Oana Alina Tomescu, MSc

Integrative Analysis of Omics Data

Enhancement of Existing Methods and Development of a Novel Gene Set Enrichment Approach

Bioinformatics
Institute for Knowledge Discovery
Graz University of Technology, Austria

DOCTORAL THESIS

to achieve the university degree of

Doktorin der technischen Wissenschaften

submitted to

Graz University of Technology

Advisors:

Gerhard G. Thallinger
Aèdin C. Culhane

Reviewers:

Rudolf Stollberger
Diethard Mattanovich
Zlatko Trajanoski

Examiners:

Rudolf Stollberger
Diethard Mattanovich

Graz, May 2015

AFFIDAVID[1]

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. The document uploaded into TUGRAZonline is identical to the present dissertation.

Graz, ____________________________ _________________________________

Date Signature

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Acknowledgments

"There are no shortcuts in evolution."
Louis D. Brandeis

Finishing my doctoral studies is the hardest thing I have ever done. It was challenging, frustrating and tough but at the same time it was an incredible experience full of making mistakes and learning from them, meeting and working with outstanding scientists and discovering that I want to dedicate my life to doing good science.

First and foremost, I would like say thank you to my supervisor Dr. Gerhard Thallinger. I am very grateful for the opportunity to work in his group, for his guidance and his support. I would like to say thank you to Prof. Diethard Mattanovich and the whole team working on ACIB Project 3.5 for the discussions and suggestions during our meetings in Vienna. I express my gratitude to Prof. Rudolf Stollberger for accepting to be the official supervisor of this thesis. A special thank you goes to Dr. Aedin Culhane for her mentorship during my stay in Boston as well as afterwards and for the opportunity to experience being part of a new research group and living in a new country. I would also like to thank you our collaborators for their assistance and for providing the data sets analyzed in this thesis. I say thank you to all my colleagues for their help, intriguing discussions and the numerous lunch breaks we spent together.

The past five years have been tough not only for me but also for everybody else in my life. I am grateful to my parents for making all this possible in the first place and to my family and friends for always believing in me. A special thank you goes to Karl Schumacher who was the first one to point me towards bioinformatics more than 10 years ago. The person I owe most thank-yous is my best friend and partner, Roman. He was always there for me, supported me every step of the way, motivated me when I most needed it and helped me whenever possible. A special thank you for Schnurrbärt. You are my life and none of this would make sense without you!

Abstract

As high-throughput data measurements in biomedical studies have become routine, the challenges shifted from data generation to data analysis. In particular, the integration of multiple omics data sets is an intriguing but ambitious task.

This thesis comprised three specific aims. First, integrative analysis methods shall be applied to multiple omics data sets. Second, existing integrative analysis methods shall be investigated. Third, the development of a novel integrative pathway enrichment approach (IPEA).

For the first aim, (multiple) co-inertia analysis (MCIA) was applied to *A. gambiae* and to the cross-species comparison of the expression systems *P. pastoris* and Chinese hamster ovary (CHO) cells. In the former case, a high structural concordance between the hemocyte transcriptome and the granulocytic proteome could be shown. In the latter case, a number of secretion and ribosome relevant target genes and proteins were identified by a detailed characterization of four production strains.

For the second aim, three integrative analysis methods (MCIA, generalized singular value decomposition (GSVD) and integrative biclustering (IBC)) were applied to the transcriptome and proteome of the parasite *P. falciparum.* From the intersection of these results, a network of biological processes was derived which characterizes the parasite's life cycle stages and unifies numerous findings from the past 25 years of research in a single analysis. Additionally, a traditional gene set enrichment analysis (GSEA) was applied to validated target genes of two sets of human microRNAs. The 36, respectively 35, enriched neuron related biological processes were almost identical between the two sets, although the overlap in the corresponding miRNA lists was below 50%.

For the third aim, in order to overcome flat gene list limitations of the traditional GSEA, we developed a novel integrative pathway enrichment analysis (IPEA). Our IPEA approach combines scores from a multivariate analysis with pathway specific scores based on network topology. Enriched pathways computed by IPEA are characterized by biologically relevant concordance between the measured data and the intrinsic structure of the pathways. IPEA visualizes the results as a double bipartite graph of activated features and enriched pathways. Applied to 38 matched tumor and stroma samples from ovarian cancer patients, IPEA reveals an unprecedented view of the cross-talk between tumor and stroma, suggesting new targets for the treatment of ovarian cancer, e.g. CTNNB1, ERBB4 and SMAD4 which have already shown their potential in the therapy of other cancers.

List of Publications

This thesis is based on articles published in peer-reviewed journals, on manuscripts that are currently under review and on unpublished work. The full text of the articles and manuscripts is included in this thesis starting with page 134.

Tomescu OA, Mattanovich D and Thallinger GG: Integrative analysis of omics data: A method comparison. *Biomed Tech 2013.* 58:Suppl 1.

Patz S, Trattnig C, Grünbacher G, Ebner B, Gülly C, Novak A, Rinner B, Leitinger G, Absenger M, **Tomescu OA**, Thallinger GG, Fasching U, Wissa S, Archelos-Garcia J and Schäfer U: More than cell dust: microparticles isolated from cerebrospinal fluid of brain injured patients are messengers carrying mRNAs, miRNAs, and proteins. *Journal of Neurotrauma* 2013. 30(14):1232–1242

Tomescu OA, Mattanovich D and Thallinger GG: Integrative omics analysis. A study based on *P. falciparum* mRNA and protein data. *BMC Systems Biology* 2014. 8(Suppl 2):S4

Smith RC[*], King JG[*], Tao D[**], **Tomescu OA**[**], Brando C, Thallinger GG and Dinglasan RR: Proteomic analysis of mosquito macrophage-like blood cells reveals an anticipatory innate immune response in the absence of malaria parasite challenge. *PLoS Pathogens* 2015. In preparation.

Meng C[*], **Tomescu OA**[*], Thallinger GG, Gholami AM and Culhane AC: Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics* 2015. Under review.

[*],[**] Authors contributed equally.

Contents

1 Introduction

The ultimate goal of science is the understanding of the world by discovering the underlying laws that govern it. This represents a complex and challenging endeavor. Human kind has achieved major breakthroughs but there still remain various unanswered questions. For example, physicists can explain the world of large objects like planets or the universe with Einstein's theory of relativity but if they have to explain the world of atoms and molecules they need the quantum theory. Is there a theory that is able to unify these two? Light is another well known example: Sometimes it is considered to be a weave and other times a particle. Even if these examples point to unanswered questions, they illustrate the need to observe a system under various conditions in order to completely understand it.

The same effect governs biology. According to the central dogma of molecular biology, in order to understand an organism as a whole, one has to have knowledge about at least three levels of abstraction: deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins. Only by integrating these three data types it is possible to better understand the organism under study. And these are probably the minimal requirements: detailed questions can only be addressed by tailored measurements on specific levels of abstraction.

Until not so long ago it was not possible to characterize a biological system on different levels. The experiments that would have been needed for it were time consuming and expensive. Due to this, each research group focused on one kind of omics data. This approach was the best at that time as it was the only one available but it is similar to trying to understand how a car functions by analyzing only the motor or only the braking system.

Integrative analysis, as it is understood by bioinformaticians today, refers to the process of combining data which originated from diverse sources, such as different subjects, species, tissues and cells; various levels of regulation including DNA, RNA, proteins, metabolites and epigenomic patterns; different experimental platforms, such as Agilent and Affymetrix or multiple time points. The purpose of integrating multiple data sets is to reveal more information than the sequential traditional analysis of the same data sets does.

Integrative analysis is a rapidly growing research field today. This is due to the unprecedented wealth of available data which is a result of technological improvements and, at the same time, dropping costs of experiments. While at the end of the year 2000, according to [1], there were only 1760 published articles on integrative analysis, in September 2014 the number exploded to 18500 publications.

The goal of an integrative analysis is knowledge discovery on one hand and data exploitation on the other hand. Some of the methods used provide also the opportunity of data visualization which promotes the overall understanding of the problem.

Integrative Analysis or the Story of the Blind Men and the Elephant

To emphasize the indispensability of integrative analysis I would like to bring to the reader's attention The Story of the Blind Men and the Elephant (see Figure 1.1). This story originates in the Indian culture where different versions are known and was introduced to the western world by the American poet John Godfrey Saxe I. In this story a king sends a group of blind men to touch an elephant and to describe what it feels like. Depending on what part of the elephant was examined, the men arrive to different conclusions: it feels like a wall, a spear, a snake or a tree. The moral of the story is the need for different observations which only together are able to correctly describe the whole system. The situation is similar in molecular biology: the complete understanding of an organism is based on measurements done on different layers of regulation such as DNA and RNA.

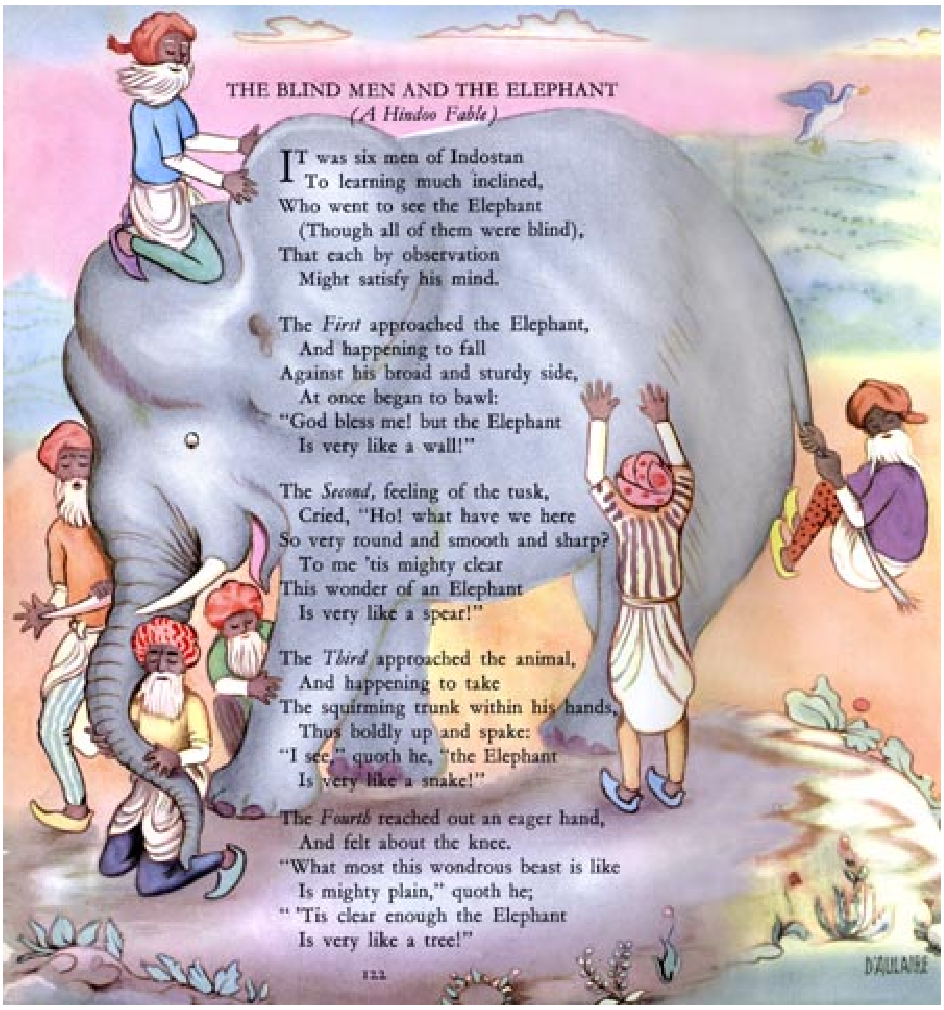
Currently it is possible to observe and measure biological systems on many different levels, such as DNA, RNA, protein or metabolite level. If only one of these levels is considered, the researcher's conclusions are similar to those of the blind men in Saxe's poem. Only by integrating more and more levels, the derived knowledge about the system mirrors more and more the biological truth and will eventually lead to the complete understanding of the biological system.

In order to have different point of views or various observations of the system under study one has to have access to the corresponding data sets. These sets can be either publicly available or they must be generated within the study. As mentioned above, one driving force of the development of integrative data analysis is the increasing amount of available data sets. These sets would not be as abundant if the technology needed to generate them would not have developed so fast and the corresponding costs would not have dropped as quickly. Section 1.1 introduces the different types of data that are currently used in integrative analyses while section 1.2 provides an overview of various types of integrative analysis methods. Section 1.3 summarizes the objectives of this doctoral thesis.

1.1 Omics Data

Omics data refers to data generated by omics technologies such as gen*omics*, transcript*omics*, prote*omics* and metabol*omics*. These technologies received their names due to their study of the gene*ome*, transcript*ome*, prote*ome* or metabol*ome*. The suffix *ome* is used in molecular biology to form nouns having the meaning "all constituents considered collectively" [2].

The different types of omics data are presented in the following subsections. Genomics is the first data type that is introduced followed by transcriptomics, proteomics and a short overview of other omics data. The descriptions of the different data types are presented in the context of molecular biology history, emphasizing the most important events that led to the research filed as we know it today.



THE BLIND MEN AND THE ELEPHANT
(A Hindoo Fable)

IT was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

The *First* approached the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me! but the Elephant
Is very like a wall!"

The *Second*, feeling of the tusk,
Cried, "Ho! what have we here
So very round and smooth and sharp?
To me 'tis mighty clear
This wonder of an Elephant
Is very like a spear!"

The *Third* approached the animal,
And happening to take
The squirming trunk within his hands,
Thus boldly up and spake:
"I see," quoth he, "the Elephant
Is very like a snake!"

The *Fourth* reached out an eager hand,
And felt about the knee.
"What most this wondrous beast is like
Is mighty plain," quoth he;
" 'Tis clear enough the Elephant
Is very like a tree!"

122

Figure 1.1: The story of the blind men and the elephant. Illustration by D'Aulaire and poem by John Godfrey Saxe I.

1.1.1 Genomics

According to the World Helth Organisation (WHO) [3]

> *"Genomics is the study of the structure and action of the genome, i.e. the sum total of genetic material present in an organism. This includes both the DNA present in the chromosomes and that in subcellular organelles (e.g. mitochondria or chloroplasts). It also includes the RNA genomes of some viruse*s".

The first experiments on what we call today genes were performed by the father of modern genetics Gregor Mendel. As a monk he used the monastery gardens to conduct experiments in which he crossed various pea plants with different colors, shapes and heights. He observed [4] that traits are passed down to the children and children's children in a predictable way through, what today are called, genes.

The next important milestone was the discovery of DNA by Friedrich Miescher in 1869 [5]. Unfortunately the did not know that the new molecule he had isolated from white blood cells, which contained hydrogen, oxygen as well as a stable phosphorus to nitogen proportion and which he called "nuclein" was actually the DNA.

In 1952 Rosalind Franklin used X-ray crystallography to study DNA structure. She took pictures of crystallized DNA fibers with phosphates on the outside of what appeared to be a helical structure. She published her findings [6] together with the famous "photograph 51" (see Figure 1.2) in the same issue of Nature [7] where Watson and Crick presented their 3D model of the DNA.

After various hypotheses regarding the structure of the DNA, such as the three chains model of Pauling and Corby [8], Watson and Crick proposed their 3D model for the DNA [7] as we know it today: double helix structure with antiparallel strands; sugars and phosphates on the outside; paired bases on the inside with hydrogen bounds linking adenine (A) to thymine (T) and cytosine (C) to guanine (G). Additionally, they also noticed that the specific pairing suggested the existence of a specific copying mechanism for the DNA.

The next step in the development of our knowledge about genes was the understanding of protein synthesis from RNA. In 1961 Marshall Nirenberg designed an experiment in which synthetic mRNA containing exclusively uracil (U), a base encountered only in the RNA, was added to a cell-free *Escherichia coli* extract including DNA, RNA, ribosomes and other machinery for protein synthesis. Deoxyribonuclease (DNase) was added to brake down the DNA and to ensure that only the synthetic poli-U mRNA was used for protein synthesis. By radioactive labeled amino acids they discovered [9] that the genetic code (see Figure 1.3) for phenylalanine was UUU (three consecutive uracil bases). This was the starting point for elucidating the other codes on which protein synthesis is based.

The next mystery waiting to be solved was the base sequence in the DNA. In 1975 Sanger *et al.* [10] proposed a method in which the DNA was denaturated through exposure to high temperatures which leads to the separation of the two strands. His procedure continues with four parallel and similar steps in which polymerase and dideoxynucleotides triphosphates (ddNTP) are added to the mixture. In each of the four parallel processes a different chain-inhibitor of the DNA polymerase is used: ddGTP, ddATP, ddTTP and ddCTP; one for each base. All ddNTP lack the 3′-OH group leading to the termination of the elongation process. In this way each of the four parallel processes yields sequences ending in the same base. In order to read the sequenced DNA piece one has to use electrophoresis. The method was published [11] in the same year as Sanger *et al.* sequenced the bacteriophage Φ X174 [10] followed by the bacteriophage λ [12] in 1982.

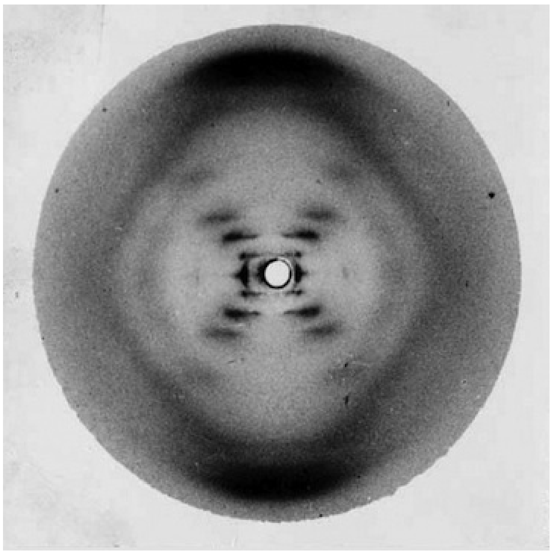


Figure 1.2: Photograph 51 by Rosalind Franklin showing an X-ray image of the DNA. Figure published in [6]

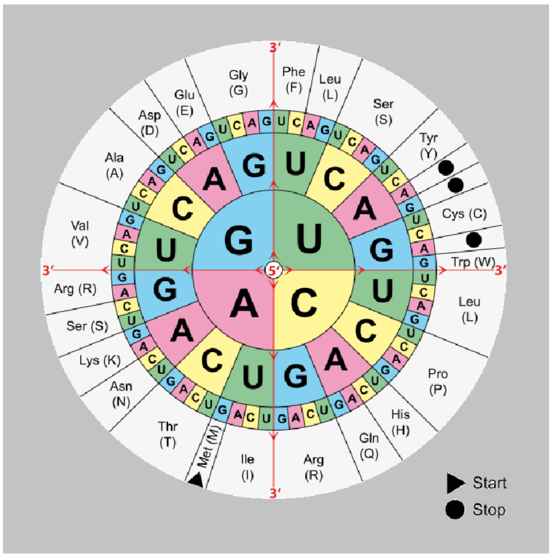


Figure 1.3: Aminoacids table. The direction of reading for the genetic code of the proteins starts at 5′ and goes to 3′. Public domain figure.

Another key tool for molecular biology is the polymerase chain reaction (PCR). Developed in 1983 by Mullis Kary, the PCR [13] is used to amplify the DNA *in vitro*. The chain reaction refers to the cyclic structure of the amplification by using the product of one round as the starting point for the next amplification cycle. This also implies the exponential nature of the reaction. Today, PCR is a widely used technique for: diagnosis of genetic diseases; identification of viruses and bacteria and validation of genetic fingerprints.

Almost 20 years later, Fleischmann *et al.* sequenced the first free living organism *Haemophilus influenza Rd.* [14] which marked the beginning of the omics era. This is also the moment when molecular biology started to change from a data poor to a data rich research field.

Sanger sequencing, with a series of enhancements, was the method of choice until the mid 2000s. Automation was probably one of the most important developments leading to the sequencing of the first human genome in 2001 within The Human Genome Project.

The Human Genome Project started in 1990 and was the result of various discussions that originated in 1984 when the US Department of Energy (DOE), the National Health Institute (NIH) and a number of international groups started discussions about the study of the human genome. Two years later a recommendation about the development of a human genome map was made by the US National Research Council. 15 years were allocated for the completion of the project and in 1990 the plan for the first years was published. A budget of 3 billion dollars was allocated. The major goals were: development of technologies to study the DNA, mapping and sequencing of the human genome and the study of the intrinsically related ethical, legal and social issues. In 2001 The Human Genome Consortium [15] (see

Figure 1.4: Cover image of the Nature and Science issues where the human genome was published. The first draft of the human genome was published simultaneously by two teams: one in Nature and one in Science

Figure 1.4) as well as Venter *et al.* [16] from Celera Genomics Corporation published, at the same time, the first draft of the human genome.

One year before the human genome was published, the joint efforts of groups at the University of California, Berkeley, and Lawrence Berkeley National Laboratory as well as Craig Ventor from Celera Genomics Corporation resulted in the report [17] of the genome sequence of the model organism fruit fly (*Drosophila melanogaster*). The fruit fly is very important as a model organism for the identification of human gene functions.

The genome of yet another important model organism, the mouse, was published [7] in 2002 by the Mouse Genome Sequencing Consortium. The mouse (*Mus musculus*) plays a very important role in the study of human disease due to the 90% similarity [7] to the human genome.

The Human Genome Project was announced to be finished in 2003. This was two and a half years before the planed end with approximately 10% of the project's budget not having been spent.

Our knowledge about the genes almost exploded compared to it's beginnings in a monastery garden where a monk crossed peas with different phenotypes. A large amount of the discovered information was concentrated in Crick's Central Dogma of Molecular Biology [18].

In his publication, Crick describes the genetic information flow in a biological system by stating that genetic information (sequential information) can not be transfered from protein

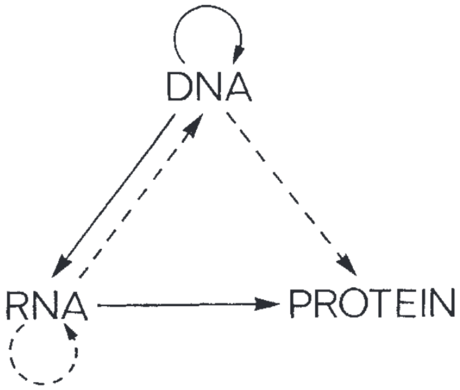


Figure 1.5: Central dogma of molecular biology. Diagram by Francis Crick as it was published in [18].

to protein or back to DNA. The article also included a diagram (see Figure 1.5) of the possible and probable direction of genetic information.

According to Crick, there are three types of possible information transfers in a biological system: general (DNA → DNA, DNA → RNA, DNA → protein), special (RNA → DNA, RNA → RNA, DNA → protein) and unknown (protein → DNA, protein → RNA, protein → protein) transfers. The general transfers were believed to normally occur in most of the cells, the special transfers were observed only under special conditions and the unknown transfers were believed to be impossible. The positive formulation of this theory is known as: "DNA makes RNA, RNA makes protein" which emphasizes the two processes that govern the protein production in a biological system: transcription and translation.

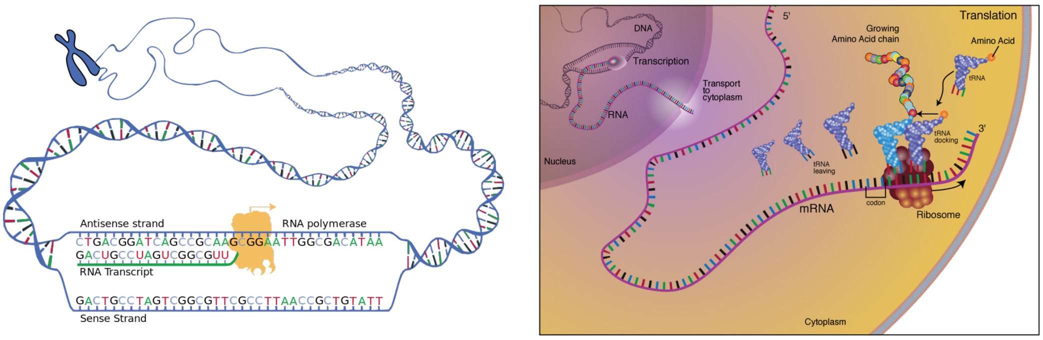


Figure 1.6: The two main processes involved in gene regulation: transcription (left) and translation (right). Public domain graphics from the National Institute of Health (www.genome.gov)

The synthesis of RNA by using DNA as a template is called transcription [19]. Through this process, in which the DNA bases A,T,C,G are translated to A,U,C and G, four types of RNA are created: messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA) and non-coding RNA (ncRNA). In case of mRNA, the process of transcription is divided in two subprocesses: synthesis and processing. A graphical representation is shown in Figure 1.6.

The synthesis of proteins based on an mRNA template is called translation [19]. A protein is

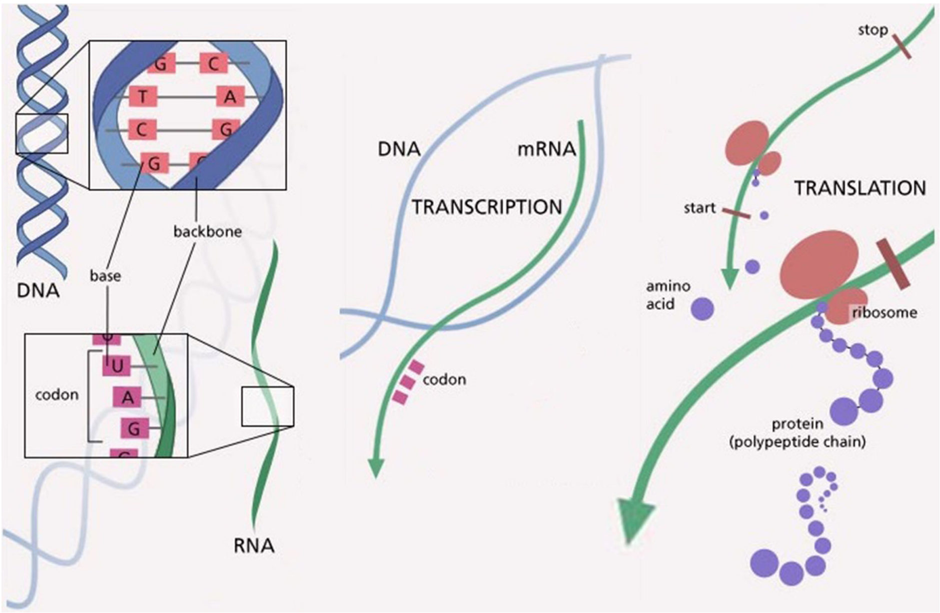


Figure 1.7: Illustration of the Central dogma of molecular biology (adapted from [20]).

created by the translation of the mRNA bases (A,U,C,G) into the corresponding sequence of amino acids of a polypeptide. This process is graphically shown in Figure 1.6.

The translation of DNA into RNA and RNA into proteins constitutes the process of gene regulation. Even if the human genome is completely sequenced, the exact functions of the genes as well as their interplay are by far not elucidated. The mechanism of gene regulation is so complex that it has to be studied for each gene or gene family separately. The most promising way would be to measure the genes of interest on all available levels (DNA, RNA and protein) and integrate theses data sets into a common analysis.

All these milestones in the history of molecular biology led to the research field as we know it today and made it possible to access large amounts of information that is needed to address current research questions.

1.1.2 Transcriptomics

Transcriptomics is the technology used to study the transcriptome which is defined by Velculescu *et al.* in [21] as

> *"the entirety of all expressed genes and their expression level for a defined population of cells."*

They also emphasize that due to the mostly static nature of the genome, as opposed to the transcriptome which changes depending on cell types, tissues and measurement time points, the transcriptome is the link between the genome of an organism and its phenotype.

Early technologies used to asses gene expression at mRNA level included: Northern blotting [22], differential display [23] or dotblot analysis [24]. One drawback shared by all of the

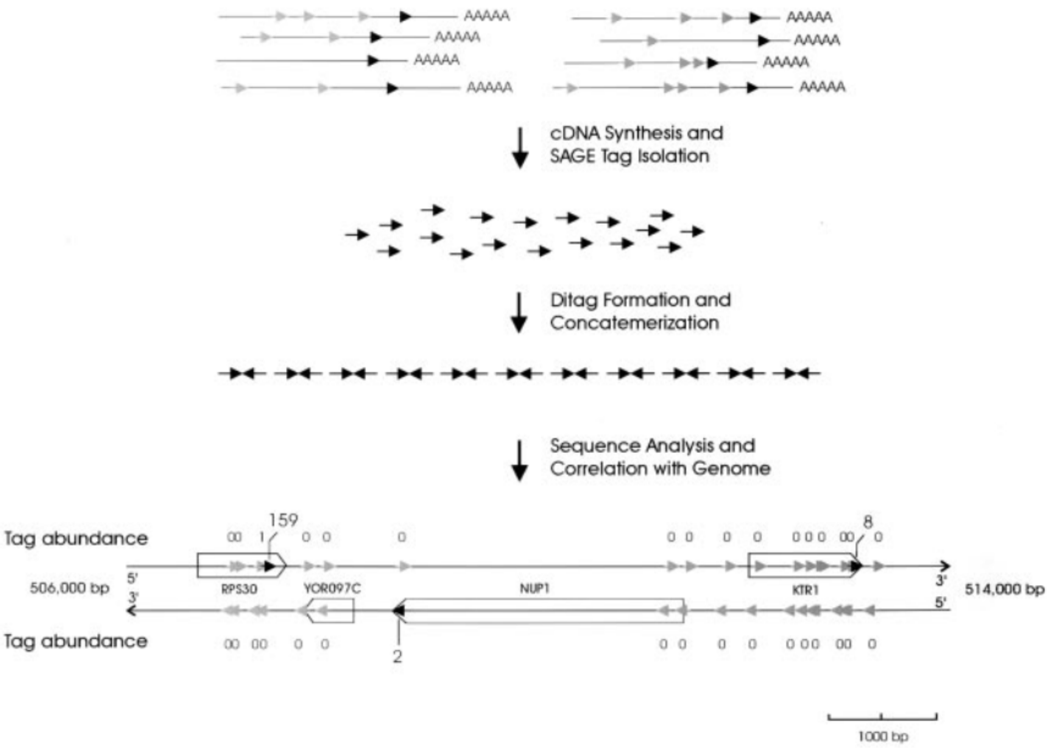

AAAAA
cDNA Synthesis and SAGE Tag Isolation
Ditag Formation and Concatemerization
Sequence Analysis and Correlation with Genome
Tag abundance
506,000 bp
514,000 bp
5'
3'
RPS30
YOR097C
NUP1
KTR1
159
8
2
1000 bp

Figure 1.8: Serial analysis of gene expression: method used for the caracterization of the first mammalian transcriptome. Figure adapted from [11]

above is their inability to measure large amounts of transcripts simultaneously which is the key requirement for transcriptome profiling.

The first mammalian transcriptome was profiled in 1991 by Craig Venter's group at NIH [25] by using serial analysis of gene expression (SAGE). It represented one of the earliest application of the Sanger sequencing method [11] and was composed of two steps as described in Figure 1.8:

> *"First, a short sequence tag (9–11 bp) is generated that contains sufficient information to identify uniquely a transcript, provided that it is derived from a defined location within that transcript. Second, many transcript tags can be concatenated into a single molecule and then sequenced, revealing the identity of multiple tags simultaneously."*

SAGE was also used to conduct a global analysis of the pancreas transcriptome [26] including 1000 manually sequenced tags.

This is the time when microarrays were born. One of the earliest publications shows the microarray analysis of *Arabidopsis thaliana* which included 48 cDNAs (complementary DNA) with an average length of 1.0 kb. Microarrays, which are based on complementary probe hybridization, developed into the method of choice for transcriptome analysis and dominated the next twenty years of molecular biology research.

According to the Glossary of Genetic Terms [27] provided by the National Human Genome Research Institute of NIH, microarrays are defined as:

> *"Microarray technology is a developing technology used to study the expression of many genes at once. It involves placing thousands of gene sequences in known locations on a glass slide called a gene chip. A sample containing DNA or RNA is placed in contact with the gene chip. Complementary base pairing between the sample and the gene sequences on the chip produces light that is measured. Areas on the chip producing light identify genes that are expressed in the sample."*

In general, microarray technology is based on the following steps: probe purification, reverse transcription of mRNA to cDNA, labeling, hybridization, washing steps, scanning of the array, normalization and analysis. Figure 1.9 provides an overview on different microarray technologies.

Microararys can be divided in spotted and in *in situ* synthesized arrays. While in spotted microarrays the probes are oligonucleotides, cDNA or PCR products that correspond to mRNAs which are synthesized and afterwards spotted onto a glass slide, in the synthesized version the probes are short sequences designed to match parts of an open reading frames (ORF) which are directly synthesized on the array surface.

Additional disjoint categories are one-channel and two-channel microarrays. In two-channel or two-color microarrays, two samples can be compared. For this, the arrays are hybridized with cDNA from the samples that were previously labeled with two fluorescent dyes. Afterwards the array is scanned with the dye-corresponding wavelengths and the ratio of the two intensities can be used to identify differentially expressed genes.

Although the name might suggest it, one-channel or one-color microarrays do not measure expression levels of a gene but rather two one-colour microarrays are used to measure ratios between two samples that were processed in the same experiment. This is at the same time an advantage of this microarray category: easier comparison of samples from different experiments. Another advantage is that an erroneous sample does not affect raw data from non-erroneous samples. Nevertheless, this technology has a disadvantage as well: compared to the two-color microarrays, twice as many arrays are needed to conduct the same experiment.

Microarrays are widely used and their applications include but are not limited to: gene expression profiling [29, 30], mutational analysis [31], drug discovery and development [32], cancer research [33–36], microbial applications [37, 38].

Microarrays dominated the research community because they stand for high throughput technology at a very reasonable price [39]. However, there are limitations that have to be taken into account: cross-hybridization can lead to high background levels which will cause erroneous data [40]; the dynamic detection range is limited by saturated and background signals; comparison between distinct microarrays requires detailed knowledge and the use of fancy normalization techniques; the most striking disadvantage being the requirement of an already existing genome sequence [39].

With the development of second-generation or next-generation sequencing (NGS) [41] alternatives to microarrays started to appear. As emphasized by Wang and colleagues in [39],

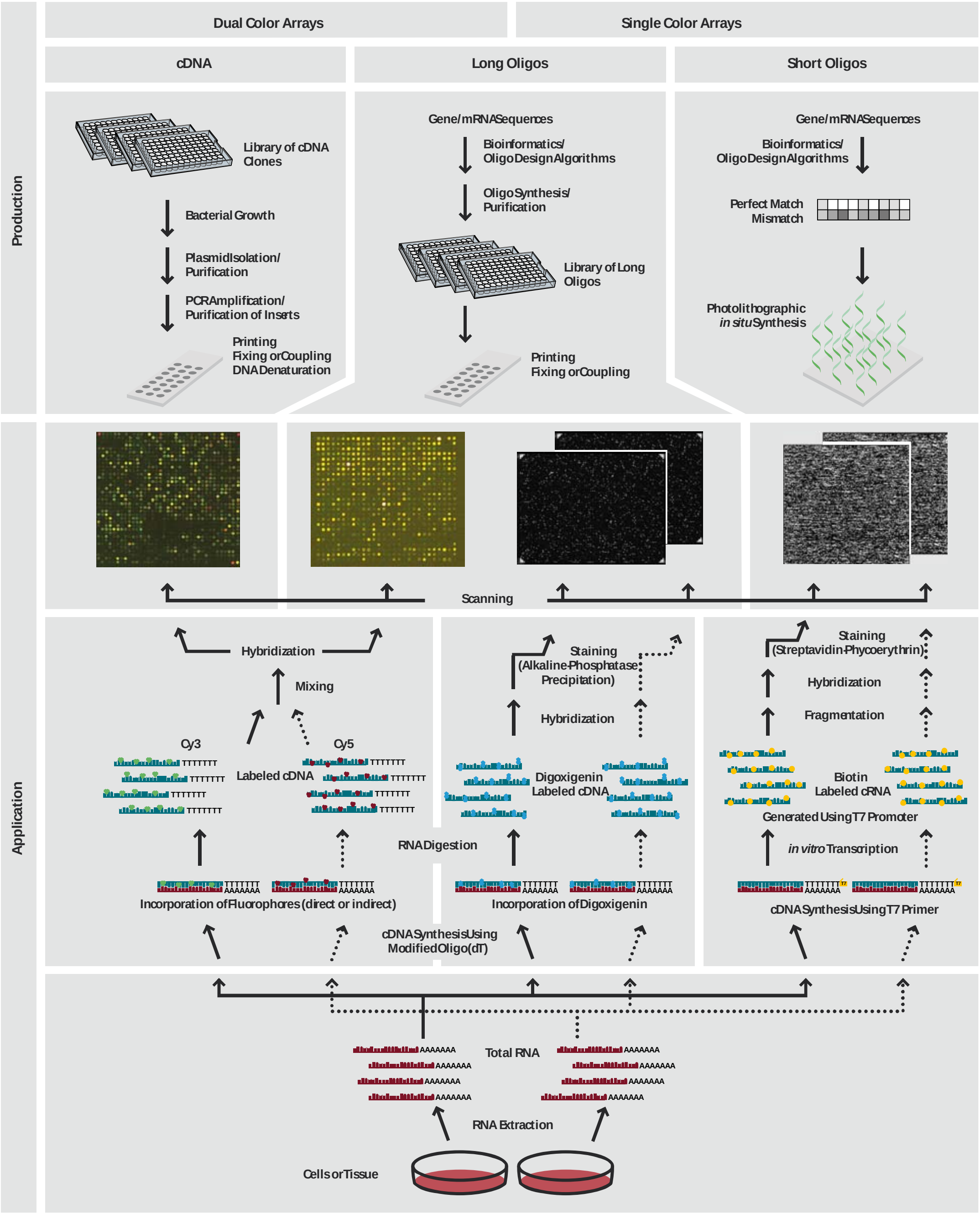

Dual Color Arrays
Single Color Arrays
cDNA
Long Oligos
Short Oligos
Production
Library of cDNA Clones
Bacterial Growth
Plasmid Isolation/ Purification
PCR Amplification/ Purification of Inserts
Printing Fixing or Coupling DNA Denaturation
Gene/mRNA Sequences
Bioinformatics/ Oligo Design Algorithms
Oligo Synthesis/ Purification
Library of Long Oligos
Printing Fixing or Coupling
Gene/mRNA Sequences
Bioinformatics/ Oligo Design Algorithms
Perfect Match Mismatch
Photolithographic in situ Synthesis
Application
Scanning
Hybridization
Mixing
Cy3
Cy5
Labeled cDNA
RNA Digestion
Incorporation of Fluorophores (direct or indirect)
Staining (Alkaline Phosphatase Precipitation)
Hybridization
Digoxigenin Labeled cDNA
Incorporation of Digoxigenin
Staining (Streptavidin-Phycoerythrin)
Hybridization
Fragmentation
Biotin Labeled cRNA
Generated Using T7 Promoter
in vitro Transcription
cDNA Synthesis Using T7 Primer
cDNA Synthesis Using Modified Oligo(dT)
Total RNA
RNA Extraction
Cells or Tissue

Figure 1.9: Microarray work flow. Figure adapted from [28].

NGS based approaches directly determine the cDNA sequence in contrast to microarray based methods that use already existing genome information.

Sanger sequencing was used for cDNA [42] and expressed sequence tag (EST) sequencing [43]. The low throughput, high costs and being non-quantitative led to the development of tag based methods such as SAGE [26], cap analysis of gene expression (CAGE) [44, 45] and massively parallel signature sequencing (MPSS) [46] which are high throughput and provide precise gene expression levels. Nevertheless, these technologies also suffer from limitations such as high costs, use of short read tags that can not be uniquely mapped to the genome as well as non-isoform specificity[39]. As a response to these demands, next generation sequencing was developed.

Generally speaking, the process of next generation sequencing can be divided into the following steps: template preparation, sequencing and imaging, and data analysis. Grada and Weinbrecht describe these steps in detail and provide additional information on this technology in [47].

An outstanding review [48] of NGS technology was written by Mardis summarizing the history of sequencing and providing a detailed list of advantages of NGS over Sanger sequencing such as: The DNA to be sequenced is used to construct a library of fragments that have synthetic and platform specific adapters covalently bound through DNA ligase making cloning unnecessary. The fragment amplification is digital and happens *in situ* on a solid surface, a bead or flat glass microfluidic channel rather than in microtiter plate wells. Sequencing and detection are simultaneous processes in NGS as opposed to Sanger sequencing. Additionally, the capacity of these steps, of hundreds of thousands of billions of reactions, enables the generation of huge data sets. Another crucial difference between the two technologies is the read length which was determined by gel-related factors in Sanger sequencing while in NGS it is a function of signal-to-noise ratio. This is specific for each NGS platform [49–51] but in general one can state that NGS produces shorter reads than Sanger sequencing. Additional information on NGS technology and platforms can be found in [48, 52–56].

Based on NGS, a new method was developed for the identification and quantification of transcriptomes: RNA-sequencing (RNA-seq) [57, 58]. Generally speaking, the work flow of RNA-seq is composed of the following steps [39]: RNA is converted to a cDNA library containing fragments with adapters attached to one or both ends; each molecule undergoes a high throughput (single- or paired-end) sequencing step resulting in 30-400bp long reads; alignment of the reads to a reference transcript or *de novo* assembly which results in a genome-scale transcription map including the transcriptional structure and the gene expression level. During the sequencing step, NGS technologies such as Illumina (formally known as Solexa) [49], Applied Biosystems SOLiD [50] and Roche 454 Life Sciences [51] are used, although Illumina seems to be the most used [59] platform.

Some of the most noteworthy advantages [39, 50, 60] of RNA-seq are: single-base level reconstruction of new and already known transcripts, broad dynamic range and reproducibility.

The applications of such a powerful technique are wide and include [60]: transcriptome profiling of non-model organisms [61, 62], model transcripts identification [63], study of

RNA modification [64, 65] and quantification of allele-specific gene expression [66].

1.1.3 Proteomics

In order to include the next level of regulation into an integrative analysis one has to interrogate not only genes but also their products: the proteins. In this way, the analysis will capture the results of transcription and translation.

The term proteomics was defined [67] as the

> *"large-scale characterization of the entire protein complement of a cell line, tissue, or organism"*

and began to be used starting with 1995 [68–70]. Nevertheless, studies that deserved the name proteomics have been conducted since 1975, when the two dimensional gel, developed by O'Farrell [71], was used in studies in which mouse [72] and guinea pig [73] protein mappings were conducted. An example of a two dimensional gel of proteins from *Bacillus subtilis* can be seen in Figure 1.1.3.

A huge limitation of the two dimensional gel was that the proteins could not be identified, just separated and visualized. One of the earliest attempts to overcome this disadvantage was the Edman degradation [74] used for the sequencing of proteins. Later, the group around Stephen Kent developed microsequencing techniques [75–77] for electroblotted proteins which represented a huge step forward.

The next major breakthrough in protein identification was the development of the Mass Spectrometry (MS) technology [78]. This breakthrough was achieved by the ability to quantify and identify proteins which was used in the study of protein interaction networks [79] and to reveal the protein composition of cellular organelles [80, 81]. Figure 1.1.3 shows a schematic view of a simple mass-spectrometer.

In general, proteomics involves the identification of proteins from a mixture. A detailed description of possible applications is given in [82]: identification of the coding sequence, computation of differential expression or further characterization such as detection of post-translational modifications. Any additional characterization is performed by MS with study dependent fractionation: electrophoretic in case of intact proteins or chromatographic for peptides. The major MS platforms currently used are *matrix-dependent laser desorption/ionization* (MALDI) and *electrospray ionization*. Downstream analysis includes protein identification through search engines like Mascot [83] which generates statistically significant peptides matches but also peptide quantification through isotope-labeling or label-free comparisons. As examples we mention here the chemical labeling [84] (iTRAQ) and stable isotope labeling with amino acids in culture (SILAC) [85].

Similar to the transcriptome and in contrast to the genome which is believed to be more or less constant, the proteome is highly variable and changes depending on time point and cell type resulting in a wide dynamic range [86]. This variability is obvious when one thinks about a caterpillar and a butterfly: they share the same genome but their appearances are distinct due to differences in the proteome. These differences are not only due to the translational process but also to post-translational modifications such as: phosphorylation, ubiquitination, methylation, acetylation, glycosylation, oxidation and nitrosylation.

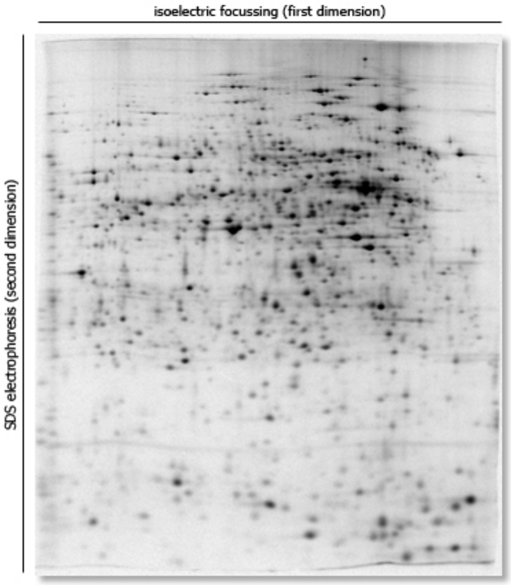

isoelectric focussing (first dimension)
SDS electrophoresis (second dimension)

Figure 1.10: Example of a two dimensional gel. Figure released under GNU Free Documentation License.

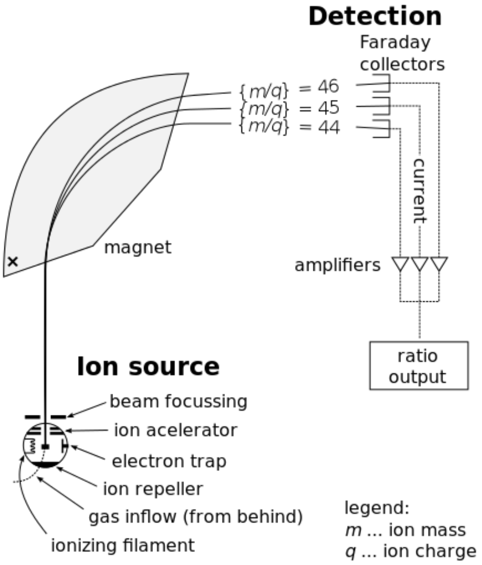


Figure 1.11: Principals of a simple mass-spectrometer. Public domain figure.

Similar to the human genome project there also exists a human proteome project (HPP). HPP is coordinated by the Human Proteome Organization and it's goal is to study all of the proteins produced by the human genome. HPP has been divided into two subprojects: the chromosome-centric HPP [87] and the biological/disease driven HPP [88].

Applications of proteomics include drug discovery such as crizotinib [89] which is successfully used in the treatment of lung cancer, biomarkers discovery for various diseases such as schizophrenia [90] or breast cancer [91] and comparative proteogenomics [92] with focus on improving gene prediction and identification of rare post-translational modifications.

Recently, two major studies of the human proteome were published [93, 94]. While Kim *et al.* report the identification of 17.294 proteins resulted from high-resolution Fourier-transform mass-spectrometry profiling of 30 histologically normal samples, Wilhelm *et al.* present a mass-spectrometry-based draft of the human proteome through the analysis of human tissues, cell lines and body fluids.

1.1.4 Other Omics Data Sets

The most well studied omics data types were presented in the previous sections. Other omics data types include, but are not limited to epigenomics, metabolomics, glycomics, kinomics, lipidomics and localizomics. An overview is shown in Figure 1.12.

Of particular interest is epigenomics. This omics type refers to the study of the epigenome which is described in [27] as follows:

"The term epigenome is derived from the Greek word epi which literally means above the genome. The epigenome consists of chemical compounds that modify, or mark, the genome in a way that tells it what to do, where to do it, and when to do it."

Similar to the other omics types, metabolomics refers to the study of the complete set of metabolites or the metabolome. This set of metabolites constitutes the response of the cell, tissue, organ or organism to the transcriptome and proteome [95].

Lipidomics refers to the study of the complete set of lipids present at a certain time point in a cell, tissue, organ or organism. Additionally, kinomics have to be mentioned which study the complete kinome (all kinases, enzymes responsible for the catalysis of phosphorylation reactions, in the genome).

Through technological improvements new technologies will emerge that will enable us to measure the complete microbiology and biochemistry of an organism. In this way an unprecedented view of a system under study will be possible.

1.2 Integrative Data Analysis

Data generation and availability is not a problem anymore. The latest technological improvements put only days between a scientist and the genome sequence, the gene expression, protein and epigenetic profile of interest. In this way, the scientist can characterize a system on different regulatory levels, compare biological systems to each other on multiple levels, investigate a subset of systems sharing a characteristic of interest on various levels or even use different platforms to measure the same regulatory level.

The challenges are shifted when high-throughput data generation becomes routine. This huge amount of data has to be managed. First is has to be properly stored - the size alone is a challenge. But the central aspect that has to be clarified is the question of interpretation. How can a scientist maximally exploit the huge amount of generated data? Individual tables and charts are not enough anymore. The current opinion on this matter is the combination of the data which leads to a detailed and more complete view of the system under study. This need for the "bigger picture", which is defined by data integration, appeared together with the first microarrays [96] more than 20 years ago. This combined analysis of multiple omics data sets is defined as *integrative data analysis.*

The integration of multiple omics data sets is a promising but at the same time challenging task [97]. The new opportunities that were made possible by integrative analysis are numerous and diverse. There are studies that integrate, for example, gene expression and methylation data [98], somatic mutations, copy number and gene expression data[99], chromatin maps and gene expression profiles [100], genotypic variation at DNA level and gene expression data [101], CHIP-seq and RNA-seq data [102], transcriptomics and proteomics data [103–105].

In the last few years different efforts were made on an international level to exploit the benefits of integrative data analysis by providing access to large collections of omics data sets. One of the first big studies on omics data was the 1000 human genomes project [106] that aimed to characterize human genetic variation. Ross *et al.* used cDNA microarrays in

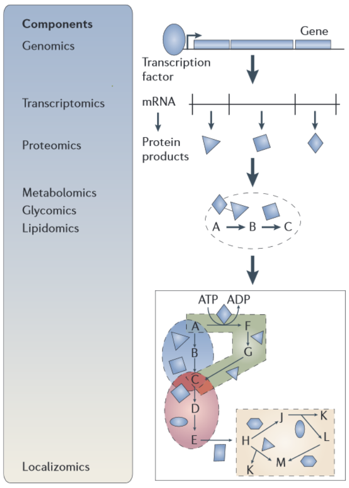


Figure 1.12: Overview of omics data. Figure adapted from [95].

[107] to measure the systematic variation among the 60 cell lines which are used by the American National Cancer Institute in anti-cancer drug studies. The next level of regulation, the human proteome, was recently characterized by two groups. While Kim *et al.* profiled 30 histologically normal human samples which included 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells and identified 17.294 proteins [93], Wilhelm *et al.* presented in [94] a draft of the human proteome through the analysis of human tissues, cell lines and body fluids. The most recent large scale omics study was done on the epigenome and integrated 111 epigenomes with additional data sets [108].

The most common application scenario is the integrative data analysis of transcriptomic and proteomic data. In this case, most analysis techniques are based on the direct correlation between transcripts and proteins such as those illustrated in [103, 104, 109–114]. [115, 116] use integrative analysis techniques that combine statistical methods with the correlation approach. As highlighted in [105, 117], the assumption of a direct correlation between transcripts and proteins is not valid in eucaryotic organisms; the reason for this being post-transcriptional and post-translational regulation. Due to this impairment alternatives were developed.

A large part of integrative analysis methods is represented by network based techniques. Piruzian *et al.* used in [118] over-connection analysis, hidden node analysis and rank aggregation to identify similarities in regulation at transcriptomic and proteomic levels, potential key transcription factors and new signaling pathways for psoriasis. Another example is presented by Tanay *et al.* in [119] and reveals modularity and organization in the yeast molecular network by the integrated analysis of highly heterogeneous genome-wide data. Patil and Nielsen uncovered in [120] the transcriptional regulation of metabolism in *S. cerevisiae* by using metabolic network topology. Perco *et al.* integrated transcriptomics and proteomics [121] on the level of protein interaction networks and amplified in this way the joint functional interpretation of the omics data sets. Another network based integrative analysis method is PANDA [122] in which messages are passed between biological networks to refine predicted interactions.

Other approaches are based on clustering. Hahne *et al.* employed analysis of variation, k-means clustering and functional annotation to the transcriptome and proteome data from salt-stressed *B. subtilis* cells and were able to show a well-coordinated induction of gene expression and changes of the protein levels as the result of a severe salt shock [123]. Verhoef *et al.* combined clustering with pathway analysis in order to characterize the changes associated with ρ-hydroxybenzoate production in the engineered *P. putida* strain S12 [124]. Biclustering is a specialized form of clustering methods adapted for integrative analysis. A binary biclustering was implemented by Gusenleitner *et al.* to identify groups of gene sets that are coordinately associated with groups of phenotypes across multiple studies [125]. Another biclustering approach was developed by Kaiser who addressed the need of an ensemble method due to the variations caused by changes in parameters and data sets of traditional biclustering methods [126].

Another large subset of the integrative analysis methods are based on singular value decomposition (SVD) such as the generalized singular value decomposition (GSVD). Alter *et al.* implemented in [127] a version of this method which allowed the simultaneous decomposition of two data sets. A generalized version which can be applied to more that two data sets is shown in [128]. The GSVD decomposes simultaneously all data sets into a product of three matrices. The decomposition is subject to the following constraint: one of the matrices has to be identical for all decompositions. Another method based on SVD is the non-negative matrix decomposition [129]. This method finds modules in the data sets that show high coordination.

The co-inertia analysis [130, 131] is also based on SVD but can be formulated in terms of the duality diagram [132, 133]. The duality diagram is a concept of the French School of Data Analysis and has rarely been presented in English. It is a very powerful concept which can be used to formulate a number of methods such as principal component analysis (PCA) [134], correspondence analysis (CA) [135], canonical correlation analysis (CCA) [136] and co-inertia analysis (CIA) in the same setup. A generalized version of CIA, which initially could be applied only two data sets, was shown by Meng *et al.*.

Other integrative analysis approaches are based on regression [138] and [139], on machine learning approaches [140], Bayesian theory [141] or logical modeling [142].

Most analyses, integrative or not, will end with a gene set enrichment (GSE) analysis that is based on gene ontology (GO) terms, pathways or functional groups. The question is how can this approach be improved and tailored for analyses in which more data sets where measured.

Several techniques were presented as solutions: PARADIGM [143] is based on statistical inference and was designed to predict the degree to which a pathway is altered in a patient; it was extended to learn subgroup-specific regulatory interactions and regulator independence [144]; Sass *et al.* developed a model-based Bayesian method [141] for inferring interpretable GO term probabilities in a modular framework; SPIA [145] was developed by Tarca *et al.* and combines the evidence obtained from the classical enrichment analysis with a novel type of evidence, which measures the actual perturbation on a given pathway under a given condition.

1.3 Objectives

The large number of integrative analysis approaches and their heterogeneity make it difficult to decide which method is best suited for a certain study. Due to this, the first goal of this doctoral thesis is a method comparison study. Subsequently, multiple co-inertia analysis was used to measure the concordance between proteomic and transcriptomic profiles of *Anopheles gambiae*. The next part is dedicated to the application of a traditional gene set enrichment analysis to validated target genes of human microRNAs. Motivated by the drawbacks of the traditional gene set enrichment approach, a new integrative pathway enrichment analysis which was developed during this thesis. Finally, an extensive cross-species comparison of two expression systems was performed.

The specific goals of this doctoral thesis are summarized below:

- Study of three integrative analysis methods: co-inertia analysis, generalized singular valued decomposition and biclustering,
- Application of multiple co-inertia analysis to proteomic and transcriptomic profiles of *Anopheles gambiae*,
- Traditional gene set enrichment analysis for microRNAs,
- Development of an integrative pathway enrichment analysis method,
- Cross-species comparison of expression systems.

1.3.1 Study of Three Integrative Analysis Methods

As described in section 1.2, there are various integrative analysis techniques: some are based on network models, some on singular value decomposition and others on logical models. In this work we have decided to apply three different integrative approaches to the same data set and evaluate the results.

The methods we decided to use are: co-inertia analysis, generalized singular value decomposition and biclustering. The choice of these methods was based on the following criteria: (i) they are based on a clear mathematical formulation, (ii) they are based on different mathematical concepts and, the most important argument, (iii) they allow the analysis of all measured features (not limited to pairs of genes and proteins).

These integrative techniques were applied to mRNA and protein abundance data from the six life cycle stages of *P. falciparum*: merozoite, ring, trophozoite, schizont, gametocyte and sporozoite.

1.3.2 Application of Multiple Co-Inertia Analysis to *Anopheles gambiae*

To determine the co-structure between measured proteomic and existing hemocyte transcriptomic profiles, candidate genes responsive to granulocyte-enrichment during sugar-feeding (SF), blood-feeding (BF), or *P. falciparum*-infection (PF) were examined by multiple co-inertia analyses (MCIA). The goal was to quantify the concordance between proteomic and transcriptomics profiles.

Using published hemocyte transcriptome data, MCIA was used to examine the degree of agreement between transcript and protein abundance in our granulocyte proteomes (SF, BF, PF). Comparisons were made between transcriptional profiles of non-selected hemocytes from sugar-fed naïve mosquitoes, 24 hours after feeding with a non-invasive CTRP mutant *Plasmodium berghei* (comparable to a non-infectious blood meal), or 24 hours after feeding with wild-type *P. berghei*.

1.3.3 Traditional Gene Set Enrichment Analysis for microRNAs

The traditional gene set enrichment analysis of microRNAs represents a part of a study in which it was shown that microparticles isolated from cerebrospinal fluid of traumatic brain injured patients are potent, injury-specific messengers carrying mRNAs, miRNAs and proteins.

The goal of the analyis was to identify neuron related biological process enriched in two groups of microRNAs. One group was associated with microparticales of the cerebrospinal fluid of brain injured patients while the second one included the complete set of microRNAs identified in the cerebrospinal fluid of brain injured patients.

Validated target genes were computed for each microRNA in each group. Subsequently, a traditional gene set enrichment analysis based on Gene Ontology terms was performed on each group.

1.3.4 Development of an Integrative Pathway Enrichment Analysis Method

Motivated by the drawbacks of the traditional gene set enrichment analysis of microRNAs, where only a flat gene list could be analyzed, a new method was designed for this task. This approach is focused on pathways which are, similar to gene ontology terms, functional groups. This method can be easily applied to one or more data sets.

The integrative pathway analysis is based on the combination of two gene scores: one score is determined by a multivariate analysis and the second score quantifies the importance of a gene in a pathway. The first gene score can be the log fold change of differential expressed genes but it can also come from any integrative analysis methods such as co-inertia analysis. The second gene score is computed for each pathway separately and it is based on the network topology of the pathway. All functional groups that incorporate network information can be used in this approach.

1.3.5 Cross-species Comparison of Expression Systems

The general concept of this project is based on a comparative genome-wide analysis of different eukaryotic host species with different capacities for protein expression with a main focus on functional, structural and regulatory processes involved in the expression of recombinant proteins.

Due to its characteristics, such as the ability to integrate more than two data sets, maximization of common trends, visualization of samples and features as well as projection of additional information into the computed space, multiple co-inertia analysis was the method of choice for this study.

In more detail, the expression system of Chinese hamster ovary (CHO) cells and *Pichia pastoris* were chosen to be compared. For each one, strains expressing two heterologous proteins were engineered. For each expressed protein, two different expression clones were selected, one with a very high expression level, and another with a low expression level.

Transcriptomic and proteomic profiles were measured for each expression system, each protein and each expression clone. The resulted data sets were analyzed with multiple co-inertia analysis, generalized singular valued decomposition and biclustering.

2 Methods

Due to the large number of available integrative analysis methods and their heterogeneity it is not always obvious which method is best suitable for a certain study. A large part of this doctoral thesis is dedicated to the comparison of three different integrative analysis approaches: co-inertia analysis, generalized singular value decomposition and biclustering. The first three sections of this chapter will cover the detailed presentation of these methods and are based on the publications that emerged from this comparison [146, 147].

Very often, one of the last steps in an integrative analysis study is to perform some kind of gene set enrichment analysis. The fourth section is dedicated to the traditional approach while the fifth section presents an integrative gene set enrichment method focused on pathway analysis which was developed during this doctoral thesis. Finally, the last section of this chapter will present all used data sets.

2.1 Co-Inertia Analysis and Multiple Co-Inertia Analysis

Co-Inertia Analysis (CIA) is an integrative analysis method that was introduced by Dolèdec and Chessel to the ecology research community in 1994 [148]. It took almost 10 years until its potential for bioinformatic analyses was discovered [130, 131]. CIA was developed for the analysis of two data sets but was extended to multiple co-inertia analysis (MCIA) for the analysis of multiple data sets. A detailed mathematical description of MCIA is provided by Hanafi and Chessel in [149]. For the non-French speaking reader, MCIA was recently introduced to the bioinformatics community by Meng *et al.* in [137].

MCIA was implemented in the R package *omicade4* [137] and is usually applied to two or more omics data sets. In order to use the R package, each data set has to be formulated as a matrix having more rows than columns. In general, the rows hold the features of a data set while the columns hold the conditions. MCIA can be applied to a list of matrices subject to having matched conditions. The measured features are not subject to any constraint.

MCIA is performed in two main steps: i) application of a dimension reduction technique on each data set and ii) computation of the co-inertia axes from all data sets.

2.1.1 Dimension Reduction

Various dimension reduction techniques (DRT) exist. During this thesis a comprehensive review on dimension reduction techniques in the context of integrative data analysis was written [150] which represents the basis of this chapter. The main goal of a DRT is the graphical display of high dimensional data into a low dimensional space by retaining as much as possible of the initial data structure. Principal component analysis (PCA) is probably the most well known and most often applied approach. Alternatives include correspondence analysis (CA) and non-symmetrical correspondence analysis (NSCA) as well as multidimensional scaling (MDS). All these methods can be computed by singular value decomposition (SVD) [151] but they differ in how the data is transformed prior to SVD. MDS or principal coordinate analysis [152] is a PCA applied to a distance matrix.

While PCA [134, 153, 154] can be applied to continuous data, CA [135, 155] and NSCA [135, 156] were designed for the analysis of categorical data summarized in form of contingency tables. Nevertheless, CA and NSCA have already been successfully applied to continuous data such as gene expression and protein profiles [105, 130, 157]. As Fellenberg *et al.* point out, gene and protein expression can be seen as an approximation of the number of corresponding molecules present in the cell during a certain measured condition. Additionally, Greenacre emphasized in [155, 158] that the descriptive nature of CA and NSCA allows their application on data tables in general, not only categorical data. These two arguments support the suitability of CA and NSCA as analysis methods for omics data.

Although CA is based on the pioneer work on categorical data of the brilliant Englishmen Ronald Fisher, Karl Pearson, Frank Yates and George Yule, the method itself [159] was developed by the Frenchmen Benzècri and it is recognized to be a French approach. According to Beh and Lombardo in [135] the name of the method is a direct translation of the French *l'analyse de correspondances* which literally means the analysis of correspondences and relationships/associations in the data. This method was slowly accepted by the international research community and, when employed, the main goal was data visualization. Forty years after its development, Fellenberg *et al.* introduced CA to the bioinformatics community by applying it to microarray data [157].

The concepts of CA and NSCA are similar to PCA in that they project the data into a lower dimensional space while trying to maintain as much as possible from the captured data structure. While PCA uses a covariance matrix to account for data structure, CA and NSCA use the χ^2 statistic to measure the deviation from independence of the considered variables.

While CA investigates symmetric associations between two variables, NSCA, an alternative CA, captures asymmetric relations between variables. NSCA [160] was developed approximately twenty years after CA at the University of Naples in Italy by Lauro and D'ambra.

2.1.2 The Duality Diagram

A detailed description of the NSCA will be shown in the context of the duality diagram (DD). The DD approach was developed in the French school of data analysis in the 1970's but was rarely published in English [132, 133, 161]. DD is a unique and unifying concept because most of the multivariate analysis methods such as PCA, CA, NSCA, discriminant analysis and canonical correlation analysis, which are well known and very often used, can be formulated as a DD problem.

Based on [133] and [149], a mathematical description of DD, NSCA and MCIA is summarized bellow. The DD framework is based on the statistical triplet (X, Q, D) where X is assumed to be a matrix with n rows (observations) and p columns (variables). D is a diagonal matrix and its elements are the weights associated with the n observations. Q, a $p \times p$ symmetric and positive definite matrix, defines a neighborhood relation between the observations. From a geometrical point of view, Q and D define geometries or inner products in $\mathbb{R}^p$ respectively $\mathbb{R}^n$:

$$x^t Q y = < x, y >_Q \qquad x, y \in \mathbb{R}^p \text{ and represent observations} \tag{2.1}$$

$$x^t D y = < x, y >_D \qquad x, y \in \mathbb{R}^n \text{ and represent variables} \tag{2.2}$$

Additionally, Q and D can be seen as linear functions from $\mathbb{R}^p$ to $\mathbb{R}^{p*} = L(\mathbb{R}^p)$ respectively from $\mathbb{R}^n$ to $\mathbb{R}^{n*} = L(\mathbb{R}^n)$ with $L(\mathbb{R}^p)$ being the space of scalar linear functions on $\mathbb{R}^p$. The association of an operator from the space of observations $\mathbb{R}^p$ to the dual space of variables $\mathbb{R}^{n*}$ was proposed by Escoufier and summarized graphically in Figure 2.1 which is addopted from [133].

$$\begin{tikzcd}
\mathbb{R}^{p*} \arrow[r, "X"'] \arrow[d, shift left=3, "V"] & \mathbb{R}^{n} \arrow[d, shift right=3, "D"'] \\
\mathbb{R}^{p} \arrow[u, shift left=3, "Q"] & \mathbb{R}^{n*} \arrow[l, "X^t"] \arrow[u, shift right=3, "W"']
\end{tikzcd}$$

Figure 2.1: Duality Diagram. Figure adapted from [133]

By defining $V = X^t D X$ and $W = X Q X^t$, the DD is made commutative. The name "Duality Diagram" comes from the property of the diagram that the decomposition of the operator $VQ = X^T D X Q$ leads to the decomposition of the operator $WD = X Q X^T D$ which leads further to an easy transition between principal components and principal axes. Another important property of the diagram is that the eigenvalues of VQ are equal to those of WD.

From a geometrical point of view, analyzing the statistical triplet (X, Q, D), which is equivalent to analyzing the DD (X, Q, D), can be formulated as either finding the inertia axes (principal axes) of a data set containing n points in $\mathbb{R}^p$ or as finding the inertia axes (principal components) of p points in $\mathbb{R}^n$.

For this, the inertia operators $WD = X Q X^T D$ and $VQ = X^T D X Q$ have to be diagonalized:

$$Q = E^T E \quad D = B^T B \quad \text{Cholesky decomposition} \tag{2.3}$$

$$\Omega = B X^T \Rightarrow \Omega^T \Omega = E X^T B^T B X E^T \Rightarrow \Omega^T \Omega = V \Lambda V^T \tag{2.4}$$

$$\Omega = B X^T \Rightarrow \Omega \Omega^T = B X E^T E X^T B^T \Rightarrow \Omega \Omega^T = U \Lambda U^T . \tag{2.5}$$

The next step is the computation of principal axis and principal components:

$$F = E^T V \quad A = E^{-1} V \qquad \text{A are the principal axis} \tag{2.6}$$

$$G = B^T U \quad K = B^{-1} U \qquad \text{K are principal components.} \tag{2.7}$$

The last step is the projection of the rows of X onto the principal axis:

$$L = XQA \tag{2.8}$$

and of the columns of X onto the principal components:

$$C = X^T DK. \tag{2.9}$$

Comparing Two Duality Diagrams

The comparison of two or multiple DDs is of interest whenever the co-structure captured by them has to be measured. This comparison can only be done if the data sets have either common rows or columns. The key element of such a comparison is the vectorized version of the R squared from Pearson's correlation, the RV coefficient which was defined by Robert and Escoufier in [162] as follows.

Given two symmetric matrices A and B with the same size, the RV coefficient is defined as:

$$RV(A, B) := \frac{tr(AB)}{\sqrt{tr(AA)tr(BB)}}, \tag{2.10}$$

where $tr(X)$ is the trace of the matrix X which is defined as the sum of the eigenvalues of the matrix X.

When comparing two DDs, one has to assume that one dimension is equal, for example the number of variables p_1 and p_2. In this case the RV coefficient is computed between the operators $W_1 D$ and $W_2 D$ which are symmetric and of the same size. Due to this, we define the RV coefficient of two DDs (X_1, Q_1, D) and (X_2, Q_2, D) as:

$$RV((X_1, Q_1, D), (X_2, Q_2, D)) := RV(W_1 D, W_2 D). \tag{2.11}$$

The RV coefficient [162] is a generalization of the squared Pearson correlation coefficient for matrices and has values in the interval $[0; 1]$. A RV value close to one indicates a high co-structure between the two data sets.

Despite its wide usage, the traditional RV coefficient defined above suffers from dimension bias. The smaller the data sets, the higher the RV coefficient. To overcome this problem, a modified version of the RV coefficient was used, as described by Smilde *et al.* in [163]. Given two matrices X and Y, the modified RV coefficient is computed as:

$$RV(X, Y) = \frac{Vec(\widetilde{XX'})Vec(\widetilde{YY'})}{\sqrt{Vec(\widetilde{XX'})'Vec(\widetilde{XX'}) \times Vec(\widetilde{YY'})'Vec(\widetilde{YY'})}}, \tag{2.12}$$

where $\widetilde{XX'} = XX' - diag(XX')$ with $diag(XX')$ being the matrix containing only the diagonal elements of XX' and $Vec(X)$ symbolizes the vector constructed by concatenating all rows of X. The modified RV coefficient ranges between -1 and 1. The interpretation of this value is similar to the interpretation of a correlation coefficient.

2.1.3 Non-symmetrical Correspondence Analysis

The analysis of the statistical triplet (X, Q, D) becomes a NSCA when the following transformations are performed: Let x_{ij} be the value of the matrix $\hat{X}$ from row i and column j. Additionally, $x_{i.}$ denotes the sum of row i and $x_{.j}$ the sum of column j. The sum of all elements of $\hat{X}$ is denoted by $x_{..}$. The frequency of x_{ij} is defined as $f_{ij} = \frac{x_{ij}}{x_{..}}$. The weight of row i is defined as $r_i = \frac{x_{i.}}{x_{..}}$ and the weight of column j is defined as $c_j = \frac{x_{.j}}{x_{..}}$. The row weights are gathered in matrix $D = diag(r_i)$ and the column weights are gathered in matrix $Q = diag(c_j)$.

The matrix $\hat{X}$ is standardized by $\frac{f_{ij}}{r_i} - c_j$ which holds the differences between the conditional prediction of the j-th column category (row profile) and the unconditional marginal prediction c_j (column marginal proportion). By considering these centered row profiles one can determine the predictive value of the column (response) values given the row (predictor) values [135].

In CA, where there are no predictors and no responses, the following standardization is used: $x_{ij} = \frac{f_{ij}}{r_i \cdot c_j} - 1$.

On of the main advantages of formulating a NSCA as a DD is the ability to plot the rows and the columns of the initial matrix in the same plot. The interpretation of the resulted plot is based on the following rules [135]: While one can directly interpret the distance between two rows as well as the distance between two columns, one can not interpret the distance between a row and a column. Nevertheless, one can interpret the angle between a row and a column. The smaller this angle is, i.e. the row and the column are projected in the same direction from the origin, the higher the predictive value of the row for that column is. The origin of the plot marks the point where the increase in predictability of the rows is zero. Due to this, one can say the rows that are projected far away from the origin have an increase of predictive power for the columns.

Figure 2.2 shows an example of a NSCA performed on data from Skelikoff's asbestos study in which duration of exposure to asbestos is associated with the diagnosed grade of Asbestos [135]. Here one can see there is a small distance between the exposure *20-29* and *30-39* (both shown in blue) as well as between *Grade 2* and *Grade 3* (both shown in red). In order to determine, e.g. if exposure *0-9* (blue) is associated to grade *none* one has to inspect the angle between them. In this case the angle is acute which means that there is an association. In contrast, the angle between *none* and exposure *40+* is obtuse resulting in no association.

	Occupational exposure (years)					
Asbestos grade						
Diagnosed	**0–9**	**10–19**	**20–29**	**30–39**	**40+**	**Total**
None	310	212	21	25	7	575
Grade 1	36	158	35	102	35	366
Grade 2	0	9	17	49	51	126
Grade 3	0	0	4	18	28	50
Total	346	379	77	194	121	1117

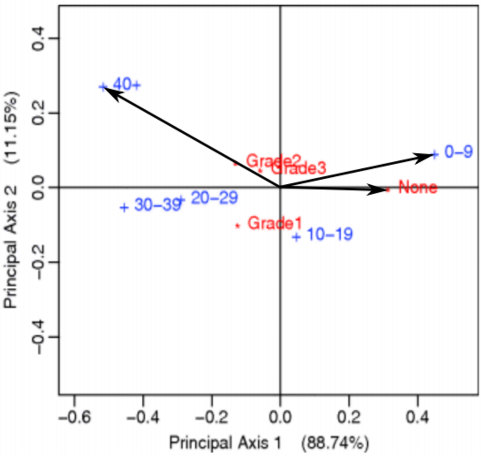


Figure 2.2: NSCA Analysis of Skelikoff's asbestos data. Example adapted from [135].

2.1.4 Mathematical Description of the (Multiple) Co-Inertia Analysis

The mathematical description of MCIA is based on the work of Hanafi and Chessel from [149]. The general concept is graphically summarized in the flowchart from Figure 2.3 which is extracted from [146]. CIA is a special case of MCIA where only two data sets are analyzed.

