prednisolone exclusively in one cell type. Several genes were identified that play crucial roles in metabolic pathways and that might be good candidates for further studies aimed at identifying biomarkers for GC-induced adverse events and the development of improved GCs.

**Materials & methods**

**Study design**

The study was a single-centre, double blind, randomized, placebo-controlled study, consisting of two distinct parts which were performed in parallel. In the first part of the study the acute effects of prednisolone treatment were assessed (acute protocol). Subjects were treated for 1 day with either two doses of 5 mg prednisolone (30 mg group; n = 6) or one dose of 75 mg prednisolone (n = 6). Prednisolone and cortisol levels were monitored at several time points (from 7:55 AM to 8:00 AM the next day) at day 0 and day 1 (predose) and day 2 (24 h or 12 h after the last dosing). Blood was sampled at day 1 (predose) and day 2 (24 h or 12 h after the last dosing) and used for whole-genome expression profiling. In the second part of the study, subjects were treated for 15 days with a placebo (n = 11) or 30 mg/day prednisolone (n = 12) and blood was sampled for expression analysis at day 0 (predose) and day 15 (predose) (24 h after the last dose on day 14) (2-week protocol). In this study, we focused on the results from the acute protocol and used results from the 2-week cohort for validation. In both parts, medication was taken in the morning (8:00 AM) and for the 2 ×15 mg prednisolone group also at 8.00 PM. Prednisolone tablets were obtained from Pfizer AB (Sollentuna, Sweden), placebo tablets were provided by MSD (Oss, The Netherlands). Tablets were encapsulated in order to allow the treatment to be blinded. Blood prednisolone and cortisol levels were
determined by using the normal-phase HPLC procedure as described previously [16]. After sampling, the blood was separated into CD4+ T lymphocytes and CD14+ monocytes. Subsequently, RNA was isolated and used for whole-genome expression analysis. All participants provided written informed consent. The study was approved by an independent ethics committee and the study was conducted in accordance with the Declaration of Helsinki, using good clinical practice.

**Study population**
Both the acute and 2-week protocols enrolled healthy male volunteers (age range between 20 and 45 years; BMI 22–30 kg/m²). Health status was confirmed by medical history, physical and laboratory examination, ECG and vital signs recordings. Participants were excluded if they had a clinically relevant history or presence of a medical disorder known to influence the investigational parameters.

**Isolation of CD4+ & CD14+ cells**
Human peripheral blood mononuclear cells (PBMCs) were isolated by Ficoll density gradient centrifugation. Next, CD4+ T cells were isolated from PBMCs by the use of a MACS® CD4+ T-Cell Isolation Kit II according to protocol (Miltenyi Biotec, Utrecht, The Netherlands). In short, T cells are isolated by depletion of non-T cells (negative selection). Non-T cells are indirectly magnetically labeled with a cocktail of biotin-conjugated monoclonal antibodies, as primary labeling reagent, and antibiotin monoclonal antibodies conjugated to Microbeads® (MicroBeads, Cateno, AS, Norway), as secondary labeling reagent. The magnetically labeled non-T cells are depleted by retaining them on a MACS Column in the magnetic field of a MACS Separator, while the unlabeled T cells pass through the column. CD14+ monocytes were isolated from PBMCs by using the MACS Monocyte Isolation Kit II, which uses a positive selection procedure.

**RNA isolation**
RNA was isolated using the RNeasy® midi kit according to the manufacturer’s protocol (Qiagen Benelux BV, Venlo, The Netherlands). To remove residual traces of genomic DNA, the RNA was treated with DNase I (Invitrogen, Leek, The Netherlands) while bound to the RNeasy column. Quality and quantity of the purified RNA was controlled using a NanoDrop® spectrophotometer (Nanodrop technologies, DE, USA). RNA integrity was investigated by using the 2100 Bioanalyzer (Agilent technologies, PA, USA).

**Expression analysis using Affymetrix GeneChip Human Genome U133 Plus 2.0**
Amplification of 20 ng total RNA was performed with the Two-Cycle Eukaryotic Target Labeling kit (Affymetrix, CA, USA). Briefly, after the first round of cDNA synthesis, an unlabeled ribonucleotide mix (MEGAscript® T7 kit, Ambion, TX, USA) was used to generate unlabeled cRNA according to the protocol of Affymetrix. After clean up of the cRNA with
a GeneChip® Sample Cleanup Module IVT Column (Affymetrix, CA, USA), the unlabeled cRNA concentration was determined and 150 ng was reverse transcribed using random primers. Subsequently, the T7-Oligo(dT) Promoter Primer was used in the second-strand cDNA synthesis to generate double-stranded cDNA containing T7 promoter sequences. The resulting double-stranded cDNA was then amplified and biotin labeled using the IVT Labeling kit (Affymetrix). Biotin-labeled cRNA was fragmented at 1 μg/μl following the manufacturer’s protocol. After fragmentation, cRNA (10 μg) was hybridized at 45°C for 16 h to the Human Genome U133 2.0 Plus array (Affymetrix). Following hybridization, the arrays were washed, stained with phycoerythrin-streptavidin conjugate (Molecular Probes, OR, USA), and the signals were amplified by staining the array with biotin-labeled antistreptavidin antibody (Vector Laboratories, CA, USA) followed by phycoerythrin-streptavidin. The arrays were laser scanned with a GeneChip Scanner 3000 7G (Affymetrix) according to the manufacturer’s instructions. Data was saved as a raw image file and quantified using AGCC version 1.0 (Affymetrix). Raw data files are available on request.

Synthesis of cDNA & quantitative PCR
Total RNA (1 μg) was reverse transcribed using a commercially available cDNA synthesis kit (iScript®, BioRad Laboratories, CA, USA). Quantitative PCR was performed by SYBR® Green-based quantification according to the manufacturer’s protocol (Applied Biosystems, CA, USA). Primers were developed for FKBPS, DDIT4, KLF9, ZBTB16, FLT3 and ABCA1 using Primer3 (Supplementary Table 1; www.futuremedicine.com/doi/suppl/10.2217/pgs.11.34) [17]. All primer pairs were exon-spanning. The genes HUWE1 and FXBO9 were chosen as endogenous controls since the expression arrays did not demonstrate any differences in expression of these genes between the experimental groups. PCR products were selected to be between 80 and 120 bp long. Samples were run on the 7900 HT Real-Time PCR System (Applied Biosystems) using the following protocol: 10 min denaturation at 95°C, and 40 cycles of 15 s denaturation at 95°C, 1 min, annealing and extension at 60°C. All primer pairs were validated in triplicate using serial cDNA dilutions that resulted in a final concentration of the equivalent of 800, 400, 200, 100 and 50 pg/μl input of total RNA in the first strand synthesis. Primer pairs that were 100 ± 10% efficient, which implies a doubling of PCR product in each cycle, were used to quantify mRNA levels. Threshold cycle numbers (referred to as Ct) were obtained using the 7900 HT SDS software version 2.3 (Applied Biosystems). All samples were measured twice and duplicate samples with a standard deviation larger than 0.5 were excluded from the analysis. The relative quantity of the gene-specific mRNA was calculated from the average value of the ΔCt (target gene Ct endogenous control gene Ct) for each sample. Differences in expression between two samples were calculated by the 2^ΔCt method [18,19].
Statistical analysis

Raw data were analyzed in R using packages from the BioConductor library [101]. Expression data were normalized using guanine cytosine robust multiarray analysis. Identification of regulated genes was carried out using the limma package from Bioconductor. For identifying differentially expressed genes after treatment with prednisolone, the following selection criteria were used: fold change >2 and p-value <0.05 after correction for multiple testing using the Benjamini–Hochberg correction.

Results

Comparison between subjects of both study protocols

Subjects of the acute and the 2-week protocol were matched regarding age, BMI and physical activity at baseline. No physical complaints were reported in both protocols when compared with the placebo group. The completion rate of the study was 100%, none of the healthy volunteers withdrew owing to prednisolone-induced adverse events [15].

Changes in lymphocyte & monocyte cell counts after treatment with prednisolone in the acute protocol

Glucocorticoid treatment will affect the number of immune cells in the blood. Leukocyte count (10^9/l) changed from 6.37 (± 2.21) at baseline to 8.77 (± 2.06) after treatment with 2 × 15 mg prednisolone at day one and from 5.62 (± 0.85) at baseline to 8.67 (± 0.66) after treatment with 1 × 75 mg prednisolone at day one. Lymphocyte count (10^9/l) changed from 2.63 (± 0.93) to 2.18 (± 0.63) in the prednisolone 2 × 15 group and from 2.23 (± 0.34) to 2.67 (± 0.10) in the prednisolone 1 × 75 group after treatment. Monocyte count (10^9/l) changed from 0.47 (± 0.197) to 0.52 (± 0.172) in the prednisolone 2 × 15 group and from 0.45 (± 0.084) to 0.53 (± 0.052) in the prednisolone 1 × 75 group after treatment.

Identification of prednisolone induced gene-expression profiles in CD4^+ T lymphocytes & CD14^+ monocytes

In order to study GC effects on distinct subpopulations in human blood, CD4^+ and CD14^+ cells were isolated from PBMCs as described in the methods section from the samples obtained from the acute protocol. These purification steps yielded >98% purified CD4^+ T cells and >90% CD14^+ cells. In total, 892 probe sets were found to be significantly regulated by prednisolone (in CD4^+ and/or CD14^+ cells). Figure 1 shows the distribution of fold changes for the differentially expressed probe sets under each condition. The maximum fold changes that are observed are between four- and eight-fold, the majority of the genes being upregulated. It appeared that the fold changes of the regulation were much larger in the CD4^+ T cells compared with CD14^+ monocytes. Surprisingly, the classical GC target genes such as FKB5, ZBTB16, DDIT4, KLF9 and FLT3 were downregulated in this
study (*FKBP5*: 2.1-fold; *ZBTB16*: 2.9-fold; *DDIT4*: threefold; *KLF9*: 3.8-fold) whereas most studies reported an upregulation of these genes after GC treatment [12–14,20–22]. These observed downregulations were also found by microarray analysis in the independent validation cohort (2-weeks protocol) and by quantitative PCR validation. It is known that expression of these genes is not only induced by GCs, which are administered as a drug, but also by the endogenous GC cortisol. We observed that in the acute protocol, cortisol levels decreased after treatment with prednisolone. Even after 24 h of prednisolone intake, cortisol levels are still reduced when compared with baseline despite that prednisolone levels are decreased to zero after a few hours (Figure 2). These reduced cortisol levels after 24 h may provide an explanation for the observed downregulation of *FKBP5*, *ZBTB16*, *DDIT4*, *KLF9* and *FLT3*.

In order to identify probe sets that are specifically regulated in CD4⁺ or CD14⁺ cells, we followed two approaches. The first approach consisted of identifying probe sets that demonstrated the largest difference in prednisolone effects between both cell types. It appeared that a large number of the probe sets that were identified using this criteria demonstrated a regulation in both cell types, but differed only in the strength of the regulation. *GAS7* is an example of such a gene (Figure 3, left panel). This indicates that the difference between the cell types is caused more by a difference in magnitude of the prednisolone effect rather than an on–off effect in which a prednisolone effect in one of the cell types is completely absent. To identify the latter category, which we regard as a better representation of cell-specific probe sets, we selected probe sets that demonstrated regulation in one cell type (using the criteria stated earlier) and an absence of regulation in the other (defined as a log2 ratio of <0.1). This analysis yielded 62 CD4⁺ specific probe sets and 14 CD14⁺ specific probe sets (Table 1, Supplementary Tables 2 & 3). Both cell-specific probe sets included many genes related to inflammation and immune regulation. However, a link between gene function and cell specificity could be made for a number of genes. For example, *FCER1G* and *IFNGR1*, regulated in CD4⁺ cells but not in CD14⁺ cells, are involved in T-helper cell signaling and differentiation [23–25],

**Figure 1** Distribution of ratios of differentially expressed probe sets in CD4⁺ and CD14⁺ subsets under both conditions (30 mg and 75 mg prednisolone). The ordered gene index shown on the x-axis presents all the probe sets that were differentially regulated on day 2 compared with day 1. These probe sets were ordered on the basis of the fold change between day 2 and day 1.
Chapter 4

Figure 2 Cortisol and prednisolone levels in acute protocol. The solid lines show the prednisolone levels during the course of the trial. Prednisolone was administered at 8 am on day 1 at two doses: 2 × 15 mg (second dose at 8 pm; lighter colored line) and 75 mg (darker colored line) as indicated by the arrows. Cortisol levels (dashed lines) were suppressed for almost 20 h by these doses of prednisolone compared with the levels at day 0 after placebo administration. Samples for gene-expression analysis were taken predose at day 1 and after 24 h.

Figure 3 Examples of genes which show differences in prednisolone effects concerning magnitude or cell specificity. (A) Example of a gene which is regulated by prednisolone in both CD4+ and CD14+ cells but shows a difference in the magnitude of expression after treatment. (B) Example of a gene which demonstrates expression in both cell types (cut-off: intensity >20) but is regulated by prednisolone only in CD4+ cells (CD4+ specific). (C) Example of a gene which demonstrates expression in both cell types (cut-off: intensity >20) but is regulated by prednisolone only in CD14+ cells (CD14+ specific).

Table 1 Number of cell-specific regulated probe sets. Cut-off for determining whether a probe set was expressed or not was set at an average intensity of 20. Numbers denote probe sets that demonstrated a log ratio of <0.1 in the other cell type (corresponding to a fold change of <1.07)

<table>
<thead>
<tr>
<th>Cells in which specific regulation occurs</th>
<th>Total number of regulated genes</th>
<th>Genes expressed in other cell type</th>
<th>Genes not expressed in other cell type</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD4+</td>
<td>62</td>
<td>33</td>
<td>29</td>
</tr>
<tr>
<td>CD14+</td>
<td>15</td>
<td>1</td>
<td>14</td>
</tr>
</tbody>
</table>
processes that are predominantly mediated by CD4+ T lymphocytes. The genes *PECAM1* and *CXCL9*, regulated in CD14+ cells but not in CD4+ cells, are known to have a role in antigen presentation and phagocytosis, which are functions mainly performed by CD14+ cells [26,27].

The relatively small sets of cell-type specific regulated genes indicate that only a minority of the genes demonstrate true cell specific behavior and that the majority of prednisolone-regulated genes are regulated in both cell types.

**Table 2** Top five over-represented pathways as identified by Ingenuity Pathway Analysis and CoPub Analysis. The enrichment was calculated for 892 probe sets that were differentially regulated in CD4+ or C14+ samples. The total list of over-represented pathways is outlined in Supplementary Table 4.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ingenuity Pathway Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>TREM1 Signaling</td>
<td>2.45E-09</td>
</tr>
<tr>
<td>Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses</td>
<td>7.41E-08</td>
</tr>
<tr>
<td>B Cell Receptor Signaling</td>
<td>2.00E-07</td>
</tr>
<tr>
<td>Production of Nitric Oxide and Reactive Oxygen Species in Macrophages</td>
<td>5.75E-07</td>
</tr>
<tr>
<td>Toll-like Receptor Signaling</td>
<td>2.51E-06</td>
</tr>
<tr>
<td><strong>CoPub Analysis</strong></td>
<td></td>
</tr>
<tr>
<td>cell adhesion</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>neutrophil activation</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>respiratory burst</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>t-cell proliferation, t-cell homeostatic proliferation</td>
<td>9.79E-11</td>
</tr>
<tr>
<td>cell development</td>
<td>9.79E-11</td>
</tr>
</tbody>
</table>

**Pathways affected by prednisolone treatment**

In order to investigate which biological pathways were mostly affected by prednisolone, we used Ingenuity Pathway Analysis software version 8.8 (Ingenuity systems, CA, USA) and CoPub ([28,102]) to map the regulated probe sets to canonical pathways. The top five over-represented pathways as identified by Ingenuity Pathway Analysis and CoPub analysis is outlined in Table 2, the total list is outlined in Supplementary Table 4. The top scoring terms appeared to have a strong immunological component, exemplified by pathways such as Toll-like receptor signaling, acute phase response, T-cell and B-cell signaling. Also a link to autoimmune diseases such as systemic lupus erythematosus and rheumatoid arthritis was observed. In addition, a link to specific signaling pathways
via NF-κB, IL-8, IL-6 and GM-CSF was observed, all of which are signaling pathways that play an important role in inflammatory processes. This indicates that the main effects that are induced by prednisolone in CD4+ and CD14+ are related to the pharmacological, anti-inflammatory properties of prednisolone. Interestingly, these effects can thus be observed in healthy, unchallenged volunteers.

Text-mining analysis to identify genes that are involved in prednisolone-induced metabolic side effects

The above mentioned results demonstrated that we can clearly identify GC signatures in both CD4+ and CD14+ cell types and that these signatures have a strong link to immunological processes. The above analysis is, however, an enrichment analysis in which only the most affected pathways are identified. In the case that only a small, but possibly crucial, number of genes in a pathway are regulated, this pathway will not receive a high enrichment score and might therefore easily be missed when the results are interpreted.

To make a link to possible prednisolone-induced metabolic side effects, we set out to identify genes that are regulated by prednisolone and are involved in metabolic pathways and processes. To perform this in an unbiased way we used CoPub to identify from the set of regulated genes, all genes with a strong literature link to keywords describing glucose and fatty-acid metabolism. Since we were also interested in the interplay between immunological and metabolic processes and the possible relation with GC regulation, we also examined for each of the genes whether there was a strong known link with keywords describing GR-mediated processes and immunological pathways. The keywords that were used are presented in Supplementary Table 5.

We identified 253 genes with a link to one or more of the metabolic keywords that were used for searching. The top scoring metabolic genes (34 genes) are shown in Figure 4, together with their literature scores. A short description regarding the function of ten of these 34 genes is outlined in Table 3; in Supplementary Table 6 the function of all 34 genes is summarized. It appeared that a large number of these genes had not only a strong link to metabolic processes but were also connected to inflammatory processes. For example, MAPK14 (p38 MAP kinase) is involved in signal transduction of inflammatory stimuli, but has also been reported to play a role in insulin-induced glucose transport in muscle [29]. Other genes with a strong link to both metabolic and inflammatory processes are MAPK1 and IL-1B. However, for a number of genes an immunological function is less obvious, and these may represent a set of genes that have a predominantly metabolic function and are less important for the anti-inflammatory action of prednisolone. Examples of such genes are genes related to fatty acid transport and metabolism: ABCA1, ACOX1, PDK4, ACSL1 and EPHX2 [30–36], and genes related to (regulation of) glucose and glycogen metabolism: PYGL, GYG1, SLC16A3, PFKFB3, PFKFB2, AQP9, GK and SLC2A3 (also known as GLUT3) (Figure 4) [32,35,37–42]. These results clearly indicate that the glucose and
lipid metabolic signaling pathways, in which these genes play a role, are affected by GC treatment.

Some of the above mentioned genes, for example \textit{PYGL}, \textit{GYG1} and \textit{AQP9} display a weak literature link with GC. A detailed search in PubMed confirms that these genes indeed have not been reported to be under control of GCs. Interestingly, in promoter regions of human and pig \textit{AQP9}, putative GC response elements have been identified \cite{43,44} indicating that this gene might be under the direct control of GCs. A metabolic role for \textit{GYG1} and \textit{PYG1} in PBMCs is supported by the fact that these genes were also strongly

Figure 4 Text-mining analysis to identify genes involved in prednisolone-induced metabolic side effects. Genes depicted in this list are sorted according to high literature co-occurrence score for keywords involved in metabolism (cut-off ≥0.5).
Table 3 Function of ten of 34 genes that were identified by text mining. Information was obtained from two sources. Information on the general function of the gene was obtained from the EntrezGene database from the NCBI (http://www.ncbi.nlm.nih.gov/gene). The link between the gene and metabolic processes was obtained from manual inspection of the Pubmed abstracts returned by the CoPub text mining strategy, i.e., those abstracts in which the genes and metabolic keywords co-occur which is the underlying evidence for the literature scores depicted in figure 4. The total gene list of 34 genes is outlined in supplementary table 6.

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene name</th>
<th>Function</th>
<th>Role in metabolism</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PYGL</strong></td>
<td>Phosphorylase, glycogen, liver</td>
<td>Encodes a homodimeric protein that catalyses the cleavage of alpha-1,4-glucosidic bonds to release glucose-1-phosphate from liver glycogen stores</td>
<td>Humans have three glycogen phosphorylase isozymes that are primarily expressed in liver, brain and muscle, respectively. The liver isozyme serves the glycemic demands of the body in general while the brain and muscle isozymes supply just those tissues. In glycogen storage disease type VI, or Hers disease, mutations in liver glycogen phosphorylase inhibit the conversion of glycogen to glucose and results in moderate hypoglycemia, mild ketosis, growth retardation and hepatomegaly</td>
</tr>
<tr>
<td><strong>GYG1</strong></td>
<td>Glycogenin 1</td>
<td>Encodes a member of the glycogenin family. Glycogenin is a glycosyltransferase that catalyzes the formation of a short glucose polymer from uridine diphosphate glucose in an autoglucosylation reaction</td>
<td>This gene is expressed in muscle and other tissues and has a key role in glycogen formation. Mutations in this gene result in glycogen storage disease XV</td>
</tr>
<tr>
<td><strong>AQP9</strong></td>
<td>Aquaporin 9</td>
<td>The aquaporins are a family of water-selective membrane channels. The protein encoded by this gene allows passage of a wide variety of noncharged solutes</td>
<td>AQP9 stimulates urea transport and osmotic water permeability; there are contradicting reports about its role in providing glycerol permeability. It may also play a role in specialized leukocyte functions such as immunological response and bactericidal activity</td>
</tr>
<tr>
<td><strong>PDK4</strong></td>
<td>Pyruvate dehydrogenase kinase, isozyme 4</td>
<td>PDK4 is located in the matrix of the mitochondria and inhibits the pyruvate dehydrogenase complex by phosphorylating one of its subunits</td>
<td>By inhibiting the pyruvate dehydrogenase complex it is contributing to the regulation of glucose metabolism. Expression of this gene is regulated by glucocorticoids, retinoic acid and insulin</td>
</tr>
<tr>
<td><strong>GLUL</strong></td>
<td>Glutamate-ammonia ligase</td>
<td>The protein catalyzes the synthesis of glutamine from glutamate and ammonia</td>
<td>Glutamine is a main source of energy and is involved in cell proliferation, inhibition of apoptosis, and cell signaling</td>
</tr>
<tr>
<td><strong>ACSL1</strong></td>
<td>Acyl-CoA synthetase long-chain family member 1</td>
<td>Isozyme of the long-chain fatty-acid-coenzyme A ligase family. All isozymes of this family convert free long-chain fatty acids into fatty acyl-CoA esters</td>
<td>Plays a key role in lipid biosynthesis and fatty acid degradation</td>
</tr>
</tbody>
</table>
### Table: Prednisolone-induced changes in healthy volunteers

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>Gene name</th>
<th>Function</th>
<th>Role in metabolism</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEBPB</td>
<td>CCAAT/enhancer binding protein (C/EBP), beta</td>
<td>The protein encoded by this intronless gene is a bZIP transcription factor which can bind as a homodimer to certain DNA regulatory regions</td>
<td>Polymorphisms in CEBPB are associated with abdominal obesity and related metabolic abnormalities like type 2 diabetes</td>
</tr>
<tr>
<td>IL1B</td>
<td>Interleukin 1, beta</td>
<td>This cytokine is an important mediator of the inflammatory response, and is involved in a variety of cellular activities, including cell proliferation, differentiation, and apoptosis</td>
<td>Inflammation is associated with obesity, the metabolic syndrome, and diabetes. Weight reduction results in a decrease in the mRNA expression of IL-1beta (IL1B), IL-1 receptor antagonist, and tumor necrosis factor alpha. Decrease in IL1B expression is correlated with an increase in insulin sensitivity index. The strong correlation between the decrease in IL1B expression and the increase in insulin sensitivity suggest a contribution of IL1B to insulin-resistant states found in obesity and the metabolic syndrome.</td>
</tr>
<tr>
<td>MAPK14</td>
<td>Mitogen-activated protein kinase 14</td>
<td>Member of the MAP kinase family. MAP kinases act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development</td>
<td>Regulates the activities of many transcription factors and proteins/enzymes and thus has a wide-spectrum of biological effects. Patients with insulin resistance and/or type 2 diabetes have high levels of plasma free fatty acids, inflammatory cytokines, and/or glucose, and over-activation of the cardiovascular renin-angiotensin system, all are capable of activating MAPK14. MAPK14 plays a central role in hepatic glucose and lipid metabolism, leading to increased hepatic glucose production and decreased hepatic lipogenesis. The role of MAPK14 in insulin-mediated glucose uptake in skeletal muscle and adipose tissue remain controversial</td>
</tr>
<tr>
<td>MAPK1</td>
<td>Mitogen-activated protein kinase 1</td>
<td>Member of the MAP kinase family. MAP kinases act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development</td>
<td>Recent evidence suggests that common stress-activated signaling pathways such as nuclear factor-kappaB, MAPK1, and NH2-terminal Jun kinases/stress-activated protein kinases underlie the development of diabetic complications. In type 2 diabetes, there is evidence that the activation of these same stress pathways by glucose and possibly FFA leads to both insulin resistance and impaired insulin secretion</td>
</tr>
</tbody>
</table>
regulated in PBMCs of healthy volunteers after a high calorie breakfast [45].

**Discussion**

It is well known that GCs are very efficacious drugs in humans and that prolonged use of GCs induces metabolic side effects in humans. However, knowledge of the effects of GCs on immune-related and metabolic pathways in an *in vivo* setting is relatively scarce. Therefore, we conducted genome-wide profiling of gene-expression in human healthy volunteers that were treated with prednisolone as a typical, broadly used GC. To our knowledge this is the first study in which gene-expression profiles in prednisolone-treated healthy volunteers is described. As a first aim of this study, we identified prednisolone-induced gene signatures in CD4+ T lymphocytes and CD14+ monocytes, cells which are known to be peripheral immune target cells for GC action. These expression signatures will provide insight into the effects of GCs in an *in vivo* setting. We observed a much stronger induction of gene-expression in CD4+ T lymphocytes than in CD14+ monocytes with respect to fold changes, but the number of truly cell-specific genes regulated by prednisolone was relatively low. Tissue-specific regulation is known for GCs and large differences between expression profiles of liver, muscle and kidney can be observed [46–48]. Genes such as *TAT*, *G6PC* and *PCK1* are specifically regulated in the liver, and the interplay between GR- and tissue-specific cofactors have been studied in detail [49–51]. However, to the best of our knowledge, no literature studies exist in which *in vivo* GC regulation in monocytes and T cells have been compared. Apparently, the tissue specificity of GC is less pronounced between cells from the same organ and lineage. It should, however, be noted that, although we found a strong prednisolone effect on immune-related pathways, these cells were derived from healthy volunteers; in diseased people the effect of GCs might well be larger and more cell specific than observed in our study.

We found that classical GC target genes *FKBP5* and *ZBTB16* were downregulated. The product of *FKBP5* functions as a co-chaperone molecule of the GR that affects the transport of GR into the nucleus. *ZBTB16* encodes a zinc finger transcription factor that contains nine Kruppel-type zinc finger domains at the carboxyl terminus. This protein is located in the nucleus and is involved in cell cycle progression. Expression of both genes have previously been reported as being strongly induced shortly after GC administration in both cell lines and primary patient cells [12,52–55]. The main difference between those studies and our study is that we studied prednisolone-induced gene-expression in a complex *in vivo* setting in contrast to other studies in which most often cell lines, animals or cultured primary cells, were investigated [12–14,20–22]. Interestingly, other known target genes of GR (for example *GLUL*) demonstrated an upregulation by prednisolone after 24 h, whereas other known GC target genes such as *DUSP1* and *GILZ* were not regulated by prednisolone in this study. The exact interplay between exogenous and
endogenous cortisol levels (which were strongly reduced after 24 h), kinetics of induction of gene-expression and mRNA stability determines the effect of GC on its target genes in a complex \textit{(in vivo)} setting and can only partly be replicated in isolated \textit{in vitro} studies. This demonstrates that when using the expression of GR target genes as biomarkers for GR-induced processes, either time resolved measurements are needed or markers should be chosen that are relatively independent of the time window at which sampling is taking place.

In our study, we used purified CD4$^+$ and CD14$^+$ cells to study the effect of prednisolone. For large scale clinical trials, gene-expression studies with whole-blood samples are preferred because of ease of handling and reproducibility of the results [56]. A potential problem in measuring prednisolone-induced changes in gene expression in whole blood is that direct effects of prednisolone on gene transcription may be confounded by the shifts in T cell and monocyte subsets that can be induced by prednisolone. In the case that prednisolone would act on largely distinct gene sets for each cell type, changes in whole-blood gene expression could be unambiguously assigned to a specific cell type. However, the fact that we observed a large overlap in gene sets regulated by prednisolone in CD4$^+$ and CD14$^+$ cells indicates that when measuring prednisolone-induced gene-expression in whole blood, cell counts of leukocyte subsets populations or gene-expression markers in these subsets need to be measured to account for shifts in cell populations [57].

