High Throughput Screening of Gene Expression for the Investigation of Multifactorial Dermatological Disorders

by

Máté Manczinger MD.

Supervisors:

Lajos Kemény MD, PhD, DSc and Lóránt Lakatos PhD

PhD Thesis

Department of Dermatology and Allergology
University of Szeged, Hungary

2016

Introduction

Most common diseases are caused by numerous internal and environmental factors. Multiple intracellular discrepancies in gene expression and in protein abundance can be found causing the phenotype of these diseases. Thus, the investigation of the system instead of one gene or protein have become necessary to understand disease mechanisms and find the most effective treatments. I am going to demonstrate high-throughput gene expression studies for two multifactorial diseases: the infectious vulvovaginal candidiasis (VVC) and the non-infectious psoriasis. In the first study, we constructed an in vitro model of VVC and carried out RNA-Seq analysis to identify virulence factors of the fungus. In the second one, we have done a meta-analysis of published microarray data about psoriasis, which was further processed in silico with a network-based approach. In both studies, we looked for potential proteins, which may have role in pathogenesis and could be modulated with drugs and we identified potential drug candidates as well for psoriasis.

The physiologic and pathogenic role of the microbiome has become an important field of medical sciences in the recent years. *Candida albicans* (*C. albicans*) is a dimorphic fungus exhibiting non-pathogenic yeast morphology as a member of the normal human flora as well as pathogenic hyphal morphology. *C. albicans* is adapted to its mammalian host in several ways and hyphal growth can be induced by several factors. Everything that makes the environment unfavorable for the fungus, induces the hyphal transcriptional program. Filamentous form causes cell damage by penetration. Induced phagocytosis of the yeast is also prevalent. Several invasin molecules mediate the uptake of the fungus by host cells. Filamentous growth is then initiated in the phagosome and hyphae cause cell lysis.

Differentiation between the two forms is present on the level of innate recognition. Pattern recognition receptors recognize glycoproteins on the surface of the fungus in a non-specific way. Cell wall constitution of yeast is different from hyphae and cell wall remodeling is also initiated in the beginning of filamentation. Filamentous form induces immune-response, while the commensal form is less immunogenic. The particular role of adaptive immunity against *C. albicans* is less well described. Several studies reported the presence of *C. albicans*-specific antibodies in the circulation. Induction of the cellular adaptive immune response, which has also been described, contributes more significantly to defense mechanisms against candidal infection, than humoral factors. While adaptive immune system has a significant role during

the course of the disease on the mucosal surfaces of the gastrointestinal tract, only the role of innate immunity is supposed in case of VVC.

Huge amount of large-scale data about microorganisms in the normal human flora is available. Genetic, genomic, metagenomic, epigenomic, transcriptomic, proteomic, metabolomic and evolutionary analysis have become main-stream methods in biomedical research. High throughput sequencing made it possible to analyze genomes, epigenomes, transcriptomes and the constitution of microbes during infection. As a result, more effective identification of virulence factors and, thus, potential drug targets have become possible.

Psoriasis is one of the most studied skin diseases. More than 39000 hits are available currently in PubMed for the keyword „psoriasis" and the number is increasing. Patients have hyperkeratotic, scaly patches on their skin. Predilection sites are the elbows, knees and the sacral area, but it can also affect the scalp, nails and the whole skin surface. Genetic predisposition and environmental factors are both important factors in disease etiology. Several genome-wide association studies have been carried out and until now 36 susceptibility loci have been identified. Environmental triggers like drugs, smoking, mental stress, skin injury, Streptococcal infection, hormonal changes etc. are reported. Psoriasis is an immune-mediated disease. Important immune cells and cytokines have been identified in disease pathogenesis such as IL6, IL17A, TNF etc. Autoimmune basis for chronic inflammation is supposed, although no consistent antigen has been found. Keratinocyte hyperproliferation is present in lesional phenotype and is responsible for scale formation. Patients with psoriasis have higher risk for metabolic syndrome, and risk increases with disease severity. Both diseases have immunological basis with common cytokines and genetic risk loci like CDKAL1.

Numerous treatment options are available. Recommendations depend on disease severity, co-morbidities and treatment history. Treatment options can be classified into local, UV-light and systemic groups. Most common local therapies are steroids, vitamin D3 analogues, dithranol, retinoids and calcineurin inhibitors. Mainstream systemic therapeutic options are methotrexate, calcineurin inhibitors and biological therapy. A general consideration is that different patients respond to treatment options with different extent. In some cases, treatment worsens the disease instead of improving it. There are also treatment-resistant cases. All of these facts suggest, that psoriasis is much more complex than how we are now understanding it and systematic analysis is required to reveal further aspects of pathogenesis.