According to [132] and [133], the inertia in CIA and MCIA is defined as the trace of the operator WD which is equal to the trace of the operator VQ. In general the inertia with regards to a point A of a set of n weighted points is defined as $\sum_{i=1}^{n} p_i d^2(x_i, A)$ where p_i is the weight of the point x_i and $d(x_i, A)$ is the distance between the points x_i and A. For CA and NSCA, the inertia is proportional to the χ^2 statistic while for ordinary PCA it is equal to the total variance of all the variables.

MCIA operates on K statistical triplets (X_k, Q_k, D) with $k = 1, ..., K$. X_k are a set of transformed matrices, Q_k are a set of $(p_k \times p_k)$ diagonal matrices containing the row weights of X_k. D is a $n \times n$ identity matrix. Matrix X is created by merging all matrices X_k: $X = [\omega_1 X_1 | \omega_2 X_2 | ... | \omega_K, X_K]$, where ω_k is the inverse sum of the eigenvalues of X_k. Please note that the eigenvalues of a matrix can be computed as the square roots of the matrix's singular values.

MCIA is defined as the analysis that computes k vectors u_k^1 normed in $\mathbb{R}^{p_k}$ and an auxiliary variable v^1, D-normed in $\mathbb{R}^n$, that maximize:

$$g(u_1, u_2, ..., u_K, v) = \sum_{k=1}^{K} \omega_k cov^2(X_k Q_k u_k, v) = \sum_{k=1}^{K} \omega_k (X_k Q_k u_k | v)_D^2 \tag{2.13}$$

In a second step, the vectors u_k^2 normed in $\mathbb{R}^{p_k}$ and the auxiliary variable v^2 normed in $\mathbb{R}^n$ that maximize the same function g and are orthogonal to u_k^1 and v^1 are computed.

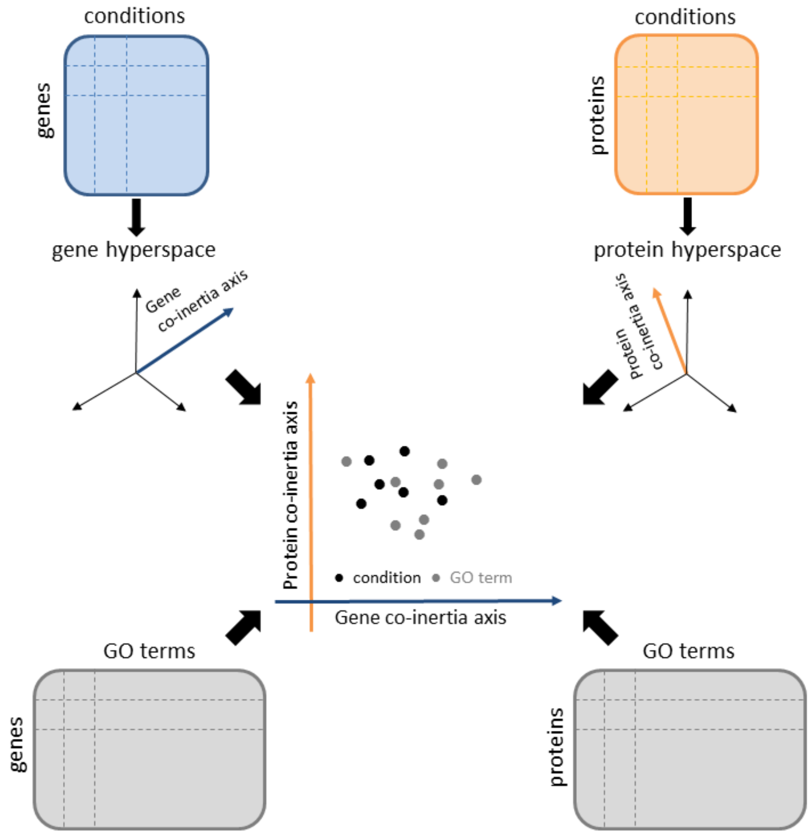


Figure 2.3: Flowchart of CIA. The gene expression matrix (blue) contains the genes in the rows and the conditions in the columns. The protein expression matrix (orange) contains the proteins in the rows and the conditions in the columns. Genes and proteins annotated to the considered GO Terms are gathered in a separate matrix (gray). The gene and the protein expression matrices are transformed into a new hyperspace. Axes maximizing the squared covariance are computed from the axes spanning the gene and protein hyperspaces. Conditions together with GO Terms and features are projected into the computed CIA space. Figure adapted from [148].

In the s step, the function g is maximized and:

$$(v^j|v^s)_D = 0 \text{ and } (u_k^j|u_k^s)_{Q_{p_k}} = 0 \quad (1 \leq j < s,\ 1 \leq k < K) \tag{2.14}$$

The set of vectors u_k^i and v^i are calculated iteratively. Each set of k vectors u_k^i and the vector v^i are computed during one iteration: the first set of vectors u_k^1 and the auxiliary variable v^1 are computed with the first order solution. The remaining sets of k vectors u_k^i and v^i, $i = 2, \ldots, S$ are computed subsequently with the second, third, ..., S order solution, where S is the number of desired MCIA eigenvectors. The details of the iterative solutions are shown below.

First order solution

For a fixed vector v, D-normed in $\mathbb{R}^n$, the use of the Cauchy-Schwartz inequality shows that $(X_k Q_k u_k | v)^2_D$ is maximized by $||X_k^T Dv||^2_{Q_k}$ for $u_k = \frac{X_k^T Dv}{||X_k^T Dv||_{Q_k}}$.

In [149] it was shown that since v maximizes g it also maximizes:

$$\sum_{k=1}^{K} \omega_k ||X_k Q_k v||^2_{Q_k} = v^T D \left(\sum_{k=1}^{K} \omega_k W_k D \right) v \tag{2.15}$$

v is the first D normed principal component of the matrix X. Additionally, the axes u_k^1, Q normed in $\mathbb{R}^{p_k}$ are the normalized vectors $\frac{X_k^T Dv}{||X_k^T Dv||_{Q_k}}$.

Second order solution

- Consider P_k^1, the Q_k orthogonal projections of u_k^1 into the vector space of $\mathbb{R}^{p_k}$.
- Define a new matrix $Z = [Z_1, Z_2, ..., Z_K]$, where $Z_k = X_k - X_k {P_k^1}^T$
- Compute the MCIA first order solution for the matrix Z.

S order solution

- Consider P_k^{s-1}, the Q_k orthogonal projections of u_k^{s-1} into the vector space of $\mathbb{R}^{p_k}$.
- Define a new matrix $Z = [Z_1, Z_2, ..., Z_K]$, where $Z_k = X_k - X_k {P_k^{s-1}}^T$
- Compute the MCIA first order solution for the matrix Z.

2.1.5 Projecting Additional Information into the MCIA Space

As already mentioned, the results of integrative analysis methods, but not only these, are often subject to a gene set enrichment analysis. This is specifically helpful when the result consists of long lists of genes. This approach is described in detail in Sections 2.4 and 3.4. Additionally, the annotation of genes to gene sets like gene ontology (GO) terms [164] or pathways from databases such as Reactome [165], can be used as an additional layer in MCIA.

Additional information such as GO annotations can be superimposed on the (M)CIA plots. This overlay was already done for CA [166], and is also possible for CIA [105]. GO Term projections are obtained by first normalizing the two GO matrices in the same way as the expression data sets and then multiplying them by the weights of the genes/proteins resulting from the NSC. The projection scores computed in this way show GO Term associated with the present features, in relation to the measured conditions.

2.2 Generalized Singular Value Decomposition

The second method that will be used for comparison is the generalized singular value decomposition (GSVD). GSVD was developed as an extension of the singular value decomposition (SVD), one of the most often applied analysis methods. In this section we start by introducing the SVD and its most well known application: the principal component analysis. The second part of the section is dedicated to the mathematical description of GSVD which was extracted from [146].

2.2.1 Singular Value Decomposition

SVD is an analysis method applied on its own [167, 168] and as part of PCA [169, 170]. SVD [171] is a linear algebra method used for matrix factorization. The SVD of a matrix $M \in \mathbb{R}^{mxn}$ is defined as:

$$M = U\Sigma V^*, \tag{2.16}$$

where U contains the left singular eigenvectors, Σ contains the singular values and V^* contains the right singular vectors. Please note that the singular vectors of M are the eigenvectors of MM^t as well as of M^tM and the the singular values of M are the square roots of the eigenvalues of MM^t or M^tM.

The interpretation of SVD is intuitive in the special case when M is a square, invertible matrix. In that case, the matrices U and V can be seen as rotation matrices while Σ can be seen as a scaling matrix. In this context, the SVD can be seen as a sequence of geometrical transformations: rotation, scaling, rotation.

One of the most popular applications of SVD is the principal component analysis (PCA). PCA was developed by Pearson in 1901 [153] and thirty years later independently rediscovered and extended by Hotelling in [154, 172].

In general, PCA is applied on a matrix M with p conditions and n variables which has been mean centered. This data can be seen as n variables in a $\mathbb{R}^p$ space. PCA tries to find orthogonal principal components that provide the best representation of the data in a lower dimensional space $\mathbb{R}^q$, $q < p$. In practice, $q \in \{2, 3\}$ to ensure a graphical representation of the high dimensional data which can be easily perceived by the human eye.

The representation quality in the new low dimensional space is measured by the percentage of explained variance of the original data. The principal components which build the orthogonal basis of the new space are computed by SVD and are the left eigenvectors of the original matrix M or of the covariance matrix M^tM. The percentage of explained variance by each principal component is given by the eigenvalues of M^tM which correspond to the squared singular values of M.

SVD is widely used and has shown its potential in numerous applications. Nevertheless, SVD has a huge disadvantage, especially in the data rich era we are currently experiencing: because it was designed for one data set it is not suitable for integrative data analysis. Due to this, an extension of the SVD which can be applied to multiple omics data sets was developed: the generalized SVD.

2.2.2 Mathematical Description of the Generalized Singular Value Decomposition

GSVD was first described by Golub and Van Loan in their book about matrix computation [171]. In a study by Alter *et al.*, GSVD was used as a comparative analysis method [127] for two gene expression data sets of cell cycle data from yeast and humans. In this comparison study, the GSVD was applied to two different data sets, gene and protein abundances.

GSVD is based on the joint decomposition of two matrices:

$$G = U_1 \Sigma_1 X^{-1} \tag{2.17}$$

$$P = U_2 \Sigma_2 X^{-1} \tag{2.18}$$

which is subject to the following constraint: the third decomposition matrix, X^{-1}, is shared by both decompositions.

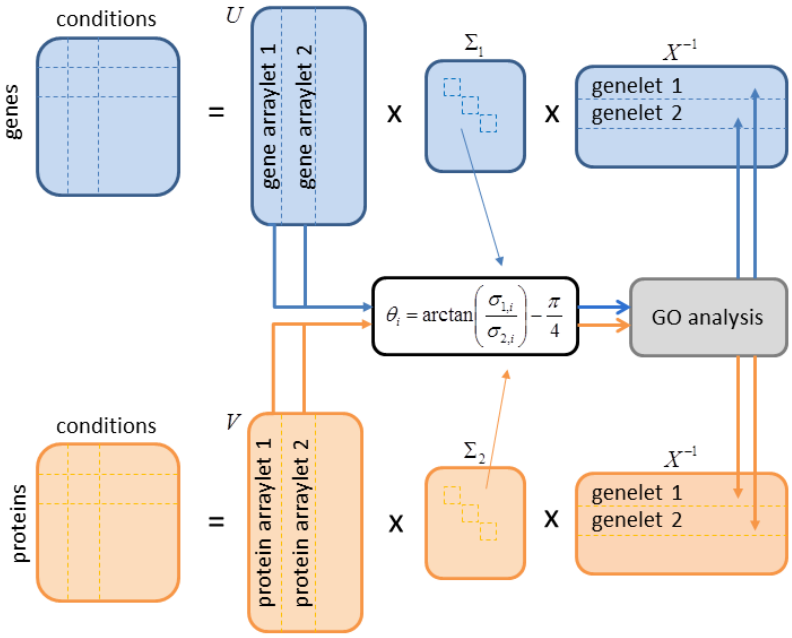


Figure 2.4: Flowchart of the GSDV. The gene expression matrix (blue) contains the genes in the rows and the conditions in the columns. The protein expression matrix (orange) contains the proteins in the rows and the conditions in the columns. Genes and proteins annotated to the considered GO Terms are gathered in a separate matrix (gray). The gene and the protein expression matrices are each decomposed in three matrices. The matrices U and V contain the arraylets, which encode for the expression of genes and proteins in the corresponding genelets X^{-1}, which represent the cellular state in the measurement conditions. According to the angular distance θ_i, which is computed from the generalized eigenvalues $\sigma_{1,2}$ and $\sigma_{1,2}$, a restricted GSE analysis is performed on the genes and/or proteins with the absolute (from a mathematical point of view) highest values in the arraylets in order to assign GO Terms to the corresponding genelets. Figure adapted from [127].

In general, the matrices G and P contain the data sets to be analyzed, e.g. gene and protein abundance data. The rows of the common matrix X^{-1} are named *genelets*. Alter *et al.* showed that these genelets can be seen as biological processes captured by both data sets. The genelets are expressed only in the corresponding arraylets (corresponding columns of U_1 and U_2) with a relative significance measured by the generalized eigenvalues $(\sigma_{1,m}, \sigma_{2,m})$ from the diagonals of Σ_1 and Σ_2. The relative significance of a genelet in the gene data set relative to the protein data set is measured by an antisymmetric angular distance calculated as:

$$\theta_m = \arctan\left(\frac{\sigma_{1,m}}{\sigma_{2,m}}\right) - \frac{\pi}{4}. \tag{2.19}$$

An angular distance between $-\pi/4$ and $-\pi/8$ represents a high significance of the m^{th} genelet in the second data set relative to the first data set. If the value of the angular distance ranges between $\pi/8$ and $\pi/4$, then the m^{th} genelet has a high significance in the first data set relative to the second data set. The m^{th} genelet shows equal significance in both data sets if the angular distance ranges between $-\pi/8$ and $\pi/8$. In this comparative study, where the first matrix contains mRNA abundance data and the second matrix protein abundance data, the relative significances are assigned as follows:

$$\theta_m \in \begin{cases} [-\pi/4, -\pi/8] & \text{protein space} \\ [-\pi/8, \pi/8] & \text{gene and protein space} \\ [\pi/8, \pi/4] & \text{gene space.} \end{cases} \tag{2.20}$$

A summary of the computation flow is shown in a block diagram in Figure 2.4.

Alter *et al.* [127] used a Mathematica implementation of a numerically robust GSVD algorithm based on [171, 173], which was reimplemented in R (see Appendix A.1) during this doctoral thesis.

In order to discover the processes captured by the genelets, a restricted gene set enrichment (for details see Section 2.4) analysis is performed on 50% of the genes and/or proteins showing the highest absolute values in the corresponding arraylets. The GSE analysis is performed with the R package *GOstats* [174], which computes the statistically significantly enriched GO Terms based on the hypergeometrical distribution. A detailed description of this traditional analysis can be found in section 2.4.

2.3 Integrative Biclustering

This section will present the third method used in the comparison study: integrative biclustering (IB). A short introduction to clustering methods is given in the first part of this section while a detailed description of IB, extracted from [146], can be found in the second part.

2.3.1 Clustering

Clustering, similar to PCA, represents one of the most well-known and often applied method in bioinformatics. Traditional methods include k-means and hierarchical clustering.

While k-means [175] starts with k randomly chosen means (mean values) and proceeds by alternating between assigning data points to the cluster yielding the least within-cluster sum of squares (intuitively the nearest mean) and recomputing the mean of the newly discovered clusters until convergence, the agglomerative hierarchical clustering approach [176] starts by assigning each data point to one cluster and proceeds by iteratively merging the clusters based on a distance metric until all points are assigned to one cluster and presents the results as a dendrogram. K-means has two disadvantages: the user has to guess the optimal number of clusters k and the algorithm's convergence is based on the Euclidean distance as a measure.

Given a data matrix M which contains n variables measured in p conditions, the results of the clustering approaches share one key aspect: the matrix is clustered either by variables or by conditions but not by variables *and* conditions. This disadvantage led to the development of biclustering approaches where variables and rows can be clustered simultaneously.

2.3.2 Detailed Description of Integrative Biclustering

The basic idea of biclustering (co-clustering or two-way clustering) was presented in [177], but it took almost thirty years until the method was applied to gene expression data [178]. In the last two decades, biclustering has become more and more popular [179–181]. In contrast to clustering, where either rows *or* columns are clustered, biclustering performs clustering of rows *and* columns simultaneously. The members of the obtained biclusters are as similar to one another and as different from the other biclusters as possible. Figure 2.5 exemplifies how mRNA abundances, protein expression and GO Terms are assembled to a complete data set and what the resulting biclusters could look like.

There are four types of possible biclusters as reviewed in [126, 182, 183]. Biclusters can have (i) equal values over rows and columns as well as (ii) equal values over rows or columns. They can also have (iii) coherent values, which means that each column or row can be computed by adding or multiplying a constant to the previous column or row. The fourth type has (iv) coherent evolutions, which means that the exact value of a matrix entry is not important, but whether the values increase or decrease over rows or columns. The biclustering algorithm used here is implemented in the R package *biclust* [184].

The types of computed biclusters vary. There are single biclusters where only one bicluster is found in the whole data set as well as exclusive rows and/or exclusive columns biclusters. Non-overlapping and non-exclusive biclusters can also be computed. The fifth type is the arbitrarily positioned overlapping biclusters. Graphical representations of the different categories of biclusters can be found in [126].

Most of the biclustering algorithms implemented depend on the starting point of the search and thus may lead to different results in consecutive runs. Additionally, biclustering does not result in a perfect data separation, as overlapping biclusters are possible. As a remedy, the *biclust* package provides a robust method that delivers stable and reliable results. This function includes the repeated use of one algorithm in combination with several parameter settings and/or subsamples of the data. A modified version of the Jaccard index [126] is used for the combination of the resulting biclusters, which in case of two biclusters takes into account the fraction of row-columns combinations in both biclusters to all row-column combinations. For more detailed mathematical definitions, please refer to [126].

Analogous to integrative clustering, we define integrative biclustering as the biclustering of two or more data sets. Integrative clustering was already applied to copy number and gene expression data in order to identify novel breast tumours subgroups [185]. Mo *et al.* describe in [99] integrative clustering of genomic, epigenomic and transcriptomic profiling.

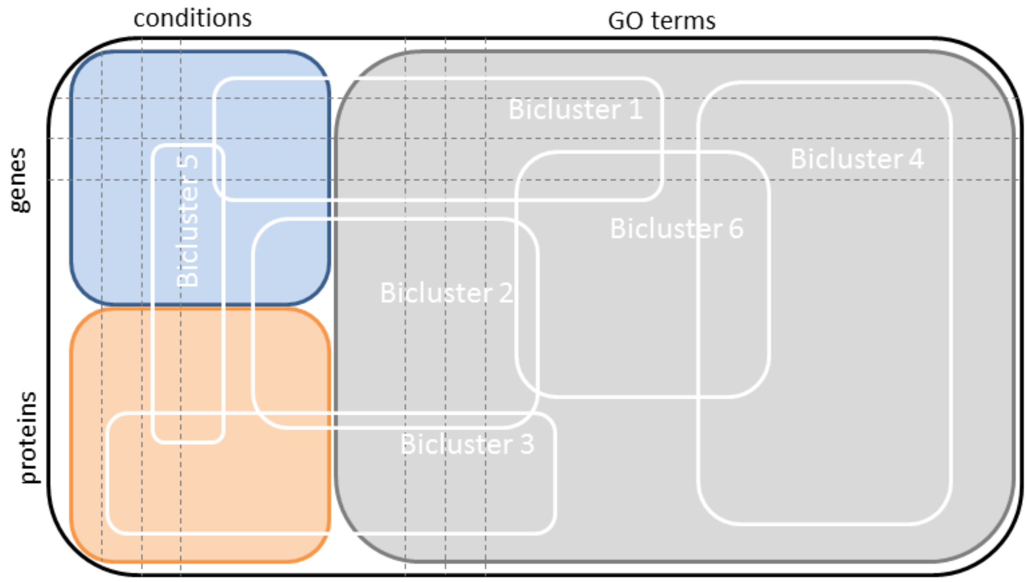


Figure 2.5: Flowchart of IBC. The gene expression matrix (blue) contains the genes in the rows and the conditions in the columns. The protein expression matrix (orange) contains the proteins in the rows and the conditions in the columns. Genes and proteins annotated to the considered GO Terms are gathered in a separate matrix (gray). The three matrices are combined to a new matrix, which is subjected to biclustering. The resulting biclusters can include genes, proteins, conditions, GO Terms or any combination of these.

Integrative biclustering was applied to gene expression, protein interaction, growth phenotype and transcription factor binding data in [186] in order to reveal modularity and organization in the yeast molecular network. Here, integrative biclustering was applied to a matrix consisting of mRNA and protein abundances as well as an additional level of complexity introduced by the corresponding GO matrices. The genes and the proteins are represented by rows, whereas the samples and the GO Terms by columns. Before biclustering can be carried out, discretization is necessary. Here, the built-in function *discretize* from the R package *biclust* [126] was used. This function computes a discrete matrix with a given number of levels of equally spaced intervals from minimum to maximum. After appropriate processing, the result of IBC was loaded into Cytoscape [187] to obtain a network view of the associations.

2.4 Traditional Gene Set Enrichment Analysis for microRNAs

In this section the gene set enrichment analysis of microRNAs is described. The section starts with a short introduction to microRNAs and ends with a detailed presentation of the gene set enrichment (GSE) approach. GSE was used on top of the three integrative analysis methods CIA (see Section 2.1), GSVD (see Section 2.2) and IB (see Section 2.3) [146] as well as a stand alone method in the analysis of miRNAs [188].

2.4.1 Short Introduction to microRNAs

MicroRNA (miRNA) are small (micro comes from the Greek word *micros* which means small), non-coding RNAs that play a very important role in gene regulation, especially in gene silencing. It was shown that miRNAs act highly specific on the post-transcriptional level and that their length is between 21 and 23 base pairs [189].

miRNAs bind to the 3'UTR (untranslated region) of their target mRNA and regulate its expression in two ways: if the binding sequences are 100% complementary then the target mRNA will be degraded while partial complementary sequences lead to the inhibition of mRNA translation [190].

Since miRNAs are non-coding RNAs they will not be translated to proteins and thus are not coding for genes in the traditional sense. Due to this it is not possible to use miRNAs directly in a gene set enrichment approach. Nevertheless, there are databases that gather information about the genes that are regulated by miRNAs. These genes are called target genes and one of the most widely used databases is miRBase [191] (`http://www.mirbase.org/`).

2.4.2 Detailed Description of the Traditional Gene Set Enrichment Analysis

The ability to measure large amounts of data led to the development of new methods for the interpretation of the analysis results which also increased in size. This process is enhanced by the use of a structured description of the vast, already existing biological knowledge. There are different resources that provide this kind of information such as Gene Ontology (GO) [164], Reactome [165] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [192].

Gene Ontology

GO is a open access database which provides information on genes and their products based on three vocabularies: biological process (BP), molecular function (MF) and cellular compartment (CC). The analysis described here will focus on biological processes. An example of a more general biological process would be *metabolic process* or *cellular proccess* and of a more specific would be *cellular aldehyde metabolic process*.

The structure of each vocabulary is a directed acyclic graph with the more general terms at the top and the more specialized at the bottom (see Figure 2.6 for an example). The genes

[1] http://amigo.geneontology.org/amigo/term/GO:0006081

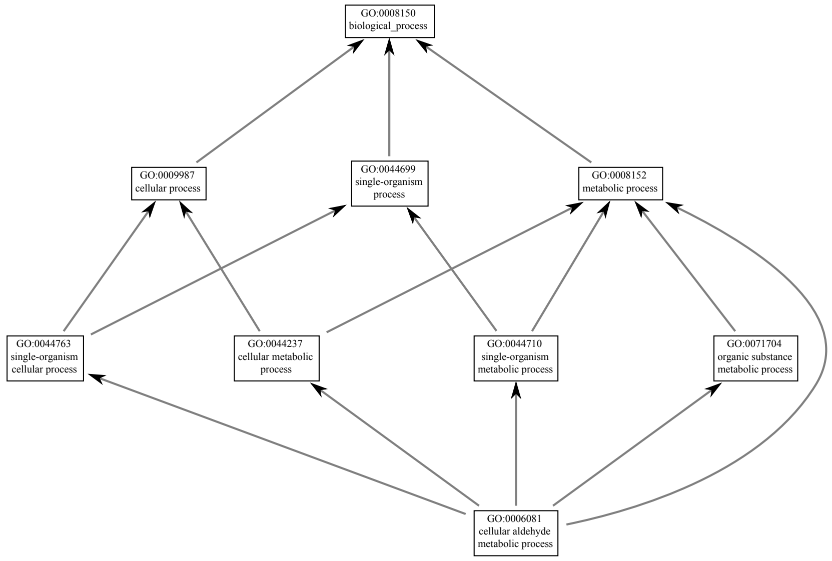


Figure 2.6: Example of GO Terms in the GO graph from the GO homepage[1]. Each term is represented by its name and its ID.

annotated to a node are also attributed to the parent nodes. Within a GO Term, the genes associated to it are equally important.

Due to the size of the three vocabularies, a reduced representation in form of GO Slim Terms has been developed. These terms can be created by GO or by researchers. The GO defined GO Slim Terms, such as the *S. cerevisiae* slim, represent a subset of the GO Terms and are particularly useful when a broad characterization of the analysis results is needed. The second category of GO Slim Terms are organism specific and often created to answer a particular research question. Our collaborators from the Austrian Centre for Industrial Biotechnology defined a set of GO Slim Terms for the cross-species comparison study described in this thesis (see Sections 3.6 and 4.5). This set is of particular interest for the study, e.g. relevant for protein secretion, and facilitates the comparison between *P. pastoris* and CHO on the GO level.

Mathematical Description and a Short Example

The traditional gene set enrichment analysis tells a researcher which GO Terms (or other gene sets) are enriched in his gene list. Mathematically [193], GSEA answers the following question: Given K (test set) out of N (reference set) sampled genes, what is the probability that k or more of these genes (i.e differentially expressed) belong to a functional category C shared by n of the N genes in the reference set? Figure 2.7 shows a graphical visualization of this question.

GSEA was initially introduced by Mootha *et al.* and Subramanian *et al.* in [194, 195] and the above question was answered with the Kolmogorov-Smirnov Test. Alternatively, the hyper geometric distribution or Fisher's Exact Test can be used.

Different tools were created to answer the question from above. The hyper geometric distribution is implemented in BINGO [193] which is a Cytoscape [196] plugin. Other widely used tools include DAVID [197] which employs Fisher's Exact Test and the GSEA approach developed and hosted by the Broad Institute [194, 195] which is based on the Kolmogorov-Smirnov Test. In this scenario, BINGO was chosen because, in addition to the enrichment analysis, it also provides a visualization tool.

The probability mass function (the equivalent of the probability distribution function for a discrete random variable) gives the probability of observing exact k successes and is defined as:

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \tag{2.21}$$

where N is the population size (reference set), K is the number of successes in the population (test set), n is the number of draws (genes in the gene set to be tested, by definition), k is the number of observed successes (observed genes in the gene set to be tested) and $\binom{a}{b}$ is the binomial coefficient defined as $\frac{n!}{k! \cdot (n-k)!}$.

In order to compute the statistical significance (p-value) of the gene set enrichment, the probability of observing at least k successes, on has to sum up the probability of sampling each possible success between k and n:

$$p = \sum_{\kappa=k}^{n} P(X = \kappa) \tag{2.22}$$

Given a list of 25 genes that were selected from a set of 1000 measured genes, a researcher is interested if a certain GO Term (GO:0070966) is enriched in the mentioned list if the list of selected genes includes 8 of the 163 genes annotated the GO Term of interest. This scenario is summarized in Figure 2.7.

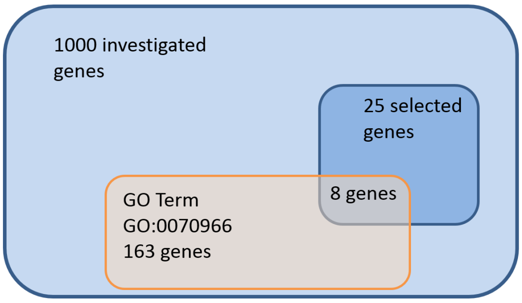


Figure 2.7: Gene set enrichment analysis constellation.

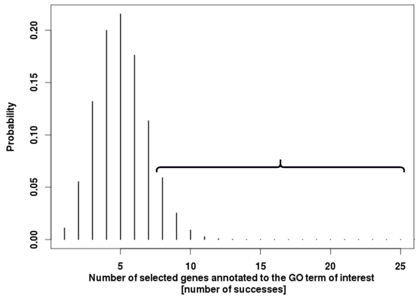


Figure 2.8: Probability mass faction of the hyper geometric distribution.

When using the hyper geometric distribution, the answer to the question is computed by summing the probability of at least 8 of the 25 genes being annotated to GO:0070966, given that 163 genes are attributed to the GO Term of interest and that the experiment was performed by measuring 1000 genes in total. The probability mass function is shown in

Figure 2.8 and the values which have to be summed in order to compute the statistical significance of the enrichment are emphasized resulting in a p-value of 0.038.

Due to the large number of GO Terms which are tested for enrichment, multiple testing correction [198] has to be performed in order to control the type I error (false positive) rate. BINGO implements the Benjamini and Hochberg [199] correction which controls the false discovery rate (FDR), i.e. the expected proportion of false positives among the positively identified tests [193].

2.5 Data Sets

During this doctoral thesis five data sets underwent integrative analysis. This section describes data generation, structure, availability and any processing steps prior to any integrative analyses.

2.5.1 *Plasmodium falciparum* Lifecycle Stages Analysis

The comparison study [146] between CIA, GSVD and IBC was performed on two publicly available data sets containing matched mRNA and protein abundance data from the six life cycle stages of *P. falciparum*. Microarray [200] and proteomic analyses [103] were carried out on *P. falciparum* clone 3D7. Gene expression levels were measured with a custom oligonucleotide array and computed with the match-only integral distribution (MOID) algorithm [201]. Proteins were detected by multidimensional protein identification technology (MudPIT), and protein abundance was estimated by the number of MS/MS spectra identified per protein. In total, 4294 genes and 2903 proteins were measured in all six life cycle stages. For each data set a matrix was created where the genes and proteins are represented as rows and the life cycle stages as columns.

Additionally, GO [164] information on biological processes in *P. falciparum* were employed. The R [202] packages *org.Pf.plasmo.db* [203] and *GO.db* [204], which provide *P. falciparum* specific mappings of genes to GO Terms as well as additional information on GO Terms, were used. Based on these two annotation databases, 3283 of 4294 genes and 2491 of 2903 proteins were associated with 614 GO Terms. For each data set, a GO matrix with the same number of rows as the corresponding expression data set was created. The columns of the GO matrix hold data describing the gene/protein affiliation to a certain GO Term. If a gene/protein is associated with that GO Term, the strength of the affiliation is computed as the ratio between 1 and the total number of genes/proteins associated with the GO Term. CIA and IBC use directly the computed GO matrix. GSVD performs a GSE analysis based on *org.Pf.plasmo.db* and *GO.db*. In this way, we make sure that all three methods are applied to the same data sets.

2.5.2 Analysis of Hemocyte and Granulocyte Immune Response of *Anopheles gambiae*

The *A. gambiae* data consists of two matched omics data sets: gene expression and protein abundance data.

The protein abundance data was generated by our collaborators: Ryan C. Smith, Jonas G. King, Dingyin Tao, Clara Brando and Rhoel R. Dinglasan from Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, USA. Based upon their ability to enrich for phagocytic granulocyte [205] populations, they performed proteomic analysis on mosquito hemocyte populations to determine the effects of sugar feeding (SF), blood feeding (BF) and malaria parasite infection (PF). LC-MS/MS was performed onto an Agilent LC-MS system consisting of a 1200 LC system coupled to a 6520 Q-TOF via an HPLC Chip Cube interface. Scaffold (Version 4.3.4, Proteome Software) was used for the curation, label-free quantification analysis and visualization of all search results. Scaffold's normalized spectral counting was employed to compare relative protein abundance between non-selected hemocytes (sugar-fed) and magnetic-bead enriched granulocytes (sugar-fed, blood-fed, and Plasmodium-infected). Scaffold calculates the quantitative spectrum count value by normalizing spectral counts across an experiment.

There are two published anopheline hemocyte transcriptome data sets that are available [206, 207]. However, the two studies used different microarray platforms. Merging the two dataset would artificially reduce the transcriptome data that can be used for MCIA to only the subset of transcripts that was measured in both analyses. This would limit the utility and primary advantage of a MCIA approach, i.e., avoid the need to subset the data sets for the analysis. As such, it was decided to focus on the Pinto *et al.* data set from [206], which provided all the necessary matching transcript data for the MCIA comparison to the granulocyte proteome. Experiments were performed on the GPL1321 GeneChip *Plasmodium/Anopheles* genome arrays (Affymetrix). Microarray analysis was performed using GCOS 1.4 (Affymetrix) and GeneSpring GX 7.3 software (Agilent Technologies). Data was downloaded from the Gene Expression Omnibus [208]: GSE17919 (naïve experiments) and GSE17866 (infection experiments).

MCIA was used to examine the degree of agreement between transcript and protein abundance in the granulocyte proteomes (SF, BF, PF). Comparisons were made between transcriptional profiles of non-selected hemocytes from sugar-fed naïve mosquitoes, 24 hours after feeding with a non-invasive CTRP mutant *Plasmodium berghei* (comparable to a non-infectious blood meal), or 24 hours after feeding with wild-type *P. berghei*. Hemocyte transcript and protein profiles analyzed by MCIA were based on the log fold change between two treatments (SF, BF and PF). Additionally, MCIA was performed on subsets of the proteomics data in order to examine the hemocyte-specific, immune-specific and proliferation-specific response of the innate immune system.

The transcriptomics data set includes 10469 genes. The hemocyte-specific proteomics data set contains 1128 proteins, the immune specific data set includes 43 proteins while 7 proteins are attributed to the proliferation specific data set. All features, i.e. genes and proteins as well as the mentioned subsets, were measured in three conditions: PFvSF, PFvsBF and BFvSF.

2.5.3 Gene Set Enrichment Analysis of Human miRNAs Following Traumatic Brain Injury

The miRNA data set introduced here represents a subset of the data generated by our collaborators from the Medical University Graz within a co-authored study [188] in which it was shown that microparticles isolated from cerebrospinal fluid of traumatic brain injured patients are potent injury specific messengers carrying mRNAs, miRNAs and proteins.

Briefly, cerebrospinal fluid (CSF) samples of patients with severe traumatic brain injuries (TBI) were collected when ventricular drainage was implemented as a measure of intensive care treatment [188].

In this study, a total of 63 human miRNAs were identified. 35 of them were associated with cerebrospinal fluid microparticles (CSF-MP). The goal of the analysis is to identify enriched, neuron-related, biological processes in the two lists of identified miRNAs. Owing to ethical considerations samples were not taken at standardized time points but rather when ventricular drainage was indicated due to acute increased intracranial pressure. Thereby a higher risk of ventricular drain infections as a result of additional study dependent interventions was avoided. 26 samples from 11 patients were collected over a time period of 2 years. CSF samples from adults who received a lumbar puncture for exclusion of subarachnoid hemorrhage or inflammatory diseases were collected for control studies (n=26). Subjects without confirmed central nervous system diseases, i.e. hemorrhage or inflammation, were considered healthy and included as controls (n=17).

The miRWalk database [209] was used to detect validated target genes of the identified human miRNAs (has-miR).

2.5.4 Integrative Pathway Enrichment Analysis of Microdissected Tumor and Stroma from Ovarian Cancer

In order to elucidate the crosstalk of tumor and stroma in ovarian cancer, a novel approach developed during this doctoral thesis, Integrative Pathway Enrichment Analysis (IPEA) is used.

We apply IPEA to the gene expression profiles of 38 pairs of microdissected tumor and stroma from high-grade serous ovarian cancers. The data set GSE40595 was downloaded from the Gene Expression Omnibus [208]. Data generation and preparation was described by Mok *et al.* in [210] and by Leung *et al.* in [211].

The samples originated from Department of Gynecologic Oncology and Reproductive Medicine at The University of Texas. Microdissection [212] was performed to separate the stromal from the tumor components.

GPL570 GeneChip Human Genome U133 Plus 2.0 microarrays (Affymetrix Inc., Santa Clara, CA, USA) were used for gene expression profiling. The resulted arrays were scanned with GeneChip Scanner 3000 7G (Affymetrix Inc.) while the normalization was performed with an invariant set of probe sets to adjust the overall signal level of the arrays to the same level.

Expression levels were calculated with a model-based PM-only approach from the dChip software [213].

Both tumor and stroma data sets contain 20184 genes which were measured in 38 matched samples.

2.5.5 Cross-Species Comparison between *Pichia pastoris* and Chinese Hamster Ovary Cells

The cross-species comparison between *P. pastoris* and CHO is based on two model proteins with different complexities which challenge the expression systems in various ways: human serum albumin (HSA), a monomeric and non-glycosylated protein and a more complex model protein, a single chain Fv-Fc fusion antibody (3D6scFv-Fc) derived from the monoclonal anti-HIV-1 antibody 3D6 which is homodimeric and contains the Fc-specific glycosylation.

The first part of the data generation was performed by our collaborators Andreas Maccani and Nils Landes under the supervision of Prof. Diethard Mattanovich from the Austrian Centre of Industrial Biotechnology and included model protein construction which resulted in a high and a low producer strain for each model protein and production system as well as the subsequent fed batch cultivation which was performed in a comparable regime allowing a quantitative comparison.

Microarray experiments were performed by our collaborators Niels Landes, Nadine Tatto and Alexandra Graf from the Austrian Centre of Industrial Biotechnology. A CHO specific DNA microarray (Agilent 4x44k design) based on the published and fully annotated genomic sequence of the CHO-K1 cell line [214] was designed. For transcriptome analysis of *P. pastoris* samples a custom DNA microarray (Agilent 8x15k design) was used. For both production systems total RNA samples from five strains (3D6scFv-Fc low and high producer, HSA low and high producer and non producer) were analyzed in biological triplicates (samples of three independent fermentations) and technical duplicates (dye-swap). For each production system, a reference sample was generated by pooling equal amounts of total RNA from all 15 biological replicates for each production system. Raw data were processed with the Agilent Feature Extraction software (v11.0). Since a clear intensity dependent dye bias was seen in the data, it was decided to use Loess normalization after correcting for low level background noise. Intensity dependent dye bias is commonly seen in 2-color arrays due to the different properties of Cy3 and Cy5. Loess normalization counteracts this systematic bias. Differentially expressed genes were determined with the *limma* Bioconductor package [215], using a linear model fit and the ebayes function to calculate significance levels. Since in microarray analysis a large amount of independent tests are made (one for each gene in the dataset), the significance values need to be corrected for multiple testing. For this purpose the Benjamini-Yekutieli correction [216] was used to compute adjusted p-values.

Proteomics experiments were performed by Clemens Gruber from the Austrian Centre of Industrial Biotechnology under the supervision of Prof. Friedrich Altmann. For both production systems, three biological and three technical replicates of each producing clone were compared with an empty vector control using 2D-LC-ESI MS/MS with prior TMT labelling.

Protein quantification and statistical evaluation was done by Gerda Modarres and Alexandra Graf from the Austrian Centre of Industrial Biotechnology. The R package *isobar* [217] provides methods for preprocessing, normalization and report generation for the analysis of quantitative mass spectrometry proteomics data tagged with isobaric labels, such as iTRAQ and TMT. In the first step, sample normalization is performed in order to standardize median intensities in all reporter channels. Subsequently, differentially expressed proteins are computed.

Finally, four data sets emerged containing: 5254 *P. pastoris* genes, 1955 *P. pastoris* proteins, 20650 CHO genes and 609 CHO proteins. These data sets were measured in all four strains: 3D6 low producer (3D6 L:WT), 3D6 high producer (3D6 H:WT), HSA low producer (HSA L:WT) and HSA high producer (HSA H:WT).

3 Results

The current chapter summarizes the results of this doctoral thesis. The analysis of the life cycle stages is presented in the first section followed by the results of the cross-species comparison between *P. pastoris* and CHO cells. Subsequently, the results of the integrative analysis of the immune response of *A. gambiae* are shown. The forth section describes the results of a traditional gene set enrichment approach applied to miRNAs enriched in patients with traumatic brain injury. The last part of this chapter is dedicated to the new pathway enrichment method that was developed and the results obtained through its application to microdissected tumor and stroma gene expression profiles.

3.1 *Plasmodium falciparum* Lifecycle Stages Analysis

This section presents the results of the triple integrative analysis of the transcriptome and proteome of *P. falciparum* as published in [146]. The results of each analysis method can be divided into method-specific associations and general associations. The general associations are used to detect results which are common to all three methods.

3.1.1 Co-Inertia Analysis Results

With CIA the six life cycle stages in the gene and protein space (see Figure 3.1) are visualized. We observe that the co-inertia x axis separates the intraerythrocytic cycle stages (trophozoite, ring, schizont, merozoite) from gametocytes and sporozoites. In the erythrocytes, the cycle begins with the ring stage, followed by the trophozoite stage. Trophozoites mature into schizonts, which cause the rupture of blood cells resulting in the release of merozoites. This exact sequence within the intraerythrocytic cycle can be observed in Figure 3.1. The sporozoites are the sexual stage of the mosquito and will be released in the blood stream of the infected organism. The ring stage can develop into a gametocyte and can be ingested by a mosquito. In addition to the life cycle stages, GO terms can also be represented through projections in the CIA plot (see Figure 3.5).

General associations

General associations resulting from CIA are distributed as follows: In gene space, GO terms in the first trigonometric quadrant are associated with trophozoites, GO terms in the second quadrant with gametocytes and GO terms in the third quadrant with sporozoites. GO terms in the first and forth quadrant, which were not identified as specific for trophozoites are

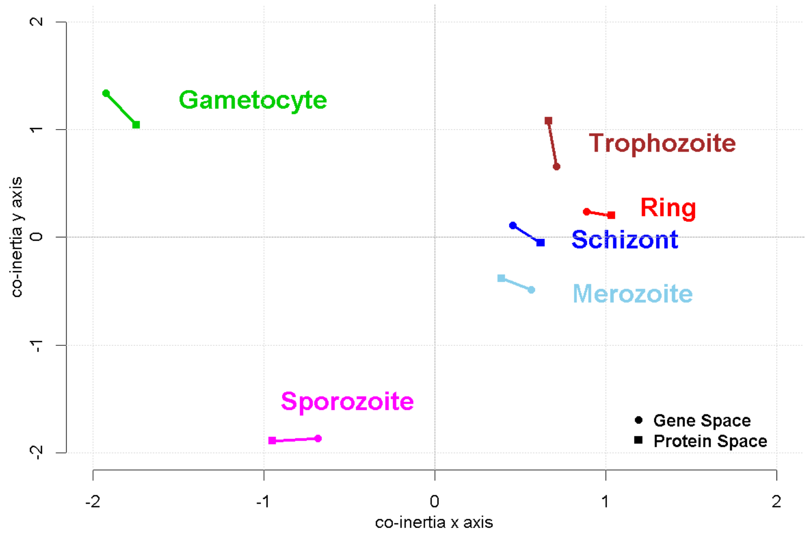


Figure 3.1: CIA offers the possibility to visualize the gene and protein space projections of the six life cycle stages of *P. falciparum* in one plot. The projection in gene space are represented by circles and in the protein space by squares. For each life cycle stage, the two corresponding projections are connected through a line. We observe that the co-inertia x axis separates the intraerythrocytic cycle from the stages gametocyte and sporozoite.

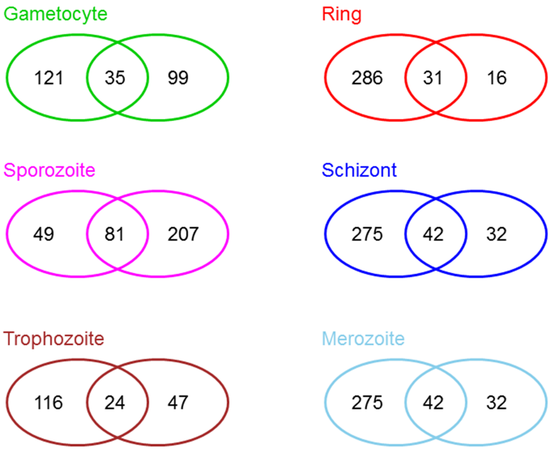


Figure 3.2: CIA general associations. Overlap between gene and protein space. For each life cycle stage, the left ellipse shows the number of general GO term associations in gene space whereas the right ellipse shows the number of general GO term associations in protein space. The amount of identical GO terms is shown in the overlapping region of the ellipses. In general, more GO term associations emerge form gene space than from protein space. Two exceptions can be observed: the sporozoite stage, where more associations are found in protein space and the gametocyte stage, where a similar number of associations is found in each space. These tendencies can be also observed in Figure 3.5.

associated with rings, schizonts and merozoites. Due to the proximity of stages in the CIA gene space, a more specific distribution to each stage is not possible.

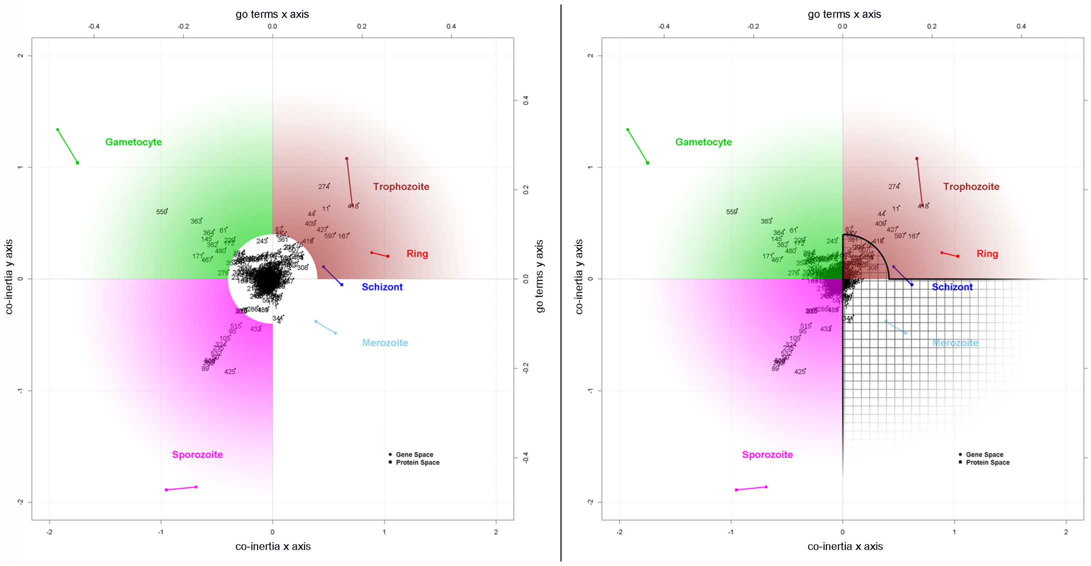


Figure 3.3: CIA division limits in gene space. CIA division limits for the general (left) and specific (right) associations in gene space. The colors of the areas correspond to the colors of the stages they are associated with.

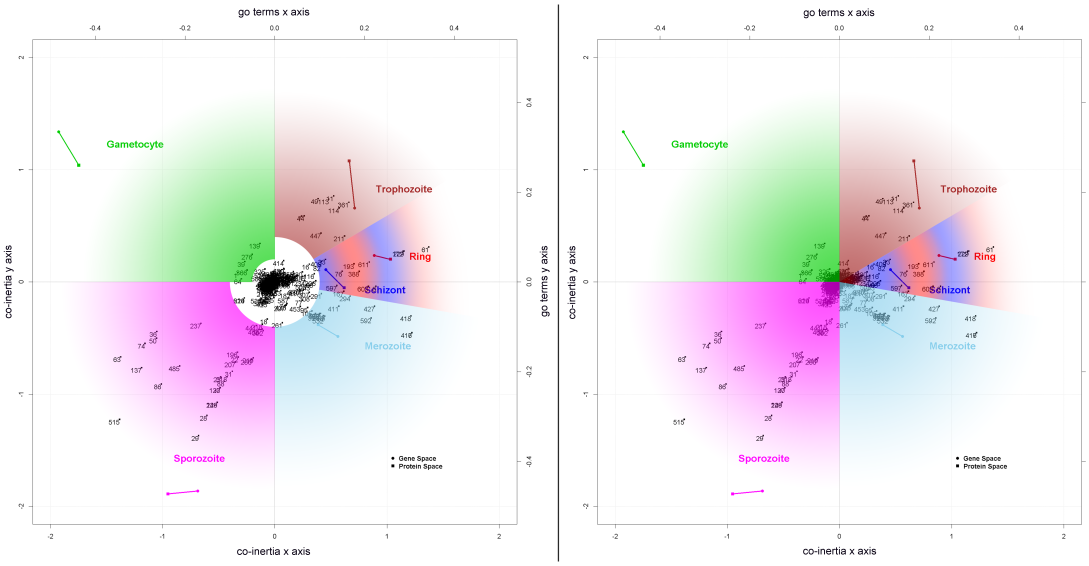


Figure 3.4: CIA division limits in protein space. CIA division limits for the general (left) and specific (right) associations in protein space. The colors of the areas correspond to the colors of the stages they are associated with.

In protein space the associations are produced as follows: For gametocytes and sporozoites, we follow the same criteria as in gene space. In order to associate GO terms to trophozoites, rings, schizonts and merozoites, we divide the first and forth quadrant in three sectors. GO terms that form angles of at least 30 degrees with the positive co-inertia x axis are associated with trophozoites. GO terms with an angle between -10 and 30 degrees are associated with rings and schizonts. GO terms with an angle wider than -10 degrees are associated

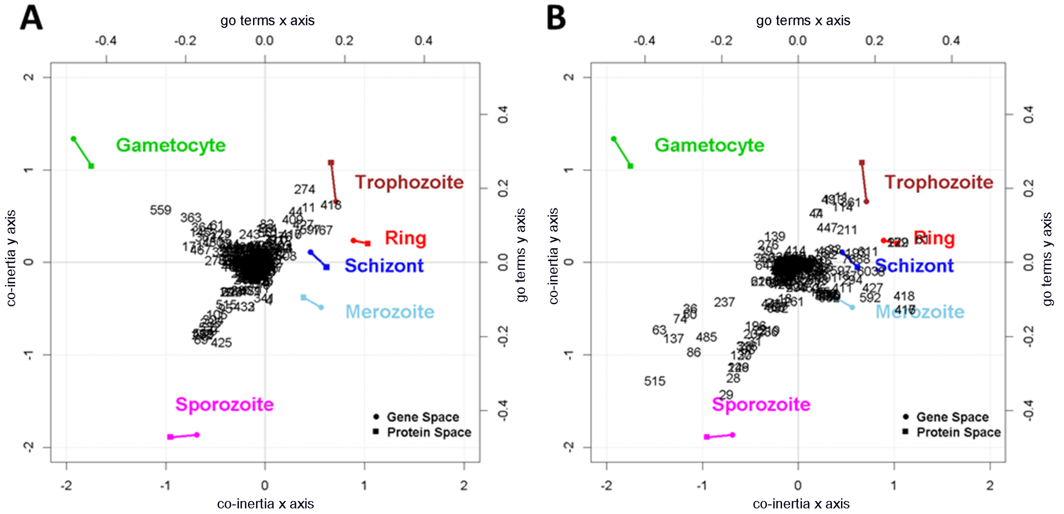


Figure 3.5: Co-inertia analysis and GO terms - results. In addition to the life cycle stages, GO terms can also be projected into the CIA plot. A) projections of the GO terms in gene space and B) projections of the GO terms in the protein space. Each GO term is represented by a number for readability reasons. Please note that the life cycle stages and the GO terms are plotted on different scales. The lower and left axes represent the life cycle stages (co-inertia x and y axis) and the upper and right axes represent for the GO terms (co-inertia go x and y axis). In gene space we observe a clear projection of the GO terms in the direction of gametocytes and sporozoites. In protein space, GO terms are projected clearly in direction of sporozoites and the intraerythrocytic cycle.

with merozoites. As GSVD and IBC discover associations only in common space (gene and protein space), the CIA associations for each life cycle are computed as the set union of the associations in gene and the associations in protein space. These general associations are shown in Additional files 2 and 3 of [146]. Detailed representations of the division limits for the specific and general associations are shown in Figures 3.3 and 3.4. The overlap between the general association in gene and protein space are shown in Figure 3.2.

Method-specific associations

In addition to the general results, method-specific associations of GO terms with life cycle stages are observed. For these associations, the direction of the projected GO terms is considered. From the general associations, those GO terms are taken that have a distance of at least 0.1 to the origin of the coordinate systems. An exception is made for gametocytes in protein space. A threshold of 0.05 is more appropriate here due to the spacial distribution of GO terms relative to the origin. These thresholds result in GO term associations with gametocyte, trophozoite and sporozoite stages in gene space. Details are presented in Table 3.1 and Additional file 5 of [146]. In the protein space, clear GO term associations with gametocyte, sporozoite, trophozoite and merozoite stages are found (Table 3.2 and Additional file 6 of [146]). This file also includes associations with the stages ring and schizont.

Based on Figure 3.5 some of the most remarkable associations in gene space are: GO:0006071 *glycerol metabolic process* (559) and GO:0002720 *positive regulation of cytokine production* (363)

Table 3.1: CIA specific GO term association to the life cycle stage gametocyte in gene space. The index corresponds to numbers in Figure 3.5A.

Index	GO Term ID and Description
61	GO:0006334 nucleosome assembly
145	GO:0006072 glycerol-3-phosphate metabolic process
171	GO:0006465 signal peptide processing
362	GO:0007131 reciprocal meiotic recombination
363	GO:0002720 regulation of cytokine production involved in immune response
364	GO:0006359 regulation of transcription from RNA polymerase III promoter
467	GO:0051604 protein maturation
480	GO:0001819 positive regulation of cytokine production
559	GO:0006071 glycerol metabolic process

Table 3.2: CIA specific GO term associations to the life cycle stage gametocyte in protein space. The index corresponds to the numbers in 3.5B.

Index	GO Term ID and Description
1	GO:0009405 pathogenesis
39	GO:0007165 signal transduction
64	GO:0007155 cell adhesion
139	GO:0045454 cell redox homeostasis
276	GO:0044262 cellular carbohydrate metabolic process
366	GO:0006103 2-oxoglutarate metabolic process

for gametocytes; GO:0006101 *citrate metabolic process* (425) and GO:0016255 *attachment of GPI anchor to protein* (89) for sporozoites; GO:0006591 *ornithine metabolic process* (274) and GO:0006094 *gluconeogenesis* (418) for trophozoites. In protein space we observe: GO:0045454 *cell redox homeostasis* (139) and GO:0044262 *cellular carbohydrate metabolic process* (276) for gametocytes; GO:0006928 *cellular component movement* (515) and GO:0015991 *ATP hydrolysis coupled proton transport* (29) for sporozoites; GO:0006412 *translation* (11) and GO:0019538: *protein metabolic process* (361) for trophozoites; GO:0006334 *nucleosome assembly* (61) and GO:0050776 *regulation of immune response* (8) for ring and schizonts; GO:0042594 *response to starvation* (592), GO:0000045 *autophagic vacuole assembly* (416) and GO:0002253 *activation of immune response* (418) for merozoites.

While the overlap of general CIA GO term associations between gene and protein space is moderate (Figure 3.2), the overlap of specific CIA GO term associations between gene and protein space is modest: Three GO terms were projected in the direction of trophozoites in gene and in protein space: GO:0006412 *translation* (11), GO:0006414 *translational elongation* (44) and GO:0044257 *cellular protein metabolic process* (114).

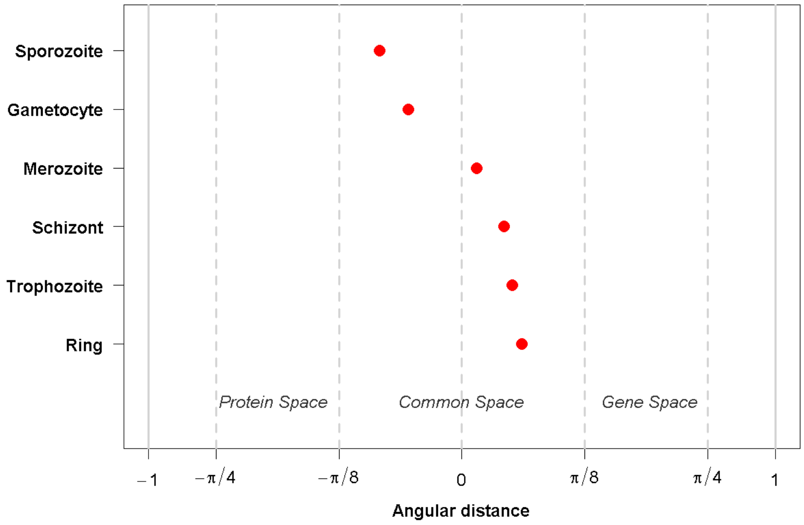


Figure 3.6: Angular distances of generalized singular value decomposition. In general, the angular distances map to the common space.

3.1.2 Generalized Singular Value Decomposition Results

As the final step of the GSVD, a restrictive gene set enrichment analysis (GSE) is performed. The type of performed GSE analysis is based on the angular distance that encodes for each life cycle stage the significance of the gene set relative to the protein set. If the angular distances are between $-\frac{\pi}{8}$ and $\frac{\pi}{8}$, then the gene and protein data sets are of equal significance, and the GSE is conducted in the common space. The common space is defined by both the gene and the protein data set. In the current analysis all life cycle stages (Figure 3.6) are assigned to the common space.

Based on the angular distance, a separation of the intraerythrocytic cycle (angular distances greater than zero) from other stages (angular distances less than zero) is possible. Following the workflow introduced by Alter *et al.* in [127], the restricted GSE performs a GSE for each life cycle stage on 50% of the genes and proteins that present the highest absolute values in the corresponding arraylets.

General associations

All resulting GO terms having a p-value smaller than 0.05 are considered to be general associations. These GO terms are shown in Additional file 7 of [146].

Method-specific associations

The method-specific GO terms are a subset of the general associations consisting of the top 15 GO terms, with the smallest p-values. The top 15 GO terms were chosen because this

Table 3.3: GSVD specific associations to gametocyte stage in the common space.

GO Term ID	GO Term Description
GO:0044238	primary metabolic process
GO:0008152	metabolic process
GO:0044237	cellular metabolic process
GO:0045017	glycerolipid biosynthetic process
GO:0043170	macromolecule metabolic process
GO:0034645	cellular macromolecule biosynthetic process
GO:0046474	glycerophospholipid biosynthetic process
GO:0009059	macromolecule biosynthetic process
GO:0022613	ribonucleoprotein complex biogenesis
GO:0044260	cellular macromolecule metabolic process
GO:0019538	protein metabolic process
GO:0046486	glycerolipid metabolic process
GO:0042254	ribosome biogenesis
GO:0006839	mitochondrial transport
GO:0009987	cellular process

number mirrors approximately the number of CIA specific associations. In this way a fair comparison can be performed.

The method-specific associations are presented in Tables 3.3 and 3.4 and in Additional File 8 of [146]. Biologically relevant associations include: GO:0051805/ GO:0051807 *evasion or tolerance of immune/defense response of other organism involved in symbiotic interaction*, GO:0051832 *avoidance or defenses of other organism involved in symbiotic interaction*, and GO:0052173 *response to defenses (immune response) of other organism involved in symbiotic interaction* for trophozoites and schizonts. The other stages are associated with more general GO terms such as GO:0044237 *cellular metabolic process*, GO:0019538 *protein metabolic process* and GO:0046474 *glycerophospholipid biosynthetic process*.

3.1.3 Integrative Biclustering Results

The IBC results include two types of biclusters: (i) biclusters containing genes, proteins, GO terms and life cycle conditions and (ii) biclusters containing genes, proteins and GO terms. Since we are interested in GO terms associations with life cycle stages, only the first type of biclusters will be used for further analysis. If a GO term is in the same bicluster as a life cycle stage, this GO term is associated with that life cycle stage. If there are more life cycle stages in a bicluster, the GO terms are associated with all these life cycle stages. If a life cycle stage is included in more than one bicluster, GO terms from all biclusters are

Table 3.4: GSVD specific associations to trophozoite stage in the common space.

GO Term ID	GO Term Description
GO:0044403	symbiosis, encompassing mutualism through parasitism
GO:0044419	interspecies interaction between organisms
GO:0051704	multi-organism process
GO:0009607	response to biotic stimulus
GO:0006952	defense response
GO:0051707	response to other organism
GO:0051805	evasion or tolerance of immune response of other organism involved in symbiotic interaction
GO:0051807	evasion or tolerance of defense response of other organism involved in symbiotic interaction
GO:0051832	avoidance of defenses of other organism involved in symbiotic interaction
GO:0051834	evasion or tolerance of defenses of other organism involved in symbiotic interaction
GO:0052173	response to defenses of other organism involved in symbiotic interaction
GO:0052564	response to immune response of other organism involved in symbiotic interaction
GO:0020033	antigenic variation
GO:0051809	passive evasion of immune response of other organism involved in symbiotic interaction
GO:0006091	generation of precursor metabolites and energy

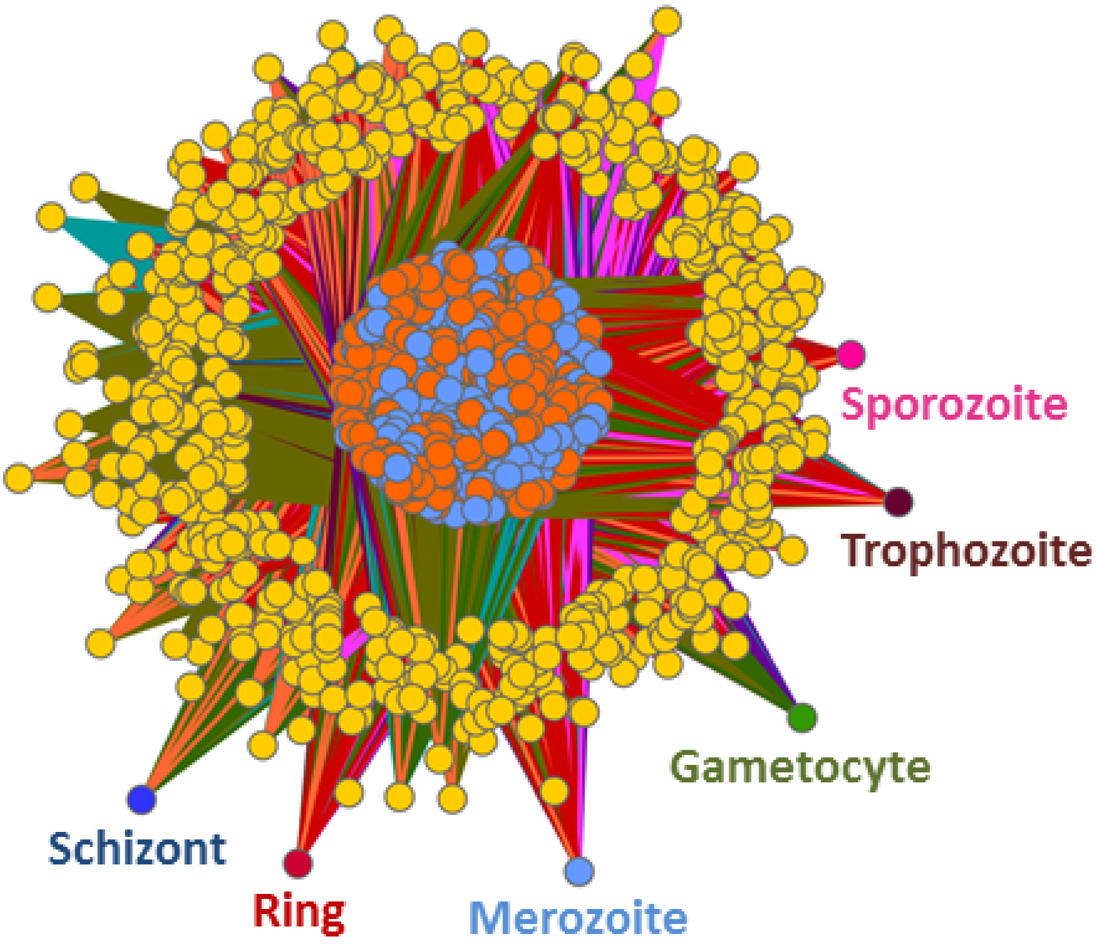


Figure 3.7: Integrative biclustering - network view of the results. The results of IBC were inspected and only biclusters including life cycle stages were considered for further analysis. IBC discovered 20 clusters of which 9 contained life cycle stages and GO terms. These 9 biclusters were processed and fed into Cytoscape. An association between a life cycle stage and a GO term is represented by an edge. Different biclusters are represented by different edge colours. The life cycle stages are shown in the same colours as those used for CIA and GSVD. Genes are coloured in orange, proteins in light blue and GO terms in yellow.

associated with that life cycle stage. IBC discovered 20 biclusters and 9 of them contained life cycle stages and GO terms. A network view of the results is shown in Figure 3.7.

General and method-specific associations

Since a life cycle stage is either included in a bicluster or not and as a consequence is either associated to a GO term or not, it is not possible to distinguish between general and method-specific associations. Figure 3.7 shows a vast amount of genes (in orange), proteins (in light blue), GO terms (in yellow) and the six life cycle stages: gametocyte (in green), sporozoite (in pink), trophozoite (in brown), ring (in red), schizont (in dark blue) and merozoite (in light blue). The 9 different biclusters which included life cycle stages can be identified through the colour of their edges (Figure 3.7). The exact associations with the life cycle stages are shown in Additional file 9 of [146].

3.1.4 Results Identified by CIA, GSVD and IBC

In this section we present the GO associations which were obtained by intersecting the results computed by CIA, GSVD and IBC. These associations are termed common associations and are shown in Figure 3.8. They are based on gene as well as protein information and are therefore considered to be in the common space.

The GO associations tables computed in R were converted into a csv-format and loaded into Cytoscape. We observe here that the gametocytes are linked to the rest of the network through only one general GO term, GO:0009987 *cellular process*. The sporozoite stage is also loosely connected to the network through two GO terms, GO:0009056 *catabolic process* and GO:0009116 *nucleoside metabolic process*.

The intraerythrocytic cycle, composed of trophozite, ring, schizont and merozoite are highly interconnected. The merozoite stage presents a high number of associations with specific GO terms such as GO:0030260 *entry into host cell* and GO:0044409 *entry into host*. Trophozoites are associated with a small number of GO terms, including GO:0050896 *response to stimulus*, GO:0006096 *glycolysis*, GO:0006006 *glucose metabolic process* and GO:0006091 *generation of precursor metabolites and energy*. The stages schizont and ring are connected through the GO terms GO:0006955 *immune response*, GO:0050776 *regulation of immune response*, GO:0006325 *chromatin organization* and GO:0006091 *generation of precursor metabolites and energy*. It is also interesting to see that merozoites and schizonts are linked only by the GO term GO:0009116 *nucleoside metabolic process*.

Relative proportions of common and methods-specific results

In the case of CIA, one can observe a high overlap between the common results and the CIA specific GO terms associations: 8 GO terms (GO:0005975 *carbohydrate metabolic process*, GO:0006644 *phospholipid metabolic process*, GO:0008654 *phospholipid biosynthetic process*, GO:0045017 *glycerolipid biosynthetic process*, GO:0006661 *phosphatidylinositol biosynthetic process*, GO:0046488 *phosphatidylinositol metabolic process*, GO:0006506 *GPI anchor biosynthetic process*,

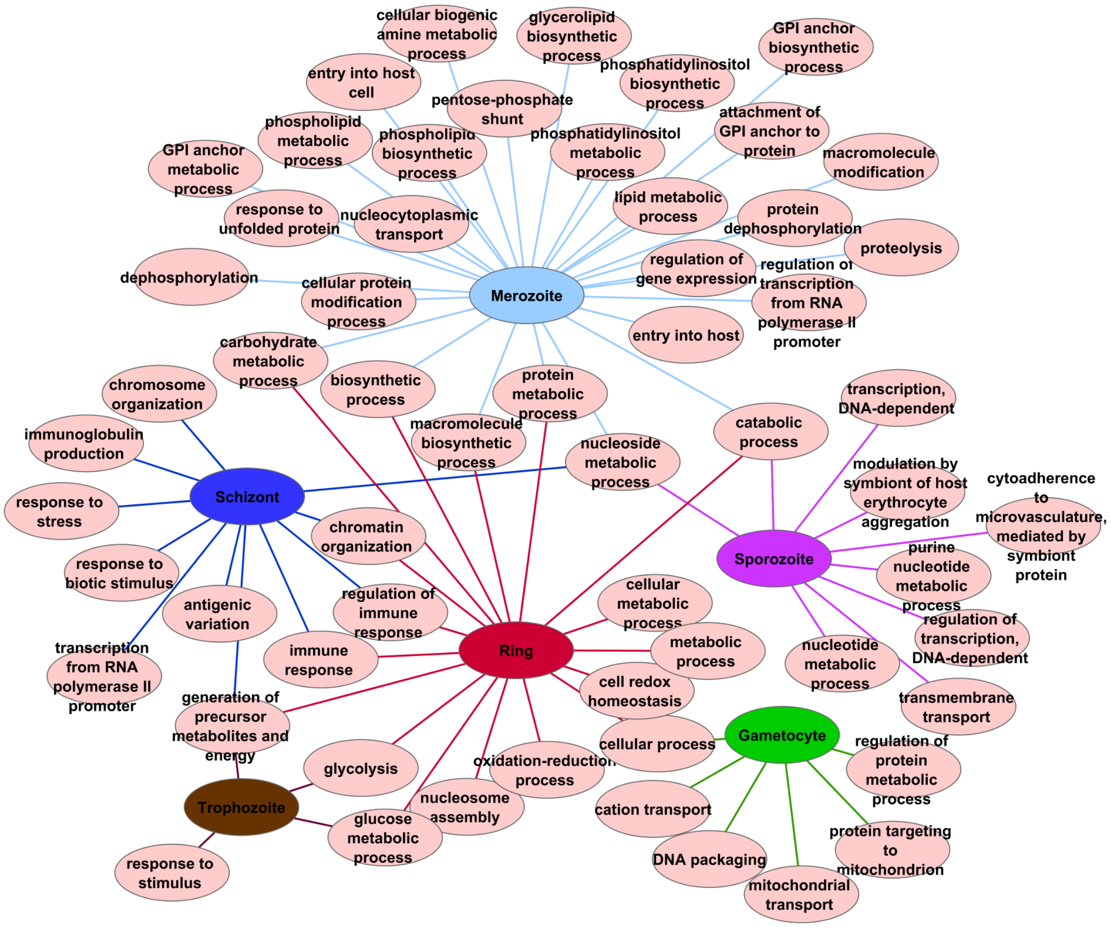

cellular biogenic amine metabolic process
glycerolipid biosynthetic process
GPI anchor biosynthetic process
entry into host cell
phosphatidylinositol biosynthetic process
pentose-phosphate shunt
phospholipid metabolic process
phospholipid biosynthetic process
phosphatidylinositol metabolic process
attachment of GPI anchor to protein
macromolecule modification
GPI anchor metabolic process
lipid metabolic process
protein dephosphorylation
response to unfolded protein
nucleocytoplasmic transport
proteolysis
regulation of gene expression
regulation of transcription from RNA polymerase II promoter
dephosphorylation
cellular protein modification process
Merozoite
entry into host
carbohydrate metabolic process
chromosome organization
biosynthetic process
protein metabolic process
transcription, DNA-dependent
immunoglobulin production
macromolecule biosynthetic process
nucleoside metabolic process
catabolic process
modulation by symbiont of host erythrocyte aggregation
cytoadherence to microvasculature, mediated by symbiont protein
Schizont
response to stress
chromatin organization
purine nucleotide metabolic process
Sporozoite
response to biotic stimulus
antigenic variation
regulation of immune response
cellular metabolic process
regulation of transcription, DNA-dependent
transcription from RNA polymerase II promoter
immune response
Ring
metabolic process
nucleotide metabolic process
cell redox homeostasis
transmembrane transport
generation of precursor metabolites and energy
cellular process
Gametocyte
regulation of protein metabolic process
glycolysis
oxidation-reduction process
Trophozoite
nucleosome assembly
cation transport
glucose metabolic process
protein targeting to mitochondrion
response to stimulus
DNA packaging
mitochondrial transport

Figure 3.8: GO terms to life cycle associations discovered by all three methods. Network view of the GO term to life cycle stage associations discovered by all three integrative analysis methods: CIA, GSVD and IBC. We observe that gametocytes and sporozoites are loosely connected to the rest of the network, underlining the separation of these stages from the intraerythrocytic cycle. Merozoites possess the largest amount of GO term associations, while trophozoites show the lowest amount of associations. Further details concerning individual stage-to-GO-term mappings are addressed in the discussion.

GO:0016255 *attachment of GPI anchor to protein*) associated by CIA with merozoites in protein space, 8 GO terms (GO:0009058 *biosynthetic process*, GO:0051276 *chromosome organization*, GO:0006325 *chromatin organization*, GO:0050776 *regulation of immune response*, GO:0006955 *immune response*, GO:0006096 *glycolysis*, GO:0006334 *nucleosome assembly*, GO:0044237 *cellular metabolic process*) associated by CIA with rings and schizonts in protein space and 3 GO terms (GO:0009117 *nucleotide metabolic process*, GO:0006163 *purine nucleotide metabolic process*, GO:0009116 *nucleoside metabolic process*) associated by CIA with sporozoites in protein space. Only two GO terms (GO:0006096 *glycolysis* and GO:0006006 *glucose metabolic process associated with trophozoites*) from gene space, coincide with GO terms from the common results. Protein activity characteristics derived from CIA show considerable similarities to the other two methods.