Clinical measurements that were carried out during this trial demonstrated that both acute and short-term exposure to different prednisolone impaired aspects of pancreatic $\beta$-cell function, such as glucose sensitivity and fasting insulin secretion, contribute to a diabetic phenotype [15]. Therefore, we used the expression profiles obtained from CD4$^+$ and CD14$^+$ cells to study whether we could find a link to pathways and mechanisms that might be involved in prednisolone-induced metabolic adverse events. Although blood cells are not the tissues most relevant for metabolic side effects (these are most likely liver, pancreas, muscle and adipose tissue), there is growing evidence that blood cells can act as sentinels of disease [58–61]. The permanent interactions between blood cells and every tissue in the body make it plausible that subtle changes occurring in relation to disease within cells or tissues of the body trigger specific changes in gene-expression in blood cells, thereby reflecting the disease state of affected cells and tissues. Indeed, Liew and colleagues demonstrated that over 80% of the genes expressed in other tissues overlapped with the genes expressed in peripheral blood [58]. These findings suggest that peripheral bloodexpression profiles can provide information concerning the health or disease state of any particular tissue in the body and may serve as a tool for diagnostic purposes.

Text-mining analysis identified 253 prednisolone-regulated genes with a metabolic link. Among the top scoring genes were a number of genes that play a crucial role in the regulation of glucose and lipid metabolism (Figure 4). From these genes, \textit{PDK4} and \textit{ACSL1} are particularly interesting. \textit{PDK4} is a crucial regulator of glucose and lipid metabolism.
In the well-fed state, the pyruvate dehydrogenase complex reduces blood glucose levels by converting it to acetyl-coenzyme A which will enter the citric acid cycle. In the fasting state, inactivation of the pyruvate dehydrogenase complex is mainly achieved by PDK4, thereby maintaining blood glucose levels [62]. Increased PDK4 activity has been implicated in the pathogenesis of insulin resistance, diabetes and obesity [62,63]. Our results demonstrate an upregulation of PDK4 in CD4+ and CD14+ cells after prednisolone treatment. These results imply that prednisolone-induced insulin resistance might partly be mediated through PDK4 regulation. ACSL1 encodes an isozyme of the long-chain fatty acid-CoA ligase family that converts free fatty acids into fatty acyl-CoA esters that can be used for lipid biosynthesis and oxidation. In adipose tissue, insulin signaling leads to the translocation of ACSL1 to the membrane, together with other enzymes involved in fatty acid binding and translocation. SNPs in this gene have been demonstrated to be associated with metabolic syndrome [32]. It was observed that ACSL1 expression in subcutaneous adipose tissue was negatively correlated with insulin resistance [64]. Whereas the regulation of PDK4 and ACSL1 in metabolic tissues such as the muscle, adipose tissue and the liver, and the effect of GC on this regulation, has been well studied [65], relatively little is known regarding the regulation of these genes in T cells or monocytes.

The results obtained by text-mining analyses identify interesting genes that might be involved in GC-induced metabolic adverse events. Expression-profiling studies demonstrated that several of the genes identified in this study are also regulated in the muscle of collagen-induced arthritic mice after treatment with prednisolone [Van Lierop et al., Unpublished Data]. Genes that were regulated after prednisolone treatment are AQP9, PDK4, GLUL, CEBPB and CD14.

Taken together, these results indeed demonstrate that prednisolone has a major impact on metabolic processes in blood cells. Whether this is caused by a direct effect of prednisolone on these cells or whether this is a reaction of the blood cells to a changed metabolic state of the primary metabolic organs such as the liver, remains to be studied.

A number of genes that were identified via text mining also revealed prednisolone-induced genes with a metabolic function that had not yet been linked to GCs in the literature. These genes might be interesting new candidate genes for further investigation of GC-induced metabolic side effects. This demonstrates that a combined transcriptomics and text-mining approach may reveal new genes and pathways involved in GC-induced metabolic side effects and provide new hypotheses for follow-up research.

**Conclusion**

By using whole-genome expression profiling, we generated expression signatures in blood CD4+ and CD14+ cells that reflect anti-inflammatory effects but also metabolic side effects. Our findings demonstrate that prednisolone induces a very specific profile
in both cell types, consisting of a large number of primary GC-responsive genes that are regulated in both cell types, as well as a set of cell-type-specific genes. Combined with the fact that prednisolone-induced pancreatic b-cell failure was observed in these volunteers, these findings suggest that peripheral blood-expression profiles can provide information on the prednisolone-induced metabolic state. The identified genes may be useful as biomarkers for GC-induced side effects or provide a starting point for the development of improved GCs that demonstrate the same anti-inflammatory effect but have a better safety profile.

Financial & competing interests disclosure
Clinical trial number: NCT00971724. This work was performed within the framework of Dutch Top Institute Pharma, project ‘Glucocorticoid-induced insulin resistance’ (project nr. T1-106-1). Ulla Nässander, Marie-José C van Lierop, Susanne Bauerschmidt, Wim HA Dokter and Wynand Alkema are all employees of Merck, Sharp and Dohme. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed. No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research
The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

Supplementary information can be found online at the website of the publisher or can be retrieved on request by the author of this thesis.
Chapter 4

References

Papers of special note have been highlighted as:
* of interest
** of considerable interest

46. Chrousos GP, Kino T: Glucocorticoid signaling in the cell. Expanding clinical implications to complex


* One of the first papers that identifies FKBP51 as a biomarker for glucocorticoid sensitivity.


** Results from this study show that peripheral blood cells reflect the disease state of other less accessible tissues. This suggests that easily accessible blood cells can be used to monitor disease state, which might be a valuable tool for diagnostic purposes.


Websites
101. The R Project for Statistical Computing
     www.r-project.org
102. CoPub
     www.copub.org
103. The National Center for Biotechnology Information
Prednisolone induces the Wnt signaling pathway in 3T3-L1 adipocytes

Wilco W.M. Fleuren1,2, Margot M.L. Linssen3, Erik J.M. Toonen4, Gerard C.M. van der Zon3, Bruno Guigas3,5 Jacob de Vlieg1,6, Wim H.A. Dokter7, D. Margriet Ouwens8,9 and Wynand Alkema1,a

1CDD, CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands, 2Netherlands Bioinformatics Centre (NBIC), Nijmegen, The Netherlands, 3Department of Molecular Cell Biology, Leiden University Medical Center, Leiden, The Netherlands, 4Department of Medicine, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands, 5Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands, 6Institute of Clinical Biochemistry and Pathobiochemistry, German Diabetes Center, Düsseldorf, Germany. 7Syntht biooharmaceuticals BV, Nijmegen, The Netherlands, 8Institute of Clinical Biochemistry and Pathobiochemistry, German Diabetes Center, Düsseldorf, Germany. 9Department of Endocrinology, Ghent University Hospital, Ghent, Belgium

aPresent address: NIZO Food Research BV, Ede, The Netherlands.

Accepted for publication in Arch Physiol Biochem. 2013
Abstract

Synthetic glucocorticoids are potent anti-inflammatory drugs but show dose dependent metabolic side effects such as the development of insulin resistance and obesity. The precise mechanisms involved in these glucocorticoid-induced side effects, and especially the participation of adipose tissue in this are not completely understood.

We used a combination of transcriptomics, antibody arrays and bioinformatics approaches to characterize prednisolone-induced alterations in gene expression and adipokine secretion, which could underlie metabolic dysfunction in 3T3-L1 adipocytes. Several pathways, including cytokine signaling, Akt signaling, and Wnt signaling were found to be regulated at multiple levels, showing that these processes are targeted by prednisolone.

These results suggest that mechanisms by which prednisolone induce insulin resistance include dysregulation of wnt signaling and immune response processes. These pathways may provide interesting targets for the development of improved glucocorticoids.

Keywords: glucocorticoids, metabolic dysfunction, gene profiling
Introduction

Synthetic glucocorticoids (GCs) such as prednisolone and dexamethasone are widely used for the treatment of inflammatory diseases such as rheumatoid arthritis, asthma, inflammatory bowel disease and psoriasis [1-3]. Despite their excellent efficacy, GC usage is hampered because of adverse (metabolic) side effects, such as insulin resistance, glucose intolerance, diabetes, central adiposity, dyslipidemia, skeletal muscle wasting and osteoporosis [4-8]. The precise mechanisms involved in these GC-induced side effects are not completely understood and are tissue specific [9-12]. One of the key tissues thought to be involved in metabolic side effects is the adipose tissue. Several studies have shown that the side-effects of GCs in adipose tissue include the development of central adiposity [7], dyslipidaemia [13, 14] and the inhibition of insulin-stimulated glucose uptake [15-18]. It is now well established that adipose tissue is not only involved in energy storage, but also functions as an endocrine organ that secretes hundreds of bioactive substances known as adipokines [19-21]. Previously, GCs have been found to affect the secretion of several adipokines, including TNF-α, IL-6, Resistin, Adiponectin and Leptin in rodents [22]. Because recent proteomic approaches have led to the characterization of numerous novel adipokines, we applied a combination of genomics, antibody arrays and bioinformatics approaches to identify prednisolone-induced alterations in (adipokine) gene expression and secretion which could potentially underlie prednisolone-induced metabolic dysfunction of 3T3-L1 adipocytes.

Materials and Methods

Set up of the study

In this study, differentiated 3T3-L1 adipocytes were treated with prednisolone or DMSO for 0, 1, 6, 24 and 48 hours and the effects of prednisolone were studied at gene expression and protein secretion level.

Cell culture

3T3-L1 pre-adipocytes obtained from ATCC were cultured and differentiated in adipocytes as previously described [23]. Cells were used seven days after completion of the differentiation process. Only cultures in which >95% of cells displayed adipocyte morphology were used. Prior to use, adipocytes were serum starved for 16 hours with DMEM supplemented with 0.5% fetal bovine serum.

Analysis of insulin signaling

Differentiated 3T3-L1 adipocytes, grown in 12-well plates, were incubated with prednisolone (1 μM) or DMSO at day 8 for 48 hours. After incubation with prednisolone or DMSO, cells were serum-starved for 2 hours, and then stimulated with insulin (100
Prednisolone induces the Wnt signaling pathway in 3T3-L1 adipocytes

nM) for 10 minutes. Following insulin stimulation, cells were washed twice with ice-cold PBS, and lysed as described previously [24]. Protein expression and phosphorylation were determined by Western blotting as described [24] using the following antibodies: phospho-insulin receptor substrate 1-Tyr1222, phospho-Akt-Thr308, phospho-Akt-Ser473 (all from Cell Signalling Technology, Danvers, MA, USA), total Akt, phospho-Akt substrate of 160-kDa (AS160), total AS160 (all from Millipore, MA, USA), glucose transporter 4 (GLUT4), and β-actin (both from Abcam, Cambridge, UK).

Assay of 2-deoxy-D-glucose (2DOG) uptake
Differentiated 3T3-L1 adipocytes, grown in 12-wells plates, were incubated with prednisolone (1 µM) or DMSO at day 8 for 6, 24 and 48 hours. After incubation with prednisolone or DMSO for the indicated times, the cells were washed once with HEPES buffer, consisting of 50 mM HEPES, 0.14 M NaCl, 1.85 mM CaCl₂, 1.3 mM MgSO₄, and 4.8 mM KCl [pH 7.4]), and then incubated in this buffer for an additional hour at 37°C. Then, cells were stimulated with insulin (100 nM) or kept untreated. After 15 min, 2DOG uptake was initiated by the addition of 2-deoxy-D-[¹⁴C]glucose (0.075 µCi per well) in 3 mM 2DOG. After 10 min, the assay was terminated by three quick washes with ice-cold PBS. Cells were lysed in 0.1 M NaOH, 0.2% SDS, whereafter incorporated 2-deoxy-D-[¹⁴C] glucose was determined by liquid scintillation counting.

RNA isolation
Differentiated 3T3-L1 adipocytes grown in 6-well plates were incubated with prednisolone (1 µM) for 1, 6, 24 and 48 hours and used for RNA isolation and whole-genome expression profiling. RNA was isolated using Tripure (Roche), and subsequently purified using the RNeasy mini kit according to the manufacturer’s protocol (Qiagen Benelux B.V. Venlo, The Netherlands). To remove residual traces of genomic DNA, the RNA was treated with DNase I (Invitrogen, Leek, The Netherlands) while bound to the RNeasy column. Quality and quantity of the purified RNA was controlled using a NanoDrop spectrophotometer (Nanodrop technologies, Montchanin, DE, USA). RNA integrity was investigated by using the 2100 Bioanalyser (Agilent technologies, Philadelphia, PA, USA).

Whole-genome expression analysis using Affymetrix GeneChip Mouse Genome 430 2.0 Array
Amplification of 20 ng total RNA was performed with the Two-Cycle Eukaryotic Target Labeling kit (Affymetrix, Santa Clara, CA). Briefly, after the first round of cDNA synthesis, an unlabeled ribonucleotide mix (MEGAscript T7 kit, Ambion) was used to generate unlabeled cRNA according to the protocol of Affymetrix. After cleanup of the cRNA with a GeneChip Sample Cleanup Module IVT Column (Affymetrix, Santa Clara, CA), the unlabeled cRNA concentration was determined and 150 ng was reverse transcribed using random primers. Subsequently, the T7-Oligo(dT) Promoter Primer was used in
the second-strand cDNA synthesis to generate double-stranded cDNA containing T7 promoter sequences. The resulting double-stranded cDNA was then amplified and biotin labeled using the IVT Labeling kit (Affymetrix, Santa Clara, CA). Biotin-labeled cRNA was fragmented at 1 μg/μl following the manufacturer’s protocol. After fragmentation, cRNA (10 μg) was hybridized at 45°C for 16 hours to the Human Genome U133 2.0 Plus array (Affymetrix, Santa Clara, CA). Following hybridization, the arrays were washed, stained with phycoerythrin-streptavidin conjugate (Molecular Probes, Eugene, OR), and the signals were amplified by staining the array with biotin-labeled anti-streptavidin antibody (Vector Laboratories, Burlingame, CA) followed by phycoerythrin-streptavidin. The arrays were laser scanned with a GeneChip Scanner 3000 7G (Affymetrix, Santa Clara, CA) according to the manufacturer’s instructions. Data was saved as raw image file and quantified using AGCC v 1.0 (Affymetrix, Santa Clara, CA).

Murine adipokine antibody arrays
This array allows simultaneous detection of 308 secreted mouse proteins in cell culture supernatants (AAM-BLM, L-series, Raybiotech, Inc, Norcross GA, USA). Differentiated 3T3-L1 adipocytes were subjected to 4 hours pred/dmso, followed by two rinsing steps and culture medium containing 0.2% FBS and Prednisolone or vehicle (1 • M) for 44 hours. To prepare samples for the antibody array, samples were subjected to centricron tubes (Millipore amicon ultra-15 ultracel 3K) and washed 3 times with PBS, pH 8.0. Finally proteins were recovered in 2 ml PBS. Samples were biotin-labelled according to manufacturer’s protocol. Then, membranes were blocked and incubated with the biotinylated samples overnight at 4ºC, and finally incubated with HRP conjugated streptavidin. Signals were detected using enhanced chemiluminescence and quantified on a LumilMager system using LumiAnalyst software version 3.1 (Roche Diagnostics, Mannheim, Germany).

Gene expression data
Microarray data were analyzed with packages from the BioConductor library[25]. Gene expression data was normalized using gcrma. Identification of regulated probes was done using limma package from Bioconductor. For identifying differentially expressed probes after incubation with prednisolone, the following selection criteria were used: fold change >2, intensity > 20, and p-value<0.05 after correction for multiple testing using the Benjamini-Hochberg correction. All statistical analysis was performed in R. Significance of regulation of the Wnt pathway has been calculated using globaltest.

Secreted protein data obtained from murine adipokine antibody arrays
Regulated proteins as a result of prednisolone treatment have been selected using a P-value < 0.05 (calculated using a standard Student's t-Test).
Keyword enrichment analysis

Keyword enrichment analysis on the microarray data was performed using CoPub, a text mining algorithm that detects co-occurring biomedical concepts in abstracts from the MedLine literature database [26], using the following settings for the threshold values: p-value <0.05, R-scaled score > 35 and abstract count > 3. A heatmap of the pathway terms was been generated using a transformation of the p-value in the following way:

\[
\text{Pval}_{\text{transform}} = \frac{1}{(1 + 6 \times \exp(-0.35 \times \log_{10}(p\text{-value})))}
\]

Additional gene annotation enrichment analysis was done using the functional annotation module of Database for Annotation, Visualization and Integrated Discovery (DAVID) tool [27, 28] with default settings against the KEGG pathway database.

Results

Effect of prednisolone on insulin action in 3T3-L1 adipocytes

Exposure of differentiated 3T3-L1 adipocytes to 1 µM prednisolone had no effect on the expression of key signaling molecules, like the insulin receptor β-subunit, Akt, AS160 and GLUT4. Prednisolone significantly impaired the phosphorylation of Akt-Thr308 and Akt-Ser473 in response to insulin, whereas no effects were observed on IRS1-Tyr1222 (Figure 1). Intriguingly, but in line with a study on the effects of dexamethasone on insulin action in 3T3-L1 adipocytes [29], prednisolone did not impact on insulin-induced phosphorylation of AS160-Thr642 (Figure 1). The mechanism(s) underlying this discordance in signaling from Akt to AS160 is not understood. Nevertheless, prednisolone was found to impair insulin-stimulated glucose uptake already after 6 hours of incubation, and this inhibition was still present at 48 hours after incubation (Figure 1). These findings are in line with earlier observations using dexamethasone in 3T3-L1 adipocytes [15, 29] and confirm the induction of insulin resistance by prednisolone in our experimental conditions.

Whole genome expression profiling

We next studied prednisolone-induced changes in gene expression following exposure of 3T3-L1 adipocytes to prednisolone or DMSO for 1, 6, 24 and 48 hours. Principle Component Analysis (PCA) showed a separation between prednisolone- and DMSO-treated samples after 6, 24 and 48 hours (Supplementary Figure 1). At each time-point, the significantly differentially expressed probes were identified between cells treated with prednisolone and DMSO (Supplementary Table 2). Comparison of probes significantly affected by prednisolone at 6 hours with probes significantly affected by prednisolone at 48 hours
showed that probes with a high fold change are found to be regulated at both time points (Figure 2). Probes which are only regulated at 6 hours or at 48 hours have a lower fold change.

The same trend was observed when comparing prednisolone induced probes at 6 hours with prednisolone induced probes at 24 hours (Figure 2B). At all time points, 25 genes represented by 34 probes (in fact 26 genes were found, but for probe 1434025_at we could not find annotation), were found to be significantly regulated by prednisolone
Most of these genes were upregulated and showed a sustained response to prednisolone at all time points, starting after one hour of incubation, with the highest response after 6 hours, and still present after 48 hours of incubation.

**Biological pathways targeted by prednisolone**

To identify biological pathways that were affected by prednisolone, we mapped the regulated genes to biological pathway terms using the text mining algorithm CoPub [26] (Figure 3). It appeared that the most significant pathway terms could be divided into three categories: immune system/inflammation (eicosanoid metabolism, prostaglandin metabolism, cytokine receptors, cytokine network, immune system, platelet activation), general metabolism (fatty acid metabolism, glucose metabolism, gluconeogenesis, lipid metabolism, fatty acid metabolism, glycolysis) and signaling (insulin signaling, TGF-β signaling, Akt signaling, Wnt signaling, MAPK signaling). Additional functional analysis of these regulated genes using DAVID, an annotation server based on GeneOntology and KEGG pathways [27], gave similar results (results not shown).

We are interested in mechanisms and pathways that could underlie prednisolone-induced metabolic dysfunction and more specific prednisolone-induced insulin resistance in adipocytes. Therefore for the remainder of the paper, we focused on pathways that are known to play a role in the disturbance of insulin signaling that eventually could lead to insulin resistance, i.e. Akt/insulin signaling, cytokine signaling, TGF-β signaling and MAPK signaling. Furthermore, we also focused on pathways that play a role in the disturbance of adipocyte differentiation that could lead to dyslipidaemia and obesity, i.e. Wnt signaling and cytokine signaling. In Figure 4 the prednisolone-induced genes have been categorized according to these pathways. Genes that have an effect on insulin signaling, e.g. growth-
Table 1: Regulated probes (genes) by prednisolone after 1, 6, 24 and 48 hours of incubation. For each probe the log fold change at each time point is shown. Probes that are up-regulated are indicated with a red color and probes that are down-regulated are indicated with a green color.

<table>
<thead>
<tr>
<th>Affymetrix ID</th>
<th>EntrezGene ID</th>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>PRED T1 - DMSO T1</th>
<th>PRED T6 - DMSO T6</th>
<th>PRED T24 - DMSO T24</th>
<th>PRED T48 - DMSO T48</th>
</tr>
</thead>
<tbody>
<tr>
<td>1451021_a_at</td>
<td>12224</td>
<td>Klf5</td>
<td>Kruppel-like factor 5</td>
<td>1.92E+00</td>
<td>3.61E+00</td>
<td>2.62E+00</td>
<td>4.32E+00</td>
</tr>
<tr>
<td>1423233_at</td>
<td>12609</td>
<td>Cebpd</td>
<td>CCAAT/enhancer binding protein (C/EBP), delta</td>
<td>1.97E+00</td>
<td>2.97E+00</td>
<td>2.57E+00</td>
<td>1.54E+00</td>
</tr>
<tr>
<td>1448700_at</td>
<td>14373</td>
<td>G0s2</td>
<td>G0/G1 switch gene 2</td>
<td>-1.55E+00</td>
<td>-1.53E+00</td>
<td>-2.38E+00</td>
<td>-1.12E+00</td>
</tr>
<tr>
<td>1425281_a_at</td>
<td>14605</td>
<td>Tsc22d3</td>
<td>TSC22 domain family, member 3</td>
<td>1.76E+00</td>
<td>2.74E+00</td>
<td>2.37E+00</td>
<td>1.94E+00</td>
</tr>
<tr>
<td>1420772_a_at</td>
<td>14605</td>
<td>Tsc22d3</td>
<td>TSC22 domain family, member 3</td>
<td>1.60E+00</td>
<td>2.74E+00</td>
<td>2.37E+00</td>
<td>1.94E+00</td>
</tr>
<tr>
<td>1450297_at</td>
<td>16193</td>
<td>Il6</td>
<td>Interleukin 6</td>
<td>-1.32E+00</td>
<td>-2.06E+00</td>
<td>-2.86E+00</td>
<td>-2.58E+00</td>
</tr>
<tr>
<td>1422557_s_at</td>
<td>17748</td>
<td>Mt1</td>
<td>Metallothionein</td>
<td>1.13E+00</td>
<td>2.16E+00</td>
<td>1.96E+00</td>
<td>2.26E+00</td>
</tr>
<tr>
<td>1428942_at</td>
<td>17750</td>
<td>Mt2</td>
<td>Metallothionein</td>
<td>2.27E+00</td>
<td>4.23E+00</td>
<td>3.49E+00</td>
<td>3.13E+00</td>
</tr>
<tr>
<td>1421471_at</td>
<td>18166</td>
<td>Npy1r1</td>
<td>Neuropeptide Y receptor Y1</td>
<td>1.32E+00</td>
<td>1.02E+00</td>
<td>1.97E+00</td>
<td>1.77E+00</td>
</tr>
<tr>
<td>1448830_s_at</td>
<td>19252</td>
<td>Dusp1</td>
<td>Dual specificity phosphatase 1</td>
<td>1.68E+00</td>
<td>2.71E+00</td>
<td>1.40E+00</td>
<td>1.20E+00</td>
</tr>
<tr>
<td>1420380_at</td>
<td>20296</td>
<td>Ccl2</td>
<td>Chemokine (C-C motif) ligand 2</td>
<td>-1.42E+00</td>
<td>-3.12E+00</td>
<td>-1.94E+00</td>
<td>-2.00E+00</td>
</tr>
<tr>
<td>1421228_at</td>
<td>20306</td>
<td>Ccl7</td>
<td>Chemokine (C-C motif) ligand 7</td>
<td>-1.50E+00</td>
<td>-2.00E+00</td>
<td>-1.14E+00</td>
<td>-1.48E+00</td>
</tr>
<tr>
<td>1416029_at</td>
<td>21847</td>
<td>Klf10</td>
<td>Kruppel-like factor 10</td>
<td>-1.42E+00</td>
<td>-1.71E+00</td>
<td>-1.48E+00</td>
<td>-1.00E+00</td>
</tr>
<tr>
<td>1417273_at</td>
<td>27273</td>
<td>Pdk4</td>
<td>Pyruvate dehydrogenase kinase, isoform 4</td>
<td>1.55E+00</td>
<td>2.82E+00</td>
<td>2.75E+00</td>
<td>3.56E+00</td>
</tr>
<tr>
<td>1417309_at</td>
<td>21847</td>
<td>Tob2</td>
<td>Transducer of ERBB2</td>
<td>2.48E+00</td>
<td>2.98E+00</td>
<td>2.09E+00</td>
<td>2.21E+00</td>
</tr>
<tr>
<td>1417223_at</td>
<td>22733</td>
<td>Tob2</td>
<td>Transducer of ERBB2</td>
<td>2.40E+00</td>
<td>2.71E+00</td>
<td>2.14E+00</td>
<td>2.11E+00</td>
</tr>
<tr>
<td>1417233_at</td>
<td>22733</td>
<td>Tob2</td>
<td>Transducer of ERBB2</td>
<td>2.32E+00</td>
<td>2.16E+00</td>
<td>1.73E+00</td>
<td>1.61E+00</td>
</tr>
<tr>
<td>1448667_x_at</td>
<td>57259</td>
<td>Tob2</td>
<td>Transducer of ERBB2</td>
<td>2.24E+00</td>
<td>3.35E+00</td>
<td>2.04E+00</td>
<td>1.99E+00</td>
</tr>
<tr>
<td>1448666_s_at</td>
<td>57259</td>
<td>Tob2</td>
<td>Transducer of ERBB2</td>
<td>2.24E+00</td>
<td>3.35E+00</td>
<td>2.04E+00</td>
<td>1.99E+00</td>
</tr>
<tr>
<td>1448830_x_at</td>
<td>57259</td>
<td>Tob2</td>
<td>Transducer of ERBB2</td>
<td>2.24E+00</td>
<td>3.35E+00</td>
<td>2.04E+00</td>
<td>1.99E+00</td>
</tr>
<tr>
<td>1435595_at</td>
<td>69068</td>
<td>Rik10011</td>
<td>RIKEN cDNA 1810011O10 gene</td>
<td>2.38E+00</td>
<td>3.42E+00</td>
<td>1.99E+00</td>
<td>2.18E+00</td>
</tr>
<tr>
<td>1451415_at</td>
<td>69068</td>
<td>Rik10011</td>
<td>RIKEN cDNA 1810011O10 gene</td>
<td>2.24E+00</td>
<td>3.35E+00</td>
<td>2.04E+00</td>
<td>1.99E+00</td>
</tr>
<tr>
<td>Affymetrix ID</td>
<td>Ensembl ID</td>
<td>Gene Name</td>
<td>PRED_T1</td>
<td>DMSO_T1</td>
<td>PRED_T6</td>
<td>DMSO_T6</td>
<td>PRED_T24</td>
</tr>
<tr>
<td>---------------</td>
<td>------------</td>
<td>-----------</td>
<td>---------</td>
<td>--------</td>
<td>---------</td>
<td>--------</td>
<td>----------</td>
</tr>
<tr>
<td>143407237.0</td>
<td>2143320</td>
<td>A00001DK</td>
<td>1.62E+00</td>
<td>2.29E+00</td>
<td>1.88E+00</td>
<td>3.61E+00</td>
<td>1.91E+00</td>
</tr>
<tr>
<td>143407237.0</td>
<td>2143320</td>
<td>A00001DK</td>
<td>1.62E+00</td>
<td>2.29E+00</td>
<td>1.88E+00</td>
<td>3.61E+00</td>
<td>1.91E+00</td>
</tr>
<tr>
<td>143407237.0</td>
<td>2143320</td>
<td>A00001DK</td>
<td>1.62E+00</td>
<td>2.29E+00</td>
<td>1.88E+00</td>
<td>3.61E+00</td>
<td>1.91E+00</td>
</tr>
<tr>
<td>143407237.0</td>
<td>2143320</td>
<td>A00001DK</td>
<td>1.62E+00</td>
<td>2.29E+00</td>
<td>1.88E+00</td>
<td>3.61E+00</td>
<td>1.91E+00</td>
</tr>
<tr>
<td>143407237.0</td>
<td>2143320</td>
<td>A00001DK</td>
<td>1.62E+00</td>
<td>2.29E+00</td>
<td>1.88E+00</td>
<td>3.61E+00</td>
<td>1.91E+00</td>
</tr>
<tr>
<td>143407237.0</td>
<td>2143320</td>
<td>A00001DK</td>
<td>1.62E+00</td>
<td>2.29E+00</td>
<td>1.88E+00</td>
<td>3.61E+00</td>
<td>1.91E+00</td>
</tr>
<tr>
<td>143407237.0</td>
<td>2143320</td>
<td>A00001DK</td>
<td>1.62E+00</td>
<td>2.29E+00</td>
<td>1.88E+00</td>
<td>3.61E+00</td>
<td>1.91E+00</td>
</tr>
<tr>
<td>143407237.0</td>
<td>2143320</td>
<td>A00001DK</td>
<td>1.62E+00</td>
<td>2.29E+00</td>
<td>1.88E+00</td>
<td>3.61E+00</td>
<td>1.91E+00</td>
</tr>
<tr>
<td>143407237.0</td>
<td>2143320</td>
<td>A00001DK</td>
<td>1.62E+00</td>
<td>2.29E+00</td>
<td>1.88E+00</td>
<td>3.61E+00</td>
<td>1.91E+00</td>
</tr>
<tr>
<td>143407237.0</td>
<td>2143320</td>
<td>A00001DK</td>
<td>1.62E+00</td>
<td>2.29E+00</td>
<td>1.88E+00</td>
<td>3.61E+00</td>
<td>1.91E+00</td>
</tr>
</tbody>
</table>

**Chapter 5**

Prednisolone induces the Wnt signaling pathway in 3T3-L1 adipocytes.
factor insulin-like growth factor I (Igf1), insulin receptor substrate 1 (Irs1) and platelet derived growth factor receptor, beta polypeptide (Pdgfrb) and genes from the TGF-β pathway such as MAD homolog 3 (Drosophila) (Smad3), gremlin 1 (Grem1), transforming growth factor, beta 2 (Tgfb2) and transforming growth factor, beta induced (Tgfbi) were down-regulated. Phosphatase and tensin homolog (Pten) an important regulator of the Akt pathway was up-regulated.