"Omics" data gives the opportunity to examine the disease with systems biology approach. With this, it has become possible to reveal new treatment options and to understand the disease better. The way we handle huge datasets is crucial to get reliable results. The proper use of bioinformatics tools for filtering, normalization and statistical processing is essential to avoid false inferences. In silico model construction is possible only after these steps

Stationary changes in gene expression are responsible for fixing phenotypes like the lesional skin area in psoriasis. DNA microarray is a widely used method for large-scale screening of gene expression. Assays for splicing, single nucleotide polymorphisms (SNP), transcription factor binding sites, fusion genes etc. have become available with the improvement of microarray technology. In silico normalization and filtering of data is crucial. Microarray is relatively cheap and broad spectrum of bioinformatical methods and libraries are available for the analysis of microarray data. One disadvantage opposed to high-throughput sequencing is, that microarray cannot reveal new transcripts, only predestinated sequences can be detected. Several microarray studies have been carried out to characterize gene expression in healthy and psoriatic skin samples. Microarray meta-analysis gives the opportunity to evade biological, regional, and study design-caused variation between studies.

High throughput sequencing technologies emerged in the 1990s. Researchers are able to get sequences of whole genomes and transcriptomes, but these technologies can also be used for different other purposes like characterizing DNA-protein interactions (ChIP-sequencing) and the epigenome. The cost of these assays were high, but the price is constantly getting lower allowing their use in basic research. Compared to the classical "chain termination" based Sanger sequencing, these methods are much faster, have higher sequencing capacity and lower cost. The most popular types are ligation based sequencing (SOLiD), synthesis-based sequencing (Illumina), pyrosequencing (454) and ion torrent sequencing methods. We used SOLiD sequencing in our VVC study. Beside the already mentioned advantages of next generation sequencing, it is also capable to identify novel transcripts. Although the cost of these methods is getting lower, it can be still very expensive to use biological replicates. If this is the case, validation of results is essential.

Network analysis is a novel and highly developing area of systems biology. Considering gene expression data, it is possible to explain alterations in intracellular processes with the analysis of protein-protein interaction (PPI) and protein-DNA interaction (PDI or gene regulatory) networks. These networks consist of proteins and/or regulated genes as nodes and undirected or directed edges between them. Centralities, like degree or stress, are suitable for

ranking nodes. Interconnected nodes make up network motifs. These elements have important role in network dynamics It is important to investigate intracellular proteins as members of an intracellular network. Since the description of scale-free networks in the late 90s, rapid development can be seen in the field of network-based analysis of large-scale datasets. Multiple in silico methods and software are developed and are widely used. We can consider biological pathways as networks and analyze processes instead of individual proteins, genes or metabolites. We can also construct custom networks based on protein-protein and protein-DNA interactions. Network analysis is also prevailing in drug discovery. Identification of new drug targets is essential. For this, first, the integration of all information about the model biological system is needed. The most plausible way is to create networks. Network based approaches help us to predict drug side-effects, drug-drug interactions and can be used in numerous other fields of drug discovery.

Aims

I. We managed to construct an in vitro model of VVC in our first study. We hypothesized that the characteristics of hyphae growing in the presence of human cells is markedly different from control hyphae, which grow without human cells. We also supposed, that genes, that are solely differentially expressed in the presence of human cells are potential virulence factors.

II. Our goal was to construct reliable networks consisting of differentially expressed genes (DEG) between lesional and non-lesional skin samples of psoriatic patients and/or the encoded proteins. We hypothesized, that it could be possible to find novel elements in psoriasis pathogenesis and potential drug candidates with the detailed analysis of these networks.

Methods I

Strains, growth conditions and cell culturing

C. albicans clinical isolate SC5314 was grown on YPD medium at 30°C, cultured under standard conditions. The immortalized human vaginal epithelial cell line (VECL) PK E6/E7 was cultured in serum-free complete keratinocyte medium (CKM) (Life Technologies) in a CO_2 thermostat at 37°C. A total of 10^5 PK E6/E7 VECL cells were seeded in 6-well plates and incubated for 24 hours in serum-free CKM. At 24 hours prior to infection with *C. albicans*, the medium was changed to serum-free CKM without antibiotic/antimycotic solution. Fungal cells

were added to wells with a multiplicity of infection (MOI) of 3:1 to infect PK E6/E7 VECL cells. For hyphal control, fungal cells were resuspended only in complete CKM without antibiotic/antimycotic solution. In order to induce hyphal growth, plates were incubated in a CO_2 thermostat at 37°C. Plates were incubated for 3 hours in a humidified atmosphere containing 5% CO_2 at 37°C; fungal cells rapidly switch to filamentous growth under such circumstances. Yeast control cells were harvested at 0 hour time point

Viability test

The effect of *C. albicans* infection on the viability of PK E6/E7 VECL cells was measured by Real-Time Cell Analysis (RTCA; ACEA Biosciences), as described previously. 10^4 PK E6/E7 cells per well were seeded in 96-well E-plates (ACEA Biosciences). Epithelial cells were allowed to attach to the bottom of the wells and grow for 3 days. Cells were then treated with *C. albicans* hxk1Δ or DIC185 cells. Triton-X (Sigma) treatment was used as a positive control to kill the vaginal epithelial cells. Real-time measurements of impedance were done with the xCELLigence System RTCA HT Instrument (ACEA Biosciences); The cell index at each time point was defined as $(R_n - R_b)/15$, where R_n is the cell-electrode impedance of the well when it contains cells and R_b is the background impedance of the well with the medium alone. CI values reflect cell number, adherence, cell growth, and health. Statistical significance between treatment groups was determined using one-way and two-way ANOVA following pairwise T tests with Bonferroni correction. Experiments were repeated three times and the number of biological replicates varied between 3 and 6.