Six specific results of GSVD for the life cycle stage ring coincide with the common GO term associations to this stage (GO:0009058 *biosynthetic process*, GO:0019538 *protein metabolic* *process*, GO:0044237 *cellular metabolic process*, GO:0008152 *metabolic process*, GO:0055114 *oxidation-reduction process*, GO:0006091 *generation of precursor metabolites and energy*). There are three identical associations for the stage merozoite (GO:0019538 *protein metabolic process*, GO:0016311 *dephosphorylation* and GO:0006470 *protein dephosphorylation*). For each of the other stages, only one GO term from the common associations coincides with the method-specific associations (GO:0009987 *cellular process* for gametocytes, GO:0006091 *generation of precursor metabolites and energy* for trophozoites, GO:0020033 *antigenic variation* for schizonts and GO:0009056 *catabolic process* for sporozoites). In conclusion, the ring stage is very well characterized by the GSVD, which is almost in complete agreement with the other methods. The properties of the other stages do not coincide with the common results but should definitely be considered for further analysis as they are highly significant.

3.2 Analysis of Hemocyte and Granulocyte Immune Response of *Anopheles gambiae*

In this section the structural concordance between the proteomic profiles of the granulocyte cell subset and the transcriptomic profile of the general hemocyte population is measured and the results are shown as described in the co-authored publication [205].

To determine the co-structure between the proteomic and existing hemocyte transcriptomic profiles, candidate genes responsive to granulocyte-enrichment during sugar-feeding (SF), blood-feeding (BF), or *P. falciparum* infection (PF) were further examined by multiple co-inertia analyses (MCIA) (Figure 3.9 - 3.11). Using published hemocyte transcriptome data [206], MCIA was used to examine the degree of agreement between transcript and protein abundance in the granulocyte proteomes (SF, BF, PF). Comparisons were made between transcriptional profiles of non-selected hemocytes from sugar-fed naive mosquitoes, 24 hours after feeding with a non-invasive CTRP mutant *Plasmodium berghei* (comparable to a non-infectious blood meal), or 24 hours after feeding with wild-type *P. berghei*.

Hemocyte transcript and protein profiles analyzed by MCIA were based on the log fold change between two treatments (SF, BF and PF) and are displayed as the end points of a segment. The more similar the two profiles are, the shorter the segment; since the length is proportional to the divergence between the two datasets. If the two profiles were identical, the length of the segment would be zero. Individual transcripts or proteins that define the comparison are highlighted for each MCIA plot (Figures 3.9, 3.10 and 3.11).

The MCIA plots are based on the first two MCIA axes (Sample Axis 1 and Sample Axis 2). MCIA Sample Axis 1 accounted for 84% to 91% of the explained covariance. Therefore, the shown plots represent a good approximation of two data sets. Additionally, transcripts and proteins (features of these data sets) were projected into the same plane (Feature Axis 1 and Feature Axis 2). The scaling of the Sample Axes is different from the scaling of the Feature Axes. This was done in order to obtain an easier to read plot and to better illustrate the associations of features to samples.

Despite differences in sample collection, sampling time points, and the species of malaria parasite used, our MCIA analysis revealed a high level of concordance between the

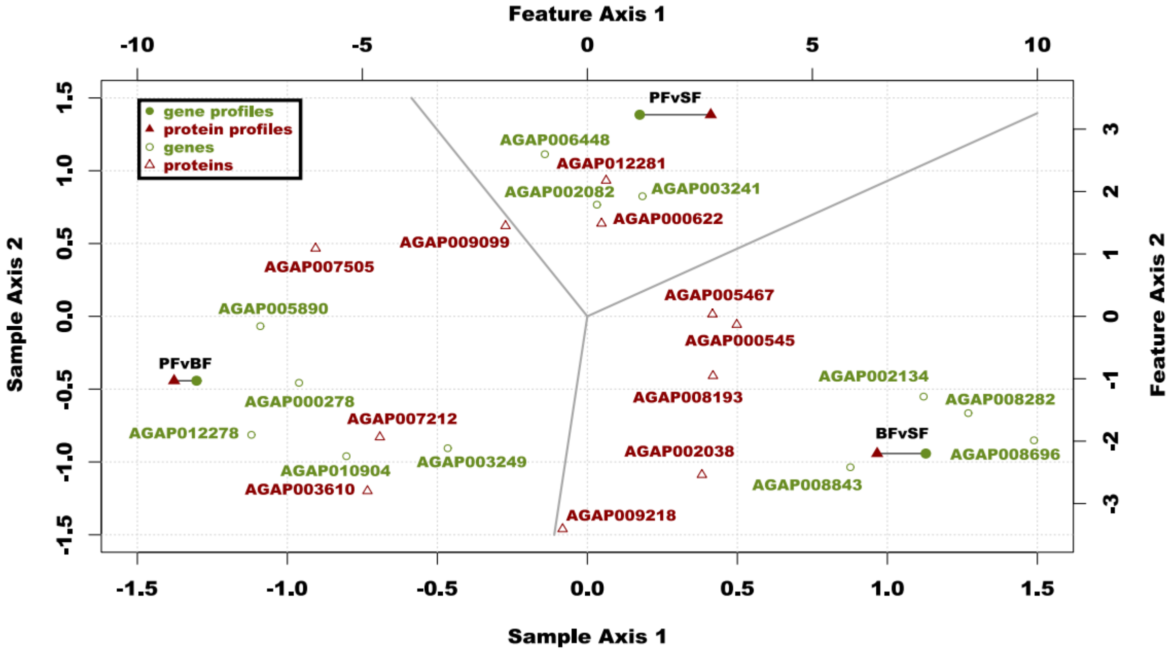


Figure 3.9: Hemocyte-specific MCIA. The hemocyte transcriptome is compared to the granulocyte proteome (Table S3, tab A). RV-coefficient = 0.97. Transcriptome (green circle) and proteome (red triangle) profiles are displayed for each sample comparison (*P. falciparum* infection (PF), blood-feeding (BF), and sugar-feeding (SF)). The samples in this analysis were computed as log fold changes between two treatments: *P. falciparum* infection referenced to blood-feeding (PFvBF), *P. falciparum* infection referenced to sugar-feeding (PFvSF) and blood-feeding referenced to sugar-feeding (BFvSF). Additionally, the most highly expressed features (genes and proteins with the greatest distance from the origin) are projected in the MCIA result plots. Due to differences between the coordinates of the comparisons and of the most expressed features plots, different axes scaling was used.

previously published hemocyte transcriptomes and the enriched granulocyte protein profiles. Comparisons were performed to identify global similarities between hemocyte transcript and protein abundance profiles (Figure 3.9), and to examine sub-populations of immune-specific (Figure 3.10) or proliferation-specific (Figure 3.11) protein profiles. An overview with the highlighted genes and proteins is provided in Table 3.5 and 3.6 for the hemocyte specific analysis while the immune-specific and proliferation-specific associations are shown in the Appendix in Tables B.1 - B.4.

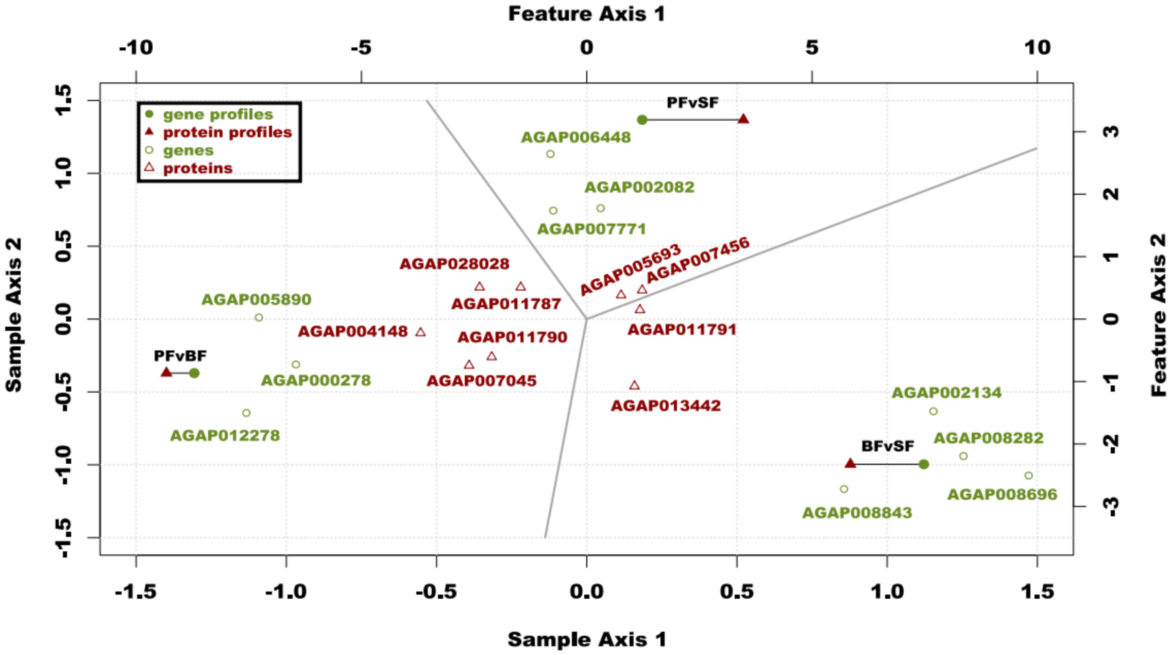


Figure 3.10: Immune-specific MCIA. The hemocyte transcriptome is compared to the immune-specific granulocyte proteome (Table S3, tab B). RV-coefficient = 0.96. For additional details please consult Figure 3.9.

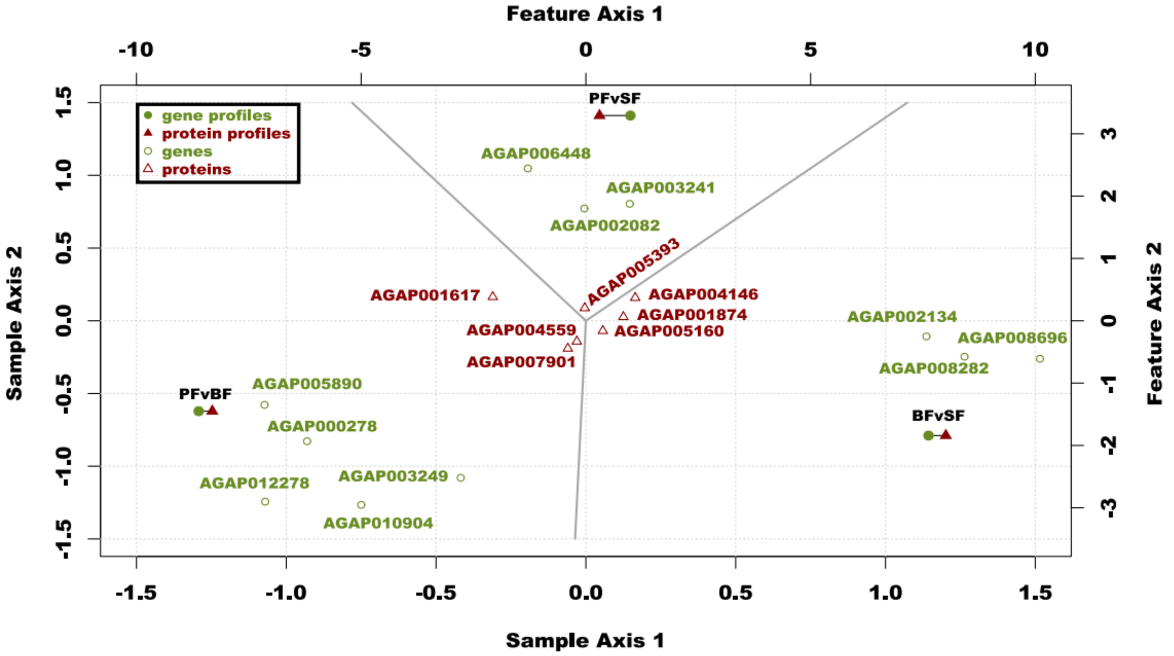


Figure 3.11: Proliferation-specific MCIA. The hemocyte transcriptome is compared to the proliferation-specific granulocyte proteome. RV-coefficient = 0.99. For additional details please consult Figure 3.9.

Table 3.5: Hemocyte-specific gene associations

Genes	Annotation	Feature Axis 1 [AU]	Feature Axis 2 [AU]
AGAP005890	no metadata	-7,27	-0,16
AGAP002134	no metadata	7,47	-1,29
AGAP012278	no metadata	-7,47	-1,90
AGAP008282	no metadata	8,46	-1,55
AGAP008696	no metadata	9,92	-1,99
AGAP000278	OBP9 odorant binding protein	-6,41	-1,07
AGAP006448	cAMP-dependent protein kinase regulator	-0,94	2,60
AGAP003249	CLIP3 Clip-domain serine protease	-3,10	-2,11
AGAP002082	no metadata	0,21	1,79
AGAP003241	no metadata	1,22	1,92
AGAP010904	CPFL3 cuticular protein 3 from CPFL family	-5,35	-2,24
AGAP008843	aquaporin 1	5,84	-2,42

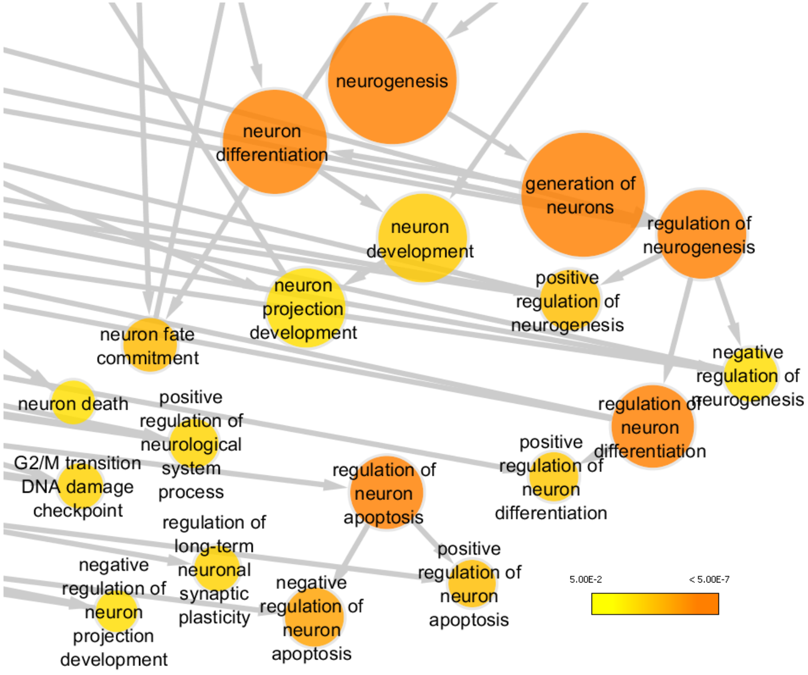

neurogenesis
neuron differentiation
generation of neurons
neuron development
regulation of neurogenesis
neuron projection development
positive regulation of neurogenesis
neuron fate commitment
negative regulation of neurogenesis
neuron death
positive regulation of neurological system process
regulation of neuron differentiation
G2/M transition DNA damage checkpoint
regulation of neuron apoptosis
positive regulation of neuron differentiation
regulation of long-term neuronal synaptic plasticity
positive regulation of neuron apoptosis
negative regulation of neuron projection development
negative regulation of neuron apoptosis
5.00E-2
< 5.00E-7

Figure 3.12: Overrepresented biological processes in the set of validated target genes of CSF-MP associated miRNAs

Table 3.6: Hemocyte-specific protein associations

Proteins	Annotation	Feature Axis 1 [AU]	Feature Axis 2 [AU]
AGAP000545	no metadata	3,32	-0,13
AGAP007212	ATP-dependent RNA helicase DHX8/PRP22	-4,61	-1,94
AGAP005467	vigilin	2,78	0,03
AGAP007505	vitellogenic carboxypeptidase-like protein	-6,04	1,09
AGAP003610	no metadata	-4,89	-2,80
AGAP008193	nidogen (entactin)	2,79	-0,95
AGAP009218	no metadata	-0,71	-3,41
AGAP002038	no metadata	2,55	-2,54
AGAP000622	no metadata	0,31	1,49
AGAP012281	Eukaryotic translation initiation factor 3 subunit M	0,42	2,18
AGAP009099	transglutaminase	-1,71	1,45

3.3 Gene Set Enrichment Analysis of Human miRNAs Following Traumatic Brain Injury

In this section the results of the GSEA applied to two sets of human miRNAs are presented. As described in [188], the 35 overrepresented biological processes in the set of validated target genes of cerebrospinal fluid microparticles (CSF-MP) associated miRNAs are shown. Additionally, enriched biological processes were assessed for all 63 identified miRNAs.

For the 35 CSF-MPs associated miRNAs, the validated target genes were identified with miRWalk. A total of 1363 experimentally validated target genes were found (see Supplement 4 of [188] for a complete list).

The analysis of the target gene associated GO terms yielded a hierarchical network consisting of 1761 nodes and 3131 edges. The nodes represent biological processes that are connected through directed edges. The 35 overrepresented neuron related biological processes and their computed p-values are presented in Table 4. The subset of the processes with a p-value smaller than 0.001 are shown in Figure 3.12.

For all 63 miRNAs (Supplement 3 of [188]), identified by Bioanalyser profiling, a total of 1659 validated target genes were identified (Supplement 5 of [188]). Validated target genes were not found for the following miRNAs: hsa-miR-1180, hsa-miR-1300, hsa-miR-923, hsa-miR-1182, hsa-miR-374a* and hsa-miR-431. The GO analysis predicted 1997 significantly enriched GO Terms, 36 of them being neuron related. Identified overrepresented neuron related biological processes, in CFSF-MP associated miRNAs as well as in all identified miRNAs, together with the corresponding p-values, can be found in Table C.1.

3.4 Development of an Integrative Gene Set Enrichment Analysis Method

This section describes the novel pathway centric gene set enrichment approach, suitable for integrative data analysis that was developed in the course of this thesis. The first part motivates the need for such an approach while the second part provides a detailed mathematical description.

While multiple approaches which test for pathway enrichment in a given list of genes (or gene products) have been described, the simplest being a one tailed Fisher's Exact Test or a test based on the hyper geometric distribution, others account for the rank of genes [195]. Traditional methods for determining functional pathway enrichment treat pathways as a list of elements, while ignoring their inherent connectivity (see 2.4 for details). Integrative approaches include: FunNet [218] which computes similarity between gene sets based on the co-expression networks of the measured data, clusters relevant sets and displays the results in a network overlaid with gene co-expression information; SPIA [145] which combines the enrichment of a pathway with its perturbation which is computed by propagating expression changes throughout the pathway; PARADIGM [143, 144] uses a factor graph with a set of interconnected variables encoding expression and activity of genes as well as other known gene products to predict patient centric pathway alterations and MONA [141] a multi-level ontology analysis based on a Bayesian network with two layers: an ontology term layer connected to a layer of hidden gene products as defined by the ontology used. These methods provide an approach to integrate multiple data types in the context of pathway enrichment but are computationally intensive and do not take into account the topology of a pathway. This information can be used to compute the information flow based importance of a gene in a pathway and, later in the analysis, account for it.

In this doctoral thesis an integrative, network based and pathway centric gene set enrichment approach is developed. This method is termed Integrative Pathway Enrichment Analysis (IPEA). IPEA integrates multiple data sets and/or data types via MCIA, collects and combines the most variant set of features from each set and correlates them with their biological pathway importance.

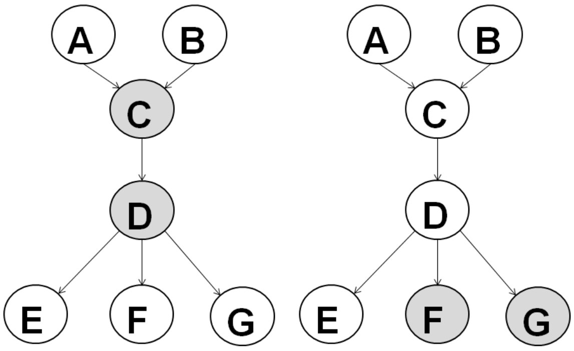


Figure 3.13: Effect of network topology on the activation of a network.

Using the pathways curated by the Reactome project [165], scores will be assigned to pathway elements based on their contribution to the information flow in the network. A toy network shown in Figure 3.13 emphasizes the fact that the topology of a network is very important when one wishes to compute its activation. All information flows through the nodes C and D in this toy network. Activation of C and D has a higher impact on the overall activation of the pathway than the activation of F and G which are end nodes and have no effect on other network nodes. This flow based approach [219] rewards both highly linked hubs and bottleneck nodes which may have few connections but bridge different clusters within a network. The enrichment score of a given pathway will then be quantified by correlating the scores from the network analysis with the MCIA scores.

3.4.1 Detailed Description of the Integrative Pathway Enrichment Analysis

IPEA is a network based approach for the enrichment of pathways which can be applied to the results of an integrative analysis method. In the scenario presented here, MCIA is employed as an integrative approach due to its ability to capture the most covariant features in the analyzed data sets. In practice, MCIA can be substituted by other integrative analysis methods appropriate for a specific study aim. Additionally, IPEA can be applied to individual analysis results such as the list of differentially expressed genes between two conditions. In general, IPEA can be applied to any gene list that has gene scores associated to it.

IPEA can be divided into the following steps (see Figure 3.14):

- computation of Reactome gene scores,
- computation of MCIA gene scores and
- computation of pathway enrichment scores from the Reactome and MCIA gene scores.

Computation of Reactome Gene Scores

Reactome [165] is an open source, manually curated and peer-reviewed database that summarizes the rich and already existent information on biological pathways in a computationally accessible format. The core entity of the database is the Reactome reaction. Distinct biological entities such as nucleic acids, proteins, complexes, vaccines, anti-cancer therapeutics and small molecules can participate in a reaction. One noteworthy advantage of Reactome is that each pathway can be represented as a network which allows the use of the corresponding network topology.

Access to the Reactome pathways is facilitated by the Bioconductor package *graphite* [220] which organizes the pathways in a list [221]. Each element of the list represents one pathway and includes its nodes (genes) and edges (connections between genes). From this, the adjacency matrix of the pathway is computed.

The information flow based scores of the genes in a pathway are defined by their dynamical importance (DI) which was shown to best characterize the importance of nodes in a network [219]. I_k, the dynamical importance I of node k is defined as the change (Δ) in the largest eigenvalue λ of the corresponding adjacency matrix upon removal of node k:

$$I_k \equiv \frac{\Delta_k}{\lambda}. \tag{3.1}$$

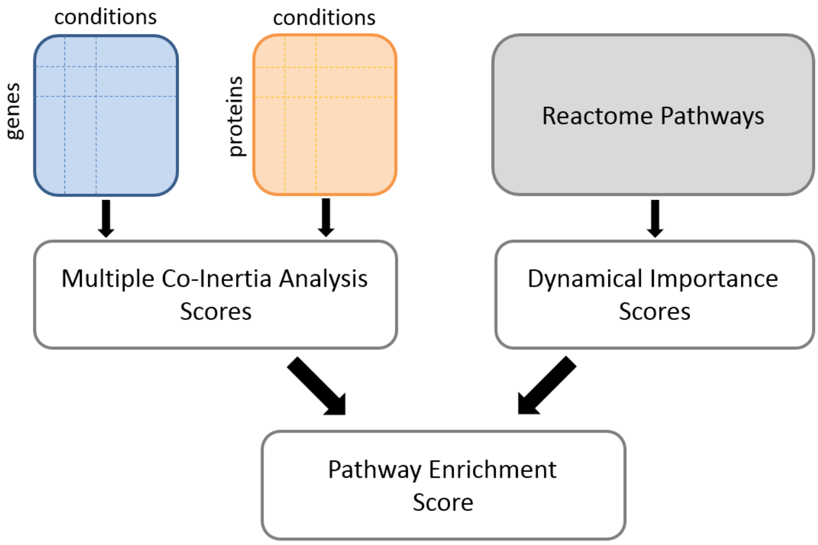


Figure 3.14: IPEA Workflow

Additionally, [219] provides an approximation for I_k, $\hat{I}_k$, that decreases the computation time for large networks:

$$\hat{I}_k = \frac{v_k u_k}{v^T u}, \tag{3.2}$$

where v_k and u_k are the k^{th} components of the left and right eigenvectors v and u corresponding to the largest eigenvalue λ.

Computation of MCIA Gene Scores

MCIA is implemented in the Bioconductor package *omicade4* [137]. The MCIA gene scores are defined as the coordinates of genes projected on the first MCIA axis which is associated to the highest eigenvalue and captures the most covariant features. Genes projected far away from the origin have a higher importance. Due to this, genes projected at the ends of the axis are of interest and, thus higher (absolute) MCIA scores are equivalent to more important genes.

In addition to the first MCIA axis, subsequent axes can be taken into account. Since each axis can capture different aspects of the data under study, conducting IPEA on these axes can reveal pathways associated with these aspects.

The genes of each data set analyzed by MCIA are characterized by their own MCIA gene scores. The corresponding Reactome scores, on the other hand, are equal for the studied data sets. These scores mirror the biological network topologies of the pathways which are independent of the measured data sets.

Computation of the Pathway Enrichment Score

The enrichment scores of each pathway in Reactome are calculated from the Reactome gene scores and the MCIA gene scores. These enrichment scores mirror whether the genes which are biologically driving a pathway (high DI) were also found to be most covariant in the data sets under study (high MCIA scores) and thus account for the captured co-structure.

The most suitable way to compute this agreement is Spearman's correlation. In contrast to Pearson's correlation, Spearman's correlation, which uses the ranks of the scores instead of the scores directly, is not effected by outliers known to artificially inflate the correlation.

By looking at the negative and the positive side of the MCIA axes separately, up and down regulated pathways can be computed. If the pathway was enriched on the negative side of the axis, the pathway is believed to be down regulated whereas a pathway which is enriched on the positive side of the axis is up regulated.

Statistical Significance of the Pathway Enrichment Score

In order to exclude high enrichment scores due to chance, the significance of the enrichment score is computed with a permutation test. This approach is motivated by the fact that the distribution of the enrichment scores under the null hypothesis (pathway is not enriched) is unknown.

The test constructs the null distribution by the permutation of the MCIA gene scores. From this distribution the p-value (probability of an equal or higher enrichment score) of the enrichment score can be computed as the number of permuted scores which are higher than the pathway score computed from the real data, divided by the total number of permutations.

3.5 Integrative Pathway Enrichment Analysis Applied to Ovarian Cancer

In this subsection the results of IPEA applied to microdissected tumor and stroma from ovarian cancer are presented. The section begins by exemplifying the ranking by DI in the Reactome pathway *TGF-beta receptor signaling in EMT*. The second part of the section is dedicated to the enriched pathways and corresponding genes in the microdissected tumor and stroma ovarian cancer samples.

3.5.1 Dynamical Importance Ranking in the Pathway TGF-beta receptor signaling in EMT

In order to examine in more detail the effects of the ranking imposed by the DI, the Reactome pathway *TGF-beta receptor signaling in EMT* is used as a example. The structure of the network can be seen Figure 3.15. It can be noticed that this is a strongly connected network with 15

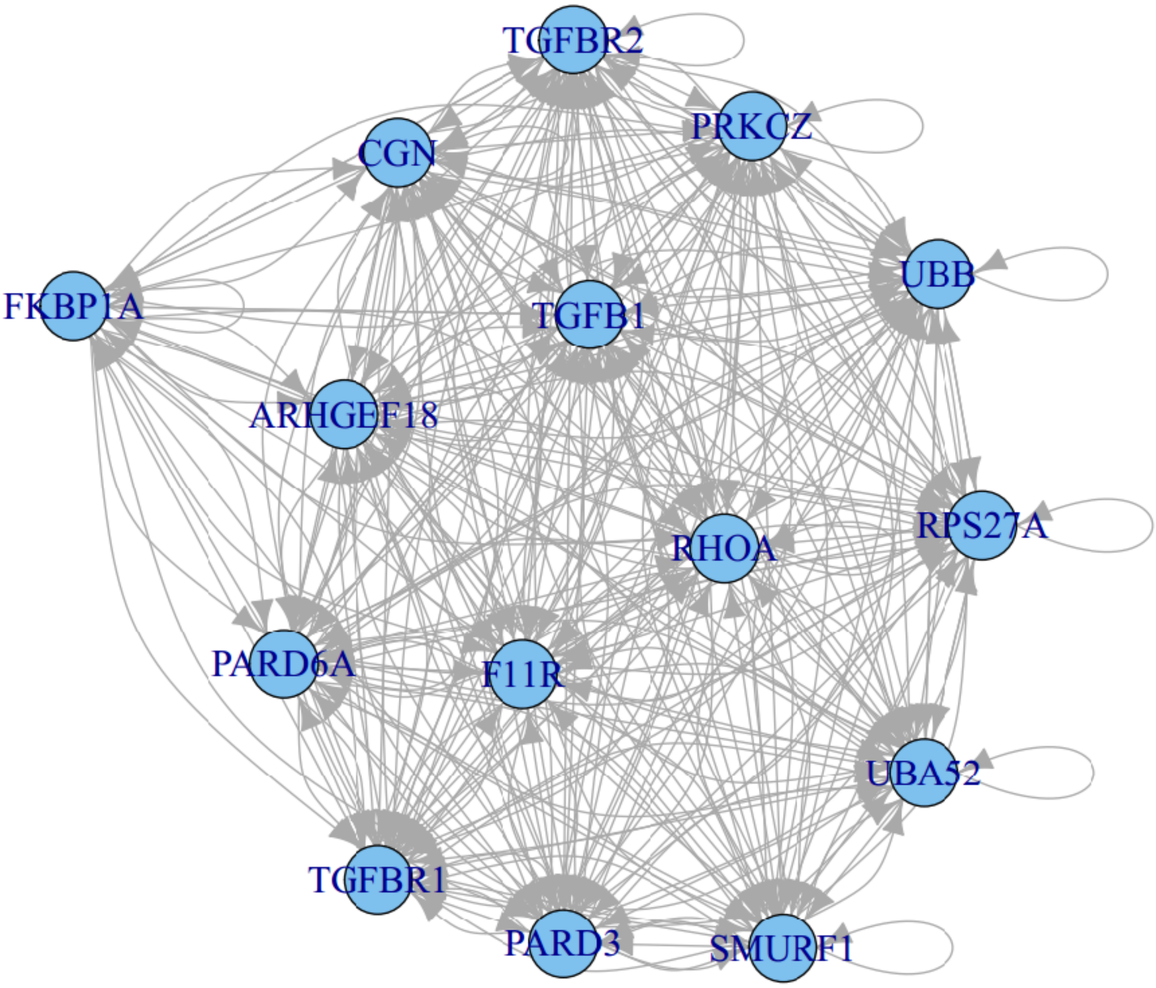


Figure 3.15: The TGF-beta receptor signaling in EMT pathway

nodes and 313 edges. The pathway was chosen because it is a strongly connected, cancer related pathway with a small enough number of nodes that can be visualized easily. A short description of the pathway is available in Reactome:

> *"In normal cells and in the early stages of cancer development, signaling by TGF-beta plays a tumor suppressive role, as SMAD2/3:SMAD4-mediated transcription inhibits cell division by downregulating MYC oncogene transcription and stimulating transcription of CDKN2B tumor suppressor gene. In advanced cancers however, TGF-beta signaling promotes metastasis by stimulating epithelial to mesenchymal transition (EMT). TGFBR1 is recruited to tight junctions by binding PARD6A, a component of tight junctions. After TGF-beta stimulation, activated TGFBR2 binds TGFBR1 at tight junctions, and phosphorylates both TGFBR1 and PARD6A. Phosphorylated PARD6A recruits SMURF1 to tight junctions. SMURF1 is able to ubiquitinate RHOA, a component of tight junctions needed for tight junction maintenance, leading to disassembly of tight junctions, an important step in EMT."*

Table 3.7 summarizes the results of the ranking based on DI. We notice that the highest ranked gene is TGFBR1. 7 genes are ranked as second most important: RHOA, PRKCZ, ARHGEF18, PARD3, PARD6A, CGN and F11R. The next ranked genes are TGFBR2, TGFB1, SMURF1 and RPS27A. UBB and UBA52 are ranked 7th followed by FKBP1A.

In order to have a better understanding of how the DI works, a plot of the dependence of the DI rank of a node on the corresponding geometric mean between in and out degree is shown in Figure 3.16. Here it can be noticed that the DI captures information about the centrality of a node in a network empathizing its suitability for our purposes, i.e., to account for hubs and bottlenecks in biological pathways.

Rank	Gene	DI
1	TGFBR1	0.07227
2	RHOA	0.07206
2	PRKCZ	0.07206
2	ARHGEF18	0.07206
2	PARD3	0.07206
2	PARD6A	0.07206
2	CGN	0.07206
2	F11R	0.07206
3	TGFBR2	0.07084
4	TGFB1	0.07077
5	SMURF1	0.06521
6	RPS27A	0.06038
7	UBB	0.05966
7	UBA52	0.05966
8	FKBP1A	0.03679

Table 3.7: Ranking of the nodes in the TGF-beta receptor signaling in EMT based on the dynamical importance.

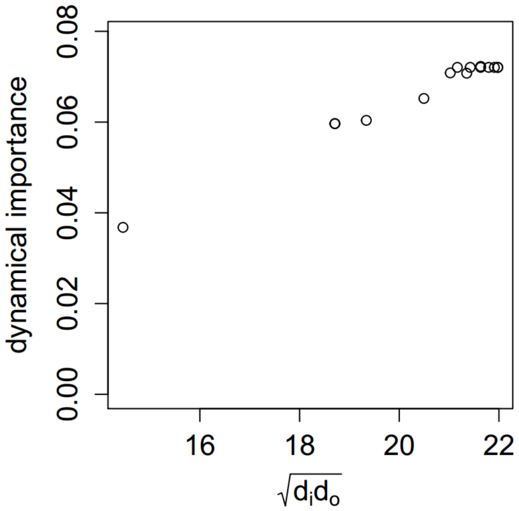


Figure 3.16: Dependence of the dynamical importance on the geometric mean of the in and out degree of the nodes exemplified in the TGF-beta receptor signaling in EMT.

3.5.2 IPEA of Matched Tumor and Stroma Samples

Here, IPEA is applied to microdissected tumor and stroma samples in order to better elucidate tumor and stroma cross-talk and to eventually discover new target genes for ovarian cancer therapy.

The results of IPEA applied to 38 matched microdissected tumor and stroma samples from patients with high grade serous ovarian cancer are shown. The first step of IPEA is the MCIA analysis of the tumor and stroma data set. In Figure 3.17 it can be observed that the matched tumor and stroma samples are very different. While some show similar tumor and stroma profiles, such as sample 13 and 34, other tumor and stroma profiles are very different, e.g. sample 26 and 21. This emphasizes the fact that stroma reacts differently to tumor cells and could be a new target in ovarian cancer therapy.

The enriched pathways in tumor and stroma are computed based on the scores of the tumor and stroma genes on the first MCIA axis. These scores are correlated with the DI scores of each pathway in Reactome. The value of the Spearman correlation is used as an enrichment

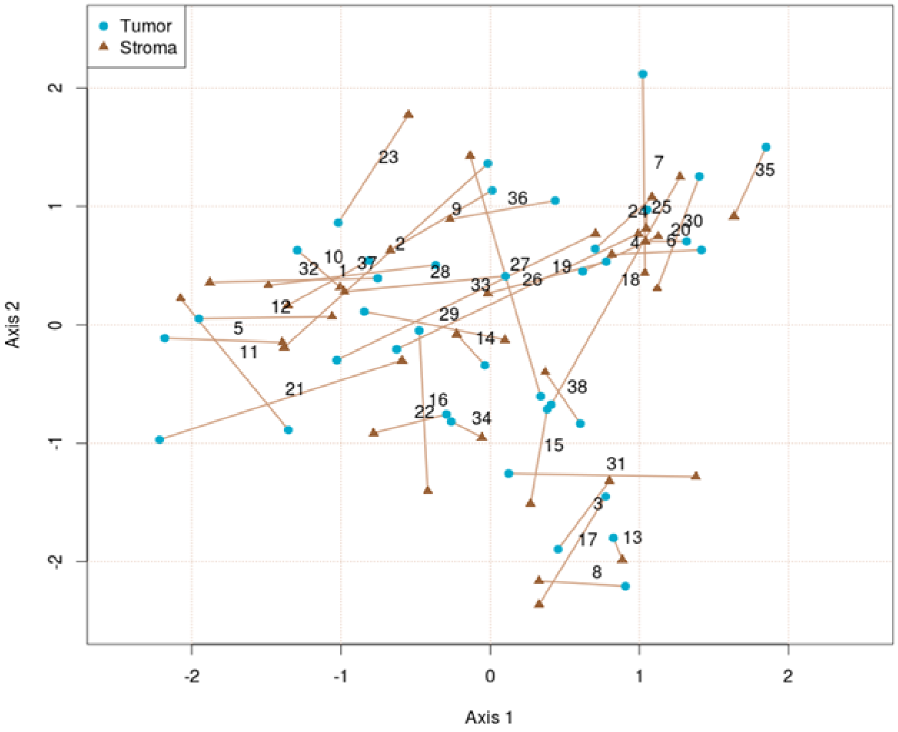


Figure 3.17: MCIA of microdissected and matched tumor and stroma samples

score. Additionally, the enrichment scores are calculated separately for the negative side and for the positive side of the MCIA axis resulting in down and up regulated pathways.

The enriched pathways are displayed in Figure 3.18. In order to obtain an easy to interpret result, only pathways that include genes with MCIA scores higher than the 75% quantile are shown, i.e, projected as far as possible from the origin on the first MCIA axis. The computed network contains genes as well as the pathways they belong to. Genes/pathways that are active in tumor are displayed as blue ellipses while genes/pathways that are active in stroma are displayed as red ellipses. Gray ellipses represent genes/pathways that are active in tumor and stroma. Genes belonging to the same biological pathways are linked. Each enriched pathway is represented by an ellipse which is linked to the genes belonging to it by dashed lines. Figure 3.18 shows a double bipartite graph: one can distinguish between tumor and stroma but also between up and down regulated pathways/genes. In addition to the active genes and enriched pathways, actionable gene targets [222] are mapped on the resulted network and are displayed as triangles. These target genes were downloaded from the homepage of the Broad Institute, Boston, MA, USA.

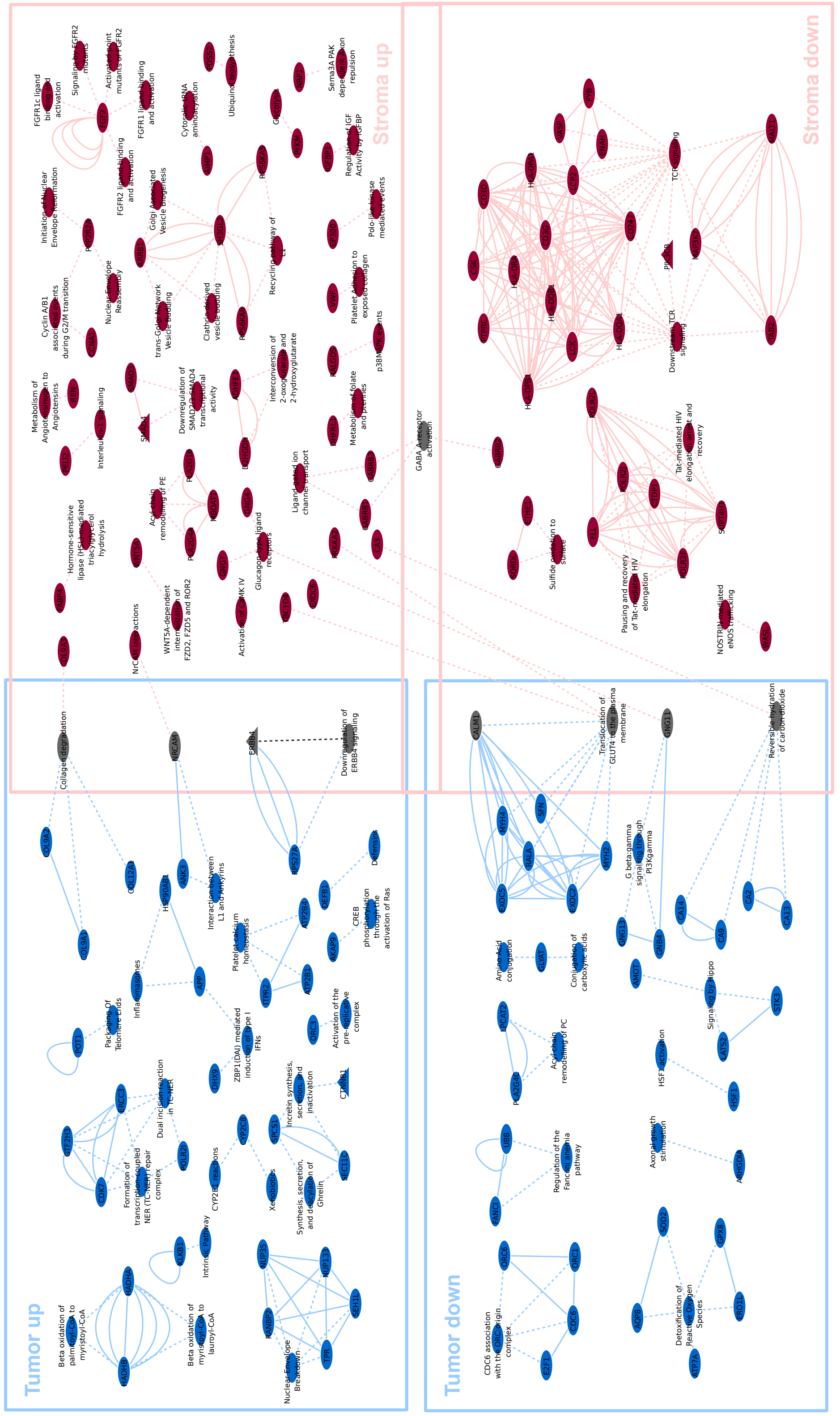

Stroma up
Stroma down
Tumor up
Tumor down
FGFR1c ligand binding and activation
Signaling by FGFR2 mutants
Activated point mutants of FGFR2
FGFR1 ligand binding and activation
FGFR2 ligand binding and activation
Initiation of Nuclear Envelope Reformation
Cyclin A/B1 associated events during G2/M transition
Nuclear Envelope Reassembly
Golgi Associated Vesicle Biogenesis
trans-Golgi Network Vesicle Budding
Clathrin derived vesicle budding
Recycling pathway of L1
Cytosolic tRNA aminoacylation
Ubiquinol biosynthesis
Sema3A PAK dependent Axon repulsion
Regulation of IGF Activity by IGFBP
Polo-like kinase mediated events
Platelet Adhesion to exposed collagen
p38MAPK events
Metabolism of Angiotensinogen to Angiotensins
Interleukin-1 signaling
Downregulation of SMAD2/3:SMAD4 transcriptional activity
Interconversion of 2-oxoglutarate and 2-hydroxyglutarate
Metabolism of folate and pterines
Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis
Acyl chain remodeling of PE
Ligand-gated ion channel transport
GABA A receptor activation
Glucagon-type ligand receptors
WNT5A-dependent internalization of FZD2, FZD5 and ROR2
Activation of CaMK IV
NrCAM interactions
TCR signaling
Downstream TCR signaling
Tat-mediated HIV elongation arrest and recovery
Sulfide oxidation to sulfate
Pausing and recovery of Tat-mediated HIV elongation
NOSTRIN mediated eNOS trafficking
Collagen degradation
NRCAM
ERBB4
Downregulation of ERBB4 signaling
CALM1
Translocation of GLUT4 to the plasma membrane
GNG11
Reversible hydration of carbon dioxide
Interaction between L1 and Ankyrins
Platelet calcium homeostasis
CREB phosphorylation through the activation of Ras
Inflammasomes
Packaging Of Telomere Ends
ZBP1(DAI) mediated induction of type I IFNs
Activation of the pre-replicative complex
Dual incision reaction in TC-NER
Formation of transcription-coupled NER (TC-NER) repair complex
Incretin synthesis, secretion, and inactivation
CYP2E1 reactions
Xenobiotics
Synthesis, secretion, and deacylation of Ghrelin
Intrinsic Pathway
Beta oxidation of palmitoyl-CoA to myristoyl-CoA
Beta oxidation of myristoyl-CoA to lauroyl-CoA
Nuclear Envelope Breakdown
G beta:gamma signalling through PI3Kgamma
Amino Acid conjugation
Conjugation of carboxylic acids
Signaling by Hippo
Acyl chain remodelling of PC
HSF1 activation
Axonal growth stimulation
Regulation of the Fanconi anemia pathway
CDC6 association with the ORC:origin complex
Detoxification of Reactive Oxygen Species

Figure 3.18: Enriched pathways and corresponding genes resulting from IPEA computed from the first MCIA axis of the tumor and stroma data sets. Notably, there are no edges linking tumor up to tumor down nor stroma down to tumor up.

Table 3.8: The co-structure between the four data sets is measured with the modified RV coefficient

	Pichia Transcriptome	Pichia Proteome	CHO Transcriptome	CHO Proteome
Pichia Transcriptome	1.000	0.755	0.939	0.826
Pichia Proteome	0.755	1.000	0.607	0.811
CHO Transcriptome	0.939	0.607	1.000	0.806
CHO Proteome	0.826	0.811	0.806	1.000

3.6 Cross-Species Comparison Between *Pichia pastoris* and Chinese Hamster Ovary Cells

In order to better characterize the two production systems *P. pastoris* and CHO, MCIA was applied to the four data sets: *P. pastoris* transcriptomics, *P. pastoris* proteomics, CHO transcriptomics and CHO proteomics. The data was measured in four conditions: 3D6 L:WT, 3D6 H:WT, HSA L:WT, HSA H:WT, resulting in 5254 *P. pastoris* genes, 1955 *P. pastoris* proteins, 20650 CHO genes and 609 CHO proteins. Additionally, four comparisons of the initial conditions were computed: 3D6 H:3D6 L, HSA H:HSA L, 3D6 L:HSA L, 3D6 H:HSA H.

In addition to the overall analysis, secretion and ribosome specific analyses are of particular interest as these feature subsets play key roles in the recombinant protein production (see Appendix D).

3.6.1 Co-Structure of the Measured Data Sets

The pairwise co-structure of the four data sets has been computed with the traditional RV coefficient. Additionally, a permutation test was performed (n = 1000 repetitions) in order to compute the significance of the computed RV values. The values for the traditional RV coefficient were very high (≥ 0.84) while the computed p-values were equal to 1, i.e. the RV coefficient was statistically not significant. The traditional RV coefficient suffers from dimension bias. The smaller the data sets, the higher the RV coefficient. To overcome this problem, the modified version of the RV coefficient (see Equation 2.12) was used.

Table 3.8 summarizes the pairwise co-structure between the four data sets measured with the modified RV coefficient. In order to asses the significance of the displayed values a permutation test was performed (n = 1000 repetitions). This test resulted in highly significant p-values for all pairwise RV coefficients. Compared to the traditional RV coefficients the modified RV coefficient is lower but statistically significant.

One can notice that the largest co-structure is shared by the two transcriptome data sets followed by CHO proteome and *P. pastoris* transcriptome which is similar to the co-structure between the two proteomes. The lowest co-structure was measured between

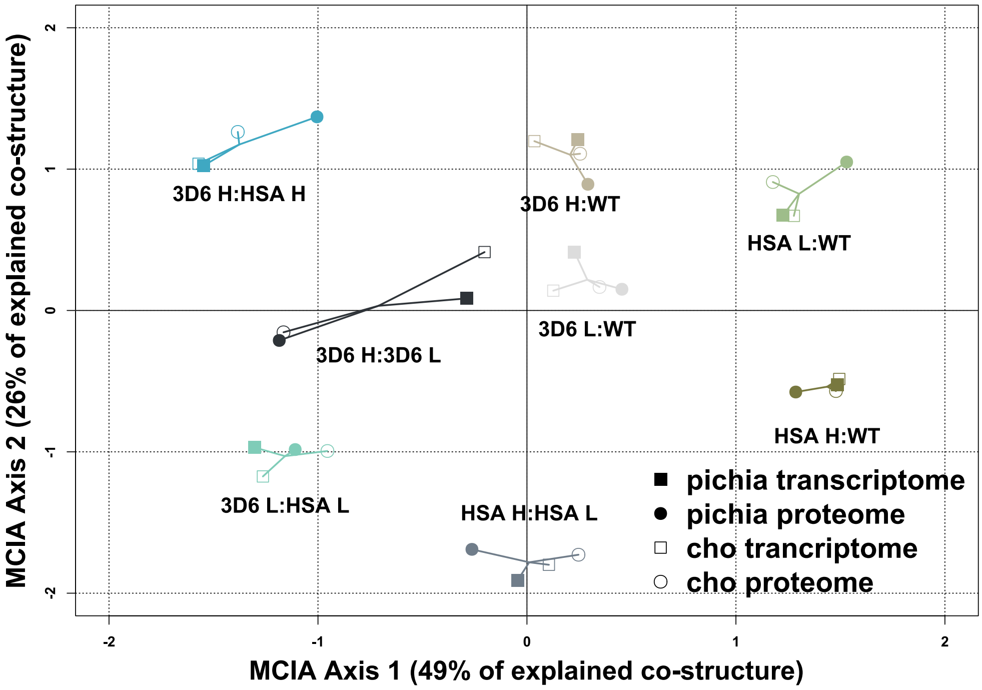


Figure 3.19: MCIA result of the cross-species comparison between *P. pastoris* and CHO cells.

P. pastoris proteome and CHO transcriptome followed by *P. pastoris* transcriptome and *P. pastoris* proteome as well as CHO transcriptome and CHO proteome.

3.6.2 Vizualization of the Eight Conditions

One advantage of MCIA is the ability to visualize all measured conditions and all data sets in one plot. The resulted graphic is shown in Figure 3.19. The MCIA axes 1 and 2 cover together 75 % of the captured co-structure (49 % and 26 %).

Each condition is displayed as four segments sharing one end point. The shared point was computed as the coordinate-wise mean value of the other four points. Each of the four points represents one data set: filled square represents the *P. pastoris* transcriptome, filled circle represents the *P. pastoris* proteome, empty square represents the CHO transcriptome and the empty circle represents the CHO proteome.

In Figure 3.19 it can be noticed that the first MCIA axis separates the comparisons to wild type (WT) from the other four conditions, excepting the CHO data from HSA H:HSA L which is on the same side with the WT comparisons. Additionally, the second and the first axes separate HSA H:WT from the other comparisons to WT. Based on the length of the segments, it can be observed that 3D6 H:3D6 L feature the lowest agreement between transcriptomes and proteomes.

Since we are interested in the characterization of the four engineered strains, the subsequent analyses will be limited to the following comparisons: 3D6 L:WT, 3D6 H:WT, HSA L:WT, HSA H:WT. As the other four conditions are computed as linear combinations of the four engineered strains, taking them into account does not add any new information. Furthermore, the interpretation of the results is easier as one can compare one strain to another. The comparison of 3D6 H:HSA H to 3D6 H:3D6 L is much more difficult to interpret. Nevertheless, secretion and ribosome relevant analyses were performed also on all eight conditions. The data is presented in the Appendix D showing an overall agreement between the MCIA results of all eight conditions and MCIA results of the four engineered strains. This emphasizes the legitimacy of conducting all further analyses on 3D6 L:WT, 3D6 H:WT, HSA L:WT and HSA H:WT.

3.6.3 Detailed Analyses of 3D6 L:WT, 3D6 H:WT, HSA L:WT and HSA H:WT

In order to characterize the four engineered strains: 3D6 L:WT, 3D6 H:WT, HSA L:WT and HSA H:WT, MCIA was applied and the results are shown in Figure 3.20A. In addition to the overall analysis, secretion and ribosome specific analyses are of particular interest as these sets of features play key roles in the recombinant protein production.

The complete data sets containing 5354 *P. pastoris* genes, 1961 *P. pastoris* proteins, 20650 CHO genes and 735 CHO proteins were subset to:

- secretion relevant data sets (see Figure 3.20B) with 405 *P. pastoris* genes, 236 *P. pastoris* proteins, 534 CHO genes and 45 CHO proteins;
- non ribosome data sets (see Figure 3.20C) containing 4860 *P. pastoris* genes, 1673 *P. pastoris* proteins, 20353 CHO genes and 675 CHO proteins and
- ribosome relevant data sets (see Figure 3.20D) which included 495 *P. pastoris* genes, 288 *P. pastoris* proteins, 297 CHO genes and 60 CHO proteins.

In general the four strains are separated by the first two MCIA axes which cover 51%-56%, respectively 25%-30% of the captured variance. In all analyses, the strains are projected in the same quadrants: HSA H:WT in the first quadrant, 3D6 H:WT in the second quadrant, 3D6 L:WT in the third quadrant and HSH L:WT in the forth quadrant (Figure 3.20).

All four MCIA analyses were further investigated. First the modified RV coefficient was computed for all pairwise comparisons and is displayed in Tables 3.9, 3.10, 3.11 and 3.12. Additionally, the p-values of the RV coefficients were computed by a permutation test with $n = 1000$ repetitions. All p-values were significant with only one exception (p-value of RV coefficient between *P. pastoris* proteome and CHO transcriptome from MCIA of non ribosome relevant data sets) which featured a p-value of 0.105.

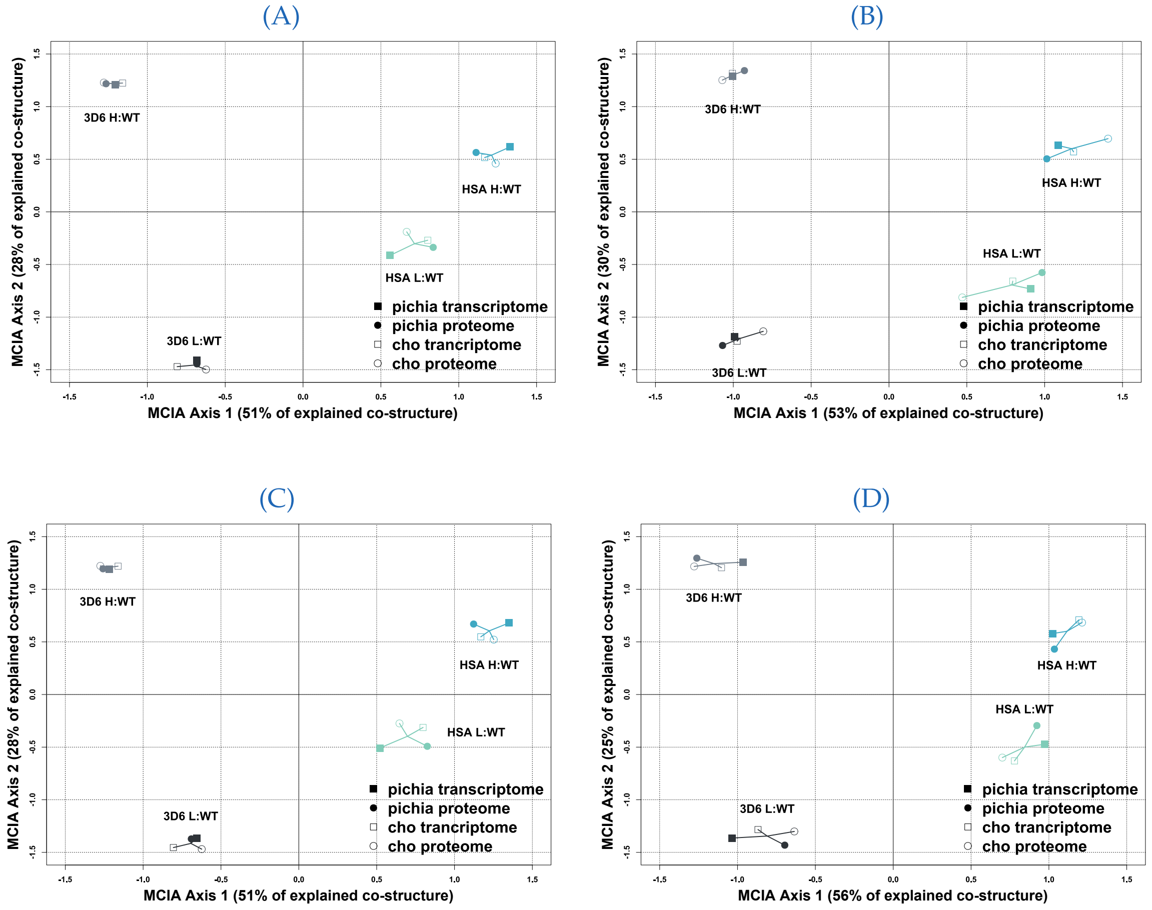


Figure 3.20: MCIA of the four engineered strains computed from (A) All Features, (B) Secretion Relevant Features, (C) All Features Except Ribosome Relevant Features and (D) Ribosome Relevant Features.

Table 3.9: RV coefficient of all features[1]

	Pichia T'ome	Pichia P'ome	CHO T'ome	CHO P'ome
Pichia T'ome	1.000	0.612	0.935	0.802
Pichia P'ome	0.612	1.000	0.588	0.896
CHO T'ome	0.935	0.588	1.000	0.795
CHO P'ome	0.802	0.896	0.795	1.000

Table 3.10: RV coefficient of secretion relevant Features[1]

	Pichia T'ome	Pichia P'ome	CHO T'ome	CHO P'ome
Pichia T'ome	1.000	0.601	0.991	0.755
Pichia P'ome	0.601	1.000	0.659	0.786
CHO T'ome	0.991	0.659	1.000	0.826
CHO P'ome	0.755	0.786	0.826	1.000

Table 3.11: RV coefficient of all features except ribosome specific features[1]

	Pichia T'ome	Pichia P'ome	CHO T'ome	CHO P'ome
Pichia T'ome	1.000	0.597	0.922	0.819
Pichia P'ome	0.597	1.000	0.577	0.886
CHO T'ome	0.922	0.577	1.000	0.809
CHO P'ome	0.819	0.886	0.809	1.000

Table 3.12: RV coefficient of ribosome specific features[1]

	Pichia T'ome	Pichia P'ome	CHO T'ome	CHO P'ome
Pichia T'ome	1.000	0.523	0.965	0.724
Pichia P'ome	0.523	1.000	0.572	0.826
CHO T'ome	0.965	0.572	1.000	0.795
CHO P'ome	0.724	0.826	0.795	1.000

Next, the distance to origin of all four strains was computed. More precise, a data set specific as well as a mean distance (of data set specific distances) were calculated. The results are summarized in Tables 3.13, 3.14, 3.15 and 3.16.

Additionally, the distances between all possible combinations of two out of the four strains of interest were computed. The calculations are based on the first three MCIA axes. These distances are displayed in Tables 3.17, 3.18, 3.19 and 3.20. The distances were computed between corresponding data sets, e.g. distance between 3D6 L:WT and 3D6 H:WT based on the *P. pastoris* proteome, as well as a mean distance computed as the mean value between the data set specific distances between two strains.

In addition to the strains, one can also investigate corresponding MCIA plots of the measured features, e.g. *P. pastoris* genes, *P. pastoris* proteins, CHO genes and CHO proteins (Figure 3.21 and 3.22). Tables which contain data set specific associated features were derived from the MCIA applied to all features (see Figure 3.21A) and can be seen in Tables 3.21 - 3.24.

Furthermore, a set of 67 GO (Slim) Terms (Table D.4) relevant for the comparison of *P. pastoris* and CHO expression systems were chosen and projected as additional information into the resulted MCIA space (Figure 3.24 and 3.24). Tables which contain data set specific associated GO Terms were derived from the MCIA applied to all features (see Figure 3.23A). These associations can be seen in Tables 3.25 - 3.28.

[1] Due to space reasons the types of data sets were abbreviated as T'ome for transcriptome and P'ome for proteome.

Table 3.13: Distances to origin computed from all features[2]

	Mean Dist	Pichia T'ome	Pichia P'ome	CHO T'ome	CHO P'ome
3D6 L:WT	2.202	2.143	2.167	2.358	2.140
3D6 H:WT	2.822	2.776	2.888	2.704	2.919
HSA L:WT	2.006	1.810	2.186	2.120	1.909
HSA H:WT	2.689	2.925	2.513	2.601	2.718

Table 3.14: Distances to origin computed from secretion relevant features[2]

	Mean Dist	Pichia T'ome	Pichia P'ome	CHO T'ome	CHO P'ome
3D6 L:WT	2.521	2.558	2.744	2.559	2.222
3D6 H:WT	2.624	2.618	2.533	2.640	2.706
HSA L:WT	2.097	2.294	2.361	2.064	1.670
HSA H:WT	2.665	2.507	2.330	2.678	3.146

Table 3.15: Distances to origin computed from all features except ribosome relevant features[2]

	Mean Dist	Pichia T'ome	Pichia P'ome	CHO T'ome	CHO P'ome
3D6 L:WT	2.179	2.086	2.139	2.358	2.134
3D6 H:WT	2.816	2.788	2.864	2.705	2.908
HSA L:WT	1.995	1.782	2.198	2.115	1.886
HSA H:WT	2.714	2.966	2.548	2.597	2.744

Table 3.16: Distances to origin computed from ribosome relevant features[2]

	Mean Dist	Pichia T'ome	Pichia P'ome	CHO T'ome	CHO P'ome
3D6 L:WT	2.307	2.716	2.164	2.366	1.980
3D6 H:WT	2.860	2.499	3.098	2.744	3.101
HSA L:WT	2.212	2.447	2.317	2.116	1.967
HSA H:WT	2.692	2.499	2.490	2.871	2.906

[2] Due to space reasons the types of data sets were abbreviated as T'ome for transcriptome and P'ome for proteome.

Table 3.17: Distances between two strains computed from all features [3]

	Mean	Pichia	Pichia	CHO	CHO
	Dist	T'ome	P'ome	T'ome	P'ome
3D6 L:WT 3D6 H:WT	3.210	3.141	3.229	3.123	3.349
HSA L:WT HSA H:WT	2.690	2.990	2.556	2.555	2.660
3D6 L:WT HSA L:WT	3.638	3.305	3.780	3.965	3.503
3D6 H:WT HSA H:WT	5.091	5.264	4.965	4.875	5.258

Table 3.18: Distances between two strains computed from secretion relevant features [3]

	Mean	Pichia	Pichia	CHO	CHO
	Dist	T'ome	P'ome	T'ome	P'ome
3D6 L:WT 3D6 H:WT	3.116	3.066	3.236	3.147	3.017
HSA L:WT HSA H:WT	2.520	2.386	2.142	2.394	3.156
3D6 L:WT HSA L:WT	3.997	4.271	4.606	4.037	3.075
3D6 H:WT HSA H:WT	4.756	4.572	4.324	4.793	5.334

Table 3.19: Distances between two strains computed from all features except ribosome relevant [3]

	Mean	Pichia	Pichia	CHO	CHO
	Dist	T'ome	P'ome	T'ome	P'ome
3D6 L:WT 3D6 H:WT	3.181	3.128	3.143	3.124	3.330
HSA L:WT HSA H:WT	2.736	3.081	2.640	2.536	2.690
3D6 L:WT HSA L:WT	3.591	3.193	3.740	3.969	3.464
3D6 H:WT HSA H:WT	5.105	5.319	4.956	4.879	5.268

Table 3.20: Distances between two strains computed from ribosome relevant features[3]

	Mean	Pichia	Pichia	CHO	CHO
	Dist	T'ome	P'ome	T'ome	P'ome
3D6 L:WT 3D6 H:WT	2.884	2.750	3.110	2.669	3.006
HSA L:WT HSA H:WT	2.332	2.187	2.061	2.511	2.569
3D6 L:WT HSA L:WT	4.070	4.796	4.065	4.013	3.407
3D6 H:WT HSA H:WT	5.204	4.608	5.294	5.244	5.669

[3] Due to space reasons the types of data sets were abbreviated as T'ome for transcriptome and P'ome for proteome.

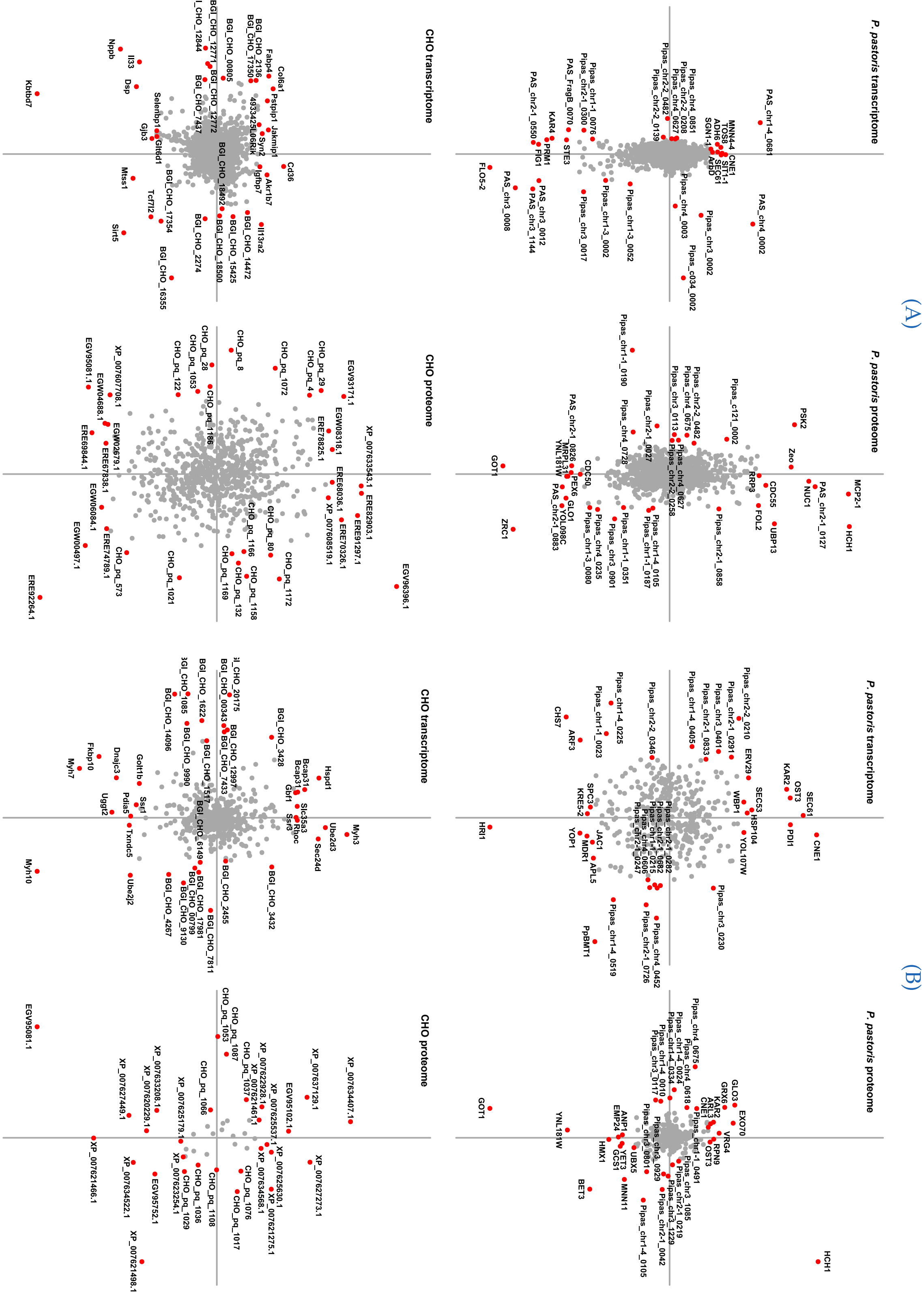


Figure 3.21: MCIA associated features. Feature associations based on MCIA of (A) all features and (B) secretion relevant features.

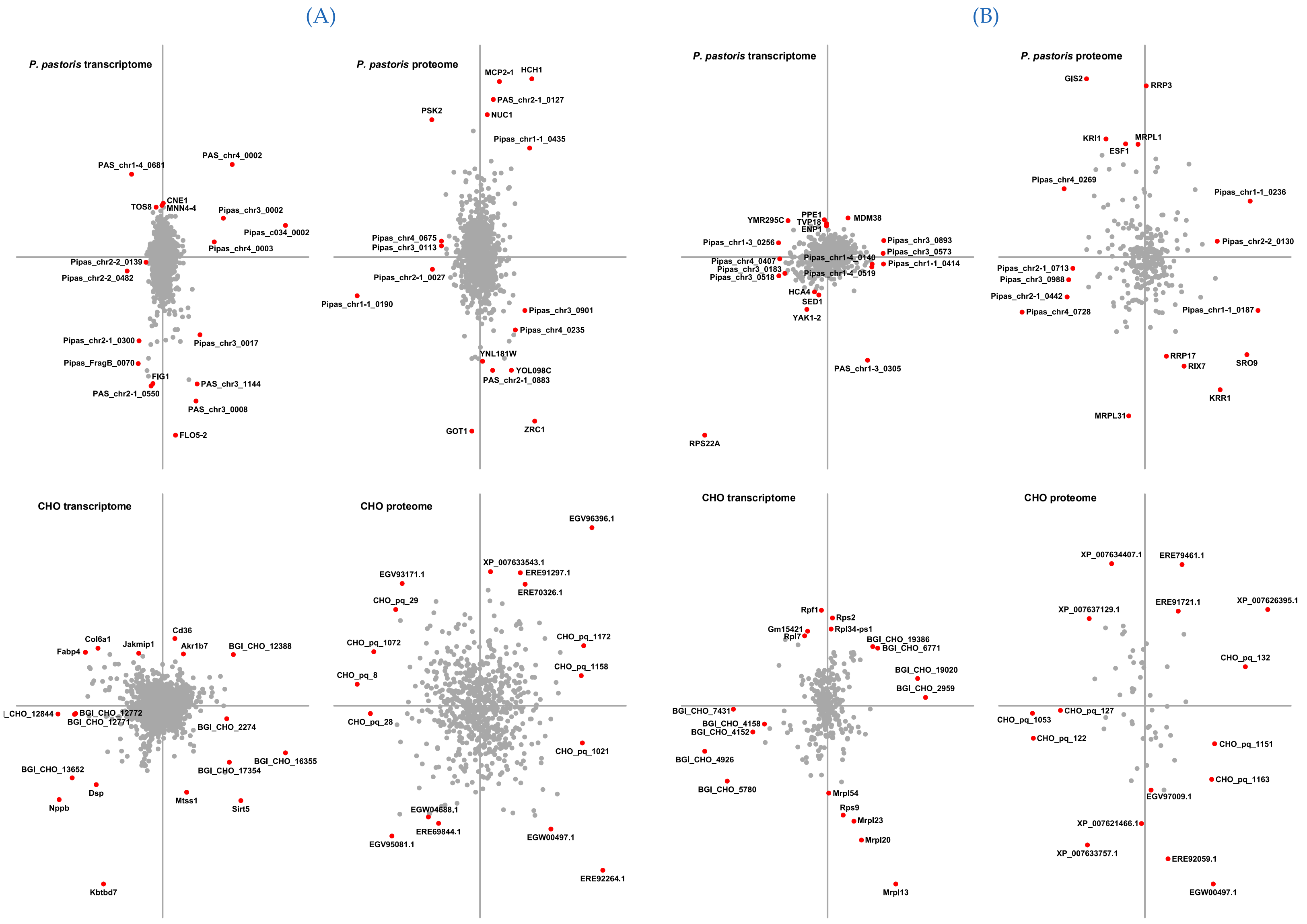


Figure 3.22: MCIA associated features. Feature associations based on MCIA of (A) all features without the ribosome relevant and (B) ribosome relevant features.