The majority of the genes coding for cytokines, such as Interleukin-6 (Il-6), chemokine (C-X-C motif) ligand 5 (Cxcl5), chemokine (C-X-C motif) ligand 10 (Cxcl10) and chemokine (C-C motif) ligand 7 (Ccl7), were down regulated by prednisolone. Other cytokines such as interleukin 1 receptor, type I (IL1r1) and interleukin 1 receptor accessory protein (IL1rap) were up regulated.

Many Wnt signaling pathway members were significant down regulated by prednisolone in our study. We observed down regulation of frizzled receptors Fzd1, Fzd2, Fzd4, Fzd5 together with ligand wingless-related MMTV integration site 5A (Wnt5A) and antagonist secreted frizzled-related protein 2 (Sfrp2). Also several down-stream molecules such as Wisp1, Wisp2, Id2 were down regulated (Figure 4).
Chapter 5

Regulation of adipokines measured after 48 hours of incubation with prednisolone

The microarray data showed several transcripts that encode hormones secreted by 3T3-L1 adipocytes. The microarray data showed several transcripts that encode hormones secreted by 3T3-L1 adipocytes. Prednisolone-induced up-regulated genes are shown in red, prednisolone-induced down-regulated genes are shown in green. For clarity genes have been assigned to major pathways. However, one should keep in mind that genes can belong to multiple pathways depending on the definition of a pathway.

Figure 4: Prednisolone induced genes after 1 hour (T1), 6 hours (T6), 24 hours (T24) and 48 hours (T48). Prednisolone-induced up-regulated genes are shown in red, prednisolone-induced down-regulated genes are shown in green. For clarity genes have been assigned to major pathways. However, one should keep in mind that genes can belong to multiple pathways depending on the definition of a pathway.
adipocytes, like cytokines, and regulators of the Wnt, Akt and TGF-β pathway. We therefore decided to study to what extent the altered levels of gene expression would lead to changes in protein expression, by analyzing the effects of prednisolone on adipokine secretion by 3T3-L1 adipocytes treated with DMSO or prednisolone for 48 hours. For this purpose, the supernatants of the 3T3-L1 adipocytes that were used in the gene expression study were analyzed. Out of the 308 proteins on the array, 25 proteins were significant regulated by prednisolone (Table 2, p-value < 0.05. For all proteins, see supplementary table 1). The majority of these proteins are cytokines involved in inflammatory and immune response related processes (indicated with a purple background in Table 2). Cytokines such as CXCL2, CCL3, CCL9, IL31 were down regulated by prednisolone while CCR6, IL12B, TNFSF8, TNFRSF7 and TNFRSF17 showed an up regulation. Also several components of the Wnt signaling pathway were significantly regulated at protein level (Table 2, indicated with a blue background). DKK4 an inhibitor of the Wnt signaling pathways was up-regulated. Furthermore downstream regulators of the Wnt signaling pathway, LRP6 and WISP1 are down regulated. Additionally, initiation of insulin signaling is represented by the down regulation of IGFBP5, IGFBP3 and Resistin. These results indicate that the regulation of the Akt/insulin signaling, cytokine signaling and Wnt signaling pathways by prednisolone is not only found on the transcriptome level, but is also translated into altered proteins levels in these pathways.

Regulation of the Wnt signaling pathway by prednisolone

Individual Wnt signaling components were found to be significantly regulated by prednisolone at both gene and protein levels. Therefore, we next investigated whether the entire Wnt signaling pathway was affected by prednisolone. The KEGG database was used to select genes upstream in the canonical Wnt signaling pathway that are more exclusively linked to the Wnt signaling in comparison with down-stream genes like Smad3, Lef, Tak1 and Pparδ, which are also linked to TGF-β signaling, MAPK signaling and Cell cycle (Supplementary Table 3). We used the expression profiles of these 51 Wnt signaling genes to test whether there is a difference in expression for the entire Wnt signaling pathway between prednisolone and DMSO treated samples. In this method, genes that do not meet the FC>2-fold and Pval <0.05 cut offs but may have nevertheless a changed expression level were taken into consideration. A significant difference in expression for the entire Wnt signaling pathway was observed after 6 hours (p-value = 0.031). In Figure 5, the genes that had the most influence on this difference are shown. In this figure, the gene expression ratios (open dots) and protein ratios (black dots) as a result of prednisolone treatment are also shown. Genes that had an influence on the difference between DMSO and prednisolone treated samples and that showed a higher expression in DMSO treated samples (red bars) were down-regulated because of prednisolone treatment (open dots). The same effect on the Wnt signaling pathway was also seen after 48 hours (p-value = 0.064) of treatment (Supplementary Figure 3). This
### Table 2: Regulated proteins after 48 hours of incubation with prednisolone.

<table>
<thead>
<tr>
<th>Name</th>
<th>Protein Symbol</th>
<th>Mean log 2 ratios</th>
<th>P-value</th>
<th>Function description derived from Uniprot</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIF</td>
<td>LIF</td>
<td>0.64</td>
<td>0.016</td>
<td>LIF has the capacity to induce neuronal cell differentiation, and the induction of neuronal cell differentiation in neural and neural embryonic progenitors is affected by LIF. LIF also has the capacity to induce ERBB2-dependent signaling in breast cancer cells.</td>
</tr>
<tr>
<td>IL-12 p70</td>
<td>IL12B</td>
<td>0.64</td>
<td>0.024</td>
<td>Cytokine that can act as a growth factor for activated T cells and can also act as a growth factor for activated B cells. It promotes the differentiation of monocytes/macrophages and enhances the cytotoxic activity of NK cells.</td>
</tr>
<tr>
<td>Dkk-4</td>
<td>DKK4</td>
<td>0.57</td>
<td>0.024</td>
<td>Antagonizes canonical Wnt signaling by inhibiting LRP5/6 receptor activation.</td>
</tr>
<tr>
<td>CD27/L / TNFRSF7</td>
<td>TNFRSF7</td>
<td>0.51</td>
<td>0.040</td>
<td>Receptor for CD70/CD27. May play a role in survival and proliferation of B cells.</td>
</tr>
<tr>
<td>CD27/L / TNFRSF7</td>
<td>TNFRSF7</td>
<td>0.37</td>
<td>0.010</td>
<td>Cytokine that binds to TNFSF13B/BLyS/BAFF and TNFSF13/APRIL. Promotes B-cell survival and plays a role in the regulation of B-cell survival and apoptosis.</td>
</tr>
<tr>
<td>CD30 L / TNFSF8</td>
<td>TNFSF8</td>
<td>0.37</td>
<td>0.010</td>
<td>Cytokine that binds to TNFRSF8/CD30. Induces proliferation of B cells and enhances B-cell survival.</td>
</tr>
<tr>
<td>Gremlin</td>
<td>GREM1</td>
<td>0.24</td>
<td>0.019</td>
<td>Cytokine that may play an important role during bone formation.</td>
</tr>
<tr>
<td>WISP-1 / CCN4</td>
<td>WISP1</td>
<td>0.17</td>
<td>0.033</td>
<td>Downstream regulator in the Wnt/β-catenin signaling pathway.</td>
</tr>
<tr>
<td>MIP-1 gamma</td>
<td>CCL9</td>
<td>0.32</td>
<td>0.007</td>
<td>Monokine with inflammatory, pyrogenic and chemokinetic properties. Circulates at high concentrations in the blood of healthy animals.</td>
</tr>
</tbody>
</table>

**Legend:**
- Inflammatory related proteins are indicated with a purple background.
- Wnt signaling related proteins are indicated with a blue background.
- Insulin signaling related proteins are indicated with a green background.
- Up-regulated proteins are indicated with a red color.
- Down-regulated proteins are indicated with a green color.
analysis indicates that in 3T3-L1 adipocytes, the entire Wnt signaling pathway is down-regulated by prednisolone.
Discussion and conclusions

The prolonged use of glucocorticoids is hampered by metabolic side-effects such as insulin resistance, which eventually leads to diabetes and obesity. Adipose tissue is thought to play an important role in GC-induced development of metabolic side effects. The mechanisms behind these side-effects are not completely understood. Key findings of our study indicate that prednisolone reduces insulin stimulated glucose uptake in 3T3-L1 cells, which is reflected by changes in expression of genes and proteins from the Akt/insulin, cytokine and Wnt signaling pathways. To our knowledge, this is the first work that uses a combination of gene expression and antibody arrays to study effects of prednisolone in adipocyte cells.

The set of 25 genes that showed regulation by prednisolone at all time points includes known target genes of the glucocorticoid receptor, such as Ddit4, Dusp1 and Cebpd [30, 31], and genes such as Fam107a and PDK4, that were also found to be regulated by prednisolone in in vivo studies of mouse and human [30, 32]. This indicates that prednisolone acts via its target genes and initiates a GC specific response in these 3T3-L1
Multiple cytokine molecules were regulated by prednisolone. Increasing evidence suggests that a chronic low-grade state of inflammation in adipose tissue contributes to the development of systemic insulin resistance and diabetes [33-35]. In the absence of IL1, mice are protected against high fat diet induced insulin resistance which is accompanied with a reduction in local adipose tissue inflammation [36, 37]. Our results showed up-regulation of IL1 signaling by means of IL1r1 and IL1rap. This suggests a role of IL-1 signaling in GC induced development of insulin resistance in 3T3-L1 cells. Also other studies reported about the up-regulation of cytokines such as Il6, Cxcl5, Ccl2 and Cxcl10 in insulin resistance and obesity (Figure 6) [35, 38-42]. In contrast to these studies, we observed a down regulation of expression of these genes by prednisolone. Based on these results it is difficult to deduce whether prednisolone-induced regulation of cytokines in adipocytes contribute to the development of insulin resistance or are the result of the immune suppressing properties of prednisolone or a combination of both. Furthermore our results indicated that the reduction in insulin stimulated glucose uptake by prednisolone acts partly via the down regulation of insulin signaling. This regulation might be mediated by Pten, a known suppressor of insulin signaling that acts via the PI3K/ Akt signaling pathway [43]. Additional studies confirm this observation and suggest an important role of Pten in the development of insulin resistance and diabetes [44-46]. Again an apparent contrasting effect of prednisolone is the down regulation of proteins like Igfb3 and resistin that reduce glucose uptake in adipose tissue [47, 48]. Inhibition of TGF-β/ Smad3 signaling results in diminished adiposity, improved glucose tolerance and insulin sensitivity and in the protection from insulin resistance, diabetes and obesity [49-51]. The observed down regulation of Smad3 suggests that disturbance of glucose uptake and adiposity by prednisolone in 3T3-L1 cells acts via additional mechanisms. Activation of the Wnt signaling pathway has been linked to inhibition of adipogenesis [52-55]. Recent studies toward the characterization of the adipocyte secretome have shown that multiple regulators of the activity the Wnt signaling are secreted by adipocytes, such as WISP2 and SFRP5 [54, 56]. Interestingly, expression and circulating levels of these factors are altered in insulin resistance and obesity [54, 57-59], underlying the suggestion that a dysregulated secretion of Wnt regulators could lead to metabolic disturbances and eventually into the development of obesity and diabetes [54, 60-63]. Yet, in the context of our study, it remains to be investigated to what extent the factors affected by prednisolone interfere with insulin action in adipocytes. Prednisolone induced Pten, Smad3 and Lef1 which are able to mediate cross-talk between the signaling pathways [64]. Via the alteration of the Pten function, TGF-β is able to influence the Akt activity, and via Smad3 the Akt pathway is able to restrict the TGF-β pathway. Also Smad activity is involved in the cross-talk between Wnt signaling and TGF-β signaling, as part of the Smad/β-catenin/Lef protein complex in the nucleus. These
Prednisolone-induces the Wnt signaling pathway in 3T3-L1 adipocytes

Chapter 5

genes could play a central role in the development of prednisolone induced metabolic effects because of their ability to connect multiple pathways (Figure 6). The sometimes opposing effects of prednisolone on individual pathway members of Smad3 and Akt in comparison with earlier work, requires additional in vivo studies in which the systemic effect of prednisolone on these pathways could be determined.

To our knowledge, for the first time it is shown that prednisolone affects the expression and secretion of Wnt regulators in adipocytes. The fact that Wnt signaling also participates in the development of metabolic disturbances in other tissues [65, 66] may direct future work towards a dedicated study of the Wnt signaling pathway in GC-induced metabolic effects. In an in vivo setting the Wnt signaling mediated GC effects in multiple glucose responsive tissues such as muscle tissue, liver tissue and adipose tissue could be determined. Hence new GC compounds with an improved efficacy/ side effects ratio should have a reduced effect on components in the Wnt signaling pathway.
Declaration of interest
This work was performed within the framework of Dutch Top Institute Pharma (TIP) project “Glucocorticoid-induced insulin resistance” (project nr. T1-106-1). DMO is supported by the Federal Ministry of Health, the Ministry of Innovation, Science, Research and Technology of the German State of North-Rhine Westphalia. WF is supported by the Biorange project (BR4.2) “A Systems Bioinformatics Approach For Evaluating And Translating Drug-Target Effects In Disease Related Pathways” of NBIC. From 1 October 2012, WF is supported by the Netherlands eScience project (027.011.306) “Creation of food specific ontologies for food focused text mining and knowledge management studies”. The authors Wynand Alkema, Wim Dokter and Jacob de Vlieg were employed by Merck at which part of this work was initiated. The authors Erik J.M. Toonen, Wilco W.M. Fleuren, Margot M. L. Linssen, Gerard C.C. van der Zon, Bruno Guigas and D. Margriet Ouwens have no conflict of interest. Wim Dokter and Jacob de Vlieg own stock of Merck.

Supplementary information can be retrieved on request by the author of this thesis or can be found at the website of the journal.
References

26. Fleuren, W.W., et al., CoPub update: CoPub 5.0 a text mining system to answer biological questions.


35. Bastard, J.P., et al., Recent advances in the relationship between obesity, inflammation, and insulin resistance. Eur Cytokine Netw, 2006. 17(1): p. 4-12.


Prednisolone induces the Wnt signaling pathway in 3T3-L1 adipocytes

80. Garcia-Diaz, D.F., et al., Glucose and insulin modify thrombospondin 1 expression and secretion in...


Prednisolone-induces the Wnt signaling pathway in 3T3-L1 adipocytes
CHAPTER 6
Prednisolone-induced differential gene expression in mouse liver carrying wild type or a dimerization-defective glucocorticoid receptor

Raoul Frijters¹, Wilco W.M. Fleuren¹, Erik J.M. Toonen³, Jan P Tuckermann⁵, Holger M Reichardt⁶, Hans van der Maaden⁴, Andrea van Elsas³, Marie-Jose van Lierop³, Wim Dokter³, Jacob de Vlieg¹,² and Wynand Alkema²

¹Computational Drug Discovery (CDD), Nijmegen Centre for Molecular Life Sciences (NCMLS), Radboud University Nijmegen Medical Centre, Geert Grooteplein Zuid 26-28, 6525 GA Nijmegen, The Netherlands. ²Department of Molecular Design & Informatics, Schering-Plough, Molenstraat 110, 5342 CC Oss, The Netherlands. ³Department of Immunotherapeutics, Schering-Plough, Molenstraat 110, 5342 CC Oss, The Netherlands. ⁴Molecular Pharmacology Department, Schering-Plough, Molenstraat 110, 5342 CC Oss, The Netherlands. ⁵Tuckermann Lab, Leibniz Institute for Age Research - Fritz Lipmann Institute, Beutenbergstraße 11, 07745 Jena, Germany. ⁶Department of Cellular and Molecular Immunology, University of Göttingen Medical School, Humboldtallee 34, 37073 Göttingen, Germany.

BMC Genomics 2010, Jun 5;11:359
Abstract

Background
Glucocorticoids (GCs) control expression of a large number of genes via binding to the GC receptor (GR). Transcription may be regulated either by binding of the GR dimer to DNA regulatory elements or by protein-protein interactions of GR monomers with other transcription factors. Although the type of regulation for a number of individual target genes is known, the relative contribution of both mechanisms to the regulation of the entire transcriptional program remains elusive. To study the importance of GR dimerization in the regulation of gene expression, we performed gene expression profiling of livers of prednisolone-treated wild type (WT) and mice that have lost the ability to form GR dimers (GR\textsuperscript{dim}).

Results
The GR target genes identified in WT mice were predominantly related to glucose metabolism, the cell cycle, apoptosis and inflammation. In GR\textsuperscript{dim} mice, the level of prednisolone-induced gene expression was significantly reduced compared to WT, but not completely absent. Interestingly, for a set of genes, involved in cell cycle and apoptosis processes and strongly related to Foxo3a and p53, induction by prednisolone was completely abolished in GR\textsuperscript{dim} mice. In contrast, glucose metabolism-related genes were still modestly upregulated in GR\textsuperscript{dim} mice upon prednisolone treatment. Finally, we identified several novel GC-inducible genes from which \textit{Fam107a}, a putative histone acetyltransferase complex interacting protein, was most strongly dependent on GR dimerization.

Conclusions
This study on prednisolone-induced effects in livers of WT and GR\textsuperscript{dim} mice identified a number of interesting candidate genes and pathways regulated by GR dimers and sheds new light onto the complex transcriptional regulation of liver function by GCs.
**Background**

Naturally occurring glucocorticoids (GCs), such as cortisol, play an important role in the regulation of cardiovascular, metabolic and immunological processes. GCs are potent suppressors of inflammatory indices and are widely used to treat chronic inflammatory diseases such as rheumatoid arthritis and asthma [1-3]. However, chronic use of GCs induces side effects, such as diabetes mellitus, fat redistribution, osteoporosis, muscle atrophy, glaucoma and skin thinning [4].

One of the aspects influencing the balance between the desired anti-inflammatory effects and the side effects of GCs is the activation and repression of gene expression. After binding of GCs to the cytosolic GC receptor (GR), the receptor-ligand complex translocates to the nucleus where it alters gene expression. Ligand-bound GR can influence the expression of target genes, either by binding as a dimer to palindromic GC response elements (GRE) or tethering to other DNA-bound transcription factors [5, 6].

It is generally assumed that the anti-inflammatory actions of GCs are mainly driven by transrepression, in which ligand-bound GR binds to the pro-inflammatory transcription factors NF-κB (Nuclear factor kappa-light-chain-enhancer of activated B cells), AP-1 (Activator protein 1), IRF-3 (Interferon regulatory factor 3) or other factors, forming an inactive transcription machinery complex, which prevents expression of pro-inflammatory genes [7-16]. However, recent studies demonstrated that under some inflammatory conditions DNA dimer-dependent gene expression could also contribute to the anti-inflammatory activities of the GR [17]. The occurrence of side effects has been linked mainly to transactivation of gene transcription. The well-documented example of GC-induced upregulation of \( \text{Pck1} \) (Phosphoenolpyruvate carboxykinase 1; also known as \( \text{Pepck} \)) and \( \text{G6pc} \) (Glucose-6-phosphatase), two genes encoding key enzymes that control gluconeogenesis, is one example linking transactivation to metabolic side effects [18-20].

One of the mechanisms by which GR activates gene transcription is by binding as a homodimer to a GRE in the promoter region of a target gene. GR dimerization involves the D loop located in the DNA-binding domain of the GR, in which several amino acids interact to facilitate receptor dimer formation. It was shown that an A458T point mutation, introduced in the D loop of the GR (GR\(^{\text{dim}}\)), impairs homodimerization and ablates DNA binding [21]. Mice carrying this GRdim mutation are almost as effective as wild type (WT) mice in repressing AP-1 and NF-κB-modulated gene transcription, whereas GR-mediated \( \text{Tat} \) (Tyrosine aminotransferase) mRNA induction is largely abolished [21, 22].

The mechanism of dimer-dependent transactivation by the GR has been studied in detail for a limited number of GR target genes, including \( \text{Fkbp5} \) (FK506 binding protein 5), \( \text{Tat} \), \( \text{Pck1} \) and \( \text{Dusp1} \) (Dual specificity phosphatase 1) [23-26], and tethering-facilitated transrepression of \( \text{Mmp1/Mmp13} \) (Matrix metallopeptidase 1/13), \( \text{IL-8} \) (Interleukin 8) and others [27-29]. In order to perform a comprehensive study of the genes and cellular processes that are affected by GC-treatment and influenced by GR dimerization, we have performed genome wide gene expression profiling in liver of prednisolone-treated WT...
Prednisolone-induced differential gene expression in mouse liver

We found that prednisolone predominantly influenced expression of genes involved in glucose metabolism, inflammation, the cell cycle and apoptosis. In general, activation of transcription, including transactivation of known GR marker genes such as \textit{Fkbp5} and \textit{Tat}, was significantly reduced, but not totally absent in the GR\textsuperscript{dim} mutant. However, a total absence of transcriptional transactivation upon GC-treatment was observed in GR\textsuperscript{dim} mice for a subset of genes, all related to regulation of the cell cycle. In contrast, a small subset of genes was found exclusively regulated by GCs in GR\textsuperscript{dim} mice. Furthermore, we identified \textit{Fam107a} (\textit{Family with sequence similarity 107, member a}; also known as \textit{Tu3a} and \textit{Drr1}) as a novel GC-inducible gene that completely relies on GR dimerization for transactivation.

Results

Clustering of individual samples

In order to chart the gene expression profile induced by prednisolone in WT and GR\textsuperscript{dim} mice, 6 mice per group were treated with prednisolone or vehicle. A principal component analysis (PCA) on the normalized intensity data showed that the samples clustered into four distinct groups (Figure 1). The largest separation, represented by PC1, was observed between samples derived from male versus female mice. The genes with the highest gender specificity belong to the family of cytochrome P450 proteins, such as \textit{Cyp3a41}, \textit{Cyp3a44} and \textit{Cyp4a12}, which are known for their gender-specific expression [30, 31]. In addition to the gender-based separation, a clear separation was observed between prednisolone-treated and vehicle-treated WT mice, whereas prednisolone-treated GR\textsuperscript{dim} mice clustered closely together with vehicle-treated GR\textsuperscript{dim} and WT mice (PC2). These

Figure 1 Principal component analysis on gene expression intensity data.

A principal component analysis (PCA) was performed on normalized gene expression intensity data from 12 wild type (WT) mice (vehicle and prednisolone treatment) and 12 GR\textsuperscript{dim} mice (vehicle and prednisolone treatment). Bullets: WT mice; Triangles: GR\textsuperscript{dim} mice. White symbols: vehicle-treated mice; Black symbols: prednisolone-treated mice. The largest separation was seen between samples derived from male versus female mice (PC1).
results indicate that there is a significant prednisolone effect in WT mice, which is strongly reduced in prednisolone-treated GRdim mice. Genes with high loading scores on PC2 include Tat and Fkbp5, which are known GR-regulated genes in liver [32, 33] (Figure 2). These results show that the transcriptional program that is induced by prednisolone in WT mice is strongly reduced in mice carrying the GRdim mutation. This supports the hypothesis that the loss of GR dimerization leads to a reduction in transcriptional transactivation by the GR and underscores the importance of GR dimerization in induction of gene expression. The effect of prednisolone on up and downregulated genes and pathways will be discussed in more detail below.

Figure 2 Intensity profiles of Tat and Fkbp5 in vehicle and prednisolone-treated wild type versus GRdim mice. The expression intensities of Tat (Tyrosine aminotransferase) and Fkbp5 (FK506 binding protein 5) in vehicle-treated (white bars) and prednisolone-treated (grey bars) wild type (WT) and GRdim mice are plotted. Symbols: Bullets: male mice; Triangles: female mice. Both Tat and Fkbp5 are known GR marker genes and show strong differential regulation in prednisolone-treated WT mice.

Prednisolone-induced changes in gene expression
To quantify the differences that were seen in the multivariate analysis, differentially expressed probe sets were identified for several treatment comparisons (e.g. prednisolone versus vehicle-treated WT mice and prednisolone vs. vehicle-treated GRdim mice). Probe sets were marked as differentially regulated when the p-value, corrected for multiple testing, was below the 0.01 cutoff value.

Prednisolone treatment resulted in a significant differential regulation of 518 probe sets in WT mice (347 upregulated and 171 downregulated), whereas in prednisolone-treated GRdim mice only 34 probe sets were differentially regulated (29 upregulated and 5 downregulated). Notably, no differentially regulated probe sets were found in the comparison between vehicle-treated GRdim and vehicle-treated WT mice, suggesting that
the GR\textsuperscript{dim} mutation itself does not cause differential gene regulation and that the effects between WT and GR\textsuperscript{dim} mice only become apparent after treatment with prednisolone. The observation that only 34 probe sets are differentially regulated in GR\textsuperscript{dim} mice upon prednisolone treatment, suggests that there is almost no regulation of gene expression by prednisolone in GR\textsuperscript{dim} mice. However, a large number of the probe sets that are differentially regulated in WT mice showed the same direction of regulation by prednisolone in GR\textsuperscript{dim} mice, but did not meet the cutoff value of 0.01 for the p-value. This indicates that induction of gene expression by prednisolone in GR\textsuperscript{dim} mice is strongly reduced but not completely absent (see below). 
Taken together these data are in line with the hypothesis that transactivation of gene expression through the GR is significantly reduced in mice carrying the GR\textsuperscript{dim} mutation.

**Biological processes targeted by prednisolone in WT mice**

In order to identify cellular processes on which prednisolone has a prominent effect, a gene set enrichment analysis was conducted using the 518 probe sets that were differentially regulated in WT mice by prednisolone. For this analysis we used CoPub, a text mining tool that calculates which biological processes are significantly associated to a set of genes [34]. It appeared that the differentially regulated genes are predominantly involved in three major processes: (glucose) metabolism, cell cycle/apoptosis and immune/inflammatory response (Table 1).

In order to gain insight into the relationships between genes and cellular processes a network representation was created, in which differentially regulated genes are plotted together with enriched keywords. In this representation, cellular processes that are affected by prednisolone can be appreciated as separate areas in the network, such as cell cycle and apoptosis, acute-phase response, and metabolism-related processes such as amino acid metabolism, gluconeogenesis and lipid metabolism. The most influential genes appear as highly connected hubs (Figure 3). Important factors that seem to connect the gluconeogenesis and the cell cycle and apoptosis subnetworks are the transcription factors Foxo1 and Foxo3a (Forkhead box O1 and Forkhead box O3a; see below).

Strong effects of prednisolone treatment on gluconeogenesis, the cell cycle and apoptosis are in line with the anticipated effect of the GR in the liver. The fact that many inflammation and acute phase response-related genes, such as IL6r (Interleukin 6 receptor) and Cxcl12 (Chemokine (C-X-C motif) ligand 12) were regulated (Figure 3) is interesting, given that we analyzed livers from healthy mice, not challenged with inflammatory stimuli. The direction and magnitude of prednisolone-induced differential regulation of genes involved in gluconeogenesis and the cell cycle was studied in more detail to get more insight into these prednisolone-affected processes.
Table 1 CoPub keyword enrichment analysis on prednisolone-regulated genes. CoPub was used to calculate keyword enrichment in the set of prednisolone-regulated probe sets in wild type mice. The analysis revealed that the differential regulated genes (non-redundant set) are predominantly involved in three main processes: metabolism, cell cycle and apoptosis, and immune/inflammatory response. The right column shows the number of differential regulated genes associated with a specific enriched keyword.

<table>
<thead>
<tr>
<th>Biological process</th>
<th># of associated genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Metabolism</strong></td>
<td></td>
</tr>
<tr>
<td>Glucose metabolism / transport</td>
<td>22</td>
</tr>
<tr>
<td>Gluconeogenesis</td>
<td>21</td>
</tr>
<tr>
<td>Lipid metabolism / glycosylation</td>
<td>21</td>
</tr>
<tr>
<td>Glycolysis</td>
<td>14</td>
</tr>
<tr>
<td>Carbohydrate metabolism / transport</td>
<td>14</td>
</tr>
<tr>
<td>Amino acid metabolism / transport</td>
<td>13</td>
</tr>
<tr>
<td><strong>Cell cycle and apoptosis</strong></td>
<td></td>
</tr>
<tr>
<td>Apoptosis</td>
<td>77</td>
</tr>
<tr>
<td>Cell proliferation</td>
<td>72</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>65</td>
</tr>
<tr>
<td>Cell growth and—or maintenance, cell growth</td>
<td>63</td>
</tr>
<tr>
<td>Homeostasis</td>
<td>58</td>
</tr>
<tr>
<td>Cell differentiation</td>
<td>49</td>
</tr>
<tr>
<td>Cell cycle arrest</td>
<td>37</td>
</tr>
<tr>
<td><strong>Immune / inflammatory response</strong></td>
<td></td>
</tr>
<tr>
<td>Inflammatory response</td>
<td>31</td>
</tr>
<tr>
<td>Acute-phase response</td>
<td>20</td>
</tr>
</tbody>
</table>

Gluconeogenesis
Table 2 shows the prednisolone-regulated genes in WT mice identified by CoPub as being associated with gluconeogenesis. Among the upregulated genes are Pck1, encoding the rate limiting enzyme in gluconeogenesis [35], Ppargc1a *Peroxisome proliferative activated receptor, gamma, coactivator 1 alpha*, a transcriptional coactivator that coordinates the expression of genes involved in gluconeogenesis and ketogenesis [36], and Sds *Serine dehydratase* and Aass *Aminoadipate-semialdehyde synthase*, two genes encoding enzymes involved in amino acid catabolism and amino acid utilization for gluconeogenesis [37, 38].