C. albicans adherence assay

For adherence assay, PK E6/E7 VECL cells were grown until confluency was reached. A total of 10^5 hxk1Δ mutant and the parental strain (DIC185) cells suspended in CKM were used to infect vaginal epithelial cells. Supernatant was aspirated after 1,5 hours and the wells were washed two times with 1× PBS. The monolayers with attached *C. albicans* were fixed by 3.7% (v/v) paraformaldehyde in PBS. Quantitation of *C. albicans* adherence was performed by light microscopy. Ten randomly chosen fields of view covered with epithelial cells were counted. Significance was calculated with a two-sample T test and a p value of less than 0.05 was considered to be significant. Experiments were performed with three biological replicates.

High throughput sequencing, bioinformatical and statistical analysis

Whole transcriptome sequencing was performed as described previously. Bioinformatical analysis of the whole transcriptome sequencing was performed using Genomics Workbench (CLC Bio). Sequences were mapped in a strand specific way onto the *C. albicans* SC5314 assembly 19 reference genome. Normalized gene expression was calculated using the "scaling" normalization method. Differentially expressed genes from the RNA-Seq output were determined using the R package DEGSeq. Gene expression was considered significantly different between two conditions if the false discovery rate (FDR) corrected probability (p) value was less than 0.05 and the absolute fold change value was more than 2.

Quantitative reverse transcriptase polymerase chain reaction (QRT-PCR)

cDNA was synthesized from at least 100 ng of total RNA. SybrGreen technology-based real-time quantitative PCR was used to quantify the relative abundance of the selected mRNAs. As controls, we used reaction mixtures without cDNA. Relative expression of the given gene in the yeast-like form was set to 1 and the expression in control hyphae and hyphae developed in the presence of PK E6/E7 vaginal epithelial cells was calculated by comparing the values to the yeast form. Two technical and three biological replicates were used. The ratio of each mRNA relative to the 18S rRNA was calculated using the $2^{-\Delta\Delta CT}$ method; Differential expression was assessed with one-way ANOVA followed by pairwise t-test and Bonferroni p value correction.

Methods II

Microarray meta-analysis

Six microarray studies examining lesional and non-lesional skin biopsy samples of psoriatic patients were found in GEO. Only non-lesional and lesional samples from affected individuals were used for analysis, samples of healthy individuals were excluded. Quality of each sample was assessed with the R package arrayQualityMetrics, and outliers were excluded. Raw data normalization of remaining samples was carried out with the R package Easy Microarray data Analysis (EMA). The R package MetaQC was used for filtering out low quality studies. MetaDE package was used to determine DEGs in lesional samples compared to non-lesional ones. Genes with FDR less than 0.001 and with fold change higher than 1.5 or less than -1.5 were accepted as DEGs.

Construction of protein-protein, protein-DNA and chemical-protein interaction networks

STRING database 9.0 was used as resource for protein-protein interactions (PPI). Both directed and undirected networks were created by selecting all interactions between DEG – coded proteins in raw STRING data. Interactions with a confidence score higher, than were used in case of undirected and 800 in case of directed network construction. Chemical-protein interactions between potential drugs, intra- and extracellular compounds and DEG-coded proteins were collected from STITCH database 3.1. Confidence score calculation is the same in this database as in STRING thus interactions with the described confidence score cutoff values were selected for network construction. Protein-DNA interaction (PDI) network consisting of DEGs and DEG-coded transcription factors (TF) was created using cis-Regulatory Element Database (CisRED). Merged DEG-derived network containing PPI and PDI interactions was also generated. Complete PPI, PDI, merged and chemical-protein interaction networks were created for controls using all available interactions in databases with the same statistical threshold as in DEG-derived network construction.

General network analysis, identification of central nodes and motif detection

General network analysis and node centrality value calculation were carried out with NetworkAnalyzer Cytoscape plugin. Power law distribution of these node centralities was proven, Power law distribution cannot be described with classical statistics used for Gaussian distribution. The sum of first and second moment (mean value and variance) was used as cutoff for centralities with distribution exponent of $\gamma > 3$. Expression of variance becomes infinite, if $\gamma \leq 3$, thus only first moment (mean value) was used as cutoff for centralities with distribution exponent $2 < \gamma < 3$. Expression of mean value becomes infinite, if $\gamma \leq 2$. In this case weighted mean was used to assess cutoff. NetMODE software was used for network motif statistical analysis. NetMODE p value indicates the number of random networks in which a motif occurred more often than in the input network, divided by total number of random networks. $p < 0.05$ was used as cutoff. Respective sub-networks of enriched motifs were identified with NetMatch Cytoscape plugin.