Table 3.21: *P. pastoris* genes associated to the four production strains derived from the MCIA of all features

ID	Symbol	Description
HSA H:WT		
Pipas_c034_0002	FLO100	contains GLEYA adhesin domain
Pipas_chr4_0002	PAS_chr4_0002	Hypothetical protein_not annotated in NCBI
Pipas_chr3_0002	FLO101	Lectin-like protein
Pipas_chr4_0003	PAS_chr4_0003	hypothetical protein
Pipas_chr2-2_0006	PAS_chr2-2_0006	hypothetical protein
3D6 H:WT		
Pipas_chr1-4_0681	PAS_chr1-4_0681	Luciferase-like monooxygenase
Pipas_chr4_0851	PAS_chr4_0851	hypothetical protein
Pipas_chr4_0627	HSP12	Plasma membrane protein involved in maintaining membrane organization in stress conditions
Pipas_chr2-2_0208	PAS_chr2-2_0208	hypothetical protein
Pipas_chr4_0576	ADH6	NADPH-dependent medium chain alcohol dehydrogenase
3D6 L:WT		
Pipas_chr2-2_0482	FLO11	GPI-anchored cell surface glycoprotein (flocculin)
Pipas_FragB_0070	PAS_FragB_0070	in frame stop codons
Pipas_chr2-1_0300	PAS_chr2-1_0300	Hypothetical protein not annotated in NCBI
Pipas_chr2-1_0550	PAS_chr2-1_0550	similarity to cell wall endo-beta-1,3-glucanase
Pipas_chr1-1_0332	PRM1	Pheromone-regulated multispanning membrane protein involved in membrane fusion during mating
HSA L:WT		
Pipas_chr3_0008	PAS_chr3_0008	Ion channel regulatory protein UNC-93; MFS general substrate transporter
Pipas_chr3_1144	PAS_chr3_1144	(not correctly annotated sequence)
Pipas_chr3_0017	PAS_chr3_0017	similarity to bacterial 3-hydroxyisobutyrate dehydrogenase
Pipas_chr3_1145	FLO5-2	Lectin-like cell wall protein (flocculin) involved in flocculation
Pipas_chr3_0012	PAS_chr3_0012	Fungal Zn(2)-Cys(6) binuclear cluster domain

Table 3.22: *P. pastoris* proteins associated to the four production strains derived from the MCIA of all features

ID	Symbol	Description
HSA H:WT		
Pipas_chr3_0170	HCH1	Heat shock protein regulator; binds to Hsp90p and may stimulate ATPase activity
Pipas_chr2-1_0581	MCP2-1	Putative protein of unknown function
Pipas_chr1-1_0435	UBP13	Putative ubiquitin-specific protease that cleaves Ub-protein fusions
Pipas_chr2-1_0127	PAS_chr2-1_0127	TB2/DP1, HVA22 family
Pipas_chr4_0864	NUC1	Major mitochondrial nuclease, has RNAse and DNA endo- and exonucleolytic activities
3D6 H:WT		
Pipas_chr4_0076	PSK2	PAS-domain containing serine/threonine protein kinase
PDKT_016	Zeo	Antibiotic resistance Zeo
Pipas_c121_0002	ALG2	Mannosyltransferase that catalyzes two consecutive steps in the N-linked glycosylation pathway
Pipas_chr2-1_0780	ADA2	Transcription coactivator, component of the ADA and SAGA transcriptional adaptor
Pipas_chr3_0113	KEI1	Component of inositol phosphorylceramide (IPC) synthase
3D6 L:WT		
Pipas_chr1-1_0190	PAS_chr1-1_0190	Non-catalytic subunit of N-terminal acetyltransferase
Pipas_chr1-1_0271	GOT1	Homodimeric protein that is packaged into COPII vesicles and cycles between the ER and Golgi
Pipas_chr2-1_0027	YPL191C	Putative protein of unknown function; diploid deletion strain exhibits high budding index
Pipas_chr4_0728	MAK5	Essential nucleolar protein, putative DEAD-box RNA helicase
Pipas_chr1-4_0133	PEX6	AAA-peroxin that heterodimerizes with AAA-peroxin Pex1p
HSA L:WT		
Pipas_chr4_0896	ZRC1	Vacuolar membrane zinc transporter; transports zinc from cytosol to vacuole for storage
Pipas_chr4_0546	YOL098C	Putative metalloprotease
Pipas_chr3_0901	PAS_chr3_0901	Arrestin (or S-antigen), N-terminal domain
Pipas_chr4_0842	GLO1	Monomeric glyoxalase I, catalyzes the detoxification of methylglyoxal
Pipas_chr1-3_0080	BET3	Hydrophilic protein that acts in conjunction with SNARE proteins

Table 3.23: CHO genes associated to the four production strains derived from the MCIA of all features

ID	Symbol	Description
HSA H:WT		
BGI_CHO_12388	Il13ra2	interleukin-13 receptor subunit alpha-2-like (LOC100754617), mRNA
BGI_CHO_15425	Fgf18	fibroblast growth factor 18-like (LOC100761504), mRNA
BGI_CHO_14472	Ces1f	liver carboxylesterase 4-like (LOC100769145), mRNA
BGI_CHO_18492	Uts2	urotensin-2-like (LOC100768033), mRNA
BGI_CHO_18500		UDP-glucuronosyltransferase 1-8-like (LOC100750842), partial mRNA
3D6 H:WT		
BGI_CHO_13520	Fabp4	fatty acid-binding protein, adipocyte-like (LOC100760812), mRNA
BGI_CHO_2136	LOC100773771	hypothetical protein LOC100773771 (LOC100773771), mRNA
BGI_CHO_17350	Sprr1a	cornifin-A-like (LOC100760951), mRNA
BGI_CHO_00805	Ccl2	c-C motif chemokine 2-like (LOC100763833), mRNA
BGI_CHO_5046	Col6a1	collagen, type VI, alpha 1 (Col6a1), mRNA
3D6 L:WT		
BGI_CHO_18336	Kbtbd7	kelch repeat and BTB domain-containing protein 7-like (LOC100764033), partial mRNA
BGI_CHO_17111	Nppb	natriuretic peptides B-like (LOC100773766), mRNA
BGI_CHO_12844	Ldhc	L-lactate dehydrogenase C chain-like (LOC100751672), mRNA
BGI_CHO_13652	Il33	interleukin-33-like (LOC100761971), mRNA
BGI_CHO_12771	Fn1	fibronectin 1, transcript variant 1 (Fn1), mRNA
HSA L:WT		
BGI_CHO_16355	Ugt1a1	UDP glucuronosyltransferase 1 family, polypeptide A1 (Ugt1a1), mRNA
BGI_CHO_2716	Sirt5	NAD-dependent deacetylase sirtuin-5-like (LOC100767069), mRNA
BGI_CHO_15306	Tcf7l2	transcription factor 7-like 2-like, transcript variant 2 (LOC100758400), mRNA
BGI_CHO_17354	Olfr1350	olfactory receptor 1038-like (LOC100762985), mRNA
BGI_CHO_2274	Lmo2	rhombotin-2-like (LOC100773201), mRNA

Table 3.24: CHO proteins associated to the four production strains derived from the MCIA of all features

ID	Symbol	Description
HSA H:WT		
CHO_pq_1206	EGV96396.1, XP_007624367.1	Oxysterol-binding protein-related protein 11
CHO_pq_1172	XP_007626395.1	2-hydroxyacyl-CoA lyase 1 isoform X1
CHO_pq_1158	ERE49129.1, XP_007607559.1	protein ETHE1, partial
CHO_pq_821	ERE91297.1	platelet glycoprotein 4
CHO_pq_691	ERE82903.1, XP_007634912.1	protein ERGIC-53-like protein
3D6 H:WT		
CHO_pq_8	EGW01221.1, XP_007649516.1	Fermitin family-like 2
CHO_pq_74	EGV93171.1	UPF0480 protein C15orf24-like
CHO_pq_1072	XP_007629953.1	transcriptional activator protein Pur-alpha
CHO_pq_29	ERE87936.1, XP_007636792.1	peroxisomal carnitine O-octanoyltransferase isoform 2
CHO_pq_4	EGV95701.1	Lanosterol 14-alpha demethylase
3D6 L:WT		
CHO_pq_102	EGV95081.1, ERE80723.1	Splicing factor 45
CHO_pq_28	ERE87480.1	ADP-ribosylation factor 4-like protein
CHO_pq_929	XP_007607708.1	suppressor of G2 allele of SKP1 homolog isoform X3
CHO_pq_517	ERE69844.1, XP_007616531.1	zinc finger CCCH domain-containing protein 6
CHO_pq_262	EGW04688.1	Phosphoribosyl pyrophosphate synthetase-associated protein 1
HSA L:WT		
CHO_pq_833	ERE92264.1, XP_007612600.1	metaxin-1-like protein
CHO_pq_1219	EGW00497.1, ERE74919.1, ERE74920.1	U2-associated protein SR140
CHO_pq_1021	XP_007621498.1	cytochrome c oxidase subunit 6C-2 isoform X2
CHO_pq_573	ERE74027.1, XP_007625813.1	NADH dehydrogenase [ubiquinone] iron-sulfur protein 2
CHO_pq_585	ERE74789.1, XP_007643559.1	DNA/RNA helicase, DEAD/DEAH box type containing protein

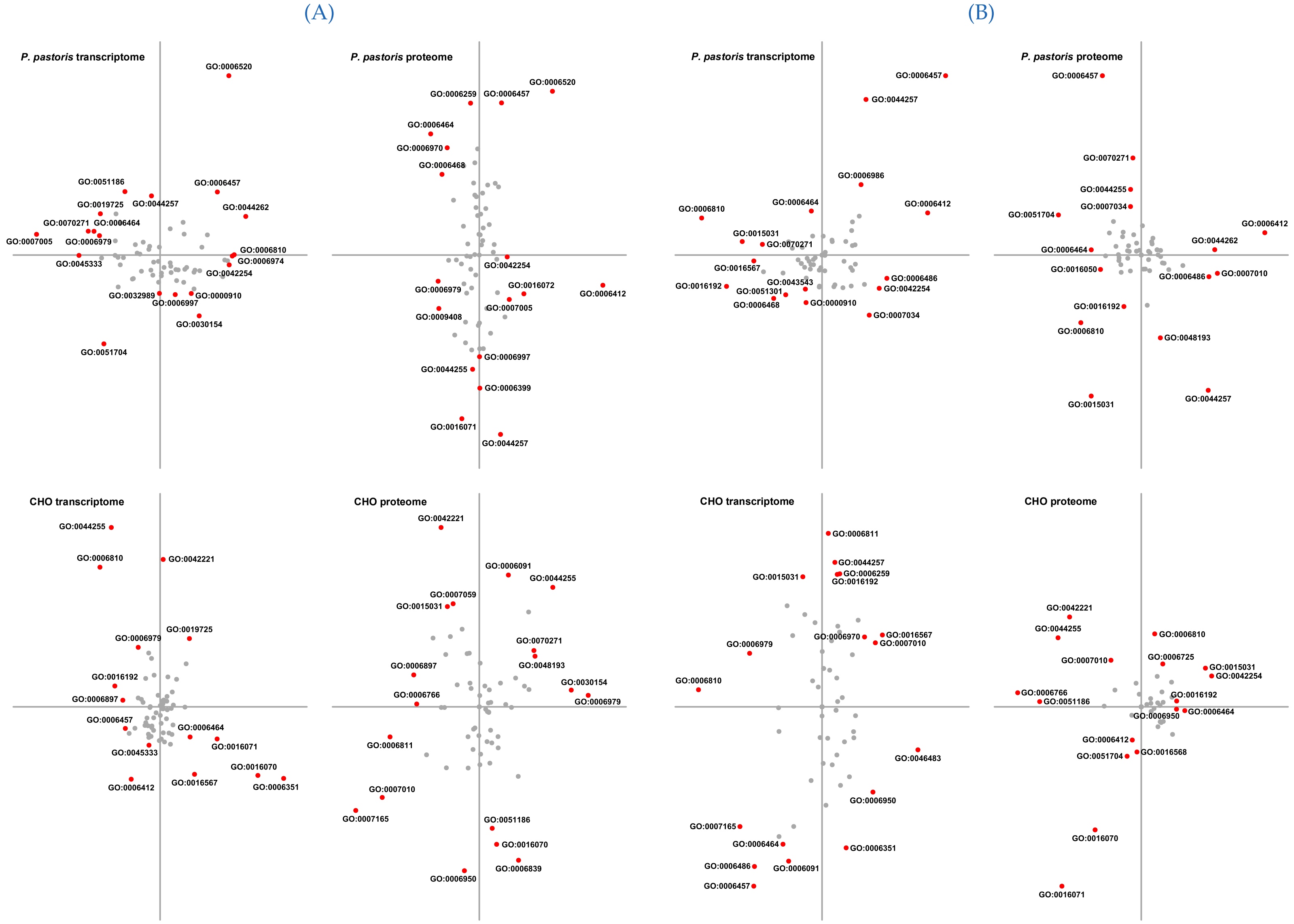


Figure 3.23: MCIA associated GO Terms. GO Terms associations based on MCIA of (A) all features and (B) secretion relevant features.

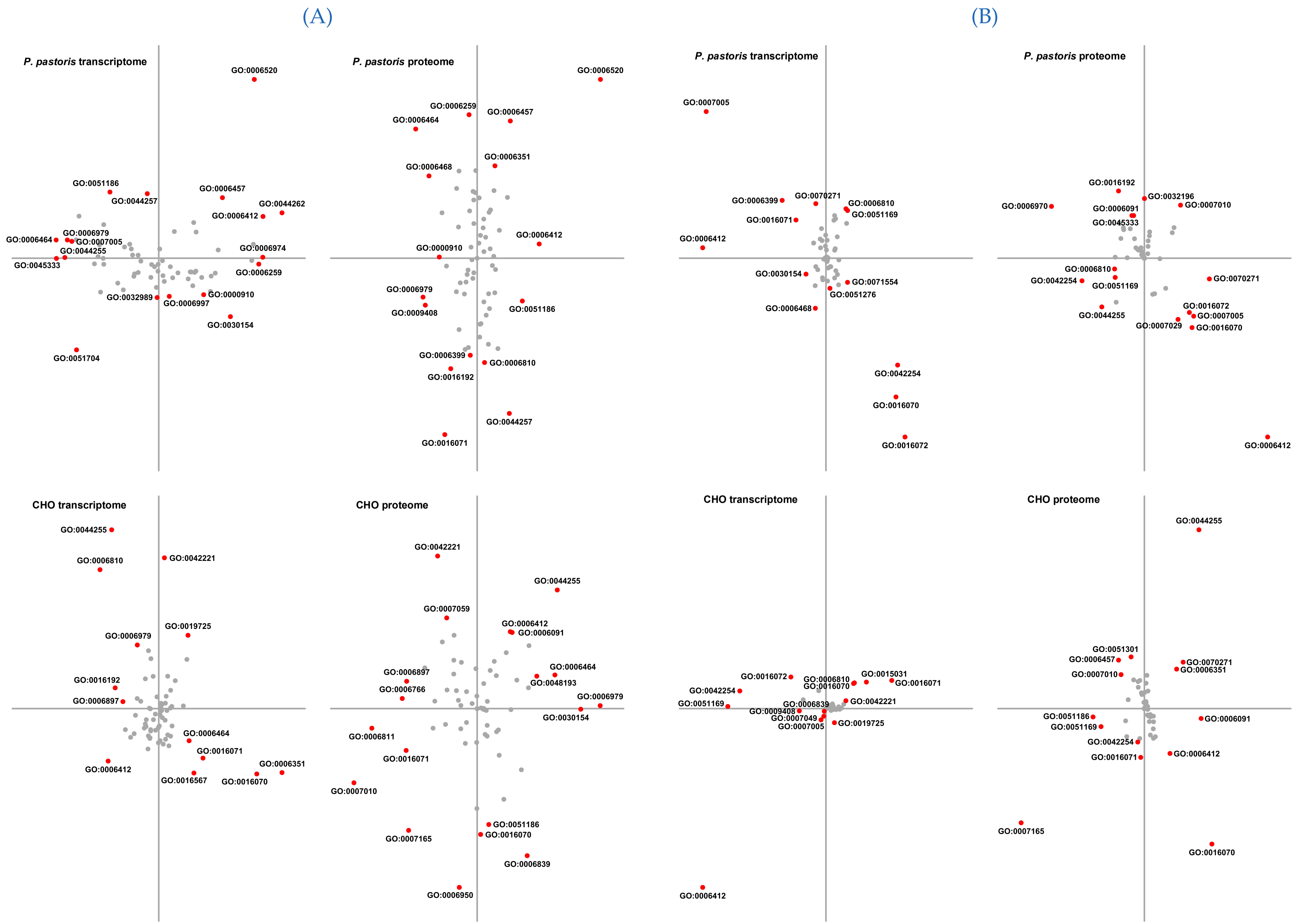


Figure 3.24: MCIA associated GO Terms. GO Terms associations based on MCIA of (A) all features without the ribosome relevant and (B) ribosome relevant features.

Table 3.25: Associated GO Terms derived from *P. pastoris* transcriptome and the MCIA of all features

GO ID	GO Name
HSA H:WT	
GO:0006520	cellular amino acid metabolic process
GO:0006457	protein folding
GO:0044262	cellular carbohydrate metabolic process
GO:0007031	peroxisome organization
GO:0006810	transport
3D6 H:WT	
GO:0051186	cofactor metabolic process
GO:0007005	mitochondrion organization
GO:0044257	cellular protein catabolic process
GO:0019725	cellular homeostasis
GO:0006412	translation
3D6 L:WT	
GO:0051704	multi organism process
GO:0032989	cellular component morphogenesis
GO:0007165	signal transduction
GO:0045333	cellular respiration
GO:0007049	cell cycle
HSA L:WT	
GO:0030154	cell differentiation
GO:0000910	cytokinesis
GO:0006997	nucleus organization
GO:0006811	ion transport
GO:0016072	rRNA metabolic process

Table 3.26: Associated GO Terms derived from *P. pastoris* proteome and the MCIA of all features

GO ID	GO Name
HSA H:WT	
GO:0006520	cellular amino acid metabolic process
GO:0006457	protein folding
GO:0006839	mitochondrial transport
GO:0006351	transcription
GO:0006486	protein glycosylation
3D6 H:WT	
GO:0006464	cellular protein modification process
GO:0006468	protein phosphorylation
GO:0006970	response to osmotic stress
GO:0006259	DNA metabolic process
GO:0007049	cell cycle
3D6 L:WT	
GO:0016071	mRNA metabolic process
GO:0009408	response to heat
GO:0006979	response to oxidative stress
GO:0016192	vesicle mediated transport
GO:0000910	cytokinesis
HSA L:WT	
GO:0006412	translation
GO:0044257	cellular protein catabolic process
GO:0016072	rRNA metabolic process
GO:0006399	tRNA metabolic process
GO:0007005	mitochondrion organization

Table 3.27: Associated GO Terms derived from CHO transcriptome and the MCIA of all features

GO ID	GO Name
HSA H:WT	
GO:0042221	response to chemical stimulus
GO:0019725	cellular homeostasis
GO:0009408	response to heat
GO:0006970	response to osmotic stress
GO:0032989	cellular component morphogenesis
3D6 H:WT	
GO:0044255	cellular lipid metabolic process
GO:0006810	transport
GO:0016192	vesicle mediated transport
GO:0006979	response to oxidative stress
GO:0006897	endocytosis
3D6 L:WT	
GO:0006412	translation
GO:0006457	protein folding
GO:0030154	cell differentiation
GO:0051186	cofactor metabolic process
GO:0045333	cellular respiration
HSA L:WT	
GO:0006351	transcription
GO:0016070	RNA metabolic process
GO:0016071	mRNA metabolic process
GO:0016567	protein ubiquitination
GO:0006464	cellular protein modification process

Table 3.28: Associated GO Terms derived from CHO proteome and the MCIA of all features

GO ID	GO Name
HSA H:WT	
GO:0006979	response to oxidative stress
GO:0044255	cellular lipid metabolic process
GO:0030154	cell differentiation
GO:0006091	generation of precursor metabolites and energy
GO:0006412	translation
3D6 H:WT	
GO:0042221	response to chemical stimulus
GO:0006897	endocytosis
GO:0006766	vitamin metabolic process
GO:0015031	protein transport
GO:0007059	chromosome segregation
3D6 L:WT	
GO:0007165	signal transduction
GO:0007010	cytoskeleton organization
GO:0006811	ion transport
GO:0006950	response to stress
GO:0016071	mRNA metabolic process
HSA L:WT	
GO:0006839	mitochondrial transport
GO:0016070	RNA metabolic process
GO:0051186	cofactor metabolic process
GO:0007005	mitochondrion organization
GO:0043543	protein acylation

4 Discussion

The results of this doctoral thesis are discussed in this chapter. The first and second section review the results of the *P. falciparum* and *A. gambiae* analyses. Subsequently, the results of the traditional GSEA applied to human miRNAs are discussed. Next, the development of the novel pathway enrichment approach which was applied to microdissected tumor and stroma gene expression profiles is reviewed. Finally, the results of the cross-species comparison between *P. pastoris* and CHO cells are discussed.

4.1 *Plasmodium falciparum* Life Cycle Stages Analysis

In this section results of the analysis of the life cycle stages of *P. falciparum*, as described in [146], are discussed. In this study, we have applied three integrative analysis methods to a data set containing mRNA and protein abundances from the six life cycle stages of *P. falciparum.* The use of integrative analysis methods allows considering all annotated and measured genes (3283) and proteins (2491) and is not limited to the 2230 pairs of genes and proteins as when it was first published in [103]. The integration of knowledge on different levels allows linking of the data sets based on samples and not on variables.

Three different integrative analysis methods were introduced, each with its own justification: CIA discovers biological processes on the basis of maximal covariance. GSVD decomposes the data sets into genelets and arraylets and conducts a modified GSEA analysis on them. IBC computes biclusters according to the distance between genes, proteins and GO terms.

Method-specific results as well as results common to all three analysis methods were shown. In the case of CIA, associations in protein space presented a high overlap with the common results. This was not the case for associations in gene space. In case of the sporozoite stage, GSVD associations are very similar to the common results. For the other stages, GSVD yielded different mappings compared to the common results. As a GO term is associated or not with a life cycle stage, only general but no method-specific results were computed for IBC.

For CIA, it is important to consider that GO term associations are done through projection, whereas GSVD maps GO terms to individual stages through restricted GSE analysis and IBC assigns GO terms to life cycle stages through the distance to the corresponding life cycle stage. Another important aspect is that with CIA it is not possible to associate one GO term to more than one life cycle stage, while this is possible with GSVD and IBC. Due to the heterogeneity of the computational methods, we proposed taking the intersection of the three obtained results.

In the three-fold validated network view of the biological processes (Figure 3.8), the separation of the intraerytrocytic cycle (merozite, ring, trophozoite and schizont) from sporozoites and gametocytes can be seen. While the stages of the intraerytrocytic cycle are tightly connected to one another, sporozoites share two biological processes and gametocytes share only one biological process with the rest. Gametocyte and sporozoite stages do not possess any common processes, reflecting the profound differences between these stages. Gametocytes are released into the blood stream, from where they travel to the liver, while sporozoites represent the sexual stage and lie dormant in cell cycle arrest until ingestion by a mosquito.

The data used here was initially gathered in order to investigate the role of post-transcriptional regulation in *P. falciparum* [103]. For this, only pairs of mRNA and corresponding proteins were considered, resulting in the exploitation of 89% of the proteins and 60% of the genes that were experimentally measured. By employing integrative analysis methods, we were able to take all measured data into account.

Le Roch *et al.* [103] mention that there is a

> *"bias in proteomic analysis of whole-cell lysates, in that such methods may fail to detect secreted or membrane proteins present in low abundance, such as GPI anchors."*

Due to the integrative approach, our analysis associates several GO terms related to GPI anchors proteins (GO:0006506 *GPI anchor biosynthetic process*, GO:0016255 *attachment of GPI anchor to protein*, GO:0006661 *phosphatidylinositol biosynthetic process*, GO:0046488 *phosphatidylinositol metabolic process*) with the merozoite stage, prevailing over this shortcoming. These associations are in agreement with [223], where distinct protein classes, with a focus on merozoite surface antigens, are discussed. The importance of GPI anchor proteins in the merozoite stage is well known and very important in immune evasion [224, 225].

Other biological processes mentioned in [103] such as *glycolysis* and *cell invasion* without any life cycle mapping were also found in the resulting network such as GO:0044409 *entry into host* and GO:0030260 *entry into host cell*, both associated with the merozoite stage. This network assigns GO:0006096 *glycolysis* to the stage trophozoite, in concordance to [226] where the transcriptome of *P. falciparum* was characterized.

Simmilar to the current findings, cell invasion was associated with merozoites in [224], where a proteomic view of the *P. falciparum* life cycle was presented. Other concordances with [224] include the assignment of GO:0006508 *proteolysis* to the merozoite stage. During trophozoite stage, digestion of haemoglobin takes place. The computed network maps GO:0006091 *generation of precursor metabolites and energy* to trophozoites, confirming the importance of energy production during this stage. As mentioned by Florens *et al.* [224], sporozoites are injected into the blood stream where they have to survive in a hostile environment. Based on the here combined results, sporozoites are associated with GO:0020013 *modulation by symbiont of host erythrocyte aggregation* and GO:0020035 *cytoadherence to microvasculature, mediated by symbiont protein*, which reflects the process of survival. Additionally, sporozoites are associated with metabolism and transcription, as was shown in Figure 5 of [224]. The current results reflect these findings by mapping GO:0006163 *purine nucleotide metabolic* *process*, GO:0009117 *nucleotide metabolic process*, GO:0006351 *transcription, DNA dependent* and GO:0006355 *regulation of transcription, DNA dependent* to the sporozoite stage.

During gametocyte stage, DNA processing and energy production is highly regulated, as mentioned in [224]. In agreement, the here discussed results assign GO:0006323 *DNA packaging*, GO:0006839 *mitochondrial transport* and GO:0006626 *protein targeting to mitochondrion* to the gametocytes.

The analysis of the *P. falciparum* proteome by LaCount *et al.* in [227] associated the intraerythrocytic cycle with chromatin modification, transcriptional regulation, mRNA stability/processing, ubiquitination, nucleic acid metabolism and invasion of host cells. Since the here performed analysis corresponds to individual life cycle stages, it is possible to associate biological processes to a certain stage of the intraerythrocytic cycle, providing a more detailed description of *P. falciparum.* According to our findings, chromatin modification takes place during schizont stage (GO:0006325 *chromatin organization*, GO:0051276 *chromosome organization*); merozoites are associated with GO:0006357 *regulation of transcription from RNA polymerase II promotor* and schizonts with GO:0042795 *transcription from RNA polymerase II promotor*; merozoites are associated with GO:0009116 *nucleoside metabolic process*; invasion of host cells can be observed during merozoite stage (GO:0044409 *entry into host* and GO:0030260 *entry into host cell*). Ubiquitination was only detected through its parent term GO:0044267 *cellular protein metabolic process*, which was associated with merozoites.

Fagan *et al.* in [105] conducted CIA on a slightly different data set taking *P. berghei* orthologues into account and showed that GO:0006412 *biosynthesis* is associated to the intraerythrocytic cycle. In the here computed network, several more specialized biosynthetic processes are associated with the merozoite stage: GO:0009059 *macromolecule biosynthetic process*, GO:0008654 *phospholipid biosynthetic process*, GO:0045017 *glycerolipid biosynthetic process*, GO:0006661 *phosphatidylinositol biosynthetic process*, GO:0006506 *GPI anchor biosynthetic process*, as well as the GO term GO:0006412 *biosynthetic process* itself.

The importance of immune evasion through antigenic variation was highlighted by Winzeler in [228]. The results discussed here show that this process is related to the schizont stage, as the current analysis associates GO:0020033 *antigenic variation*, GO:0006955 *immune response*, GO:0050776 *regulation of immune response*, GO:0002377 *immunoglobulin production*, GO:0006950 *response to stress* and GO:0009607 *response to biotic stimulus* with this stage.

The role of lipids during merozoite stage was already shown in 1988 by Mikkelsen *et al.* [229]. The here computed network associates merozoites with GO:0006644 *phospholipid metabolic process*, GO:0008654 *phospholipid biosynthetic process*, GO:0046486 *glycerolipid metabolic process* and GO:0006629 *lipid metabolic process*, reflecting this early finding.

Phosphorilation and dephosphoryliation processes play an important role in the internalization step of meroziotes [230], a fact that is also reflected our results, as merozoites are associated with GO:0016311 *dephosphorylation* and GO:0006470 *protein dephosphorylation*.

The role of the pentose phosphate pathway in *P. falciparum* was disscused in [231], without a clear life cycle stage assignment. The computed network view maps GO:0006098 *pentose-phosphate shunt* to merozoites.

As shown in [232], REDOX complexes play an important role during ring stage, which is in agreement with the association of GO:0045454 *cell redox homeostasis* and GO:0055114 *oxidation-reduction process* to the ring stage.

Roth showed in [233] that carbohydrate metabolism is a key metabolic process connecting the host cells with *P. falciparum.* The findings discussed here also assign GO:0005975 *carbohydrate metabolic process* to merozoite and ring stages.

Most of the network associations are in concordance with several publications dealing with the characterization of *P. falciparum*, based on transcriptome [225, 226] and proteome [224, 227] data. A considerable amount of the findings in the above publications are reflected in the results of the used integrative analysis methods. These findings are more detailed through the association with a specific life cycle stage rather than, e.g. the whole intraerythrocytic cycle as well as through the association of a child GO term instead of a parent GO term to the corresponding stage. This study unifies individual findings from several publications of the past 25 years of research. Not all results from the publications mentioned above are present in the resulting network. This could be due to the fact that none of the cited publications, except [103], used the same data sets as in this integrative scenario. Llinás *et al.* in [225] compared the three *P. falciparum* strains 3D7, Dd2 and HB3 through the measurement of the gene expression profiles of 6287, 5294 and 6415 genes during the intraerythrocytic cycle. Bozdech *et al.* in [226] considered in their analysis of the intraerythrocytic cycle transcriptome the expression of 5508 genes. LaCount *et al.* analysed in [227] 1267 proteins for their protein interaction network of *P. falciparum.* In [224], Florens *et al.* use approximately 2400 proteins in order to create a proteomic view of the *P. falciparum* life cycle. The other studies are based on lab experiments on smaller groups of genes or proteins [223, 228–230, 232].

All in all, the combined network view of life cycle stage dependent GO term association provides a new overview for vaccine research and offers novel insight in the interdependencies between life cycle stages. Key biological processes were identified which may be potential targets in further *P. falciparum* vaccine researcher.

4.2 Analysis of Hemocyte and Granulocyte Immune Response of *Anopheles gambiae*

This section is dedicated to the discussion of the structural concordance between the hemocyte and granulocyte immune response and is based on the co-authored publication [205].

The MCIA results across all three comparisons suggest that the greatest degree of post-transcriptional regulation occurs after an infectious blood meal (PFvSF), followed by the effects of blood-feeding (BFvSF). The segment corresponding to PFvSF is the longest in this comparison showing the lowest degree of co-structure between transcript and corresponding protein profile. This could be the result of an intense post-transcriptional regulation program induced by the infectious blood meal. In contrast, the effects of parasite infection when compared to blood-feeding alone (PFvBF) show the highest concordance (Figure 3.9). Based on the pair-wise RV-coefficient, the highest post-transcriptional regulation can be observed for the immune-specific proteome (Figure 3.10), implying that components of the mosquito immune response are more likely to undergo translational regulation. Similar analyses of the proliferation-specific proteome suggest that transcript and protein expression are tightly linked (Figure 3.11).

4.2.1 Hemocyte-specific MCIA

In our global analysis of mosquito hemocyte transcripts/proteins, we identified a very high co-structure between the two data sets (RV-coefficient of 0.97). Notably, several unique proteins featured prominently in our MCIA comparisons (Figure 3.9) that were also independently identified in the enrichment analysis. A Vitellogenic Carboxypeptidase (VCP)-like protein (AGAP007505) in response to *P. falciparum* infection (PFvBF) was identified that was also significantly enriched in the PF sample. The mammalian ortholog of VCP-like has been implicated in the maturation of monocytes into macrophages [234], and may have similar roles in mosquito hemocyte activation. Additional proteins, Vigilin (AGAP005467) and a von Willebrand factor A-domain containing protein (AGAP000545), were also identified in enriched BF samples. The remainder of proteins highlighted in the global hemocyte MCIA analysis did not show a significant enrichment across the sample treatments.

4.2.2 Immune-specific MCIA

Similar to Figure 3.9, the MCIA analysis of immune-specific transcripts/proteins indicate a very high agreement between the two data sets with a RV-coefficient of 0.96 (Figure 3.10). The proteins that are specifically associated with the PFvBF comparison include LRIM16A (AGAP028028), CLIPB5 (AGAP004148), CLIPA2 (AGAP011790), and LRIM15 (AGAP007045), with LRIM16A and LRIM15 showing significant enrichment in PF samples. Both are members of a leucine-rich repeat family of proteins implicated in mosquito immunity [234] and contain predicted transmembrane domains, suggesting that these LRIM proteins could be candidate surface markers of activated granulocytes following *P. falciparum* infection. Previously implicated in the melanization response [235], recent reports have identified that CLIPA2 serves as a negative regulator of TEP1 function to avoid hyper-immune activation in response to pathogen challenge [236]. In addition, CLIPA5 (AGAP011787), LRIM17 (AGAP005693), and LRIM8B (AGAP0007456) define the PFvSF comparison, while CLIPA1 (AGAP011791) and CLIPB9 (AGAP0013442) highlight the response to blood-feeding (BFvSF).

4.2.3 Proliferation-specific MCIA

Although the proteins cluster far from the genes, the proliferation-specific (Figure 3.11) RV-coefficient is 0.99, which is the highest agreement found in our MCIA analyses (Figure 3.11) and is likely due to the relatively few genes/proteins used in the analysis. The three proteins that cluster with the PFvBF comparison are Ras-related Rab7A (AGAP001617), Ras homology gene family member A (AGAP005160), and Ras-related Rab5C (AGAP007901). Of these three, only the latter two are enriched in the PF and only Rab5C falls below the $p < 0.01$ stringency cutoff. However, the MCIA Features Axis 1 value of -2.07 for Rab7A (Figure 3.11, Table B.4) is high and projected in the direction of the PFvBF comparison. This suggests that it may be a potential marker for a nuanced granulocyte proliferation response to parasite infection and thus deserves future examination. This data is in agreement with the observation that Ras superfamily GTPases were down-regulated at the protein level among granulocytes in general, but increased in protein abundance in response to blood–feeding and *P. falciparum* infection.

4.3 Traditional Gene Set Enrichment Analysis of Human miRNAs

Examination of the results of the traditional GSEA shows that the two sets of enriched neuron related biological processes are very similar. The GO analysis of the CSF-MPs associated miRNAs detected 35 neuron related biological processes. These are similar to the 36 computed for all miRNAs identified by Bioanalyzer profiling. The GO analysis of the CSF-MPs associated miRNAs revealed two additional neuron related biological processes as statistically highly significant: *regulation of neurological system process* and *neuron fate specification*. In the group of all identified miRNAs three processes were detected that are not present in the GO analysis of the CSF-MPs associated miRNAs: *neuron projection regeneration*, *cell morphogenesis involved in neuron differentiation* and *axongenesis*. These differences result from the additional 296 validated targets of all miRNAs compared to the validated target genes of the CSF-MPs associated miRNAs.

Although the two sets of miRNAs differed by approximately 50%, the enriched neuron related biological processes were almost identical. The reasons why this might be the case include unvalidated or still undiscovered miRNA target genes. Another reason may be the fact that the traditional gene set enrichment analysis works on flat lists of genes, not taking into account any additional information like network topology.

4.4 Integrative Pathway Enrichment Analysis of Tumor and Stroma in Ovarian Cancer

Ovarian cancer is the fifth leading cause of cancer death in women world wide. Most women are diagnosed with advanced stage disease and consequently have a poor probability of survival after five years. Tumor-debulking surgery followed by platinum-taxane combination chemotherapy has remained first-line standard of care for decades, and although the initial response rate is good (70-80%) most will recur and succumb to chemoresistant disease.

Studies which have sought to define the molecular subtypes of high grade serous ovarian cancer [237–241] report that ovarian cancers often exhibit multiple subtype gene expression signatures, are often assigned to more than one subtype [137, 237, 242] and that these subtypes are not reproducibly associated with outcome [240, 242, 243]. Molecular subtypes defined from gene or protein expression data from the same tumor samples (TCGA study) share weak overlap [237, 240, 242]. While the lack of robust molecular subtype classifiers may be due to weaknesses in study design [244–246], our collaborator Aedin Culhane from Dana-Farber Cancer Institute and Harvard School of Public Health has found that variability in the proportion of tumor stroma in molecular profiling studies biases gene signature and subtype discovery [manuscript in preparation]. Only a few gene expression studies microdissect tumor tissue and the proportion of tumor stroma varies considerably both within and between studies.

Casey *et al.* emphasize the role of the tumor microenvironment by its double role as a cause of tumorigenesis and as a tumor interaction partner during subsequent development. As pointed out in [247], indirect and direct cellular interplay leads to a dynamical symbiotic development based on the manipulation of the cellular proliferation, growth, and metabolism, as well as angiogenesis and hypoxia and innate and adaptive immunity.

After realizing the shortcomings of the traditional gene set enrichment approach and their effects, a novel integrative pathway analysis is proposed and applied on 38 matched and microdissected tumor and stroma samples from ovarian cancer patients. The results of the traditional GSEA are presented in Section 3.3.

The most remarkable advantage of IPEA over traditional GSEA is its ability to account for the network topology of a biological pathway. At the same time, this means that IPEA can only be applied to gene sets with a known network topology. Here, Reactome database is used. Alternatively, the KEGG database could be employed. However, accessing KEGG information is tedious as each pathway has to be downloaded separately.

The use of MCIA as a first step of IPEA is optional. Any other integrative analysis method can be used instead. MCIA was chosen due to its ability to capture common trends between the data sets. As a matter of fact, IPEA can be applied to any list of features that are ranked, e.g. differentially expressed genes or proteins. This makes IPEA versatile and suitable for a wide range of applications.

The result of IPEA is a network of enriched pathways and activated genes, displayed as a double bipartite graph. Actionable target genes (validated set of genes involved in other cancers and linked to drugs which are already used in cancer therapy) are superimposed on the resulting network. One can notice that these actionable target genes are not only active in tumor but also in stroma, emphasizing the role played by stroma in ovarian cancer. Four actionable target genes were present in the resulting network: CTNNB1 active in up-regulated tumor pathways, ERBB4 active in up-regulated tumor and stroma pathways, SMAD4 active in up-regulated stroma pathways and PIK3CB active in down-regulated stroma pathways. In the resulting network only one target gene was attributed to tumor while all others are linked to the stroma, underlining the key role played by stroma in ovarian caner.

Actionable Target Genes

CTNNB1 was shown to have prognostic effects in colorectal [248] and gastric [249] cancer. According to the TARGET database [222], its activation may mediate resistance to EGFR TKIs, PI3K inhibitors and AKT inhibitors. Additionally, the activation of CTNNB1 may predict sensitivity to inhibitors of WNT signaling. Please note that according to MCIA, CTNNB1 is highly activated (high score on the positive side of the first MCIA axis) in tumor and thus, is a promising target gene.

ERBB4 mutations predict sensitivity to Lapatinib [222] and were identified by IPEA as highly active in tumor and stroma. This drug is already successfully used in the treatment of breast cancer [250]. Signaling by ERBB4 (enriched in tumor and stroma) triggers a rich network of pathways, culminating in responses ranging from cell division to cell death and from motility to adhesion [251]. Due to this, diverse other targets from these pathways could also be interesting for further investigation.

SMAD4 was shown to play an important role in the development of some gastrointestinal tumors [252]. It was also shown that PIK3CB inhibition produces synthetic lethality when combined with estrogen deprivation in estrogen receptor positive breast cancer [253] which is in concordance with IPEA findings where PIK3CB is highly repressed in stroma. Additionally, the role of the PI3K pathway as a drug target in human cancer was emphasized in [254]. I.e., SMAD4 and PIK3CB both exhibit potential as actionable target genes in ovarian cancer therapy.

The advantage of inspecting known actionable target genes lies in the fact that there already exists detailed knowledge such as matching drugs, which can be repurposed.

IPEA also identified other potential target genes. E.g. NRCAM which is activated in tumor and stroma, was linked to different cancers [255, 256] and, of particular interest, has a crucial role in tumor progression, especially during cell invasion and metastasis [257]. These functions are in complete agreement with the finding shown here where NRCAM links tumor to stroma suggesting a key role in stroma cell invasion by tumor cells.

Enriched Pathways

A detailed inspection of the enriched pathways emphasizes that the first MCIA axis is partly driven by DNA repair, which is known (respectively its down regulation) to be one of the key risk factors in ovarian cancer. Pathways involved in DNA repair are: *Regulation of the Fanconi Anemia Pathway* (tumor down) which is also known as the BRCA pathway (gene FANCD1 is BRCA2) [258], *Formation of transcription coupled nucleotide excision repair (TC-NER)* and *Dual incision reaction in TC-NER* (tumor up) [259]. As pointed out by Helleday *et al.* in [260], DNA repair pathways can enable tumor cells to survive chemotherapeutic induced DNA damage while alterations in DNA repair pathways that arise during tumor development can make some cancer cells reliant on a reduced set of DNA repair pathways for survival.

As described by Wang *et al.* in [261], the interaction between immune system and tumor is complex and dynamic. Major components of anti-tumor immunity are T and T effector cells together with B cells and macrophages, natural killer (NK) cells and NK-T cells. These cells recognize and kill tumor cells inducing a complex adaptive and innate immune response. In order to overcome the immune response of the cell, tumor cells manipulate their microenvironment and induce the development of immunosuppressive cells. IPEA identified several immune system pathways: *T cell receptor (TCR) signaling* and *Downstream*

TCR signaling (both in stroma down) as well as *Inflammasomes*, *ZBP1(DAI) mediated induction of type I IFNs* (tumor up) and *Interleukin-1 signaling* (stroma up).

Wang *et al.* describe another intracellular metabolic interaction which is based on the metabolic derangement of the tumor environment where tumor-surrounding cells may either compromise or support highly metabolic demanding tumor cells by competing nutrients or by forming a metabolic symbiosis [261]. While a high BMI is a risk factor [262] for cancer, in [263] was shown that ovarian cancer prefers to metastasize to the omentum, an organ primarily composed of adipocytes. According to Nieman *et al.* adipocytes act as an energy source for the cancer cells and their role is mediated by fatty acid-binding protein 4 (FABP4, also known as aP2) [263]. A number of relevant pathways were found to be enriched in this analysis: activation of *Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis* and FABP4 in stroma, activation of *Beta oxidation of palmitoyl-CoA to myristoyl-CoA* and *Beta oxidation of myristoyl-CoA to lauroyl-CoA* in tumor (both participate in fatty acid metabolism).

In her doctoral thesis, Walpole investigated the role of the hormones ghrelin and obestatin in ovarian cancer and could show that ghrelin and obestatin (two isoforms of preproghrelin) lead to an increase in cell migration and may play a role in cancer progression [264]. *Synthesis, secretion, and deacylation of Ghrelin* is activated in tumor according to IPEA. Additionally, Granata *et al.* showed that obestatin enhanced the *Translocation of GLUT4 to the plasma membrane* [265], a pathway repressed in tumor and stroma in the resulting network.

Justus *et al.* summarize the molecular connections between cancer cell metabolism and the tumor microenvironment [266]. One key aspect is the *Warburg effect* which is described as the preference of cancer cells to utilize glycolysis instead of oxidative phosphorylation for metabolism, even in the presence of oxygen. Dang showed that hypoxia, mainly mediated through the hypoxia-inducible factors (HIFs), enhances the Warburg effect by up regulation of glycolytic genes such as hexokinases, LDH-A, and GLUT [267]. Additional regulators include genetic factors such as oncogenes and tumor suppressors and microenvironmental factors such as acidosis, beside spatial hypoxia. Interestingly, in [266] it was pointed out that altered cancer cell metabolism can modulate the tumor microenvironment which plays important roles in cancer cell somatic evolution, metastasis, and therapeutic response.

One mechanism to repress hypoxia is the *Detoxification of Reactive Oxygen Species* [268] which is down-regulated in the microdissected tumor. Another regulator of hypoxia is *Reversible hydration of carbon dioxide* which according to IPEA is enriched in tumor and stroma and is mediated by carbonic anhydrases [269]. Kang *et al.* showed that hypoxia activated heat shock factor-1 (HSF1) [270] which was found enriched in tumor down. Additionally, recent studies [271] link HSF1 to overexpression of the human epidermal growth factor receptor-2 (HER2) which strongly correlates with tumor aggressiveness and poor prognosis in breast cancer.

Another pathway which is repressed in tumor in the resulting network is *Signaling by Hippo*. The importance of the Hippo pathway in human cancer was reviewed in [272] by Harvey *et al.*. They emphasized the function of the Hippo pathway in the control of organ size as well as the key role played by its deregulation in inducing tumors in model organisms. Additionally, it was pointed out that changes of this pathway were observed in a broad range of human carcinomas, including lung, colorectal, ovarian and liver cancer. Zhang *et al.*

showed that YAP (Yes-associated protein in the Hippo pathway) can enhance the transformed phenotype of ovarian cancer cell lines and can induce resistance to chemotherapeutic agents [273]. Additionally, it was shown that high nuclear YAP expression correlated with poor patient survival.

In their review, Branzei and Foiani describe regulative mechanisms of DNA repair within the cell cycle [274]. Numerous pathways involved in the cell cycle are enriched in the current analysis: *Activation of the pre-replicative complex* (tumor up), *Packaging of telomere ends* (tumor up) which is discussed in the context of cancer also in [275], *CDC6 association with the ORC: origin recognition complex* (tumor down), *Nuclear envelop breakdown* (tumor up), *Cyclin A/B1 associated events during G2/M transition* (stroma up), *Initiation of Nuclear Envelope Reformatio*n (stroma up) and *Nuclear envelope reassembly* (stroma up).

Collagen degradation is a key pathway, up-regulated in tumor and stroma, which participates in the degradation of the extracellular matrix. The role of the extracellular matrix degradation in cancer metastasis and cell invasion is emphasized in [276] while its importance in ovarian cancer is shown in [277, 278]. Four genes involved in collagen degradation are active in tumor COL9A2, COL9A1 and COL12A1 and stroma COL93A, representing possible entry points for the deactivation of this pathway.

De Alvaro *et al.* showed in [279] that tyrosine mediated impariment of tumor necrosis factor α (TNF-α) phosphorylation leads to insulin resistance on glucose uptake and to GLUT4 translocation to the plasma membrane. Repression of *GLUT4 translocation to plasma membrane* in tumor and stroma suggests that TNF-α, which is already used as a target in chemotherapy of melanoma [280], could also be a potential target in ovarian cancer treatment.

One biological pathway is enriched due to the combined effect of stroma up and stroma down: *GABA A receptor activation*. The GABA A receptor was shown to be highly differentially expressed in breast cancer tumor epithelium [281]. More recently it was discovered that over-expression of GABA stimulates pancreatic cancer growth [282]. This suggests that GABA A could be a potential target for the treatment of ovarian cancer.

In summary, IPEA identifies enriched up- and down-regulated pathways as well as highly activated or repressed genes. The double bipartite network provides an unprecedented view on the cross-talk between tumor and stroma in ovarian cancer. Already studied actionable target genes were superimposed on the resulting network. Additionally, other genes and pathways were identified which were already linked to other cancers and due to this could be potential targets in ovarian cancer therapy.

4.5 Cross-Species Comparison of *Pichia pastoris* and CHO Cells

Biopharmaceutical human protein therapeutics have become increasingly important in the treatment of various diseases since the first recombinant protein (human insulin) was engineered over thirty years ago. Established production systems [283] include *Escherichia coli*, *Saccharomyces cerevisiae* and Chinese hamster ovary (CHO) cells. The complexity of the desired product in combination with the required post-translational modifications (PTM) are key factors in the selection of the appropriate expression system. Human-like PTMs can be achieved by mammalian cell lines which, although they have been used for a few decades, suffer from disadvantages such as: low growth rate, low biomass density, high media costs and time-consuming cell line development, as it has been pointed out by Maccani *et al.* in [284]. Another expression system capable of protein folding and proper PTMs (especially glycosylation) is the yeast *Pichia pastoris*. In general, the use of yeast as an expression system has numerous advantages [285], such as protein secretion into the culture stock, absence of endotoxins and viral DNA.

In order to further investigate similarities and differences between *P. pastoris* and CHO cells, a cross-species comparison study was performed on gene expression and protein abundance data sets measured in CHO cells, the established expression system for human heterologous protein production, and *P. pastoris* which features diverse advantages over CHO cells. The comparison is based on two model proteins with different complexities which challenge the expression systems in various ways: i) human serum albumin (HSA), a monomeric and ii) non-glycosylated protein and a more complex model protein, a single chain Fv-Fc fusion antibody derived from the monoclonal anti-HIV-1 antibody 3D6 which is homodimeric and contains the Fc-specific glycosylation.

Using MCIA, a detailed analysis of the four strains was performed (see Figure 3.20A). To our knowledge, such a comparison was not done before. The integrative analysis was performed on the complete data sets, but also on subsets. To better understand the processes induced by the four strains, ribosome, ribosome deficient and secretion relevant subsets were investigated. These subsets are believed to be of particular use in elucidating the effects of the two model proteins and of the different protein production regimes. Comparison between all strains, between high and low producers and between 3D6 and HSA producers are performed. Additionally, the comparisons are performed on a data set basis: *P. pastoris* transcriptome and proteome, CHO transcriptome and proteome. In this way, a detailed strain, data set and feature subset specific results are discussed.

4.5.1 MCIA Axis 1 Is Driven by the Different Model Proteins While MCIA Axis 2 Is Driven by their Produced Amount

In all analyses, MCIA Axis 1 comprising the largest variance (51%-56%), separates the four strains by the protein they are producing: 3D6 H:WT and 3D6 L:WT on the negative side and HSA H:WT together with HSA L:WT on the positive side of MCIA Axis 1, suggesting that the most prominent differences in the measured data sets are induced by the different proteins.

MCIA Axis 2 (25%-30% of the variance) separates the strains by the amount of produced protein: the high producers are located on the positive side, while the low producers are on the negative side of MCIA Axis 2. This observation suggests that that this is the second most prominent reason leading to differences between the strains.

These observations emphasize the suitability of MCIA for the characterization of the four engineered strains. 71%-91% of the variance captured by the analysis is caused by the produced proteins and their amount. This is not surprising, given the way the study was designed, nevertheless, it is reassuring that the performed analyses captures exactly the required information.

4.5.2 Protein Production Challenges *P. pastoris* More Than CHO

In all four analyses, the co-structure between the measured data sets (*P. pastoris* transcriptome, *P. pastoris* proteome, CHO transcriptome and CHO proteome) was calculated with the RV coefficient. The resulted values are summarized in Table 4.1. For all computed RV values, correponding p-values were calculated with a permutation test (n = 1000 repetitions). All RV coefficients proved to be significant.

Table 4.1: Summary of all RV coefficients.

Sample Set	All features	Secretion	Without Ribosomes	Ribosomes
Transcriptomes	0.935	0.991	0.922	0.965
Proteomes	0.896	0.786	0.886	0.826
CHO	0.795	0.826	0.809	0.795
P. pastoris	0.612	0.601	0.597	0.523

Notably, the highest co-structure was found between the two transcriptomes followed by the two proteomes, regardless whether complete data sets were inspected or only subsets. Additionally, in all performed analyses can be observed that the co-structure between CHO transcriptome and CHO proteome is higher that between *P. pastoris* transcriptome and proteome.

In the analysis of the complete data sets a high co-structure can be observed between the two transcriptomes and the two proteomes, as well as within CHO. The lowest agreement was computed between *P. pastoris* transciptome and proteome. The situation is similar to the data sets without the ribosome relevant features. It is worth to mention that in this scenario the lowest agreement was measured between the two transcriptomes.

The CHO transriptome and proteome share the highest co-structure in the secretion relevant analysis. In this case only the agreement between the transcriptomes of *P. pastoris* and CHO show a higher RV. Furthermore, this is the highest level of agreement measured in all performed analyses. Interestingly, in the secretion relevant analysis, the agreement within CHO omics is higher than within *P. pastoris* omics. This may suggest that protein secretion challenges *P. pastoris* more than CHO cells.

In the ribosome relevant analysis, it can be noticed that the co-structure within *P. pastoris* is much lower than within CHO, representing the lowest RV value computed in this study. While the agreement between the two transcriptomes is lower, the co-structure between the proteomes is higher than in the secretion relevant analysis. Exactly the opposite (higher RV for the transcriptomes and lower RV for the proteomes) happens when compared to the analysis of the complete data sets. These results suggest again that complex processes have to be induced in *P. pastoris* for the model protein expression.

In summary, all analyses seem to suggest that protein production challenges *P. pastoris* more than CHO cells and that their ribosomes seem to play a key role in this complex process.

4.5.3 Effect of the Expressed Proteins and their Produced Quantities

In order to characterize the effect of the different proteins and their produced quantity, the distances between the four strains and the origin of the MCIA coordinate system were calculated. This is done on a per data set basis and as a mean value of the different data sets. The calculation is based on the first three MCIA axes and was performed in all four analyses (see Tables 3.13-3.16).

Descending order of the mean distances results in an consistent ranking across all analyses: 3D6 H:WT followed by HSA H:WT and 3D6 L:WT followed by HSA L:WT, suggesting that the high producers induce more changes in the data sets than the low producers, and if compared within the high or the low regime, 3D6 induces more changes than HSA. Inspection of the data set specific distances reveals more details.

P. pastoris Transcriptome

Table 4.2 summarizes the distances from the four strains to the origin of the plot measured for *P. pastoris* transcriptome. When all features are considered, HSA H:WT induces the highest changes, followed by 3D6 H:WT, 3D6 L:WT and finally HSA L:WT. The secretion relevant transcriptome is almost equally challenged by all four strains. The 3D6 strains induce slightly (first decimal place) more changes but the differences between high and low producers of the same protein are almost negligible (second decimal place). When all features without the ribosome relevant one are considered, the effects are similar to the complete data set analysis: HSA H:WT induces most of the changes, followed by 3D6 H:WT, 3D6 L:WT and finally HSA L:WT. In the ribosome relevant analysis, it can be seen that 3D6 L:WT causes the most predominant changes while 3D6 H:WT is equally challenging as HSA H:WT, followed by HSA L:WT which induces only slightly (second decimal place) less predominant changes.

Table 4.2: *P. pastoris* Transcriptome - Distances to Origin

Comparison	All features	Secretion	Without Ribosomes	Ribosomes
3D6 L:WT	2.143	2.558	2.086	2.716
3D6 H:WT	2.776	2.533	2.788	2.499
HSA L:WT	1.810	2.361	1.782	2.447
HSA H:WT	2.925	2.330	2.966	2.499

The findings can be summarized as follows: For the *P. pastoris* transcriptome, when all features or all features without the ribosome associated ones are considered, the largest difference can be observed between the high and the low producers. These differences are not so pronounced in the secretion and ribosome relevant analyses. These results can also be seen in the *P. pastoris* transcriptome relevant region of Figures 3.21A–3.22B.

The *P. pastoris* transcriptome relevant region of Figures 3.21A and 3.22A show that the genes are mostly projected along the first MCIA axis and therefore can be associated to high (PAS_chr1-4_0681, PAS_chr4_0002, MNN4, TOS8, ADH6, SGN1-1, CNE1, SIT1-1, SEC61, ArbD) and to low producers (FLO5-2, PAS_chr3_0008, PAS_chr3_1144, PAS_chr3_0012, PAS_chr2_-1_0550, FIG1, PRM1, KAR4, STE3).

Notably, MNN4 mutations are linked to human-like N-glycosylation of proteins in *P. pastoris* [286].

In the secretion relevant analysis (upper right of Figure 3.21B), where it was observed that all strains challenge the *P. pastoris* transcriptome similarly, it can be seen that *P. pastoris* genes are projected in the direction of 3D6 H:WT: Pipas_chr1-4_0405, Pipas_chr2-1_0835, Pipas_chr3-0401, Pipas_chr2-1_0291, Pipas_chr2-2_0210, ERV29; HSA H:WT: Pipas_chr3-4_0230; 3D6 L:WT: Pipas_chr2-2_0346, Pipas_chr1-4_0225, Pipas_chr1-1_0023, ARF3, CHS7; HSA L:WT: Pipas_chr4-0452, Pipas_chr2-1_0726, Pipas_chr1-4_0519, PpBMT1; high producers: CNE1, SEC61, OST3, KAR2, PDI1; low producers: HRI1.

Gasser *et al.* report a strong enhancement in the transcription of CNE1 (calnexin) and SEC53 (phosphomannomutase), genes involved in ER quality control and glycosylation, in response to Hac1 overexpression in *P. pastoris* [287].

The feature plot of the *P. pastoris* transcriptome in the ribosome related analysis is dominated by the gene RPS22A and PAS_chr1-3_0305 which are projected in the direction of 3D6 L:WT and HSA L:WT, respectively. Additionally, YAK1-2, SED1 and HCA4 can be associated to low producers while PPE1, TVP18 and ENP1 are associated to high producers. Pipas_chr1-1_0414, Pipas_chr3_0573 and Pipas_chr3_0893 are associated to HSA producing strains while Pipas_chr1-3_0256, Pipas_chr4_0407, Pipas_chr3_0183 and Pipas_chr3-1_0518 are associated to 3D6 producer. YMR295C is associated to 3D6 H:WT and MDM38 to HSA H:WT.

Additionally, Figure 3.24 shows the corresponding projections of the investigated GO Terms for each data set and for each analysis. Associations of GO terms to the four strains resulting from Figure 3.23A are summarized in Table 3.27.

P. pastoris Proteome

Table 4.2 summarizes the distances from the four strains to the origin. When all features are considered, 3D6 H:WT induces the most prominent changes, followed by HSA H:WT, 3D6 L:WT and finally, HSA L:WT. Looking at the *P. pastoris* proteome relevant part of Figures 3.21A and 3.22A, one can observe that most proteins are associated to the high (MCP2-1, HCH1, PAS_chr2-1_0127, NUC1, Zeo) and to the low producers (GOT1, ZRC1, PAS_chr2-1_0883, YNL181W, MRPL31, YOL09810, GLO1, PEX6, CDC50).

Nathan *et al.* identified HCH1 as a multicopy suppressors of a *Saccharomyces cerevisiae* Hsp90 loss-of-function mutation in [288]. Hsp90 participates in a multicomponent chaperone system which is in charge of a set of target proteins that play key roles in the regulation of cell growth and development.

Similar to the *P. pastoris* transcriptome, the secretion relevant analysis shows that the strains induce high changes relative to the subset analyses: 3D6 strains slightly larger than HSA strains. Interestingly, 3D6 L:WT seems to challenge the *P. pastoris* proteome more than 3D6 H:WT. Considering the *P. pastoris* proteome relevant part of Figure 3.21B, one can observe that two proteins dominate the plot: GOT1 and YNL181W which can be associated to low producers and HCH1 which is projected in the direction of HSA H:WT. No other proteins were projected in the directions of HSH H:WT and 3D6 L:WT. Additional associations to HSA L:WT may be: MNN11 and Pipas_chr1-4_0105 while GLO3, GRX6 and Pipas_chr4_0675 are projected in the direction of 3D6 H:WT.

Table 4.3: *P. pastoris* Proteome - Distances to Origin

Comparison	All features	Secretion	Without Ribosomes	Ribosomes
3D6 L:WT	2.167	2.744	2.139	2.164
3D6 H:WT	2.888	2.533	2.864	3.098
HSA L:WT	2.186	2.361	2.198	2.317
HSA H:WT	2.513	2.330	2.548	2.490

The ribosome relevant analysis shows that 3D6 H:WT induces the highest changes in the *P. pastoris* proteome, followed by HSA H:WT similar to HSA L:WT and 3D6 L:WT. Examining the *P. pastoris* proteome relevant part of Figure 3.22B, one first notices GIS2 associated to 3D6 H:WT followed by KRI1 and Pipas_chr4_0269. Furthermore, RRP3, MRPL1 and ESF1 can be associated with high producers in general. Pipas_chr1-1_0236 is projected in the direction of HSA H:WT while Pipas_chr1-1_0187, RRP17, RIX7, SR09 and KRR1 are projected in the direction of HSA L:WT. MRPL31 is highly associated with low producers.

In summary, the *P. pastoris* proteome is almost equally challenged by all four strains. Interestingly, 3D6 H:WT induces the largest overall changes in the ribosome relevant proteome.

Additionally, Figure 3.24 shows the corresponding projections of the investigated GO Terms for each data set and for each analysis. Associations of GO Terms to the four strains resulting from Figure 3.23A are summarized in Table 3.28.

CHO Transcriptome

Table 4.4 summarizes the distances from the four strains to the origin. When all features are considered, the high producers induce more changes than the low producers, and 3D6 more than HSA, similar to the results from the analysis done without the ribosome relevant features.

Inspection of the CHO transcriptome relevant region of Figures 3.21A and 3.22A shows CHO gene associations to: 3D6 H:WT (Col6a1, Fabp4, Pstpip1, BGI_CHO_2136, BGI_CHO_17350), 3D6 L:WT (Nppb, Il33, Dsp), HSA L:WT (Sirt5, BGI_CHO_16355, BGI_CHO_17354, Tcf7l2), HSA H:WT (Ill13ra2, Akr1b7, BGI_CHO_14472), 3D6 strains in general (BGI_CHO_12844, BGI_CHO_12771, BGI_CHO_12772) and HSA strains in general (BGI_CHO_18500, BGI_CHO_18492, BGI_CHO_15425, BGI_CHO_2274).

Fabp4 was linked to the regulation of phosphofructokinase in a multi omics comparison study between protein producing Hek293 cells and the parental cell line. Hek293 cells are employed for stable expression of proteins where PTMs performed by CHO cells are inadequate [289].

Table 4.4: CHO Transcriptome - Distances to Origin

Comparison	All features	Secretion	Without Ribosomes	Ribosomes
3D6 L:WT	2.358	2.559	2.358	2.366
3D6 H:WT	2.704	2.640	2.705	2.744
HSA L:WT	2.120	2.064	2.115	2.116
HSA H:WT	2.601	2.678	2.597	2.871

Inspection of the secretion related CHO transcriptome distances to origin reveals similar challenges induced by (in descending order) HSA H:WT, 3D6 H:WT, 3D6 L:WT and less prominent changes caused by HSA L:WT. The CHO transcriptome relevant region of Figure 3.21B reveals numerous CHO gene associations tp: HSA H:WT (BGI_CHO_3432), high producers in general (Myh3, Ube2d3, Sec24d, Slc35a3, Rhoc, Ssrc), 3D6 H:WT (Hspd1, Bcap31, Gbf1, BGI_CHO_3428), 3D6 L:WT (Myh7, Fkbp10, Dnajc3, Golt1b, BGI_CHO_14096, BGI_CHO_1085, BGI_CHO_9990) and HSA L:WT (Ube2j2, BGI_CHO_4267, BGI_CHO_9130).

The ranking of the strains in the ribosome relevant analysis is equal to the ranking in the secretion relevant analysis (in descending order): HSA H:WT, 3D6 H:WT, 3D6 L:WT and HSA L:WT. The CHO transcriptome relevant region of Figure 3.22B reveals numerous CHO gene associations to: HSA H:WT (BGI_CHO_2959, BGI_CHO_19020, BGI_CHO_6771, BGI_CHO_19386), 3D6 H:WT (Gm15421, Rpl7), 3D6 L:WT (BGI_CHO_5780, BGI_CHO_4926, BGI_CHO_4152, BGI_CHO_4158) and HSA L:WT (Mrpl13, Mrpl20, Mrpl23, Rps9).

Additionally, Figure 3.24 shows the corresponding projections of the investigated GO Terms for each data set and for each analysis. Associations of GO Terms to the four strains resulting from Figure 3.23A are summarized in Table 3.28.

CHO Proteome

Table 4.5 summarizes the distances from the four strains to the origin. If all features are considered, the high producers induce more prominent changes than the low producers and 3D6 more than HSA. These results correspond to the analysis done without ribosome relevant features.

The CHO proteome relevant region of Figures 3.21A and 3.22A exhibit CHO proteins associated to: HSA H:WT (EGV96396.1, ERE91297.1, ERE70326.1, CHO_pq_1172, CHO_pq_80), HSA L:WT (ERE92264.1, EGW00497.1, EGW06084.1, ERE74789.1, CHO_pq_573, CHO_pq_1021), 3D6 H:WT (EGV93171.1, EGW08318.1, ERE78825.1, CHO_pq_29, CHO_pq_4, CHO_pq_1072) and 3D6 L:WT (EGV95081.1, EGW04688.1, ERE69844.1, XP_007607708.1, EGW02679.1, ERE67838.1, CHO_pq_122, CHO_pq_1053).

EGV96396.1 is described as a oxysterol-binding protein-related protein (see Table 3.24). Fang *et al.* showed that Kes1p (a oxysterol-binding protein in yeast) participates in a regulatory pathway for yeast Golgi-derived transport vesicle biogenesis [290].

The secretion related CHO proteome distances to origin reveal major changes induced by HSA H:WT and 3D6 H:WT and minor changes induced by 3D6 L:WT and HSA L:WT. The CHO proteome relevant region of Figure 3.21B reveals numerous CHO protein associations to: 3D6 L:WT (EGV95081.1, XP_007633208.1, XP_007627449.1), 3D6 H:WT (XP_007637129.1, XP_007622928.1, CHO_pq_1037), HSA H:WT (XP_007621275.1, CHO_pq_1017) and HSA L:WT (XP_007621498.1, EGV95752.1).

According to the ribosome relevant analysis, 3D6 production induces more changes than HSA production, and high producers challenge the strains more than low producers. The CHO proteome relevant region of Figure 3.22B reveals numerous CHO protein associations to: high producers (XP_007634407.1, ERE79461.1), HSA H:WT (XP_007626395.1, CHO_pq_132), to 3D6 H:WT (XP_007637129.1), 3D6 L:WT (XP_007633757.1, CHO_pq_122) and HSA H:WT (EGW00497.1, ERE92059.1, CHO_pq_1163, CHO_pq_1151).

Table 4.5: CHO Proteome - Distances to Origin

Comparison	All features	Secretion	Without Ribosomes	Ribosomes
3D6 L:WT	2.140	2.222	2.134	1.980
3D6 H:WT	2.919	2.706	2.908	3.101
HSA L:WT	1.909	1.670	1.886	1.967
HSA H:WT	2.718	3.146	2.744	2.906

Additionally, Figure 3.24 shows the corresponding projections of the investigated GO Terms for each data set and for each analysis. Associations of GO Terms to the four strains resulting from Figure 3.23A are summarized in Table 3.28.

4.5.4 Further Characterization

To further characterize the four engineered strains, distances between any two of them were computed in the MCIA space. Data set specific and mean distances were calculated. The results are summarized in Tables 3.17–3.20.

Inspection of the four tables results in the following observations:

- The distance between 3D6 producers is always larger than the distance between the HSA producers.
- The distance between the high producers is always larger than the distance between the low producers.

Shifting 3D6 Expression from Low to High Is More Challenging than the Corresponding HSA Shift, Except in the CHO Secretion Relevant Proteome

In order to assess the impact of changing from a low to a high producer, the difference between the distance 3D6 L:WT to 3D6 H:WT and the distance HSA L:WT to HSA H:WT was calculated based on the Tables 3.17–3.20 and summarized in Table 4.6.

In all analyses the distance between 3D6 producers is larger than the distance between the HSA producers. The largest differences can be observed between the pichia proteomes, in the secretion and ribosome relevant analysis.

In all analyses and for all data sets, increasing the produced amount of 3D6 is more challenging than expressing more HSA product. One exception was found: in the secretion relevant analysis, when the distance between the CHO proteomes is considered, the distance between HSA L:WT and HSA H:WT is higher than the distance between 3D6 L:WT and 3D6 H:WT. This may suggest that, in the CHO proteome, increasing the amount of HSA results in more prominent changes than increasing the amount of 3D6.

Table 4.6: Difference between the distance {3D6 L:WT to 3D6 H:WT} and the distance {HSA L:WT to HSA H:WT}

Data set	All features	Secretion	Without Ribosomes	Ribosomes
Mean	0.526	0.596	0.445	0.552
Pichia Transcriptomes	0.151	0.680	0.047	0.563
Pichia Proteomes	0.673	1.094	0.503	1.049
CHO Transcriptomes	0.568	0.753	0.585	0.158
CHO Proteomes	0.689	-0.139	0.640	0.437

P. pastoris Secretion Relevant Proteome and Ribosome Relevant Transcriptome Are More Sensitive to Changes Between the Low Producers than to Changes Between the High Producers

In order to assess the amount of changes resulting from producing different proteins in the same amount, the difference between the distance 3D6 H:WT to HSA H:WT and the distance 3D6 L:WT to HSA L:WT was computed for all analyses based on the Tables 3.17–3.20 and summarized in Table 4.7.

It can be observed that the difference between the two calculated distances is always positive indicating that the changes resulted from producing different proteins in a high amount induces a larger spectra of biological processes than producing different proteins in a low amount. The largest differences were found in the CHO secretion and ribosome relevant proteomes. Nevertheless, two exceptions were found.

The first case occurs when one inspects the distances between the *P. pastoris* proteoms in the secretion relevant analysis. There, the studied difference is negative. This could indicate that changing from producing 3D6 in a low amount to HSA in the same regime induces more changes in the *P. pastoris* secretion relevant proteome than when moving from producing 3D6 in a high amount to producing HSA in the same regime.

The second case occurs when one inspects the distances between the *P. pastoris* transcriptomes in the ribosome relevant analysis.There, the studied difference is negative. This could indicate that changing from producing 3D6 in a low amount to HSA in the same regime induces more changes in the *P. pastoris* ribosome relevant transcriptome than when moving from producing 3D6 in a high amount to producing HSA in the same regime.

Table 4.7: Difference between the distance {3D6 H:WT to HSA H:WT} and the distance {3D6 L:WT to HSA L:WT}

Data set	All features	Secretion	Without Ribosomes	Ribosomes
Mean	1.453	0.759	1.514	1.137
Pichia Transcriptomes	1.959	0.301	2.126	-0.188
Pichia Proteomes	1.185	-0.283	1.216	1.229
CHO Transcriptomes	0.910	0.756	0.910	1.231
CHO Proteomes	1.755	2.259	1.804	2.262

In summary: the distance between the high producers is always larger than the distance between the low producers except in the *P. pastoris* secretion relevant proteome and in the ribosome relevant transcriptome.

Comparison of Table 4.6 to Table 4.7 suggests that switching from 3D6 to HSA when the amount of product does not vary induces larger changes than shifting from low to high producers.

5 Conclusion

As high-throughput data measurements in biomedical studies have become routine, the challenges shifted from data generation to data analysis. In particular, the integration of multiple omics data sets is a promising but at the same time difficult to accomplish task.

In this thesis, three integrative analysis methods were studied and applied to several omics data sets: (multiple) co-inertia analysis (MCIA), generalized singular value decomposition (GSVD) and integrative biclustering (IBC). Additionally, a traditional gene set enrichment analysis (GSEA) was applied to human microRNAs. Finally, a novel integrative pathway enrichment approach (IPEA) was developed and employed to characterize tumor-stroma cross-talk in ovarian cancer omics data.

Three integrative analysis methods (MCIA, GSVD and IBC) were applied to the transcriptome and proteome of *P. falciparum*, the parasite causing malaria in humans. From the intersection of their results, a network of biological processes was derived which characterizes the parasite's life cycle stages and unifies numerous findings from the past 25 years of research in a single analysis.

The innate immune response of *A. gambiae*, the primary malaria mosquito vector in Sub-Saharan Africa, to sugar, blood and to *P. falciparum* infected blood feeding was analyzed. With MCIA a high structural concordance between the hemocyte transcriptome and the granulocytic hemocyte-specific, immune-specific and proliferation-specific proteome could be shown. The results across all three comparisons suggest that the highest degree of post-transcriptional regulation occurs after an infectious blood meal, followed by the effects of blood-feeding.

MCIA was employed for the cross-species comparison of the expression systems *P. pastoris* and Chinese hamster ovary (CHO) cells, challenged by the production of the two model proteins, HSA and 3D6. Detailed characterization of the strains results in three hypotheses: i) protein production challenges *P. pastoris* more than CHO, ii) production of 3D6 induces more changes than production of HSA and iii) producing a model protein in a high regime is more challenging than producing the same protein in a low regime. A number of secretion and ribosome relevant target genes and proteins were identified.

Traditional GSEA was applied to validated target genes of all microRNAs identified in samples from patients following traumatic brain injury as well as to the miRNAs associated with cerebrospinal fluid microparticles. The 36, respectively 35, enriched neuron related biological processes were almost identical between the two sets, although the overlap in the corresponding miRNA lists was below 50%.

As traditional GSEA is limited to flat gene lists, a novel integrative pathway enrichment approach (IPEA) was developed. IPEA combines scores from a multivariate analysis with pathway specific scores based on network topology. Enriched pathways computed by IPEA are characterized by a biologically relevant agreement between the measured data and the intrinsic structure of the pathways. IPEA visualizes the results as a double bipartite graph of activated features and enriched pathways. Applied to 38 matched tumor and stroma samples from ovarian cancer patients, IPEA reveals an unprecedented view of the cross-talk between tumor and stroma suggesting new targets, e.g. CTNNB1, ERBB4 and SMAD4 which have already shown their potential in the therapy of other cancers, for the treatment of ovarian cancer.

5.1 Challenges

Integrative data analysis is still in its infancy. Although a relatively large number of approaches are available, various aspects of integrative data analysis still have to be addressed. The co-authored review [150] summarized these challenges. While biostatisticians and computational biologists are extending and developing integrative analysis methods, few gold standard or canonical test data sets exist and therefore it is often difficult to compare the performance of different methods. The community needs to define a set of test datasets for this purpose.

Although several integrative analysis approaches have been applied to molecular data, little consideration is often given to the underlying data structure. For example PCA is frequently applied to count data with many zeros, when CA is more appropriate.

Most visualization approaches were designed for datasets with fewer features, and visualization and interpretation of plots with thousands of features can be complex.

Finally, interpretation of long lists of biological features (genes, proteins, miRNAs) remains a challenge and often one needs to search dispersed data sources to annotate these features. Within R, the Bioconductor annotation project greatly facilitates quick and easy access to them.

An attractive feature of decomposition-based integrative analysis methods is that feature annotation can be projected into the same space to determine a score for Gene Ontology terms such as biological processes but also for pathways from databases like Reactome.

Simultaneous analysis of omics data sets will produce ranked lists of features that are the most co-variant, features that are highly associated with a sample or condition, and features that are grouped together. These are on the same scale and can be concatenated to increase the power of gene set or pathway analysis. A challenge addressed in this thesis is the shortcoming of traditional GSEA to operate on flat lists of genes without taking into account other available information.

After performing an integrative GSEA such as IPEA, it would be informative to visualize, for example, the enriched pathway as a network by emphasizing the variables that contributed to its enrichment and the omics levels on which these were measured.

Bibliography

[1] Google Scholar Search: **integrative analysis**. 2014. [scholar.google.at online accessed 15-September-2014].

[2] Oxford English Dictionary Online: **-ome, comb. form.** Oxford University Press, Oxford, 1989. [www.oed.com online accessed 16-September-2014].

[3] World Health Organization: **genomics**. 2015. [http://www.who.int online accessed 16-September-2014].

[4] Mendel G: **Versuche über pflanzen-hybriden**. In **Verhandlungen des Naturforschenden Vereines in Brünn**. Naturforschender Verein, Brünn, 1866.