The set of downregulated genes includes Irs1 *Insulin receptor substrate 1*, a downstream mediator of the growth factor/insulin signaling pathway that negatively regulates gluconeogenesis [39] and Pklr *Pyruvate kinase*, which encodes a glycolysis associated
Prednisolone-induced differential gene expression in mouse liver

Chapter 6

enzyme known to catalyze the production of pyruvate from phosphoenolpyruvate [40]. The direction of regulation and the function of these genes, suggests that in the liver of prednisolone-treated WT, glucose metabolism is balanced towards gluconeogenesis, which is in line with the reported gluconeogenic effect of glucocorticoids [41, 42].

Table 2 Prednisolone-regulated genes associated with ‘gluconeogenesis’ in literature. CoPub identified 21 prednisolone-regulated genes in wild type (WT) mice that are associated in literature with gluconeogenesis. The identified genes share at least 3 publications and have an R-scaled score [34] of at least 35 with keyword ‘gluconeogenesis’ in Medline abstracts. For each gene, the fold change (Fc) in prednisolone versus vehicle-treated WT mice (Fc WT) is given. Additionally, for each gene also the Fc in prednisolone versus vehicle-treated GRdim mice (Fc GRdim) is shown. *It should be noted that in GRdim mice these genes did not meet the p-value cutoff of 0.01 for significance after correction for multiple testing, and are mentioned to illustrate a small effect of prednisolone on gene expression in GRdim mice.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Symbol</th>
<th>Fc WT</th>
<th>Fc GRdim*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insulin-like growth factor binding protein 1</td>
<td>Igfbp1</td>
<td>9.7</td>
<td>3.6</td>
</tr>
<tr>
<td>Tyrosine aminotransferase</td>
<td>Tat</td>
<td>4.1</td>
<td>1.6</td>
</tr>
<tr>
<td>6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3</td>
<td>Pfkfb3</td>
<td>3.9</td>
<td>2.3</td>
</tr>
<tr>
<td>Peroxisome proliferative activated receptor, gamma,</td>
<td>Ppargc1a</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td>coactivator 1 alpha</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCAAT/enhancer binding protein (C/EBP), beta</td>
<td>Cebpb</td>
<td>3.1</td>
<td>1.6</td>
</tr>
<tr>
<td>Protein-tyrosine sulfotransferase 2</td>
<td>Tpst2</td>
<td>2.7</td>
<td>1.5</td>
</tr>
<tr>
<td>Solute carrier family 2 (facilitated glucose transporter), member 1</td>
<td>Slc2a1</td>
<td>2.6</td>
<td>1.9</td>
</tr>
<tr>
<td>Forkhead box O1</td>
<td>Foxo1</td>
<td>2.2</td>
<td>1.7</td>
</tr>
<tr>
<td>Serine dehydratase</td>
<td>Sds</td>
<td>2.1</td>
<td>1.5</td>
</tr>
<tr>
<td>Aminoadipate-semialdehyde synthase</td>
<td>Aass</td>
<td>2.1</td>
<td>1.3</td>
</tr>
<tr>
<td>Phosphoenolpyruvate carboxykinase 1, cytosolic</td>
<td>Pck1</td>
<td>1.9</td>
<td>1.3</td>
</tr>
<tr>
<td>Tryptophan 2,3-dioxygenase</td>
<td>Tdo2</td>
<td>1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>Aconitase 2, mitochondrial</td>
<td>Aco2</td>
<td>1.4</td>
<td>1.3</td>
</tr>
<tr>
<td>Transketolase</td>
<td>Tkt</td>
<td>-1.2</td>
<td>-1.1</td>
</tr>
<tr>
<td>Protein kinase, AMP-activated, beta 1 non-catalytic subunit</td>
<td>Prkab1</td>
<td>-1.5</td>
<td>1.0</td>
</tr>
<tr>
<td>Adiponectin receptor 2</td>
<td>Adipor2</td>
<td>-1.6</td>
<td>-1.3</td>
</tr>
<tr>
<td>Purinergic receptor P2Y, G-protein coupled, 5</td>
<td>P2ry5</td>
<td>-1.6</td>
<td>-1.3</td>
</tr>
<tr>
<td>Pyruvate kinase liver and red blood cell</td>
<td>Pklr</td>
<td>-2.1</td>
<td>-1.4</td>
</tr>
<tr>
<td>CCAAT/enhancer binding protein (C/EBP), alpha</td>
<td>Cebpa</td>
<td>-2.3</td>
<td>-1.2</td>
</tr>
<tr>
<td>Sterol regulatory element binding factor 1</td>
<td>Srebf1</td>
<td>-3.0</td>
<td>-1.6</td>
</tr>
<tr>
<td>Insulin receptor substrate 1</td>
<td>Irs1</td>
<td>-4.0</td>
<td>-1.5</td>
</tr>
</tbody>
</table>
The reduced fold changes in GRdim mice compared to WT mice (Table 2) indicate that glucocorticoid-induced gluconeogenesis is reduced in GRdim mice.

Cell cycle

Table 3 shows the prednisolone-regulated genes in WT mice that CoPub identified as being associated with the cell cycle. Upregulated genes include *Gadd45b, Gadd45g* (*Growth arrest and DNA-damage-inducible 45 beta and gamma*), *Cdkn1a* (*Cyclin-dependent kinase inhibitor 1a*; also known as *p21*) and *Plk3* (*Polo-like kinase 3*), which
are stress sensors involved in cell cycle arrest, DNA repair and apoptosis [43-45]. Other upregulated genes are Bcl2l1 (also known as Bcl2-like1 and Bcl-xl), an anti-apoptotic protein that is enhanced by binding to the Gadd45 family [46, 47], and Dusp1, a p53 target gene involved in cell cycle regulation [48].

Amongst the downregulated genes are Ccnd2 (Cyclin D2) and Cks2 (Cdc28 protein kinase regulatory subunit 2), both involved in cell cycle progression [49, 50]. Overall, the direction of the prednisolone-induced differential regulation of genes associated with the cell cycle suggests that prednisolone induces a cytostatic response in liver of WT mice.

**Foxo transcription**

From the literature-based network, the two transcription factors Foxo1 and Foxo3a appear to be intermediates between the gluconeogenesis and cell cycle and apoptosis subnetworks (Figure 3). Foxo transcription factors are key mediators of cell cycle progression, apoptosis, glucose metabolism, reactive oxygen species detoxification and DNA damage repair [51-54]. Their activity is tightly controlled by the insulin and growth factor-inducible Pi3k/Akt pathway. Akt-mediated phosphorylation of Foxo transcription factors promotes their translocation from the nucleus to the cytosol and thereby their inactivation through binding of 14-3-3 proteins [54-56]. Foxo transcription factors regulate gene expression of enzymes that are important regulators of gluconeogenesis, such as Pck1 and G6p [57, 58]. In addition, they suppress genes that are involved in glycolysis, the pentose shunt and lipogenesis [59]. Analysis of Foxo1-overexpressing mice revealed an increased expression of Igfbp1 (Insulin-like growth factor binding protein 1), Pck1, Tat and Tdo (Tryptophan 2,3-dioxygenase) and downregulation of Srebf1 (Sterol regulatory element binding factor 1) and Adipor2 (Adiponectin receptor 2) in liver [59]. Interestingly, these genes are similarly regulated in prednisolone-treated WT mice (Figure 3 and Table 2).

Finally, Foxo transcription factors are regulators of cell cycle progression and were shown to induce Cdkn1a and Gadd45b [60] and to suppress Ccnd2 [54]. Also these genes show the same direction of regulation in prednisolone-treated WT mice (Figure 3 and Table 3). These observations suggest that GCs in WT mice induce Foxo transcription factors, which in turn may synergize with the GR to modulate the expression of their respective target genes.

**Prednisolone-induced differential gene expression in WT versus GRdim mice**

Table 2 and 3 list a number of genes that are significantly regulated by prednisolone in WT mice. In the GRdim mice these genes do not meet the significance cutoff for differential regulation (p<0.01), but for most of them a small effect of prednisolone on the expression level can be observed in GRdim mice. To analyze these differences quantitatively, we plotted the log2 ratios of gene expression after prednisolone treatment in WT versus GRdim mice (Figure 4). This figure shows that on average, genes in WT mice are more
Figure 3 Literature-based network of glucocorticoid-induced effects. A network representation of the enrichment results was generated, in which differentially expressed genes together with enriched keywords are plotted. Genes are shown as circles (Red: upregulated; Green: downregulated), whereas enriched keywords are shown as squares. Connections between genes and keywords represent co-publications in Medline abstracts. To avoid an over-complex network, thresholds were set to simplify the interpretation of the results. Only keywords and genes that share at least 5 publications and have an R-scaled score of at least 45 are plotted in the network [34]. Several higher-order biological processes that are affected by prednisolone can be appreciated: cell cycle and apoptosis, acute-phase response, stress response, amino acid metabolism, gluconeogenesis and lipid metabolism.
strongly regulated than in GR\textsuperscript{dim} mice after prednisolone treatment and confirms that induction of gene expression by prednisolone is not abrogated in GR\textsuperscript{dim} mice but reduced to on average 33% of the level in WT mice, as indicated by the slope of the dotted line in Figure 4.

![Figure 4 Log2 ratio plot of prednisolone-induced differential gene expression in wild type versus GR\textsuperscript{dim} mice. Genes on the solid black line show an equal induction of differential gene expression by prednisolone in wild type (WT) versus GR\textsuperscript{dim} mice. The black-dotted line indicates the average slope of the data and shows that the level of gene regulation in prednisolone-treated GR\textsuperscript{dim} mice is 33% of that observed in WT mice. A subset of genes that show obvious differences in the magnitude of regulation by prednisolone in WT versus GR\textsuperscript{dim} mice is highlighted, as well as genes that show equal magnitude of regulation in WT versus GR\textsuperscript{dim} mice.](image)

**Identification of genes that differentially respond to prednisolone in GR\textsuperscript{dim} mice as compared to WT mice**

In Figure 4, several genes are highlighted that show a different response upon prednisolone treatment in WT mice than in GR\textsuperscript{dim} mice. Unexpectedly, a small subset of these genes shows stronger upregulation by prednisolone in GR\textsuperscript{dim} mice compared to WT mice. These genes include Car8 (Carbonic anhydrase 8), a protein associated with proliferation and invasiveness of colon cancer cells [61], Zdhhc23 (Zinc finger, DHHC-type containing 23), a Nitric oxide synthase-binding and activation protein [62], Flvcr2 (Feline leukemia virus subgroup C cellular receptor family, member 2),
a calcium-chelate transporter [63], and Amy2 (Amylase 2), an amylase that catalyzes the endohydrolysis of 1,4-alpha-D-glucosidic linkages in oligosaccharides.

Genes that appear to strictly rely on GR dimerization include Dusp1, Gadd45b, Cdkn1a, Foxo3a, Plk3 and Fam107a (Figure 4 and 5). As discussed above, Foxo3a, Dusp1, Gadd45b, Cdkn1a and Plk3 all participate in cell cycle regulation, apoptosis and DNA damage repair [43-45, 52]. The most striking difference in the level of regulation by prednisolone between WT and GRdim mice was observed for Fam107a (also known as Tu3a and Drr1) (Figure 4, 5 and 6). The expression profile suggests that transcription of Fam107a is under direct control of the GR and strictly depends on its dimerization. Fam107a encodes a ubiquitously expressed protein that was first described in the context of renal cell carcinoma, in which Fam107a expression is reduced or absent due to promotor hypermethylation [64-66]. Overexpression of Fam107a in Fam107a-negative cell lines leads to growth retardation and apoptosis, indicating that Fam107a might act as a tumor suppressor [65, 67].

We also identified a subset of genes that showed an equal induction of gene expression by prednisolone in WT and GRdim mice (located on the diagonal black line in Figure 4); Dst (Dystonin), a cytoskeleton-interacting protein postulated to cross-link cytoskeletal filaments and thereby maintain cellular integrity [68], Gtf2a2 (General transcription factor II A, 2), a subunit of the transcription initiation factor TFIIA [69], Agpat6 (1-acylglycerol-3-phosphate O-acyltransferase 6), an enzyme involved in triglyceride synthesis [70], Marcks (Myristoylated alanine rich protein kinase C substrate), cytoskeletal protein involved in cell adhesion and cell motility [71], Obfc2a (Oligonucleotide/oligosaccharide-binding fold containing 2A), a single stranded nucleic-acid-binding protein [72], and Ppargc1a.

Prednisolone-induced changes in gene expression in WT and GRdim mice were validated by Q-PCR for Fam107a, Dusp1, Cdkn1a and Car8. Figure 6 shows the fold-changes of the selected genes in prednisolone versus vehicle-treated WT and GRdim mice. For all four genes, the regulation by prednisolone in WT and GRdim mice was qualitatively similar to what was found by microarray analysis.

**Discussion and conclusion**

In this study we performed liver gene expression profiling of WT and GRdim mice after prednisolone administration. Our aim was to chart the biological processes in the liver that are affected by GCs and to study their dependence on DNA-binding and dimerization of the GR. CoPub keyword enrichment analysis with prednisolone-regulated genes in WT mice showed enrichment of keywords associated with glucose, lipid and amino acid metabolism, the cell cycle and apoptosis (Table 1 and Figure 3). This is in agreement with cellular processes known to be affected by GCs [73, 74]. Interestingly, the forkhead transcription factors Foxo1 and Foxo3a are regulators of these cellular processes and induced by prednisolone in WT mice. Together with the observation that a subset of
Chapter 6

Figure 5 Intensity profiles of *Plk3*, *Dusp1*, *Gadd45b*, *Cdkn1a*, *Foxo3a* and *Fam107a* in vehicle and prednisolone-treated wild type and GR<sup>dim</sup> mice. Several genes were identified that showed differences between prednisolone-treatment in wild type (WT) and GR<sup>dim</sup> mice. Amongst those genes are *Plk3*, *Dusp1*, *Gadd45b*, *Cdkn1a*, *Foxo3a* and *Fam107a* were differential regulated in prednisolone-treated WT mice, but not in GR<sup>dim</sup> mice. White bars represent vehicle-treated mice, whereas grey bars represent prednisolone-treated mice. Symbols; Triangles: female mice, Bullets: male mice.

Figure 6 Validation of microarray-obtained gene expression by Q-PCR. The relative expression of *Fam107a*, *Dusp1*, *Cdkn1a* and *Car8* in prednisolone-treated versus vehicle-treated wild type (WT) and GR<sup>dim</sup> mice was validated by Q-PCR. Differences in expression between two samples were calculated by the 2ΔΔCt method [103, 104]. Mean differences in gene expression between WT and GR<sup>dim</sup> mice were analyzed using the Student’s t-test. No expression was observed for *Fam107a* and *Cdkn1a* in vehicle and prednisolone-treated GR<sup>dim</sup> mice (Ct>30). Asterisks denote p-values as follows: *p < 0.05, **p < 0.01 and ***p < 0.001.
the prednisolone-induced genes are Foxo1 and Foxo3a target genes and are similarly regulated in Foxo1 and Foxo3a expression profiling experiments [54, 59, 60], this suggests that the GR synergizes with these transcription factors in mouse liver to control lipid and glucose metabolism and the cell cycle. This concept is further supported by the recent finding that the Murf1 (Muscle RING finger 1) promoter contains adjacent binding sites for the GR and Foxo transcription factors and is synergistically activated when both are co-expressed [75]. However, it is likely that a subset of the prednisolone-induced genes is under direct control of Foxo1 and Foxo3a without the need for synergy with the GR. Other studies, addressing small gene sets or individual genes, had also identified a role for Foxo1 and Foxo3a in GR-induced gene expression [58, 75-77].

The list of differentially regulated genes by prednisolone in WT mice overlaps with that of an earlier study by Phuc Le et al., which combined chromatin immunoprecipitation (ChIP) data with gene expression data to identify direct GR target genes in mouse liver [78]. Most of the genes found in our study displayed the same up and downregulation of gene expression, such as Tat, Foxo1 and Fkbp5 that are upregulated and Adipor2, CCAAT/enhancer binding protein alpha (C/EBPα) and Tkt that are downregulated. From a candidate gene set of 302 genes, Phuc Le et al. identified metabolism, cell proliferation and programmed cell death as important processes in their GO-term analysis, which is in agreement with our findings.

Interestingly, in contrast to our study, Phuc Le et al. did not observe upregulation of bona fide GR target genes, such as Pck and Igfbp in fed CD1 mice and explain this by lack of response due to a potential inhibition by insulin signaling in the fed state [79]. The fact that we observed upregulation of these genes indicates that in our experimental setup, which differs in several aspects from the setup used by Phuc Le et al, such as mouse strain used for the study, GC dosage and the fed state of the mice, the inhibitory effects of fed-induced insulin signaling do not play a role.

The overall induction of gene expression was strongly reduced in prednisolone-treated GRdim mice compared to WT mice (Figure 4). Nevertheless, in many cases residual gene induction by prednisolone was still observed. This indicates that GR dimerization is indeed an important mechanism for activation of these genes, although some regulation can take place even in the absence of GR homodimers. In fact, GR monomers are in principle capable of binding to GC response elements (GREs) and evoking a basal induction of gene expression. However, due to the lack of cooperative binding they are less potent than GR homodimers. This is in line with the regulation reported for Pnmt (Phenylethanolamine-N-methyltransferase) and Amy2 in which binding of GR monomers to GREs or a GRE half-site (i.e. only one-half of the classical palindrome) was sufficient to confer induction by GCs [80, 81]. In case of the Pnmt gene, multiple GRE half-site have been identified in the promoter region allowing receptor clustering and thereby stable binding of GR monomers independent of the DNA-binding domain (DBD) interface [80]. This also explains why expression of the Pnmt gene is not compromised in GRdim mice despite the
lack of GR dimer formation [21]. Another explanation for the residual gene induction by prednisolone in GR\textsuperscript{dim} mice is that the mutant GR still forms homodimers but that these are far less stable than in WT mice.

A recent study showed that the specific sequence of a GRE differently affects the conformation of the GR and thereby its activity towards specific target genes [82]. Mutation analysis of overexpressed GR domains that are involved in transcriptional activation, namely the dimerization region in the DBD (Dim) as well as AF1 and AF2 (activation function 1 and 2), showed that the dependency on each of them was specific for the sequence of the GR binding site and that genes differed in their dependence on dimerization [82]. The fold induction of \textit{Tat} and \textit{Fkbp5} by dexamethasone in the Dim mutant was around 30% for \textit{Tat} and 50% for \textit{Fkbp5} of that for WT GR, which is in gross agreement with our own observations \textit{in vivo} (Figure 4).

The effect of the GR\textsuperscript{dim} mutation was found in all major pathways that were induced by prednisolone, i.e. for genes in inflammatory pathways, gluconeogenesis and cell cycle. Moreover, the attenuating effect of the GR\textsuperscript{dim} mutation was found for genes that were upregulated as well as for genes that were downregulated by prednisolone. This suggests that genes, which fail to be repressed by prednisolone in GR\textsuperscript{dim} mice, are either regulated through GR binding to negative GRE elements (nGREs) or indirectly regulated via other GR target genes. Interestingly, analysis of the magnitude of gene regulation by prednisolone in WT mice compared to GR\textsuperscript{dim} mice showed that the cell cycle-related genes are more dependent on the dim interface than genes related to gluconeogenesis (Table 2 and 3). This can also be appreciated in Figure 4; several genes that are on the x-axis (i.e. show no differential gene expression upon prednisolone treatment in GR\textsuperscript{dim} mice), are cell cycle-related.

We identified several genes that showed strong upregulation by prednisolone in GR\textsuperscript{dim} mice compared to WT mice, such as \textit{Amy2}, \textit{Car8} and \textit{Zdhhc23}, and several genes that showed equal induction of gene expression by prednisolone in WT and GR\textsuperscript{dim} mice, amongst them are \textit{Dst} and \textit{Ppargc1a} (Figure 4). These genes are potential candidates for having GRE half-sites in their promotor regions, which could explain why these genes show equal or even higher induction of gene expression by prednisolone compared to WT mice. As mentioned earlier, the presence of GR half-sites in the promotor region for \textit{Amy2} was indeed experimentally confirmed [81]. For the other genes however, we did not find literature evidence for the presence of GR half-sites in their promotor regions. Therefore, follow-up studies on these genes to determine the functional GRE sites in their regulatory region would be of interest to study the significance of dimer-interface independent GRE binding. Genes that showed equal or stronger upregulation by prednisolone in GR\textsuperscript{dim} mice compared to WT mice can also be secondary response genes; under transcriptional control of transcription factors other than the GR. Keyword enrichment analysis performed with this set of genes did not identify enrichment for a particular cellular process.
We identified several genes that are absolutely dependent on GR dimerization for the induction of gene expression by prednisolone, such as *Cdkn1a*, *Gadd45b*, *Dusp1*, *Plk3* and *Foxo3a* (Figure 4). Interestingly, all of these have a strong relationship with p53: some are under direct transcriptional control of p53 (*Cdkn1a* [83], *Gadd45b* [84] and *Dusp1* [48]) or physically interact with p53 (*Foxo3a* [85, 86] and *Plk3* [87]). The cellular responses mediated by Foxo3a and p53 are highly similar, share some of their target genes (e.g. *Cdkn1a* and *Gadd45b*) and use similar mechanisms to regulate post-translational modification [51, 88]. Therefore, Foxo3a and p53 can be regarded as partners that positively as well as negatively regulate each other, depending on the context [51]. This observation might indicate that GC-induced activation of Foxo3a and/or p53 is hampered in GR<sup>dim</sup> mice. In line with this hypothesis Foxo3a was recently shown to be required for GC-induced apoptosis in lymphocytes [89]. Interestingly, this process is defective in GR<sup>dim</sup> mice highlighting a possible link between the GR, Foxo3a and induction of lymphocyte apoptosis [21].

*Fam107a* showed the largest induction of gene expression by prednisolone in WT mice (Figure 4 and 5). Analysis of protein-protein interactions revealed that Fam107a interacts with Tada2a [90, 91], a protein that together with binding partner Tada3a (Transcriptional adaptor 2 and 3 alpha) are core proteins of the histone acetyltransferase (HAT) complex [92]. HAT complexes are involved in chromatin structure modification for initiation of gene transcription, but can also acetylate non-histone proteins to modify their activity and stability [92, 93]. Interestingly, Fam107b a paralog of Fam107a with 84% protein similarity (not regulated in mouse liver by prednisolone) was shown to interact with Tada3a [94].

The observation that Fam107a inhibits cell proliferation and induces apoptosis when overexpressed [65, 67], suggests that Fam107a, like Foxo3a, Dusp1, Gadd45b, Cdkn1a and Plk3 play a role in regulating the cell cycle. Furthermore, the association of Fam107a with a core protein of the HAT complex might indicate that Fam107a may serve as a cofactor in the transcription machinery complex.

The activity and function of Foxo3a and p53 are strongly modulated by acetylation [51, 88, 95]. Hence, *Fam107a* is an interesting candidate gene for follow up experiments to study whether it modulates GR-induced gene expression and/or acetylation of GR-associated transcription factors such as Foxo3a, p53, C/EBPα and C/EBPβ (CCAAT/enhancer binding protein, beta).

For a more comprehensive view on GC regulated process in the liver, experiments with multiple time points, different doses of GCs and using one or more inflammatory stimuli, could be considered.
Methods

Animals

All mice (WT and GR\textsuperscript{dim}; Balb/c) were bred at the Institute of Virology and Immunobiology at the University of Würzburg. In total 24 mice (8 male WT, 4 female WT, 8 male GR\textsuperscript{dim} and 4 female GR\textsuperscript{dim}) were included in the study. Mice were treated subcutaneously with 1 mg/kg prednisolone (5% DMSO and 5% Chremophor in manitol, 10 ml/kg) once and sacrificed 150 minutes later by cervical dislocation, which was approved by the responsible authorities in Bavaria (Regierung von Unterfranken). All experiments were performed in the morning between 9 and 10 AM. The mice were exposed to a regular dark-light-cycle (lights on between 6 AM and 6 PM) and had access to water and food \textit{ad libitum} at any time.

RNA isolation

Liver biopsy specimens were collected into aluminum containers, snap freezed in liqN\textsubscript{2} and stored at -80°C before use. RNA isolation was done using Trizol, followed by RNeasy clean-up to enhance the $A_{260}/A_{230}$ ratio. RNA quantity and quality was determined using the NanoDrop Spectrophotometer and Agilent Bioanalyzer. For all samples subjected to microarray hybridization, the RIN (RNA integrity number) was 9.0 - 10.

Microarray data processing

Processed RNA samples were hybridized on GeneChip Mouse Genome 430 2.0 arrays (MOE430-2) from Affymetrix [96]. Processing and downstream statistical analysis of the microarray data was done using the R-Statistics package [97]. Data were normalized using the gcrma algorithm, pair-wise ratios between treatments were built using the limma package and annotation for the probe sets was derived from the mouse430-2 library, all as provided in BioConductor [98]. In all contrast matrices, a correction for gender type was applied. Data were deposited in the NCBI Gene Expression Omnibus (GEO), accession number GSE21048.

Keyword enrichment analysis

Keyword enrichment analysis on the microarray data was performed using CoPub [34] with default settings as provided by the web server [99]. The literature-network between enriched keywords and genes (nodes) and their co-publications (edges) were visualized using Cytoscape software [100, 101].

Validation of microarray results with Q-PCR

Total RNA (1 µg) was reverse transcribed using a commercially available cDNA synthesis kit (iScript, BioRad Laboratories, Hercules, CA, USA). Q-PCR was performed by SYBR Green-
based quantification according to the manufacturer’s protocol (Applied Biosystems, Foster City, CA, USA). Primers were developed for *Fam107a* (*Family with sequence similarity 107, member a*; Fw: TCATCAAACCCAAGAAGCTG; Rev: CTCAGGCTTGCTGCCATAC), *Car8* (*Carbonic anhydrase 8*; Fw: CACACCATTCAAGTCATCCTG; Rev: ACCACGCTGGTTTCTCTTC), *Cdkn1a* (*Cyclin-dependent kinase inhibitor 1a*; Fw: ATCTGCGCTTGGAGTGTAG; Rev: CTCTGCTCAGTGTTAGCTG), and *Dusp1* (*Dual specificity phosphatase 1*; Fw: GTGCGCTGACAGTGCAGAATC; Rev: CCAGGTACAGGAAGGACAGG) using Primer3 [102]. All primer pairs were exon-spanning. The gene *Azgp1* (*alpha-2-glycoprotein 1, zinc-binding*; Fw: AAGGAAAGCCAGCTTCAGAG; Rev: ACCAAACATTCCCTGAAAGG) was chosen as endogenous control since the expression arrays did not show any differences in expression of this gene between the two experimental groups (WT vs GRdim). PCR products were selected to be between 80 and 120 bp long. Samples were run on the 7500 Fast Real-Time PCR System (Applied Biosystems) using the following protocol: 10 min. denaturation at 95°C, and 40 cycles of 15 sec. denaturation at 95°C, 60 sec. annealing and extension at 60°C. All primer pairs were validated in triplicate using serial cDNA dilutions. Primer pairs that were 100+/-10% efficient, which implies a doubling of PCR product in each cycle, were used to quantify mRNA levels. Threshold cycle numbers (referred to as Ct) were obtained using the 7900 HT System SDS software version 2.3 (Applied Biosystems). All samples were measured for three times and samples with a standard deviation (SD) larger than 0.5 were excluded from the analysis. The relative quantity (RQ) of the gene-specific mRNA was calculated from the average value of the ΔCt (target gene Ct – endogenous control gene Ct) for each of the 24 analyzed samples. Differences in expression between two samples were calculated by the 2ΔΔCt method [103, 104]. For Q-PCR, mean differences in expression between groups were analyzed using the Student’s t-test. A p-value of <0.05 was considered statistically significant in each situation.

**Protein – protein interaction data**

Protein-protein interaction data were retrieved from the Biological General Repository for Interaction Datasets (BioGRID) database [105, 106].

**Abbreviations**

GR: Glucocorticoid receptor, WT: Wild type, GRdim: GR with the A458T point mutation in the dimerization region of the DNA-binding domain, GCs: Glucocorticoids, GRE: Glucocorticoid response element.

**Authors’ contributions**

RF, AvE, MJvL, JPT, HMR, WD, JdV and WA designed research. JPT and HMR generated and provided GRdim and WT mice. HvdM performed RNA isolations and microarray hybridization experiments. EJMT designed and performed Q-PCR experiments. RF, WF
and WA analyzed the microarray data and wrote the paper. All authors read and approved the final manuscript.