Results I

Vaginal epithelial cell - C. albicans co-culture as a model of vulvovaginal infection

Infection of epithelial cells by *C. albicans* requires adhesion of yeast form cells to the surface of epithelium. We used the immortalized PK E6/E7 VECL and infected them with *C. albicans* SC5314 yeast form cells. Since we aimed to monitor the primary effect of human cells on hyphae formation, we sampled the cells at 3 hours post infection. *C. albicans* cells adhered to the surface of PK E6/E7 VECL and developed hyphae and approximately 5% of hyphae penetrated into epithelial cells. We call this sample "model hyphae" for simplification. *C. albicans* cells were cultured in serum-free CKM without human cells for 3 hours as control. Microscopic examination showed that at this time point *C. albicans* cells are adhering to the surface of the culture chamber and develop hyphae. We call this sample "control hyphae" for simplification.

Primary analysis of transcriptome data

To study the early and specific molecular events occurring upon hyphae formation in the absence or presence of vaginal epithelial cells, global transcriptional changes of *C. albicans* cells were monitored using RNA-Seq. Genes with 2-fold absolute difference of gene expression and a FDR less than 0.05 were considered to be DEGs. Pairwise comparison resulted in 1283 DEGs between control hyphae and yeast. Surprisingly, almost twice as much, 2537 DEGs were found between model hyphae and yeast. We identified 1574 DEGs between the two hyphal forms. RNA-Seq data allowed us to identify 384 DEGs with higher expression in both hyphae compared to yeast and no difference between the two hyphal forms. These genes might be considered as effector genes of hyphae formation as a response of culturing *C. albicans* in serum-free CKM. We found 376 DEGs upregulated in both hyphal forms compared to yeast and differentially expressed between control and model hyphae. 1205 DEGs were exclusively found between model hyphae and yeast. These genes may play an important role in the virulence of *C. albicans* after contact with vaginal epithelial cells. QRT-PCR analysis was performed to validate the expression pattern of 22 genes. This gene set includes representatives of all the identified expression patterns. The results of QRT-PCR analysis are in complete agreement with the RNA-Seq expression data.

Functional analysis of RNA-Seq data

Analysis of signal transduction pathways involved in hyphal morphogenesis

Our culturing conditions induced strong hyphal morphogenesis of *C. albicans* with or without PK E6/E7 VECL. We sought to determine if different signal transduction pathways leading to hyphal morphogenesis are responding to these conditions at the level of transcription. Our results suggested, that the ratio of transcriptional activators and repressors is crucial in the regulation of the hyphae specific genes, as upregulation of both activators and repressors could be seen.

N-acetyl-glucosamine (GlcNAc) is known to induce hyphal morphogenesis and white opaque switching in *C. albicans*. Interestingly, we found that the NGT1 gene representing the transporter gene in the GlcNAc transporter was solely significantly upregulated in model hyphae, indicating the specificity of this response to epithelial cells. As GlcNAc induces hyphal morphogenesis in *C. albicans*, we sought to monitor the expression of GlcNAc catabolic genes in our in vitro model. Since the RNA-Seq experiment did not provide sufficient number of unique reads for statistical analysis of this group, the expression of several GlcNAc catabolic genes was tested by QRT-PCR. For this, the following conditions were used: control hyphae, model hyphae, and control hyphae supplemented with 10 mM of GlcNAc. Expression of GlcNAc deacetylase (DAC1), hexokinase 1 (HXK1) and GlcNAc deaminase (NAG1) were repressed in control hyphae as compared to the yeast form *C. albicans*; but the expression of NGT1 remained unaltered. Our results showed, that the expression of all four genes (NGT1, DAC1, HXK1, and NAG1) genes involved in GlcNAc catabolism were significantly upregulated in model hyphae. Administration of 10 mM GlcNAc caused definite expression of the GlcNAc catabolic genes that is probably due to the high concentration of the substrate for catabolism.

We next sought to determine the importance of the GlcNAc metabolic pathway in the virulence of the fungus. Nag1Δ and dac1Δ mutants could not grow on glucose if the medium contained GlcNAc; hence we have chosen to use a hxk1Δ mutant strain in our subsequent experiments. We used an RTCA assay, which provides real-time, quantitative information about the number of the living, attached cells by measuring electrode impedance. Vaginal epithelial cells were treated with different numbers of yeast form *C. albicans* parental (DIC185) or mutant (hxk1Δ) strains, the impedance was measured for 24 hours and the data was converted to cell index (CI). Our results showed, that when PK E6/E7 VECL cells were infected with low

numbers (2000 and 5000) of *C. albicans*, the hxk1Δ mutant exhibited lower cytotoxicity as compared to wild type. When the number of infecting *C. albicans* cells was increased (10000 and 20000 cells), the hxk1Δ mutant no longer exhibited a reduced cytotoxic effect.