[5] Miescher F: **Über die chemische Zusammensetzung der Eiterzellen**. *Hoppe-Seyers Medizinisch-Chemische Untersuchungen* 1871. 4(4):441–460.

[6] Franklin RE and Gosling RG: **Molecular configuration in sodium thymonucleate**. *Nature* 1953. 171(4356):740–741.

[7] Watson JD and Crick FH: **A Structure for Deoxyribose Nucleic Acid**. *Nature* 1953. 171(4356):737–738.

[8] Pauling L and Corby RB: **Structure of the Nucleic Acids**. *Nature* 1953. 171(4356):364.

[9] Nirenberg M and Matthaei H: **The dependendence of cell-free protein synthesis in *E. Coli* upon naturally occurring or synthetic polyriconucleotides**. *Proceedings of the National academy of Sciences of the United States of America* 1961. 47(10):1588–1602.

[10] Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocombe PM and Smith M: **Nucleotide Sequence of Bacteriophage φX174**. *Journal of Molecular Biology* 1978. 125(2):225–246.

[11] Sanger F, Nicklen S and Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National academy of Sciences of the United States of America* 1977. 74(12):5364–5367.

[12] Sanger F, Coulson AR, Hong G, Hill D and Petersen Gd: **Nucleotide sequence of bacteriophage λ DNA**. *Journal of Molecular Biology* 1982. 162(4):729–773.

[13] Saiki R, Scharf S, Fred F, Millis K, Horn G, Erlich H and Norman A: **Enzymatic Amplification of β-Globulin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia**. *Science* 1985. 230(4732):1350–1354.

[14] Fleischmann RD, Adams MD, White O, Clayton R, Kirkness EF, Kerlavage R, Bult CJ, Tomb JF, Dougherty B and Merrick JM: **Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd.*** *Science* 1995. 269(5223):496–512.

[15] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.*: **Initial sequencing and analysis of the human genome**. *Nature* 2001. 409(6822):860–921.

[16] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans C, Holt R *et al.*: **The sequence of the human genome**. *Science* 2001. 291(5507):1304–1351.

[17] Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF *et al.*: **The genome sequence of *Drosophila melanogaster***. *Science* 2000. 287(5461):2185–2195.

[18] Crick F: **Central Dogma of Molecular Biology**. *Nature* 1970. 227(5258):561–563.

[19] Knippers R: **Molekulare Genetik**. Georg Thieme Verlag, Stuttgart, 2006.

[20] NIH History: **Deciphering the Genetic Code**. 2014. Http://history.nih.gov/exhibits/nirenberg/glossary.htm.

[21] Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Jr DEB, Hieter P, Vogelstein B and Kinzler KW: **Characterization of the Yeast Transcriptome**. *Cell* 1997. 88(2):243–251.

[22] Alwine JC, Kemp DJ and Stark GR: **Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes**. *Proceedings of the National academy of Sciences of the United States of America* 1977. 74(12):5350–5354.

[23] Liang P and Pardee AB: **Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction**. *Science* 1992. 257(5072):967–971.

[24] Lennon GG and Lehrach H: **Hybridization analyses of arrayed cDNA libraries**. *Trends in Genetics* 1991. 7(10):314–317.

[25] Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF *et al.*: **Complementary DNA sequencing: expressed sequence tags and human genome project**. *Science* 1991. 252(5013):1651–1656.

[26] Velculescu VE, Zhang L, Vogelstein B, Kinzler KW *et al.*: **Serial analysis of gene expression**. *Science* 1995. 270(5235):484–487.

[27] National Institutes of Health: National Human Genome Research Institute: **Talking glossary of genetic terms**. 2014. Http://www.genome.gov/glossary/.

[28] Thallinger GG: **Data Management and Applications for Biomolecular Arrays**. Ph.D. thesis, Graz University of Technology, Institute for Genomics and Bioinformatics, 2007.

[29] Duggan DJ, Bittner M, Chen Y, Meltzer P and Trent JM: **Expression profiling using cDNA microarrays**. *Nature Genetics* 1999. 21:10–14.

[30] Brown PO and Botstein D: **Exploring the new world of the genome with DNA microarrays**. *Nature Genetics* 1999. 21:33–37.

[31] Hacia JG: **Resequencing and mutational analysis using oligonucleotide microarrays**. *Nature Genetics* 1999. 21:42–47.

[32] Debouck C and Goodfellow PN: **DNA microarrays in drug discovery and development**. *Nature Genetics* 1999. 21:48–50.

[33] Benoit GR, Tong JH, Balajthy Z and Lanotte M: **Exploring (novel) gene expression during retinoid-induced maturation and cell death of acute promyelocytic leukemia**. In **Seminars in Hematology**, volume 38. Elsevier, 2001 pages 71–85.

[34] Triche TJ, Schofield D and Buckley J: **DNA microarrays in pediatric cancer**. *Cancer Journal* 2000. 7(1):2–15.

[35] Cooper CS: **Applications of microarray technology in breast cancer research**. *Breast Cancer Research* 2001. 3(3):158–175.

[36] Grouse LH, Munson PJ and Nelson PS: **Sequence databases and microarrays as tools for identifying prostate cancer biomarkers**. *Urology* 2001. 57(4):154–159.

[37] Soini H and Musser JM: **Molecular diagnosis of *Mycobacteria***. *Clinical Chemistry* 2001. 47(5):809–814.

[38] Diehn M and Relman Da: **Comparing functional genomic datasets: lessons from DNA microarray analyses of host–pathogen interactions**. *Current Opinion in Microbiology* 2001. 4(1):95–101.

[39] Wang Z, Gerstein M and Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature Reviews Genetics* 2009. 10(1):57–63.

[40] Okoniewski MJ and Miller CJ: **Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations**. *BMC Bioinformatics* 2006. 7(1):276–290.

[41] Shendure J: **The beginning of the end for microarrays?** *Nature methods* 2008. 5(7):585–587.

[42] Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P *et al.*: **The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC)**. *Genome Research* 2004. 14(10B):2121–2127.

[43] Boguski MS, Tolstoshev CM, Bassett Jr DE *et al.*: **Gene discovery in dbEST**. *Science* 1994. 265(5181):1993–1994.

[44] Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T *et al.*: **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.** *Proceedings of the National Academy of Sciences of the United States of America* 2003. 100(26):15776–15781.

[45] Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M *et al.*: **CAGE: cap analysis of gene expression.** *Nature methods* 2006. 3(3):211–222.

[46] Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, Mccurdy S, Foy M, Ewan M *et al.*: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays**. *Nature Biotechnology* 2000. 18(6):630–634.

[47] Grada A and Weinbrecht K: **Next-Generation Sequencing: Methodology and Application**. *Journal of Investigative Dermatology* 2013. 133(8):e11.

[48] Mardis ER: **Next-Generation Sequencing Platforms**. *Annual Review of Analytical Chemistry* 2013. 6:287–303.

[49] Bennett ST, Barnes C, Cox A, Davies L and Brown C: **Toward the $ 1000 human genome**. *Phamacogenomics* 2005. 6(4):373–382.

[50] Cloonan N, Forrest ARR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G *et al.*: **Stem cell transcriptome profiling via massive-scale mRNA sequencing**. *Nature Methods* 2008. 5(7):613–619.

[51] Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben L, Berka J, Braverman MS, Chen YJ, Chen Z *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005. 437(7057):376–380.

[52] Shendure J and Ji H: **Next-generation DNA sequencing.** *Nature Biotechnology* 2008. 26(10):1135–1145.

[53] Reis-Filho JS: **Next-generation sequencing.** *Breast Cancer Research* 2009. 11(Suppl 3):S12.

[54] Pettersson E, Lundeberg J and Ahmadian A: **Generations of sequencing technologies**. *Genomics* 2009. 93(2):105–111.

[55] Hawkins RD, Hon GC and Ren B: **Next-generation genomics: an integrative approach.** *Nature Reviews Genetics* 2010. 11(7):476–486.

[56] Metzker ML: **Sequencing technologies - the next generation**. *Nature Reviews Genetics* 2010. 11(1):31–46.

[57] Mortazavi A, Williams B, McCue K, Schaeffer L and Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nature Methods* 2008. 5(7):621–628.

[58] Marioni JC, Mason CE, Mane SM, Stephens M and Gilad Y: **RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Research* 2008. 18(9):1509–1517.

[59] Van Verk MC, Hickman R, Pieterse CMJ and Van Wees SCM: **RNA-Seq: revelation of the messengers**. *Trends in Plant Science* 2013. 18(4):175–179.

[60] Finotello F and Di Camillo B: **Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis**. *Briefings in Functional Genomics* 2014. 14(2):130–142.

[61] Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I and Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing**. *Molecular Ecology* 2008. 17(7):1636–1647.

[62] Crawford JE, Guelbeogo WM, Sanou A, Traoré A, Vernick KD, Sagnon N and Lazzaro BP: **De novo transcriptome sequencing in *Anopheles funestus* using illumina rna-seq technology**. *PloS One* 2010. 5(12):e14202.

[63] Roberts A, Pimentel H, Trapnell C and Pachter L: **Identification of novel transcripts in annotated genomes using RNA-Seq**. *Bioinformatics* 2011. 27(17):2325–2329.

[64] Peng Z, Cheng Y, Tan BCM, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X *et al.*: **Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome**. *Nature Biotechnology* 2012. 30(3):253–260.

[65] Bahn JH, Lee JH, Li G, Greer C, Peng G and Xiao X: **Accurate identification of A-to-I RNA editing in human by transcriptome sequencing**. *Genome Research* 2012. 22(1):142–145.

[66] Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, Leng J, Bjornson R, Kong Y, Kitabayashi N *et al.*: **AlleleSeq: analysis of allele-specific expression and binding in a network framework**. *Molecular Systems Biology* 2011. 7(1):522.

[67] Graves PR and Haystead TJ: **Molecular Biologist's Guide to Proteomics**. *Microbiology and Molecular Biology Reviews* 2002. 66(1):39–63.

[68] Anderson NG and Anderson NL: **Twenty years of two-dimensional electrophoresis: Past, present and future**. *Electophoresis* 2000. 17(3):443–453.

[69] Wasinger VC, Cordwell SJ, Cerpa-poljak A, Gooley AA, Wilkins MR, Duncan MW, Williams KL and Humphery-smith I: **Progress with gene-product mapping of the mollicutes: *Mycoplasma genitalium***. *Electophoresis* 1995. 16:1090–1094.

[70] Wilkins MR, Sanchez JC, Gooley Aa, Appel RD, Humphery-Smith I, Hochstrasser DF and Williams KL: **Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It**. *Biotechnology and Genetic Engineering Reviews* 1996. 13(1):19–50.

[71] O'Farrell PH: **High Resolution Two-Dimensional Electrophoresis of Proteins**. *The Journal of Biological Chemistry* 1975. 250(10):4007–4021.

[72] Klose J: **Protein Mapping by Combined Isoelectrie Focusing and Electrophoresis of Mouse Tissues**. *Humangenetik* 1975. 26(3):231–243.

[73] Scheele GA: **Two-Dimensional Gel Analysis of Soluble Proteins. Chracterization of Guinea Pig Exocrine Pancreatic Proteins**. *The Journal of Biological Chemistry* 1975. 250(14):5375–5385.

[74] Edman P: **Method for Determination of the Amino Acid Sequence in Peptides**. *Acta Chemica Scandinavica* 1950. 4(7):283–293.

[75] Aebersold RH, Teplow B, Hood LE and Kent BH: **Electroblotting onto Activated Glass**. *The Journal of Biological Chemistry* 1986. 261(9):4229–4238.

[76] Aebersold RH, Leavitt J, Saavedra RA, Hood LE and Kent S: **Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after in situ protease digestion on nitrocellulose**. *The Journal of Biological Chemistry* 1987. 84(20):6970–6974.

[77] Aebersold RH, Pipes G, Hood LE and Kent SB: **N-terminal and internal sequence determination of microgram amounts of proteins separated by isoelectric focusing in immobilized pH gradients**. *Electrophoresis* 1988. 9(9):520–530.

[78] Andersen JS and Mann M: **Functional genomics by mass spectrometry**. *FEBS Letters* 2000. 480(1):25–31.

[79] Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FCP, Weissman JS and Krogan NJ: **Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae***. *Molecular & Cellular Proteomics* 2007. 6(3):439–450.

[80] Au CE, Bell AW, Gilchrist A, Hiding J, Nilsson T and Bergeron JJ: **Organellar proteomics to create the cell map**. *Current Opinion in Cell Biology* 2007. 19(4):376–385.

[81] Yates JR, Gilchrist A, Howell KE and Bergeron JJM: **Proteomics of organelles and large cellular structures**. *Nature Reviews Molecular Cell Biology* 2005. 6(9):702–714.

[82] Ceciliani F, Eckersall D, Burchmore R and Lecchi C: **Proteomics in veterinary medicine: applications and trends in disease pathogenesis and diagnostics.** *Veterinary Pathology* 2014. 51(2):351–362.

[83] Hirosawa M, Hoshida M, Ishikawa M and Toya T: **MASCOT: multiple alignment system for protein sequences based on three-way dynamic programming**. *Computer Applications in the Biosciences* 1993. 9(2):161–167.

[84] Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S *et al.*: **Multiplexed Protein Quantitation in *Saccharomyces cerevisiae* Using Amine-reactive Isobaric Tagging Reagents**. *Molecular & Cellular Proteomics* 2004. 3(12):1154–1169.

[85] Ong S: **Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics**. *Molecular & Cellular Proteomics* 2002. 1(5):376–386.

[86] Corthals GL, Wasinger VC, Hochstrasser DF and Sanchez JC: **The dynamic range of protein expression : A challenge for proteomic research Proteomics and 2-DE**. *Electrophoresis* 2010. 21(6):1104–1115.

[87] Paik YK, Jeong SK, Omenn GS, Uhlen M, Hanash S, Cho SY, Lee HJ, Na K, Choi EY, Yan F *et al.*: **The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome**. *Nature Biotechnology* 2012. 30(3):221–223.

[88] Aebersold R, Bader GD, Edwards AM, van Eyk JE, Kussmann M, Qin J and Omenn GS: **The Biology/Disease-driven Human Proteome Project (B/D-HPP): Enabling Protein Research for the Life Sciences Community**. *Journal of Proteome Research* 2013. 12(1):23–27.

[89] Vaidyanathan G: **Redefining clinical trials: the age of personalized medicine.** *Cell* 2012. 148(6):1079–1080.

[90] Lakhan SE: **Schizophrenia proteomics: biomarkers on the path to laboratory medicine?** *Diagnostic Pathology* 2006. 1(11):B113.

[91] Li J, Zhang Z, Rosenzweig J, Wang YY and Chan DW: **Proteomics and Bioinformatics Approaches for Identification of Serum Biomarkers to Detect Breast Cancer**. *Clinical Chemistry* 2002. 48(8):1296–1304.

[92] Gupta N, Benhamida J, Bhargava V, Goodman D, Kain E, Kerman I, Nguyen N, Ollikainen N, Rodriguez J, Wang J *et al.*: **Comparative proteogenomics: Combining mass spectrometry and comparative genomics to analyze multiple genomes**. *Genome Research* 2008. 18(7):1133–1142.

[93] Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S *et al.*: **A draft map of the human proteome**. *Nature* 2014. 509(7502):575–581.

[94] Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H *et al.*: **Mass-spectrometry-based draft of the human proteome.** *Nature* 2014. 509(7502):582–587.

[95] Joyce AR and Palsson BO: **The model organism as a system: integrating 'omics' data sets**. *Nature Reviews Molecular Cell Biology* 2006. 7(3):198–210.

[96] Kohane IS, Butte AJ and Kho A: **Microarrays for an integrative genomics**. MIT press, Cambridge, 2002.

[97] Choi H and Pavelka N: **When one and one gives more than two: challenges and opportunities of integrative omics**. *Frontiers in Genetics* 2011. 2:105.

[98] Wang KS and Liu X: **Integrative Analysis of Genome-wide Expression and Methylation Data**. *Journal of Biomedical Biostatistics* 2013. 4:4–6.

[99] Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M and Shen R: **Pattern discovery and cancer gene identification in integrated cancer genomic data**. *Proceedings of the National Academy of Sciences of the United States of America* 2013. 110(11):4245–4250.

[100] Kockmann T, Gerstung M, Schlumpf T, Xhinzhou Z, Hess D, Beerenwinkel N, Beisel C and Paro R: **The BET protein FSH functionally interacts with ASH1 to orchestrate global gene activity in *Drosophila***. *Genome Biololgy* 2013. 14:R18.

[101] Chen Z and Zhang W: **Integrative Analysis Using Module-Guided Random Forests Reveals Correlated Genetic Factors Related to Mouse Weight**. *PLoS Computational Biology* 2013. 9(3):e1002956.

[102] Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R *et al.*: **Architecture of the human regulatory network derived from ENCODE data**. *Nature* 2012. 489(7414):91–100.

[103] Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder A, Carucci DJ *et al.*: **Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle**. *Genome Research* 2004. 14:2308–2318.

[104] Cox B, Kislinger T and Emili A: **Integrating gene and protein expression data: pattern analysis and profile mining**. *Methods* 2005. 35(3):303–314.

[105] Fagan A, Culhane AC and Higgins DG: **A multivariate analysis approach to the integration of proteomic and gene expression data**. *Proteomics* 2007. 7(13):2162–2171.

[106] 1000 Genomes Project Consortium and others: **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012. 491(7422):56–65.

[107] Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M *et al.*: **Systematic variation in gene expression patterns in human cancer cell lines**. *Nature Genetics* 2000. 24(3):227–235.

[108] Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ *et al.*: **Integrative analysis of 111 reference human epigenomes**. *Nature* 2015. 518(7539):317–330.

[109] Cagney G, Park S, Chung C, Tong B, Dushlaine CO, Shields DC and Emili A: **Human Tissue Profiling with Multidimensional Protein Identification Technology**. *Journal of Proteome Research* 2005. 4(5):1757–1767.

[110] Corbin RW, Paliy O, Yang F, Shabanowitz J, Platt M, Lyons CE, Root K, McAuliffe J, Jordan MI, Kustu S *et al.*: **Toward a protein profile of Escherichia coli: comparison to its transcription profile**. *Proceedings of the National Academy of Sciences of the United States of America* 2003. 100(16):9232–9237.

[111] Chen Y, Juan H, Huang H, Huang H, Lee Y, Liao M, Tseng C, Lin L, Chen J, Wang M *et al.*: **Quantitative Proteomic and Genomic Profiling Reveals Metastasis-Related Protein Expressio Patterns in Gastric Cancer Cells research articles**. *Journal of Proteome Research* 2006. 5(10):2727–2742.

[112] Griffin TJ: **Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae***. *Molecular & Cell Proteomics* 2002. 1(4):323–333.

[113] Mootha VK, Bunkenborg J, Olsen JV, Hjerrild M, Wisniewski JR, Stahl E, Bolouri MS, Ray HN, Sihag S, Kamal M *et al.*: **Integrated Analysis of Protein Composition, Tissue Diversity, and Gene Regulation in Mouse Mitochondria**. *Cell* 2003. 115(5):629–640.

[114] Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, Deciu C, Winzeler E and Yates JR: **Protein pathway and complex clustering of correlated mrna and protein expression analyses in *Saccharomyces cerevisiae***. *Proceedings of the National Academy of Sciences of the United States of America* 2003. 100(6):3107–3112.

[115] Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT *et al.*: **Global Survey of Organ and Organelle Protein Expression in Mouse: Combined Proteomic and Transcriptomic Profiling**. *Cell* 2006. 125(1):173–186.

[116] Nie L, Wu G, Brockman FJ and Zhang W: **Integrated analysis of transcriptomic and proteomic data of *Desulfovibrio vulgaris*: zero-inflated poisson regression models to predict abundance of undetected proteins**. *Bioinformatics* 2006. 22(13):1641–1647.

[117] Haider S and Pal R: **Integrated Analysis of Transcriptomic and Proteomic Data**. *Current Genomics* 2013. 14(2):91–110.

[118] Piruzian E, Bruskin S, Ishkin A, Abdeev R, Moshkovskii S, Melnik S, Nikolsky Y and Nikolskaya T: **Integrated network analysis of transcriptomic and proteomic data in psoriasis**. *BMC Systems Biology* 2010. 4(1):41–53.

[119] Tanay A, Sharan R, Kupiec M and Shamir R: **Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data**. *Proceedings of the National academy of Sciences of the United States of America* 2004. 101(9):2981–2986.

[120] Patil KR and Nielsen J: **Uncovering transcriptional regulation of metabolism by using metabolic network topology**. *Proceedings of the National academy of Sciences of the United States of America* 2005. 102(8):2685–2689.

[121] Perco P, Mühlberger I, Mayer G, Oberbauer R, Lukas A and Mayer B: **Linking transcriptomics and proteomic data on the level of protein interaction networks**. *BMC Systems Biololgy* 2010. 31(11):1780–1789.

[122] Glass K, Huttenhower C, Quackenbush J and Yuan GC: **Passing messages between biological networks to refine predicted interactions**. *PloS One* 2013. 8(5):e64832.

[123] Hahne H, Mäder U, Otto A, Bonn F, Steil L, Bremer E, Hecker M and Becher D: **A comprehensive proteomics and transcriptomics analysis of *Bacillus subtilis* salt stress adaptation**. *Journal of Bacteriology* 2010. 192(3):870–882.

[124] Verhoef S, Ballerstedt H, Volkers RJM, de Winde JH and Ruijssenaars HJ: **Comparative transcriptomics and proteomics of p-hydroxybenzoate producing *Pseudomonas putida* s12: novel responses and implications for strain improvement**. *Applied Microbiology and Biotechnology* 2010. 87(2):679–690.

[125] Gusenleitner D, Howe EA, Bentink S, Quackenbush J and Culhane AC: **iBBiG: iterative binary bi-clustering of gene sets**. *Bioinformatics* 2012. 28(19):2484–2492.

[126] Kaiser S: **Biclustering: Methods, Software and Application**. Ph.D. thesis, Ludwig-Maximilians-Universität München, Institut für Statistik, 2011.

[127] Alter O, Brown PO and Botstein D: **Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms**. *Proceedings of the National Academy of Sciences of the United States of America* 2003. 100(6):3351–3356.

[128] Ponnapalli SP, Saunders M, Van Loan CF and Alter O: **A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms**. *PloS One* 2011. 6(12):e28072.

[129] Zhang S, Liu CC, Li W, Shen H, Laird PW and Zhou XJ: **Discovery of multi-dimensional modules by integrative analysis of cancer genomic data**. *Nucleic Acids Research* 2012. 40:9379–9391.

[130] Culhane AC, Thioulouse J, Perrière G and Higgins DG: **MADE4: an R package for multivariate analysis of gene expression data**. *Bioinformatics* 2005. 21(11):2789–2790.

[131] Fagan A, Culhane A and Higgins D: **A multivariate analysis approach to the integration of proteomic and gene expression data**. *Proteomics* 2007. 7(13):2162–2171.

[132] De la Cruz O and Holmes SP: **The duality diagram in data analysis: Examples of modern applications**. *Annals of Applied Statistics* 2011. 5(4):2266–2277.

[133] Holmes S: **Multivariate data analysis: the French way**. In **Probability and statistics: Essays in honor of David A. Freedman**, pages 219–233. Institute of Mathematical Statistics, 2008.

[134] Jolliffe IT: **Principal Component Analysis**. Springer-Verlag, New York, Berlin, Heidelberg, 2002.

[135] Beh EJ and Lombardo R: **Correspondence Analysis: Theory, Practice and New Strategies**. John Wiley & Sons, Chichester, United Kingdom, 2014.

[136] Witten D, Tibshirani R and Hastie T: **A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis**. *Biostatistics* 2009. 10(3):515–534.

[137] Meng C, Kuster B, Culhane AC and Moghaddas Gholami A: **A multivariate approach to the integration of multi-omics datasets**. *BMC Bioinformatics* 2014. 15(1):162–75.

[138] Lê Cao KA, Rossouw D, Robert-Granié C and Besse P: **A sparse PLS for variable selection when integrating omics data**. *Statistical Applications in Genetics and Molecular Biology* 2008. 7(1):Article 35.

[139] Kohl M, Megger Da, Trippler M, Meckel H, Ahrens M, Bracht T, Weber F, Hoffmann AC, Baba Ha, Sitek B *et al.*: **A practical data processing workflow for multi-OMICS projects**. *Biochimica et Biophysica Acta - Proteins and Proteomics* 2014. 1844(1):52–62.

[140] Reif DM, Motsinger AA, McKinney BA, Crowe JE and Moore JH: **Feature selection using a random forests classifier for the integrated analysis of multiple data types**. In **Computational Intelligence and Bioinformatics and Computational Biology**. 2006 pages 1–8.

[141] Sass S, Buettner F, Mueller NS and Theis FJ: **A modular framework for gene set analysis integrating multilevel omics data**. *Nucleic Acids Research* 2013. 41(21):9622–9633.

[142] Arisi I, D'Onofrio M, Brandi R, Felsani A, Capsoni S, Drovandi G, Felici G, Weitschek E, Bertolazzi P and Cattaneo A: **Gene expression biomarkers in the brain of a mouse model for Alzheimer's disease: Mining of microarray data by logic classification and feature selection**. *Journal of Alzheimer's Disease* 2011. 24(4):721–738.

[143] Vaske C, Benz S, Sanborn J, Earl D, Szeto C, Zhu J, Haussler D and Stuart J: **Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM**. *Bioinformatics* 2010. 26(12):237–245.

[144] Sedgewick A, Benz S, Rabizadeh S, Soon-Shiong P and Vaske C: **Learning subgroup-specific regulatory interactions and regulator independence with PARADIGM**. *Bioinformatics* 2013. 29(13):62–70.

[145] Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim Js, Kim CJ, Kusanovic JP and Romero R: **A novel signaling pathway impact analysis**. *Bioinformatics* 2009. 25(1):75–82.

[146] Tomescu OA, Mattanovich D and Thallinger GG: **Integrative omics analysis. a study based on *Plasmodium falciparum* mrna and protein data**. *BMC Systems Biology* 2014. 8(Suppl 2):S4.

[147] Tomescu OA, Mattanovich D and Thallinger GG: **Integrative Analysis of Omics Data: A Method Comparison**. *Biomedical Engineering/Biomedizinische Technik* 2013. 58:Suppl 1.

[148] Dolèdec S and Chessel D: **Co-inertia analysis: an alternative method for studying species-environment relationships**. *Freshwater Biology* 1994. 31:277–294.

[149] Hanafi M and Chessel D: **Analyses de la co-inertie de K nuages de points**. *Revue de Statistique Appliquée* 1996. 2:35–60.

[150] Meng C, Tomescu OA, Thallinger GG, Gholami AM and Culhane AC: **Dimension reduction techniques for the integrative analysis of multi-omics data**, 2015. In review.

[151] Wall ME, Rechtsteiner A and Rocha LM: **Singular value decomposition and principal component analysis**. In **A practical approach to microarray data analysis**, pages 91–109. Springer, Berlin, Heidelberg, 2003.

[152] Borg I and Groenen P: **Modern multidimensional scaling: theory and applications**. *Journal of Educational Measurement* 2003. 40(3):277–280.

[153] Pearson K: **On lines and planes of closest fit to systems of points in space**. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 1901. 2(11):559–572.

[154] Hotelling H: **Analysis of a complex of statistical variables into principal components**. *Journal of Educational Psychology* 1933. 24(6):417.

[155] Greenacre M: **Theory and Applications of Correspondence Analysis**. Academic Press, London, United Kingdom, 1983.

[156] Gimaret-Carpentier C, Chessel D and Pascal J: **Non-symmetric correspondence analysis: an alternative for species occurrences data**. *Plant Ecology* 1998. 138(1):97–112.

[157] Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD and Vingron M: **Correspondence analysis applied to microarray data**. *Proceedings of the National academy of Sciences of the United States of America* 2001. 98(19):10781–10786.

[158] Greenacre M: **Correspondence analysis in practice**. CRC Press, Boca Raton, FL, USA, 2007.

[159] Benzècri JP: **L'Analyse des Données. Volume II. L'Analyse des Correspondances**. Dunod, Paris, France, 1973.

[160] Lauro CN and D'ambra L: **L'analyse non symétrique des correspondances**. *Data Analysis and Informatics* 1984. 3:433–446.

[161] Escoufier Y: **The duality diagram: a means for better practical applications**. In **Develoments in Numerical Ecology**, pages 139–156. Springer, Berlin, Heidelberg, 1987.

[162] Robert P and Escoufier Y: **A unifying tool for linear multivariate statistical methods: the RV-coefficient**. *Applied Statistics* 1976. 25(3):257–265.

[163] Smilde AK, Kiers HA, Bijlsma S, Rubingh C and Van Erk M: **Matrix correlations for high-dimensional data: the modified RV-coefficient**. *Bioinformatics* 2009. 25(3):401–405.

[164] The Gene Ontology Consortium: **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000. 25(1):25–29.

[165] Croft D, O'Kelly G, G.and Wu, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G and Jassal B: **Reactome: a database of reactions, pathways and biological processes**. *Nucleic Acids Research* 2011. 39(Suppl 1):D691–D697.

[166] Busold CH, Winter S, Hauser N, Bauer A, Dippon J, Hoheisel JD and Fellenberg K: **Integration of GO annotations in Correspondence Analysis: facilitating the interpretation of microarray data**. *Bioinformatics* 2005. 21(10):2424–2429.

[167] Alter O, Brown PO and Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling**. *Proceedings of the National Academy of Sciences of the United States of America* 2000. 97(18):10101–10106.

[168] Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, Connell JXO, Zhu S, Fero M, Sherlock G, Pollack JR *et al.*: **Mechanisms of disease Molecular characterisation of soft tissue tumours: a gene expression study**. *The Lancet* 2002. 359(9314):1301–1307.

[169] Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL and Somogy R: **Large-scale temporal gene expression mapping of central nevous system development**. *Proceedings of the National Academy of Sciences of the United States of America* 1998. 95(1):334–339.

[170] Hilsenbeck SG, William E, Schiff R, Connell O, Hansen RK, Osborne K and Fuqua SAW: **Statistical Analysis of Array Expression Data as Applied to the Problem of Tamoxifen Resistance**. *Journal of the National Cancer Institute* 1999. 91(5):453–459.

[171] Golub GH and Van Loan CF: **Matrix Computation**. Johns Hopkins University Press, Baltimore and London, 1996.

[172] Hotelling H: **Relations between two sets of variates**. *Biometrika* 1936. pages 321–377.

[173] Paige CC and Saunders MA: **Towards a generalized singular value decomposition**. *SIAM Journal on Numerical Analysis* 1981. 18(3):398–405.

[174] Falcon S and Gentleman R: **Using GOstats to test gene lists for GO term association**. *Bioinformatics* 2007. 23(2):257–258.

[175] Lloyd S: **Least squares quantization in PCM**. *IEEE Transactions on Information Theory* 1982. 28(2):129–137.

[176] Maimon O and Rokach L: **Data mining and knowledge discovery handbook**, volume 2. Springer, New York, 2005.

[177] Hartigan JA: **Direct Clustering of a Data Matrix**. *Journal of the American Statistical Association* 1972. 67(337):123–129.

[178] Cheng Y and Church GM: **Biclustering of expression data.** In **Proceedings of the International Conference on Systems and Molecular Biology**, volume 8. 2000 pages 93–103.

[179] Getz G, Levine E and Domany E: **Coupled two-way clustering analysis of gene microarray data**. *Proceedings of the National academy of Sciences of the United States of America* 2000. 97(22):12079–12084.

[180] Ben-Dor A, Chor B, Karp R and Yukhini Z: **Descovering local structure in gene expression data: The order preserving submatrix problem**. *Journal of Computational Biology* 2003. 10(3-4):373–384.

[181] Murali T and Kasif S: **Extracting conserved gene expression motifs from gene expression data**. In **Pacific Symposium on Biocomputing**, volume 8. World Scientific, 2003 pages 77–88.

[182] Madeira SC and Oliveira AL: **Biclustering algorithms for biological data analysis: a survey**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2004. 1:24–45.

[183] Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L and Zitzler E: **A systematic comparison and evaluation of biclustering methods for gene expression data**. *Bioinformatics* 2006. 22(9):1122–1129.

[184] Kaiser S, Santamaria R, Tatsiana, Khamiakova, Sill M, Theron R, Quintales L and Leisch F: **biclust: BiCluster Algorithms**, 2013. R package version 1.0.2.

[185] Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y *et al.*: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature* 2012. 486(7403):346–352.

[186] Tanay A, Sharan R, Kupiec M and Shamir R: **Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data**. *Proceedings of the National academy of Sciences of the United States of America* 2004. 101:2981–2986.

[187] Smoot M, Ono K, Ruscheinski J, Wang PL and Ideker T: **Cytoscape 2.8: new features for data integration and network visualization**. *Bioinformatics* 2011. 27(3):431–432.

[188] Patz S, Trattnig C, Grünbacher G, Ebner B, Gülly C, Novak A, Rinner B, Leitinger G, Absenger M, Tomescu OA *et al.*: **More than cell dust: microparticles isolated from cerebrospinal fluid of brain injured patients are messengers carrying mRNAs, miRNAs, and proteins**. *Journal of Neurotrauma* 2013. 30(14):1232–1242.

[189] He L and Hannon GJ: **MicroRNAs: small RNAs with a big role in gene regulation**. *Nature Reviews Genetics* 2004. 5(7):522–531.

[190] Wienholds E and Plasterk RH: **MicroRNA function in animal development**. *FEBS letters* 2005. 579(26):5911–5922.

[191] Griffiths-Jones S, Saini HK, van Dongen S and Enright AJ: **miRBase: tools for microRNA genomics**. *Nucleic Acids Research* 2008. 36(Suppl 1):D154–D158.

[192] Ogata H, Goto S, Sato K, Fujibuchi W, Bono H and Kanehisa M: **KEGG: Kyoto encyclopedia of genes and genomes**. *Nucleic Acids Research* 1999. 27(1):29–34.

[193] Maere S, Heymans K and Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks**. *Bioinformatics* 2005. 21(16):3448–3449.

[194] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E *et al.*: **PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nature Genetics* 2003. 34(3):267–273.

[195] Subramanian A, Tamayo P, Mootha V, Mukherjee S, Ebert B, Gillette M, Paulovich A, Pomeroy S, Golub T and Lander E: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005. 102(43):15545–15550.

[196] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B and Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Research* 2003. 13(11):2498–2504.

[197] Huang DW, Sherman BT and Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists**. *Nucleic Acids Research* 2009. 37(1):1–13.

[198] Ge Y, Dudoit S and Speed TP: **Resampling-based multiple testing for microarray data analysis**. *Test* 2003. 12(1):1–77.

[199] Benjamini Y and Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing**. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995. 57(1):289–300.

[200] Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder A, Batalov S, Carucci DJ *et al.*: **Discovery of gene function by expression profiling of the malaria parasite life cycle**. *Science* 2003. 301(5639):1503–1508.

[201] Zhou Y and Abagyan R: **Match-only integral distribution (MOID) algorithm for high-density oligonucleotide array analysis**. *BMC Bioinformatics* 2002. 3(1):3.

[202] R Development Core Team: **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[203] Carlson M: **org.Pf.plasmo.db: Genome wide annotation for Malaria**, 2012. R package version 2.8.1.

[204] Carlson M: **GO.db: A set of annotation maps describing the entire Gene Ontology**, 2012. R package version 2.8.0.

[205] Smith RC, King JG, Tao D, Tomescu OA, Brando C, Thallinger GG and Dinglasan RR: **Proteomic analysis of mosquito macrophage-like blood cells reveals an anticipatory innate immune response in the absence of malaria parasite challenge**, 2015. In review.

[206] Pinto SB, Lombardo F, Koutsos AC, Waterhouse RM, McKay K, An C, Ramakrishnan C, Kafatos FC and Michel K: **Discovery of *Plasmodium* modulators by genome-wide analysis of circulating hemocytes in *Anopheles gambiae***. *Proceedings of the National Academy of Sciences of the United States of America* 2009. 106(50):21270–21275.

[207] Baton LA, Robertson A, Warr E, Strand MR and Dimopoulos G: **Genome-wide transcriptomic profiling of *Anopheles gambiae* hemocytes reveals pathogen-specific signatures upon bacterial challenge and *Plasmodium berghei* infection**. *BMC Genomics* 2009. 10(1):257.

[208] Edgar R, Domrachev M and Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Research* 2002. 30(1):207–210.

[209] Dweep H, Sticht C, Pandey P and Gretz N: **miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes**. *Journal of Biomedical Informatics* 2011. 44(5):839–847.

[210] Mok SC, Bonome T, Vathipadiekal V, Bell A, Johnson ME, Park DC, Hao K, Yip DK, Donninger H, Ozbun L *et al.*: **A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: microfibril-associated glycoprotein 2**. *Cancer Cell* 2009. 16(6):521–532.

[211] Leung CS, Yeung TL, Yip KP, Pradeep S, Balasubramanian L, Liu J, Wong KK, Mangala LS, Armaiz-Pena GN, Lopez-Berestein G *et al.*: **Calcium-dependent FAK/CREB/TNNC1 signalling mediates the effect of stromal MFAP5 on ovarian cancer metastatic potential**. *Nature Communications* 2014. 5.

[212] Yeung TL, Leung CS, Wong KK, Samimi G, Thompson MS, Liu J, Zaid TM, Ghosh S, Birrer MJ and Mok SC: **TGF-β modulates ovarian cancer invasion by upregulating CAF-derived versican in the tumor microenvironment**. *Cancer Research* 2013. 73(16):5016–5028.

[213] Li C and Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection**. *Proceedings of the National academy of Sciences of the United States of America* 2001. 98(1):31–36.

[214] Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, Chen W, Xie M, Wang W, Hammond S *et al.*: **The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line**. *Nature Biotechnology* 2011. 29(8):735–741.

[215] Ritchie M, Phipson B, Wu D, Hu Y, Law C, Shi W and Smyth G: **limma powers differential expression analyses for RNA-sequencing and microarray studies**. *Nucleic Acids Research* 2015. 43(7):e47.

[216] Dudoit S, Shaffer JP and Boldrick JC: **Multiple hypothesis testing in microarray experiments**. *Statistical Science* 2003. 18(1):71–103.

[217] Breitwieser FP, Müller A, Dayon L, Köcher T, Hainard A, Pichler P, Schmidt-Erfurth U, Superti-Furga G, Sanchez JC, Mechtler K *et al.*: **General statistical modeling of data from protein relative expression isobaric tags**. *Journal of Proteome Research* 2011. 10(6):2758–2766.

[218] Prifti E, Zucker J, Clement K and Henegar C: **FunNet: an integrative tool for exploring transcriptional interactions**. *Bioinformatics* 2008. 24(22):2636–2638.

[219] Restrepo J, Ott E and Hunt B: **Characterizing the Dynamical Importance of Network Nodes and Links**. *Physical Review Letters* 2006. 97(9):094102.

[220] Sales G, Calura E and Romualdi C: **graphite: GRAPH Interaction from pathway Topological Environment**, 2014. R package version 1.10.1.

[221] Sales G, Calura E, Martini P and Romualdi C: **Graphite Web: Web tool for gene set analysis exploiting pathway topology**. *Nucleic Acids Research* 2013. 41(W1):W89–W97.

[222] Broad Institute, Boston, MA, USA: **TARGET DB**. online resource, 2015. www.broadinstitute.org/cancer/cga/target.

[223] Sanders PR, Gilson PR, Cantin GT, Greenbaum DC, Nebl T, Carucci DJ, McConville MJ, Schofield L, Hodder AN, Yates JR *et al.*: **Distinct protein classes including novel merozoite surface antigens in raft-like membranes of *Plasmodium falciparum***. *Journal of Biologocal Chemistry* 2005. 280(48):40169–40176.

[224] Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL *et al.*: **A proteomic view of the *Plasmodium falciparum* life cycle**. *Nature* 2002. 419(6906):520–526.

[225] Llinás M, Bozdech Z, Wong ED, Adai AT and DeRisi JL: **Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains**. *Nucleic Acids Research* 2006. 34(4):1166–1173.

[226] Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J and DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum***. *PLoS Biology* 2003. 1(1):85–100.

[227] LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C *et al.*: **A protein interaction network of the malaria parasite *Plasmodium falciparum***. *Nature* 2005. 438(7064):103–107.

[228] Winzeler EA: **Malaria research in the post-genomic era**. *Nature* 2008. 455(7214):751–756.

[229] Mikkelsen RB, Kamber M, Wadwa KS, Lin PS and Schmidt-Ullrich R: **The role of lipids in *Plasmodium falciparum* invasion of erythrocytes: a coordinated biochemical and microscopic analysis**. *Proceedings of the National academy of Sciences of the United States of America* 1988. 85(16):5956–5960.

[230] Ward G, Fujioka H, Aikawa M and Miller L: **Staurosporine Inhibits Invasion of Erythrocytes by Malarial Merozoites**. *Experimental Parasitology* 1994. 79(3):480–487.

[231] Bozdech Z and Ginsburg H: **Data mining of the transcriptome of *Plasmodium falciparum*: the pentose phosphate pathway and ancillary processes**. *Malaria Journal* 2005. 4(1):17.

[232] Mok S, Imwong M, Mackinnon MJ, Sim J, Ramadoss R, Yi P, Mayxay M, Chotivanich K, Liong KY, Russell B *et al.*: **Artemisinin resistance in *Plasmodium falciparum* is associated with an altered temporal pattern of transcription**. *BMC Genomics* 2011. 12(1):391.

[233] Roth EJ: ***Plasmodium falciparum* carbohydrate metabolism: a connection between host cell and parasite**. *Blood Cells* 1990. 16(2-3):453–466.

[234] Mahoney JA, Ntolosi B, DaSilva RP, Gordon S and McKnight AJ: **Cloning and characterization of CPVL, a novel serine carboxypeptidase, from human macrophages**. *Genomics* 2001. 72(3):243–251.

[235] Yassine H, Kamareddine L, Chamat S, Christophides GK and Osta MA: **A serine protease homolog negatively regulates tep1 consumption in systemic infections of the malaria vector *Anopheles gambiae***. *Journal of Innate Immunity* 2013. 6(6):806–818.

[236] Lemaitre B and Hoffmann J: **The host defense of *Drosophila melanogaster***. *Annual Review of Immunology* 2007. 25:697–743.

[237] Yang JY, Yoshihara K, Tanaka K, Hatae M, Masuzaki H, Itamochi H, Takano M, Ushijima K, Tanyi JL, Coukos G *et al.*: **Predicting time to ovarian carcinoma recurrence using protein markers**. *The Journal of Clinical Investigation* 2013. 123(9):3740.

[238] Bosquet JG, Marchion DC, Chon H, Lancaster JM and Chanock S: **Analysis of Chemotherapeutic Response in Ovarian Cancers Using Publicly Available High-Throughput Data**. *Cancer Research* 2014. 74(14):3902–3912.

[239] Bentink S, Haibe-Kains B, Risch T, Fan JB, Hirsch MS, Holton K, Rubio R, April C, Chen J, Wickham-Garcia E *et al.*: **Angiogenic mRNA and microRNA gene expression signature predicts a novel subtype of serous ovarian cancer**. *PloS One* 2012. 7(2):e30269.

[240] Cancer Genome Atlas Research Network: **Integrated genomic analyses of ovarian carcinoma**. *Nature* 2011. 474(7353):609–615.

[241] Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, Johnson DS, Trivett MK, Etemadmoghadam D, Locandro B *et al.*: **Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome**. *Clinical Cancer Research* 2008. 14(16):5198–5208.

[242] Verhaak RG, Tamayo P, Yang JY, Hubbard D, Zhang H, Creighton CJ, Fereday S, Lawrence M, Carter SL, Mermel CH *et al.*: **Prognostically relevant gene signatures of high-grade serous ovarian carcinoma**. *The Journal of Clinical Investigation* 2013. 123(1):517.

[243] Sfakianos GP, Iversen ES, Whitaker R, Akushevich L, Schildkraut JM, Murphy SK, Marks JR and Berchuck A: **Validation of ovarian cancer gene expression signatures for survival and subtype in formalin fixed paraffin embedded tissues**. *Gynecologic Oncology* 2013. 129(1):159–164.

[244] Hsu YH, Zillikens MC, Wilson SG, Farber CR, Demissie S, Soranzo N, Bianchi EN, Grundberg E, Liang L, Richards JB *et al.*: **An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits**. *PLoS Genetics* 2010. 6(6):e1000977.

[245] Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA *et al.*: **Inferring tumour purity and stromal and immune cell admixture from expression data**. *Nature Communications* 2013. 4.

[246] Waldron L, Haibe-Kains B, Culhane AC, Riester M, Ding J, Wang XV, Ahmadifar M, Tyekucheva S, Bernau C, Risch T *et al.*: **Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer**. *Journal of the National Cancer Institute* 2014. 106(5):dju049.

[247] Casey SC, Amedei A, Aquilano K, Azmi AS, Benencia F, Bhakta D, Bilsland AE, Boosani CS, Chen S, Ciriolo MR *et al.*: **Cancer prevention and therapy through the modulation of the tumor microenvironment**. In **Seminars in cancer biology**. Elsevier, 2015 page Epub ahead of print.

[248] Morikawa T, Kuchiba A, Yamauchi M, Meyerhardt JA, Shima K, Nosho K, Chan AT, Giovannucci E, Fuchs CS and Ogino S: **Association of CTNNB1 (β-catenin) alterations, body mass index, and physical activity with survival in patients with colorectal cancer**. *The Journal of the American Medical Association* 2011. 305(16):1685–1694.

[249] Cho JY, Lim JY, Cheong JH, Park YY, Yoon SL, Kim SM, Kim SB, Kim H, Hong SW, Park YN *et al.*: **Gene expression signature–based prognostic risk score in gastric cancer**. *Clinical Cancer Research* 2011. 17(7):1850–1857.

[250] Bilancia D, Rosati G, Dinota A, Germano D, Romano R and Manzione L: **Lapatinib in breast cancer**. *Annals of Oncology* 2007. 18(suppl 6):vi26–vi30.

[251] Yarden Y and Sliwkowski MX: **Untangling the ErbB signalling network**. *Nature Reviews Molecular Cell Biology* 2001. 2(2):127–137.

[252] Howe JR, Roth S, Ringold JC, Summers RW, Järvinen HJ, Sistonen P, Tomlinson IP, Houlston RS, Bevan S, Mitros FA *et al.*: **Mutations in the SMAD4/DPC4 gene in juvenile polyposis**. *Science* 1998. 280(5366):1086–1088.

[253] Crowder RJ, Phommaly C, Tao Y, Hoog J, Luo J, Perou CM, Parker JS, Miller MA, Huntsman DG, Lin L *et al.*: **PIK3CA and PIK3CB inhibition produce synthetic lethality when combined with estrogen deprivation in estrogen receptor–positive breast cancer**. *Cancer Research* 2009. 69(9):3955–3962.

[254] Courtney KD, Corcoran RB and Engelman JA: **The PI3K pathway as drug target in human cancer**. *Journal of Clinical Oncology* 2010. 28(6):1075–1083.

[255] Gorka B, Skubis-Zegadło J, Mikula M, Bardadin K, Paliczka E and Czarnocka B: **NrCAM, a neuronal system cell-adhesion molecule, is induced in papillary thyroid carcinomas**. *British Journal of Cancer* 2007. 97(4):531–538.

[256] Aitkenhead M, Wang SJ, Nakatsu MN, Mestas J, Heard C and Hughes CC: **Identification of endothelial cell genes expressed in an in vitro model of angiogenesis: induction of ESM-1, βig-h3, and NrCAM**. *Microvascular Research* 2002. 63(2):159–171.

[257] Cavallaro U and Christofori G: **Cell adhesion and signalling by cadherins and Ig-CAMs in cancer**. *Nature Reviews Cancer* 2004. 4(2):118–132.

[258] D'Andrea AD and Grompe M: **The Fanconi anaemia/BRCA pathway**. *Nature Reviews Cancer* 2003. 3(1):23–34.

[259] Taron M, Rosell R, Felip E, Mendez P, Souglakos J, Ronco MS, Queralt C, Majo J, Sanchez JM, Sanchez JJ *et al.*: **BRCA1 mRNA expression levels as an indicator of chemoresistance in lung cancer**. *Human Molecular Genetics* 2004. 13(20):2443–2449.

[260] Helleday T, Petermann E, Lundin C, Hodgson B and Sharma RA: **DNA repair pathways as targets for cancer therapy**. *Nature Reviews Cancer* 2008. 8(3):193–204.

[261] Wang T, Liu G and Wang R: **The intercellular metabolic interplay between tumor and immune cells**. *Frontiers in Immunology* 2014. 5:358.

[262] Gilbert CA and Slingerland JM: **Cytokines, obesity, and cancer: new insights on mechanisms linking obesity to cancer risk and progression**. *Annual Reviews of Medicine* 2013. 64:45–57.

[263] Nieman KM, Kenny HA, Penicka CV, Ladanyi A, Buell-Gutbrod R, Zillhardt MR, Romero IL, Carey MS, Mills GB, Hotamisligil GS *et al.*: **Adipocytes promote ovarian cancer metastasis and provide energy for rapid tumor growth**. *Nature Medicine* 2011. 17(11):1498–1503.

[264] Walpole CM: **The function and mechanisms of action of ghrelin and obestatin in ovarian cancer**. Ph.D. thesis, Queensland University of Technology, 2012.

[265] Granata R, Gallo D, Luque RM, Baragli A, Scarlatti F, Grande C, Gesmundo I, Córdoba-Chacón J, Bergandi L, Settanni F *et al.*: **Obestatin regulates adipocyte function and protects against diet-induced insulin resistance and inflammation**. *The FASEB Journal* 2012. 26(8):3393–3411.

[266] Justus CR, Sanderlin EJ and Yang LV: **Molecular Connections between Cancer Cell Metabolism and the Tumor Microenvironment**. *International Journal of Molecular Sciences* 2015. 16(5):11055–11086.

[267] Dang C: **The interplay between MYC and HIF in the Warburg effect**. In **Oncogenes Meet Metabolism**, pages 35–53. Springer, Berlin, Heidelberg, 2008.

[268] Sun L, Zang WJ, Wang H, Zhao M, Yu XJ, He X, Miao Y and Zhou J: **Acetylcholine Promotes ROS Detoxification Against Hypoxia/reoxygenation-Induced Oxidative Stress Through FoxO3a/PGC-1α Dependent Superoxide Dismutase**. *Cellular Physiology and Biochemistry* 2014. 34(5):1614–1625.

[269] Potter C and Harris AL: **Hypoxia Inducible Carbonic Anhydrase IX, Marker of Tumour: Hypoxia, Survival Pathway and Therapy Target**. *Cell Cycle* 2004. 3(2):159–162.

[270] Kang MJ, Jung SM, Kim MJ, Bae JH, Kim HB, Kim JY, Park SJ, Song HS, Kim DW, Kang CD *et al.*: **DNA-dependent protein kinase is involved in heat shock protein-mediated accumulation of hypoxia-inducible factor-1α in hypoxic preconditioned HepG2 cells**. *FEBS Journal* 2008. 275(23):5969–5981.

[271] Schulz R, Streller F, Scheel A, Rüschoff J, Reinert M, Dobbelstein M, Marchenko N and Moll U: **HER2/ErbB2 activates HSF1 and thereby controls HSP90 clients including MIF in HER2-overexpressing breast cancer**. *Cell Death & Disease* 2014. 5(1):e980.

[272] Harvey KF, Zhang X and Thomas DM: **The Hippo pathway and human cancer**. *Nature Reviews Cancer* 2013. 13(4):246–257.

[273] Zhang X, George J, Deb S, Degoutin J, Takano E, Fox S, Bowtell D and Harvey K: **The Hippo pathway transcriptional co-activator, YAP, is an ovarian cancer oncogene**. *Oncogene* 2011. 30(25):2810–2822.

[274] Branzei D and Foiani M: **Regulation of DNA repair throughout the cell cycle**. *Nature Reviews Molecular Cell Biology* 2008. 9(4):297–308.

[275] Patel DJ, Phan AT and Kuryavyi V: **Human telomere, oncogenic promoter and 5UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics**. *Nucleic Acids Research* 2007. 35(22):7429–7455.

[276] Stetler-Stevenson W, Liotta L and Kleiner D: **Extracellular matrix 6: role of matrix metalloproteinases in tumor invasion and metastasis**. *The FASEB Journal* 1993. 7(15):1434–1441.

[277] Moser TL, Young TN, Rodriguez GC, Pizzo SV, Bast RC and Stack MS: **Secretion of extracellular matrix-degrading proteinases is increased in epithelial ovarian carcinoma**. *International Journal of Cancer* 1994. 56(4):552–559.

[278] Fishman DA, Bafetti LM, Banionis S, Kearns AS, Chilukuri K and Stack MS: **Production of extracellular matrix-degrading proteinases by primary cultures of human epithelial ovarian carcinoma cells**. *Cancer* 1997. 80(8):1457–1463.

[279] De Alvaro C, Teruel T, Hernandez R and Lorenzo M: **Tumor necrosis factor α produces insulin resistance in skeletal muscle by activation of inhibitor κB kinase in a p38 MAPK-dependent manner**. *Journal of Biological Chemistry* 2004. 279(17):17070–17078.

[280] Lejeune FJ, Rüegg C and Liénard D: **Clinical applications of TNF-α in cancer**. *Current Opinion in Immunology* 1998. 10(5):573–580.

[281] Ma XJ, Dahiya S, Richardson E, Erlander M and Sgroi DC: **Gene expression profiling of the tumor microenvironment during breast cancer progression**. *Breast Cancer Research* 2009. 11(1):R7.

[282] Takehara A, Hosokawa M, Eguchi H, Ohigashi H, Ishikawa O, Nakamura Y and Nakagawa H: **γ-aminobutyric acid (GABA) stimulates pancreatic cancer growth through overexpressing GABAA receptor π subunit**. *Cancer Research* 2007. 67(20):9704–9712.

[283] Ferrer-Miralles N, Domingo-Espín J, Corchero JL, Vázquez E and Villaverde A: **Microbial factories for recombinant pharmaceuticals**. *Microbial Cell Factories* 2009. 8(1):17.

[284] Maccani A, Landes N, Stadlmayr G, Maresch D, Leitner C, Maurer M, Gasser B, Ernst W, Kunert R and Mattanovich D: ***Pichia pastoris* secretes recombinant proteins less efficiently than Chinese hamster ovary cells but allows higher space-time yields for less complex proteins**. *Biotechnology Journal* 2014. 9(4):526–537.

[285] Porro D, Sauer M, Branduardi P and Mattanovich D: **Recombinant protein production in yeasts**. *Molecular Biotechnology* 2005. 31(3):245–259.

[286] Bretthauer RK: **Genetic engineering of *Pichia pastoris* to humanize n-glycosylation of proteins**. *Trends in Biotechnology* 2003. 21(11):459–462.

[287] Gasser B, Maurer M, Rautio J, Sauer M, Bhattacharyya A, Saloheimo M, Penttilä M and Mattanovich D: **Monitoring of transcriptional regulation in *Pichia pastoris* under protein production conditions**. *BMC Genomics* 2007. 8(1):179.

[288] Nathan DF, Vos MH and Lindquist S: **Identification of ssf1, cns1, and hch1 as multicopy suppressors of a *Saccharomyces cerevisiae* hsp90 loss-of-function mutation**. *Proceedings of the National Academy of Sciences of the United States of America* 1999. 96(4):1409–1414.

[289] Dietmair S, Hodson MP, Quek LE, Timmins NE, Gray P and Nielsen LK: **A multi-omics analysis of recombinant protein production in hek293 cells**. *PLoS One* 2012. 7(8):e43394.

[290] Fang M, Kearns BG, Gedvilaite A, Kagiwada S, Kearns M, Fung M and Bankaitis VA: **Kes1p shares homology with human oxysterol binding protein and participates in a novel regulatory pathway for yeast golgi-derived transport vesicle biogenesis.** *The EMBO Journal* 1996. 15(23):6447.

Appendices

Appendix A

Generalized Singular Value Decomposition

R-File containing the function used for the computation of the generalized singular value decomposition.

```
my_gsvd = function(yeast,human){
matrix = rbind(yeast,human)

# QR decomposition
qrv = qr(matrix)
Q = qr.Q(qrv)
R = qr.R(qrv)


genes1 = dim(yeast)[1]
arrays1 = dim(yeast)[2]

q1 = Q[1:genes1,]

# computation of the genelets with singular value decomposition
svdv = svd(q1)
genelets = t(svdv$v)%*%R

# normalization of the genelets
for(i in 1:dim(genelets)[1])
genelets[i,] = genelets[i,]/sqrt(genelets[i,]%*%genelets[i,])


# computation of the arraylets
arraylets1 = yeast%*%solve(genelets)
arraylets2 = human%*%solve(genelets)

# arraylets normalization
for(i in 1:dim(arraylets1)[1])
arraylets1[i,] = arraylets1[i,]/sqrt(arraylets1[i,]%*%arraylets1[i,])


for(i in 1:dim(arraylets2)[1])
```

```r
arraylets2[i,] = arraylets2[i,]/sqrt(arraylets2[i,]%*%arraylets2[i,])

# computation of the generalized eigenvalues
library("corpcor")
d1 = diag(pseudoinverse(arraylets1)%*%yeast%*%solve(genelets))
d2 = diag(pseudoinverse(arraylets2)%*%human%*%solve(genelets))

# setup of the result
res = list(arraylets1 = arraylets1,
           arraylets2 = arraylets2,
           d1 = d1,
           d2 = d2,
           genelets = genelets)

return(res)
}
```

Listing A.1: R script to compute the generalized singular value decomposition for two data sets

Appendix B

Analysis of Hemocyte and Granulocyte Immune Response of *Anopheles Gambiae*

Table B.1: Immnue-specific gene associations.

Genes	Annotation	Feature Axis 1 [AU]	Feature Axis 2 [AU]
AGAP005890	no metadata	-7,28	0,02
AGAP002134	no metadata	7,69	-1,48
AGAP012278	no metadata	-7,55	-1,5
AGAP008282	no metadata	8,36	-2,19
AGAP008696	no metadata	9,81	-2,51
AGAP000278	OBP9 odorant binding protein	-6,45	-0,73
AGAP006448	cAMP-dependent protein kinase regulator	-0,81	2,64
AGAP002082	no metadata	0,31	1,78
AGAP008843	aquaporin 1	5,71	-2,72
AGAP007771	no metadata	-0,74	1,74

Table B.2: Immnue-specific protein associations.

Proteins	Annotation	Feature Axis 1 [AU]	Feature Axis 2 [AU]
AGAP028028	leucine-rich immune protein	-2,38	0,51
AGAP007456	leucine-rich immune protein (Short)	1,23	0,46
AGAP011791	Clip-Domain Serine Protease	1,18	0,15
AGAP004148	CLIPB5, easter-like	-3,69	-0,22
AGAP007045	leucine-rich immune protein - LRIM15, insulin-like growth factor binding protein complex	-2,61	-0,74
AGAP013442	Clip-Domain Serine Protease	1,06	-1,07
AGAP005693	leucine-rich immune protein (Coil-less)	0,76	0,38
AGAP011787	CLIPA5	-1,47	0,51
AGAP011790	Clip-Domain Serine Protease, CLIPA2	-2,11	-0,61

Table B.3: Proliferation-specific gene associations.

Genes	Annotation	Feature Axis 1 [AU]	Feature Axis 2 [AU]
AGAP005890	no metadata	-7,15	-1,35
AGAP002134	no metadata	7,58	-0,25
AGAP012278	no metadata	-7,14	-2,9
AGAP008282	no metadata	8,42	-0,58
AGAP008696	no metadata	10,1	-0,61
AGAP000278	OBP9 odorant binding protein	-6,2	-1,93
AGAP006448	cAMP-dependent protein kinase regulator	-1,29	2,44
AGAP003249	CLIP3 Clip-domain serine protease	-2,78	-2,52
AGAP002082	no metadata	-0,03	1,8
AGAP003241	no metadata	0,98	1,88
AGAP010904	CPFL3 cuticular protein 3 from CPFL family	-5	-2,95

Table B.4: Proliferation-specific protein associations.

Proteins	Annotation	Feature Axis 1 [AU]	Feature Axis 2 [AU]
AGAP005160	Ras homolog gene family, member A	0,38	-0,16
AGAP004146	Ras-related protein Rab-1A	1,1	0,37
AGAP005393	Ras-related protein Rab-2A	-0,08	0,2
AGAP007901	Ras-related protein Rab-5C	-0,4	-0,45
AGAP001617	Ras-related protein Rab-7A	-2,07	0,38
AGAP001874	Ras-related protein Rap-1A	0,83	0,06
AGAP004559	Ras-related protein	-0,04	-0,37

Appendix C

Traditional Gene Set Enrichment Analysis for microRNAs

Table C.1: Overrepresented biological processes in the set of validated target genes of CSF-MP associated miRNAs and all miRNAs. Non-significance is marked by n.s.

	p values	
Overrepresented neuron related biological processes	CSF-MP miRNAs	all miRNAs
neurogenesis	1,3803E-16	1,3881E-16
generation of neurons	1,1113E-14	3,5503E-15
regulation of neuron apoptosis	3,1342E-11	9,0896E-10
regulation of neurogenesis	3,4301E-11	1,3565E-10
regulation of neuron differentiation	5,5376E-10	2,0373E-09
neuron differentiation	4,5311E-08	1,0827E-08
negative regulation of neuron apoptosis	9,2085E-07	3,2024E-06
neuron fate commitment	0,000010986	0,000092647
positive regulation of neurogenesis	0,000014836	0,000018025
positive regulation of neuron apoptosis	0,000015613	0,000078275
regulation of neurological system process	0,000018074	n.s.
positive regulation of neuron differentiation	0,000040784	0,000038682
neuron development	0,000056759	6,8055E-06
regulation of long-term neuronal synaptic plasticity	0,00013011	0,000078275
positive regulation of neurological system process	0,00015663	0,00079919
neuron projection development	0,00037089	0,000019934
negative regulation of neurogenesis	0,00037164	0,0024235
negative regulation of neuron projection development	0,00037367	0,0011074
neuron death	0,00042698	0,000027475
negative regulation of axonogenesis	0,0010188	0,000018025
response to axon injury	0,0013414	0,000038682
regulation of neuron projection development	0,0013688	0,0026152
central nervous system neuron differentiation	0,0013739	0,009209
neuron apoptosis	0,0013963	0,000078444
regulation of neuron projection regeneration	0,0014606	0,0031111
regulation of axon regeneration	0,0014606	0,0031111
negative regulation of neuron projection regeneration	0,0022784	0,0041026
positive regulation of neuroblast proliferation	0,0027254	0,0057133
neuron projection morphogenesis	0,0027974	0,00030567
negative regulation of neuron differentiation	0,0028146	0,0025622
negative regulation of neuroblast proliferation	0,0053363	0,0094649
negative regulation of neurotransmitter transport	0,0071839	0,010739
regulation of neuronal synaptic plasticity	0,0077517	0,0022518
neuron fate specification	0,0085902	n.s.
neuron projection regeneration	n.s.	0,0017688
cell morphogenesis involved in neuron differentiation	n.s.	0,0042632
axonogenesis	n.s.	0,0046638

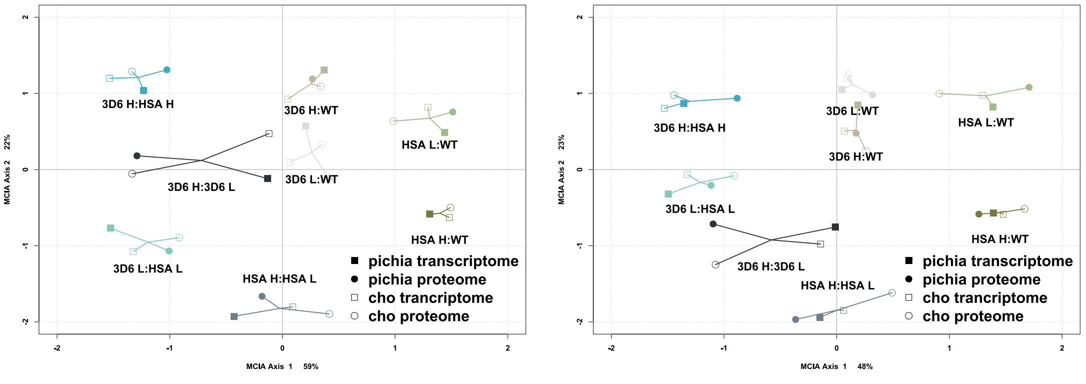
Appendix D

Cross-species Comparison of all Eight Conditions

D.1 Secretion and Ribosome Relevant Analysis

In this section, the results of MCIA performed on the secretion and ribosomes related features are shown. The ribosome relevant feature subsets include 495 *P. pastoris* genes, 288 *P. pastoris* proteins, 297 CHO genes and 53 CHO proteins while the secretion relevant subset included 405 *P. pastoris* genes, 236 *P. pastoris* proteins, 534 CHO genes and 40 CHO proteins. The results are displayed in Figure D.1. The computed RV coefficients are summarized in Tables D.1 and D.2. Due to display reasons transcriptome and proteome are abbreviated as T'ome, respectavly P'ome in the tables.

A permutation test was used to asses the significance of the eigenvalues of MCIA. The test revealed that the first three eigenvalues are statistically significant. Due to this, the first three MCIA axis were used to compute distances between different measured conditions. The distances were computed between each data set separately and a mean value of these distances is shown additionally. For comparison reasons the same distances were computed in the MCIA of ribosome relevant and secretion relevant features and displayed Table D.3.



(A)Ribosomes-Specific

(B)Secretion-Specific

Figure D.1: MCIA analysis of ribosome (A) and secretion (B) specific data sets.

Table D.1: Ribosome-specific RV coefficient.

	Pichia T'ome	Pichia P'ome	CHO T'ome	CHO P'ome
Pichia T'ome	1.000	0.566	0.915	0.631
Pichia P'ome	0.566	1.000	0.581	0.815
CHO T'ome	0.915	0.581	1.000	0.726
CHO P'ome	0.631	0.815	0.726	1.000

Table D.2: Secretion-specific RV coefficient.

	Pichia T'ome	Pichia P'ome	CHO T'ome	CHO P'ome
Pichia T'ome	1.000	0.506	0.966	0.633
Pichia P'ome	0.506	1.000	0.601	0.774
CHO T'ome	0.966	0.601	1.000	0.760
CHO P'ome	0.633	0.774	0.760	1.000

D.2 P Values of the MCIA Eigenvalues

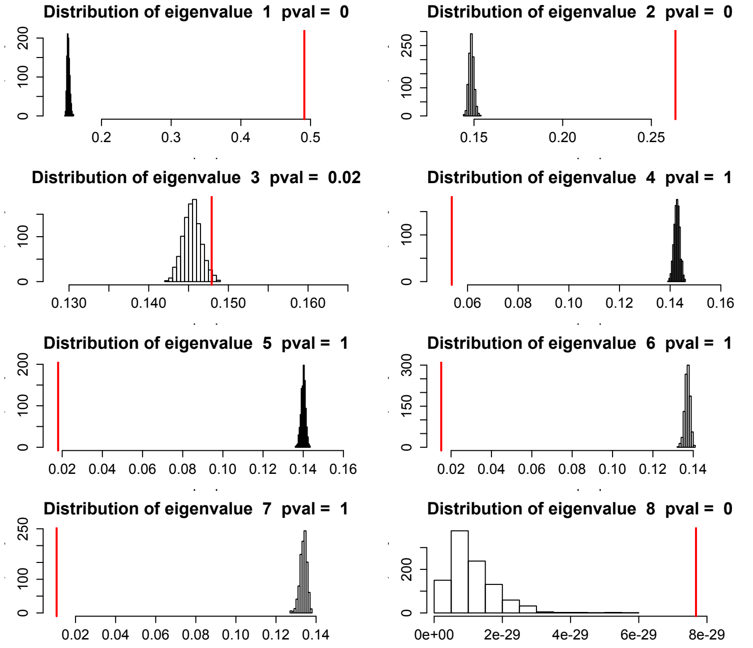


Figure D.2: P values for eigenvalues of MCIA including all conditions and all measured features.

Table D.3: Distances between different conditions computed from the MCIA of all features, ribosome relevant as well as secretion relevant features.

	Mean Distance	Pichia Transcriptome	Pichia Proteome	CHO Transcriptome	CHO Proteome
			Overall Distances		
3D6 L:WT to 3D6 H:WT	1.603	1.437	1.617	1.576	1.780
HSA L:WT to HSA H:WT	1.549	1.402	1.802	1.311	1.682
3D6 L:WT to HSA L:WT	2.142	1.997	2.326	2.343	1.903
3D6 H:WT to HSA H:WT	2.886	2.954	2.425	3.263	2.902
3D6 H:3D6 L to HSA H:HSA L	2.527	2.206	2.325	2.444	3.135
3D6 L:HSA L to 3D6 H:HSA H	2.456	2.205	2.581	2.444	2.593
			Ribosome Relevant Distances		
3D6 L:WT to 3D6 H:WT	1.288	1.143	1.537	1.217	1.255
HSA L:WT to HSA H:WT	1.280	1.002	1.260	1.375	1.483
3D6 L:WT to HSA L:WT	2.425	2.776	2.575	2.797	1.552
3D6 H:WT to HSA H:WT	2.913	2.597	2.736	3.438	2.880
3D6 H:3D6 L to HSA H:HSA L	2.691	1.706	2.900	1.983	4.174
3D6 L:HSA L to 3D6 H:HSA H	1.965	1.707	2.066	1.983	2.104
			Secretion Relevant Distances		
3D6 L:WT to 3D6 H:WT	1.722	1.663	1.792	1.592	1.842
HSA L:WT to HSA H:WT	1.590	1.243	1.712	1.447	1.957
3D6 L:WT to HSA L:WT	2.364	2.555	2.759	2.386	1.756
3D6 H:WT to HSA H:WT	2.721	2.751	2.397	2.904	2.832
3D6 H:3D6 L to HSA H:HSA L	2.206	1.915	2.006	1.865	3.039
3D6 L:HSA L to 3D6 H:HSA H	2.154	1.915	2.300	1.865	2.535
	Mean Distance	Pichia Transcriptome	Pichia Proteome	CHO Transcriptome	CHO Proteome

D.3 GO Slim Terms

Table D.4: GO Slim Terms

GO Slim Term ID	GO Slim Term Description
GO:0000910	cytokinesis
GO:0006091	generation of precursor metabolites and energy
GO:0006259	DNA metabolic process
GO:0006351	transcription, DNA-dependent
GO:0006399	tRNA metabolic process
GO:0006412	translation
GO:0006457	protein folding
GO:0006464	cellular protein modification process
GO:0006468	protein phosphorylation
GO:0006486	protein glycosylation
GO:0006520	cellular amino acid metabolic process
GO:0006725	cellular aromatic compound metabolic process
GO:0006766	vitamin metabolic process
GO:0006810	transport
GO:0006811	ion transport
GO:0006839	mitochondrial transport
GO:0006897	endocytosis
GO:0006950	response to stress
GO:0006970	response to osmotic stress
GO:0006974	response to DNA damage stimulus
GO:0006979	response to oxidative stress
GO:0006986	response to unfolded protein
GO:0006997	nucleus organization
GO:0007005	mitochondrion organization
GO:0007010	cytoskeleton organization
GO:0007029	endoplasmic reticulum organization
GO:0007031	peroxisome organization
GO:0007033	vacuole organization
GO:0007034	vacuolar transport
GO:0007049	cell cycle
GO:0007059	chromosome segregation
GO:0007126	meiosis
GO:0007165	signal transduction
GO:0009408	response to heat
GO:0009409	response to cold
GO:0015031	protein transport
GO:0016044	cellular membrane organization
GO:0016049	cell growth
GO:0016050	vesicle organization
GO:0016070	RNA metabolic process
GO:0016071	mRNA metabolic process
GO:0016072	rRNA metabolic process
GO:0016192	vesicle-mediated transport
GO:0016567	protein ubiquitination
GO:0016568	chromatin modification
GO:0016570	histone modification
GO:0019725	cellular homeostasis
GO:0030154	cell differentiation
GO:0032196	transposition
GO:0032989	cellular component morphogenesis
GO:0042221	response to chemical stimulus
GO:0042254	ribosome biogenesis
GO:0042594	response to starvation
GO:0043543	protein acylation
GO:0044255	cellular lipid metabolic process
GO:0044257	cellular protein catabolic process
GO:0044262	cellular carbohydrate metabolic process
GO:0045333	cellular respiration
GO:0046483	heterocycle metabolic process
GO:0048193	Golgi vesicle transport
GO:0051169	nuclear transport
GO:0051186	cofactor metabolic process
GO:0051276	chromosome organization
GO:0051301	cell division
GO:0051704	multi-organism process
GO:0070271	protein complex biogenesis
GO:0071554	cell wall organization or biogenesis

List of Figures

List of Tables

Publications

Biomed Tech 2013; 58 (Suppl. 1) © 2013 by Walter de Gruyter · Berlin · Boston. DOI 10.1515/bmt-2013-4311

INTEGRATIVE ANALYSIS OF -OMICS DATA: A METHOD COMPARISON

Oana A. Tomescu[1,2], Diethard Mattanovich[3,4] and Gerhard G. Thallinger[1,2]

[1]Corefacility Bioinformatics, Austrian Centre for Industrial Biotechnology, Graz, Austria
[2]Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria
[3]Cell Design and Cell Engineering, Austrian Centre for Industrial Biotechnology, Graz, Austria
[4]Department of Biotechnology, BOKU-VIBT University of Natural Resources and Life Sciences, Vienna, Austria

oana.tomescu@acib.at

Abstract: *Technological improvements have shifted the focus from data generation to data analysis. The availability of huge amounts of data like transcriptomics, protemics and metabolomics raise new questions concerning suitable integrative analysis methods. We compare three integrative analysis techniques (co-inertia analysis, generalized singular value decomposition and integrative biclustering) by applying them to gene and protein abundance data from six life cycle stages of Plasmodium falciparum. We create a network view of the GO terms associated to cell cycle stages by all three methods.*

Keywords: *Integrative analysis, comparison, genes, proteins.*

Introduction

Continuous technological improvements facilitate the availability of huge amounts of data resulting from the simultaneous characterization of the same organism or experimental condition. It is possible to measure the activity of thousands of genes, hundreds of proteins and hundreds of metabolites. Only the integrative analysis of all data types yields a deeper understanding of the system under study.