Acknowledgements
WF is sponsored by a grant received from the Netherlands Bioinformatics Centre (NBIC) under the BioRange program.
References


96. Affymetrix [http://www.affymetrix.com]
Prednisolone-induced differential gene expression in mouse liver
CHAPTER 7
Identification of T lymphocyte signal transduction pathways

Wilco W.M. Fleuren²,³, Ruben L Smeets¹,⁶,⁹, Xuehui He⁷, Paul M Vink¹, Frank Wijnands¹, Monika Gorecka¹, Henri Klop¹, Susanne Bauerschmidt⁴, Anja Garritsen¹, Hans JPM Koenen⁷, Irma Joosten⁷, Annemieke MH Boots¹,⁸ and Wynand Alkema²,³,⁴,⁵

¹Department of Immune Therapeutics, Merck Research Laboratories (MRL), MSD, Oss, the Netherlands. ²Computational Drug Discovery (CDD), CMBI, NCMLS, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. ³Netherlands Bioinformatics Centre (NBIC), Nijmegen, The Netherlands. ⁴Department of Molecular Design and Informatics, Merck Research Laboratories (MRL), MSD, Oss, the Netherlands. ⁵NIZO Food Research, Ede, The Netherlands. ⁶Department of Laboratory Medicine, laboratory for Clinical Chemistry, Radboud University Medical Centre, Nijmegen, The Netherlands. ⁷Department of Laboratory Medicine, laboratory for Medical Immunology, Radboud University Medical Centre, Nijmegen, The Netherlands. ⁸Department of Rheumatology and Clinical Immunology, University Medical Centre Groningen, University of Groningen, Groningen, The Netherlands. ⁹Department of Laboratory Medicine, laboratory for Clinical Chemistry, Radboud University Medical Centre, Nijmegen, Geert Grooteplein 10, Postbus 9101, 6500 HB Nijmegen, The Netherlands.

This work is adapted from ‘Molecular pathway profiling of T lymphocyte signal transduction pathways; Th1 and Th2 genomic fingerprints are defined by TCR and CD28-mediated signaling.’ Ruben L Smeets¹,⁶,⁹, Wilco W.M. Fleuren¹,³, Xuehui He¹, Paul M Vink¹, Frank Wijnands¹, Monika Gorecka¹, Henri Klop¹, Susanne Bauerschmidt⁴, Anja Garritsen¹, Hans JPM Koenen⁷, Irma Joosten⁷, Annemieke MH Boots¹,⁸ and Wynand Alkema²,³,⁴,⁵

Abstract

Background: T lymphocytes are orchestrators of adaptive immunity. Naïve T cells may differentiate into Th1, Th2, Th17 or iTreg phenotypes, depending on environmental co-stimulatory signals. To identify genes and pathways involved in differentiation of Jurkat T cells towards Th1 and Th2 subtypes we performed comprehensive transcriptome analyses of Jurkat T cells stimulated with various stimuli and pathway inhibitors. Results from these experiments were validated in a human experimental setting using whole blood and purified CD4+ T cells.

Results: Calcium-dependent activation of T cells using CD3/CD28 and PMA/CD3 stimulation induced a Th1 expression profile reflected by increased expression of *T-bet, RUNX3, IL-2,* and *IFNγ,* whereas calcium-independent activation via PMA/CD28 induced a Th2 expression profile which included *GATA3, RXRA, CCL1* and *Itk.* Knock down with siRNA and gene expression profiling in the presence of selective kinase inhibitors showed that proximal kinases *Lck* and *PKCθ* are crucial signaling hubs during T helper cell activation, revealing a clear role for Lck in Th1 development and for *PKCθ* in both Th1 and Th2 development. Medial signaling via MAPkinases appeared to be less important in these pathways, since specific inhibitors of these kinases displayed a minor effect on gene expression. Translation towards a primary, whole blood setting and purified human CD4+ T cells revealed that PMA/CD3 stimulation induced a more pronounced Th1 specific, Lck and PKCθ dependent *IFNγ* production, whereas PMA/CD28 induced Th2 specific *IL-5* and *IL-13* production, independent of Lck activation. PMA/CD3-mediated skewing towards a Th1 phenotype was also reflected in mRNA expression of the master transcription factor Tbet, whereas PMA/CD28-mediated stimulation enhanced *GATA3* mRNA expression in primary human CD4+ T cells.

Conclusions: This study identifies stimulatory pathways and gene expression profiles for in vitro skewing of T helper cell activation. PMA/CD3 stimulation enhances a Th1-like response in an Lck and PKCθ dependent fashion, whereas PMA/CD28 stimulation results in a Th2-like phenotype independent of the proximal TCR-tyrosine kinase Lck. This approach offers a robust and fast translational in vitro system for skewed T helper cell responses in Jurkat T cells, primary human CD4+ T cells and in a more complex matrix such as human whole blood.

Keywords: Signal transduction pathways, Gene expression profiling, T lymphocytes, Th1 and Th2 development
**Background**

Activation of T helper 0 (Th0) cells leads to differentiation into several lineages. These lineages include the Th1 and Th2 subsets as well as the more recently described subsets such as induced T regulatory cells and Th17 cells. The Th1 cells protect against intracellular pathogens and are in general characterized by their ability to produce \( \text{IFN}\gamma \), \( \text{IL-2} \) and \( \text{TNF}\alpha \) and express the Th1-specific transcription factor T-bet. The Th2 subset, which is involved in the defense against extracellular pathogens, is characterized by the production of \( \text{IL-4}, \ \text{IL-5} \) and \( \text{IL-13} \) and is controlled by the master transcription factor \( \text{GATA3} \) [1,2].

In a proper functioning immune system, these different T helper subsets are well-balanced and co-operate to eliminate invading pathogens and to maintain homeostasis. Hyper activation of one T helper subset, however, can tip the balance from health towards disease, in which Th2-overshoot can lead to inappropriate immune responses leading to diseases like allergy and asthma. Alternatively, overshoot towards a Th1 or Th17-phenotype can cause autoimmune diseases, like rheumatoid arthritis and multiple sclerosis [3,4]. For effective CD4 T cell activation, the antigen presenting cell (APC) provides a key contact point to facilitate T cell activation and polarization towards different T helper subsets. A crucial event in this process is the interaction between the antigen presented via the MHCII receptor and the TCR receptor (signal 1). The nature of activation, defined by the strength of the TCR stimulation, can affect T helper cell polarization towards Th1 or Th2, in which a high affinity interaction favors Th1 development and low affinity drives Th2 development [5-8]. Besides the TCR signal transduction, an additional signal is provided by the APC in the form of a co-stimulatory signal (signal 2). This signal is provided via CD28-B7 interaction and has been shown to be important for effective T cell activation [9]. Furthermore, CD28-mediated co-stimulation has been implicated in effective polarization of T cells towards a Th2 phenotype [10,11]. Also other co-stimulatory molecules, including ICOS and OX40, have been positively correlated with Th2 differentiation [12,13]. The results from these studies underline the importance of both signal 1 and signal 2, but also underline the complexity of these integrated signaling pathways.

The cascade of biochemical events, linking cell surface receptor engagement to cellular responses has been a focus of many studies. Detailed investigation of these signal transduction events has led to identification and functional characterization of many kinases and phosphatases downstream of the TCR and CD28-receptor. TCR ligation results in the recruitment of \( p56Lck \) (Lck), a proximal TCR Src family kinase, which kick-starts the signal transduction cascade leading to phosphorylation of the ITAM motifs in the TCR, which recruits and activates \( \text{ZAP70} \) [14]. This initial step leads to the activation of \( \text{PLC}\gamma \) that hydrolyzes \( \text{PIP}_2 \) into \( \text{IP}_3 \), which is the second messenger molecule responsible for the sustained intracellular calcium flux in T cells. CD28-ligation on T cells results in the recruitment of \( \text{PI}_3\text{K} \), with \( \text{PIP}_2 \) and \( \text{PIP}_3 \), which serve as pleckstrin homology (PH) domain membrane anchors. Via this mechanism \( \text{PDK1} \) and \( \text{PKB/Akt} \) are recruited and regulate...
Identification of T lymphocyte signal transduction pathways

several pathways that increase cellular metabolism [15]. Additionally, CD28-signaling has been shown to initiate NFB signaling, via a mechanism that is functionally linked through recruitment of PKCθ to CD28 in the immunological synapse [16-18]. Members of the Mitogen-activated protein kinase family, which can be activated via TCR signaling, also play a role in the differentiation of Th1 and Th2 subsets. In a thorough review by Dong et al., the role of p38, JNK and ERK in T helper cell differentiation has been outlined [19]. ERK is important for Th2 differentiation, whereas p38 and JNK2 appear to be involved in Th1 development. TCR/CD3 stimulation and CD28 stimulation alone are weak activators of T cell signaling. It is generally conceived that CD28 signaling merely acts as a signal potentiator on top of the initiator signal mediated via the TCR/CD3. Ledbetter and June et al. described that CD28 stimulation in the absence of cross-linking on top of PMA stimulation can activate T cells, without increasing calcium flux [20,21]. This suggests that co-stimulatory pathways synergize with biochemical pathways induced via the TCR. Whether CD28 ligation, in the absence of TCR signaling, leads to activation and differentiation has not been fully explored.

These findings show that effective T cell activation and differentiation towards effector subsets is the result of precise integration of multiple signaling routes. To explore the pathways underlying these distinct routes towards T cell activation and differentiation we used comprehensive biochemical characterization and gene expression profiling of Jurkat T cells that were activated with various co-stimulatory signals in the presence of various inhibitors of specific signaling routes. With this approach we identified specific PMA/CD3 and PMA/CD28 signal transduction and genomic fingerprints. PMA/CD3 stimulation induced a Th1 phenotype, dependent on both Lck and PKCθ, whereas PMA/CD28 stimulation, which is independent of TCR-mediated activation of Lck, resulted in a profound activation of T cells, skewing towards a Th2 phenotype.

Results and discussion

Activation of Jurkat T cells by various stimuli leads to differential signaling fingerprints

Jurkat T cells were activated by anti-CD3, anti-CD28, PMA, or ionomycin or combinations of these single stimuli, in order to map the contribution of these stimuli towards the activation of proximal, medial and distal signal transduction pathways. As shown in Figure 1A, CD3-stimulation and ionomycin/PMA were able to increase intracellular levels of Ca\(^{2+}\). Interestingly, neither CD28 nor PMA stimulation alone, affected intracellular Ca\(^{2+}\) levels. As expected, CD3-signaling resulted in an Lck-dependent phosphorylation of ZAP70 (Figure 1B). Stimuli containing PMA directly activated the MAPK pathway, which is reflected by the phosphorylation of ERK, P38 and JNK (Figure 1B). Furthermore, PMA addition directly activated PKC which was not reflected in the autophosphorylation of PKCθ, but was clearly detectable on the phosphorylation of the PKC substrate MARCKS (Figure 1B). CD3-mediated stimulations and PMA-induced stimulations resulted both in
the activation of AP1 family transcription factors c-Jun and ATF2 (Figure 1B). Analysis of nuclear translocation of NFATc1 and c-Jun (AP1)/NFκB p65, as part of the distal signaling events revealed that indeed CD3-mediated signaling induced both NFAT and c-Jun/NFκB, of which the latter pathways were potentiated by CD28-mediated signaling (Figure 1C). In line with the calcium release from the ER, PMA or PMA/CD28-mediated signaling did not induce NFAT nuclear translocation but highly activated the CD28 responsive element transcription factors c-Jun and NFκB p65 (Figure 1C). These results indicate that two distinct co-stimulatory profiles can be identified. A CD3/28 and PMA/CD3 stimulus that signals via Lck, increasing intracellular Ca\textsuperscript{2+} and activating NFAT, and a PMA/CD28 calcium independent (co)-stimulatory activation signaling via PKC\textgreek{theta} and MARCKS. Next, the molecular mechanisms involved in these signaling pathways were further explored in genomics studies.

**Differential regulation of genes after PMA/CD3 and CD3/28 Vs PMA/CD28 stimulation**

In order to further characterize the different signal transduction events induced by different (co)-stimulatory signals, we performed a first gene expression experiment with Jurkat T cells that were stimulated for 1 or 8 hours with PMA/CD3, CD3/28 and PMA/CD28. It appeared that after 1 hour a limited response on transcription level was seen, whereas after 8 hours of stimulation, several hundreds of genes were regulated. Furthermore, PMA/CD3 and PMA/CD28 regulate more genes compared to CD3/28, reflecting the strength of the stimuli used (Figure 2). Multivariate analysis by principal component analysis and hierarchical clustering showed that the 3 stimuli lead to clearly distinct gene expression profiles. At both time points the profile induced by PMA/CD28 is clearly distinct from the profiles induced by PMA/CD3 and CD3/28 (Figure 2).

**Differentially regulated genes **CCL1** and **IL-2** are profilespecific secreted proteins**

Gene profiles of the differential stimuli were ranked on the level of induction and evaluated on whether or not the translated protein is secreted. This resulted in the identification of the PMA/CD28-specific transcript CCL1 (Figure 3A), the CD3-specific transcripts IL-2 (Figure 3B) and XCL1/2 (data not shown). Small but significant inductions of these genes were observed after 1 hour of stimulation for both CCL1 and IL-2. However, both genes were highly induced after 8 hours of stimulation. The secretion of the protein 24 hours after stimulation reveals an identical profile compared to the mRNA (Figure 3A/B, right hand panel).
Figure 1 Signal transduction in Jurkat T cells. Jurkat T cells were stimulated with different combinations of stimuli in order to elucidate the different signal transduction pathways. A; Jurkat T cells were stimulated as indicated and intracellular Ca2+ release was monitored over/in time. B; Intracellular signal transduction routes were charted via phospho-analysis using western blot. Jurkat T cells were stimulated for 15 min using different stimulations. Proximal (Lck, ZAP70, PKCθ and the PKC substrate MARCKs), medial (MAPK phosphorylation) and distal (c-Jun and ATF2) signaling was monitored based on the phosphorylation status of the described proteins. C; Nuclear translocation of the transcription factors NFAT, NFκB and c-Jun was evaluated 15 minutes after stimulation.

Pathway profiling with multiple stimuli and inhibitors

To investigate the contribution of proximal, medial and distal signaling events on the CD3/28, PMA/CD3 and PMA/CD28 stimuli, we performed a second gene expression profiling experiment with different selective inhibitors, including proximal kinase inhibitors.
Chapter 7

Lck (A420983), PKCθ (AEB071), medial MAPK inhibitors PD98059 (MEK/ERK), SP600125 (pan JNK), Org 48762-0 (P38) and the distal Calcineurin (Cn) inhibitor Cyclosporin A (CsA). Jurkat T cells were stimulated with PMA, CD28 and CD3 alone and combinations thereof in the presence of the above mentioned inhibitors. Based on the results of the first gene expression profiling experiment, we chose evaluated gene expression after 8 hours of stimulation.

A principal component analysis on the ratio data set is shown in Additional file 1: Figure S1. It appeared that the PMA/CD28, CD3/CD28 and PMA/CD3 co-stimuli induced several hundreds of genes, whereas the effect of a single stimulus was smaller, with the exception of the PMA single stimulus (Additional file 1: Figure S1, Additional file 2: Table S1). The gene set induced by a PMA stimulus showed a larger overlap with the genes induced by the PMA/CD28 stimulus than with the CD3/PMA induced gene set. Whereas PMA and CD3 as a single stimulus induce a large number of genes, CD28 elicits only a minor effect. It can be observed that CsA, AEB071 and A420983 induce the largest effects on gene regulation, whereas the inhibitors of the MAPK pathway only have a minor effect on gene expression. This finding is corroborated by the number of regulated genes, showing that the MAPK inhibitors only regulate a small number of genes whereas A420983, CsA and AEB071 regulate many genes (Additional file 2: Table S1). A420983 and CsA only show a significant effect on the PMA/CD3 and CD3/CD28 pathways, in which the effect of CsA

Figure 2 PMA/CD3 and CD3/28 stimulations differ from PMA/CD28 stimulation. The table shows the number of regulated probe sets and their overlap at 1 and 8 hrs following stimulation with CD3/PMA, PMA/CD28 and CD3/CD28 (using a fold change cut off of 2 and a p-value cut off of <1.10^-6). The right panel shows a hierarchical clustering of the data using the data for the regulated probe sets. Most of the variation is caused by the time difference (1hr vs 8 hrs). At both time points the PMA/CD28 stimulus was clearly different from the CD3/CD28 and CD3/PMA stimulus. Repeating this analysis with different values for the fold change and p-value cut off yielded essentially the same results for the multivariate analysis.
is smaller than the effect of A420983. AEB071 is the only compound that shows also a significant effect on the PMA/CD28 induced pathway. These analyses were rerun with different settings for the thresholds used for gene selection. In all cases similar results were obtained indicating that the results were not critical dependent on the thresholds settings that were used.

Inspection of the profiles of CCL1 and IL-2 revealed that CCL1 mRNA is highly induced via the PMA/CD28 pathway. This induction is depending on PKC signaling and negatively regulated via Lck signaling. Apparently this effect was upstream of Cn, since the inhibitor CsA did not increase CCL1 mRNA induction (Figure 4A).

![Figure 3 Differential regulation of IL-2 and CCL1.](image)

**Figure 3** Differential regulation of IL-2 and CCL1. Jurkat T cells were activated for 1 and 8 hours using different stimuli (as indicated). Differentially regulated genes of which the proteins were secreted were identified and the highest ranking genes were selected. A; regulation of the CCL1 mRNA is highly and differentially regulated after PMA/CD28 stimulation. Right panel shows the validation of the mRNA levels on protein level, secreted by activated Jurkat T cells. B; regulation of IL-2 mRNA is highly and differentially regulated after CD3/28 and PMA/CD3 stimulation. Right panel shows the validation of the mRNA levels on protein, secreted by activated Jurkat T cells.
As expected, PMA/CD3 and CD3/28 stimuli and to a lesser extent PMA/CD28 resulted in a marked expression of IL-2, which is highly depending on the Lck/Cn signal transduction pathway, and the PKC pathway. Interestingly inhibition of MAPK signaling (with the
exception of the MEK/ERK pathway) does not affect IL-2 mRNA induction (Figure 4B). These effects on CCL1 and IL-2 production by inhibitors of the Lck/Cn and PKC pathway were further substantiated in a full dose-response experiment. Figure 5A shows that indeed AEB071 dose-dependently inhibited PMA/CD28-induced CCL1 production, which is slightly enhanced in the presence of an Lck inhibitor and which is not affected by Cn inhibition via CsA. IL-2 production can be blocked by inhibition of both the Lck/Cn and PKC pathway (Figure 5B). The involvement of Lck in the CD3-mediated pathway and PKC in the CD3 and CD28-mediated pathways was further confirmed by the knock-down of both kinases under the distinct stimuli. Knock-down of Lck did not affect PMA/CD28-induced CCL1 production, whereas knock-down of PKCθ resulted in significant inhibition of both IL-2 and CCL1 (Figure 5C). These results clearly show that PMA/CD28-induced gene profiles are highly depending on PKCθ signaling pathways but are independent of Lck/Cn and MAPK signaling pathways, whereas CD3-mediated signaling pathways are dependent on both Lck/Cn and PKCθ signal transduction and independent on MAPK signaling events.

PMA/CD3, PMA/CD28 and CD3/CD28 induce distinct genomic fingerprints

The above analysis indicated that treating Jurkat T-cells with multiple combinations of stimuli and inhibitors highlights pathways that are regulated by specific combinations of stimulus and inhibitor, revealing the involvement of certain kinases as signaling hub under specific stimulatory conditions. In order to identify additional genes in the pathways that are exemplified by CCL1 and IL-2, we searched for genes with similar profiles to these pathway genes. Figure 6 shows genes related to CCL1, identified by a strong up regulation following PMA/CD28 stimulation only and a down regulation by AEB071. Interestingly, besides CCL1, which is a chemo attractant for Th2 cells, many Th2-associated genes co-clustered with CCL1, including GATA3, Itk, RXRA, c-FLIP (CFLAR), ICOS and the IL-31 receptor and also other genes that are associated with Th2 development (See Figure 6; CCL1 gene cluster).

Likewise a very specific IL-2 profile can be constructed by selecting genes that are up regulated under all three conditions and down regulated by all three inhibitors, and by which CsA is the weakest inhibitor (see Figure 6; IL-2 cluster). In this gene cluster appeared to be Th1- associated genes, including the Th1 master transcription factor Tbet (TBX21), Th1 chemokine XCL1/2, IFNγ, granzyme, RUNX3, FASL, OX40L (TNFRSF4), CD27, and the IL-21 receptor. Of note, inhibition of both Lck and Cn under PMA/CD3 stimulation enhanced the expression of Th2 master transcription factors GATA3 and RXRA, but also peptidoglycan recognition protein 4 (PGLYRP4) and G protein-coupled receptor 84 (GPR84).

A third example is provided by genes clustering together with EGR1, which show an up regulation by all stimuli but are specifically regulated by AEB071 in all conditions and only by A420983 after CD3/CD28 stimulus. Although IL-2 and EGR1 show a similar regulation
by the stimuli used, the profiles can clearly be discriminated by the effects of the various inhibitors on their expression. The list of genes that are shown in Figure 7 together with their annotation and the correlation score to the CCL, IL-2 and EGR1 profiles are shown in Additional file 1: Figure S1, Additional file 2: Table S2, Additional file 3: Table S3. This analysis shows that by applying multiple stimuli and selective compound treatments, pathways can be unraveled at high resolution.

Figure 5 PMA/CD28-induced CCL1 production is not dependent on the Lck/Cn pathway. Jurkat T cells were stimulated using PMA/CD3 and PMA/CD28 in the presence of Lck (A420983), Cn (CsA) and PKC (AEB071) pathway inhibitors for 24 hours. Dose-response effects of the inhibitors were evaluated on the production of CCL1, after PMA/CD28 stimulation (A) and IL-2 after PMA/CD3 stimulation (B) in supernatant of the cell cultures. The data are representative for 3 independent experiments. C; Knock down of Lck and PKCθ resulted in a clear dose-dependent reduction of the protein (500, 100, 20 nM siRNA). 24 hour culture supernatants were collected after stimulation with PMA/CD28 and PMA/CD3 and the effect of knock down on respectively CCL1 and IL-2 was determined. Data of two independent experiments are presented as mean + SEM. Significance of differences are indicated by ** p<0.01, *p<0.05, # no difference using a one-way ANOVA with a Bonferroni's post-hoc test. N.b the values shown in this figure are log2 values no to be confused with the intensity values as shown in figure 3.
Identification of T lymphocyte signal transduction pathways

Chapter 7

A  CCL1 related genes

B  IL2 related genes
In order to assess whether the stimulation profiles and signal transduction profiles identified in Jurkat T-cells were also relevant in a primary human setting, the stimulation protocols were adapted and a primary assay was established using healthy human donor whole blood. The effect of differential stimulation was evaluated using IFNγ and IL-17 as Th1- and Th17-associated read-outs respectively and both IL-5 and IL-13 as Th2-associated read-outs. Stimulation of human whole blood cells with CD3/CD28 was unsuccessful: no cytokine release was detected. However, PMA/CD3 stimulation of human blood cells, resulted in a high production of IFNγ (Figure 7A). Interestingly, INFγ production levels were significantly lower after PMA/CD28 stimulation. A similar observation was seen

Figure 6 Heatmap of stimulation dependent gene clusters. A: CCL1 cluster characterized by induction of genes exclusively following CD28/PMA stimulation and subsequent repression by AEB071. B: IL-2 cluster characterized by induction of genes by all stimulations, and down-regulated by A420983 and AEB071 and to a lesser extent by CsA. C: EGR1 cluster characterized by induction of genes by all stimulations, but down-regulated only by AEB071 when stimulated with PMA/CD28 or PMA/CD3. A red color denotes an up-regulation, a green color a down-regulation. The MAPK inhibitors induced very little regulation and have been omitted from this figure for clarity. The gene lists shown in this figure, with extended annotation and their distance to the CCL1, IL-2 and EGR1 profiles are provided in supplemental files 1, 2 and 3.

Translation of PMA/CD3 and PMA/CD28 stimulations; differential modulation of primary T cell cytokine responses in human donor blood

In order to assess whether the stimulation profiles and signal transduction profiles identified in Jurkat T-cells were also relevant in a primary human setting, the stimulation protocols were adapted and a primary assay was established using healthy human donor whole blood. The effect of differential stimulation was evaluated using IFNγ and IL-17 as Th1- and Th17-associated read-outs respectively and both IL-5 and IL-13 as Th2-associated read-outs. Stimulation of human whole blood cells with CD3/CD28 was unsuccessful: no cytokine release was detected. However, PMA/CD3 stimulation of human blood cells, resulted in a high production of IFNγ (Figure 7A). Interestingly, INFγ production levels were significantly lower after PMA/CD28 stimulation. A similar observation was seen
when analyzing IL-17 production. Thus, higher production levels of both IFNγ and IL-17 were seen following PMA/CD3 stimulation when compared to PMA/CD28 stimulation. Furthermore, when analyzing Th2-associated IL-5 and IL-13 production, we found that PMA/CD28 stimulation was superior to PMA/CD3 stimulation in enhancing production of these cytokines (Figure 7B). Of note, CCL1 production could not be detected in this assay system (data not shown). In aggregate, the data suggests that PMA/CD28 stimulation favours Th2 responsiveness in this assay. Since PMA/CD28 signaling was shown to be independent of Lck, but mainly dependent on PKCθ whereas PMA/CD3 signaling was both Lck and PKCθ dependent we evaluated the effect of both proximal kinases in this human whole blood assay and evaluated IFNγ and IL-13 production since these cytokines were most readily produced. Figure 7C shows that indeed PMA/CD3-induced IFNγ production is depending on both Lck and PKCθ signaling, whereas PMA/CD28-induced IL-13 production is Lck-independent and PKCθ dependent. These results clearly show that the differential stimulations identified in the Jurkat assay can be translated towards a primary human cellular assay and are depending on the same proximal signaling hubs. Furthermore, also in this setting it can be observed that PMA/CD3 stimulation diverges more towards a Th1-like phenotype, whereas PMA/CD28 stimulation skews more towards a Th2-like response.

**PMA/CD3 stimulation of purified human CD4+ T cells enhances Th1 activation, whereas PMA/CD28 potentiates Th2 activation**

Using purified human CD4+ T cells we validated the effects observed on Jurkat T cells and in a primary human whole blood assay setting. Of all stimuli used PMA/CD3 appeared to be the most powerful stimulus able to induce IFNγ production. Also in this setting inhibition of either Lck using A-420983, or PKC using AEB071 completely inhibited PMA/CD3 induced IFNγ production. CD3/CD28-mediated stimulation, which can be successfully applied in this assay format induced IFNγ production, which was dependent on both Lck and PKC mediated signal transduction pathways. Of interest and comparable to the effects observed on CCL1 production by Jurkat T cells, Lck inhibition under PMA/CD28 stimulation did not inhibit IFNγ production and even appears to slightly enhance IFNγ production (Figure 8A). The observed effect on IFNγ production of the different stimuli is in line with the effects observed on the induction of the Th1 master transcription factor Tbet (Figure 8B). Both inhibition of Lck and PKC reduced CD3/28 and PMA/CD3 mediated induction of Tbet, whereas Lck inhibition did not affect PMA/CD28-induced expression of Tbet. PMA/CD28 was the most profound inducer of IL-13 in CD4+ cells (Figure 8B). Interestingly IL-13 production under al stimulatory conditions used is dependent on PKC, whereas Lck inhibition does not affect IL-13 production under PMA/CD3 or PMA/CD28 culture conditions. Under all culture conditions inhibition of PKC reduced IL-13, which was paralleled with reduced GATA3 expression (Figure 8D), whereas inhibition of Lck appeared
to promote Th2 development under all stimuli used, which was reflected by enhanced expression of GATA3.
Identification of T lymphocyte signal transduction pathways

Chapter 7

Text mining of IL-2 cluster genes and CCL1 cluster genes.

The gene expression study showed that PMA/CD3 induced an IL-2 gene profile which has been associated with a Th1 type of response. Similarly, PMA/CD28 induces a CCL1 gene profile which is associated with a Th2 response. In order to find additional evidence for the Th1 cell specificity of IL-2 and the Th2 cell specificity of CCL1, we used CoPub to search in Medline abstracts for other studies which support these results.

First we performed a keyword enrichment analysis of the genes that showed a similar gene expression profile to CCL1 (CCL1 cluster) and of the genes that showed an expression profile similar to IL-2 (IL-2 cluster).

Figure 8 PMA/CD3 and PMA/CD28 effects on purified primary human CD4+ T cells. Purified CD4+ T cells were stimulated with anti-CD3/CD28, PMA/CD3 and PMA/CD28 and cultured for 24 hours in the presence or absence of pathway inhibitors for Lck (A420983; 1µM) and PKC(AEB071; 1µM). Culture supernatants were analyzed for IFNγ production (A), mRNA expression of the Th1 master transcription factor T-bet (B), the production of the Th2 cytokine IL-13 (C) and expression of the Th2 master transcription factor GATA3 (D). Message RNA expression was normalized against ribosomal house hold gene RPL19. Data are presented of three independent experiments using purified human CD4+ T cells from three different healthy donors. Data are presented as mean + SD. Significance of differences are indicated by ** p<0.01, * p<0.05, using a mann-withney U-test.

Text mining of IL-2 cluster genes and CCL1 cluster genes.

The gene expression study showed that PMA/CD3 induced an IL-2 gene profile which has been associated with a Th1 type of response. Similarly, PMA/CD28 induces a CCL1 gene profile which is associated with a Th2 response. In order to find additional evidence for the Th1 cell specificity of IL-2 and the Th2 cell specificity of CCL1, we used CoPub to search in Medline abstracts for other studies which support these results.

First we performed a keyword enrichment analysis of the genes that showed a similar gene expression profile to CCL1 (CCL1 cluster) and of the genes that showed an expression profile similar to IL-2 (IL-2 cluster).
Table 1 Gene set enrichment analysis of CCL1 cluster and IL2 cluster with cell terms from the CoPub database. P-value < 0.01.