Numerous published data prove, that GlcNAc has a significant role in cell wall construction. We hypothesized, that cell wall, and glycosylated cell wall proteins may have a key role in the adhesion to host cells and this could be an important factor in the decreased virulence of hxk1Δ strain. We carried out an adherence assay. Monolayers of PK E6/E7 vaginal epithelial cells were treated with *C. albicans* parental (DIC185) and mutant (hxk1Δ) strains. After 90 min of contact, non-adhered cells were washed away and the number of adherent *C. albicans* cells was counted. Our results showed, that significantly less hxk1Δ mutant remained attached to the surface of the PK E6/E7 cells compared to the DIC185 parental strain. This data indicates the importance of HXK1 gene and therefore the GlcNAc pathway in the adherence of *C. albicans* to vaginal epithelial cells. Noteworthy, it was published parallel with us, that GlcNAc could have a role in adhesion in other organisms, such as *Staphylococcus aureus.*

Results II

Detection of DEGs with microarray meta-analysis

In order to get reliable data about gene expression in lesional psoriatic skin samples, microarray meta-analysis was carried out. The study by Johnson-Huang et al. was already excluded after sample quality analysis with arrayQualityMetrics package. The overall quality of each study was assessed with MetaQC as described before. All five studies were defined as being non-problematic based on quality values. DEGs were identified by MetaDE. 2307 upregulated and 3056 downregulated genes were found in lesional skin samples compared to non-lesional ones. DEGs were used for network construction. The high number of DEGs (5363) in our study may be surprising, but it can be caused by the lower gene expression fold change cutoff (1.5 and -1.5 instead of 2 and -2) and by pre - filtering of samples, which can decrease variance and, thus, can increase the number of DEGs.

General Network analysis

Undirected and directed PPI networks with DEG – coded proteins, directed PDI networks with DEG – coded TFs and regulated DEGs and merged directed networks containing both PPIs and PDIs were created. DEG – derived networks had higher diameter (i. e. the length of the longest shortest path in the network) and average shortest path length, than control full

networks. This may be caused by the inverse correlation of node degree and fold change. Nodes with lower fold change has higher degree. As genes with a fold change under cutoff are filtered out from DEG derived networks, remaining nodes has smaller average degree, therefore connectivity of the network is lower resulting in higher diameter and average shortest path length value.

Determination of hubs in DEG-derived networks

Most important nodes of DEG-derived networks were determined using degree and/or stress centralities. Numerous psoriasis-associated protein-coding genes, which are already published, were found. CCNA2, FYN and PIK3R1 proteins are yet unpublished in association with the disease and can be found in top rated hubs in undirected PPI network. PDI network contained DEG-coded TFs and regulated DEGs as nodes and directed edges pointing from the TFs to the regulated genes. TFs were ranked using the out-degree centrality. Androgen receptor (AR), TFDP1, MECOM and MEF2A, which are yet not associated with psoriasis were above centrality cutoff.

Motif analysis in DEG-derived networks

Motifs consisting of 3 or 4 nodes were analyzed in directed DEG-derived and control networks as well. Analysis found motifs, which were enriched in directed DEG-derived but were absent from control networks or vice versa. Some were already generally described in biological systems like convergent, divergent and bifan motifs. An interesting result of motif analysis is the enrichment of feedback loops in merged networks (PDI + PPI), but not in simple ones. A motif was enriched in DEG-derived merged network, which consists of a feedback loop and all nodes of the loop are controlled by another separated node.

Controller sub-network construction

Both lesional and non-lesional skin areas can be found on patients at the same time. We wanted to highlight nodes which may be important in the “all or none” switch in lesional skin areas and sustain this phenotype for a long time. To do this, we considered the following: it is reported, that positive feedback loops have fundamental role in maintaining autoimmune and autoinflammatory disease states; it has also been argued that hubs can be found in much more positive feedback loops than negative ones; this is published that in biological systems, interlinked slow and fast positive feedback loops allow systems to convert graded inputs (like several environmental and genetic factors in a psoriatic individual) into decisive all or none

outputs (like lesional skin phenotype). Enrichment of feedback loops also suggests the central role of feedback in lesional skin. In order to find the most important positive feedback loops containing 2, 3 or 4 nodes, a merged PPI-PDI network was constructed from proteins with centralities above cutoff value. All feedback loops were identified. All positive feedback loops had common nodes, thus a merged network was generated containing interlinked positive feedback loops. Transcriptional changes of all nodes and influence of all edges maintained the actual state in this sub-network, which has been proven by Boolean analysis.

Analysis of chemical-protein interaction networks

Undirected and directed chemical-protein interaction networks were constructed using STITCH database, which contained interactions between proteins and chemical compounds. Drugs or potential drugs were filtered out from chemicals and ranked by degree and/or stress centralities. Top ranked drugs were grouped into Anatomical Therapeutic Chemical (ATC) classes. Best rated drugs were retinoic acid, cholecalciferol, costicosteroids, methotrexate, sirolimus and tacrolimus, which can be already found in psoriasis guidelines and large clinical trials have proved their effectiveness.