Methods

In this work we concentrate on gene and protein data. Most of the current analysis techniques are based on the assumption of a direct correlation between genes and proteins. This assumption does not hold due to post-transcriptional and post-translational expression regulation processes.

Here we compare three alternatives to conservative analysis techniques. Co-inertia analysis (CIA) is an integrative analysis method used to visualize and explore gene and protein data [1]. The generalised singular value decomposition (GSVD) [2] has shown its potential in the analysis of two transcriptome data sets. Integrative Biclustering (IBC) applies Biclustering [3] to gene and protein data.

We compare CIA, GSVD and IBC by applying them to gene and protein abundance data of *Plasmodium falciparum* [4]. The data was gathered from samples in six life cycle stages of the parasite: merozoite, ring, trophozoite, schizont, gametocyte and sporozoite. For the comparison we add additional information in from of gene ontology terms related to biological processes.

Results

Using CIA we visualize in Figure 1 the six life cycle stages and GO terms in a 2D plane. Each cell cycle stage is represented by it's projection in gene (circles) and in protein (squares) space connected through a line. The smaller the line between the two projections, the higher the concordance between the gene and protein data sets. Here we observe a very good agreement between the two data sets for all cell cycle stages. We notice the strict separation of the intraerythrocytic life cycle (ring, trophozoite, schizont and merozoite) from sporozoite and gametocyte stages.

Additionally, CIA offers the possibility of projecting GO terms (represented by numbers) onto the CIA plot and to associate them to a certain cell cycle stage (see Figure 1). The association of GO terms to the life cycle stages was done as follows. GO terms with positive x coordinates were associated to the stages of the intraerythrocytic life cycle. Due to the very close spatial position of these four stages no further discrimination between GO terms was possible. GO terms with negative x coordinates are associated to the gametocyte stage if they have positive y coordinates and to the sporozoite stage if they have negative y coordinates.

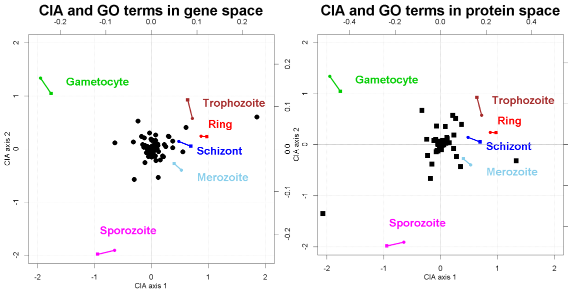


Figure 1: Co-inertia analysis and GO terms association.

With GSVD we decompose the data sets in matrices with biologically meaningful interpretations (arraylets, generalized eigenvalues and genelets) and explore the processes captured by them. The genelets represent biological processes captured by the data sets and expressed in the corresponding arraylets with a relative significance measured by the generalized eigenvalues. GO terms are associated to cell cycle stages through gene/protein set enrichment analysis based on the cell cycle stage depended angular distances (see Figure 2). All cell cycle stages are associated with both gene and protein space resulting in a gene and protein set enrichment analysis of genes and proteins showing the highest absolute values in the corresponding arraylets. Biclustering was applied to the gene, protein, life cycle stages and GO terms. The six life cycle stages are represented in Figure 3 by the left set containing the green (gametocyte), brown (trophozoite), red (ring), dark blue (schizont), light blue (merozoite) and pink (sporozoite) squares. Different biclusters are represented by distinct edge colors. The genes are colored in orange, the proteins in light blue and the GO terms in yellow. One can see that there are gene and proteins that strictly belong to one cluster (all edges of these nodes have the same color) as well as others that are associated to more than one bicluster (edges of these nodes have more than one color). GO terms are related to genes, proteins and cell cycle stages and they are connecting different biclusters.

We compare the results of the three integrative analysis methods showing in Figure 4 GO terms association to cell cycle stages common to all methods. Cell cycle stages not belonging to the intraerythrocytic cycle are either completely disconnected from the other stages (sporozoite) or connected by only one node (gametocyte linked through glycolysis) to the rest. The cell cycle stages of the intraerythrocytic cycle are densely interconnected.

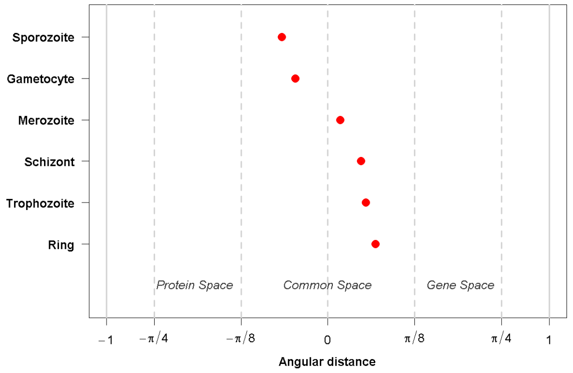


Figure 2: Generalized singular value decomposition - cell cycle depended angular distances.

Discussion

We have started this analysis with 4294 genes, 2903 proteins and 248 GO terms measured and annotated during six cell cycle stages of *P. falciparum*. The results of all methods were examined and GO terms associated to cell cycle stages by all methods were summarized in a GO term/cell cycle association network with 34 nodes (cell cycle stages and GO terms) and 42 edges. In concordance with the literature we observe a strong connectivity between the intraerythrocytic cell cycles and a low or non connectivity to the other stages. Each method produces a vast amount of results which are tedious to interpret. Inspection of the common associations is not only faster but it is more reliable and relevant because the results undergo a triple validation.

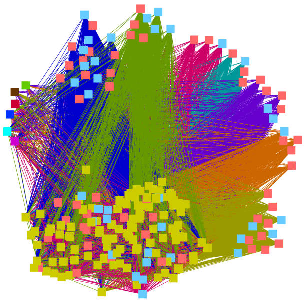


Figure 3: Biclustering of genes, proteins, life cycle stages and GO terms. The colors of the edges represent the different clusters. The genes are colored in orange, the proteins in light blue and the GO terms in yellow.

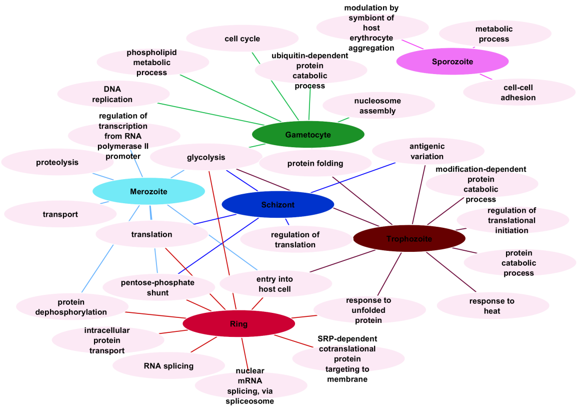

modulation by symbiont of host erythrocyte aggregation
metabolic process
cell cycle
phospholipid metabolic process
ubiquitin-dependent protein catabolic process
Sporozoite
cell-cell adhesion
DNA replication
nucleosome assembly
regulation of transcription from RNA polymerase II promoter
Gametocyte
proteolysis
glycolysis
protein folding
antigenic variation
modification-dependent protein catabolic process
Merozoite
Schizont
transport
regulation of translational initiation
translation
regulation of translation
Trophozoite
protein catabolic process
pentose-phosphate shunt
entry into host cell
protein dephosphorylation
response to unfolded protein
response to heat
Ring
intracellular protein transport
SRP-dependent cotranslational protein targeting to membrane
RNA splicing
nuclear mRNA splicing, via spliceosome

Figure 4: GO terms association to cell cycle stages common to all methods.

Acknowledgment

This work was funded by the Austrian Centre of Industrial Biotechnology (ACIB) supported by the Austrian BMWFJ, BMVIT, SFG, Standortagentur Tirol and ZIT through the Austrian FFG-COMET- Funding Program [FFG Grant 824186].

Bibliography

[1] A. Fagan, A. C. Culhane, and D. G. Higgins, “A multivariate analysis approach to the integration of proteomic and gene expression data.,” *Proteomics*, vol. 7, pp. 2162–71, 2007.

[2] O. Alter, P. O. Brown, and D. Botstein, “Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.,” *Proc. Natl. Acad. Sci. USA*, vol. 100, pp. 3351–6, 2003.

[3] S. Kaiser, *Biclustering: Methods, Software and Application.* PhD thesis, Ludwig-Maximilians-Universität München, 2011.

[4] K. G. Le Roch, J. R. Johnson, L. Florens, Y. Zhou, A. Santrosyan, M. Grainger, S. F. Yan, K. C. Williamson, A. Holder, D. J. Carucci, J. R. Yates, and E. Winzeler, “Global analysis of transcript and protein levels across the plasmodium falciparum life cycle.,” *Genome Res.*, vol. 14, pp. 2308–18, 2004.

JOURNAL OF NEUROTRAUMA 30:1232–1242 (July 15, 2013)
© Mary Ann Liebert, Inc.
DOI: 10.1089/neu.2012.2596

More than Cell Dust: Microparticles Isolated from Cerebrospinal Fluid of Brain Injured Patients Are Messengers Carrying mRNAs, miRNAs, and Proteins

Silke Patz,[1] Christa Trattnig,[1] Gerda Grünbacher,[1] Birgit Ebner,[2] Christian Gülly,[2] Alexandra Novak,[2] Beate Rinner,[2] Gerd Leitinger,[3] Markus Absenger,[2] Oana A. Tomescu,[4] Gerhard G. Thallinger,[4] Ulrike Fasching,[1] Sonja Wissa,[5] Juan Archelos-Garcia,[6] and Ute Schäfer[1]

Abstract

Microparticles are cell-derived, membrane-sheathed structures that are believed to shuttle proteins, mRNA, and miRNA to specific local or remote target cells. To date best described in blood, we now show that cerebrospinal fluid (CSF) contains similar structures that can deliver RNAs and proteins to target cells. These are, in particular, molecules associated with neuronal RNA granules and miRNAs known to regulate neuronal processes. Small RNA molecules constituted 50% of the shuttled ribonucleic acid. Using microarray analysis, we identified 81 mature miRNA molecules in CSF microparticles. Microparticles from brain injured patients were more abundant than in non-injured subjects and contained distinct genetic information suggesting that they play a role in the adaptive response to injury. Notably, miR-9 and miR-451 were differentially packed into CSF microparticles derived from patients versus non-injured subjects. We confirmed the transfer of genetic material from CSF microparticles to adult neuronal stem cells *in vitro* and a subsequent microRNA-specific repression of distinct genes. This first indication of a regulated transport of functional genetic material in human CSF may facilitate the diagnosis and analysis of cerebral modulation in an otherwise inaccessible organ.

Key words: cerebrospinal fluid; microparticles; microRNA; traumatic brain injury

Introduction

Microparticles (MPs) are cell-derived, membrane-sheathed structures that harbor a concentrated set of proteins, mRNA, and microRNA[1–5] and appear to transfer these components to adjacent and remote cells. MPs primarily termed ''platelet dust'' were first detected in blood in the late 1960s.[6] MPs have since been shown to be released by both vascular and non-vascular cells under a variety of physiological and pathophysiological conditions.[7–10] MP size and the molecules they carry vary greatly, often according to the specific shedding process and the cellular and subcellular shedding localizations.[9,11,12] This is reflected in the nomenclature. Membrane enclosed particles have been termed microvesicles, membrane particles, ectosomes, argosomes, exosomes, or nanospheres. Here, MP is used as a hypernym for these terms.

MP-mediated transcellular delivery is hypothesized to be of special importance in the blood by enabling the specific targeting of distal cells[8] and by preventing dilution of information in this expansive circulating system. Reports of MPs in cerebrospinal fluid (CSF), which is pumped around an analogous circulatory system comprising the ventricles, the subarachnoid space, and spinal cord, are in contrast limited.[10,13–16] The rationale for MP-based communication is nonetheless equally strong.

The morphological and proteomic analysis performed by Harrington and coworkers[17] that demonstrated the presence in CSF of discrete spheres associated with biochemically distinct components, such as prostaglandin H synthase, acetylcholine, or adenosine diphosphate-ribosylation factor protein, presently provides the best evidence that CSF contains MPs. In parallel intercellular RNA transfer by exosomal microvesicles was considered to be a possible mode of signaling within the nervous system.[18,19] This assumption was, however, focused mainly on intersynaptic communications.

This study is based on the hypothesis that CSF harbors a communication network comparable to the system described in blood.

[1] Research Unit for Experimental Neurotraumatology, Department of Neurosurgery, Medical University of Graz, Graz, Austria.
[2] Center of Medical Research, Medical University of Graz, Graz, Austria.
[3] Institute of Cell Biology, Histology and Embryology, Medical University of Graz, Graz, Austria.
[4] Austrian Centre of Industrial Biotechnology (ACIB), Institute for Genomics and Bioinformatics, Graz University of Technology, Graz, Austria.
[5] Department of Neurosurgery, Medical University of Graz, Graz, Austria.
[6] Clinical Division of General Neurology, Department of Neurology, Medical University of Graz, Graz, Austria.

It was our aim to establish the presence of MPs in the CSF (CSF-MPs) derived from non-injured subjects (NIS) and traumatic brain-injured patients (TBIP) and to analyze their potential as vehicles that shuttle distinct biological and genetic information.

Methods

Ethics statement

The ethics committee of the Medical University of Graz specifically approved this study on February 10, 2009. The study and the potential to participate were explained to patients before they were discharged from the clinic. At 4 to 6 months later, letters were sent to the patients with the same information and the option to call for more information. Only patients who consented to the study by returning a signed form by mail at this time point were included. Six of the trauma patients called back for more information before sending a signed form. In two cases, guardians consented on behalf of participants.

Clinical parameters of TBIPs and healthy controls included in the study

CSF samples were collected from patients with severe TBI traumatic brain injuries (Glasgow Coma Scale [GCS] ≤ 8) when ventricular drainage was implemented as a measure of intensive care treatment. Because of ethical considerations, samples were not taken at standardized time points but rather when ventricular drainage was indicated from acute increased intracranial pressure. A higher risk of ventricular drain infections resulting from additional study dependent interventions was thereby avoided.

There were 26 samples collected from 11 patients over 2 years. The number of samples per patient and the time points of sample collection relative to the time point of primary injury varied considerably (time of sampling corresponding to the time of injury is indicated in the respective figures). Further, differences in the location and type of brain injury as well as additional injuries to different body parts and organ systems render the group of brain-injured patients a rather heterogeneous study population (Table 1). Hence, statistical analysis of results is a mere attempt to draw some general conclusions, but must be interpreted with caution. The analyzed parameters are additionally presented for each subject, thereby allowing for a more personalized interpretation of data.

CSF samples from adults who received a lumbar puncture for diagnostic purposes were collected for control studies ($n=26$). Only subjects without a confirmed subarachnoid hemorrhage or infection (confirmed central nervous system [CNS] diseases [CCD]) were included as non-injured controls (NIS, $n=17$; Table 1; see online supplementary Fig. S1,S2 at ftp.liebertpub.com)

Preparation of CSF

CSF was obtained from patients with TBI by ventricular drainage and from NIS by lumbar puncture. Samples were taken directly from the external ventricular drain tube. Sample volumes varied between 3–7 mL (TBIP) and 0.2–1 mL (NIS). Samples were stored no longer than 30 min at 4°C before further processing. CSF samples were centrifuged at 400 g for 5 min and filtered through a 0.45 μm syringe filter; 1 μL of Protease Inhibitor Cocktail (Sigma)/mL of CSF was added. The CSF was then stored at −80°C until further analysis. MPs were isolated by thawing CSF on ice, followed by ultracentrifugation (Optima L-90k; Beckman Coulter) for 40 min at 170,000 g at 4°C. Supernatant was carefully discarded, and the resulting MP pellets were processed according to the intended experiment.

Flow cytometric detection of MPs

There were 100 μL MP suspensions transferred into Trucount^{TM}Tubes (BD Bioscience) containing a known number of fluorescent beads (diameter: 1 μm) to enumerate MPs. Flow cytometry were performed using the LSR II BD Bioscience and the BD FACSDiva^{TM} 6.0 software. Forward and side scatter were set in logarithmic scale. MPs were differentiated from signal noise by threshold settings of 200 forward scatter and 200 side scatter. The acquisition was terminated after 2000 bead counts. The number of MPs were calculated by the following formula described by Shet and associates[20]: MP [counts/mL] = [(total beads per tube/beads counted) × events counted] × dilution factor.

microRNA (miRNA) preparation, miRNA array, and bioinformatic analysis

For miRNA isolation, at least 50 mL of CSF pooled from 10 patients with TBI was centrifuged as described above. The MP pellet was prepared according to the protocol of the Qiagen miRNeasy Kit. Rat brain RNA and miRNA were isolated for use as internal controls. RNA and miRNA quality and concentrations were determined using a Bioanalyzer 2100 (Agilent).

MP samples and two miRNA rat brain samples were analyzed using the Affymetrix GCS300 (Affymetrix) with FlashTag Biotin HSR (Gensiphere).

The miRWalk (http://www.ma.uni-heidelberg.de/apps/zmf/mirwalk/; effective 3/30/2011) database was used to detect validated target genes of the identified CSF-MP-associated human miRNAs (hsa-miR). A gene ontology (GO) analysis was conducted to determine the neuron-related biological processes overrepresented in the computed set of genes. To this end, the validated target genes were imported to Cytoscape.[21] The GO analysis itself was performed with the BiNGO[22] plug-in. The hypergeometric test was used with the Benjamini & Hochberg False Discovery Rate correction and a significance level of 0.05.

Overrepresented neuron-related biological processes with a p value less than 0.001 were used for visualization. These processes were searched for in the initial GO network and rearranged without altering their connections to the other processes.

PCR-mRNA

For the polymerase chain reaction (PCR), the RNA concentration in the CSF was determined using the Ribogreen assay kit (Invitrogen) in accordance with the manufacturer's protocol; RNA content was measured using a POLARstar optima fluorometer. To detect mRNAs for activity-regulated cytoskeleton-associated protein (ARC), ß-ACTIN, microtubule-associated protein 2 (MAP2), and LIM domain kinase 1 mRNA (LIMK1), Dicer, fibroblast growth factor receptor 1 (FGFR1), and CD133 cDNA was synthesized with the First Strand cDNA Synthesis Kit (Fermentas) according to the manufacturer's protocol using 0.5 ng of total RNA. Real-time PCR (RT-PCR) was performed using the FastStart PCR Master Mix (Roche) and specific primers (Eurogentec) (ARC [412 bp]; ß-ACTIN [327 bp]; MAP2 [319 bp], LIMKI [264 bp]). PCR conditions were kept within the linear range determined for every product. Expression levels were normalized to the corresponding PCR product from NT2 cell cDNA. For primer sequences, see online supplementary Table S1 at ftp.liebertpub.com.

PCR miRNA

For miRNA analysis, NCode™ miRNA qRT-PCR Kits (Invitrogen) were used according to the manufacturer's protocol. cDNA was synthesized using 0.5 ng of total RNA. Universal Primer was provided in the reaction kit, specific miRNA primers were designed using the NCode™ miRNA Database (Invitrogen), and RT-PCR was performed using a Roche LightCycler® 480 according to the manufacturer's instructions. PCR for mRNA and miRNA was performed three times per cDNA sample; cDNA was prepared three times from each CSF sample.

Table 1. Clinical Parameters of Traumatic Brain Injured Patients and Non-Injured Subjects

	Traumatic brain injury patients (TBIP)		*Non-injured subjects (NIS)*	
Number of patients (female; male) Number of samples	n = 11 (f: n = 5; m: n = 6) n = 26		n = 17 (f: n = 8; m: n = 9) n = 17	
Age (years) (female; male)	47.85 ± 16.88 (f: 55.76 ± 11.37; m: 41.26 ± 18.77)		44.54 ± 12.21 (f: 41.64 ± 12.46; m: 45.80 ± 12.10)	
Body height (cm) (female; male)	174.88 ± 14.90 (f: 167.33 ± 7.37; m: 179.40 ± 17.13)		175.56 ± 8.53 (f: 171.33 ± 7.02; m: 177.67 ± 8.98)	
Body weight (kg) (female; male)	78.38 ± 20.27 (f: 73.33 ± 17.95; m: 81.40 ± 22.96)		74.67 ± 14.11 (f: 68.33 ± 8.51; m: 77.83 ± 15.92)	
Body mass I index (kg/m2) (female; male)	25.26 ± 3.43 (f: 25.93 ± 4.62; m: 24.85 ± 3.06)		24.11 ± 3.28 (f: 23.31 ± 2.93; m: 24.51 ± 3.64)	
Glasgow Coma Scale (GSC)	GSC 3:	*n* = 5		
	GSC 4:	*n* = 2		
	GSC 5:	*n* = 1		
	GSC 7:	*n* = 1		
	GSC 11*:	*n* = 2		
Outcome	9 survived, 2 died			
Ø Duration of stay (days)	22			
Additional diagnoses	Exogenous noxa	n = 5	Headache	n = 6
	(Hemato)pneumotheorax	n = 5	Vertigo	n = 4
	Rib fractures	n = 3	Tinnitus	n = 3
	Dysphagia	n = 3	Spinal Pain	n = 2
	Face/skull fractures	n = 2	Arthrosis	n = 2
	Multiple injuries	n = 1	Brain tumor	n = 2
	Internal injuries	n = 1	Optic nerve damage	n = 2
	Lung contusion	n = 1	Vascular obliteration	n = 2
	Polytrauma	n = 1	Concussion	n = 1
	Hand injury	n = 1	Myopia	n = 1
	Hip injury	n = 1	Preeclampsia	n = 1
	Spine injury	n = 1	Pregnancy	n = 1
	Pelvic fracture	n = 1	Facial nerve paresis	n = 1
	Clavicle fracture	n = 1	Abdominal pain	n = 1
			Vomitus	n = 1
			Common cold	n = 1
			Stent control	n = 1
			Cerebral infarction	n = 1
			Epistaxis	n = 1
			Circulatory collapse	n = 1
			Proliferative retinopathy	n = 1
			Macula degeneration	n = 1

*The GCS initially diagnosed is provided.

PCR reaction products were separated on a 2–4% agarose gel, scanned, and band intensities quantified by densitometry using Photoshop CS5 software. For primer sequences see online supplementary Table S1 at ftp.liebertpub.com.

Electron microscopy

MP pellets were resuspended in 100 μL phosphate buffered solution (PBS) for electron microscopy. Then 10 μL of the sample were transferred to pioloform-and-carbon-coated grids, blot dried, and rinsed with water. Negative staining was performed by applying 2% uranyl acetate for 1 min. Samples were then air-dried and viewed with a Zeiss EM 901 transmission electron microscope.

Western blot

CSF pellets derived from CSF samples pooled from five patients were homogenized in 100 μL radioimmunoprecipitation assay buffer containing 5% sodium dodecyl sulfate (SDS), phenylmethanesulfonylfluoride (PMSF), iodoacetamide, and aprotinine (each 1 mM). Protein content was determined using a BCA Protein Assay Kit (Novagen®). Ten μg of protein was loaded on a 10% SDS-PAGE gel followed by transfer to nitrocellulose (Schleicher & Schuell; BA85) in a Tank Blotter (Biorad). Blots were blocked with 1% non-fat milk (Sigma) in Tris-buffered saline (TBS) for 3 h. Filters with specific antibodies were incubated overnight in blocking solution (rb α eIF2C 1:100 [argonaute2 (AGO2); Santa Cruz]; gt α Staufen (STAU2) 1:50 [Santa Cruz]; mou α GFI-1 1:500 [Sigma]), followed by 2 × TBST and 3 × TBS washing steps and incubation with biotinylated antibodies for 1 h. For visualization, blots were washed again and developed by the ABC-horseradish peroxidase (Vectashield) method using diaminobenzidine as chromogen.

Cytochemistry

Ten mL of CSF was incubated with 2 mg EZ-Link® Sulfo-NHS-Biotin (Thermo Scientific) for 30 min at room temperature (RT), followed by ultracentrifugation as described above. Pellets were resuspended in 400 μL PBS and streptavidin Cy5-conjugated (1:200; Invitrogen), and RiboGreen (1:100) was added and incubated for 30 min at RT. The fraction was then washed with 10 mL PBS and again ultracentrifuged as described above. Ten μL of the MP suspension was transferred to an object slide and air-dried at 37°C. The MPs were finally washed (3 × PBS) and mounted with Aquatex (Merck) and examined under the microscope (amplification 40 ×; images were also digitally enlarged).

For antibody staining, MPs were treated with 1% bovine serum albumin (BSA)-PBS blocking solution for 1h. Antibody gt α hnRNP A2/B1 1:100 (Santa Cruz) was then incubated overnight followed by 3 × PBS washing and incubation with biotinylated antibodies for 1 h. For visualization, slides were again washed and developed using the ABC-horseradish peroxidase method using diaminobenzidine as chromogen. Slides were mounted with Eukitt® (Kindler).

Incubation of NT-2 cells with MPs

NTERA2 clone D1 (NT2.cl.D1) embryonal carcinoma stem cells were grown and maintained in Dulbecco's Modified Eagle's Medium, supplemented with 10% fetal calf serum and 2 mM L-glutamine at 37°C in 5% CO_2. NT2 cells were seeded on a glass object slide for further investigation. Ten mL CSF was incubated with 2 mg EZ-Link Sulfo-NHS-Biotin for 30 min at RT, followed by ultracentrifugation as described above. Pellets were resuspended in 400 μl PBS and Cy5 conjugated streptavidin (1:200), and RiboGreen (1:100) was added and incubated for 30 min at RT. The fraction was then washed with 10 mL PBS and again ultracentrifuged as described above. The pellet was resuspended in 100 μL PBS and 50 μL applied to the NT2 cells in a six-well dish and incubated for 30 min. Glass slides were then removed from the wells subsequent to a wash step with 10 mL PBS, mounted with Aquatex (Merck), and examined with a TCS SP2 (Leica) confocal microscope.

For PCR analysis, cells were incubated with MPs containing 5 ng or 10 ng RNA. The RNA amount was measured using the RiboGreen assay. After 1 h and 3 h incubation time, cells were washed twice with 10 mL PBS harvested and RNA was isolated using the RNeasy micro Kit. RNA amount was measured using Nanodrop,® and cDNA libraries were prepared as previously described, followed by PCR for Dicer1, CD133, FGFR1, miR-451, and U6 as housekeeping gene as already described above.

Silencing of miR-451

NTERA2 cells were cultured as described above to a confluence about 30–50%. Cells were transfected using X-tremeGENE siRNA transfection reagent (Roche) combined with a final concentration of 50 nM of either hsa-miR-451 inhibitor (miRCURY LNA microRNA inhibitor) or scrambled (Negative Control A) (Exiqon) following the manufacturer's protocol and preincubated for 2 h. MPs were prepared as described above, and RNA content was measured using the RiboGreen assay. MPs with a content of 10 ng RNA were then supplied to the cells for 1 h. Cells were washed, harvested, and analyzed as described above.

Statistical analyses

Comparison of CSF-MPs (n/100 μL) and RNA content (ng/μL) in TBIPs and NIS as well as comparison of content of ARC, LIMKI, MAP2, and β-ACTIN in TBIPs and NIS were performed with the non-parametrical Mann-Whitney *U* test for independent samples. Statistical significance was set at the level of $\alpha = 0.05$.

Frequency of occurrence of miRNA species in CSF samples of TBIPs and NIS was evaluated statistically with the logistic regression analysis. NIS were defined as 0, TBI patients as 1; miRNA not present was defined as 0, miRNA present as 1. The Nagelkerke's R^2 and the omnibus-test for statistical significance were performed in SPSS 18 (PASW). Statistical significance was set again at the level of $\alpha = 0.05$.

Comparison of CSF-MPs (n/100 μL) and RNA content (ng/μL) in NIS and CCD patients (see online supplementary Fig. S1,S2 at ftp.liebertpub.com) was again performed with the non-parametrical Mann-Whitney *U* test for independent samples.

Results

Quantification of MPs in CSF from patients with TBI and healthy subjects

CSF was collected from TBIP by ventricular drainage when indicated as a measure of intensive care treatment. Control CSF from NIS was obtained after lumbar puncture implemented for diagnostic purposes. CSF samples of subjects without subarachnoid hemorrhage or CSF infections were included as non-injured controls (NIS, Table 1). Patients diagnosed for subarachnoid hemorrhage or CNS infection (CCD) were not included in the comparative analysis of CSF-MPs. CCD patient diagnosis as well as the number of MPs and RNA content of respective samples are presented in Figure S1 (see online supplementary Fig. S1 at ftp.liebertpub.com.

CSF was collected and stored by a standardized protocol (see Methods). Putative MP-containing fractions were derived from CSF samples by serial centrifugation and filtration (< 400 nm). High magnification transmission electron microscopy showed the samples to contain a heterogeneous population of intact particles ranging in size from 50 to 400 nm (Fig. 1A). An increasing number of reports suggest MPs to be membrane sheathed shuttles for a variety of RNA molecules and proteins.[11] Staining of the CSF-MPs with RiboGreen (RNAs) and membrane-specific Sulfo-NHS-biotin confirmed the presence of membrane covered particles that carry RNA molecules (Fig. 1B–D). Furthermore, mRNA in CSF was protected from RNaseI digestion underlining the finding that RNA is shuttled by membranous particles (see online supplementary Fig. S3 at ftp.liebertpub.com).

The particles were quantified by flow cytometry and shown to be significantly more abundant in TBIPs than in NIS (Fig. 1E,F). They were also more abundant in CCDs, suggesting an association between CNS damage or disease and elevated numbers of CSF-MPs (see online supplementary Fig. S1 at ftp.liebertpub.com). The RNA levels in CSF-MPs did not differ significantly in TBIPs or NIS (Fig. 1G). It has to be taken into account, however, that the relative numbers of CSF-MPs and the relative RNA levels varied considerably between individual patients or NIS samples (see online supplementary Fig. S1,S2 at ftp.liebertpub.com).

CSF-MPs are enriched in mature mRNAs and proteins that are associated with neuronal RNA-granules

There is evidence that MPs carry ribonucleoproteins, which are associated with the transport of RNA to distal cellular sites, such as dendrites, to enable localized translation.[23] In accordance with this observation, we detected typical neuronal granule proteins such as MAP2, ARC, and LIMK1 mRNA, as well as ß-ACTIN in CSF-MPs from both patients and controls (Fig. 2). A semi-quantitative comparison of the relative levels of distinct mRNAs in single samples that would allow a correlation of RNA levels with disease progression was disregarded, because validated normalization candidates were not available for CSF-MPs derived mRNA. An attempt to evaluate the average level of these mRNAs in CSF-MPs derived from patient or NIS samples was implemented by using external signals derived from NT-2 cell cDNA amplification for normalization (see online supplementary Fig. S4A–D at ftp

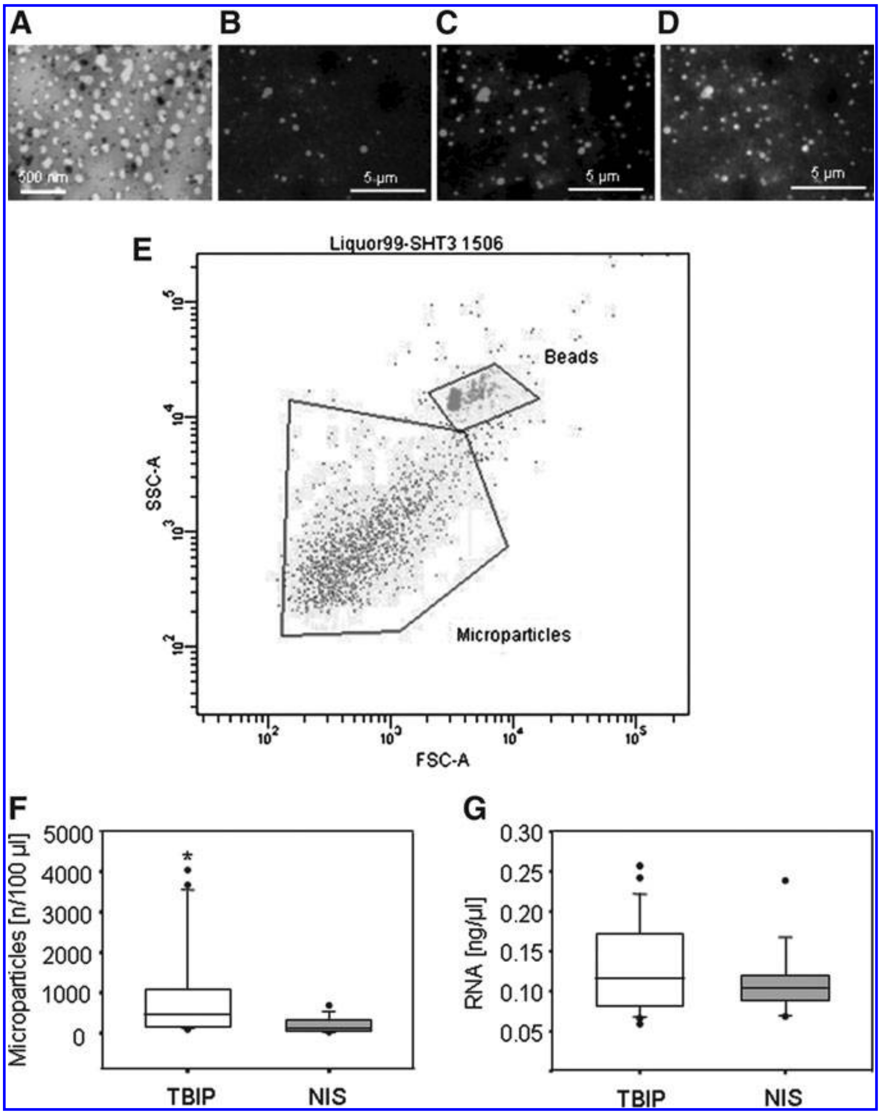


FIG. 1. Detection and quantification of microparticles derived from cerebrospinal fluid (CSF-MPs). (**A**) Electron microscopy of MP pellet derived from CSF of brain injured patients; (**B**) fluorescent microscopy of MP pellet after biotinylation of membrane proteins and subsequent visualization with anti-biotin antibodies; (**C**) fluorescent microscopy of MP pellet after incubation with RiboGreen; (**D**) overlay of images B and C; (**E**) Representative FACS image; 1 μm beads; MPs analyzed in CSF sample (**F**) number of microparticles in cerebrospinal fluid * $p \leq 0.04$; (**G**) concentration of RNA in CSF-MPs; traumatic brain injured patients (TBIP) (samples: $n = 26$); non-injured subjects (NIS) (samples: n = 17); box blots: 50% of ratings have values within the box. Twenty-five percent are more and 25% are less than the values within the box. Horizontal line inside the box: median. Upper boundary of whisker: largest observed value that is not an outlier. Lower boundary of whisker: smallest observed value that is not an outlier.

.liebertpub.com). This approach has to be considered a compromise and has thereby to be interpreted with caution.

We also verified the presence of RNA granule-associated proteins in CSF-MPs derived from TBIP. AGO2 and STAU2 were detected by Western blotting of CSF-MPs proteins isolated from pooled CSF (Fig. 2B). This pool was also used to visualize hnRNP in CSF-MPs by immunocytochemistry.

The very small volumes of CSF available precluded a direct comparison of the protein levels in CSF-MPs from NIS by Western blotting.

Profiling of CSF-MP microRNAs

A bioanalyzer profile of CSF-MP RNA revealed the presence of a broad range of RNA sizes including a prominent peak of small RNA species (51% of total RNA) (Fig. 3A) consistent with an enrichment in miRNAs. Based on this result, we hybridized RNA from pooled CSF-MPs from 10 brain-injured patients with Affymetrix GeneChip miRNA arrays. Eighty-one distinct miRNAs were identified (see online supplementary Table S2 at ftp.liebertpub.com). Twenty-three of the identified miRNAs have been previously indicated to be differentially regulated in rat brains after TBI[24,25] Furthermore, eight of the identified miRNA species have been demonstrated to play a role in the regulation of neurological function. Five miRNA species share a role in neurological function and experimental injury induced regulation (Fig. 3B).

A total of 1659 target genes were linked to the identified miRNAs using miRWalk database (see online supplementary Table S3 at ftp.liebertpub.com). Biological processes overrepresented by the identified target genes were examined by GO analysis.

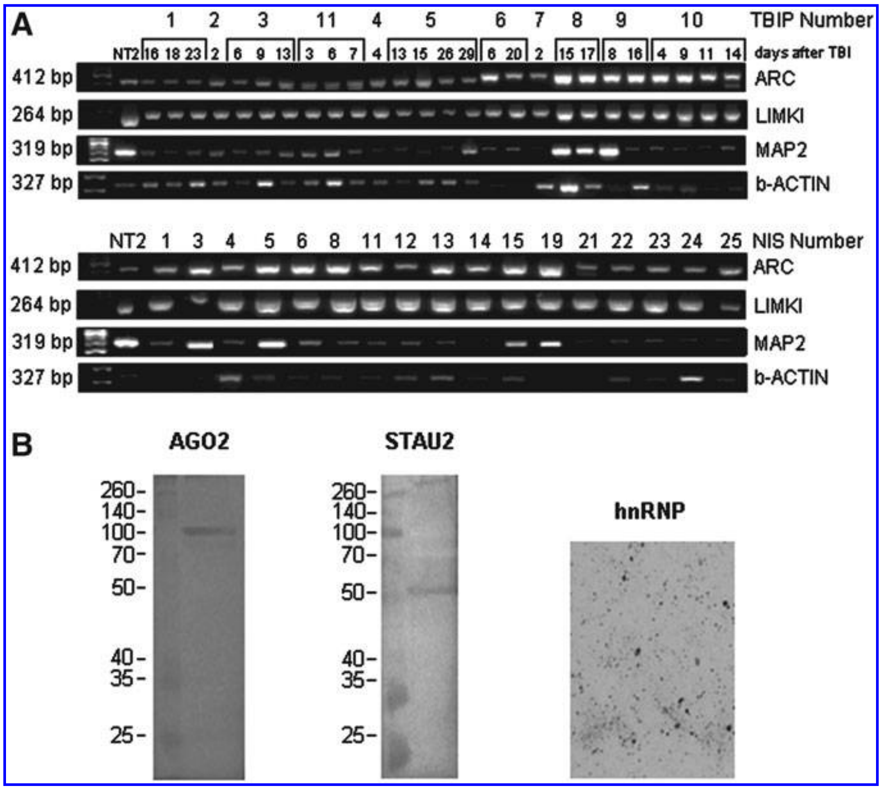

A
1 2 3 11 4 5 6 7 8 9 10 TBIP Number
NT2 16 18 23 2 6 9 13 3 6 7 4 13 15 26 29 6 20 2 15 17 8 16 4 9 11 14 days after TBI
412 bp ARC
264 bp LIMKI
319 bp MAP2
327 bp b-ACTIN
NT2 1 3 4 5 6 8 11 12 13 14 15 19 21 22 23 24 25 NIS Number
412 bp ARC
264 bp LIMKI
319 bp MAP2
327 bp b-ACTIN
B
AGO2
STAU2
hnRNP
260- 140- 100- 70- 50- 40- 35- 25-
260- 140- 100- 70- 50- 40- 35- 25-

FIG. 2. Detection of mRNAs shuttled by cereb rospinal fluid microparticles (CSF-MPs). (**A**) Representative agarose gel images of ARC, LIMK1, MAP2, and ß-ACTIN; traumatic brain injured patients (TBIP) (samples: $n=26$); non-injured subjects (NIS) (samples: $n=17$); experiments were replicated three times. (**B**) Western blots of Argonaute 2, Staufen 2, and immunostaining of hnRNP protein in CSF-MPs from TBIP ($n=5$, pooled)

Thirty-seven overrepresented biological processes seem to be neuron related (see online supplementary Table S4 at ftp.liebertpub.com). Thirty-four of the neuron related processes are associated with genes targeted by the subset of miRNA summarized in Fig. 3B. A subset of these processes with a p value smaller than 0.001 is represented in Fig. 3C.

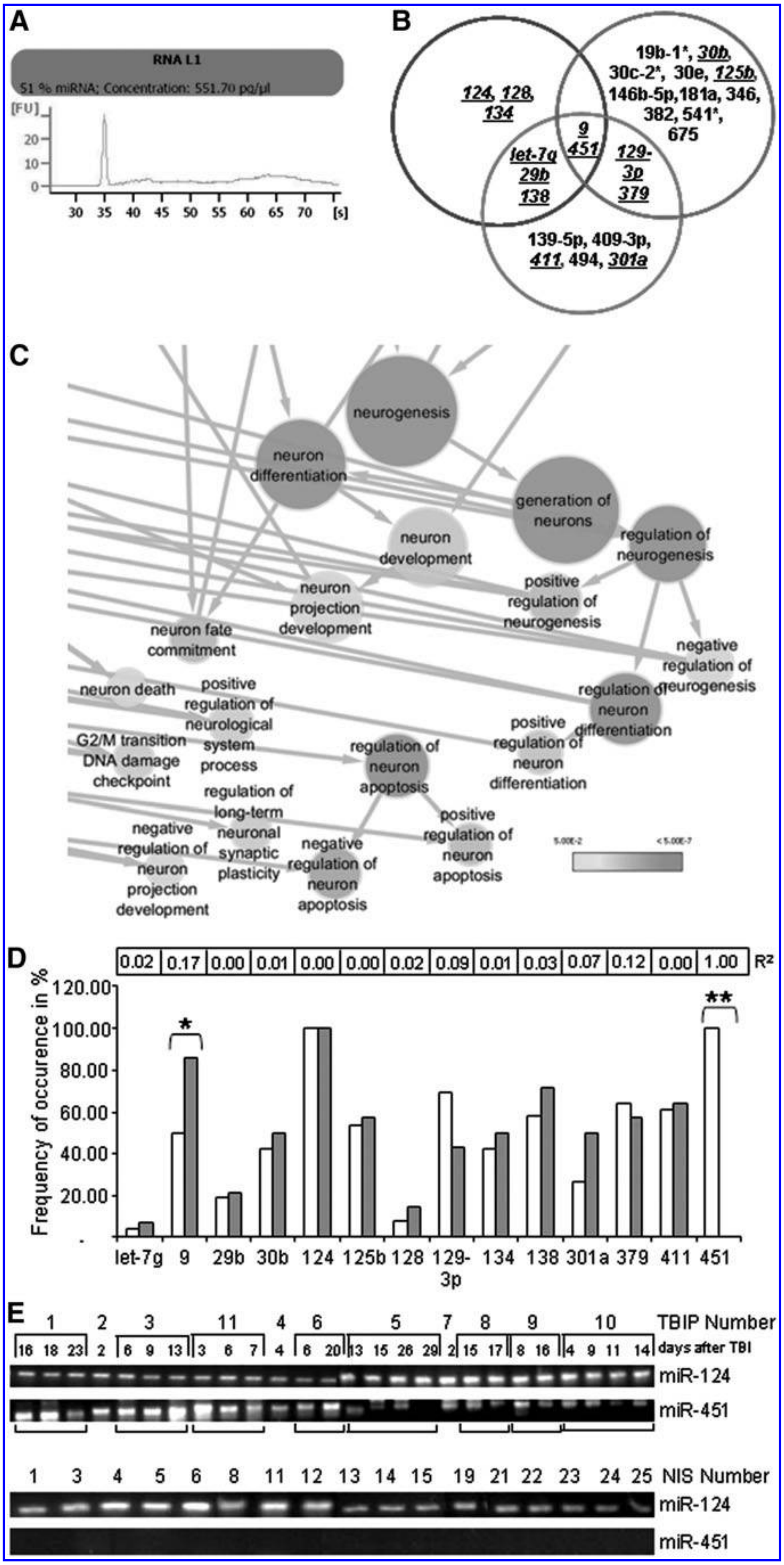
CSF derived MPs from TBIP and healthy subjects contain a distinct pattern of microRNAs

The presence of 14 distinct miRNA species in CSF-MPs derived from individual TBIPs and NIS was verified by qRT-PCR (Fig. 3B, underlined). A miRNA was deemed to be present in the CSF-MPs of a given patient or NIS when the intensity of the respective miRNA RT-PCR signal was reproducibly 20% higher than the background signal. The frequency of occurrence of distinct miRNAs was determined by logistic regression analysis (Fig. 3D-E). Using this approach, we verified the presence of miR-124 in all samples (R^2: 0.00). miR-451 in contrast was only detected in CSF-MPs from TBIPs (R^2: 1, $p=0.00$) and could therefore play a regulatory role associated with TBI. The presence of miR-9 on the other hand was shown to be more prevalent in CSF-MPs derived from NIS (R^2: 0.172, $p \leq 0.025$). A statistically significant or predictive frequency of occurrence was not observed for any of the other miRNA species analyzed.

Functional transfer of CSF-MP components to undifferentiated NTERAs

The ability of CSF-MPs to deliver their constituents to target cells was investigated by incubating membrane-specific biotinylated, RiboGreen-stained MPs with undifferentiated NTERAs (NT-2), a neuronally committed human teratocarcinoma cell line.[26] Undifferentiated NTERAs express nestin and vimentin, intermediate filament (IF) proteins expressed in neuroepithelial precursor cells, as well as MAP1b, expressed in human neuroepithelium.[27] Uptake of RiboGreen stained RNA was shown by fluorescence microscopy (Fig. 4A–E). The non-superposition of green

FIG. 3. Profiling of microRNAs derived from cerebrospinal fluid microparticles (CSF-MPs). (**A**) Range of miRNA sizes in CSF-MPs as determined by Bioanalyser 2100. (**B**) Identified miRNAs that have been demonstrated to regulate neuronal function or to be regulated during experimental traumatic brain injury demonstrated in two different studies[24,25]; underlined: miRNA presence in CSF-MPs verified by real-time polymerase chain reaction. (**C**) Hierarchical view of overrepresented neuron related biological processes presenting a p value <0.001. The size of the nodes is proportional to the number of genes that were annotated to the biological process. (**D**) Frequency of occurrence of specific miRNAs in CSF-MPs; $n=26$: traumatic brain injured patients (TBIP) (white bars); $n=17$: non-injured subjects (NIS): grey bars. Frequency of occurrence: R^2 values are indicated in the upper row, $*p \leq 0.03$; $**p \leq 0.001$. (**E**) Representative agarose gel images of the occurrence of miRNA-451 ($R^2=1.0$) and miR-124 ($R^2=0$) in CSF-MP samples; experiments were replicated three times.


A
RNA L1
51 % miRNA; Concentration: 551.70 pg/µl
[FU]
20
10
0
30 35 40 45 50 55 60 65 70 [s]
B
124, 128, 134
9 451
let-7g 29b 138
129-3p 379
19b-1*, 30b, 30c-2*, 30e, 125b, 146b-5p, 181a, 346, 382, 541*, 675
139-5p, 409-3p, 411, 494, 301a
C
neurogenesis
neuron differentiation
generation of neurons
regulation of neurogenesis
neuron development
neuron projection development
positive regulation of neurogenesis
neuron fate commitment
negative regulation of neurogenesis
neuron death
positive regulation of neurological system process
regulation of neuron differentiation
G2/M transition DNA damage checkpoint
regulation of neuron apoptosis
positive regulation of neuron differentiation
regulation of long-term neuronal synaptic plasticity
negative regulation of neuron projection development
negative regulation of neuron apoptosis
positive regulation of neuron apoptosis
5.00E-2
< 5.00E-7
D
0.02 0.17 0.00 0.01 0.00 0.00 0.02 0.09 0.01 0.03 0.07 0.12 0.00 1.00 R²
Frequency of occurence in %
120.00
100.00
80.00
60.00
40.00
20.00
-
*
**
let-7g 9 29b 30b 124 125b 128 129-3p 134 138 301a 379 411 451
E
1 2 3 11 4 6 5 7 8 9 10 TBIP Number
16 18 23 2 6 9 13 3 6 7 4 6 20 13 15 26 29 2 15 17 8 16 4 9 11 14 days after TBI
miR-124
miR-451
1 3 4 5 6 8 11 12 13 14 15 19 21 22 23 24 25 NIS Number
miR-124
miR-451

fluorescent RNA and red fluorescent membranes might indicate that RNA is released from MPs during uptake (Fig. 4E).

On addition of CSF-MPs derived from TBIPs to NTERAs, a rapid (1 h) concentration and time-dependent miR-451 mediated decrease of Dicer, FGFR1, and CD133 mRNA expression was observed (Fig. 4F–H). These genes are listed as putative miR-451 target genes in the miRWalk database. The decrease in putative target gene expression was accompanied by an increase in miR-451 in these cells (see online supplementary Fig. S1B at ftp.liebertpub.com). These results are in accordance with the detection of miR-451 in patient CSF-MPs. Down-regulation of putative miR-451 target genes was not observed when CSF-MPs from NIS were added to the culture. Three hours after incubation, gene expression of Dicer, FGFR1, and CD133 was comparable to untreated cells. A significant increase in ß-ACTIN mRNA was mediated by CSF-MPs derived from patients or NIS (Fig. 4I). The increase in cellular ß-ACTIN mRNA levels is probably due to the transfer of ß-ACTIN mRNA from MPs to cells.

miR-451 was inhibited by a LNA miR-451 inhibitor but not by LNA scrambled when added to the cells before incubation of NTERAs with CSF-MPs derived from TBIPs (Fig. 4J). These results demonstrate the down-regulation of Dicer, FGFR1, and CD133 mRNA to be specifically mediated by miR-451.

Discussion

In this study, we show for the first time that human CSF contains membrane-sheathed MPs that carry genetic information and proteins. We demonstrate that the genetic information comprises mRNAs associated with RNA granules and miRNAs implicated in the regulation of neuronal processes. CSF-MPs are more abundant in subjects with brain injury and shuttle a distinct set of miRNAs, including miR-9 and miR-451, both of which have previously been shown to be regulated in cerebral tissue after experimental traumatic brain injury.[24,25] We confirm a transfer of genetic material from CSF-MPs to cultured NTERAs and a subsequent decrease in putative miR-451 target gene expression, suggesting CSF to harbor a transcellular delivery system that contributes to the signaling between cells.

Experimental design

Analyzing CSF samples derived from TBIP is to some extent restricted by a limiting study design. A primary impediment is the absence of true controls; i.e., CSF of healthy subjects. Observed differences between CSF samples derived from TBIP and NIS should therefore be interpreted with care. Further, samples are taken from different locations (ventricle/subarachnoid space lumbar region). To exclude a potential rostro-caudal gradient of CSF-MPs and the associated content, a study analyzing MPs and RNA content in lumbar versus ventricular CSF samples in the same patient population is necessary.[28–33]

MPs in CSF

We confirmed the presence of CSF- MPs in human CSF samples by electron and fluorescent microscopy. Consistent with descriptions in body fluids and in cell cultures, CSF-MPs were highly heterogeneous with sizes ranging from 50–400 nm, which suggests that they are generated by a combination of membrane budding and exocytosis (MPs in the lower size range).[13,14,17,34,35]

Whereas the range of CSF-MP sizes was comparable in patients and NIS, the abundance of CSF-MPs was clearly elevated in the former group (see above), consistent with reports in other non-CNS disease states,[7,8] albeit the numbers detected were considerably below those reported in a previous CNS study. The discrepancy, however, most likely reflects the use of different isolation protocols; i.e., sample filtration through a 0.45 micron mesh after high speed centrifugation might result in significantly reduced numbers of particles.[17,36]

RNA carried by MPs derived from CSF

The amount of RNA per MP sample varied considerably between samples. The composition and function of MPs are dependent on their cellular origin, the agonist responsible for MP formation, and the microenvironment of the parental cell.[37,38] The heterogeneity of the MP population in CSF samples might be the reason why a correlation between RNA levels and the health status of the donor was not detected.

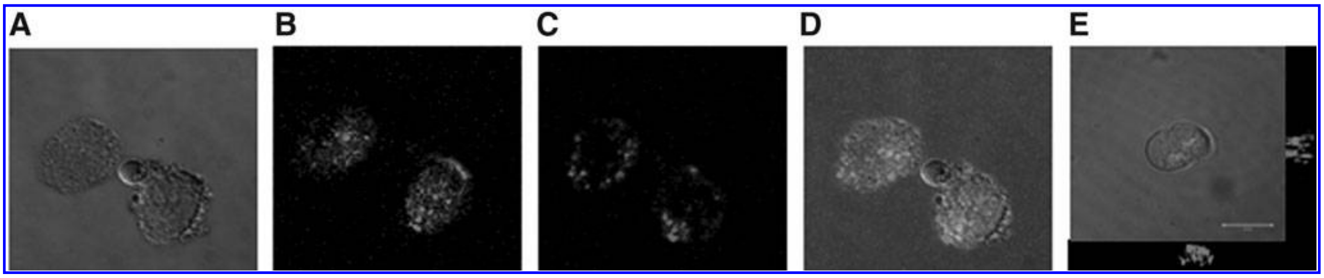
The variable total RNA content of CSF-MPs comprises ß-ACTIN, MAP2, LIMK1, and ARC mRNA as well as Staufen-2 and Argonaute-2 protein, molecules that have been associated with the regulated intercellular transport of RNA.[39,40] Our findings thereby confirm the results of a previous study that showed circulating MPs and MPs derived from cultured stem cells to carry RNA granule associated ribonucleoproteins.[23,41] Collino and associates[23] have suggested a role of ribonucleoproteins in RNA transport and stability in MPs derived from cultured mesenchymal stem cells. A similar role of ribonucleoproteins and respective mRNAs has still to be established in CSF-MPs.

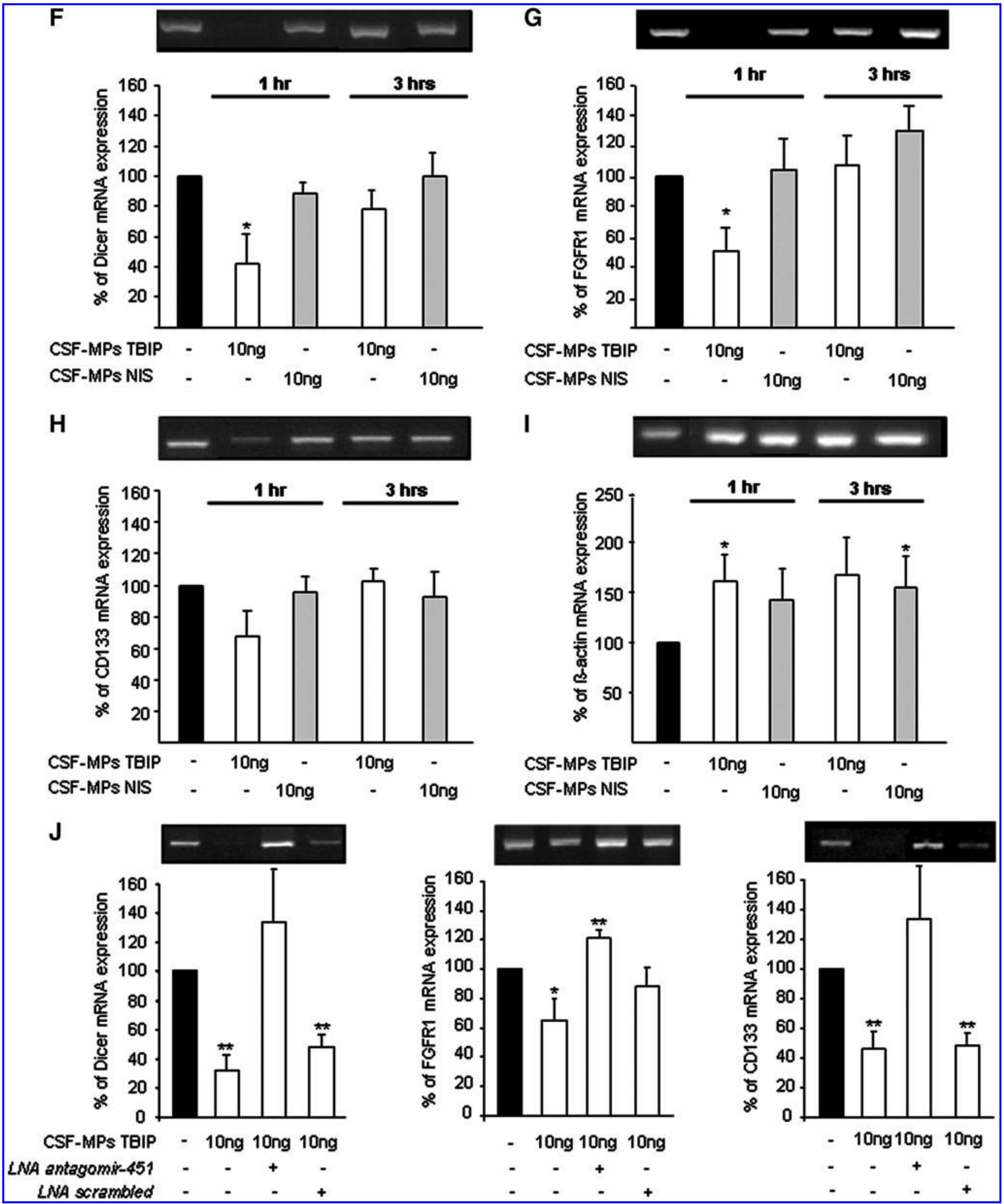
microRNAs shuttled by MPs

microRNAs constitute a high percentage of ribonucleic acids shuttled by CSF-MPs. We determined a predictive value for the presence of miR-451 in CSF-MPs of brain injured patients. miR-9 was detected with a higher frequency of occurrence in CSF-MP samples derived from NIS than from TBIPs. Intriguingly, miR-9 and miR-451 were the only two miRNA species independently identified in rat brains after TBI.[24,25] Also, other disease-dependent released MPs differ significantly in their miRNA content profile, as shown by Diehl and colleagues.[42]

miR-9 is specifically expressed in the mammalian nervous system and has been implicated in the regulation of a variety of

FIG. 4. Transfer of CSF-MP content to undifferentiated NTERAs. (**A–D**) Cedrebrospinal fluid microparticles (CSF-MPs) were stained with Sulfo-NHS-Biotin (660 nm) and RiboGreen (485 nm) and subsequently incubated with undifferentiated NTERAs. (A) Light microscopy of undifferentiated NTERas; (B) visualization of RiboGreen labeled CSF-MP derived RNA in NTERAs; (C) visualization of Cy5 positive CSF-MPs membranes in NTERAs (D) overlay of images A,B, and C; (**E**) LSM image of RiboGreen particles inside cells; amplification 40×. (**F–J**) Dicer, FGFR1, CD133, ß-ACTIN mRNA expression, respectively, in cultured NTERAs after incubation with CSF-MPs; inlays: representative agarose gel images; (I) Silencing of CSF-MP TBIP mediated regulation of Dicer, fibroblast growth factor receptor 1 (FGFR1), and CD133 mRNA expression in undifferentiated NTERAs by LNA miR-451 inhibitor (50 nM) or LNA scrambled (50 nM); inlays: representative agarose gel images; CSF-MP traumatic brain injured patients (TBIP) derived CSF-MPs (pooled: $n=5$), white bars; CSF-MP non-injured subjects (NIS) derived CSF-MP (pooled: $n=5$), grey bars; untreated cells, black bars; 10 ng: amount of CSF-MPs carrying an equivalent of 10 ng total RNA. Experiments were repeated three times with different patient or NIS pools and replicated twice per experiment; $^*p \leq 0.05$; $^{**}p \leq 0.001$.


A
B
C
D
E


F
% of Dicer mRNA expression
160
140
120
100
80
60
40
20
1 hr
3 hrs
*
CSF-MPs TBIP - 10ng - 10ng -
CSF-MPs NIS - - 10ng - 10ng
G
% of FGFR1 mRNA expression
160
140
120
100
80
60
40
20
1 hr
3 hrs
*
CSF-MPs TBIP - 10ng - 10ng -
CSF-MPs NIS - - 10ng - 10ng
H
% of CD133 mRNA expression
160
140
120
100
80
60
40
20
1 hr
3 hrs
CSF-MPs TBIP - 10ng - 10ng -
CSF-MPs NIS - - 10ng - 10ng
I
% of ß-actin mRNA expression
250
200
150
100
50
1 hr
3 hrs
*
*
CSF-MPs TBIP - 10ng - 10ng -
CSF-MPs NIS - - 10ng - 10ng
J
% of Dicer mRNA expression
160
140
120
100
80
60
40
20
0
**
**
% of FGFR1 mRNA expression
160
140
120
100
80
60
40
20
0
*
**
% of CD133 mRNA expression
160
140
120
100
80
60
40
20
0
**
**
CSF-MPs TBIP - 10ng 10ng 10ng
LNA antagomir-451 - - + -
LNA scrambled - - - +

neuronal processes such as neuron development[43] or axis formation.[44] The prevalence of miR-9 in CSF-MPs from NIS might indicate a role in the homeostasis of neuronal regulation.

miR-451 was detected in patient CSF-MPs only. A potential role of miR-451 as a damage associated regulator of gene expression was also verified by *in vitro* studies. In undifferentiated adult neuronal stem cells, expression of putative miR-451 target genes Dicer, FGFR1, and CD133 was silenced within 30 min of scratch induced cell damage (see online supplementary Fig. S5 at ftp.liebertpub.com). Down-regulation of these genes was prevented by addition of LNA miR-451 inhibitor, but not by addition of LNA scrambled.

Taking into consideration the role of miR-451 in erythropoiesis, however, it should be noted that the prevalence of miR-451 in CSF-MPs of brain injured patients might indicate the presence of vascular MPs in the CSF. As previously discussed, transfer of MPs from the nervous to the cardiovascular system and in this case *vice versa* might constitute an additional novel transborder communication channel.[18]

The emerging role of the identified miRNAs in cerebral function was further affirmed by GO analysis. The result suggests 34 of 37 neuron related cellular processes to be overrepresented in target genes regulated by the distinct set of miRNAs (see online supplementary Table S4 at ftp.liebertpub.com).

Transfer of miRNA from MPs to cultured cells

We confirmed the ability of CSF-MPs to transfer their ''RNA cargo'' to target cells *in vitro.* Attachment of plasma derived MPs, and the transfer of their miRNA to human umbilical vein endothelial cells *in vitro* has been previously demonstrated.[42] Our observations are also consistent with results showing miRNA transfer from stem cell MPs to fibroblasts,[34] from mesenchymal stem cells to tubular epithelial cells or tumor cell lines,[23,45] and from embryonic stem cell microvesicles to mouse embryonic fibroblasts.[46]

The uptake of CSF-MP content by cultured NTERAs was paralleled by the regulation of gene expression. The decrease of putative miR-451 target genes in NTERAs incubated with CSF-MPs from brain injured patients demonstrates a functional transfer of messages and a role of this miRNA in damage induced cerebral regulation. A MP mediated reprogramming of target cells has been shown for cell culture derived MPs.[5,47–50]

Conclusion

The results of the present study provide strong evidence for the shuttling and cell-to-cell transfer of brain injury associated miRNA and mRNA by CSF-MPs. The transport of genetic information in CSF and subsequent reprograming of target cells signifies a communication network that promotes signal transduction between adjacent and distal cells in the ventricular system. The uncovering of the extracellular transport of signals in the CSF will facilitate a more detailed analysis of the regulation of cerebral function and might thereby support the identification of novel diagnostic markers or even therapeutic strategies for damage associated or neurodegenerative cerebral modulations.

Acknowledgments

We thank Chris Wrighton for editing the article and helpful comments. This work was supported by a grant from the Styria government.

Author Disclosure Statement

No competing financial interests exist.

References

1. Ratajczak, J., Wysoczynski, M., Hayek, F., Janowska-Wieczorek, A., and Ratajczak, M.Z. (2006). Membrane-derived microvesicles: important and underappreciated mediators of cell-to-cell communication. Leukemia 20, 1487–1495.
2. Hunter, M.P., Ismail, N., Zhang, X., Aguda, B.D., Lee, E.J., Yu, L., Xiao, T., Schafer, J., Lee, M.L., Schmittgen, T.D., Nana-Sinkam, S.P., Jarjoura, D., and Marsh, C.B. (2008). Detection of microRNA expression in human peripheral blood microvesicles. PLoS ONE 3, e3694.
3. Garcia, B.A., Smalley, D.M., Cho, H., Shabanowitz, J., Ley, K., and Hunt, D.F. (2005). The platelet microparticle proteome. J. Proteome Res. 4, 1516–1521.
4. Gibbings, D.J., Ciaudo, C., Erhardt, M., and Voinnet, O. (2009). Multivesicular bodies associate with components of miRNA effector complexes and modulate miRNA activity. Nat. Cell Biol. 11, 1143–1149.
5. van der Vos, K.E., Balaj, L., Skog, J., and Breakefield, X.O. (2011). Brain tumor microvesicles: insights into intercellular communication in the nervous system. Cell Mol. Neurobiol. 31, 949–959.
6. Wolf, P. (1967). The nature and significance of platelet products in human plasma. Br. J. Haematol. 13, 269–288.
7. VanWijk, M.J., VanBavel, E., Sturk, A., and Nieuwland, R. (2003). Microparticles in cardiovascular diseases. Cardiovasc. Res. 59, 277–287.
8. Burnier, L., Fontana, P., Kwak, B.R., and Angelillo-Scherrer, A. (2009). Cell-derived microparticles in haemostasis and vascular medicine. Thromb. Haemost. 101, 439–451.
9. De Maio, A. (2011). Extracellular heat shock proteins, cellular export vesicles, and the Stress Observation System: a form of communication during injury, infection, and cell damage. It is never known how far a controversial finding will go! Cell Stress Chaperones 16, 235–249.
10. Marzesco, A.M., Janich, P., Wilsch-Bräuninger, M., Dubreuil, V., Langenfeld, K., Corbeil, D., and Huttner, H.B. (2005). Release of extracellular membrane particles carrying the stem cell marker prominin-1 (CD133) from neural progenitors and other epithelial cells. J. Cell Sci. 118, 2849–2858.
11. Mause, S.F., and Weber, C. (2010). Microparticles: protagonists of a novel communication network for intercellular information exchange. Circ. Res. 107, 1047–1057.
12. Simons, M., and Raposo, G. (2009). Exosomes—vesicular carriers for intercellular communication. Curr. Opin. Cell Biol. 21, 575–581.
13. Wetterberg, L., Nybom, R., Bratlid, T., Fladby, T., Olsson, B., and Wigzell, H. (2002). Micrometer-sized particles in cerebrospinal fluid (CSF) in patients with schizophrenia. Neurosci. Lett. 329, 91–95.
14. Ekelund, J., Wahlbeck, K., and Back, N. (2003). No association between micrometer-sized particles in human cerebrospinal fluid and schizophrenia. Neurosci. Lett. 349, 68–70.
15. Morel, N., Morel, O., Petit, L., Hugel, B., Cochard, J.F., Freyssinet, J.M., Sztark, F., and Dabadie, P. (2008). Generation of procoagulant microparticles in cerebrospinal fluid and peripheral blood after traumatic brain injury. J. Trauma 64, 698–704.
16. Huang, M., Hu, Y.Y., and Dong, X.Q. (2009). High concentrations of procoagulant microparticles in the cerebrospinal fluid and peripheral blood of patients with acute basal ganglia hemorrhage are associated with poor outcome. Surg. Neurol. 72, 481–489.
17. Harrington, M.G., Fonteh, A.N., Oborina, E., Liao, P., Cowan, R.P., McComb, G., Chavez, J.N., Rush, J., Biringer, R.G., and Hühmer, A.F. (2009). The morphology and biochemistry of nanostructures provide evidence for synthesis and signaling functions in human cerebrospinal fluid. Cerebrospinal Fluid Res. 6, 10.
18. Smalheiser, N.R. (2009). Do neural cells communicate with endothelial cells via secretory exosomes and microvesicles? Cardiovasc. Psychiatry Neurol. 2009, 383086.
19. Smalheiser, N.R. (2007). Exosomal transfer of proteins and RNAs at synapses in the nervous system. Biol. Direct. 2, 35.
20. Shet, A.S., Aras, O., Gupta, K., Hass, M.J., Rausch, D.J., Saba, N., Koopmeiners, L., Key, N.S., and Hebbel, R.P. (2003). Sickle blood contains tissue factor-positive microparticles derived from endothelial cells and monocytes. Blood 102, 2678–2683.
21. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504.

22. Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21, 3448–3449.
23. Collino, F., Deregibus, M.C., Bruno, S., Sterpone, L., Aghemo, G., Viltono, L., Tetta, C., and Camussi, G. (2010). Microvesicles derived from adult human bone marrow and tissue specific mesenchymal stem cells shuttle selected pattern of miRNAs. PLoS ONE 5, e11803.
24. Lei, P., Li, Y., Chen, X., Yang, S., and Zhang, J. (2009). Microarray based analysis of microRNA expression in rat cerebral cortex after traumatic brain injury. Brain Res. 1284, 191–201.
25. Redell, J.B., Liu, Y., and Dash, P.K. (2009). Traumatic brian injury alters expression of hippocampal microRNAs: Potential regulators of multiple pathophysiological processes. J. Neursci. Res. 87, 1435–1448.
26. Langlois, A., and Duval, D. (1997). Differentiation of the human NT2 cells into neurons and glia. Methods in Cell Sci. 19, 213–219.
27. Pleasure, S.J., and Lee, V.M. (1993). NTera 2 cells: a human cell line which displays characteristics expected of a human committed neuronal progenitor cell. J. Neurosci. Res. 35, 585–602.
28. Sommer, J.B., Gaul, C., Heckmann, J., Neundörfer, B., and Erbguth, F.J. (2002). Does lumbar cerebrospinal fluid reflect ventricular cerebrospinal fluid? A prospective study in patients with external ventricular drainage. Eur. Neurol. 47, 224–232.
29. Atack, J.R., May, C., Kaye, J.A., and Rapoport, S.I. (1990). Cerebrospinal fluid gradients of acetylcholinesterase and butyrylcholinesterase activity in healthy aging. Neurochem. Int. 16, 533–538.
30. Bach, F.W., Schmidt, J.F., and Faber, T. (1992). Radioimmunoassay of beta-endorphin in ventricular and lumbar cerebrospinal fluid. Clin. Chem. 38, 847–852.
31. Facchinetti, F., Petraglia, F., Cicero, S., Nappi, G., Valentini, M., and Genazzani, A.R. (1987). No gradient exists between lumbar and ventricular cerebrospinal fluid beta-endorphin. Neurosci. Lett. 77, 349–352.
32. Faull, K.F., Rafie, R., Pascoe, N., Marsh, L., and Pfefferbaum, A. (1999). N-acetylaspartic acid (NAA) and N-acetylaspartylglutamic acid (NAAG) in human ventricular, subarachnoid, and lumbar cerebrospinal fluid. Neurochem. Res. 24, 1249–1261.
33. Tarnaris, A., Toma, A.K., Chapman, M.D., Petzold, A., Keir, G., Kitchen, N.D., and Watkins, L.D. (2011). Rostrocaudal dynamics of CSF biomarkers. Neurochem. Res. 36, 528–532.
34. Skog, J., Würdinger, T., Van Rijn, S., Meijer, D.H., Gainche, L., Sena-Esteves, M., Curry, W.T., Jr., Carter, B.S., Krichevsky, A.M., and Breakefield, X.O. (2008). Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. Nat. Cell Biol. 10, 1470–1476.
35. Sadallah, S., Eken, C., and Schifferli, J.A. (2008). Erythrocyte-derived ectosomes have immunosuppressive properties. J. Leukoc. Biol. 84, 1316–1325.
36. Mrvar-Brecko, A., Sustar, V., Jansa, V., Stukelj, R., Jansa, R., Mujagic, E., Kruljc, P., Iglic, A., Hägerstrand, H., and Kralj-Iglic, V. (2010). Isolated microvesicles from peripheral blood and body fluids as observed by scanning electron microscope. Blood Cells Mol. Dis. 44, 307–312.
37. Sinauridze, E.I., Kireev, D.A., Popenko, N.Y., Pichugin, A.V., Panteleev, M.A., Krymskaya, O.V., and Ataullakhanov, F.I. (2007). Platelet microparticle membranes have 50- to 100-fold higher specific procoagulant activity than activated platelets. Thromb. Haemost. 97, 425–434.
38. Pluskota, E., Woody, N.M., Szpak, D., Ballantyne, C.M., Soloviev, D.A., Simon, D.I., and Plow, E.F. (2008). Expression, activation, and function of integrin alphaMbeta2 (Mac-1) on neutrophil-derived microparticles. Blood 112, 2327–2335.
39. Schratt, G.M., Tuebing, F., Nigh, E.A., Kane, C.G., Sabatini, M.E., Kiebler, M., and Greenberg, M.E. (2006). A brain-specific microRNA regulates dendritic spine development. Nature 439, 283–289.
40. Kiebler, M.A., and Bassell, G.J. (2006). Neuronal RNA granules: movers and makers. Neuron 51, 685–690.
41. Li, L., Zhu, D., Huang, L., Zhang, J., Bian, Z., Chen, X., Liu, Y., Zhang, C.Y., and Zen, K. (2012). Argonaute 2 complexes selectively protect the circulating microRNAs in cell-secreted microvesicles. PLoS ONE 7, e46957.
42. Diehl, P., Fricke, A., Sander, L., Stamm, J., Bassler, N., Htun, N., Ziemann, M., Helbing, T., El-Osta, A., Jowett, J.B., and Peter, K. (2012). Microparticles: major transport vehicles for distinct microRNAs in circulation. Cardiovasc. Res. 93, 633–644.
43. Shibata, M., Nakao, H., Kiyonari, H., Abe, T., and Aizawa, S. (2011). MicroRNA-9 regulates neurogenesis in mouse telencephalon by targeting multiple transcription factors. J. Neurosci. 31, 3407–3422.
44. Bonev, B., Pisco, A., and Papalopulu, N. (2011). MicroRNA-9 reveals regional diversity of neural progenitors along the anterior-posterior axis. Dev. Cell 20, 19–32.
45. Bruno, S., Collino, F., Deregibus, M.C., Grange, C., Tetta, C., and Camussi, G. (2013). Microvesicles derived from human bone marrow mesenchymal stem cells inhibit tumor growth. Stem Cells Dev. 22, 758–771.
46. Yuan, A., Farber, E.L., Rapoport, A.L., Tejada, D., Deniskin, R., Akhmedov, N.B., and Farber, D.B. (2009). Transfer of microRNAs by embryonic stem cell microvesicles. PLoS ONE 4, e4722.
47. Ratajczak, J., Miekus, K., Kucia, M., Zhang, J., Reca, R., Dvorak, P., and Ratajczak, M.Z. (2006). Embryonic stem cell-derived microvesicles reprogram hematopoietic progenitors: evidence for horizontal transfer of mRNA and protein delivery. Leukemia 20, 847–856.
48. Aliotta, J.M., Pereira, M., Johnson, K.W., de Paz, N., Dooner, M.S., Puente, N., Ayala, C., Brilliant, K., Berz, D., Lee, D., Ramratnam, B., McMillan, P.N., Hixson, D.C., Josic, D., and Quesenberry, P.J. (2010). Microvesicle entry into marrow cells mediates tissue-specific changes in mRNA by direct delivery of mRNA and induction of transcription. Exp. Hematol. 38, 233–245.
49. Waldenstrom, A., Gennebäck, N., Hellman, U., and Ronquist, G. (2012). Cardiomyocyte microvesicles contain DNA/RNA and convey biological messages to target cells. PLoS ONE 7, e34653.
50. Cantaluppi, V., Gatti, S., Medica, D., Figliolini, F., Bruno, S., Deregibus, M.C., Sordi, A., Biancone, L., Tetta, C., and Camussi, G. (2012). Microvesicles derived from endothelial progenitor cells protect the kidney from ischemia-reperfusion injury by microRNA-dependent reprogramming of resident renal cells. Kidney Int. 82, 412–427.