<table>
<thead>
<tr>
<th>CCL1 cluster</th>
<th>IL2 cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell name</td>
<td>Nr. of genes</td>
</tr>
<tr>
<td>macrophage</td>
<td>16</td>
</tr>
<tr>
<td>epithelial cell</td>
<td>15</td>
</tr>
<tr>
<td>monocyte</td>
<td>14</td>
</tr>
<tr>
<td>fibroblast</td>
<td>14</td>
</tr>
<tr>
<td>Thymus derived lymphocyte</td>
<td>13</td>
</tr>
<tr>
<td>Lymph cell</td>
<td>13</td>
</tr>
<tr>
<td>Neoplastic cell</td>
<td>13</td>
</tr>
<tr>
<td>Cancer cell</td>
<td>12</td>
</tr>
<tr>
<td>dendritic cell</td>
<td>12</td>
</tr>
<tr>
<td>hepatocyte</td>
<td>11</td>
</tr>
<tr>
<td>endothelial cell</td>
<td>11</td>
</tr>
<tr>
<td>Th2 cell</td>
<td>10</td>
</tr>
<tr>
<td>Keratinocyte</td>
<td>10</td>
</tr>
<tr>
<td>Bursa-equivalent lymphocyte</td>
<td>10</td>
</tr>
<tr>
<td>leukocyte</td>
<td>10</td>
</tr>
<tr>
<td>Peripheral blood mononuclear cell</td>
<td>9</td>
</tr>
<tr>
<td>Antigen-presenting cell</td>
<td>9</td>
</tr>
<tr>
<td>mast cell</td>
<td>9</td>
</tr>
<tr>
<td>Neutrophilic leukocyte</td>
<td>9</td>
</tr>
<tr>
<td>granulocyte</td>
<td>9</td>
</tr>
<tr>
<td>eosinophil</td>
<td>8</td>
</tr>
<tr>
<td>Effector cell</td>
<td>8</td>
</tr>
<tr>
<td>Breast cancer cell</td>
<td>7</td>
</tr>
<tr>
<td>Th1 cell</td>
<td>7</td>
</tr>
<tr>
<td>adipocyte</td>
<td>7</td>
</tr>
<tr>
<td>splenocyte</td>
<td>7</td>
</tr>
<tr>
<td>thymocyte</td>
<td>7</td>
</tr>
<tr>
<td>T-lymphocyte,cytotoxic</td>
<td>7</td>
</tr>
<tr>
<td>natural killer cell</td>
<td>7</td>
</tr>
<tr>
<td>Melanoma cell</td>
<td>6</td>
</tr>
<tr>
<td>Muscle cell</td>
<td>6</td>
</tr>
<tr>
<td>stromal cell</td>
<td>6</td>
</tr>
<tr>
<td>Stem cell</td>
<td>6</td>
</tr>
<tr>
<td>T4 lymphocyte</td>
<td>5</td>
</tr>
</tbody>
</table>
Identification of T lymphocyte signal transduction pathways

Both the enrichment analysis with the IL-2 cluster and the CCL1 cluster resulted in similar enriched terms such as ‘macrophage’, ‘lymph cell’, ‘Thymus derived lymphocyte’ and ‘monocyte’ (Table 1). These results are not surprising since both clusters of genes were found to be expressed in Th1 and Th2 cells. Th1 and Th2 are T helper cells which play important roles in the immune system in recruiting other cells. This analysis confirms that these clusters of genes are involved in immune responses.

Both the CCL1 cluster and the IL-2 cluster are enriched with the terms ‘Th1 cell’ and ‘Th2 cell’. Therefore we studied into more detail the relation of each cluster with Th1 and Th2 cell types. We created a literature network, searching for co-occurrences between the genes in each cluster and identifying the link of the genes in the network with Th1 and Th2 cells in Medline abstracts. In Figure 9A the network of the IL-2 cluster is shown. In the network we can identify 7 hub-genes, i.e. IL-2, IFNG, IL2RA, GZMB, FASLG, TNFRSF4 and TNFRSF7, which are connected to several other genes in the network. With the exception of GZMB, all these hub genes have connections to both Th1 and Th2. From the network it is also clear that Th1 has a slightly higher connectivity (9 connections) than Th2 (8 connections). In Figure 9B the network of the CCL1 cluster is shown. Also in this network highly connected genes such as CCL1, CXCL10, CXCL9, GATA3 and ISG20 have connections to both Th1 and Th2. Hence from the literature network, the specificity of the genes for either Th1 or Th2 is not immediately clear. However, in the network (Figure 9B) Th2 has more connections (9 connections) than Th1 (7 connections).

![Figure 9 Literature networks of the IL-2 cluster (A) and the CCL1 cluster (B). In both networks the link to Th1 cell and Th2 cell is shown. The thickness of the line between co-occurring genes or between genes and cell terms indicates the strength by means of the R-scaled score. In the networks IL-2 and CCL1 are highlighted with a red border color.](image-url)
Conclusions

In this study we systematically explored pathways involved in T cell activation by molecular profiling. We showed that TCR (both CD3/28 and PMA/CD3) driven stimulation profiles are truly distinct from co-stimulatory profiles mediated via PMA/CD28. Secondly, using selective inhibitors and siRNA we found that the proximal kinase Lck is involved in CD3 and not PMA/CD28 activation, whereas PKCθ appears to be a crucial central signaling kinase in both TCR and PMA/CD28 (co)-stimulatory activation of T cells. Finally, stimulations involving TCR/CD3 appear to preferentially induce a Th1-like fingerprint, whereas lack of TCR/CD3 signaling in the presence of PMA/CD28 stimulation diverts T cells towards a Th2-like state.

It has been suggested that the strength of TCR-signaling can regulate the fate determination of naive T cells; high potency signals skew towards Th1 differentiation, whereas low potency signals promote Th2 differentiation [6,22]. Although TCR and co-stimulatory pathways have been the focus of many studies in the previous decades, the direct contribution of TCR stimulation vs. co-stimulatory signals towards Th differentiation is not fully understood. By stimulating T cells with PMA/CD3 and PMA/CD28 we dissected signaling pathways and explored the activation profiles. CD3-mediated signaling rapidly increased intracellular Ca²⁺, a second messenger to activate many enzymes including Calcineurin, which resulted in an increased nuclear translocation of NFATc1. Interestingly, PMA/CD28 stimulation did not result in a Ca²⁺-mediated response (and was therefore marked a calcium-independent stimulation) but enhanced many of the co-stimulatory mediators including MAPK/AP1 and NFB signal transduction. These results are in line with earlier studies that showed differential effects of cyclosporine A (CsA) and dexamethasone on CD3 vs CD28-mediated signaling, which revealed that PMA/CD28 stimulation was insensitive towards CsA-mediated Calcineurin inhibition in contrast to PMA/CD3 stimulation [20,23].

Gene expression induced by combinations of stimulatory signals revealed pathway-specific biomarkers or fingerprints. PMA/CD3-induced gene profiles included IL-2, IFNγ, XCL1, granzyme B, and FASL, which have been associated with a Th1 type of response [24,25]. Also, sustained NFAT signaling, which is also induced via PMA/CD3 stimulation, has been shown to promote Th1-like gene transcripts, including IFNγ, FasL and P-selectin glycoprotein ligand 1 [26]. Our results are further substantiated by the finding that T-bet (TBX21), the Th1 master transcription factor [27], and RUNX3, which together with Tbet are crucial for inducing IFNγ and repressing IL-4 [28], were highly expressed under PMA/CD3-stimulatory conditions.

PMA/CD28 stimulation does not induce a Ca²⁺ flux nor does it increase nuclear translocation of NFAT. However it provides the cell with a high level of co-stimulatory signaling, and induces a completely distinct genomic fingerprint compared to PMA/CD3 stimulation. Following PMA/CD28 stimulation, Jurkat T cells highly expressed CCL1/
IL309, a chemokine which is highly expressed during a Th2-eosinophil response in allergic airway diseases [29,30]. Lymphotoxin (LT), a cytokine which is associated with a Th2-type of response controlling IgE production [31], was also highly expressed under PMA/CD28 stimulation. In conjunction with this finding, the master transcription factors for Th2, GATA3 [32] and the Retinoid X Receptor (RXR) [33], were induced under the PMA/CD28 stimulatory condition. Notably, Th2-associated cytokines like IL-4, IL-5 and IL-13 were not induced in Jurkat T cells after PMA/CD28-stimulation, this in contrast with PMA/CD28 stimulation of human whole blood and purified CD4+ T cells, which could be due to the developmental blockage of Jurkat T cells. Additional file 6: Figure 2 shows a schematic overview summarizing the involvement of the signaling pathways and genes induced under differential stimulation as observed in this study and highlights their relation towards T helper 1 and -2 development.

Our results are in line with the notion that high calcium levels drive Th1 and CTL responses and low calcium levels drive Th2 responses [7,34,35], which was further substantiated by our results using inhibitors for Lck and Cn, which modulate Calcium signaling in T cells. These inhibitors repressed Th1-associated genes under PMA/CD3-stimulation, but induced Th2 transcription factors GATA3 and RXRA, revealing a skewing of Th1 towards Th2 profiles. In contrast, PMA/CD28 stimulation in the presence of Lck and Cn inhibition, Th2-associated genes, e.g. CCL1 or IL-13 in CD4+ T cells, were not affected or even induced. The crucial role of Calcium and Lck in driving Th1 response is in line with the observation that knock down of Lck affects the virus-specific Th1/CTL response in mice and Lck deficiency increases Th2 associated cytokine production [36,37]. Interestingly, lack of Calcium signaling can give rise to an anergic T cell phenotype (reviewed in [38]). Therefore it would be of interest to further explore the role of Lck in calcium-dependent activation via PMA/CD3 on Th1/CTL responses and calcium independent activation of T cells via PMA/CD28 on the induction of anergy in more detail.

CD28 signaling has been functionally linked with PKCθ induced activation of NFB [39], which was also validated using PMA/CD28 as stimulus [40]. Previously it has been reported that CD28-costimulation induces GATA3 expression and Th2 differentiation via the activation of NFB [41,42]. Additional studies in mice revealed that PKCθ is involved in mounting both Th2- and Th1-mediated lung inflammation, although Th2-mediated inflammation is more PKCθ-dependent [43]. Our studies show that inhibition of PKCθ can indeed inhibit a PMA/CD28 stimulation, which was reflected by the effect of PKCθ inhibition on the PMA/CD28-induced Th2-like gene expression profile. These observations are in line with the results from CD28 knock-out mice and inhibition of CD28 signaling using CTLA4Ig, showing that the CD28 co-stimulatory signaling is crucial for mounting a proper Th2 response. In contrast, Th1 and CTL responses were found to be less dependent on CD28 signaling [44,45]. Of interest, PKCθ inhibition in our hands, also affected PMA/CD3-induced Th1-like expression profiles. These results underline the duality of PKCθ in the integration of TCR and CD28-mediated signaling events which is evident from PKCθ
KO mice experiments.

Text mining analysis has been used to find additional evidence for the observed specificity of the *IL-2* cluster for Th1 cells and the *CCL1* cluster for Th2 cells. Keyword enrichment analysis with both gene clusters did not show this specificity for either Th1 or Th2. However enrichment analysis resulted in immune related cells for both gene clusters, thus confirming the experimental observed results presented in this paper. Literature networks for both the *IL-2* and the *CCL1* cluster showed that the highly connected genes in each network are connected to both cell types. Although in the *IL-2* network, Th1 has a higher connectivity than Th2 and in the *CCL1* network Th2 a higher connectivity than Th1, the cell specificity of both clusters is not clearly visible. Moreover in the literature networks is shown that both Th1 cell and Th2 cell are highly connected to each other.

The reason that the cell specificity could not be detected using co-occurrence based text mining could lie in the fact that in abstracts in which Th1 occurs, in more than 30% of these abstracts also Th2 occurs and vice versa. This makes it impossible to find a specificity of the genes for one of the cell types using co-occurrence based methods such as CoPub. Hence in order to detect this type of specificity in text, other types of information extraction methods such as natural language processing (NLP) are needed. NLP is able to identify the type of relation between a gene and a cell type en enables to search for specific relations.

Finally, our results also show that this differential stimulation does not only occur in Jurkat T cells, but also plays a role in primary human T cells. These cells were found to secrete a Th1-like response (*Tbet-IFNg*) via PMA/CD3 stimulation, whereas PMA/CD28 stimulation led to a Th2 activation profile (*GATA3-IL-5/IL-13*). In these cells inhibition of the Lck/Cn/NFAT pathway was only effective after PMA/CD3 stimulation whereas inhibition of PKCθ inhibited both PMA/CD3-induced IFNg production and PMA/CD28-induced *IL-13* production. These results illustrate that the findings in the Jurkat T cell line were successfully translated and relevant to a human primary cellular setting. Interestingly, PMA/CD3 stimulation also enhanced *IL-17* production in the primary human whole blood assay and increased the expression of the *IL-21* receptor, which is crucial for Th17 induction [46,47], in Jurkat T cells. These results suggest that additional signals, like *IL-21* in conjunction with *TGFβ* and *IL-6*, might be necessary to differentiate from a Th1-like phenotype towards a Th17 phenotype, whereas the absence of TGFβ in the presence of high levels of *IL-2* will favor Treg development or stabilization. Therefore further exploration of these differential stimulations in the presence of defined/different cytokine stimuli could further elucidate T helper cell differentiation and establish sub-set specific genome profiles. The findings described in this paper offer a robust platform for in vitro activation of T cells, in which observed responses can be easily translated form Jurkat T cells, towards purified CD4+ T cells and even human whole blood. This can be of interest for efficiency and selectivity profiling of kinase inhibitors or for pathway-specific biomarker identification for future drug development and clinical studies.
Chapter 7

Methods

Compounds
Inhibitors selectively targeting defined pathways used in this study were A-420983 (1 M; Lck inhibitor) [48], AEB-071 (10 M; PKCθ inhibitor) [49] and Cyclosporin A (1 µM; Calcineurin inhibitor). Additionally, inhibitors of the MAPK pathway, SP600125 (10 µM; pan JNK), PD98059 (10 µM; MEK1/2), Org 48762-0 (10 M; P38) [50] were used. All compounds were dissolved in 100% DMSO. Maximal and final concentration of DMSO used in the culture assays was 0.1% v/v.

Cell culture
Jurkat E6.2.11 T cells were cultured in DMEM F12 medium (#041-94895 M, Gibco) supplemented with 10% FBS (#10099-141, Invitrogen) and 80 U/ml penicillin/80 g/ml streptomycin (#15140-122 Gibco). Cells were cultured at concentrations between 1-2 × 10^5 cells/ml at 37°C/5% CO₂. Cells were stimulated for 15 minutes up to 24 hours with anti-CD3 (1 µg/ml, OKT3), anti-CD28 (1 µg/ml, pericluster CD28 #M1456 Sanquin, the Netherlands), PMA (10 ng/ml, Sigma, USA) and ionomycin (1 µg/ml, Sigma, USA) or combinations thereof.
For gene expression profiling Jurkat T cells were seeded in T25 culture flasks at a concentration of 1 × 10^6 cells/ml (1 × 10^7 cells in total) and cultured overnight at 37°C/5%CO₂, one day prior to stimulation. On the day of the experiment cells were preincubated with the compound of interest for 30 minutes, followed by a stimulation with either CD3/CD28, PMA/CD28 or PMA/CD3, at concentrations of 10 ng/ml PMA, 1 µg/ml CD3 and 1 µg/ml CD28. Jurkat T cells were cultured in the presence or absence of stimulation for one or eight hours in total, after which the cells were washed in ice cold PBS. Thereafter cell pellets were collected and snap frozen at -80°C. Cell pellets were stored until further processing.

Isolation and quality check of mRNA
Total RNA was isolated from Jurkat T cells using the RNeasy mini extraction kit (Qiagen # 74106) according to the manufactures’ protocol. RNA was dissolved and diluted in RNase free water and the RNA concentration was determined via Nanodrop analysis. The quality of total RNA was evaluated by capillary electrophoresis using an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, Calif.)
Double-stranded cDNA was synthesized from 1.5 µg total RNA using the One-Cycle Target Labeling Kit (Affymetrix Santa Clara, CA), and used as a template for the preparation of biotin-labeled cRNA using the GeneChip IVT Labeling Kit (Affymetrix Santa Clara, CA). Biotin-labeled cRNA was fragmented at 1 µg/l following the manufacturer’s protocol. After fragmentation, cRNA (10 µg) was hybridized at 45°C for 16-17 hours to the Human Genome U133A 2.0 Array or the Human Genome U133 Plus 2.0 Array (Affymetrix, Santa
Chapter 7

Clara, CA). Following hybridization, the arrays were washed, stained with phycoerythrin-streptavidin conjugate (Molecular Probes, Eugene, OR), and the signals were amplified by staining the array with biotin-labeled anti-streptavidin antibody (Vector Laboratories, Burlingame, CA) followed by phycoerythrin-streptavidin. The arrays were laser scanned with a GeneChip Scanner 3000 6 G (Affymetrix, Santa Clara, CA) according to the manufacturer’s instructions. Data was saved as raw image file and quantified using GCOS (Affymetrix).

**Statistical analysis**

The .CEL files were analyzed with the R http://www.rproject.org and the BioConductor software package http://www.bioconductor.org. Normalization was done using gcrma. Building of the experimental design and calculation of the ratios was done with the limma package. Regulated probe sets were selected on basis of the fold change and the adjusted p-value (Benjamini-Hochberg correction). Multivariate data analysis and clustering was done with standard methods in the R software package http://www.r-project.org. For the principal component analysis and hierarchical clustering, ratio data were used. The ratio data were calculated for each treatment to its corresponding control. For the treatment with the stimuli, the untreated cells were taken as a control. For the treatment with stimulus + compound combinations, the treatment with the stimulus alone was taken as a control. Results were expressed as mean ± SEM. Significance of differences was determined using a one-way ANOVA followed by post-hoc testing as indicated. Data sets can be found in GEO http://www.ncbi.nlm.nih.gov/gds/ under accession number GSE30678.

**FLIPR calcium flux assay**

96-wells plates were coated with poly-L-lysine in PBS for 1 h at 37°C. Jurkat T cells were seeded at a concentration of 7 × 10^5 in culture medium and rested for 1 hour at 37°C/5%CO2. Thereafter cells were incubated for 1 hour in the dark with FLIPR calcium buffer, according to the manufacturer’s protocol. Stimuli were added via the Flex station384 and calcium release was monitored in time (Molecular devices, Sunnyvale, USA).

**Western blotting and nuclear translocation assay**

Cells were washed in ice-cold PBS and pellets were lysed on ice in lysis buffer (Biosource, cat nr FNN0011, supplemented with 1 × protease inhibitor cocktail Roche cat no 11873580051, 1 mM PMSF Fluka cat no 93482, phosphatase inhibitor cocktail I, Sigma cat no P2850, phosphatase inhibitor cocktail II, Sigma cat no P5726) followed by an incubation for 30 min on ice. The lysates were stored at -80°C until further analysis. Phosphorylation of proteins from stimulated Jurkat cells were evaluated via western blot analysis. Briefly, samples were run on a 4-12% NuPage gels (#NP321BOX, Invitrogen) for 35 min on 200 V in 1 × MES buffer (#NP0002, Invitrogen, USA) and subsequently transferred
to a PVDF membrane (#162-0184, Biorad). The blots were blocked in PBS/0.05% Tween-20 with 1% skim milk (#232100, Difco) and 1% BSA. Blots were incubated O/N at 4°C in a roller bottle with the primary antibody diluted 1:1000 in block buffer, followed by incubation with a secondary detection antibody. Thereafter blots were incubated in ECL (#RPN2106V1 and RPN2106V2 Amersham Pharmacia) and hyperfilms (#RPN3103K, Amersham Pharmacia) were exposed and developed. For the detection the following antibodies were used: pLck/Src (Tyr416 Cat N° 2101 L), Lck (Cat N° 2752), pZAP70 (Tyr192 Cat N° 2701), pPKCθ (Thr538, Cat N° 9377), pMARCKS (Ser152/156 Cat N° 4273) and ATF-2 (Thr71 Cat N° 9221) (all cell signaling technology, Danvers, USA). pP38 (Thr180/Tyr182 Cat N° 44-684 G), pERK (Thr185/Tyr187 Cat N° 44-680 G), pJNK1/2 (Thr183/Tyr185 Cat N° 44-682 G), JNK1 (Cat N° 44-690 G) were obtained from Invitrogen (Carlsbad, USA) and c-Jun (pSer73) was obtained from calbiochem (cat no 420114). For the analysis of nuclear translocation of the transcription factors NFAT, NFkBp65 and c-JUN, nuclear fractions of activated Jurkat T cells were isolated via hypotonic shock and levels of activated transcription factors in the nuclear lysates was tested in a TransAM transcription factor ELISA according to the manufacturers’ protocol (Active Motif, Carlsbad, USA).

**Knock down of PKCθ and lck in Jurkat T cells**

Jurkat T cells (2 × 10e7 cells/ml) were mock transfected or electroporated (250 V/975 μF) with siRNA targeting Lck (RefSeq accession-number NM-005356; sense 5’-GCACGCUGCUCAUCCGAAAdTdT) and PKCθ (RefSeq accession-number NM-006257; sense 5’-GCAGCAAUUUCGACAAAGAdTdT) in different concentrations of 500, 100 and 20 nM or scrambled control siRNA (Thermo scientific, Dharmacon Inc, Lafayette, USA). Electroporated cells were cultured in M505 supplemented with 10% FCS. Cells were stimulated 72 hours after electroporation and knockdown efficiency of the specified proteins was checked via western blot analysis. Culture supernatants of PMA/CD28 or PMA/CD3-stimulated cells were collected after 24 hours and production of respectively CCL1 and IL-2 was determined.

**Experiments with WBA**

Peripheral whole blood was obtained by venipuncture from healthy adults (male/female) and was collected into lithium heparinised tubes. Blood was obtained from healthy volunteers and the time between puncture and processing was less than 1 hour. Blood was diluted 1:4 with RPMI 1640 (Life Technologies, cat no. 32404-014) supplemented with penicillin/streptomycin (GibcoBRL, cat 15140-122) and 2 mM L-glutamine (GibcoBRL, cat 25030-024) and distributed (200 μl/well) into 96-wells plates (Nunc, Cat 167008). Blood cultures were stimulated with soluble αCD3/PMA and soluble αCD28/PMA, or left unstimulated. Blood (200 μl/well) in RPMI 1640 medium was incubated with 25 μl compound (maximum of 0.1% v/v DMSO).
Cytokine determination
Cytokines and chemokines IL-2, CCL1/I309 and XCL1, secreted into the supernatant of stimulated Jurkat T cells were determined via ELISA (R&D systems, USA). Cytokines IL-17, IFNγ, IL-13 and IL-5, produced by human CD4+ T-cells activated in whole blood, cultured in the presence or absence of compounds were determined in the culture supernatant, using a bead-based human cytokine multiplex kit (Bioplex-system; Bio-Rad, Veenendaal, The Netherlands) according to the manufacturer’s instructions. Culture supernatants were collected at day 1 of culture. Samples were analyzed using a Luminex-100 analyzer (Luminex, Austin, USA) with Bio-plex Manager Software 3.0 (Bio-Rad). Proteins were discriminated based on the fluorescent label of the bead and the PE levels were corrected for background levels of negative controls. The sensitivity of the cytokine assay was less than 5 pg/ml for all cytokines measured.

cDNA synthesis and q-PCR
Primary human CD4+ T cells were isolated from buffy coats from three healthy donors using MACS negative CD4+ purification technology (Miltenyi biotech, Germany), yielding a overall 96% pure CD4+ T cell population. RNA from stimulated CD4+ T cells was isolated using an RNeasy minikit (QIAGEN Gmbh, Germany). RNA content of samples was analyzed using Nanodrop (Agilent technologies, Ca, USA) and purity was analyzed using the Agilent RNA 6000 nanokit protocol on the RNA nano labchip using the Agilent 2100 bioanalyzer (Agilent technologies, Waldborn, Germany). Three microgram of RNA was used for cDNA synthesis using random hexamer primer mix (invitrogen), 10 mM dNTP, M-MLV RT buffer and M-MLV Reverse transcriptase (Promega). RT reaction was performed at 42°C for 1 hour followed by a deactivation for 5 minutes at 90°C. cDNA of the Th1 master transcription factor Tbx21 (Tbet), Th2 transcription factor GATA3 or the control household gene RPL19 was amplified using Power SYBR green mastermix (Applied Biosystems, Warrington, UK) and expression was monitored on the ABI prism 7900 HT sequence detection system (Applied Biosystems, Warrington, UK), Ct values were normalized for the expression of the RPL19 gene.

Additional material can be found at the online version of this paper at the publishers website or can be retrieved on request by the author of this thesis.

Abbreviations
AP1: Activator protein 1; APC: Antigen presenting cell; ATF: Activating transcription factor; CCL1: CC chemokine ligand 1; Cn: Calcineurin; CsA: Cyclosporin A; CTL: Cytotoxic T cell; CTLA4: Cytotoxic lymphocyte antigen 4; ERK: Extracellular-signal-regulated kinase; FASL: TNFSF6 Fas ligand; GATA3: T cell specific transcription factor binding to DNA sequence GATA; IFNγ: Interferon gamma; IL: Interleukin; ITAM: Immuno-tyrosine based activation motif; Itk: IL-2-inducible T cell kinase; JNK: c-JUN N-terminal Kinase; Lck: p56
Lymphocyte-specific protein tyrosine kinase; MAPK: Mitogen activated protein kinase; MARCKS: Myristoylated alanine-rich C-kinase substrate; MHC: Major histocompatibility complex; NFAT: Nuclear factor of activated T cells; NFκB: Nuclear factor kappa B; OX40: CD134; P38: p38 Mitogen-activated protein kinase; PCA: Principle component analysis; PDK1: 3-phosphoinositide dependent protein kinase-1; PI3K: Phosphatidylinositol-3-kinase; PIP: Phosphatidylinositolphosphate; PLCγ: Phospholipase C gamma; PKB/AKT: Protein kinase B; PKCθ: Protein kinase C theta; PMA: Phorbol 12-myristate 13-acetate; RUNX3: Runt-related transcription factor 3; RXRA: Retinoid X Receptor alpha; T-bet: T-box transcription factor 21 (TBX21); TCR: T cell receptor; Th: T helper cell; TNF: Tumor necrosis factor; ZAP70: Zeta-chainassociated protein 70.

Acknowledgements
The authors would like to thank Hans van de Maaden for performing the PKCθ and Lck knock down experiments and Roselinde van Ravestein-van Os, Roger van de Wetering (Merck Research Laboratories, Oss, the Netherlands) for their contribution to the microarray analyses. WF was supported by a grant from NBIC, the Netherlands.

Authors’ contributions
Contribution: RLS, WA, SB, PV, XH, JPMK, IJ, AG and AMHB designed research; RLS, FW, MG, HK, and WA performed research; RLS, WF, WA, XH, JPMK, IJ, SB, AMHB analyzed data; and RLS, WF, WA, SB, XH, JPMK, IJ, and AMHB, wrote the paper. All authors read and approved the final manuscript.

Competing interests
The authors of this manuscript have no conflicts of interest to disclose as described by the Journal. The work performed here was funded by NV Organon, now part of MSD, Merck.
References


48. Borhani DW, Calderwood DJ, Friedman MM, Hirst GC, Li B, Leung AK, McRae B, Ratnofsky S, Ritter K,


CHAPTER 8
Functional annotation of microorganisms using text mining

Wilco W.M. Fleuren¹, Marijn Sanders², Jos Boekhorst³,⁴, Jacob de Vlieg¹,² and Wynand Alkema¹,⁴.

¹ CDD, CMBI, NCMLS, Radboud University Medical Centre, Nijmegen, The Netherlands.
² Netherlands eScience Center, Amsterdam, The Netherlands.
³ Bacterial genomics, CMBI, NCMLS, Radboud University Medical Centre, Nijmegen, The Netherlands.
⁴ NIZO Food Research BV, Ede, The Netherlands.

In preparation
Abstract

Background
Metagenomics by means of 16S rRNA sequencing allows identifying microbial communities that are directly retrieved from environmental samples. This technique is used in so-called community profiling studies and results in lists of organisms that are affected by a certain treatment or process. Knowledge of the biological properties of these organisms is important to understand their role in these processes. This knowledge is for a large part contained in the scientific literature. Because these organism lists contain dozens of organisms, retrieving the biological properties from the literature is labor-intensive and therefore requires methods that can do this automatically.

Method
To this end we developed a method which is able to automatically retrieve biological properties of microorganisms from Medline abstracts.

Results
We analyzed lists of microorganisms derived from metagenomics experiments that were aimed at the characterization of microbiota in female urine and the characterization of the oral fungal microbiome in healthy individuals. With the organisms from both lists, we searched for associations with diseases in Medline abstracts. Several specific disease associations were found indicating the pathogenic potential of these microorganisms. Furthermore, we automatically identified biological properties and industrial applications for groups of lactic acid bacteria (LAB). We identified a cluster of LAB with opposite roles in the process of wine making and a cluster of LAB that are involved in food spoilage as well as in fermentation processes to make sauerkraut and kimchi.

Conclusions
Automatic information extraction on microorganisms gives rapid insight into the pathogenic potential of microorganisms by linking to diseases in Medline abstracts. Furthermore word clouds of highly occurring keywords in these abstracts give an overview of the biological properties of LAB and the processes in which they play a role. This overview can help to select LAB for fermentation processes in order to improve the characteristics of fermented products.

Keywords: Microorganism annotation, text mining, community profiling, metagenomics data, lactic acid bacteria.
Background

Microorganisms play crucial roles in a number of biological processes ranging from the decomposition of waste products [1, 2], to industrial processes where they are used for the development of fermented foods and beverages [3-5]. Furthermore, pathogenic microorganisms such as *Mycobacterium tuberculosis*, *Salmonella* and *Clostridium tetani* can be harmful, causing tuberculosis, food poisoning and tetanus respectively.