Psoriasis studies are available for numerous potential drugs with high centralities. “Blood glucose lowering drugs” are promising drug candidates. Many studies are available about “Thiazolidinedione” antidiabetic drugs. The “HMG CoA reductase inhibitor” drug simvastatin was effective in a pilot study, although atorvastatin in the same class showed only a non-significant improvement in a different study. The “Antineoplastic agent” methotrexate is a well-known medication for psoriasis. Several additional drugs in the same class were found in our analysis. Mahonia aquafolium extract - containing berberine among others - is not classified into ATC classes, but three clinical trials already indicated improvement of psoriasis with the use of this substance.

The efficacy of several drugs in results are supposed by in vitro experiments, like the “Lipid modifying agent” clofibrate, fluvastatin and pravastatin, novel COX2 inhibitors, N-acetyl-cysteine,the histone – deacetylase inhibitor trichostatin A, arsenic compounds, the phosphodiesterase inhibitor rolipram and the natural polyphenolic compound rottlerin. Case reports are available about psoriasis induction by clonidine, two “agents acting on the renin-angiotensin system” like captopril or losartan; the “protein kinase inhibitor” and “antineoplastic agent” imatinib; diclofenac, olanzapine, fluoxetine and chloroquine. Also, case reports are available about the beneficial effects of ritonavir; “antineoplastic agents” like cytarabine,

doxorubicin, and cysplatin; gefitinib, colchicine, lidocaine and nicotine. In summary, studies are available for 34 drugs, experimental evidence is available for 24 drugs, case reports suggest beneficial or disease-inductor effect of 21 drugs and we also found 98 unpublished drug candidates for the treatment of psoriasis.

Effective drugs predominantly act on proteins of the controller sub-network

Targets of the 32 effective drugs in the "Studies available" group were analyzed. All target proteins got an in-degree value reflecting the number of effective drugs acting on it. The group of proteins forming the controller sub-network was compared with the group of other targets. The controller sub-network protein group got significantly higher median value than the other one.

Discussion

In vitro modelling vs. biopsy specimens

We used in vitro modelling in our *C. albicans* study. We hypothesized, that although hyphal transition is an obligate step during pathogenesis, many other genes and processes are needed for virulence and, thus, hyphae growing in the presence of human cells may be markedly different from control hyphae, which is triggered only by physical environmental factors. One of the biggest challenge was to model VVC in vitro trustworthily.

Vaginal fluid creates a pH of approximately 4.5. However, lactic acid concentration and pH similar to that of the vaginal fluid greatly inhibited cell division and germ tube formation of *C. albicans* in previous reports. Thus, in our experimental model, *C. albicans* was cultured in CKM at pH 8.0. Culturing medium contained glucose at 5.6 mM concentration. Glucose concentration of the vaginal fluid contains ~5.2 mM glucose as a final concentration. Noteworthy, 0.1% (w/v) glucose (5.6 mM) strongly induced hypha development of *C. albicans* on solid media. Starvation to glucose may be one factor that drives the yeast to hyphae transition of *C. albicans* in our in vitro system. Additionally, we used the temperature of the human body (37 C) in our experiments, which also induces filamentation. Although we tried to consider as many factors as possible during in vitro modelling, it had some limitations. First, hyphal growth of *C. albicans* is regulated by other microorganisms in the vaginal microbiome with yet unknown mechanisms. Additionally, we do not know the concentration of *C. albicans* cells relative to human cells in vivo, thus we used only empirical MOI. We used only one human cell type in our in vitro system, although there are numerous cell types in vivo. Though these

drawbacks, we were able to find important factors in the fungal-host interaction with the use of adequate controls (control hyphae and yeast).

Several microarray studies were available for our in silico analysis of psoriasis. All of them assessed gene expression of punch biopsy specimens. Biopsy specimens are suitable for the investigation of complex diseases like psoriasis. Although the phenotype can be analyzed with its own complexity, there are also drawbacks available. First, in case of in vitro modelling, standardization is relatively easy, but in case of biopsy specimens it is cumbersome. We can select patients of the same age, gender, disease severity, treatment, but there can be numerous personal variations, which may have significant influence on gene expression. The goal of microarray meta-analysis is to avoid these variations and find uniform expression patterns among studies. Second, cellular-level gene expression analysis is difficult. There are in silico methods, which are able to determine cell-specific expression data from high throughput data of tissues, but the resolution of results is low and it is sometimes biased. Single cell RNA sequencing is promising, but the dissection of different cell types from tissue specimens is problematic.