Address correspondence to:
Ute Schäfer, PhD
Research Unit for Experimental Neurotraumatology
Medical University of Graz
Auenbruggerplatz 2^2
8036 Graz
Austria

E-mail: ute.schaefer@medunigraz.at

BMC Systems Biology

RESEARCH Open Access

Integrative omics analysis. A study based on *Plasmodium falciparum* mRNA and protein data

Oana A Tomescu[1,2], Diethard Mattanovich[3,4], Gerhard G Thallinger[1,2,5*]

From High-Throughput Omics and Data Integration Workshop
Barcelona, Spain. 13-15 February 2013

Abstract

Background: Technological improvements have shifted the focus from data generation to data analysis. The availability of large amounts of data from transcriptomics, protemics and metabolomics experiments raise new questions concerning suitable integrative analysis methods. We compare three integrative analysis techniques (co-inertia analysis, generalized singular value decomposition and integrative biclustering) by applying them to gene and protein abundance data from the six life cycle stages of *Plasmodium falciparum*. Co-inertia analysis is an analysis method used to visualize and explore gene and protein data. The generalized singular value decomposition has shown its potential in the analysis of two transcriptome data sets. Integrative Biclustering applies biclustering to gene and protein data.

Results: Using CIA, we visualize the six life cycle stages of *Plasmodium falciparum*, as well as GO terms in a 2D plane and interpret the spatial configuration. With GSVD, we decompose the transcriptomic and proteomic data sets into matrices with biologically meaningful interpretations and explore the processes captured by the data sets. IBC identifies groups of genes, proteins, GO Terms and life cycle stages of *Plasmodium falciparum*. We show method-specific results as well as a network view of the life cycle stages based on the results common to all three methods. Additionally, by combining the results of the three methods, we create a three-fold validated network of life cycle stage specific GO terms: Sporozoites are associated with transcription and transport; merozoites with entry into host cell as well as biosynthetic and metabolic processes; rings with oxidation-reduction processes; trophozoites with glycolysis and energy production; schizonts with antigenic variation and immune response; gametocyctes with DNA packaging and mitochondrial transport. Furthermore, the network connectivity underlines the separation of the intraerythrocytic cycle from the gametocyte and sporozoite stages.

Conclusion: Using integrative analysis techniques, we can integrate knowledge from different levels and obtain a wider view of the system under study. The overlap between method-specific and common results is considerable, even if the basic mathematical assumptions are very different. The three-fold validated network of life cycle stage characteristics of *Plasmodium falciparum* could identify a large amount of the known associations from literature in only one study.

Background

Continuous technological improvements facilitate the availability of large amounts of omics data, resulting from the simultaneous characterization on different levels (genome, transcriptome, proteome and metabolome) of an organism or an experimental condition. Regulatory mechanisms captured in this way provide a complex multi-level view of the system under study. In order to exploit the measured data to the maximum, one has to integrate all available data sets into a single analysis framework. Methods that apply analysis techniques simultaneously to more than one data set are called integrative analysis methods. The data sets can characterize one organism on different levels [1], or they can be measured on the same omics level but on different organisms/platforms [2,3]. Here we focus on the first scenario.

* Correspondence: gerhard.thallinger@tugraz.at
[1]Core Facility Bioinformatics, Austrian Centre of Industrial Biotechnology, Graz, Austria
Full list of author information is available at the end of the article

Integrative analysis methods provide a deeper understanding of the system under study through the meaningful combination of multi-level omics data. The integrated omics data differ from study to study. There are studies that integrate, for example, gene expression and methylation data [4], somatic mutations, copy number and gene expression data [5], chromatin maps and gene expression profiles [6], genotypic variation at DNA level and gene expression data [7], CHIP-seq and RNA-seq data [8], transcriptomics and proteomics data [1,9,10]. In this study we apply integrative analysis to transcriptomics and proteomics data.

With transcriptomic and proteomic data, most analysis techniques are based on the direct correlation between transcripts and proteins. Cox and colleagues [10] present different approaches based on correlation and clustering. Other correlation-based studies have also been performed in [9,11-16]. Statistical methods based on correlations are presented in [17,18]. The premise of a direct correlation between transcripts and proteins is not valid in eucaryotic organisms, due to post-transcriptional and post-translational regulation [1,19]. Other approaches are based on network analysis [20,21] and statistical methods such as analysis of variation, clustering and gene set enrichment [22-24]. Piruzian *et al.* [25] revealed similarities in regulation at transcriptomic and proteomic levels and identified potential key transcription factors and new signaling pathways for psoriasis using a network based approach, which employed overconnection analysis, hidden node analysis and rank aggregation. Perco *et al.* [26] integrated transcriptomics and proteomics on the level of protein interaction networks. They started with the modest overlap between the data sets, which increased substantially on the level of protein interaction networks and in this way, amplified the joint functional interpretation of the omics data sets. In a study by Hahne and colleagues [22] analysis of variation, k-means clustering and functional annotation were applied to transcriptome and proteome data from salt-stressed *B. subtilis* cells. They showed a well-coordinated induction of gene expression and changes of the protein levels as the result of a severe salt shock. Verhoef *et al.* characterized the changes associated with ρ-hydroxybenzoate production in the engineered *P. putida* strain S12, integrating genes and proteins as well as cluster and pathway analysis [23]. In [24], Takemasa *et al.* applied gene ontology analysis (GO) to transcriptome and proteome data from human colorectal cancer samples, which led to a better understanding of functional inference at the physiological level and to potential drug targets. Other integrative approaches can be found in [27-29] for omics data in general and in [19] for transcriptome and proteome data in particular.

In this article, we focus on the comparison of three integrative analysis techniques of mRNA and protein abundance data. We selected methods meeting the following criteria: (i) they are based on a clear mathematical formulation, (ii) they are as different as possible from one another and, the most important argument, (iii) they allow the analysis of all measured data (not limited to pairs of genes and proteins). Based on these criteria we have chosen: Co-inertia analysis (CIA), which is an integrative analysis method used to visualize and explore gene and protein data [1,30], Generalized singular value decomposition (GSVD), which has shown its potential in the analysis of two transcriptome data sets [3] and Integrative biclustering (IBC), which applies biclustering to gene and protein data [31].

We compare CIA, GSVD and IBC by applying them to mRNA and protein abundance data from a study of *Plasmodium falciparum*, [9], the parasite causing malaria in humans. The data in this study was gathered from samples for the six life cycle stages: merozoite, ring, trophozoite, schizont, gametocyte and sporozoite. For the comparison, we add additional information in the form of GO [32] terms for biological processes.

Using CIA, we visualize the six life cycle stages and GO terms in a 2D plane and interpret the spatial configuration. With GSVD, we decompose the data sets into matrices with biologically meaningful interpretations and explore the processes captured by the data sets. IBC identifies groups of genes, proteins, GO terms and life cycle stages revealing functional modules of *P. falciparum*.

We compare the results of the three integrative analysis methods based on the association of GO terms to the six life cycle stages and show common as well as method-specific results. The common results are presented in form of a three-fold validated network view of the biological processes activated in each life cycle stage. To the best of our knowledge such a complete, GO terms based, characterization of *P. falciparum* was not published before.

Methods

Data set

We analyse a publicly available data set containing mRNA and protein abundance data from the six life cycle stages of *P. falciparum* [9,33]. Microarray [33] and proteomic analyses [9] were carried out on *P. falciparum* clone 3D7. Gene expression levels were measured with a custom oligonucleotide array and computed with the match-only integral algorithm (MOID). Proteins were detected by multidimensional protein identification technology (MudPIT), and protein abundance was estimated by the number of MS/MS spectra identified per protein. In total, 4294 genes and 2903 proteins were measured in all six life cycle stages. For the analysis, we created a matrix for each data set where the genes and proteins are represented as rows and the life cycle stages as columns.

Additionally, GO [32] information on biological processes in *P. falciparum* were employed. We used the R [34] packages *org.Pf.plasmo.db* [35] and *GO.db* [36], which provide *P. falciparum* specific mappings of genes to GO terms as well as additional information on GO terms. Based on these two annotation databases, 3283 of 4294 genes and 2491 of 2903 proteins were associated with 614 GO terms. For each data set, a GO matrix with the same number of rows as the corresponding expression data set was created. The columns of the GO matrix hold data describing the gene/protein affiliation to a certain GO term. If a gene/protein is associated with that GO term, the strength of the affiliation is computed as the ratio between 1 and the total number of genes/proteins associated with the GO term. CIA and IBC use directly the computed GO matrix. GSVD performs a GSE analysis based on *org.Pf.plasmo.db* and *GO.db*. In this way, we make sure that all three methods are applied to the same data sets. Additional file 1 contains the GO terms with their GO names used in this study.

Our study comprises the analysis of four data sets: mRNA abundance data, protein abundance data, a GO matrix of mRNAs and a GO matrix of proteins. mRNA and protein abundances were computed with different algorithms, requiring a columnar z-score normalization of each data set. The GO matrices were computed based on the number of genes/proteins belonging to a particular GO term, which resulted in equal ranges of the entries in the matrices. Afterwards, they were joined and z-transformed. Columns (GO terms) that included only entries equal to zero (none of the associated genes or proteins was measured in the data set) were deleted before the normalization. After deletion, 614 GO terms were available for further analysis.

CIA

CIA was introduced by Dolèdec and Chessel [30] as an extension of Tucker's inter-battery method [37] for the study of species-environment relationships of ecology data. CIA was applied to genome and *agr* groups data of *S. aureus* [38] and to physico-chemical properties of amino-acids and the amino-acid composition of *E.coli* proteins [39]. Culhane *et al.* [40] applied this method to two gene expression data sets, and Fagan *et al.* [1] used it as an integrative analysis method for gene and protein data.

CIA is a multivariate analysis method that identifies relationships between two data sets by maximizing the covariance between them. CIA starts by performing a multivariate analysis like principal component analysis (PCA) [41], non-symmetrical correspondence analysis (NSC) [42] or correspondence analysis (CA) [43] on each individual data set. The produced results are a set of principal axes that maximize the projected variability (inertia) of each data set independently. Each set of axes spans a new multidimensional gene and protein space. Based on the computed axes, CIA identifies one axis in each new multidimensional space on which the projected data sets present maximal covariance and simultaneously maximal standard deviations. Thereby, CIA maximizes the covariance between the two data sets. Global correlation or co-structure between the data sets is measured by the RV coefficient [44]. For mathematical details on CIA, please refer to [30]. CIA computation steps are summarized in a concise flowchart in Figure 1. CIA is available in the R packages *made4* [45] and *ade4* [46].

CIA has two major advantages: It can be applied to data sets with considerable more variables (genes and proteins) than samples (life cycle stages), and the variables in the two data sets do not have to match one another.

Additional information such as GO annotations can be superimposed on the CIA plots. This overlay was already done for CA [47], and is also possible for CIA [1]. GO term projections are obtained by first normalizing the two GO matrices in the same way as the expression data sets and then multiplying them by the weights of the genes/proteins resulting from the NSC, followed by CIA analyses. The projection scores computed in this way show GO term associated with the measured genes/proteins in relation to the life cycle stages.

GSVD

The GSVD was developed as an extension of the singular value decomposition (SVD) that was already used directly as an analysis method [48,49] and indirectly as part of a PCA [50,51]. In a study by Golub *et al.* [52], GSVD was used as a comparative analysis method [3] for two gene expression data sets of cell cycle data from yeast and humans.

GSVD is based on the joint decomposition of both data sets as shown in equations (1) and (2):

$$G = U_1 \Sigma_1 X^{-1} \tag{1}$$

$$P = U_2 \Sigma_2 X^{-1}. \tag{2}$$

Matrices G and P contain the gene and protein abundance data. The rows of the common matrix X^{-1} are named genelets. In [3] it was shown that these genelets can be regarded as processes captured by both data sets. The genelets are expressed only in the corresponding arraylets (columns of U_1 and U_2) with a relative significance measured by the generalized eigenvalues $(\sigma_{1,i}, \sigma_{2,i})$ from the diagonals of Σ_1 and Σ_2. The relative significance of a genelet in the gene data set relative to the protein data set is measured by an antisymmetric angular distance calculated as shown in equation (3):

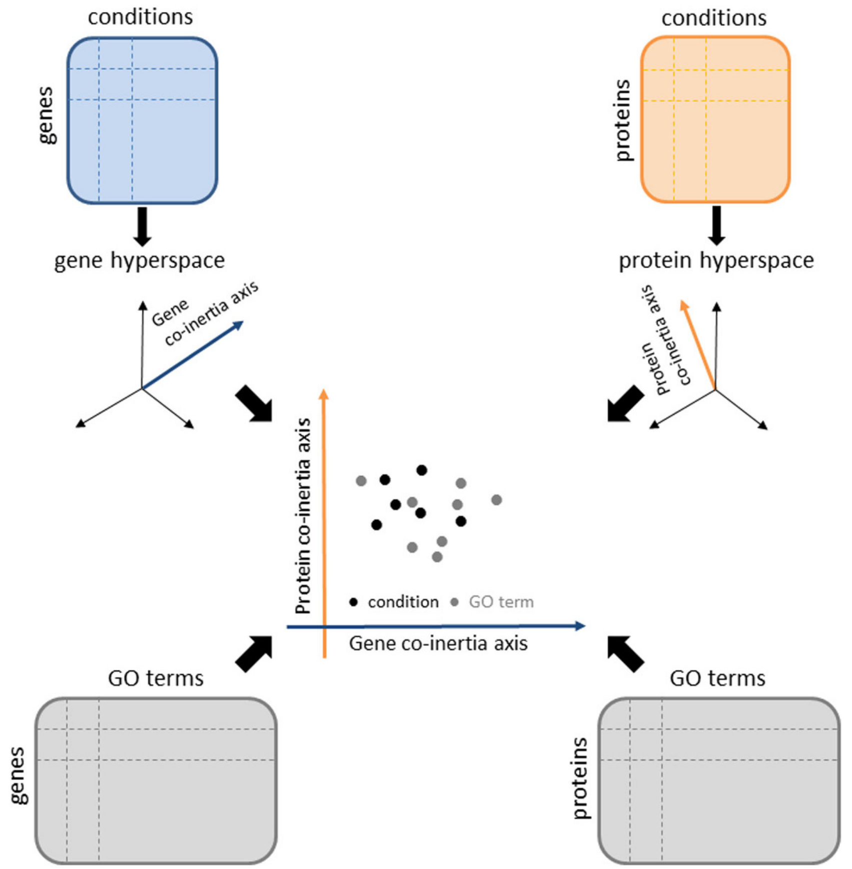


Figure 1 Flowchart of CIA. The gene expression matrix (blue) contains the genes in the rows and the conditions in the columns. The protein expression matrix (orange) contains the proteins in the rows and the conditions in the columns. The conditions in our study correspond to the life cycle stages of *P. falciparum*. Genes and proteins annotated to the considered GO terms are gathered in a separate matrix (gray). The gene and the protein expression matrices are transformed into a new hyperspace. The pair maximizing the covariance is computed from the axes spanning the two hyperspaces. These two axes (gene and proteins co-inertia axes) span a new space where the conditions (life cycle stages) can be plotted. Additionally, GO terms can be projected into the CIA space and visualized together with the life cycle stages. Figure adapted from [30].

$$\theta_i = \arctan\left(\frac{\sigma_{1,i}}{\sigma_{2,i}}\right) - \frac{\pi}{4}. \tag{3}$$

An angular distance between -$\pi/4$ and -$\pi/8$ represents a high significance of the i^{th} genelet in the second data set relative to the first data set. If the value of the angular distance ranges between $\pi/8$ and $\pi/4$, then the i^{th} genelet has a high significance in the first data set relative to the second data set. The i^{th} genelet shows equal significance in both data sets if the angular distance ranges between -$\pi/8$ and $\pi/8$ (see equation (4)). In our study, the first matrix contains mRNA abundance data, the second matrix protein abundance data and significances are assigned as follows:

$$\theta_i \in \begin{cases} [-\pi/4, -\pi/8] & \text{protein space} \\ [-\pi/8, \pi/8] & \text{gene and protein space} \\ [\pi/8, -\pi/4] & \text{gene space.} \end{cases} \tag{4}$$

A summary of the computation flow is shown in a block diagram in Figure 2.

Alter *et al.* [3] used a Mathematica implementation of a numerically robust GSVD algorithm based on [52,53], which we reimplemented in R.

In order to discover the processes captured by the genelets, a restrictive gene set enrichment (GSE) analysis is performed on 50% of the genes and/or proteins showing the highest absolute values in the corresponding arraylets. The GSE analysis is performed with the R package *GOstats* [54], which computes the statistically significantly enriched GO terms based on the hypergeometrical distribution.

IBC

The basic idea of biclustering (co-clustering or two-way clustering) was presented in [55], but it took almost thirty years until the method was applied to gene expression data [56]. In the last two decades, biclustering has

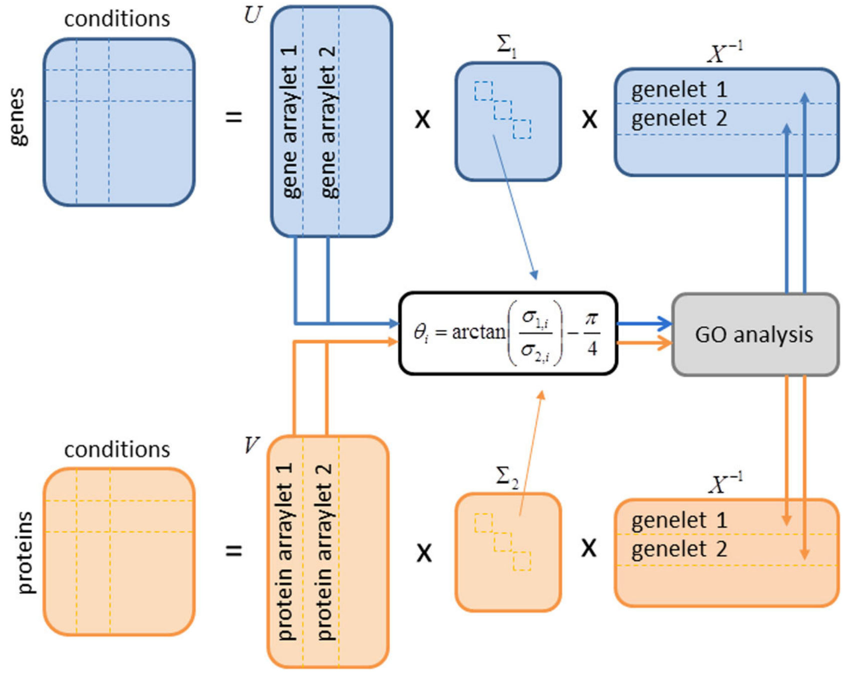


Figure 2 Flowchart of GSVD. The gene expression matrix (blue) contains the genes in the rows and the conditions in the columns. The protein expression matrix (orange) contains the proteins in the rows and the conditions in the columns. The conditions in our study correspond to the life cycle stages of *P. falciparum*. Genes and proteins annotated to the considered GO terms are gathered in a separate matrix (gray). The gene and the protein expression matrices are each decomposed in three matrices according to equations (1) and (2). The matrices *U* and *V* contain the arraylets, which encode for the expression of genes and proteins in the corresponding genelets X^{-1}, which represent the cellular state in the measurement conditions. According to the angular distance θ_i, which is computed from the generalized eigenvalues $\sigma_{1,i}$ and $\sigma_{2,i}$, a restricted GSE analysis is performed on the genes and/or proteins with the absolute (from a mathematical point of view) highest values in the arraylets in order to assign GO terms to the corresponding genelets. Figure adapted from [3].

become more and more popular [57-59]. In contrast to clustering, where either rows *or* columns are clustered, biclustering performs clustering of rows *and* columns simultaneously. The members of the obtained biclusters are as similar to one another and as different from the other biclusters as possible. Figure 3 presents how mRNA abundances, protein expression and GO terms are assembled to a complete data set and how the resulting biclusters could look like.

There are four types of possible biclusters as reviewed in [31,60,61]. Biclusters can have (i) equal values over rows and columns as well as (ii) equal values over rows or columns. They can also have (iii) coherent values, which means that each column or row can be computed by adding or multiplying a constant to the previous column or row. The forth type of bicluster has (iv) coherent evolutions, which means that the exact value of a matrix entry is not important, but whether the values increase or decrease over rows or columns. The biclustering algorithm used here is included in the R package *biclust* [62].

The types of computed biclusters vary. There are single biclusters where only one bicluster is found in the whole data set as well as exclusive rows and/or exclusive columns biclusters. Non-overlapping and non-exclusive biclusters can also be computed. The fifth type is the arbitrarily positioned overlapping biclusters. Graphical representations of the different categories of biclusters can be found in [31].

Most of the biclustering algorithms implemented depend on the starting point of the search and thus may lead to different results in consecutive runs. Additionally, biclustering does not result in a perfect data separation, as overlapping biclusters are possible. As a remedy, the *biclust* package provides a robust method that delivers stable and reliable results. This function includes the repeated use of one algorithm in combination with several parameter settings and/or subsamples of the data. A modified version of the Jaccard index is used for the combination of the resulting biclusters, which in case of two biclusters takes into account the fraction of row-columns combinations in both biclusters to all row-column combinations. For detailed mathematical definitions, please refer to [31].

Analogous to integrative clustering, we define integrative biclustering as the biclustering of two or more data sets. Integrative clustering was already applied to copy

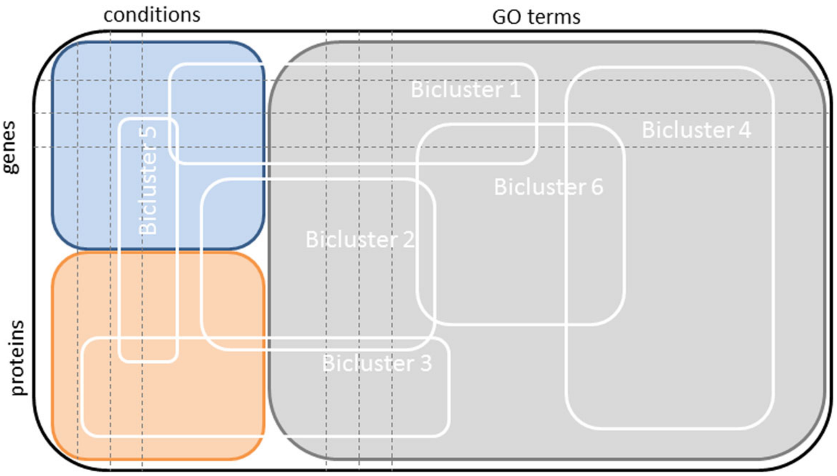


Figure 3 Block diagram of IBC. The gene expression matrix (blue) contains the genes in the rows and the conditions in the columns. The protein expression matrix (orange) contains the proteins in the rows and the conditions in the columns. The conditions in our study correspond to the life cycle stages of *P. falciparum*. Genes and proteins annotated to the considered GO terms are gathered in a separate matrix (gray). The three matrices are combined to a new matrix, which is subjected to biclustering. The resulting biclusters can include genes, proteins, conditions, GO terms or any combination of these.

number and gene expression data in order to identify novel breast tumours subgroups [63]. Mo and colleagues [5] describe integrative clustering of genomic, epigenomic and transcriptomic profiling.

Integrative biclustering was applied to gene expression, protein interaction, growth phenotype and transcription factor binding data in [64] in order to reveal modularity and organization in the yeast molecular network. We apply integrative biclustering to a matrix consisting of the mRNA and protein abundance data and of the corresponding GO matrices. The genes and the proteins are represented by rows, whereas the samples and the GO terms by columns. Before biclustering can be carried out, discretization is necessary. Here the built-in function *discretize* of the R package *biclust* [31] was used. After appropriate processing, the result of IBC was loaded into Cytoscape [65] to obtain a network view of the associations.

Results and discussion

Results of each analysis method can be divided into method-specific associations and general associations. The general associations are used to compute results common to all three methods.

CIA

With CIA we visualize the six life cycle stages in the gene and protein space (Figure 4). We observe that the co-inertia × axis separates the intraerythrocytic cycle stages (trophozoite, ring, schizont, merozoite) from gametocytes and sporozoites. In the erythrocytes, the cycle begins with the ring stage, followed by the trophozoite stage. Trophozoites mature into schizonts, which cause the rupture of blood cells resulting in the release of merozoites. In Figure 1, this exact sequence within the intraerythrocytic cycle can be observed. The sporozoites are the sexual stage of the mosquito and will be released in the blood stream of the infected organism. The ring stage can develop into a gametocyte and can be ingested by a mosquito. In addition to the life cycle stages, GO terms can also be represented through projections in the CIA plot (Figure 5).

A mapping between numbers and GO terms can be found in Additional file 1. Detailed representations of the division limits for the specific and general associations are shown in Additional file 4.

General associations

General associations resulting from CIA are distributed as follows: In gene space, GO terms in the first trigonometric quadrant are associated with trophozoites, GO terms in the second quadrant with gametocytes and GO terms in the third quadrant with sporozoites. GO terms in the first and forth quadrant, which were not identified as specific for trophozoits are associated with rings, schizonts and merozoits. Due to the proximity of stages in the CIA gene space, a more specific distribution to each stage is not possible. In protein space the associations are produced as follows: For gametocytes and sporozoites, we follow the same criteria as in gene space. For the distribution of GO terms to trophozoits, rings, schizonts and merozoits, we divide the first and forth quadrant in three sectors. GO terms that form angles of at least 30 degrees with the positive co-inertia × axis are associated with trophzoits. GO terms with an angle between -10 and 30 degrees are associated with rings and schizonts.

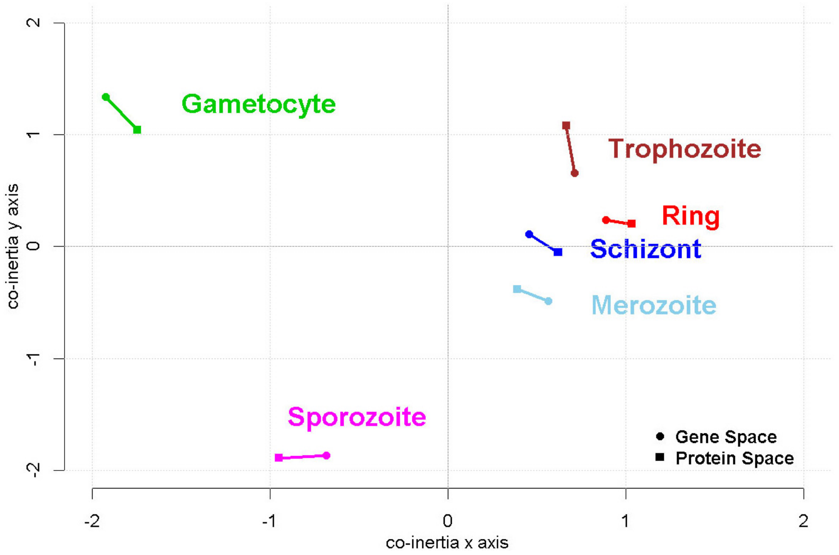


Figure 4 Co-inertia analysis - results. CIA offers the possibility to visualize the gene and protein space projections of the six life cycle stages of *P. falciparum* in one plot. The projection in gene space are represented by circles and in the protein space by squares. For each life cycle stage, the two corresponding projections are connected through a line. We observe that the y axis separates the intraerythrocytic cycle from the stages gametocyte and sporozoite.

GO terms with an angle wider than -10 degrees are associated with merozoits. Since GSVD and IBC discover associations only in common space (gene and protein space), the CIA associations for each life cycle are computed as the set union of the associations in gene and the associations in protein space. These general associations are shown in Additional files 2 and 3. The overlap between the general association in gene and protein space are shown in Figure 6.

Method-specific associations

In addition to the general results, method-specific associations of GO terms with life cycle stages are observed. For these associations, the direction of the projected GO terms is considered. From the general associations, we take those GO terms that have a distance of at least 0.1 to the origin of the coordinate systems. An exception is made for gametocytes in protein space. A threshold of 0.05 is more appropriate here due to the spacial

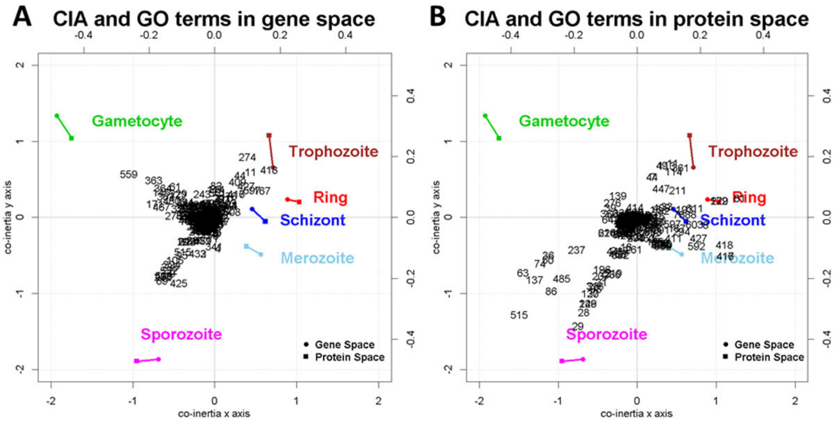


Figure 5 Co-inertia analysis and GO terms - results. In addition to the life cycle stages, GO terms can also be projected into the CIA plot. A) projections of the GO terms in gene space and B) projections of the GO terms in the protein space. Each GO term is represented by a number. Please note that the life cycle stages and the GO terms are plotted on different scales. The lower and left axes represent the life cycle stages and the upper and right axes represent for the GO terms. In gene space we observe a clear projection of the GO terms in the direction of gametocytes and sporozoites. In protein space, GO terms are projected clearly in direction of sporozoites and the intraerythrocytic cycle.

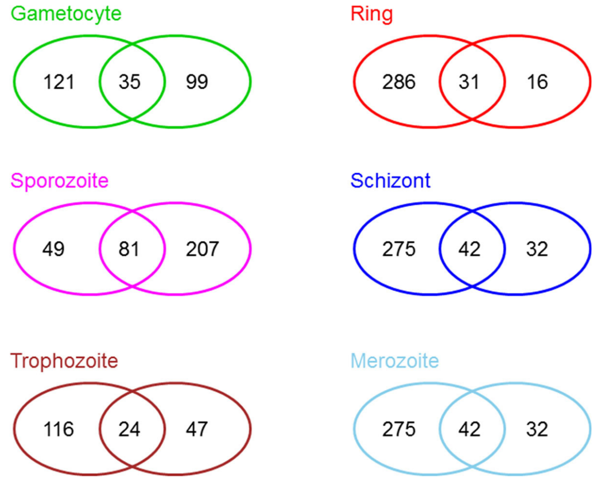


Figure 6 CIA general associations - overlap between gene and protein space. For each life cycle stage, the left ellipse shows the number of general GO term associations in gene space whereas the right ellipse shows the number of general GO term associations in protein space. The amount of identical GO terms is shown in the overlapping region of the ellipses. In general, more GO term associations emerge form gene space than from protein space. Two exceptions can be observed: the sporozoite stage, where more associations are found in protein space and the gametocyte stage, where a similar number of associations is found in each space. These tendencies can be also observed in Figure 7.

distribution of GO terms relative to the origin. These considerations result in GO term associations with gametocyte, trophozoite and sporozoite in gene space. Details are presented in Table 1 and Additional file 5. In the protein space, clear GO term associations with gametocyte, sporozoite, trophozoite and merozoite stages are found (Table 2 and Additional file 6). Additional file 6 also includes associations with the stages ring and schizont.

According to Figure 5, where for readability reasons GO terms are represented by numbers from 1 to 614, some of the most remarkable associations in gene space are: GO:0006071 *glycerol metabolic process* (559) and GO:0002720 *positive regulation of cytokine production* (363) for gametocytes; GO:0006101 *citrate metabolic process* (425) and GO:0016255 *attachment of GPI anchor to protein* (89) for sporozoites; GO:0006591 *ornithine metabolic process* (274) and GO:0006094 *gluconeogenesis* (418) for trophozoites. In protein space we observe: GO:0045454 *cell redox homeostasis* (139) and GO:0044262 *cellular carbohydrate metabolic process* (276) for gametocytes; GO:0006928 *cellular component movement* (551) and GO:0015991 *ATP hydrolysis coupled proton transport* (29) for sporozoites; GO:0006412 *translation* (11) and GO:0019538: *protein metabolic process* (361) for trophozoites; GO:0006334 *nucleosome assembly* (61) and GO:0050776 *regulation of immune response* (8) for ring and schizonts; GO:0042594 *response to starvation* (592), GO:0000045 *autophagic vacuole assembly* (416) and GO:0002253 *activation of immune response* (417) for merozoites.

The overlap between the projections in gene and protein space is modest. Three GO terms were projected in the direction of trophozoites in gene and in protein space: GO:0006412 *translation* (11), GO:0006414 *translational elongation* (44) and GO:0044257 *cellular protein metabolic process* (114).

GSVD

As the final step of the GSVD, a restrictive gene set enrichment analysis (GSE) is performed. The type of performed GSE analysis is based on the angular distance that encodes for each life cycle stage the significance of the gene set relative to the protein set. If the angular distances are between $-\frac{\pi}{8}$ and $\frac{\pi}{8}$, then the gene and protein data sets are of equal significance, and the GSE is conducted in the common space. The common space is defined by the gene and the protein data set. This is the case for all life cycle stages (Figure 7). If we compare the angular distance with zero, we obtain a separation of the intraerythrocytic cycle (angular distances bigger than zero) from other stages (angular distances smaller than zero). The restrictive GSE performs a GSE for each life cycle stage on 50% of the genes and proteins that present the highest absolute values in the corresponding arraylets.

General associations

All resulting GO terms having a p value smaller than 0.05 are considered to be general associations. These GO terms are shown in Additional file 7.

Method-specific associations

The method-specific GO terms are a subset of the general associations consisting of the top 15 GO terms, with the smallest p values. The method-specific associations are presented in Tables 3 and 4 and in Additional file 8. Biologically relevant associations include: GO:0051805/GO:0051807 *evasion or tolerance if immune/defense response of other organism involved in symbiotic interaction*, GO:0051832 *avoidance or defenses of other organism involved in symbiotic interaction*, and GO:0052173 *response to defenses (immune response) of other organism involved in symbiotic interaction* for trophozoites and schizonts. The other stages are associated with more general GO terms such as GO:0044237 *cellular metabolic process*, GO:0019538 *protein metabolic process* and GO:0046474 *glycerophospholipid biosynthetic process*.

IBC

The IBC results include two types of biclusters: (i) biclusters containing genes, proteins, GO terms and life

Table 1 CIA specific GO term association in gene space to the gametocyte stage.

CIA: Gametocyte in gene space	
61	GO:0006334: nucleosome assembly
145	GO:0006072: glycerol-3-phosphate metabolic process
171	GO:0006465: signal peptide processing
362	GO:0007131: reciprocal meiotic recombination
363	GO:0002720: positive regulation of cytokine production involved in immune response
364	GO:0006359: regulation of transcription from RNA polymerase III promoter
467	GO:0051604: protein maturation
480	GO:0001819: positive regulation of cytokine production
559	GO:0006071: glycerol metabolic process

In this table GO term association in gene space to the life cycle stage gametocyte are presented. The numbers in the left column correspond to the numbers in graphic A of Figure 5

cycle conditions and (ii) biclusters containing genes, proteins and GO terms. Since we are interested in GO terms associations with life cycle stages, we will use only the first type of biclusters for further analysis. If a GO term is in the same bicluster as a life cycle stage, this GO term is associated with that life cycle stage. If there are more life cycle stages in a bicluster, the GO terms are associated with all these life cycle stages. If a life cycle stage is included in more that one bicluster, GO terms from all biclusters are associated with that life cycle stage. IBC discovered 20 biclusters and 9 of them contained life cycle stages and GO terms. A network view of the results is shown in Figure 8.

General and method-specific associations

Since a life cycle stage is either included in a bicluster or not and as a consequence is either associated to a GO term or not, it is not possible to distinguish between general and method-specific associations. Figure 8 shows a vast amount of genes (in orange), proteins (in light blue), GO terms (in yellow) and the six life cycle stages: gametocyte (in green), sporozoite (in pink), trophozoite (in brown), ring (in red), schizont (in dark blue) and merozoite (in light blue). The different biclusters resulting from the analysis can be identified through the colour of their edges. The exact associations with the life cycle stages are shown in Additional file 9.

Table 2 CIA specific GO term association in protein space to the gametocyte stage.

CIA: Gametocyte in protein space	
1	GO:0009405: pathogenesis
39	GO:0007165: signal transduction
64	GO:0007155: cell adhesion
139	GO:0045454: cell redox homeostasis
276	GO:0044262: cellular carbohydrate metabolic process
366	GO:0006103: 2-oxoglutarate metabolic process

In this table GO term association in protein space to the life cycle stage gametocyte are presented. The numbers in the left column correspond to the numbers in graphic B of Figure 5

Common results

In this section, we present GO associations observed in all three methods. The common associations are shown in Figure 9. These associations are based on gene as well as protein information and are therefore considered to be in the common space. The GO associations computed with R were converted into a compatible format and loaded into Cytoscape. We observe here that the gametocytes are linked to the rest of the network through only one general GO term, GO:0009987 *cellular process*. The sporozoite stage is also loosely connected to the network through two GO terms, GO:0009056 *catabolic process* and GO:0009116 *nucleoside metabolic process*. The intraerythrocytic cycle, composed of trophozite, ring, schizont and merozoite are highly interconnected. The merozoite stage presents a high number of associations with specific GO terms such as GO:0030260 *entry into host cell* and GO:0044409 *entry into host.* Trophozoites are associated with a small number of GO terms, including GO:0050896 *response to stimulus*, GO:0006096 *glycolysis*, GO:0006006 *glucose metabolic process* and GO:0006091 *generation of precursor metabolites and energy*. The stages schizont and ring are connected through the GO terms GO:0006955 *immune response*, GO:0050776 *regulation of immune response*, GO:0006325 *chromatin organization* and GO:0006091 *generation of precursor metabolites and energy*. It is also interesting to see that merozoits and schizonts are linked only through the GO term GO:0009116 *nucleoside metabolic process.*

Relative proportions of common and methods-specific results

In the case of CIA, one can observe a high overlap between the common results and the CIA specific GO terms associations: 8 GO terms (GO:0005975 *carbohydrate metabolic process*, GO:0006644 *phospholipid metabolic process*, GO:0008654 *phospholipid biosynthetic process*, GO:0045017 *glycerolipid biosynthetic process*,

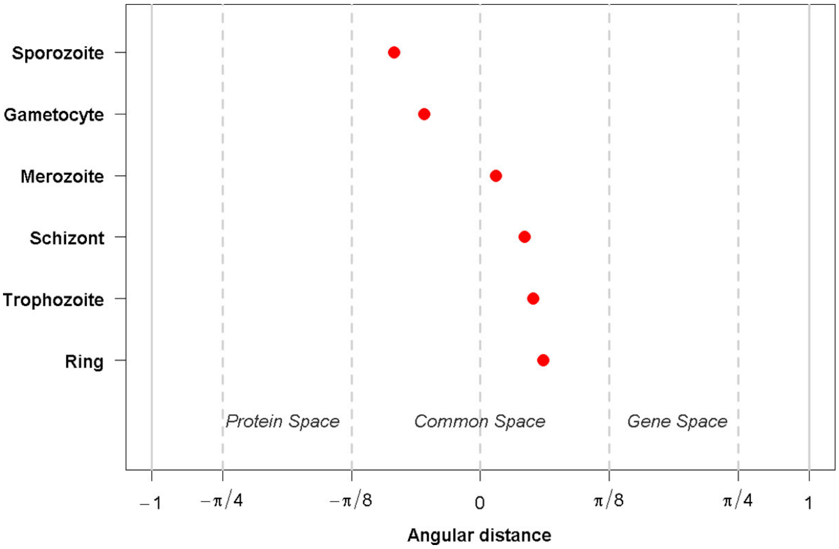


Figure 7 Generalized singular value decomposition - angular distances. GSVD computes angular distances between gene and protein space. In general, the angular distances map to the common space, for which restricted GSE analysis is performed on the gene and on the proteins arraylets. Nevertheless, while angular distances belonging to the intraerythrocytic cycle stages have positive values and show a tendency to the gene space, the angular distances of gametocytes and sporozoites have negative values and thus a tendency towards protein space. These preferences are also reflected by the amount of GO term associations emerging from the gene and from the protein space (see also Figure 6).

GO:0006661 *phosphatidylinositol biosynthetic process*, GO:0046488 *phosphatidylinositol metabolic process*, GO:0006506 *GPI anchor biosynthetic process*, GO:0016255 *attachment of GPI anchor to protein*) associated by CIA with merozoites in protein space, 8 GO terms (GO:0009058 *biosynthetic process*, GO:0051276 *chromosome organization*, GO:0006325 *chromatin organization*, GO:0050776 *regulation of immune response*, GO:0006955 *immune response*, GO:0006096 *glycolysis*, GO:0006334 *nucleosome assembly*, GO:0044237 *cellular metabolic process*) associated by CIA with rings and schizonts in protein space and 3 GO terms (GO:0009117 *nucleotide metabolic process*, GO:0006163 *purine nucleotide metabolic process*, GO:0009116 *nucleoside metabolic process*) associated by CIA with sporozoites in protein space. Only two GO terms (GO:0006096 *glycolysis* and GO:0006006 *glucose metabolic process associated with trophozoites*) from gene space, coincide with GO terms from the common results. Protein activity characteristics derived from CIA show considerable similarities to the other two methods.

Six specific results of GSVD for the life cycle stage ring coincide with the common GO terms associations with this stage (GO:0009058 *biosynthetic process*, GO:0019538 *protein metabolic process*, GO:0044237 *cellular metabolic process*, GO:0008152 *metabolic process*, GO:0055114 *oxidation-reduction process*, GO:0006091 *generation of precursor metabolites and energy*). There are three identical associations for the stage merozoite (GO:0019538 *protein metabolic process*, GO:0016311 *dephosphorylation* and GO:0006470 *protein dephosphorylation*). For each of the other stages, only one GO term from the common associations coincides with the method-specific associations (GO:0009987 *cellular process* for gametocytes, GO:0006091 *generation of precursor metabolites and energy* for trophozoites, GO:0020033 *antigenic*

Table 3 GSVD specific GO term association to gametocyte stage in common space.

GSVD: Gametocyte in common space	
GO:0044238	primary metabolic process
GO:0008152	metabolic process
GO:0044237	cellular metabolic process
GO:0045017	glycerolipid biosynthetic process
GO:0043170	macromolecule metabolic process
GO:0034645	cellular macromolecule biosynthetic process
GO:0046474	glycerophospholipid biosynthetic process
GO:0009059	macromolecule biosynthetic process
GO:0022613	ribonucleoprotein complex biogenesis
GO:0044260	cellular macromolecule metabolic process
GO:0019538	protein metabolic process
GO:0046486	glycerolipid metabolic process
GO:0042254	ribosome biogenesis
GO:0006839	mitochondrial transport
GO:0009987	cellular process

In this table GSVD based GO term association in common space to the life cycle stage gametocyte are presented.

Table 4 GSVD specific GO term association to trophozoite stage in common space.

GSVD: Trophozoite in common space	
GO:0044403	symbiosis, encompassing mutualism through parasitism
GO:0044419	interspecies interaction between organisms
GO:0051704	multi-organism process
GO:0009607	response to biotic stimulus
GO:0006952	defense response
GO:0051707	response to other organism
GO:0051805	evasion or tolerance of immune response of other organism involved in symbiotic interaction
GO:0051807	evasion or tolerance of defense response of other organism involved in symbiotic interaction
GO:0051832	avoidance of defenses of other organism involved in symbiotic interaction
GO:0051834	evasion or tolerance of defenses of other organism involved in symbiotic interaction
GO:0052173	response to defenses of other organism involved in symbiotic interaction
GO:0052564	response to immune response of other organism involved in symbiotic interaction
GO:0020033	antigenic variation
GO:0051809	passive evasion of immune response of other organism involved in symbiotic interaction
GO:0006091	generation of precursor metabolites and energy

In this table GSVD based GO term association in common space to the life cycle stage trophozoite are presented.

variation for schizonts and GO:0009056 *catabolic process* for sporozoites). In conclusion, the ring stage is very well characterized by the GSVD, which is almost in complete agreement with the other methods. The properties of the other stages do not coincide with the common results but should definitely be considered for further analysis as they are highly significant.

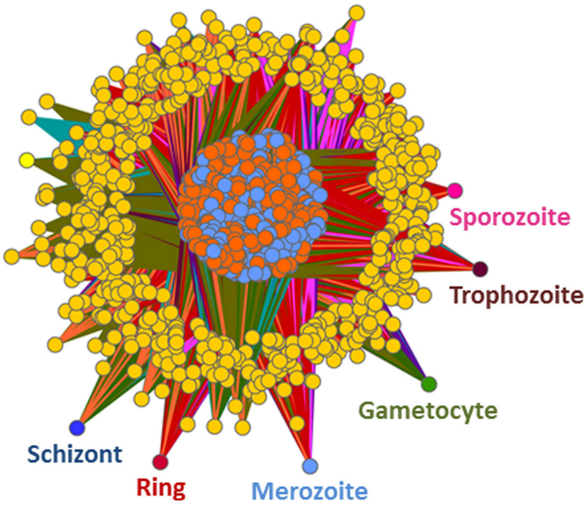


Figure 8 Integrative biclustering - network view of the results. The results of IBC were inspected and only biclusters including life cycle stages were considered for further analysis. IBC discovered 20 clusters where 9 of them contained life cycle stages and GO terms. These 9 biclusters were processed and fed into Cytoscape. An association between a life cycle stage and a GO term is represented by an edge. Different biclusters are represented by different edge colours. The life cycle stages are shown in the same colours as those used for CIA. The genes are coloured in orange, the proteins in light blue and the GO terms in yellow.

In this study, we have applied three integrative analysis methods to a data set containing mRNA and protein abundances from the six life cycle stages of *P. falciparum.* The use of integrative analysis methods allows to consider all annotated and measured genes (3283) and proteins (2491), not limited by the 2230 pairs of genes and proteins as when it was first published in [9]. The integration of knowledge on different levels allows the linking of the data sets based on samples and not on variables (genes, protein).

We presented three different integrative analysis methods, each with its own justification: CIA discovers biological processes on the basis of maximal covariance. GSVD decomposes the data sets into genelets and arraylets and conducts a modified GSE analysis on them. IBC computes biclusters according to the distance between genes, proteins and GO terms.

We have shown method-specific results as well as results common to all three analysis methods. In the case of CIA, the associations in protein space presented a high overlap with the common results. This was not the case for the associations in gene space. In case of the sporozoite stage, GSVD associations are very simmilar to the common results. For the other stages, GSVD yielded different mappings compared to the common results. As a GO term is associated or not with a life cycle stage, only general but no method-specific results were computed for IBC.

For CIA, it is important to consider that GO term associations are done through projection, whereas GSVD maps GO terms to individual stages through restricted GSE analysis and IBC assigns GO terms to life cycle stages

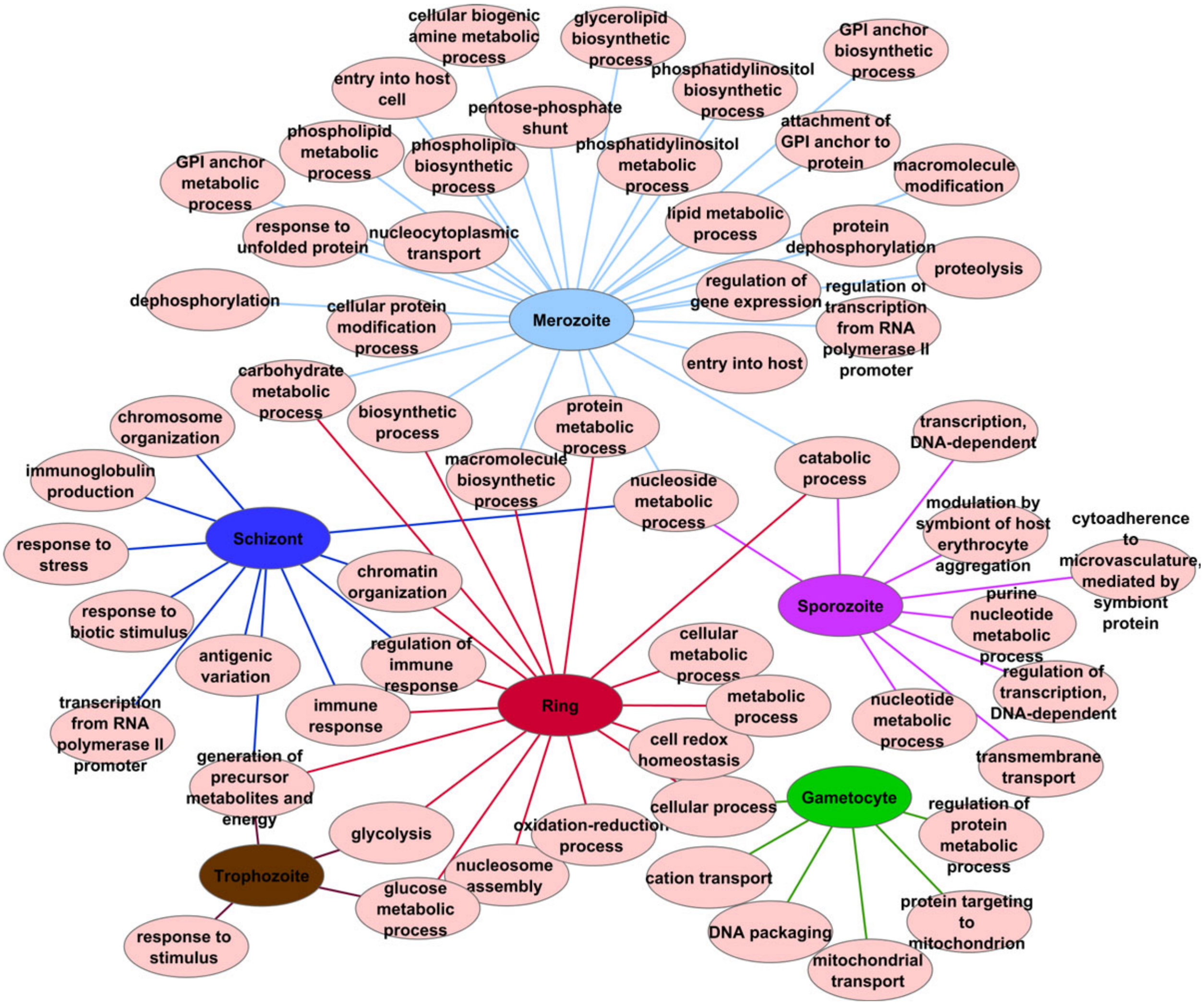

cellular biogenic amine metabolic process
glycerolipid biosynthetic process
GPI anchor biosynthetic process
entry into host cell
pentose-phosphate shunt
phosphatidylinositol biosynthetic process
phospholipid metabolic process
phospholipid biosynthetic process
phosphatidylinositol metabolic process
attachment of GPI anchor to protein
GPI anchor metabolic process
macromolecule modification
response to unfolded protein
nucleocytoplasmic transport
lipid metabolic process
protein dephosphorylation
proteolysis
dephosphorylation
cellular protein modification process
Merozoite
regulation of gene expression
regulation of transcription from RNA polymerase II promoter
entry into host
carbohydrate metabolic process
chromosome organization
biosynthetic process
protein metabolic process
macromolecule biosynthetic process
transcription, DNA-dependent
catabolic process
immunoglobulin production
nucleoside metabolic process
modulation by symbiont of host erythrocyte aggregation
cytoadherence to microvasculature, mediated by symbiont protein
Schizont
response to stress
chromatin organization
Sporozoite
response to biotic stimulus
purine nucleotide metabolic process
antigenic variation
regulation of immune response
cellular metabolic process
regulation of transcription, DNA-dependent
transcription from RNA polymerase II promoter
immune response
Ring
metabolic process
nucleotide metabolic process
generation of precursor metabolites and energy
cell redox homeostasis
transmembrane transport
cellular process
Gametocyte
regulation of protein metabolic process
glycolysis
oxidation-reduction process
Trophozoite
nucleosome assembly
cation transport
glucose metabolic process
protein targeting to mitochondrion
response to stimulus
DNA packaging
mitochondrial transport

Figure 9 GO terms to life cycle associations discovered by all three methods. Network view of the GO term to life cycle stage associations discovered by all three integrative analysis methods: CIA, GSVD and IBC. We observe that gametocytes and sporozoites are loosely connected to the rest of the network underlining the separation of these stages from the intraerythrocytic cycle. Merozoites possess the largest amount of GO term associations, while trophozoites show the lowest amount of associations. Further details concerning individual stage-to-GO-term mappings are addressed in the discussion.

through the distance to the corresponding life cycle stage. Another important aspect is that with CIA it is not possible to associate one GO term to more than one life cycle stage, while this is possible with GSVD and IBC. Due to the heterogeneous computational methods, we proposed taking the intersect of the three obtained results.

In the three-fold validated network view of the biological processes (Figure 9), we observe the separation of the intraerytrocytic cycle (merozite, ring, trophozoite and schizont) from sporozoites and gametocytes. While the stages of the intraerytrocytic cycle are tightly connected to one another, sporozoites share two biological processes and gametocytes share only one biological process with the rest. Gametocytes and sporozoites do not possess any common processes, reflecting the differences between these stages. Gametocytes are released into the blood stream, from where they travel to the liver, while sporozoites represent the sexual stage and lie dormant in cell cycle arrest until ingestion by a mosquito.

The data used here was gathered in order to investigate the role of post-transcriptional regulation in *P. falciparum* [9]. For this, only pairs of mRNA and the corresponding protein were considered, resulting in the exploitation of 89% of the proteins and 60% of the genes that were experimentally measured. By employing integrative analysis methods we were able to take all measured data into account.

LeRoch and coworkers [9] mention that there is a "bias in proteomic analysis of whole-cell lysates, in that such methods may fail to detect secreted or membrane proteins present in low abundance" such as GPI anchors. Due to the integrative approach, our analysis associates several GO terms related to GPI anchors proteins (GO:0006506 *GPI anchor biosynthetic process*, GO:0016255 *attachment of GPI anchor to protein*, GO:0006661 *phosphatidylinositol biosynthetic process*, GO:0046488 *phosphatidylinositol metabolic process*) with the merozoite stage, prevailing over this shortcoming. These associations are in agreement with [66], where distinct protein classes, with a focus on merozoite surface antigens, are discussed. The importance of GPI anchor proteins in the merozoite stage is well known and very important in immune evasion [67,68].

Other biological processes mentioned in [9] such as *glycolysis* and *cell invasion*, without any life cycle mapping, were also found in our network: GO:0044409 *entry into host* and GO:0030260 *entry into host cell*, both associated with the merozoite stage. Our network assigns GO:0006096 *glycolysis* to the stage trophozoite, in concordance to [69] where the transcriptome of *P. falciparum* was characterized.

Simmilar to our findings, cell invasion was associated with merozoites in [67], where a proteomic view of the *P. falciparum* life cycle was presented. Other concordances with [67] include the assignment of GO:0006508 *proteolysis* to the merozoite stage. During trophozoite stage, digestion of haemoglobin takes place. Our network maps GO:0006091 *generation of precursor metabolites and energy* to trophozoites, confirming the importance of energy production during this stage. As mentioned by Florens *et al.* [67], sporozoites are injected into the blood stream where they have to survive in a hostile environment. Based on our combined results, sporozoites are associated with GO:0020013 *modulation by symbiont of host erythrocyte aggregation* and GO:0020035 *cytoadherence to microvasculature, mediated by symbiont protein*, which reflects the process of survival. Additionally, sporozoites are associated with metabolism and transcription, as was shown in Figure 5 of [67]. Our results reflect these findings by mapping GO:0006163 *purine nucleotide metabolic process*, GO:0009117 *nucleotide metabolic process*, GO:0006351 *transcription, DNA dependent* and GO:0006355 *regulation of transcription, DNA dependent* to the sporozoite stage.

During gametocyte stage, DNA processing and energy production is highly regulated, as mentioned in [67]. In agreement, our results assign GO:0006323 *DNA packaging*, GO:0006839 *mitochondrial transport* and GO:0006626 *protein targeting to mitochondrion* to the gametocytes.

The analysis of the *P. falciparum* proteome by LaCount and colleagues [70] associated the intraerythrocytic cycle with chromatin modification, transcriptional regulation, mRNA stability/processing, ubiquitination, nucleic acid metabolism and invasion of host cells. Since our analysis corresponds to individual life cycle stages, we can associate biological processes to a certain stage of the intraerythrocytic cycle, providing a more detailed description of *P. falciparum*. According to our findings, chromatin modification takes place during schizont stage (GO:0006325 *chromatin organization*, GO:0051276 *chromosome organization*); merozoites are associated with GO:0006357 *regulation of transcription from RNA polymerase II promotor* and schizonts with GO:0042795 *transcription from RNA polymerase II promotor* ; merozoites are associated with GO:0009116 *nucleoside metabolic process*; invasion of host cells can be observed during merozoite stage (GO:0044409 *entry into host* and GO:0030260 *entry into host cell*). Ubiquitination was only detected through its parent term GO:0044267 *cellular protein metabolic process*, which was associated with merozoites.

Fagan *et al.* [1] conducted CIA on a slightly different data set which took *P. berghei* orthologues into account and showed that GO:0006412 *biosynthesis* is associated to the intraerythrocytic cycle. In our network, several more specialized biosynthetic processes are associated with the merozoite stage: GO:0009059 *macromolecule biosynthetic process*, GO:0008654 *phospholipid biosynthetic process*, GO:0045017 *glycerolipid biosynthetic process*, GO:0006661 *phosphatidylinositol biosynthetic process*, GO:0006506 *GPI anchor biosynthetic process*, as well as the GO term GO:0006412 *biosynthetic process* itself.

The importance of immune evasion through antigenic variation was highlighted by Winzeler [71]. Our results show that this process is related to the schizont stage, as our analysis associates GO:0020033 *antigenic variation*, GO:0006955 *immune response*, GO:0050776 *regulation of immune response*, GO:0002377 *immunoglobulin production*, GO:0006950 *response to stress* and GO:0009607 *response to biotic stimulus* with this stage.

The role of lipids during merozoite stage was already shown in 1988 by Mikkelsen *et al.* [72]. Our computed network associates merozoites with GO:0006644 *phospholipid metabolic process*, GO:0008654 *phospholipid biosynthetic process*, GO:0046486 *glycerolipid metabolic process* and GO:0006629 *lipid metabolic process*, reflecting this early finding.

Phosphorilation and dephosphoryliation processes play an important role in the internalization step of merozoites [73], a fact that is also reflected by our results. Merozoites are associated with GO:0016311 *dephosphorylation* and GO:0006470 *protein dephosphorylation*.

The role of the pentose phosphate pathway in *P. falciparum* was disscused in [74], without a clear life cycle stage assignment. Our computed network view maps GO:0006098 *pentose-phosphate shunt* to merozoites.

As shown in [75], REDOX complexes play an important role during ring stage, which is in agreement with our results that associate ring stage with GO:0045454 *cell redox homeostasis* and GO:0055114 *oxidation-reduction process*.

Roth [76] showed that carbohydrate metabolism is a key metabolic process connecting the host cells with *P. falciparum*. Our findings assign GO:0005975 *carbohydrate metabolic process* to merozoite and ring stages.

Most of our network associations are in concordance with several publications dealing with the characterization of *P. falciparum*, based on transcriptome [68,69] and proteome [67,70] characterization data. A considerable amount of the findings in the above publications are concentrated in our results of the used integrative analysis methods. Our findings are more detailed through the association with a specific life cycle stage rather than, e.g. the whole intraerythrocytic cycle as well as through the association of a child GO term instead of a parent GO term to the corresponding stage. Our study unifies individual findings from several publications of the past 25 years of research. Not all results from the publications mentioned above are present in our network. This could be due to the fact that none of the cited publications, except [9], used the same data sets as we did. Llinas *et al.* [68] compared the three *P. falciparum* strains 3D7, Dd2 and HB3 through the measurement of the gene expression profiles of 6287, 5294 and 6415 genes during the intraerythrocytic cycle. Bozdech *et al.* [69] considered in their analysis of the intraerythrocytic cycle transcriptome the expression of 5508 genes. LaCount and colleagues [70] analysed 1267 proteins for their protein interaction network of *P. falciparum*. In [67], Florens *et al.* use approximately 2400 proteins in order to create a proteomic view of the *P. falciparum* life cycle. The other studies are based on lab experiments on smaller groups of genes or proteins [66,71-73,75].

Additionally, our combined network view of life cycle stage dependent GO term association provides a new overview for the vaccine research and offers new insight in the interdependencies between life cycle stages. Possibly it could even identify key biological processes on which vaccine researchers could concentrate their work.

Conclusion

In this study we have shown the power of integrative analysis methods. We presented three very different approaches that showed significant overlap of results. We compared our findings against the past 25 years of *P. falciparum* research and showed that the obtained network unifies, on the life cycle level, results from analyses done separately on transcriptome and proteome data, as well as results from the lab, which were performed on small groups of genes or proteins. Further investigations are needed to obtain a complete map of the biological processes activated during the life cycle of *P. falciparum*. Measurement of the transcriptome and proteome of *P. falciparum*, exploiting the advantages of current high throughput technologies, would complement the spectrum of biological process presented here. An increase of our understanding of *P. falciparum* could be achieved by performing the integrative analysis methods on the molecular function and/or cellular compartment level of gene ontology. Further work could also cover the identfication of genes and proteins that play key roles during the life cycle of *P. falciparum* through integrative analysis on gene and protein level, not only on GO term level.

Additional material

Additional file 1: GO term mapping. PDF file containing the mapping between the numbers used in the CIA plots, GO terms and GO names.

Additional file 2: CIA general GO term associations in gene space. PDF file containing the CIA general GO term associations in gene space.

Additional file 3: CIA general GO term associations in gene space. PDF file containing the CIA general GO term associations in protein space.

Additional file 4: CIA division limits. PDF file containing the CIA division limits for general (left) and specific (right) associations. The colours of the areas correspond to the colours of the stages they are associated with.

Additional file 5: CIA specific GO term associations in gene space. PDF file containing the CIA specific GO term associations in gene space.

Additional file 6: CIA specific GO term associations in protein space. PDF file containing the CIA specific GO term associations in protein space.

Additional file 7: GSVD general GO term associations. PDF file containing the GSVD based general associations of GO terms to life cycle stages in common space.

Additional file 8: GSVD specific GO term associations. PDF file containing the GSVD based specific associations of GO terms to life cycle stages in common space.

Additional file 9: IBC general GO term associations. PDF file containing the IBC based general associations of GO terms to life cycle stages in common space.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

GGT and DM conceived the original idea. OAT performed the analysis, summarized and compared the results and drafted the manuscript under the supervision of GGT. All the authors have read and approved the final manuscript.

Acknowledgements

This work has been supported by the Federal Ministry of Economy, Family and Youth (BMWFJ), the Federal Ministry of Traffic, Innovation and Technology (bmvit), the Styrian Business Promotion Agency SFG, the Standortagentur Tirol and ZIT - Technology Agency of the City of Vienna through the COMET-Funding Program managed by the Austrian Research Promotion Agency FFG [COMET K2 Grant 824186, Austrian Centre of Industrial Biotechnology]. We thank Orly Alter for providing the Mathematica code of the GSVD and Julia Feichtinger for critically reading the manuscript.

Declarations

This article has been published as part of *BMC Systems Biology* Volume 8 Supplement 2, 2014: Selected articles from the High-Throughput Omics and Data Integration Workshop. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/8/S2. The publication costs for this article were funded by COST-BMBS, Action BM1006 "Next Generation Sequencing Data Analysis Network", SeqAhead and by the Austrian Centre of Industrial Biotechnology

Authors' details

[1]Core Facility Bioinformatics, Austrian Centre of Industrial Biotechnology, Graz, Austria. [2]Institute of Genomics and Bioinformatics, Graz University of Technology, Graz, Austria. [3]Cell Design and Engineering, Austrian Centre of Industrial Biotechnology, Vienna, Austria. [4]Department of Biotechnology, BOKU-VIBT University of Natural Resources and Life Sciences, Vienna, Austria. [5]Omics Center Graz, Graz, Austria.

Published: 13 March 2014

References

1. Fagan A, Culhane AC, Higgins DG: **A multivariate analysis approach to the integration of proteomic and gene expression data.** *Proteomics* 2007, **7**:2162-2171.
2. Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Weiss Solis DY, Molter C, Duque R, Bersini H, Nowe A: **GENESHIFT: a Non-Parametric Approach for Integrating Microarray Gene Expression Data Based on the Inner Product as a Distance Measure Between the Distributions of Genes.** *IEEE/ACM Trans Comput Biol Bioinf* 2013, **2**:383-292.
3. Alter O, Brown PO, Botstein D: **Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms.** *Proc Natl Acad Sci USA* 2003, **100**:3351-3356.
4. Wang KS, Liu X: **Integrative Analysis of Genome-wide Expression and Methylation Data.** *J Biom Biostat* 2013, **4**:4-6.
5. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, Shen R: **Pattern discovery and cancer gene identification in integrated cancer genomic data.** *Proc Natl Acad Sci USA* 2013, **110**:4245-4250.
6. Kockmann T, Gerstung M, Schlumpf T, Xhinzhou Z, Hess D, Beerenwinkel N, Beisel C, Paro R: **The BET protein FSH functionally interacts with ASH1 to orchestrate global gene activity in Drosophila.** *Genome Biol* 2013, **14**:R18.
7. Chen Z, Zhang W: **Integrative Analysis Using Module-Guided Random Forests Reveals Correlated Genetic Factors Related to Mouse Weight.** *PLoS Comput Biol* 2013, **9**:e1002956.
8. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R, Min R, Alves P, Abyzov A, Addleman N, Bhardwaj N, Boyle AP, Cayting P, Charos A, Chen DZ, Cheng Y, Clarke D, Eastman C, Euskirchen G, Frietze S, Fu Y, Gertz J, Grubert F, Harmanci A, Jain P, Kasowski M, Lacroute P, Leng J, Lian J, Monahan H, O'Geen H, Ouyang Z, Partridge EC, Patacsil D, Pauli F, Raha D, Ramirez L, Reddy TE, Reed B, Shi M, Slifer T, Wang J, Wu L, Yang X, Yip KY, Zilberman-Schapira G, Batzoglou S, Sidow A, Farnham PJ, Myers RM, Weissman SM, Snyder M: **Architecture of the human regulatory network derived from ENCODE data.** *Nature* 2012, **489**:91-100.
9. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder A, Carucci DJ, Yates JR, Winzeler E: **Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle.** *Genome Res* 2004, **14**:2308-2318.
10. Cox B, Kislinger T, Emili A: **Integrating gene and protein expression data: pattern analysis and profile mining.** *Methods* 2005, **35**:303-314.
11. Cagney G, Park S, Chung C, Tong B, Dushlaine CO, Shields DC, Emili A: **Human Tissue Profiling with Multidimensional Protein Identification Technology.** *J Proteome Res* 2005, **4**:1757-1767.
12. Corbin RW, Paliy O, Yang F, Shabanowitz J, Platt M, Lyons CE, Root K, McAuliffe J, Jordan MI, Kustu S, Soupene E, Hunt DF: **Toward a protein profile of Escherichia coli: comparison to its transcription profile.** *Proc Natl Acad Sci USA* 2003, **100(16)**:9232-9237.
13. Chen Yr, Juan Hf, Huang Hc, Huang Hh, Lee Yj, Liao My, Tseng Cw, Lin Ll, Chen Jy, Wang Mj, Chen Jh, Chen Yj: **Quantitative Proteomic and Genomic Profiling Reveals Metastasis-Related Protein Expressio Patterns in Gastric Cancer Cells research articles.** *J Proteome Res* 2006, **5**:2727-2742.
14. Griffin TJ: **Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in Saccharomyces cerevisiae.** *Mol Cell Proteomics* 2002, **1**:323-333.
15. Mootha VK, Bunkenborg J, Olsen JV, Hjerrild M, Wisniewski JR, Stahl E, Bolouri MS, Ray HN, Sihag S, Kamal M, Patterson N, Lander ES, Mann M: **Integrated Analysis of Protein Composition, Tissue Diversity, and Gene Regulation in Mouse Mitochondria.** *Cell* 2003, **115**:629-640.
16. Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, Deciu C, Winzeler E, Yates JR: **Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae.** *Proc Natl Acad Sci USA* 2003, **100**:3107-3112.
17. Kislinger T, Cox B, Kannan A, Chung C, Hu P, Ignatchenko A, Scott MS, Gramolini AO, Morris Q, Hallett MT, Rossant J, Hughes TR, Frey B, Emili A: **Global Survey of Organ and Organelle Protein Expression in Mouse: Combined Proteomic and Transcriptomic Profiling.** *Cell* 2006, **125**:173-186.
18. Nie L, Wu G, Brockman FJ, Zhang W: **Integrated analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: zero-inflated Poisson regression models to predict abundance of undetected proteins.** *Bioinformatics* 2006, **22**:1641-1647.
19. Haider S, Pal R: **Integrated Analysis of Transcriptomic and Proteomic Data.** *Curr Genomics* 2013, **14**:91-110.
20. Hwang D, Smith JJ, Leslie DM, Weston AD, Rust AG, Ramsey S, de Atauri P, Siegel AF, Bolouri H, Aitchison JD, Hood L: **A data integration methodology for systems biology: Experimental verification.** *Proc Natl Acad Sci USA* 2005, **102**:17302-17307.
21. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, Weston AD, de Atauri P, Aitchison JD, Hood L, Siegel AF, Bolouri H: **A data integration methodology for systems biology.** *Proc Natl Acad Sci USA* 2005, **102**:17296-17301.
22. Hahne H, Mäder U, Otto A, Bonn F, Steil L, Bremer E, Hecker M, Becher D: **A comprehensive proteomics and transcriptomics analysis of Bacillus subtilis salt stress adaptation.** *J Bacteriol* 2010, **192**:870-882.
23. Verhoef S, Ballerstedt H, Volkers RJM, de Winde JH, Ruijssenaars HJ: **Comparative transcriptomics and proteomics of p-hydroxybenzoate producing Pseudomonas putida S12: novel responses and implications for strain improvement.** *Appl Microbiol Biotechnol* 2010, **87**:679-690.
24. Takemasa I, Kittaka N, Hitora T, Watanabe M, Matsuo EI, Mizushima T, Ikeda M, Yamamoto H, Sekimoto M, Nishimura O, Doki Y, Mori M: **Potential biological insights revealed by an integrated assessment of proteomic and transcriptomic data in human colorectal cancer.** *Int J Oncol* 2012, **40**:551-559.
25. Piruzian E, Bruskin S, Ishkin A, Abdeev R, Moshkovskii S, Melnik S, Nikolsky Y, Nikolskaya T: **Integrated network analysis of transcriptomic and proteomic data in psoriasis.** *BMC Syst Biol* 2010, **4**:41-53.
26. Perco P, Mühlberger I, Mayer G, Oberbauer R, Lukas A, Mayer B: **Linking transcriptomics and proteomic data on the level of protein interaction networks.** *Electrophoresis* 2010, **31**:1780-1789.
27. Joyce AR, Palsson BO: **The model organism as a system: integrating 'omics' data sets.** *Nat Rev Mol Cell Biol* 2006, **7**:198-210.
28. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R: **Gene regulatory network inference: data integration in dynamic models-a review.** *BioSystems* 2009, **96**:86-103.
29. Zhang W, Li F, Nie L: **Integrating multiple 'omics' analysis for microbial biology: application and methodologies.** *Microbiology* 2010, **156**:287-301.
30. Dolèdec S, Chessel D: **Co-inertia analysis: an alternative method for studying species-environment relationships.** *Freshw Biol* 1994, **31**:277-294.
31. Kaiser S: **Biclustering: Methods, Software and Application.** *PhD thesis* Ludwig-Maximilians-University Munich, Department of Statistics; 2011.
32. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Rubin GM, Sherlock G: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**:25-29.
33. Le Roch KG, Zhou Y, Blair PL, Grainger M, Moch JK, Haynes JD, De La Vega P, Holder A, Batalov S, Carucci DJ, Winzeler E: **Discovery of gene function by expression profiling of the malaria parasite life cycle.** *Science* 2003, **301**:1503-1508.
34. R Development Core Team: *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria; 2013 [http://www.R-project.org], [ISBN 3-900051-07-0].
35. Carlson M: *org.Pf.plasmo.db: Genome wide annotation for Malaria* , [R package version 2.8.1].