In metagenomics studies, microbial DNA is directly extracted from environmental samples and sequenced by means of 16S RNA sequencing. This technique is well established to study the dynamics of microbial communities. One of the outcomes of these so-called community profiling studies is a list of microorganisms that are affected by a given process or treatment. Insight into the biological properties of these microorganisms is essential to better understand how they are involved in these processes.

Because these organism lists contain dozens of microorganisms, retrieval of biological properties these organisms from the scientific literature is labor-intensive and requires methods that can do this retrieval automatically.

Publications about microorganisms are growing exponentially [6] and literature databases such as the Medline database have over 1.5 million abstracts related to microbiota and microorganisms (Table 1).

<table>
<thead>
<tr>
<th>Search term</th>
<th>Number of hits found</th>
</tr>
</thead>
<tbody>
<tr>
<td>microbiota OR microbiome OR microflora OR bacteria OR microorganisms</td>
<td>1581863</td>
</tr>
<tr>
<td><em>Salmonella</em></td>
<td>72488</td>
</tr>
<tr>
<td><em>Mycobacterium tuberculosis</em></td>
<td>42774</td>
</tr>
<tr>
<td><em>lactococcus lactis</em></td>
<td>4299</td>
</tr>
<tr>
<td><em>Stretococcus thermophilus</em></td>
<td>1135</td>
</tr>
<tr>
<td><em>Clostridium tetani</em></td>
<td>995</td>
</tr>
<tr>
<td><em>Oenococcus oeni</em></td>
<td>293</td>
</tr>
</tbody>
</table>

Table 1 Number of abstracts found for each term when searching in Pubmed (till December 2012).

Depending on the microorganism, the number of abstracts range from a few hundred to tens of thousands (Table 1). This indicates that the Medline database is a suitable source for the retrieval of biological properties of microorganisms.

In earlier work we reported about CoPub [7], a publicly available text mining system that has successfully been used for the analysis of gene expression data by automatically searching for co-occurrences between genes and pathways, biological processes, diseases and drug terms in Medline abstracts [8-11].
Here we report the development of a tool that identifies organism names in any kind of text and maps these to the NCBI taxonomy ids.

With this tool, analogous to the gene list based text mining described for CoPub, lists of microorganisms can be annotated with important biological properties of these microorganisms and the processes in which they are involved.

Subsequently, this text is analyzed in two ways:

One analysis method consists of searching for significant occurrences of disease and pathway terms from thesauri in Medline abstracts. Additionally, highly occurring informative keywords are identified in these abstracts and presented as a word cloud.

We show that with both methods we can annotate lists of microorganisms from community profiling studies and find evidence about the pathogenic potential of these microorganisms, and predict biological functions for single and clusters of lactic acid bacteria.

Thus text mining is a valuable approach to quickly retrieve biological properties of microorganisms from text and can help experts to gain insight into how these microorganisms are involved in the studied processes.

**Materials and Methods**

**Detection of organism names in Medline abstracts**

A set of regular expressions was implemented in Perl and used to tag organism names in Medline abstracts. Subsequently, for organisms with a known taxonomy identifier in the NCBI Taxonomy Database [12], the hits in the Medline abstracts were stored in a database. This resulted in one to multiple organism hits per Medline abstract.

**Calculation of significant co-occurrences**

The significance of a co-occurrence between organisms or for a co-occurrence between organisms and disease terms from controlled vocabulary [7] was determined by calculating a p-value using the R implementation of Fisher’s exact test. We considered co-occurrences with a p-value < 0.05 significant. Co-occurrence matrices were clustered with the R package pvclust with “complete” setting for hierarchical clustering. For further analysis sub clusters were created (tree cut-off: 0.9).

**Selection of organism lists from community profiling studies**

Bacteria identified in female urine were selected using Table S1 from the paper by Siddiqui et al. [13]. Fungi identified in oral rinse samples were selected from Table 3 in combination with fungi from Supplementary table 3 for which the genera were presented in Figure 3 in the paper by Ghannoum et al. [14].
Generation of word clouds of highly occurring keywords
The set of abstracts for a given organism (organism_abstracts) was used to generate word clouds for the organism of interest. Keywords with less than 5 characters and keywords occurring in < 2% of the abstracts were discarded. The significance of occurrence of a keyword in the organism_abstracts was determined by calculating the relative keyword frequency for the keyword in organism_abstracts and the relative keyword frequency of the keyword in a background set. The ratio of these relative frequencies was used to rank the keywords. Wordclouds were generated using the R-package “wordcloud”.

Results

Text mining vs. manual searching in Pubmed
In order to retrieve information about microorganisms from Medline abstracts we automatically tagged 3502122 unique organism names in these abstracts. We were interested to see whether tagging of organism names, called tagging hits, yielded good results in comparison with when searching for these organism names in Pubmed, referred to as Pubmed hits.

For 100 organisms we compared the number of tagging hits with the number of Pubmed hits. The results are shown in Figure 1.

For about 40 % of the organisms the number of tagging hits was similar to the number of Pubmed hits. Differences between both search strategies can be explained by that fact there is a discrepancy of about 5% between the Medline database in Pubmed and the Medline database as provided to licensees for download, in which the tagging hits were found. Moreover, Pubmed hits were also found in Mesh Terms (Medical Subject Headings from controlled vocabulary thesaurus used for indexing articles for Pubmed) of abstracts, this in contrast to the tagging hits for which we only searched in the title and the abstract. For example for Clostridium tetani several tagging hits [15-19] were not retrieved, because the organism name is only used in the Mesh Term header.

For 35 out of the 100 organisms more tagging hits were retrieved in comparison with Pubmed hits. This difference can be explained by the fact that tagging hits were found by searching with all names and synonyms of the particular organism, while Pubmed hits were found by searching only with one name or synonym. For example in the case of Oenococcus oeni, tagging hits also included abstracts in which Leuconostoc oenos (the previous name of Oenococcus oeni), was found [20, 21]. Similarly for Lactococcus lactis subsp. cremoris additional tagging hits were found with its synonym Streptococcus cremoris [22, 23].

On the other hand for Leuconostoc, Moraxella, Listeria, Neisseria and Salmonella more Pubmed hits were retrieved. All these organisms are classified as genus in the NCBI taxonomy database and the Pubmed hits included also abstracts about organisms that belong to this genus, e.g. Leuconostoc mesenteroides. Tagging hits for Leuconostoc
resulted in abstracts about *Leuconostoc* and not in abstracts about organisms that belong to this genus. These latter abstracts were mapped to these specific organisms. Taken together, these results show that that tagging of organism names yielded good results in comparison with searching for these organism names in Pubmed and outperforms Pubmed searches for certain microorganisms.

![Graph](image.png)

**Figure 1 Retrieval of the number of abstracts for 100 microorganisms.** The x-axis represents the number of Pubmed hits, the y-axis represents the number of tagging hits. Ten percent of the microorganisms for which the most tagging hits were found are shown in green. Ten percent of the microorganisms for which the most Pubmed hits were found are shown in red. Tagging of organism names was done on Medline abstracts till October 2012. Pubmed queries were done on abstracts till the end of November 2012.

**Annotation of metagenomics experiments**

Metagenomics and community profiling studies often result in lists of microorganisms. To see if we could automatically find biologically meaningful descriptions in Medline abstracts, we used tagging to retrieve these descriptions for two sets of microorganisms: one from a study aimed at the characterization of microbiota in urine from healthy females [13] and one from a study aimed at the characterization of the oral fungal microbiome in healthy individuals [14]. For both studies we were interested to see if we could find evidence in the literature about the pathogenic potential of these microorganisms.

**Annotation of organisms derived from female urine samples**

Although urine within the urinary tract is generally considered as “sterile”, several
Table 2 Associations between organisms from urine samples of healthy volunteers and disease terms found in the literature. NA indicates no association was found.

<table>
<thead>
<tr>
<th>Tax ID</th>
<th>Organism</th>
<th>Disease term (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>165188</td>
<td>uncultured <em>Megasphaera</em> sp.</td>
<td>NA</td>
</tr>
<tr>
<td>28130</td>
<td><em>Prevotella disiens</em></td>
<td>pelvic inflammatory disease (3.12e-11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cancer (2.38e-06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cellulitis (0.000152)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>abscess (0.000311)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Endometritis (0.00066)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>furunculosis (0.0112)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>carcinoma (0.0178)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>salpingitis (0.019)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>bacterial infection (0.0419)</td>
</tr>
<tr>
<td>1376</td>
<td><em>Aerococcus urinae</em></td>
<td>urinary tract infection (5.16e-26)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>balanitis (3.17e-05)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>prostatic disease (0.0152)</td>
</tr>
<tr>
<td>361505</td>
<td><em>Anaerococcus sp. gpac155</em></td>
<td>NA</td>
</tr>
<tr>
<td>119206</td>
<td><em>Aerococcus sanguinicola</em></td>
<td>urinary tract infection (3.36e-06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>prostatic disease (0.00147)</td>
</tr>
<tr>
<td>322095</td>
<td><em>Porphyromonas somerae</em></td>
<td>NA</td>
</tr>
<tr>
<td>159274</td>
<td>uncultured <em>Porphyromonas</em> sp.</td>
<td>NA</td>
</tr>
<tr>
<td>844</td>
<td><em>Wolinella succinogenes</em></td>
<td>NA</td>
</tr>
<tr>
<td>174293</td>
<td>uncultured candidate division OP11 bacterium</td>
<td>NA</td>
</tr>
<tr>
<td>189716</td>
<td><em>Flexistipes sp. E3_33</em></td>
<td>NA</td>
</tr>
<tr>
<td>33043</td>
<td><em>Coprococcus eutactus</em></td>
<td>irritable bowel syndrome (0.00404)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inflammatory bowel disease (0.0124)</td>
</tr>
<tr>
<td>172733</td>
<td>uncultured <em>Clostridiales</em> bacterium</td>
<td>NA</td>
</tr>
<tr>
<td>38303</td>
<td><em>Corynebacterium pseudogenitalium</em></td>
<td>urinary tract infection (0.00257)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>corneal ulcer (0.00683)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Endometritis (0.00783)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>urethritis (0.0121)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nosocomial infections (0.0274)</td>
</tr>
<tr>
<td>208548</td>
<td>uncultured <em>Neisseriaceae</em> bacterium</td>
<td>NA</td>
</tr>
<tr>
<td>361501</td>
<td><em>Peptoniphilus sp. gpac121</em></td>
<td>NA</td>
</tr>
<tr>
<td>47770</td>
<td><em>Lactobacillus crispatus</em></td>
<td>urinary tract infection (0.000916)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gonorrhea (0.00145)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>trichomonas infection (0.00598)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inflammation (0.0378)</td>
</tr>
<tr>
<td>Tax ID</td>
<td>Organism</td>
<td>Disease term (p-value)</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------</td>
<td>---------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>221218</td>
<td>uncultured candidate division OD1 bacterium</td>
<td>NA</td>
</tr>
<tr>
<td>202789</td>
<td>Actinobaculum massiliense</td>
<td>urinary tract infection (0.000289)</td>
</tr>
<tr>
<td>76517</td>
<td>Campylobacter hominis</td>
<td>Gastrointestinal disease (0.00578)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>disseminated intravascular coagulation (0.00912)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gastroenteritis (0.0135)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>periodontal disease (0.0164)</td>
</tr>
<tr>
<td>195049</td>
<td>uncultured Clostridaceae bacterium</td>
<td>NA</td>
</tr>
<tr>
<td>163550</td>
<td>Eubacterium sp. oral clone BU061</td>
<td>NA</td>
</tr>
<tr>
<td>179000</td>
<td>uncultured Methylophilus sp.</td>
<td>NA</td>
</tr>
<tr>
<td>54007</td>
<td>Anaerococcus octavius</td>
<td>NA</td>
</tr>
<tr>
<td>827</td>
<td>Bacteroides ureolyticus</td>
<td>urethrits (5.13e-31)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>gastroenteritis (0.000701)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>abscess (0.00217)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>brain abscess (0.00444)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gastrointestinal disease (0.0125)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>periodontal disease (0.0127)</td>
</tr>
<tr>
<td>158755</td>
<td>uncultured Cytophagales bacterium</td>
<td>NA</td>
</tr>
<tr>
<td>178214</td>
<td>Facklamia hominis</td>
<td>NA</td>
</tr>
<tr>
<td>87541</td>
<td>Aerococcus christensenii</td>
<td>urinary tract infection (0.0276)</td>
</tr>
<tr>
<td>181675</td>
<td>Lactobacillus coleohominis</td>
<td>Crohn's disease (0.0393)</td>
</tr>
<tr>
<td>339338</td>
<td>uncultured Allisonella sp.</td>
<td>NA</td>
</tr>
<tr>
<td>187319</td>
<td>uncultured Fibrobacteres bacterium</td>
<td>NA</td>
</tr>
<tr>
<td>33012</td>
<td>Propionimicrobium lymphophilum</td>
<td>NA</td>
</tr>
<tr>
<td>359408</td>
<td>Methylocutera mobilis</td>
<td>NA</td>
</tr>
<tr>
<td>254354</td>
<td>uncultured Peptoniphilus sp.</td>
<td>NA</td>
</tr>
<tr>
<td>166587</td>
<td>uncultured Chloroflexi bacterium</td>
<td>NA</td>
</tr>
<tr>
<td>159272</td>
<td>uncultured Prevotella sp.</td>
<td>NA</td>
</tr>
<tr>
<td>159268</td>
<td>uncultured Veillonella sp.</td>
<td>NA</td>
</tr>
<tr>
<td>293422</td>
<td>uncultured Eggerthella sp.</td>
<td>NA</td>
</tr>
<tr>
<td>853</td>
<td>Faecalibacterium prausnitzii</td>
<td>Crohn's disease (1.68e-21)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inflammatory bowel disease (1.09e-10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cancer (1.69e-10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>colitis ulcerative (6.08e-08)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>irritable bowel syndrome (0.0205)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>colon cancer (0.045)</td>
</tr>
<tr>
<td>28128</td>
<td>Prevotella corporis</td>
<td>NA</td>
</tr>
<tr>
<td>1260</td>
<td>Finegoldia magna</td>
<td>Acne (4.03e-09)</td>
</tr>
</tbody>
</table>
microorganisms with a pathogenic potential were found in these samples [13]. In order to see if we could find more specific evidence that further could underline these observations; we searched for associations between the organisms and disease terms in Medline abstracts (Table 2). We found, similar to the observations and manual annotations by Siddiqui et al., disease terms related to the infection/inflammation of the urinary tract, e.g. urinary tract infection [24-26] and urethritis (inflammation of the urethra) [27, 28] and several diseases associated with female urogenital pathology, e.g. salpingitis (infection and inflammation of fallopian tubes)[29], endometritis (infection of the inner lining of the uterus) [29-31] and pelvic inflammatory disease (inflammation of the uterus, fallopian tubes, and/or ovaries) [29, 32]. In addition we found associations with intestine-related diseases, such as inflammatory bowel disease [33, 34], Crohn’s disease [35, 36], irritable bowel disease [37], gastroenteritis [38] and necrotizing fasciitis [39, 40]. These results show that tagging in combination with thesaurus based matching is as effective as manual annotation.

**Annotation of the fungal microbiome in healthy individuals**

Next we were interested to see if we also could find evidence about the pathogenic potential of fungi identified in oral rinse samples from healthy individuals. We searched for associations between these fungi and diseases (Table 3). A number of diseases related to the oral cavity and respiratory system were found such as candidiasis [41-43], aspergillosis [44-47], asthma [48, 49], rhinitis [50, 51] and respiratory tract disease [52]. Several of these pathogenic fungi are also able to infect other parts of the body such as the skin [53] and the vagina [54, 55].
The risk of getting infected by these pathogenic fungi especially increases in individuals with a compromised immune system, for instance in cancer, leukemia and diabetes mellitus patients [56-58]. Typically these patients stay in a hospital environment which also increases the risk on nosocomial infections [59].

The presence of potential pathogenic fungi in healthy individuals could be the first step towards infection by them as indicating by the authors. We demonstrated that tagging in combination with thesaurus based matching can be used to get insight into specific diseases that could be caused by these fungi.

**Table 3 Associations between fungi from orinse samples of healthy volunteers and disease terms found in the literature. NA indicates no association was found.**

<table>
<thead>
<tr>
<th>Tax ID</th>
<th>Organism</th>
<th>Disease term (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>41959</td>
<td><em>aspergillus penicillioides</em></td>
<td>NA</td>
</tr>
<tr>
<td>5516</td>
<td><em>fusarium culmorum</em></td>
<td>NA</td>
</tr>
<tr>
<td>5507</td>
<td><em>fusarium oxysporum</em></td>
<td>cancer (5.24e-235)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>leukemia (3.29e-17)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>keratitis (3.4e-13)</td>
</tr>
<tr>
<td>41960</td>
<td><em>aspergillus caesiellus</em></td>
<td>NA</td>
</tr>
<tr>
<td>5074</td>
<td><em>penicillium brevicompactum</em></td>
<td>NA</td>
</tr>
<tr>
<td>92950</td>
<td><em>cladosporium sphaerospermum</em></td>
<td>carcinoma (0.0189)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Respiratory tract disease (0.0316)</td>
</tr>
<tr>
<td>119927</td>
<td><em>alternaria tenuissima</em></td>
<td>Iga deficiency (0.0166)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>retroperitoneal fibrosis (0.0213)</td>
</tr>
<tr>
<td>4897</td>
<td><em>schizosaccharomyces japonicus</em></td>
<td>NA</td>
</tr>
<tr>
<td>4956</td>
<td><em>zygosaccharomyces rouxii</em></td>
<td>NA</td>
</tr>
<tr>
<td>279121</td>
<td><em>ophiostoma pulvinisporum</em></td>
<td>NA</td>
</tr>
<tr>
<td>4896</td>
<td><em>schizosaccharomyces pombe</em></td>
<td>cancer (0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>starvation (6.13e-79)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hypovolemic shock (5.47e-14)</td>
</tr>
<tr>
<td>4932</td>
<td><em>saccharomyces cerevisiae</em></td>
<td>cancer (0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>starvation (1.26e-243)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hypovolemic shock (9.17e-66)</td>
</tr>
<tr>
<td>415599</td>
<td><em>glomus fulvum</em></td>
<td>NA</td>
</tr>
<tr>
<td>5480</td>
<td><em>candida parapsilosis</em></td>
<td>candidiasis (9e-305)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cancer (7.29e-231)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mycoses (4.9e-39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>onychomycosis (4.73e-36)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nosocomial infections (7.65e-21)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fever (3.86e-13)</td>
</tr>
<tr>
<td>Tax ID</td>
<td>Organism</td>
<td>Disease term (p-value)</td>
</tr>
<tr>
<td>-------</td>
<td>---------------------------------</td>
<td>-----------------------------------------------------</td>
</tr>
<tr>
<td>4931</td>
<td><em>saccharomyces bayanus</em></td>
<td>NA</td>
</tr>
<tr>
<td>69773</td>
<td><em>penicillium glabrum</em></td>
<td>hypersensitivity (6.88e-05)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pneumonia (0.00391)</td>
</tr>
<tr>
<td>29917</td>
<td><em>cladosporium cladosporioides</em></td>
<td>sick building syndrome (3.66e-08)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>keratitis (0.00582)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>allergy (0.00625)</td>
</tr>
<tr>
<td>5580</td>
<td><em>aureobasidium pullulans</em></td>
<td>corneal ulcer (0.000127)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fever (0.016)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>asthma (0.0304)</td>
</tr>
<tr>
<td>5062</td>
<td><em>aspergillus oryzae</em></td>
<td>Hypovolemic shock (3.27e-09)</td>
</tr>
<tr>
<td>70808</td>
<td><em>cladosporium tenuissimum</em></td>
<td>NA</td>
</tr>
<tr>
<td>396024</td>
<td><em>aspergillus ruber</em></td>
<td>hypersensitivity immediate (0.00323)</td>
</tr>
<tr>
<td>216836</td>
<td><em>alternaria triticina</em></td>
<td>NA</td>
</tr>
<tr>
<td>749552</td>
<td><em>phoma foveata</em></td>
<td>NA</td>
</tr>
<tr>
<td>62271</td>
<td><em>saccharomyces ellipsoideus</em></td>
<td>NA</td>
</tr>
<tr>
<td>5059</td>
<td><em>aspergillus flavus</em></td>
<td>cancer (2.46e-066)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>aspergillosis (5.85e-222)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>keratitis (1.79e-12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>leukemia (7.53e-10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sinusitis (1.05e-07)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>pneumonia (1.69e-05)</td>
</tr>
<tr>
<td>89919</td>
<td><em>cryptococcus diffluens</em></td>
<td>dermatitis (2.34e-06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cryptococcosis (0.0102)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>candidiasis (0.033)</td>
</tr>
<tr>
<td>227369</td>
<td><em>zygosaccharomyces pseudorouxii</em></td>
<td>NA</td>
</tr>
<tr>
<td>260535</td>
<td><em>candida khmerensis</em></td>
<td>NA</td>
</tr>
<tr>
<td>89918</td>
<td><em>cryptococcus cellulolyticus</em></td>
<td>NA</td>
</tr>
<tr>
<td>104300</td>
<td><em>ophiostoma floccosum</em></td>
<td>NA</td>
</tr>
<tr>
<td>29918</td>
<td><em>cladosporium herbarum</em></td>
<td>allergy (5.7e-46)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>asthma (2.29e-20)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rhinitis (2.94e-12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>eczema (1.49e-07)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>sick building syndrome (3.75e-06)</td>
</tr>
</tbody>
</table>
Finding relevant information for lactic acid bacteria

We showed that automatically linking microorganisms to keywords in thesauri is as effective as manual annotation. However a lot of the properties of microorganisms are described in abstracts with keywords that are not present in thesauri. We were interested to see if we could automatically retrieve these keywords. Furthermore we evaluated whether presenting these keywords in word clouds provide sufficient information about the biological properties of microorganisms.

For *Streptococcus thermophilus*, a widely used lactic acid bacterium in the dairy industry, we identified 976 abstracts. We created a word cloud of highly occurring keywords in these abstracts (Figure 2).

The word cloud reveals the role of *Streptococcus thermophilus* in fermentation processes where it is used as starter culture for the preparation of yoghurt [60, 61]. In cheese, *Streptococcus thermophilus* is mainly responsible for the acidification which is important for the ripening and the characteristics of cheese [62, 63]. Additionally, some *Streptococcus*
thermophilus strains produce exopolysaccharides (ESP) which is an important ingredient controlling the texture of yogurt and cheese and has been suggested to be beneficial for the treatment of chronic gastritis [64].

These results show that the word cloud, constructed of keywords retrieved from abstracts about Streptococcus thermophilus give an overview of the the industrial processes and applications in which Streptococcus thermophilus is involved. Besides keywords that give insight into the application and the biological properties of Streptococcus thermophilus, the word cloud also contains (parts of) names from other bacteria, e.g. bifidobacterium, Lactobacillus acidophilus and Lactobacillus delbrueckii subsp. bulgaricus are shown. This indicates that Streptococcus thermophilus is functionally related to these LAB in fermentation processes.

**Microbial community literature study of LAB**

To better understand in which ways these LAB are related to each other and how they together are involved in biological processes we searched in the literature for evidence about which LAB are often mentioned together and in which processes and applications they are mentioned together. In order to do this, we started with Lactobacillus plantarum, Streptococcus thermophilus and Oenococcus oeni and retrieved the organisms with which they co-occur. To see if taxonomy differences are also reflected in the scientific literature we also included non-LAB Bacillus subtilis. For each of these four bacteria
we obtained a list of the most significant co-occurring organisms (p-value < 0.05). For *Bacillus subtilis* we found the highest number of co-occurring organisms, 328 organisms in 8963 abstracts, and for *Oenococcus Oeni* we found the lowest number of co-occurring organisms, 17 organisms in 133 abstracts. Subsequently, for each organism list we ranked the organisms based on the p-value and retrieved from each organism list the top ranked co-occurring organisms. This yielded a unique set of 51 organisms (organisms classified as ‘superkingdom’ or ‘genus’ were discarded). To find out how these 51 organisms are related to each other, we determined whether they co-occur in Medline abstracts. Based on these co-occurrences, the organisms were clustered (Figure 3). It appeared that *Bacillus subtilis*, which is not classified as a LAB, clusters separately from the other LAB. This indicates that the taxonomy difference between *Bacillus subtilis* and LAB is also reflected by the literature context. Furthermore *Lactobacillus plantarum*, *Streptococcus thermophilus* and *Oenococcus oeni*, with which we initially started our search with, cluster in separate, sub-clusters all of them with similar LAB.

Figure 3 Hierarchical tree of LAB related microorganisms. For each cluster (designated with a character), a word cloud of highly occurring terms in abstracts (indicated with a corresponding character) in which organisms in the cluster are mentioned, is made.
To see how LAB that clustered together are functionally related to each other, we created word clouds based on the abstracts connected to these clusters.

*Determine the biological properties for groups of LAB*

To retrieve the processes and applications for LAB that clustered together, we first divided the tree in Figure 3, into 9 clusters. For each cluster we retrieved the abstracts in which the microorganisms from the cluster are mentioned. This yielded between 14582 and 46 abstracts per cluster. We used these abstracts to generate a word cloud of highly occurring keywords (Figure 3). Cluster i in Figure 3 consists of microorganisms that are involved in the wine making process, but with different functions. *Oenococcus oeni* is a malolactic starter for the wine making process, giving the wine its characteristic taste and aroma [65]. However, glucan-producing strains of *Pediococcus damnosus* are considered as wine spoilers [66]. Similarly, *Pediococcus parvulus* is also considered as unwanted in wine because of its ability to give wines ropiness [67].

Cluster h (Figure 3) consists of LAB that are used for the preparation of yoghurt and cheese [68]. Many food fermentations are performed using mixed cultures of LAB. For example yoghurt is produced by the symbiotic growth of *Streptococcus thermophilus* and *Lactobacillus delbrueckii subsp. bulgaricus* [60]. Cluster h also contains *Lactobacillus acidophilus* and several bifidobacterium species that are used as probiotics [69]. Some dairy products that are created by using *Streptococcus thermophilus*, serve as a carrier for these probiotic bacteria [70, 71]. However, *Streptococcus thermophilus* itself is not used as a probiotic, since it does not survive the stomach acid in healthy humans.

Cluster e (Figure 3) is an interesting cluster. It consists of microorganisms that are on the one hand known to cause food spoilage and on the other hand known for their use in the production of fermentative products. *Leuconostoc mesenteroides* causes spoilage of blood-sausage and vacuum packed meat [72-74], but is also used for improving the texture of sourdough [75] and for the fermentation of sauerkraut [76]; Similarly, Lactobacillus brevis is used for the fermentation of kimchi (a fermented cabbage product originating from Korea) [77], but is also a known beer spoiler [78].

These apparent opposite functions of these microorganisms can be explained by fact that these effects are caused by different strains of the microorganism. This analysis shows that it is important to provide the underlying literature in which these effects are described to get more insight into the details about which type of strain of the microorganism is causing the effect and in which context this occurs.

Additionally, several bacteriocins, e.g. nisin, bavaricin, lactococcin and mundticin are shown in the word clouds. Bacteriocins are small proteins produced by bacteria to inhibit the growth of similar or closely related bacteria. Bacteriocins from LAB are often used to increase shelf life of fermented products. Examples of this are nisin (produced by *Lactococcus lactis*) against *Staphylococcus aureus* (food spoiler) and lactococcin for the conservation of cheese and raw milk [79-81]. These bacteriocins give insight into the
mechanisms by which LABs are used for the conservation of food products. These results show that word clouds of keywords for consortia of LAB give insight into which LAB are functionally related and for which processes and products they are important.

**Discussion and conclusions**

The presented strategy for the retrieval of biological information for microorganisms is to our best knowledge the first approach that uses automatically information extraction from scientific literature for this. A number of methods exists that are able to tag automatically organism names in text but these methods do not provide an extensive analysis of this text in order to retrieve functional keywords that describe the biological properties of these organisms [82, 83]. Some studies include text mining to extract organism related information from scientific literature, but only for specific model organisms [84, 85] or they focus on species specific extraction of genetic entities [86].

Our tagging method can be improved by also searching in Mesh terms. Because Mesh terms are partly manually curated, this is once again a proof that expert curation together with text mining is a good combination to find information in the literature.

We showed that automatic information extraction from Medline abstracts gives rapid insight into the biology of microorganisms that can be used to better understand how these microorganisms are involved in the studied processes. For two lists of organisms derived from healthy individuals we found evidence about the pathogenic potential of these microorganisms, by finding specific disease associations. This analysis showed that with text mining we are able retrieve similar results to manual annotation. Moreover this also showed that we are able to retrieve additional associations with other diseases. Linking the microorganisms to a controlled vocabulary with drug terms, would enable to study the relation between found microorganisms and known therapeutic treatments.

We also showed that text mining without using thesauri yields highly occurring keywords from abstracts in which LAB are mentioned. By presenting these keywords in word clouds, we get an overview of the biological properties and industrial processes in which LAB are used. The knowledge provided by these keywords can be used to make the right decisions in selecting LAB for fermentation processes and for finding new applications for them.

By using the NCBI taxonomy database we are limited to microorganisms in this database. Many specific bacterial strains are not present in the taxonomy database, but are still detected in the abstracts using our regular expressions. These bacterial strain hits are mapped to the closest related bacterium that is present in the NCBI database. This means that a search for abstracts in which *Streptococcus thermophilus* is mentioned can also result in abstracts in which a specific bacterial strain of *Streptococcus thermophilus* is described that is not present in the NCBI taxonomy database. Therefore careful
interpretation of the subsequently presented keywords is required since these keywords could also be related to specific bacterial strains. Although the Medline database covers over 1.5 million abstracts related to microbiology, the database is focused on biomedical related research. To increase coverage of literature in which food related studies are described, other literature sources such as the FSTA database [87] could be used.