Filtering data and identification of important genes

As it could be seen from our two studies, filtering, normalization and the use of biostatistics is essential in high throughput gene expression studies. Although the price of RNA-seq is getting lower, sometimes it is still too expensive to sequence the whole transcriptome separately of three or more biological replicates. As our goal was to identify proteins or processes consisting of several proteins potentially important for the virulence of *C. albicans*, we carried out RNA-seq for the pooled RNA of three biological replicates. The resultant data was capable for the identification of proteins or processes and the findings could be further explained and validated with other in vitro methods. As we had only pooled gene expression data for each experimental scenario, special statistical inference was needed for the detection of differentially expressed genes and most important results had to be validated with quantitative real time PCR using biological replicates. Numerous genes with a known role in pathogenesis have been detected in DEGs, but we focused on GlcNAc metabolism and investigated its potential role in adhesion.

The purpose of a meta-analysis is to create general assumptions without the bias caused by geographical, ethnical, gender, study-design or age-related differences. The analysis process for all microarray study were standardized and strict requirements were set for all of them. They

had to be carried out on the same chip or chip family, and be suited to MIAME. We made the normalization and filtering process with the same methods for all studies. Quality assessment of all individual samples in a microarray study was carried out with ArrayQualityMetrics and all low-quality samples were excluded. Low-quality or outlier studies were further filtered with MetaQC. We wanted to construct whole networks from DEGs and make inference about network dynamics, thus, the selection of fold-change cutoff for DEGs was critical. Fold change cutoff selection is usually an empirical process. The same cutoff is chosen for all genes, although the expression range is affected by their function and network centralities. We used the lower fold change cutoff of 1.5 instead of 2 to avoid the exclusion of potentially important proteins and genes. Noteworthy, lot of hubs in our analysis with highest centralities had lower, than 2 absolute fold change value, which can be caused by the negative correlation between degree and gene expression fold change values. The ideal method would be the selection of cutoff values for each gene individually considering their range of expression, which is already registered in other studies.

We generated PPI networks based on the largest PPI database (STRING) available, which contains experimentally proven interactions as well as highly reliable ones based on prediction algorithms or data mining. PDI network was also generated using not only literally proven interactions but interactions predicted by high fidelity algorithms. The selection of nodes with highest centralities wasn't straightforward too. Degree and stress node centralities represent power law distribution, thus, special statistical considerations were needed. The use of lower DEG fold-change cutoff and detailed analysis based on node centrality statistics made it possible to identify proteins yet not associated with the disease but may have remarkable role in its pathogenesis.

Putting genes in context

Changes in the microenvironment massively induced many components of several different signal transduction pathways resulted in the morphological transition of *C. albicans* both in model and control conditions, but the GlcNAc transporter NGT1 was induced significantly exclusively in response to vaginal epithelial cells. We also found that the hxk1Δ mutant exhibits reduced cytotoxicity compared to the wild type strain of *C. albicans*, which may be caused by its decreased adherence to the surface of vaginal epithelial cells. The human extracellular matrix contains a significant amount of GlcNAc. In agreement with a recent review, GlcNAc released from the extracellular matrix of human cells during membrane

remodeling might explain the induction of *C. albicans* GlcNAc catabolic genes, such as, NGT1, HXK1, NAG1, and DAC1 by vaginal epithelial cells.

Adherence to the surface of epithelial and endothelial cells and penetration of hyphae into these cells are important virulence factors contributing to the pathogenesis of *C. albicans*. The GlcNAc biosynthesis plays a key role in chitin biosynthesis. The inner layer of the cell wall of *C. albicans* consists of chitin (polymers of β-(1,3)-glucan, β-(1,6)-glucan, and GlcNAc). This scaffold binds glycosylphosphatidylinositol- (GPI-) anchor-dependent cell wall proteins, which play an important role in the adherence of *C. albicans* to the epithelial cells. This suggest, that decreased adherence in hxk1Δ could be caused by decreased chitin synthesis. This theory is supported by a study, in which Nikkomycin Z, a chitin biosynthesis inhibitor caused reduced adherence of *C. albicans* to the surface of buccal epithelial cells.

Keratinocyte hyperproliferation and inhibition of apoptosis are well-known phenomena in psoriasis and several proteins have been associated with them. Most of them could be found in the group of central proteins detected by DEG-derived network analysis. Unpublished DEG-coded proteins with potential role in hyperproliferation, like CCNA2, TFDP1 and MECOM, were also found. Psoriasis is an immune-mediated disease. Many proteins published in association with the immunopathogenesis of psoriasis were highly ranked hubs in PPI networks. The downregulation of the src kinase FYN seems to be a counteracting compensatory mechanism as this protein is important in IFN gamma action, in TNF alpha induced COX2 expression and in adipose tissue - mediated inflammation leading to insulin resistance. These processes are important in the pathomechanism of psoriasis. An important node in controller sub-network is IL8. Although its role in psoriasis pathogenesis is reported, no trial has been done with IL8 inhibitors. This is true for CCL2 and IRF1 as well. Our study confirms their basic role in the sustainment of lesional phenotype as both can be found in highly ranked hubs. CCL2 is also essential in the controller sub – network by activating two positive feedback loops related to inflammation.