Acknowledgements

Author WF was supported by the Biorange project (BR4.2) “A Systems Bioinformatics Approach For Evaluating And Translating Drug-Target Effects In Disease Related Pathways” of NBIC when this work was initiated. From 1 October 2012, WF is supported by the Netherlands eScience project (027.011.306) “Creation of food specific ontologies for food focused text mining and knowledge management studies”.
References


87. FSTA, Food Science and Technology database. http://www.ifis.org/fsta/,
CHAPTER 9
General discussion
Background

In this thesis, the application of text mining (TM) and automatic information extraction (IE), to study drug effects and microorganisms was presented. In drug discovery the effects of a drug on the expression of genes is studied in genome-wide association studies. This often results in hundreds of regulated genes that need to be analyzed. Similarly, microorganisms are studied in community profiling studies, were metagenomics is used to identify microorganisms that are affected by a processes or treatment resulting in tens to hundred of organisms that are affected by the experimental conditions. In both research disciplines annotation of the results from these studies is crucial in order to draw conclusions and to initiate follow-up research.

The Medline database contains millions of abstracts and literature citations that are used by researchers to keep up to date with the state-of the-art developments in their research area. For this reason, the Medline database is very suitable for the annotation of biological experiments, because it can be used to find what already is known in the literature about the genes and microorganisms of interest.

The number of scientific papers in the Medline database grows exponentially. Therefore TM and IE are needed to automatically extract useful information from these papers.

The concept of high-throughput screening

Literature studies to get insight into the developments of a specific research area, involve manual selection and careful reading of interesting papers. Depending on the topic and how much is written about it, this can take weeks. With the rapid accumulation of new scientific papers, it is very likely that the amount of time to read and select useful information from scientific literature will only increase in the future. Moreover manual searching has the drawback of not being all inclusive, possibly missing valuable information.

In drug discovery, high-throughput screening (HTS) is a widely used technique to rapidly screen the activity of large compound libraries to identify compounds that change the activity of a molecular target.

The concept of HTS has been used by the cheminformatics community to perform virtual screening to identify bioactive compounds through computational means [1]. Similarly, the concept of HTS can be applied to automatically screen literature for interesting information using computational methods (Figure 1). HTS of literature using TM and IE involves selection of papers about a certain topic, extraction of useful information from these papers and subsequently presentation of this information in such a way that it can be used by scientists.

HTS of literature saves time since it automatically reduces the large amount of literature to a subset of papers that needs to be analyzed. The exact screening method and subsequent selection of relevant papers is depending on the research question. For example if we are interested in information about the process of cheese making, we first
screen the literature for the occurrence of ‘cheese’ or ‘cheese making’ and select those papers in which these terms occur. This results in 7184 abstracts from over 22 million abstracts that need to be analyzed. The next step in HTS of literature is the automatically extraction of detailed information about cheese making from this subset of articles.

**Figure 1** High-throughput screening of literature involves the total package of selection, extraction and visualization of information from literature.

Extraction of information from papers can be done in different ways. In this thesis we used CoPub. CoPub has been built on the principle that keywords that co-occur in Medline abstracts have a functional relationship [2]. This principle can also be used to find hidden relations by searching for shared intermediates between two keywords which are not directly related [3, 4] or to perform ‘concept profiling’ in which two keywords or concepts which have a similar profile of matching keywords, may be related to each other [5]. Another way to extract information from text is by means of natural language processing (NLP). NLP is able to search for specific relations between two keywords and has successfully been used for the extraction of protein-protein interaction information and for pathway reconstruction [6-9]. Although HTS of literature is focused on fast retrieval of relevant literature and extraction of interesting concepts, experts are still needed to set up intelligent search strategies and to analyze the search results returned by these strategies.
Automatic extraction of gene names from text

IE and TM are well established to retrieve information about genes from literature [10-12]. In the majority of the studies described in this thesis, CoPub was used to identify higher level biological relationships for lists of genes [13-15].

One of the biggest challenges in text mining is the recognition of gene names in text, since genes in scientific literature are described by different names and symbols and some of them even share symbols and names. Results from the gene normalization task of the BioCreative II contest underlined this challenge, since none of the participating systems was able to correctly extract all human gene names from a set of expert-curated Medline abstracts [16]. These types of contests help IE systems to improve their automatic gene recognition procedures, but could also lead to overfitting of the system on the training dataset resulting in a reduced performance on ‘real’ datasets.

A possible solution for solving the problem of correctly identifying gene names in text is to include Entrezgene numbers as unique gene identifiers in papers submitted for publication [17].

In general TM and IE would greatly benefit when papers would be written according to standardized structures and rules. Part of this standardization could be that for every publication two versions are produced. One is a human readable version, but according to a standardized format. The second version is completely structured in which important relations are marked by the authors. This approach of standardization of scientific papers would demand great discipline of the authors and should be a jointed-effort of international research institutes, universities, companies, publishers and the text mining community [18-20].

Co-occurrence based approach vs NLP based approach

Although co-occurrence based TM is very powerful for the analysis of large datasets it also has drawbacks.

In chapter 7 of this thesis, CoPub was used to find additional evidence for the fact that CCL1 and genes with a similar profile show a specific Th2 cell response and IL-2 and genes with a similar profile show a specific Th1 cell response. However, CoPub was not able to find this additional evidence. This is caused by the fact that for over 30% of the abstracts in which Th1 occurs also Th2 occurs and vice versa. For a co-occurrence based system, Th1 and Th1 are too close related in order to detect specific relationships between genes and either Th1 or Th2. This requires a high resolution approach such as NLP. NLP will only detect a relation between CCL1 and Th2 in the sentence “CCL1 is expressed in Th2 cells, but CCL1 is not expressed in Th1” where co-occurrence based TM would find relations between CCL1 and Th1 and between CCL1 and Th2. However NLP is computationally intensive, resulting in complex extracted information structures which makes it less suitable for the analysis of large datasets. A combination of a co-occurrence based TM with NLP might be a good solution for the extraction of information, in which the co-
occurrence based system captures the biology at a higher level, while NLP can be used to find more specific relations.

**Text mining to study effects of synthetic glucocorticoids**

In chapters 3-6, TM was used in multiple ways to get information from the scientific literature about genes and pathways induced by glucocorticoids (GCs). Treatment with synthetic GCs results in side effects and understanding how these side effects are initiated, can help to develop new GC compounds that have reduced side effects while maintaining the desired immune suppressive properties.

In chapter 4 and 6 of this thesis, prednisolone induced genes were analyzed by searching for co-occurrences of the genes with disease- and pathway terms. Although this approach was very effective to get a general insight into the affected biological processes by prednisolone, it lacks specific information on the mechanisms by which prednisolone initiates the side effects. This requires in-depth analysis of the affected processes by experts.

A more detailed approach to study GC induced metabolic side effects was presented in chapter 3, where a network of insulin resistance (IR) related genes was created based on the co-occurrences of these IR related genes in the literature. Subsequently, on this network, associations of the IR genes with dexamethasone were mapped. This system approach identified the effects of dexamethasone on a single gene and on its neighboring genes in the network. Although the study in chapter 3 was entirely based on abstracts from published studies, it nevertheless gave new insights into the development of GC induced IR. We suggested the possible involvement of the sex steroid pathway in GC induced IR.

In chapter 5 we used TM to create a knowledge framework in which we captured the observations of GC effects on genes and proteins. This framework enabled us to identify pathways affected by prednisolone that could be involved in metabolic disturbances in adipocytes.

These studies show that TM can assist research on various levels from gene and protein level to pathway level to identify relevant information about GC biology that can be used by experts to annotate their experiments. Moreover methods such as CoPubGene (chapter 3 of this thesis) and CoPub Discovery, a method to find hidden relations between keywords by searching for shared intermediates [3], can provide novel insights in the side effects of GC treatment.

**Study the dynamics of microorganisms using literature mining**

In chapter 8, thesaurus based text mining was used to find evidence for the pathogenic potential of microorganisms derived from two community profiling studies. Co-occurrences of these microorganisms with disease terms from a thesaurus resulted in many specific disease associations. However, thesaurus based TM is limited to only
keywords in the thesaurus. Especially in food research, thesaurus based TM is not applicable because of lack of specialized thesauri that can be used to find associations in the literature between industrial important microorganisms and food components. Therefore we were interested to see if we could apply TM without using a thesaurus, to identify keywords that describe the properties of lactic acid bacteria (LAB). Moreover we evaluated whether presentation of these keywords in a word cloud would give a sufficient overview of the processes and applications in which these LAB are involved.

**Word clouds**

Word clouds are a powerful solution for quickly summarizing text by maximizing the display of the most relevant keywords in the minimum amount of space. Many web applications use word clouds to summarize news articles [21]. Also in biomedical research, word clouds have been used to summarize gene related information [22], to sum up annotation for biological networks [23], or to summarize information in Mesh descriptors [24].

The word clouds indeed provided a sufficient overview of the biological properties of LAB and the processes in which these LAB were involved. These word clouds are a starting point for specific searches by experts to gain insight into the context in which the found keywords are mentioned.

However, in the word clouds also a number of keywords were found that belong to the same concept, e.g. ‘fermentation’ and ‘fermented’ have the same meaning and ‘yogurt’ and ‘yoghurt’ are the same but different spelled. Nevertheless, the power of this TM approach lies in the fact that without prior knowledge interesting keywords about LAB can be retrieved from abstracts.

Additionally, the information in the word clouds could be improved by giving keywords which belong to the same category the same color or by indicating how keywords in the word cloud are related to each other. Moreover word clouds could be extended by maximizing the display of relevant keywords based on a calculated p-value by comparison with a background word cloud.

**Other literature sources**

TM of information about microorganisms was done on Medline abstracts. The Medline database has a strong focus on biomedical research and is therefore very suitable for finding relations between microorganisms and diseases. However, food related topics are under-represented in the Medline database. Therefore additional literature resources such as the Food Science Technology Abstracts (FSTA) database are needed [25]. FSTA covers food-related literature from 1969 till present and might be useful for the retrieval of food related information on microorganisms. It would be interesting to see if indeed such a dedicated literature resource yields more food-related abstracts (recall) or yields more abstracts that are relevant for food-related research (precision) in comparison with
the Medline database.

**Conclusion**
In this thesis TM was used for high-throughput analysis and annotation of gene expression and protein datasets by linking these sets to diseases, drug terms and biological processes in the scientific literature. This gave insight into the underlying mechanisms of GC induced metabolic side effects on multiple levels. Furthermore TM was used to retrieve disease associations for sets of microorganisms from community profiling studies and for the retrieval of biological properties of groups of LAB to get an overview of the biological processes and industrial applications in which they are involved.
In a time in which the generation of biological data is exploding, TM can have a valuable contribution to translating this data to biological knowledge.
References

Summary
Summary

Bioinformatics approaches are important for the interpretation of results from experiments that are used to study drug effects and in the microbiology. Annotation of these experiments is crucial in order to draw conclusions and to initiate follow-up research.

Automatic information extraction (IE) and text mining (TM) are important techniques to find relevant information in the ever growing (scientific) literature to aid scientific research. In this thesis IE and TM are used to study the metabolic side-effects of glucocorticoid (GC) treatment on genes and biological processes. Furthermore it is shown that these techniques also can be used to study microorganisms by annotation of lists of organisms that are the result of metagenomics experiments, to better understand in which way they influence biological processes and diseases.

In chapter 2 CoPub 5.0 is presented, a publicly available text mining system that calculates keyword co-occurrences in Medline abstracts. CoPub 5.0 is the integration of existing CoPub technology with new features and a new advanced interface that can be used to answer a variety of biological questions. With CoPub 5.0 direct and indirect relations between genes, drugs, diseases and biological processes can be found.

Chapter 3 describes CoPubGene, a method to create literature networks of disease related genes. CoPubGene has been used to create a literature network of genes related to insulin resistance (IR), a known side-effect of glucocorticoid treatment for which the mechanisms are not completely understood. Subsequent mapping of dexamethasone relations on this gene network, gave new insights into the effects of this synthetic glucocorticoid on IR related genes. The method revealed the possible involvement of sex steroid synthesis in the development of GC induced IR. Additionally, implementation of the CoPubGene technology into webservice operations enables to incorporate this technology into bioinformatics workflows which makes it possible to create a wide variety of literature networks of disease related genes.

Chapter 4 describes a genome wide expression profiling study to identify prednisolone induced gene signatures in CD4+ T lymphocytes and CD14+ monocytes derived from healthy volunteers. Identified gene signatures were linked to underlying metabolic related biological pathways. This enables to study the adverse effects of prednisolone on metabolism. TM revealed several affected genes, such as PDK4 and ACSL1 which have a predominantly metabolic function and are less important for the anti-inflammatory action of prednisolone. These genes might be suitable gen-based biomarker candidates for the unwanted metabolic effects of GC treatment.

Chapter 5 describes a study in which the integration of genome wide expression data, adipokine profiling data, and TM techniques identifies pathways and mechanisms involved in prednisolone induced metabolic disturbances in murine 3T3-L1 adipocytes. Several members of the wnt signaling pathway were found to be regulated by prednisolone both
at gene expression level and at protein secretion level. Hence the wnt signaling pathway has been suggested as a suitable pathway candidate for follow-up research to study its involvement in the development of metabolic disturbances in adipocyte cells as a result of GC treatment.

Chapter 6 presents a genome wide gene expression profiling study to examine the role of glucocorticoid receptor (GR) dimerization on the regulation of gene expression in mouse liver. Biological pathways targeted by prednisolone in the liver of wild type (WT) mice and mice lacking the ability to form glucocorticoid receptor dimers (GRdim) were identified and compared using the CoPub technology. In GRdim mice, the level of prednisolone-induced gene expression was significantly reduced compared to WT, but not completely absent. Furthermore several novel GC-inducible genes were identified from which Fam107a, that encodes for a putative histone acetyltransferase complex interacting protein, was most strongly dependent on GR dimerization.

Chapter 7 describes a study that identifies pathways and genes involved in the differentiation of Jurkat T cells towards Th1 and Th2 subtype cells, using various stimuli and pathway inhibitors. PMA/CD3 induced an IL-2 gene profile associated with a Th1 type of response while PMA/CD28 induced a CCL1 gene profile associated with a Th2 type of response. TM analysis using CoPub was used to find additional evidence in the scientific literature for these Th1 and Th2 specific gene expression patterns of IL-2 and CCL1. Analysis with CoPub technology did not result in additional evidence for this specificity. This shows the limitations of co-occurrence based TM methods. However, this type of new and challenging applications can be used to improve the CoPub technology.

In chapter 8 it is shown that TM/IE techniques used in drug discovery can be transferred to other research areas such as microbiology. TM was used for the annotation of organism lists from metagenomics experiments by searching for co-occurrences of the microorganisms and disease terms. Furthermore, analysis of abstracts in which microorganisms occur, resulted in information about the biological properties of these microorganisms and the (industrial) processes in which they are used. This information can be used to assist research towards improvement of the use of microorganisms in existing applications, but also to find new applications. Furthermore this results in new hypothesis for industrial important microorganisms such as lactic acid bacteria.
Samenvatting
Samenvatting

Het toepassen van bioinformatica-methoden zijn belangrijk voor het begrijpen en interpreteren van resultaten van experimenten die gebruikt worden in studies naar medicijnwerking en in studies naar de toepassing van micro-organismen. Annotatie van deze experimenten is van cruciaal belang om de juiste conclusies te kunnen trekken en om vervolg onderzoek te kunnen initiëren.

Automatische informatie extractie (IE) en text mining (TM) zijn belangrijke technieken voor het vinden van relevante informatie in de steeds groeiende (wetenschappelijke) literatuur. In dit proefschrift worden IE en TM gebruikt voor het bestuderen van de ongewenste effecten als gevolg van de behandeling met glucocorticoïden (GC), op genen en biologische processen die betrokken zijn bij de stofwisseling. Verder wordt gedemonstreerd dat deze technieken ook gebruikt kunnen worden voor het annoteren van lijsten van micro-organismen die het resultaat zijn van metagenomische experimenten met als doel zo beter te kunnen begrijpen wat de invloed en werking van deze micro-organismen zijn op biologische processen en ziektes.

In hoofdstuk 2 wordt CoPub 5.0 gepresenteerd, een TM systeem dat relaties tussen trefwoorden (keywords) berekent in Medline samenvattingen. Een nieuwe geavanceerde interface, die bestaande CoPub-technologie combineert met nieuwe kenmerken, kan worden gebruikt voor het beantwoorden van een aan verscheidenheid aan biologische vragen. Met CoPub 5.0 kunnen directe en indirecte relaties tussen genen, medicijnen, ziekten en biologische processen gevonden worden.

Hoofdstuk 3 beschrijft CoPubGene, een methode voor het maken van netwerken van ziekte-gerelateerde genen op basis van informatie in de literatuur. CoPubGene werd gebruikt voor het maken van een literatuur netwerk van genen, gerelateerd aan insulineresistentie (IR), een bekende maar onbegrepen bijwerking van de behandeling met GC. Het mappen van dexamethasone-relaties op het verkregen gen-netwerk gaf nieuwe inzichten in de effecten van deze synthetische GC op IR gerelateerde genen. De methode onthulde de mogelijke betrokkenheid van leden van het sex-steroid-synthese proces bij de ontwikkeling van GC geinduceerde IR. Implementatie van de CoPubGene-technologie in webservice operaties stelt ons bovendien in staat om deze technologie te gebruiken in bioinformatica-workflows, om aldus literatuur netwerken voor een verscheidenheid aan ziekte gerelateerde genen te kunnen maken en bestuderen.

In Hoofdstuk 4 wordt een genoomwijde genexpressie-profilerings-studie beschreven, waarin gen signaturen worden geïdentificeerd, die geïnitieerd worden door prednisolon in CD4+ T lymphocyten en CD14+ monocyten, afkomstig van gezonde vrijwilligers. Deze geïdentificeerde genen werden in dit onderzoek gekoppeld aan de onderliggende metabole biologische processen. Op deze wijze werden we in staat gesteld om de ongunstige effecten van prednisolon op de stofwisseling te bestuderen. Door gebruik te maken van TM, werden genen zoals PDK4 en ACSL1 gevonden die voornamelijk bij de
stofwisseling betrokken zijn, en minder bij de anti-inflammatoire effecten van prednisolon. Deze genen kunnen mogelijk gebruikt worden als gen-gebaseerde biomarker-kandidaten voor de ongewenste effecten op de stofwisseling als gevolg van de behandeling met GC. In **Hoofdstuk 5** wordt een studie beschreven die door het combineren van genoomwijze genexpressie data, adipokine-profilerings data en TM technieken, processen en mechanismen identificeert die betrokken zijn bij prednisolon-geïnduceerde metabole verstoringen in murine 3T3-L1 adipocyten. Verschillende leden van de wnt signaling pathway werden gereguleerd door predisolon op zo wel gen-expressie niveau als op eiwit niveau. Daarom is de wnt signaling pathway een goede kandidaat voor vervolgonderzoek om zo de betrokkenheid van dit proces bij de ontwikkeling van prednisolon-geïnduceerde metabole verstoringen in adipocyte cellen te bestuderen.

**Hoofdstuk 6** presenteert een genoomwijze genexpressie studie om de rol van het vormen van dimeren van de glucocorticoid receptor (GR) op de regulatie van gen expressie in muizenlever te achterhalen. Biologische processen die door prednisolon beïnvloed werden in de lever van wildtype (WT) muizen en in de lever van genetisch gemodificeerde muizen die niet meer in staat zijn om dimeren van GR (GRdim) te vormen, zijn m.b.v. de CoPub- technologie vergeleken met elkaar. In GRdim muizen was het niveau van prednisolon geïnduceerde gen-expressie significant lager in vergelijking met WT muizen, maar niet helemaal afwezig. Verder werden er verschillende nieuwe GC-induceerbare genen gevonden, waarvan *Fam107a*, dat codeert voor een eiwit dat mogelijk een binding aan gaat met een histone acetyltransferase complex, daarbij sterk afhankelijk is van GR dimerisatie.

**Hoofdstuk 7** beschrijft een studie waarin processen en genen geïdentificeerd werden, die betrokken zijn bij de differentiatie van Jurkat T cellen naar specifieke Th1 en Th2 subtypen cellen, door gebruik te maken van verschillende stimulerende stoffen en pathway remmers. PMA/CD3 induceerde een *IL-2* gen profiel dat geassocieerd word met een Th1 type response, terwijl PMA/CD28 een *CCL1* gene profiel induceerde dat geassocieerd word met een Th2 type response. Er werd vervolgens gebruik gemaakt van CoPub, om additioneel bewijs in de wetenschappelijke literatuur te vinden voor deze Th1 en Th2 specifieke expressiepatronen van *IL-2* en *CCL1*. De analyse met de CoPub technologie leverde echter geen additioneel bewijs op voor deze specificiteit en toont daarmee ook de beperkingen van co-occurrence gebaseerde TM methoden aan. Dit soort nieuwe en uitdagende toepassingen kunnen echter worden gebruikt om de CoPub technologie verder te verbeteren.

In **Hoofdstuk 8** wordt gedemonstreerd dat TM/IE-technieken zoals die in medicijnonderzoek gebruikt worden, ook in andere onderzoeksvakgebieden, waaronder bijvoorbeeld de microbiologie ingezet kunnen worden. TM kan worden gebruikt voor het annoteren van lijsten van organismen komende uit metagenomische experimenten, door in de wetenschappelijke literatuur te zoeken naar relaties tussen deze organismen en ziektes. Tevens levert automatische analyse van wetenschappelijke literatuur waarin
micro-organismen worden genoemd, informatie op met betrekking tot de biologische eigenschappen van deze organismen en de (industriele) toepassingen waarin ze worden gebruikt. Deze informatie helpt onderzoek naar het verbeteren van bestaande microbiologische toepassingen, als naar het vinden van nieuwe toepassingen. Verder roept dit nieuwe hypothesen op aangaande industrieel belangrijke micro-organismen zoals melkzuur bacteriën.
Dankwoord

En dan nu het dankwoord...

Dit proefschrift heeft alleen tot stand kunnen komen door de mensen die mij gesteund hebben tijdens mijn promotieonderzoek. Een persoonlijke dank aan jullie is hier dan ook wel op zijn plaats.

Jacob, jouw manier van toegepaste wetenschap bedrijven heeft mij altijd aangesproken. Ooit begonnen als master student bij het CMBI, hield ik een presentatie over mijn stageproject. Jij zat toen in het publiek en stelde na afloop de o zo relevante vraag “Maar wat kun je hier nou mee?” Tijdens mijn promotieonderzoek heb ik die ene vraag altijd onthouden. Ik ben dan ook dankbaar dat ik in jouw academische groep, waar toegepaste wetenschap belangrijk is, mijn promotieonderzoek heb mogen uitvoeren.

Wynand, van jou heb ik zo ontzettend veel geleerd. Jouw tomeloze inzet, enthousiasme en pragmatische manier van werken, heeft mij zeer gemotiveerd om uiteindelijk dit ‘boekje’ te schrijven. Onze samenwerking begon bij MDI in Oss waar we text mining methoden ontwikkelden en toepasten in drug discovery projecten. Na de aankondiging van MSD om de onderzoeksafdelingen te sluiten, besloot je om je carrière bij NIZO in Ede voort te zetten. Ik ben je erg dankbaar dat je ook na deze carrièreswitch mij nog steeds wilde begeleiden bij mijn promotieonderzoek, ook al betekende het dat je dit naast je werkzaamheden bij NIZO in je vrije tijd moest doen. Je was nooit te beroerd om tijd in te plannen voor overleg, dan wel telefonisch dan wel bij Starbucks op het station of later ook bij NIZO. Deze samenwerking heeft er ook in geresulteerd dat we nu samen verder werken aan text mining voor voedselonderzoek in een eScience project.

Het eerste gedeelte van mijn promotieonderzoek heb ik binnen de muren van de MDI afdeling van Schering-plough (voorheen Organon, later MSD) uitgevoerd. De prettige werksfeer die daar heerste, mede te danken aan een aantal inmiddels ex-collega’s, heeft zonder meer bijgedragen bij de totstandkoming van dit ‘boekje’. Ik wil jullie hier zeker niet vergeten: Ton, Cees, Rene, Tinka, Ruud, Ross, Scott, Rita, de Jannen, Susanne, Peter, Maarten, Thea, Karin en Ria. Speciaal wil ik hier Stefan noemen. Jouw onuitputtelijke kennis van computers en programmeren heeft er voor gezorgd dat ik altijd bij jou terecht kon wanneer ik vragen of problemen had. Ik wil je graag bedanken voor al je hulp en alles wat ik van je geleerd heb.

Ook de (ex) collega’s van de CDD groep en het CMBI verdienen het om hier genoemd te worden. Allereerst Raoul “CoPub Berry”. Samen hebben we aan text mining gewerkt, ook jouw promotie onderwerp, wat geresulteerd heeft in het behalen van je Dr.-titel in september 2011. Je was een fijne collega om mee samen te werken. Succes bij je baan
als bioinformaticus bij Rijk-Zwaan.

Simon “Koetje”. Je was al weg toen ik aan mijn promotieonderzoek begon, maar het contact dat we hebben gehouden en ook onze gemeenschappelijke historie bij het CMBI hebben er voor gezorgd dat ik je hier toch even wil noemen. De bezoeken waarbij men met jou altijd zo mooi kon kletsen over van alles en helemaal niks, waren altijd zeer ontspannend. Tegenwoordig zit je bij TNO en ben je trotse vader van dochter Isis!

Dave “Dr. D. indeed” Wood. Although you were only there for a short period, it was very nice meeting you. I wish you all the best in London with your wife Alex and son George.

Sander, naast inmiddels ex-college, ben je vooral een goede vriend. Ik ben dan ook erg trots dat je mijn paranimf wil zijn. Je was altijd geïnteresseerd in mijn werk en naast het werk stond de deur bij jou, Lisa, Marrit en Mirte altijd open voor een gezellig samen zijn. Veel succes met je carrière bij Lead pharma en het aanstaande vaderschap.

Marijn “Weet je wat ut is” Sanders. Samen waren we de AIO’s in de CDD groep en hadden we veel steun aan elkaar tijdens alle veranderingen waarmee we geconfronteerd werden tijdens onze promotie. Nu werken we samen aan text mining in het eScience project.

Juma my roommate at the CMBI: Thanks for your help and support. Jurgen, bedankt voor de lekkere koffie en de mooie verhalen (succes bij d’n Epic). Barbara voor alle administratieve zaken.

Erik “stuur de sh*t maar op” Toonen. Bedankt voor de vele discussies. Op het gebied van glucocorticoid biologie en het schrijven van wetenschappelijke artikelen heb ik veel van je geleerd. Het is ook niet voor niets dat je als mede-auteur bij vier hoofdstukken in dit proefschrift staat vermeld!

Ontspanning tijdens het promotieonderzoek is belangrijk om je gedachten even te kunnen verzetten. Waar kun je dit beter doen dan bij LOKOMOTIV NIJMEGEN? De vele uurtjes op de fiets, maar ook de gezellige biertjes bij Frowijn hebben er altijd voor gezorgd dat ik mijn promotieonderzoek in perspectief heb kunnen plaatsen. Mannen bedankt! Graag wil ik hier ook nog speciaal Walter bedanken voor de prachtige kaft van mijn proefschrift!

Niels (geweldig dat je mijn paranimf wilt zijn) en Judith bedankt voor alle plezierige etentjes en de gezellige avondjes uit. Geniet van jullie prachtige dochter Ika.

Daniel thanks for your support and the nice discussions about science and about life. I wish you and Ghislaine all the best for the future!

En dan als laatste wil ik mijn familie bedanken. Allereerst mijn zus Karin, schoonbroer Bas en mijn neefjes en nichtje Jelmar, Yulotte, Tibbe en Lauritz. Een prachtige familie, waar ik altijd terecht kan.

Meinen Schwiegereltern Peter und Heike: Es ist bei Euch in Idstein wie ein zweites
Zuhause für mich. Vielen Dank für Eure Unterstützung!
Lieve pap en mam, soms was het lastig te begrijpen wat ik nu precies voor een werk deed, maar jullie waren altijd geïnteresseerd en probeerde het te begrijpen. Jullie stonden en staan nog steeds voor mij klaar wanneer ik jullie nodig heb. Ik kan jullie dan ook niet genoeg bedanken voor de fantastische ondersteuning die mij gemotiveerd heeft om dit proefschrift te schrijven!

En dan de belangrijkste dames in mijn leven:
Lieve Lara, toen ik bijna twee jaar bezig was met mijn promotieonderzoek kwam je ter wereld. Het promotieonderzoek en het vaderschap bleek soms lastig te combineren, maar ik ben zo trots om jouw papa te mogen zijn. Je bent mijn inspiratie geweest om dit proefschrift te schrijven. Dikke kus van papa!

Liebe Tina,
Du bist das Ziel einer langen Reise
Die Perfektion der besten Art und Weise
In stillen Momenten leise
Die Schaumkrone der Woge der Begeisterung
Bergauf, mein Antrieb und Schwung
Ich wollte dir nur mal eben sagen
Dass du das Größte für mich bist
Und sichergehen, ob du denn dasselbe für mich fühlst
Für mich fühlst
(Von dem Lied "Ein Kompliment", Sportfreunde Stiller)

Wilco
Nijmegen, maart 2013
CURRICULUM VITAE

Bibliography