Comorbidity of psoriasis and metabolic syndrome is a well-known phenomenon. There is a complicated interaction between the two diseases mediated by inflammatory cytokines among other factors. Numerous DEG-coded proteins associated with both diseases could be found in central proteins like PPARG, INS-IGF2, LEP etc. Others, like PIK3R1, AR and MEF2A may have role in the development of metabolic syndrome in psoriatic patients.

The real purpose of systems biology is to construct models from high throughput data. High precision data is needed for reliability. We used data of gene expression in our psoriasis study. We didn't measure protein abundance and additionally we didn't have expression data for different time points. Thus, an accurate model with differential equations could not be constructed. However, we were able to detect a sub-network with the use of network motif analysis, which could potentially induce and maintain the lesional phenotype of the disease.

Therapeutic aspects

Chemical – protein interaction networks were created using STITCH database and we managed to predict disease – modifying drugs. Many drugs, which are already widely used in the treatment of psoriasis could be found as highly ranked nodes in chemical-protein interaction networks. Our results suggested, that the fusion of different intracellular networks with chemical-protein interaction networks can be an effective method for the detection of potentially effective drugs in the treatment of psoriasis. Our intracellular networks were constructed with the use of DEGs, which characterize the lesional phenotype in itself. The integration of drug-protein interaction data in this system seemed to be a powerful tool for drug discovery of psoriasis. The rather, that other research groups have proven the efficacy of some predicted drugs since the publication of our results.

The place of systems biology in dermatological research

Large amount of "omics" data was generated in the past decades in parallel with the improvement of screening technologies. As our VVC study showed, virulence factors can be identified with the use of experimental models and carrying out transcriptome analysis. New and important findings can be acquired by the analysis of published data too. We were able not only to identify new proteins in the pathogenesis of psoriasis, but to find new therapeutic options for the disease. We have demonstrated two examples, how large scale data of multifactorial dermatological diseases can be used. The use of large datasets, analyzing them with the tools of bioinformatics and biostatistics is inevitable in this modern era of biomedical research. We have become able to understand complex processes, construct models of diseases and treat them more effectively.

Acknowledgement

Firstly, I would like to express my sincere gratitude to my advisors, Prof. Dr. Lajos Kemény and Dr. Lóránt Lakatos for the continuous support of my PhD study and related research, for their patience, motivation, and immense knowledge.

My sincere thanks also goes to Dr. Éva Kondorosi and Dr. István Nagy, who gave access to the most modern laboratory and research facilities for *C. albicans* studies in Biological Research Centre, Szeged. Without their precious support it would not be possible to conduct this research.

I thank my fellow labmates and colleagues for the stimulating discussions and for all the fun we have had. I am grateful to Dr. Csaba Pál for giving support for my research last year.

Last but not the least, I would like to say thanks to my family and especially to my fiancée, Melinda for supporting me throughout my life in general.

Publications directly related to the subject of the thesis

I. **Manczinger M**, Bocsik A, Kocsis GF, Voros A, Hegedus Z, et al. (2015) The Absence of N-Acetyl-D-glucosamine Causes Attenuation of Virulence of Candida albicans upon Interaction with Vaginal Epithelial Cells In Vitro. Biomed Res Int 2015: 398045. **IF: 1.579**

II. **Manczinger M**, Kemeny L (2013) Novel factors in the pathogenesis of psoriasis and potential drug candidates are found with systems biology approach. PLoS One 8: e80751. **IF: 3.23**

Other publications

I. Guban B, Vas K, Balog Z, **Manczinger M**, Bebes A, et al. (2015) Abnormal regulation of fibronectin production by fibroblasts in psoriasis. Br J Dermatol. **IF: 4.225**

II. Palotai M, Bagosi Z, Jaszberenyi M, Csabafi K, Dochnal R, **Manczinger M**, et al. (2013) Ghrelin and nicotine stimulate equally the dopamine release in the rat amygdala. Neurochem Res 38: 1989-1995. **IF: 2.551**

III. Palotai M, Bagosi Z, Jaszberenyi M, Csabafi K, Dochnal R, **Manczinger M**, et al. (2013) Ghrelin amplifies the nicotine-induced dopamine release in the rat striatum. Neurochem Int 63: 239-243. **IF: 2.650**

IV. Heinzlmann A, Kiss G, Toth ZE, Dochnal R, Pal A, Sipos I, **Manczinger M,** et al. (2012) Intranasal application of secretin, similarly to intracerebroventricular administration, influences the motor behavior of mice probably through specific receptors. J Mol Neurosci 48: 558-564. **IF: 2.891**

V. **Manczinger M**, Szabo EZ, Goblos A, Kemeny L, Lakatos L (2012) Switching on RNA silencing suppressor activity by restoring argonaute binding to a viral protein. J Virol 86: 8324-8327. **IF: 5.076**

VI. Koves K, Kiss G, Heinzlmann A, Dochnal R, **Manczinger M**, et al. (2011) Secretin attenuates the hereditary repetitive hyperactive movements in a mouse model. J Mol Neurosci 43: 109-114. **IF: 2.504**