36. Carlson M: *GO.db: A set of annotation maps describing the entire Gene Ontology* , [R package version 2.8.0].
37. Tucker LR: **An inter-battery method for factor analysis.** *Psychometrika* 1958, **23**:111-136.
38. Jarraud S, Mougel C, Thioulouse J, Lina G, Meugnier H, Forey F, Etienne J, Vandenesch F, Jarraud S, Mougel C, Thioulouse J, Lina G, Nesme X, Etienne J: **Relationships between Staphylococcus aureus Genetic Background, Virulence Factors, agr Groups (Alleles), and Human Disease.** *Infect Immun* 2002, **70**:631-641.
39. Thioulouse J, Lobry J: **Co-inertia analysis of amino-acid physico-chemical properties and protein composition with the ADE package.** *Comput Appl Biosci* 1995, **11**:321-329.
40. Culhane AC, Perrière G, Higgins DG: **Cross-platform comparison and visualisation of gene expression data using co-inertia analysis.** *BMC Bioinformatics* 2003, **4**:59.
41. Jolliffe IT: *Principal Component Analysis* New York Berlin Heidelberg: Springer-Verlag; 2002.
42. Gimaret-Carpentier C, Chessel D, Pascal J: **Non-symmetric correspondence analysis: an alternative for species occurrences data.** *Plant Ecol* 1998, **138**:97-112.
43. Greenacre M: *Theory and Applications of Correspondence Analysis* London: Academic Press; 1983.
44. Robert P, Escoufier Y: **A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient.** *Appl Statist* 1976, **25**:257-265.
45. Culhane AC, Thioulouse J, Perrière G, Higgins DG: **MADE4: an R package for multivariate analysis of gene expression data.** *Bioinformatics* 2005, **21**:2789-2790.
46. Chessel D, Dufour AB, Thioulouse J: **The ade4 package - I: One table methods.** *R News* 2004, **5**:5-10.
47. Busold CH, Winter S, Hauser N, Bauer A, Dippon J, Hoheisel JD, Fellenberg K: **Integration of GO annotations in Correspondence Analysis: facilitating the interpretation of microarray data.** *Bioinformatics* 2005, **21**:2424-2429.
48. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci USA* 2000, **97**:10101-10106.
49. Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, Connell JXO, Zhu S, Fero M, Sherlock G, Pollack JR, Brown PO, Botstein D, Rijn MVD: **Mechanisms of disease Molecular characterisation of soft tissue tumours: a gene expression study.** *The Lancet* 2002, **359**:1301-1307.
50. Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogy R: **Large-scale temporal gene expression mapping of central nevous system development.** *Proc Natl Acad Sci USA* 1998, **95**:334-339.
51. Hilsenbeck SG, William E, Schiff R, Connell O, Hansen RK, Osborne K, Fuqua SAW: **Statistical Analysis of Array Expression Data as Applied to the Problem of Tamoxifen Resistance.** *J Natl Cancer Inst* 1999, **91**:453-459.
52. Golub GH, Van Loan CF: *Matrix Computation* Baltimore and London: Johns Hopkins University Press; 1996.
53. Paige CC, Saunders MA: **Towards a Generalized Singular Value Decomposition.** *SIAM J Number Anal* 1981, **18**:398-405.
54. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association.** *Bioinformatics* 2007, **23**:257-258.
55. Hartigan JA: **Direct Clustering of a Data Matrix.** *J Am Stat Assoc* 1972, **67**:123-129.
56. Cheng Y, Church M: **Biclustering of Expression Data.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:93-103.
57. Getz G, Levine E, Domany E: **Coupled two-way clustering analysis of gene microarray data.** *Proc Natl Acad Sci USA* 2000, **97**:12079-12084.
58. Ben-Dor A, Chor B, Karp R, Yukhini Z: **Descovering local structure in gene expression data: The order preserving submatrix problem.** *J Comput Biol* 2003, **10**:373-384.
59. Murali T, Kasif S: **Extracting conserved gene expression motifs from gene expression data.** *Pac Symp Biocomput* 2003, **8**:77-88.
60. Madeira SC, Oliveira AL: **Biclustering algorithms for biological data analysis: a survey.** *IEEE/ACM Trans Comput Biol Bioinf* 2004, **1**:24-45.
61. Prelić A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E: **A systematic comparison and evaluation of biclustering methods for gene expression data.** *Bioinformatics* 2006, **22**:1122-1129.
62. Kaiser S, Santamaria R, Tatsiana , Khamiakova , Sill M, Theron R, Quintales L, Leisch F: *biclust: BiCluster Algorithms* 2013 [http://CRAN.R-project.org/package=biclust], [R package version 1.0.2].
63. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale AL, Brenton JD, Tavaré S, Caldas C, Aparicio S: **The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.** *Nature* 2012, **486**:346-352.
64. Tanay A, Sharan R, Kupiec M, Shamir R: **Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data.** *Proc Natl Acad Sci USA* 2004, **101**:2981-2986.
65. Smoot M, Ono K, Ruscheinski J, Wang PL, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization.** *Bioinformatics* 2011, **27**:431-432.
66. Sanders PR, Gilson PR, Cantin GT, Greenbaum DC, Nebl T, Carucci DJ, McConville MJ, Schofield L, Hodder AN, Yates JR, Crabb BS: **Distinct protein classes including novel merozoite surface antigens in Raftlike membranes of Plasmodium falciparum.** *J Biol Chem* 2005, **280**:40169-40176.
67. Florens L, Washburn MP, Raine JD, Anthony RM, Grainger M, Haynes JD, Moch JK, Muster N, Sacci JB, Tabb DL, Witney Aa, Wolters D, Wu Y, Gardner MJ, Holder Aa, Sinden RE, Yates JR, Carucci DJ: **A proteomic view of the Plasmodium falciparum life cycle.** *Nature* 2002, **419**:520-526.
68. Llinás M, Bozdech Z, Wong ED, Adai AT, DeRisi JL: **Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains.** *Nucleic Acids Res* 2006, **34**:1166-1173.
69. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum.** *PLoS Biol* 2003, **1**:E5.
70. LaCount DJ, Vignali M, Chettier R, Phansalkar A, Bell R, Hesselberth JR, Schoenfeld LW, Ota I, Sahasrabudhe S, Kurschner C, Fields S, Hughes RE: **A protein interaction network of the malaria parasite Plasmodium falciparum.** *Nature* 2005, **438**:103-107.
71. Winzeler EA: **Malaria research in the post-genomic era.** *Nature* 2008, **455**:751-756.
72. Mikkelsen RB, Kamber M, Wadwa KS, Lin PS, Schmidt-Ullrich R: **The role of lipids in Plasmodium falciparum invasion of erythrocytes: a coordinated biochemical and microscopic analysis.** *Proc Natl Acad Sci USA* 1988, **85**:5956-5960.
73. Ward G, Fujioka H, Aikawa M, Miller L: **Staurosporine Inhibits Invasion of Erythrocytes by Malarial Merozoites.** *Exp Parasitol* 1994, **79**:480-487.
74. Bozdech Z, Ginsburg H: **Data mining of the transcriptome of Plasmodium falciparum: the pentose phosphate pathway and ancillary processes.** *Malaria J* 2005, **4**:17.
75. Mok S, Imwong M, Mackinnon MJ, Sim J, Ramadoss R, Yi P, Mayxay M, Chotivanich K, Liong KY, Russell B, Socheat D, Newton PN, Day NPJ, White NJ, Preiser PR, Nosten F, Dondorp AM, Bozdech Z: **Artemisinin resistance in Plasmodium falciparum is associated with an altered temporal pattern of transcription.** *BMC Genomics* 2011, **12**:391.
76. Roth EJ: **Plasmodium falciparum carbohydrate metabolism: a connection between host cell and parasite.** *Blood Cells* 1990, **16**:453-466.

doi:10.1186/1752-0509-8-S2-S4
Cite this article as: Tomescu *et al.*: Integrative omics analysis. A study based on *Plasmodium falciparum* mRNA and protein data. *BMC Systems Biology* 2014 **8**(Suppl 2):S4.

Proteogenomic analysis of *Anopheles gambiae* phagocytic hemocyte populations reveals an anticipatory innate immune response in the absence of pathogen challenge

Running Title: *Anopheles* phagocytic hemocyte proteome

Authors: Ryan C. Smith[1,*], Jonas G. King[1,*], Dingyin Tao[1‡], Oana A.Tomescu[2,3,‡], Clara Brando[1], Gerhard G. Thallinger[2,3,4], and Rhoel R. Dinglasan[1]

Affiliations:

[1]W. Harry Feinstone Department of Molecular Microbiology and Immunology and the Malaria Research Institute, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, Maryland 21205, USA

[2]Bioinformatics, Institute for Knowledge Discovery, Graz University of Technology, 8010 Graz, Austria

[3]Corefacility Bioinformatics, Austrian Centre of Industrial Biotechnology, 8010 Graz, Austria

[4]BioTechMed OMICS Center Graz, 8010 Graz, Austria

[*]These authors contributed equally.

[‡]These authors contributed equally.

Keywords: Innate immunity, hemocyte, phagocyte, mosquito, malaria

Abstract

The innate immune response to a broad class of pathogens is highly conserved across all eukaryotes and has been studied in great detail at the cellular and transcriptomic level in several insect species. However, the commensurate cellular proteomic response, especially of hemocytes, the primary immune cell population in insects, has remained poorly understood. We report on the comprehensive proteogenomic analysis of a phagocyte subpopulation from *Anopheles gambiae*, the primary malaria mosquito vector in Sub-Saharan Africa. We leveraged the innate phagocytic response of mosquito granulocytes to achieve targeted enrichment for these cells to facilitate the examination of their proteomic response profiles following sugar feeding, a non-infectious blood meal, and *Plasmodium falciparum* infection. A comparative integrative-OMICs analysis of existing transcriptomic profiles combined with these proteomic data permitted the delineation of the functional genome of anopheline granulocytes. This rich data resource provides the first comprehensive reference for protein profiles for mosquito granulocytes during homeostasis and pathogen challenge. We observed that phagocytosis, blood feeding, and *P. falciparum* infection induced dramatic shifts in granulocyte protein expression indicative of broad changes in cellular proliferation and innate immune response priming. Importantly, we identified a large number of hemocyte immune proteins that respond to blood feeding alone, suggesting that granulocytes may play an integral role in an anticipatory immune response prior to pathogen challenge. This integrated-OMICs dataset for anopheline granulocytes can support future quantitative immunology studies to acquire novel insights into mosquito hemocyte regulation.

To facilitate the review of the mass spectral data, we have uploaded the files to PRoteomics IDEntifications database (PRIDE). The data can be accessed through the following reviewer account details:

Username: reviewer62532@ebi.ac.uk

Password: IcSSRtVQ

Data link: http://www.ebi.ac.uk/pride/archive/projects/PXD001507

Introduction

Insects live in a microbe-rich microenvironment that has driven the evolution of an effective innate immune response and insect host defense system. For instance, mosquitoes encounter dramatically different environments during their life cycle– aquatic in their juvenile stages and terrestrial as adults. During these stages, mosquitoes encounter and must survive immune challenge from a variety of bacterial, viral, fungal, and parasitic pathogens. Mechanistically, the mosquito innate immune system can be divided into evolutionary conserved cellular and humoral responses. Mosquito 'blood cells', or hemocytes, are thought to be integral to both responses. Cellular responses include the engulfment of bacteria via phagocytosis, while humoral components produced by hemocytes and the fat body act through hemolymph-derived factors to sequester and/or kill invading pathogens.

Compared to vertebrate cell-mediated immunity, our current understanding of mosquito hemocyte function is still in its infancy. Extensive conservation in the mechanisms of hematopoiesis exist between insects and mammals (Wang et al. 2014), yet only three distinct classes of hemocytes have been distinguished by morphology and limited biochemical characterization (Castillo et al. 2006). Granulocytes are the professional phagocytes of the insect, sharing similar roles with vertebrate macrophages. Believed to share a common ancestry with vertebrate macrophages, insect granulocytes mediate immune activation and the production of humoral defense factors. Oenocytoids appear to have a primary role in the production of melanin and have been implicated in wound healing and pathogen killing in other insect systems (Wang et al. 2014). Finally, prohemocytes are thought to serve as hematopoietic progenitor cells that differentiate to produce other hemocyte cell populations (Rodrigues et al. 2010), or as recently proposed, may simply represent a smaller class of granulocytes (King and Hillyer 2013).

Although hemocytes have been implicated in several aspects of insect physiology in other systems (Fauvarque and Williams 2011; Wang et al. 2014), these have yet to be fully explored in mosquitoes. Much of our current knowledge on mosquito hemocytes is limited to observations and morphological classifications of the hemocyte sub-types, thereby leaving much of the fundamental aspects of their biology unknown

(Hillyer and Strand 2014). In recent years, new information regarding hemocyte circulation dynamics (King and Hillyer 2013), aggregation (King and Hillyer 2012), and proliferation following blood feeding (Castillo et al. 2011; Bryant and Michel 2014) have been described in mosquitoes. However, these studies remain focused on presumed and likely contributions of hemocytes, as a whole, to the innate immune response (Lavine and Strand 2002; Pinto et al. 2009; Rodrigues et al. 2010; King and Hillyer 2012; Ramirez et al. 2014). Transcriptional analysis has been performed on a heterogeneous population of mosquito hemocytes to measure the response and infer the role of these cells in the context of either bacterial or *Plasmodium* parasite challenge (Baton et al. 2009; Pinto et al. 2009). These studies revealed a partitioning of hemocyte transcriptomic profiles according to pathogen challenge and temporal pattern following infection (Baton et al. 2009; Pinto et al. 2009). Pinto et al. (2009) also identified several hemocyte factors that modulate parasite developmental success. However, these studies were limited by the generalization of the overall hemocyte response and the lack of targeted, comprehensive analyses of a specific subset of the hemocyte population in the context of homeostasis and pathogen challenge.

Here, we report on the comprehensive proteogenomic analysis of the granulocyte response to blood feeding and infection in the major African malaria vector, *Anopheles gambiae*. We developed a novel purification technique that isolates the phagocytic (granulocyte) population of mosquito blood cells (granulocytes) to permit enrichment and subsequent mass spectrometry-based proteomics analyses. By comparing the proteome profiles to existing transcriptome data, we have improved the functional annotation of this hemocyte subset. In addition, our integrative-OMICs strategy dissected the effects of phagocytosis, blood-feeding, and *Plasmodium* infection, which provides a rich resource of fundamental information on mosquito hemocyte biology. These data shed light on putative of candidate granulocyte cell surface markers that can spur additional studies to support the community's effort to uncover novel aspects of granulocyte involvement in the mosquito immune response throughout the organism's life cycle.

Results

Magnetic bead-based isolation of the phagocytic hemocyte population

Several methodologies and perfusion techniques have been used to isolate mosquito hemocytes (Abraham et al. 2005; Castillo et al. 2006; Rodrigues et al. 2010; King and Hillyer 2012). In addition to variability in hemocyte numbers (Hillyer and Strand 2014), perfusion techniques are often contaminated by mosquito fat body cells, extraneous cellular debris, or bacteria as a result of this invasive technique (Castillo et al. 2006). Leveraging the ability of granulocytes to phagocytose particulates (Hillyer et al. 2003a); we hypothesized that a highly enriched population of granulocytes could be isolated following the phagocytosis of carboxylate-coated magnetic beads. Magnetic isolation would enable granulocyte purification and enrichment and thus facilitate mass spectrometry-based proteomics analysis (Figure 1A). This methodology would produce few of the contaminants of perfusion, albeit at the expense of proteomic information for the non-phagocytic populations of mosquito blood cells (prohemocytes and oenocytoids). Based on morphological classification, microscopy analyses confirmed that mosquito granulocytes heavily phagocytosed magnetic beads (Figure 1B). In contrast, perfused fat body cells did not appear to have phagocytosed magnetic beads and were only occasionally seen to have any superficial association with the beads (Figure 1B). These visible differences in morphology and phagocytic ability were further validated by immunostaining to confirm the identity of the heavily phagocytic population as granulocytes. The previously described hemocyte markers, tubulin and the fluorescent cell tracker dye, CM-DiI (King and Hillyer 2012, 2013), identify specific differences between the staining pattern of phagocytic granulocytes and those of fat body cells (Figure 1B). Additional experiments using a Notch antibody demonstrate distinct differences between cell types, with Notch expression detected in fat body nuclei but absent from perfused granulocytes (Figure 1B). These data suggest that our method of enrichment specifically targets the granulocyte population of mosquito blood cells.

The mosquito granulocyte proteomic profile in the context of homeostasis and pathogen challenge

Based upon our ability to enrich for phagocytic granulocyte populations, we performed proteomic analysis on mosquito hemocytes populations to determine the effects of granulocyte enrichment (phagocytosis), blood feeding and malaria parasite infection. Initial comparisons were made between non-selected naïve hemocytes (all cell types) and magnetic bead (mag-bead) enriched sugar-fed granulocytes (Figure 2A). Further experiments compared mag-bead enriched cell populations 48 hours after blood feeding or *P. falciparum* infection (Figure 2B). Across all samples, a total of 1128 proteins were identified, 748 of which were identified in all mag-bead purified samples (Figure 2B). Using normalized spectral counts to conduct label-free quantification, we obtained average R^2 values from pair-wise spectral counts of 0.71 for non-selected naïve hemocytes, 0.88 between sugar-fed bead selected replicates, 0.83 between blood-fed bead selected replicates, and 0.85 between *Plasmodium*-infected selected samples (Supplemental Figure S1).

Individual proteome comparisons illustrate the effects of granulocyte enrichment (Figure 2C), blood feeding (Figure 2D), and *P. falciparum* infection (Figure 2E). Of interest, widespread changes in protein abundance (both positive and negative) are observed following granulocyte enrichment, yet the effects of blood feeding and parasite infection are largely positive (Figures 2C-E). The dynamics of proteins annotated with a transmembrane domain (Figure 2F) or predicted secreted proteins (Figure 2G) mirror these observations. To further extend this analysis to components of the innate immune response, immune components belonging to serine protease inhibitor (SRPNs), clip-domain serine protease (CLIPs), thioester protein (TEPs), or leucine rich-repeat immune protein (LRIMs) gene families identified in our proteomic analysis were also followed across each of the experimental conditions (Figure 2H). Members of these gene families were influenced by phagocytosis (granulocyte enrichment) and blood feeding as expected, but showed little response to malaria parasite infection (Figure 2H). Additional analyses of the Ras superfamily of small GTPases implicated in cell proliferation and hemocyte activation (Bryant and Michel, 2014) were down-regulated in enriched sugar-fed granulocyte populations, while blood-feeding and *Plasmodium* infection produced positive changes in protein abundance (Supplemental Figure S2).

Purified granulocyte samples have a distinct and significantly enriched protein profile as compared to unselected hemocytes following phagocytosis, blood-feeding, and parasite challenge

To identify those proteins with significant changes in abundance, strict filtering criteria were applied to each comparative data set. For brevity, only statistically significant (*P<0.01*) proteins with normalized spectral counts showing greater than two fold enrichment were considered. Functional classifications were used to further define the 57 proteins identified according to these requirements (Table 1). A complete list of all statistically significant (*P<0.05*) proteins is listed in Supplemental Table S2 for each experimental treatment.

Six proteins were significantly enriched following granulocyte mag-bead purification (sugar-fed (SF) magnetic bead samples vs. non-selected (NS) hemocytes; Table 1). With widespread predicted functions, the small sample size precludes any further conclusions regarding the concerted function of these proteins (Table 1). As a result, it is unclear if the proteins identified are indicative of proteins expressed specifically in granulocyte populations or in response to phagocytosis. However, this list of granulocyte-enriched proteins following mag bead phagocytosis may represent potential cell subtype protein markers (Supplemental Table S2). Included in this list are LRIM1, AGAP011503, AGAP002593, AGAP000806 and AGAP011765, which have been proposed as hemocyte-subtype specific protein markers (Pinto et al. 2009). Our study suggests that these proteins, when used in combination will be useful in selecting and profiling mosquito phagocytic hemocyte populations. For example, RNA-FISH co-staining of AGAP000806 and AGAP012386 transcripts were able to discriminate a specific subset of hemocyte cells (Pinto et al. 2009) and in our study we observed that these two genes were both expressed as proteins in granulocytes. Moreover, it was also observed that transcript expression of AGAP011765 is not coincident with expression of AGAP012337, as RNA-FISH labeled independent cell populations. In the same manner, we identified by MS/MS only AGAP011765 in mag bead purified granulocytes and not AGAP012337 (Supplemental Table S2).

Of interest, we identified a snake-like serine protease (AGAP003691), a putative ortholog of the *Drosophila snake* protein involved in the proteolytic cascade leading to Toll activation. In addition to the role of *snake* in dorsal-ventral patterning, *snake*-like proteins have been implicated in Toll immune signaling (Irving et al. 2001), suggesting that Toll-mediated immune responses are initiated in these phagocytic cell populations. A secreted ferritin G subunit (AGAP002464) was also identified in our analysis. With presumed roles in iron transport, ferritin expression may be increased in response to the uptake of the magnetic beads, or may have anti-microbial properties as previously implicated in mosquitoes (Paskewitz et al. 2005). In addition, we detected a dramatic increase in the levels of adenylate kinase (AGAP009317) following granulocyte enrichment. Not detected in non-selected hemocytes, this change in adenylate kinase expression may reflect changes in cellular energy metabolism following bead uptake or cell proliferation as previously described in other insect systems (Chen et al. 2012).

The response to blood-feeding (blood-fed (BF) mag-bead samples vs. sugar-fed (SF) mag-bead samples) produced the largest group of proteins identified in our analysis (Table 1). Approximately one-third of these proteins (10 of 38) are components of the mosquito innate immune response (Table 1), despite the absence of parasite challenge in the non-infected blood samples. To our surprise, several well described genes influencing *Plasmodium* development were enriched, including C-type lectin 4, CTL4 (Osta et al. 2004), lysozyme c-1, LYSC1 (Kajla et al. 2011), defensin 1, DEF1 (Dimopoulos et al. 1997; Luna et al. 2006), heme peroxidase 2, HPX2 (Oliveira et al. 2012), and thioester protein 1, TEP1 (Blandin et al. 2004; Fraiture et al. 2009; Povelones et al. 2009). In addition, several proteins have presumed functions in cell signaling and metabolism, suggesting that blood-feeding triggers extensive changes in granulocyte populations. This is supported by evidence that blood feeding promotes pervasive changes to mosquito hemocyte populations (Baton et al. 2009; Castillo et al. 2011; Bryant and Michel 2014). The remaining proteins were distributed across several presumed biological roles including proteolysis, ubiquitination, metabolic enzymes, cell signaling molecules, and those of unknown function (Table 1).

In contrast with the large number of immune components enriched with a non-infectious blood meal, only two of the 13 proteins following *Plasmodium* infection were classified as having presumed roles in the immune response (Table 1). Among these two proteins, SCRBQ2 has been previously implicated as an agonist of *Plasmodium* development (González-Lázaro et al. 2009), while LRIM16A is a yet undescribed member of a family of leucine-rich repeat proteins that are closely connected to the mosquito immune system (Waterhouse et al. 2010). Two 60s ribosomal subunits (L14 and L36a) were enriched following *Plasmodium* infection, implying that pathogen challenge initiates the assembly of specific ribosomal components required for a subset of transcripts or a generalized increased in protein translation. Additional proteins with presumed roles in protein folding and vesicular transport, including the small GTPase Rab5C, may represent a dramatic cellular reorganization or intracellular communication in response to parasite infection (Table 1).

Measuring concordance between the proteomic profiles of the granulocyte cell subset and the transcriptomic profile of the general hemocyte population

To determine the correlation of our proteomic and existing hemocyte transcriptomic profiles, candidate genes responsive to granulocyte-enrichment during sugar-feeding (SF), blood-feeding (BF), or *P. falciparum*-infection (PF) (Table 1) were further examined by multiple co-inertia analyses (MCIA) (Figure 3). Using published hemocyte transcriptome data (Pinto et al, 2009), MCIA was used to examine the degree of agreement between transcript and protein abundance in our granulocyte proteomes (SF, BF, PF). Comparisons were made between transcriptional profiles of non-selected hemocytes from sugar-fed naïve mosquitoes, 24 hours after feeding with a non-invasive CTRP mutant *Plasmodium berghei* (comparable to a non-infectious blood meal), or 24 hours after feeding with wild-type *P. berghei* (Pinto et al, 2009).

Hemocyte transcript and protein profiles analyzed by MCIA were based on the log fold change between two treatments (SF, BF and PF) and is displayed as the end points of a segment, where the more similar the two profiles are, the shorter the segment; since the length is proportional to the divergence between the two datasets. If the two profiles would be identical, the length of the segment would be zero. Individual transcripts or proteins that define the comparison are highlighted for each MCIA plot (Figure 3). Despite differences in sample collection, sample time points, and the species of malaria parasite used, our MCIA analysis revealed a high level of concordance between the previously published hemocyte transcriptomes (Pinto et al, 2009) and in our own enriched granulocyte protein profiles (Figure 3). Comparisons were performed to identify global differences between hemocyte transcript and protein profiles (Figure 3A), or to examine sub-populations of immune-specific (Figure 3B) or proliferation-specific (Figure 3C) protein profiles. A table with the highlighted genes and proteins is provided in Supplemental Table S3.

The MCIA results across all three comparisons suggest that the greatest degree of post-transcriptional regulation occurs after an infectious blood meal (PFvSF), followed by the effects of blood-feeding (BFvSF). In contrast, the effects of parasite infection when compared to blood-feeding alone (PFvBF) show the highest concordance (Figure 3). Based on the pair-wise RV-coefficient, the highest post-transcriptional regulation can be observed for the immune-specific proteome (Figure 3B), implying that components of the mosquito immune response are more likely to undergo translational regulation. Similar analyses of the proliferation-specific proteome suggest that transcript and protein expression are tightly linked (Figure 3C).

In our global analysis of mosquito hemocyte transcripts/proteins, we identified a very high co-structure between the two data sets (RV-coefficient of 0.97). Of note, several unique proteins featured prominently in our MCIA comparisons (Figure 3A) that were also independently identified in our enrichment analysis (Table 1, Supplemental Table S2). We identified a Vitellogenic Carboxypeptidase (VCP)-like protein (AGAP007505) in response to *Plasmodium* infection (PFvBF) that was also significantly enriched in our PF mag-bead sample (Table S2C). The mammalian ortholog of VCP-like has been implicated in the maturation of monocytes into macrophages (Mahoney et al, 2001), and may have similar roles in mosquito hemocyte activation. Additional proteins, Vigilin (AGAP005467) and a von Willebrand factor A – domain containing protein (AGAP000545), were also identified in enriched BF-mag bead samples (Table

1). The remainder of proteins highlighted in the global hemocyte MCIA analysis did not show a significant enrichment across the sample treatments. In addition, an ATP-dependent RNA helicase (AGAP007212), a transglutaminase (AGAP009099), a eukaryotic translational initiation factor (AGAP012281), and nidogen (AGAP008193) also define the sample comparisons (Table S3). Other proteins without preliminary annotations were also identified (AGAP003610, AGAP009218, AGAP000622, AGAP002038). AGAP003610 and AGAP009218 have no known function or description and were not enriched in any of our three samples. AGAP003610 was previously found to be 3.1-fold down-regulated at 3 h and 24 h post-blood feeding (Marinotti, et al., 2006), which suggests protein translation occurring immediately following a bloodmeal in granulocytes. To date, AGAP009218 has not been detected in any transcriptomic studies analyzing mosquito or hemocyte responses to blood feeding or pathogen challenge.

Similar to Figure 3A, the MCIA analysis of immune-specific transcripts/proteins indicate a very high agreement between the two data sets with a RV-coefficient of 0.96 (Figure 3B). However, this correlation of data is lower than that observed in our global hemocyte analysis (Figure 3A). As in Figures 3A and 3B, the highest agreement between the hemocyte transcriptome and immune specific proteome was observed for PFvBF, followed by BFvSF, while the lowest agreement was observed for PFvSF. The association of the most expressed The proteins that are specifically associated with the PFvBF comparison include LRIM16A (AGAP028028), CLIPB5 (AGAP004148), CLIPA2 (AGAP011790), and LRIM15 (AGAP007045), with LRIM16A and LRIM15 showing significant enrichment in PF mag-bead samples (Table 1, Supplemental Table S2). Both are members of a leucine-rich repeat family of proteins implicated in mosquito immunity (Waterhouse et al. 2010) and contain predicted transmembrane domains, suggesting that these LRIM proteins could be candidate surface markers of activated granulocytes following *Plasmodium* infection. Previous reports have identified that CLIPA2 transcript is induced 2.1-fold in hemocytes at 24 h following *P. berghei* infection (Pinto et al 2009), providing further evidence that CLIPA2 may also be a candidate marker for granulocyte activation. Previously implicated in the melanization response (Volz et al, 2006), recent reports have identified that CLIPA2 serves as a negative regulator of TEP1 function to avoid hyper-immune activation in response to pathogen challenge (Yassine et al, 2014). In addition, CLIPA5 (AGAP011787), LRIM17 (AGAP005693), and LRIM8B (AGAP0007456) define the PFvSF comparison, while CLIPA1 (AGAP011791) and CLIPB9 (AGAP0013442) highlight the response to blood-feeding (BFvSF).

Although the proteins cluster far from the genes, the Proliferation-Specific (Figure 3C) RV-coefficient is 0.99, which is the highest agreement found in our MCIA analyses (Figure 3) and is likely due to the relatively few genes/proteins used in the analysis. As we had observed for Figures 3A-C, the highest agreement between the hemocyte transcriptome and proliferation specific proteome is observed for PFvBF, followed by BFvSF, while the lowest agreement is observed for PFvSF. The three proteins that cluster with the PFvBF comparison are Ras-related Rab7A (AGAP001617), Ras homology gene family member A (AGAP005160), and Ras-related Rab5C (AGAP007901). Of these three, only the latter two are enriched in the PF mag-bead sample (Supplemental Table S2C), and only Rab5C falls below the $P < 0.01$ stringency cutoff (Table 1). However, the MCIA features axis 1 value of -2.07 for Rab7A (Supplemental Table S3C, Figure 3D) is high and projected in the direction of the PFvBF comparison. On the MCIA plot (Figure 3D) Rab7a also sits along the border between the PFvBF and PFvSF. These observations suggest that although Rab7a was not enriched in the PF mag bead sample, it may be a potential marker for a nuanced granulocyte proliferation response to PF, and deserves future examination. We observed that Ras superfamily GTPases were down-regulated at the protein level among granulocytes in general but increased in protein abundance in response to blood–feeding and *Plasmodium* infection (Supplemental Figure S2). In fact, both Rab5C and Ras homology gene family member A protein levels in granulocytes increased 3.7-fold and 1.6-fold, respectively, at 48 h following ingestion of an infectious as opposed to a non-infectious blood meal (Supplemental Table S2C).

Orthogonal validation of transcript-protein concordance for proteins enriched in SF, BF and PF mag-beads

To determine the concordance between transcript and protein levels, we subsequently validated a subset of proteins identified in our enrichment (Table 1, Table

S2) and MCIA analyses (Figure 3, Supplemental Table S3) across all sample treatments by qRT-PCR (Figure 4). In proteins identified in SF granulocytes (Figure 4A), transcript levels of ferritin heavy chain (AGAP002465) and snake-like (AGAP003691) closely correlate with protein levels, while adenylate kinase (AGAP009317) mRNA levels are inversely related to protein abundance. Following blood-feeding, TEP1 (AGAP010815), CTL4 (AGAP005335), and DEF1 (AGAP011294) transcript levels mirror patterns of protein expression, yet small differences in transcript levels result in large changes of protein abundance (Figure 4B).

In contrast to those proteins enriched following blood-feeding, the list of enriched proteins in granulocytes following *Plasmodium* infection was surprising in that very few immune-related proteins were identified (Table 1, Supplemental Table S2C). Of these, we profiled LRIM16A (AGAP028028) and SCRBQ2 (AGAP010133) transcripts, as well Rab5C (AGAP007901) thought to be involved in proliferation and cell signaling (Figure 4C). We also examined the expression of a thioredoxin-like gene (AGAP000044) and two genes that featured prominently in our MCIA analyses, LRIM15 (AGAP007045) and VCP-like (AGAP007505), that were enriched using the less stringent (*P* <0.05) cutoff (Supplemental Table S2C). (AGAP000044) transcripts were previously shown to be highly expressed in the hemolymph (Pinto et al. 2009) and 1.2-fold upregulated in mosquitoes with low and high *P. falciparum* infections (Mendes et al. 2011). Similar to adenylate kinase expression (Figure 4A), we observed that thioredoxin-like and SCRBQ2 transcript expression was inversely related to protein abundance (Figure 4C). LRIM15 and Rab5C, both with strong contributions to our MCIA analyses (Figure 3B and 3C), displayed little variance in transcript levels among samples suggesting that both proteins may be post-transcriptionally regulated (Figure 4C). LRIM15 had been previously shown to have the greatest transcript expression in the hemolymph and was 1.3-fold upregulated in *P. berghei* infected samples (Pinto et al. 2009). Our Immune-specific MCIA analysis identified it as a top protein in the PFvBF comparison (Figure 3C). LRIM15 transcript profiles across the four groups suggest post-transcriptional regulation (Figure 4C, middle panel) with a pronounced upregulation of protein abundance in the infected PF-mag-bead sample (Supplemental Table S2C). VCP-like protein (AGAP007505) had one of the highest weights, i.e. greatest distance from the origin and projected in the same direction as the PFvBF comparison in the MCIA plot (Figure 3B). These data suggest that the proteomic data picked up this informative feature or that the microarray platform failed to detect the transcript. VCP-like transcripts have also been found to be upregulated 2-fold in high intensity *P. berghei* infections and was upregulated 1.8-fold coincident with midgut invasion between 18-24 hrs (Mendes et al. 2011). We observed that transcript and protein profiles matched closely for the unselected and SF mag-bead selected groups, yet protein levels dropped dramatically in the BF mag-bead samples before recovery in terms of concordance with transcript levels in the PF mag-bead samples (Figure 4C). Interestingly, LRIM16A transcript slightly increased across all sample treatments, while protein levels dropped in the SF and BF mag-bead samples before increasing after *Plasmodium* infection (Figure 4C). Among these comparisons in our study, we noted that the PF mag-bead group showed the greatest discordance (opposite trends) between transcript and protein expression.

Cluster analyses revealed acute and conserved proteomic changes upon blood feeding and *P. falciparum* infection

Cluster analysis was performed on the proteomic data to better understand the association of proteins with our enriched granulocyte populations across the sample treatments. Undifferentiated data showing little change across unselected and selected granulocyte proteomes was removed from analysis, resulting in six protein clusters indicative of selection and feeding status (Figure 5). Clusters 1 to 3 comprise proteins with highest levels in naïve, sugar-fed hemocytes, while clusters 4 to 6 display responses to blood-feeding and infection (Figure 5). Proteins in cluster 1 group specifically to non-selected naïve hemocyte populations, distinct from the proteins of the granulocyte-specific populations in cluster 2. Clustering also revealed that blood-feeding produced specific responses in the absence of infection (Cluster 4), or independent of infection status (Cluster 5). Responses representative of phagocytosis of magnetic beads and *Plasmodium* infection are represented by Cluster 3 and 6 respectively (Figure 5).

Proteins in each cluster (Supplemental Table S4) were further classified based on predictive gene function for comparative analysis (Figure 5). Comparisons between non-selected and mag-bead purified hemocyte samples identified several distinct groups of proteins associated with other hemocyte sub-types, the uptake of magnetic beads (phagocytosis), blood-feeding, and *Plasmodium* infection (Figure 5). While it is difficult to interpret the non-selected cell population due to the presence of at least two hemocyte-subtypes (Cluster 1), clear differences are observed between Cluster 1 and Cluster 2 suggesting that these cell populations have distinct molecular profiles. Representing the remaining cell types found in the non-selected hemocyte population, Cluster 1 had the highest abundance of proteins involved in cell metabolism and energy transport (Figure 5), suggesting that granulocytes have a much different molecular profile than other hemocyte subtypes. With shared characteristics between the non-selected and enriched granulocyte population, the data presented in Cluster 2 likely profile of naïve, sugar-fed granulocytes independent of phagocytic function.

The components of Cluster 3 likely denote the cellular profile in response to the phagocytic uptake of the magnetic beads in naïve mosquitoes. Supporting these claims, we see the highest percentage of cytoskeletal components (Figure 5) that likely reflect changes related to an increase in cellular volume or structural rearrangements associated with phagosome maturation and lysosome fusion events that accompany pathogen destruction (Lemaitre and Hoffmann 2007). Additional proteins identified linked to immune activation and the phenoloxidase cascade also clustered with the phagocytic response, including two prophenoloxidase proteins (PPO2 and PPO4) identified within this cluster (Supplemental Table S4). This is in agreement with previous reports that suggest that melanin deposition is a critical step in phagocytosis (Hillyer et al. 2003a, 2003b). However, due to the brown color of the magnetic beads in cells undergoing phagocytosis (Figure 1), the deposition of melanin on the bead surface is not easily distinguished.

In agreement with our enrichment analysis (Table 1), large numbers of immune components were regulated by blood-feeding alone (Clusters 4 and 5), independent of infection status (Figure 5, Supplemental Table S4). These immune proteins include many well-characterized proteins with integral roles in *Plasmodium* survival and clearance, including TEP1, LRIM1, and APL1, which are core components of a complement-like immune response (Fraiture et al. 2009; Povelones et al. 2009; Smith et al. 2014). We also identified an agonist of *Plasmodium* development, LRIM9 (Cluster 5), which is strongly induced by the ecdysteroid, 20-hydroxyecdysone (Upton et al. 2014). This evidence suggests that hormonal changes associated with blood-feeding may strongly influence mosquito immunity and other proteins in these clusters (Cluster 4 and 5) in the absence of infection.

Plasmodium infection also initiated dramatic changes to the hemocyte proteome (Cluster 6), yet cluster analysis revealed very few proteins associated with immune function. Instead, we identified marked increases in the translation machinery, most notably in 40S and 60S ribosomal protein subunits and translation initiation factors (Supplemental Table S4). Although unexpected, the large number of translation components implies that *Plasmodium* infection stimuli may broadly increase the translation of mRNA and the synthesis of proteins in granulocytes. These global increases in protein synthesis may be accompanied by changes in metabolic activity that promotes granulocyte activation.

Discussion

Much of our current knowledge of hemocyte biology stems from experiments performed in other insect species, leaving several important aspects of mosquito hemocyte biology unexplored. Unlike commonly studied *Drosophila* or lepidopteran insects, mosquitoes require the need for a blood meal to complete their life cycle. Through this requirement, mosquitoes are inherently poised to encounter pathogens that influence their physiology. As a result, this makes mosquito hemocyte biology unique, serving as a model for other blood-feeding insects and for comparative biology of innate immune function and the evolution of immune cells. For these reasons, we conducted a comprehensive proteogenomic analysis of a subpopulation of mosquito hemocytes before and after a blood meal in the presence or absence of *Plasmodium* infection. Without the use of genetic tools or cellular markers to distinguish hemocyte sub-populations and the effects of their individual contributions, we employed a “low-tech” strategy to enrich for phagocytic granulocyte cell populations. The efficacy of magnetic beads to purify the phagocytic populations of hemocytes was supported by microscopic data, enabling the characterization of highly-purified granulocyte sample populations by proteomic analysis. As a result, our approach removes possible contaminants associated with perfusion, as well as simplifying the interpretation of our results to a single hemocyte sub-type after enrichment and to determine the effects of infection status. Our quantitative analysis of transcript and commensurate protein profiles provide a strong foundation for the advanced study of mosquito hemocyte biology, and shed insight into their important role in cellular and humoral immunity. In addition, we observed several instances where changes in transcript abundance did not accurately reflect the protein expression profiles revealed in our study. This highlights the importance of differences between transcript and protein profiles, especially when previous descriptions of mosquito hemocyte gene function have been solely focused on transcript abundance.

In addition to their role in pathogen clearance by phagocytosis (Hillyer et al. 2003a, 2003b; King and Hillyer 2012), mosquito hemocytes are presumed to have a major role in the secretion of proteins into the hemolymph (Blandin et al. 2004; Frolet et al. 2006; Fraiture et al. 2009; Povelones et al. 2009). An initial characterization of mosquito hemolymph components identified 26 proteins (Paskewitz and Shi 2005), of which approximately two-thirds (17/26) were identified in our proteomic data. In addition, several well-described humoral immune components such as TEP1, leucine-rich repeat immune protein 1 (LRIM1), and *Anopheles Plasmodium*- responsive leucine-rich repeat 1, APL1 (Fraiture et al. 2009; Povelones et al. 2009) were also identified in our proteomic analysis, providing confirmation that the origins of several key modulators of the immune response are produced at least in part by mosquito granulocytes. Our data also posit that many modulators of the immune response or immune-related genes measured in studies focused on midgut immunity may actually be derived from hemocytes, or in this case, granulocytes attached to the midgut basal lamina. Thought to be loosely bound to the midgut basal lamina, hemocyte attachment has been previously suggested to account for midgut *TEP1* expression (Blandin et al. 2004, Vlachou et al. 2005) and likely accounts for the detection of several more hemocyte genes of interest. Thus, our data should be helpful in delineating whether mosquito innate immune responses are granulocyte or midgut-derived.

Blood-feeding has a pleiotropic impact on mosquito physiology, requiring a concerted effort to convert the blood meal into a nutrient source for egg production (Kokoza et al. 2001). These physiological changes also undoubtedly influence mosquito hemocyte populations, as blood-feeding stimulates cell proliferation and an increase in hemocyte numbers (Baton et al. 2009; Castillo et al. 2011; Bryant and Michel 2014), while *Plasmodium* infection promotes hemocyte differentiation (Rodrigues et al. 2009, Ramirez et al. 2014). To support these observations, we noted increases in the Ras family of small GTPases in response to blood-feeding and infection, not seen in naïve phagocytes. Although speculative, we believe that these Ras-like signaling components contribute to the respective hemocyte proliferation and differentiation responses described for blood-feeding and infection. Members of the Ras superfamily regulate cell growth and proliferation in all metazoans, and over-expression of Ras85D promotes over-proliferation resulting in dramatic increases of hemocyte cell numbers in *Drosophila* (Asha et al. 2003, Zetervall et al. 2004). Furthermore, Bryant et al. (2014) imply that Ras-MAPK signaling may contribute to blood-meal induced hemocyte activation, similar to the signaling pathways that define macrophage (Cook et al. 2004) and *Drosophila* hemocyte activation (Zetervall et al. 2004, Sinenko et al. 2012). Taken together, our proteomic analysis suggests the involvement of Ras-family GTPases as mediators that lead to hemocyte proliferation, activation, and differentiation.

We used MCIA analysis to integrate our proteomic data with previously published hemocyte transcriptome data (Pinto et al. 2009), to determine the levels of concordance between transcript and proteins abundance in our granulocyte proteomes. Comparisons across datasets and feeding status revealed a high level of agreement between transcript and protein levels, yet infer some level of post-transcriptional regulation in hemocytes, most pronounced following *Plasmodium* infection. Of interest, we observed that components of the mosquito innate immune system displayed the highest levels of post-transcriptional regulation in this initial analysis. Together, these results suggest that granulocytes may store select transcripts to facilitate quick responses to stimuli such as pathogen infection. As a result, future studies to explore granulocyte-specific post-transcriptional regulation are now possible, due in part to the identification and corroboration of predicted, cell population-specific proteins, which can be used to profile mosquito immune cell populations at homeostasis, blood feeding, and following pathogen challenge.

Our analyses also provide new perspectives into the regulation of the mosquito innate immune response. To our surprise, large numbers of immune components were regulated by blood-feeding alone, independent of *Plasmodium* infection. It has recently been proposed that such a response to an uninfected blood meal comprises a preemptive or anticipatory response, to combat infection of blood-borne pathogens (Upton et al. 2014). This is in agreement with previous work demonstrating that prophenoloxidase expression is influenced by 20-hydroxyecdysone in mosquito cell lines (Müller et al. 1999), suggesting that hormonal changes associated with blood-feeding may strongly influence mosquito immunity in the absence of infection. These results challenge existing paradigms that suggest that *Plasmodium* infection triggers the mosquito immune response that partly resulted in an underestimation of the contributions of hemocytes to mosquito immunity.

In summary, we provide the first proteogenomic analysis of mosquito granulocytes and demonstrate specific changes in protein abundance that corresponds with granulocyte enrichment, blood-feeding, and *Plasmodium* infection. Moreover, using MCIA analysis and qRT-PCR analysis, we illustrate that protein levels may not necessarily be in concordance with transcript expression. Our analyses also provide significant insight into the mosquito immune system, providing strong evidence that hemocytes are integral components of an anticipatory immune response as a result of blood-feeding. These data implicate hemocytes as the primary producers of hemolymph components, and granulocyte-derived proteins account for the majority of what is secreted into the hemolymph under the various conditions. The data also highlights new perspectives for the role of granulocytes following immune challenge and taken together, provide the foundation for future avenues of study in mosquito hemocyte biology.

Materials and Methods

Mosquito Rearing

The Keele strain of *Anopheles gambiae* (Hurd et al. 2005) were maintained at 27 °C and 80% relative humidity with a 14/10 h light/dark cycle. Larvae were reared in distilled water on a diet of fish food and cat food pellets, while adults were maintained on a 10% sucrose solution and housed in 8" x 8" steel cages.

Blood Feeding and *Plasmodium* Infection

Approximate five to seven day-old *An. gambiae* mosquitoes were starved overnight prior to blood feeding. For *P. falciparum* infections, NF54 isolates of *P. falciparum* gametocyte cultures were obtained from the Johns Hopkins Malaria Research Institute Parasite Core facility and the gametocytemia was determined by microscopy analysis of Giemsa stained thin blood smears. Before feeding, gametocyte samples were diluted to 0.3% gametocytemia with human RBCs. For all blood feedings (non-infected and *P. falciparum*-infected), serum was exchanged in all blood samples with heat inactivated human serum to a 45% hematocrit. Using artificial membrane feeders maintained by a circulating water bath, mosquitoes were fed on either non-infected human blood or *P. falciparum*-infected blood and maintained at 25 °C and 80% relative humidity under standard insectary conditions until hemocyte collection approximately 48 h after feeding.

Magnetic Bead Injection and Hemocyte Collection

To collect phagocytes, mosquitoes were cold anaesthetized and individual mosquitoes were injected with 0.2 µL (2mg/ml) of a suspension of 1-2 µm diameter MagnaBind Carboxyl Derivatized Beads (Thermo Scientific). Following injection, mosquitoes were returned to insectary conditions for 2 hours. Hemocytes from both non-injected and magnetic bead-injected mosquitoes were collected using a perfusion method similar to those previously described (King and Hillyer 2012; Rodrigues et al. 2010). Briefly, a small perforation was made in the abdomen of a cold anaesthetized mosquito and approximately 5 µL of anticoagulant buffer (70% Schneider's Insect medium, and 30% citrate buffer [98 mM NaOH, 186 mM NaCl, 1.7 mM EDTA, 41 mM citric acid; pH 4.5]) containing a protease inhibitor cocktail (Sigma; P8340) was injected into the thorax, causing perfusion of the circulating hemolymph through the abdominal perforation. Perfusate was collected in a siliconized tube and kept on ice for the remainder of the collection process. Samples from mosquitoes not injected with mag-bead were then centrifuged at 2,000 rcf for five minutes to pellet the cells. Perfusate from mag-bead injected mosquitoes was placed on a collection device made from a 1/4 x 1/2 inch cylindrical Neodymium alloy magnet with a pull force of approximately 8.64 pounds for 10 minutes at 4°C. To wash the cell pellets, samples were resuspended in fresh 1xPBS with protease inhibitor cocktail on ice and repeated twice prior to the collection of the final cell fractions.

Microscopy

To verify that specific hemocyte populations were responsible for magnetic bead uptake, hemocytes were examined by bright-field, phase-contrast, or fluorescence microscopy as previously described (King and Hillyer 2012, 2013). Hemocytes and fat body cells were visualized with tubulin or CM-DiI as previously described (King and Hillyer 2012, King and Hillyer 2013), or using a *Drosophila* Notch intracellular domain antibody (Developmental Studies Hybridoma Bank, C17.9C6) at a 5 µg/ml dilution.

Protein extraction

To extract total protein, the pellet was dissolved in 30 µL SDT-lysis buffer composed of 4% (w/v) SDS, 100 mM Tris/HCl, 0.1 M DTT, pH 7.6, and then boiled at 95 °C for 5 min.

Multi-Lane Combined In-gel Digestion (MLCID)

For PAGE, each lane was loaded with 30 µL of sample. After resolving on a 4-20% precast gradient gel (BioRad, Hercules, CA), the proteins were stained with Coomassie (Gel-Code Blue). Clean and sterile razors were then used to separate individual sample lanes and cut them into 8 identical slices. These slices were further cut into 1×1 mm pieces prior to de-staining, reduction and alkylation, tryptic digestion and peptide extraction (Tao et al. 2014). The extracted peptides were lyophilized by speed-vac and re-suspended in 2% acetonitrile, 97.9% water and 0.1% formic acid buffer for LC-MS/MS analysis.

LC-MS/MS

Following in-gel digestion, tryptic peptides from each biological replicate were individually analyzed. Half of each sample was injected onto an Agilent LC-MS system consisting of a 1200 LC system coupled to a 6520 Q-TOF via an HPLC Chip Cube interface. The samples were trapped and analyzed using an Agilent Polaris-HR-Chip-3C18 chip (360 nL, 180 Å C18 trap with a 75 μm i.d., 150 mm length, 180 Å C18 analytical column). Peptides were loaded onto the enrichment column by autosampler using 97% solvent A (0.1% formic acid in water) and 3% solvent B (0.1% formic acid in 90% acetonitrile) at a flow rate of 2 μL/min. Elution of peptides from the analytical column was performed using a gradient starting at 97% A at 300 nL/min. The mobile phase was 3-10% B for 4 min, 10-35% B for 56 min, 35-99% for 2 min, and maintained at 99% B for 6 min, followed by re-equilibration of the column with 3% B for 10 min. Data dependent (autoMS2) mode was used for MS acquisition by Agilent 6520 Q-TOF at 2 GHz. Precursor MS spectra were acquired from m/z 315 to 1700 and the top 4 peaks were selected for MS/MS analysis. Product scans were acquired from m/z 50 to 1700 at a scan rate of 1.5/second. A medium isolation width (~4 amu) was used, and a collision energy of slope 3.9 V/100 Da with a 2.9 V offset was applied for fragmentation. A dynamic exclusion list was applied, with precursors excluded of 0.50 min after two MS/MS spectrum was acquired.

Mass spectrometry data search and analysis

All the LC-MS/MS raw data were converted to Mascot generic Format (.mgf) by Agilent MassHunter Qualitative Analysis B.04.00. Mascot version 2.4.1, OMSSA version 2.1.9 and X!Tandem version CYCLONE 2010.12.01.1 were used to search the *Anopheles gambiae* 3.7 (14,667 sequences) protein FASTA sequence database for peptide sequence assignments using the following parameters: precursor ion mass tolerance of 50 ppm and a fragment ion mass tolerance of 0.2 daltons. Peptides were searched using fully tryptic cleavage constraints and up to two internal cleavages sites were allowed for tryptic digestion. Fixed modifications consisted of carbamidomethylation of cysteine. Variable modifications considered were oxidation of methionine residues. All the searched results were exported and then imported into the Scaffold software (Version 4.3.4, Proteome Software) for curation, label-free quantification analysis, and visualization. Scaffold's normalized spectral counting was employed to compare relative protein abundance between non-selected hemocytes (sugar-fed) and magnetic-bead enriched granulocytes (sugar-fed, blood-fed, and *Plasmodium*-infected) cell samples in each experiment as the basis for normalization of the spectral counts for all other LC-MS/MS data in that experiment. Scaffold calculates the spectrum count quantitative value by normalizing spectral counts across an experiment. The process of calculating normalized spectral counts is as follows: (a) Scaffold takes the sum of all the Total Spectrum Counts for each MS sample; (b)The sums are then scaled to the same level; and (c) Scaffold then applies the scaling factor for each sample to each protein group to produce an output with a normalized quantitative value. Overall, protein false discovery rates of less than 1% and peptide false discovery rates of less than 0.1% were obtained with Scaffold filters, and each protein has ≥ 2 unique peptides.

Data access

Our data meets all the standards regarding the Minimum Information About a Proteomics Experiment (MIAPE), and data have been deposited to the ProteomeXchange Consortium http://www.proteomexchange.org) via the PRIDE partner repository (Vizcaíno et al. 2014) with the dataset identifier PXD001507.

Gene-expression analyses

Approximately 50 naïve, sugar-fed mosquitoes were perfused with anti-coagulant buffer directly into TRIzol reagent (Invitrogen-Life Technologies) to obtain unselected hemocyte samples for RNA isolation. For mag-bead enriched samples, mosquitoes were first injected with magnetic beads and allowed to recover under insectary conditions as described above. After perfusion, magnetic enrichment, and washing, TRIzol reagent was added for RNA isolation. Total RNA was obtained using the Direct-zol RNA Mini kit (Zymo Research) according to the manufacturer's protocol. cDNA was prepared using the RevertAid First Strand cDNA Synthesis Kit (Thermo Scientific) according to the manufacturer's protocol and used for quantitative real-time PCR as previously described (Smith et al. 2012). Hemocyte gene expression was determined using gene-specific primers and normalized to levels of ribosomal protein S7 (rpS7). All qRT-PCR primers are listed in Table S5.

Multiple Co-Inertia Analysis (MCIA)

MCIA (Meng et al. 2014) is an integrative analysis method that can be applied to multiple (omics) data sets simultaneously. The data sets are matrices in which the number of features, typically stored in the rows, is much larger than the number of samples, typically stored in the columns. MCIA can be applied to multiple data sets that have matched features or matched samples. In our case, MCIA was applied to transcriptome and proteome data sets with the most optimally matched samples acquired from independent studies. In this analysis, the features refer to genes and proteins and a sample is computed as the log fold change between two treatments in a comparison between SF, BF and PF. This log fold change is computed for both genes and proteins.in both data sets.

MCIA finds maximally co-variant axes which allow the simultaneous projection of all features and samples in the same hyperspace. MCIA is a dimension reduction technique that maximizes the covariance between the data sets and the reference space. The MCIA axis selection starts with a one table ordination method such as principal component analysis, correspondence analysis or, as in this case, non-symmetrical correspondence analysis. Afterwards, the MCIA axes that maximize the squared covariance between scores of each data set on the synthetic axes are computed. The samples that share similar trends will group together in the MCIA space. Additionally, features can be projected into the MCIA space. Features highly expressed in a sample will be projected in the direction of that sample. The strength of association of a feature to a sample is directly proportional to the distance of the feature from the origin of the plot. The overall correlation between the two data sets is measured with the RV-coefficient which is a generalization of the squared Pearson correlation coefficient (Robert et al. 1976). The RV-coefficient ranges between zero and one. A RV-coefficient of zero means there is no co-structure between the two data sets. The higher the co-structure between the data sets is, the higher the RV-coefficient.

Although there are two published anopheline hemocyte transcriptome datasets that are available (Baton et al., 2009 and Pinto et al. 2009); unfortunately, the two studies used different microarray platforms. Merging the two dataset would artificially reduce the transcriptome data that can be used for MCIA to only the subset of transcripts that was measured in both analyses. This would limit the utility and primary advantage of a MCIA approach, i.e., avoid the need to subset the data sets for the analysis. As such, we decided to focus on the Pinto et al. (2009) data set, which provided all the necessary matching transcript data for the MCIA comparison to the granulocyte proteome.

Cluster and functional analyses of protein datasets

Averaged normalized spectral counts from each group were imported into Cluster 3.0 (http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm) for analysis. An arbitrary cutoff of at least one sample mean of greater than 2.0 normalized spectra was used to filter the data. This led to the inclusion of 878 of 1140 total proteins, eliminating undifferentiated data from downstream analyses. Data were centered by median and normalized on a gene-wise basis. Hierarchical clustering was performed on both genes and dataset using standard centered correlation analyses and average-linkage clustering. Following the identification of major clusters, gene IDs within each cluster were classified based on gene ontology to identify the functional categories of proteins for comparisons between clusters as previously described (Mendes et al. 2011).

Supplemental Data

Supplemental data include Figures S1-S3 and Supplemental Tables S1-S5.

Author Contributions

RCS, JGK, CB, GGT, and RRD conceptualized the study. RCS, JGK, DT, OAT, GGT and RRD designed the experiments. RCS, JGK, DT, and OAT performed the experiments. All authors contributed to data analysis, writing and final preparation of the manuscript text and figures.

Acknowledgements

The work was supported in part by the Bloomberg Family Foundation through the Johns Hopkins Malaria Research Institute (to RRD) and the Austrian BMWFJ, BMVIT, SFG, Standortagentur Tirol, and ZIT through the Austrian FFG-COMET-Funding Program FFG grant #824186 (to GGT). Additional support was provided by a JHMRI Postdoctoral fellowship (to RCS).

Disclosure Declaration

The authors declare that there are no conflicts of interest that may have influenced the conduct of the study.

Table 1. Protein enrichment in mag-bead-enriched hemocytes.

Protein Identity	Annotation	M_r	SF vs NS	BF vs SF	PF vs BF	Fold Enrichment
Innate immunity and melanization						
AGAP003691-PA	serine protease, snake-like	94	●			16.7
AGAP005625-PA	SCRASP1	147		●		53.2
AGAP005335-PA	CTL4	20		●		42.1
AGAP007347-PA	C-Type Lysozyme, LYSC1	15		●		41.5
AGAP010730-PA	Prophenoloxidase activating factor	28		●		34.4
AGAP011294-PA	defensin anti-microbial peptide, DEF1	11		●		18.4
AGAP006327-PA	LRIM6	40		●		18.2
AGAP003251-PA	CLIPB1	41		●		12.7
AGAP009033-PA	heme peroxidase HPX2	74		●		8.1
AGAP006910-PA	SRPN3	47		●		6.7
AGAP011780-PA	CLIPA4	46		●		4.9
AGAP010815-PA	TEP1	152		●		4.5
AGAP000573-PA	Clip-Domain Serine Protease	41		●		3.9
AGAP004855-PA	CLIPB13	45		●		3.7
AGAP028028-PA	LRIM16A	81			●	4.2
AGAP010133-PA	SCRBQ2	56			●	3.9
Transcription and translation						
AGAP009737-PA	Elongation factor G	83	●			15.7
AGAP004725-PA	Eukaryotic translation initiation factor 3 subunit C	112		●		23.3
AGAP005991-PA	60S ribosomal protein L14	22			●	3.0
AGAP003538-PA	60S ribosomal protein L36a	13			●	2.6
Protease function						
AGAP004534-PA	Cathepsin B precursor	37		●		51.2
AGAP004394-PA	dipeptidyl-peptidase III	81		●		20.9
Ubiquitination						
AGAP009970-PA	Cullin-associated NEDD8-dissociated protein 1	139		●		18.2
AGAP002061-PA	26S proteasome regulatory subunit N7	45		●		7.8
Protein folding and transport						
AGAP012014-PA	ADP-ribosylation factor	21		●		36.8
AGAP010251-PA	coatomer protein complex, subunit alpha, xenin	140			●	50.4
AGAP009255-PA	Sorting nexin-2	51			●	8.6
AGAP005856-PA	nodal modulator 2	131			●	5.6
Metabolic enzymes						
AGAP009317-PB	Adenylate kinase	27	●			377.1
AGAP009278-PA	phosphorylase kinase alpha/beta subunit	122		●		29.2
AGAP009173-PC	fructose-1,6-bisphosphatase I	38		●		10.3
AGAP004802-PA	4-hydroxyphenylpyruvate dioxygenase	44		●		9.3
AGAP010174-PA	oligosaccharyltransferase complex subunit alpha	52			●	19.4
Cell signaling						
AGAP009105-PA	Serine/threonine-protein phosphatase 2A	66		●		26.2
AGAP004038-PA	Fsh-Tsh-like, G-protein coupled receptor	86		●		20.5
AGAP011765-PA	Spondin-1	87		●		13.3
AGAP001600-PA	Ser/Thr protein phosphatase/nucleotidase	63		●		13.0

AGAP007699-PA	GTP-binding nuclear protein Ran	24		●		12.9
AGAP004212-PA	Calreticulin	46		●		2.3
AGAP007901-PA	Ras-related protein Rab-5C	24			●	3.7
Miscellaneous function						
AGAP001053-PD	Wings up A, troponin	25	●			57.2
AGAP004161-PA	myofilin variant C	11	●			7.3
AGAP002464-PA	secreted ferritin G subunit	26	●			4.3
AGAP010658-PA	hexamerin-like	26		●		28.3
AGAP005467-PA	vigilin	145		●		16.8
AGAP000545-PA	von Willebrand factor A - domain containing	157		●		14.2
AGAP001127-PA	leucine-rich repeat protein, P37NB	53		●		11.8
AGAP001919-PA	protein disulfide-isomerase A6	49.0			●	3.9
AGAP011244-PA	rRNA 2'-O-methyltransferase fibrillarin	33.0			●	2.9
AGAP011334-PA	Failed axon connections protein	47.0			●	2.8
AGAP001827-PA	hypoxia up-regulated 1	108			●	2.0
Unknown function						
AGAP001718-PA		22		●		65.1
AGAP009859-PA		15		●		36.4
AGAP000604-PA		12		●		21.1
AGAP005962-PA		91		●		18.0
AGAP008439-PA		58		●		10.5
AGAP007665-PA		35		●		5.4

Figure legends

Figure 1. Purification of granulocytes using magnetic microbeads. (A) Graphical overview of granulocyte separation by phagocytosis of magnetic beads. Other hemocyte cell types and common contaminants in hemocoel perfusate are not involved in phagocytosis, enabling purification of a highly-enriched granulocyte population. **(B)** Light, Phase-contrast and Fluorescent microscopy were used to verify the uptake of magnetic beads by granulocytes, but not fat body, the main contaminant present in such samples. Cell stains and conserved *Drosophila* antibodies were used to further morphologically differentiate the two cell types.

Figure 2. Proteomic analyses of phagocytic hemocyte populations. Venn diagram comparisons of protein identities from three biological replicates of each mosquito hemocyte proteomes following selection of phagocytic cell populations **(A)** or according to feeding status (naïve sugar-fed, blood-feeding, or *Plasmodium* infection) **(B)** are shown. Volcano plots of label-free quantitative analyses of protein abundance by average normalized spectral counts from three biological replicates **(C-E)**. The three graphs depict total proteins from each sample versus the appropriate reference sample according to feeding status. Levels of predicted transmembrane proteins **(F)**, secreted proteins **(G)**, and a combination of immune gene families (SRPNs, CLIPs, TEPs and LRIMs; **F**) were measured across each of the respective treatments (phagocytosis, blood-feeding, and infection). All values are depicted as the Log_2 average of normalized spectra, while significance (*P* value) is measured as the $–\text{Log}_{10}$. Dotted lines depict significance with a *P* value cutoff of 0.05.

Figure 3. Multiple Co-Inertia Analyses (MCIA) of comparisons of hemocyte transcriptome and proteomes. Using MCIA analysis, samples corresponding to our granulocyte proteomes and previously reported hemocyte transcriptomes (Pinto et al. 2009) are displayed as the global analysis of all hemocyte proteome data (**A**), immune- (**B**) or proliferation specific (**C**) subsets. Transcriptome (green circle) or proteome (red triangle) profiles are displayed for each sample comparison (*P. falciparum*-infection (PF), blood-feeding (BF), and sugar-feeding (SF)).The samples in this analysis were

computed as log fold changes between two treatments: *P. falciparum* infection referenced to blood-feeding (PFvBF), *P. falciparum* infection referenced to sugar-feeding (PFvSF) and blood-feeding referenced to sugar-feeding (BFvSF). Additionally, the most highly expressed features (genes and proteins with the greatest distance from the origin) are projected in the MCIA result plots. Due to differences between the coordinates of the comparisons and of the most expressed features plots, different axes were generated. **(A)** *Hemocyte-specific MCIA***.** MCIA is performed between hemocyte transcriptomes and granulocyte proteomes. RV-coefficient = 0.97. **(B)** *Immune-specific MCIA.* The hemocyte transcriptome is compared to the immune-specific granulocyte proteome (Supplemental Table S1D). RV-coefficient = 0.96. **(C)** *Proliferation specific MCIA.* The hemocyte transcriptome is compared to the proliferation-specific proteome (Supplemental Table S1E). RV-coefficient = 0.99.

Figure 4. qRT-PCR validation of enriched proteins. Protein candidates with significantly increased spectral counts relative to their reference sample treatment (Table 1) were evaluated by qRT-PCR to measure correlations between transcript levels and protein abundance **(A-C)**. Candidate genes with significant enrichment in phagocytic cells **(A)**, following blood-feeding **(B)**, or after *P. falciparum* infection **(C)** are displayed with the fold change in RNA (grey) or protein (colored) across each sample treatment. Each data point is the mean (+/- SEM) of three independent biological replicates. Genes examined are shown above each graph.

Figure 5. Clustering analyses reveals expression patterns indicative of feeding status and infection. Proteomic data from each sample treatment was clustered into six distinct co-expression clusters based on protein abundance (left). The analysis clearly defined feeding status (sugar- or blood-fed) into two distinct clades. Proteins within each cluster reveal distinct distributions of different molecular function groups (right).

Supplemental Figure legends

Figure S1. Correlation of biological replicates between hemocyte samples. Venn diagrams of protein identities from independent biological replicates of mag-bead enriched hemocytes from sugar-fed, blood-fed, or *Plasmodium* infected mosquitoes. Pearson-correlation identified strong reproducibility between experiments.

Figure S2. Ras family protein expression in hemocyte populations. Average normalized spectral counts of Ras superfamily GTPases across each of the respective treatments (phagocytosis, blood-feeding, and infection). All values are depicted as the Log_2 average of normalized spectra, while significance (*P* value) is measured as the –Log_{10}. Dotted lines depict significance with a *P* value cutoff of 0.05.

References

Abraham E, Pinto S, Ghosh A, Vanlandingham D, Budd A, Higgs S, Kafatos F, Jacobs-Lorena M, Michel K. 2005. An immune-responsive serpin, SRPN6, mediates mosquito defense against malaria parasites. *Proc Natl Acad Sci U S A* **102**: 16327–16332.

Asha H, Nagy I, Kovacs G, Stetson D, Ando I. 2003. Analysis of Ras-induced overproliferation in Drosophila hemocytes. *Genetics* **163**: 203–15.

Baton L, Robertson A, Warr E, Strand M, Dimopoulos G. 2009. Genome-wide transcriptomic profiling of Anopheles gambiae hemocytes reveals pathogen-specific signatures upon bacterial challenge and Plasmodium. *BMC Genomics* **13**: 1–13.

Blandin S, Shiao SH, Moita LF, Janse CJ, Waters AP, Kafatos FC, Levashina EA. 2004. Complement-like protein TEP1 is a determinant of vectorial capacity in the malaria vector Anopheles gambiae. *Cell* **116**: 661–670.

Bryant WB, Michel K. 2014. Blood feeding induces hemocyte proliferation and activation in the African malaria mosquito, Anopheles gambiae Giles. *J Exp Biol* **217**: 1238–45.

Castillo J, Brown MR, Strand MR. 2011. Blood Feeding and Insulin-like Peptide 3 Stimulate Proliferation of Hemocytes in the Mosquito Aedes aegypti ed. D.S. Schneider. *PLoS Pathog* **7**: e1002274.

Castillo JC, Robertson AE, Strand MR. 2006. Characterization of hemocytes from the mosquitoes Anopheles gambiae and Aedes aegypti. *Insect Biochem Mol Biol* **36**: 891–903.

Chen RP, Liu CY, Shao HL, Zheng WW, Wang JX, Zhao XF. 2012. Adenylate kinase 2 (AK2) promotes cell proliferation in insect development. *BMC Mol Biol*. **13**:31.

Cook AD, Braine EL, Hamilton JA. 2004. Stimulus-dependent requirement for granulocyte-macrophage colony-stimulating factor in inflammation. *J Immunol* **173**:4643-51.

Dimopoulos G, Richman A, Muller H-M, Kafatos FC. 1997. Molecular immune responses of the mosquito Anopheles gambiae to bacteria and malaria parasites. *Proc Natl Acad Sci* **94**: 11508–11513.

Fauvarque M-O, Williams MJ. 2011. Drosophila cellular immunity: a story of migration and adhesion. *J Cell Sci* **124**: 1373–82.

Fraiture M, Baxter RRHG, Steinert S, Chelliah Y, C, Frolet C, Quispe-Tintaya W, Hoffmann J a, Blandin S a, Levashina E a. 2009. Two mosquito LRR proteins function as complement control factors in the TEP1-mediated killing of Plasmodium. *Cell Host Microbe* **5**: 273–84.

Frolet C, Thoma M, Blandin S, Hoffmann JA, Levashina EA. 2006. Boosting NF-kappaB-dependent basal immunity of Anopheles gambiae aborts development of Plasmodium berghei. *Immunity* **25**: 677–685.

González-Lázaro M, Dinglasan RR, Hernández-Hernández Fde L, Rodríguez MH, Laclaustra M, Jacobs-Lorena M, Flores-Romo L. 2009. Anopheles gambiae Croquemort SCRBQ2, expression profile in the mosquito and its potential interaction with the malaria parasite Plasmodium berghei. *Insect Biochem Mol Biol* **39**:395-402

Hillyer J, Schmidt S, Christensen B. 2003a. Rapid phagocytosis and melanization of bacteria and Plasmodium sporozoites by hemocytes of the mosquito Aedes aegypti. *J Parasitol* **89**: 62–69.

Hillyer JF, Schmidt SL, Christensen BM. 2003b. Hemocyte-mediated phagocytosis and melanization in the mosquito Armigeres subalbatus following immune challenge by bacteria. *Cell Tissue Res* **313**: 117–27.

Hillyer JF, Strand MR. 2014. Mosquito hemocyte-mediated immune responses. *Curr Opin Insect Sci* 1–8.

Hurd H, Taylor PJ, Adams D, Underhill A, Eggleston P. 2005. Evaluating the costs of mosquito resistance to malaria parasites. *Evolution* **59**: 2560–72.

Irving P, Troxler L, Heuer TS, Belvin M, Kopczynski C, Reichhart JM, Hoffmann JA, Hetru C. 2001. A genome-wide analysis of immune responses in Drosophila. *Proc Natl Acad Sci U S A* **98**:15119-15124.

Kajla MK, Shi L, Li B, Luckhart S, Li J, Paskewitz SM. 2011. A new role for an old antimicrobial: lysozyme c-1 can function to protect malaria parasites in Anopheles mosquitoes. *PLoS One* **6**: e19649.

King JG, Hillyer JF. 2012. Infection-induced interaction between the mosquito circulatory and immune systems. *PLoS Pathog* **8**: e1003058.

King JG, Hillyer JF. 2013. Spatial and temporal in vivo analysis of circulating and sessile immune cells in mosquitoes: hemocyte mitosis following infection. *BMC Biol* **11**: 55.

Kokoza VA, Martin D, Ahmed A, Mienaltowski MJ, Morton CM, Raikhel AS. 2001. Transcriptional regulation of the mosquito vitellogenin gene via a blood meal-triggered cascade. *Gene* **274**: 47–65.

Lavine MD, Strand MR. 2002. Insect hemocytes and their role in immunity. *Insect Biochem Mol Biol* **32**: 1295–309.

Lemaitre B, Hoffmann J. 2007. The host defense of Drosophila melanogaster. *Annu Rev Immunol* **25**: 697–743.

Luna C, Hoa NT, Lin H, Zhang L, Nguyen HL a, Kanzok SM, Zheng L. 2006. Expression of immune responsive genes in cell lines from two different Anopheline species. *Insect Mol Biol* **15**: 721–9.

Mahoney JA, Ntolosi B, DaSilva RP, Gordon S, McKnight AJ. 2001. Cloning and characterization of CPVL, a novel serine carboxypeptidase, from human macrophages. *Genomics* **72**:243-51.

Mendes AM, Awono-Ambene PH, Nsango SE, Cohuet A, Fontenille D, Kafatos FC, Christophides GK, Morlais I, Vlachou D. 2011. Infection intensity-dependent responses of Anopheles gambiae to the African malaria parasite Plasmodium falciparum. *Infect Immun* **79**: 4708–15.

Meng C, Kuster B, Culhane AC, Gholami AM. 2014. A multivariate approach to the integration of multi-omics data sets. *BMC Bioinformatics* 15: 162

Oliveira GDA, Lieberman J, Barillas-Mury C. 2012. Epithelial nitration by a peroxidase/NOX5 system mediates mosquito antiplasmodial immunity. *Science* **335**: 856–9.

Osta MA, Christophides GK, Kafatos FC. 2004. Effects of mosquito genes on Plasmodium development. *Science* **303**: 2030-32.

Paskewitz SM, Shi L. 2005. The hemolymph proteome of Anopheles gambiae. *Insect Biochem Mol Biol* **35**: 815–24.

Pinto SB, Lombardo F, Koutsos AC, Waterhouse RM, McKay K, An C, Ramakrishnan C, Kafatos FC, Michel K. 2009. Discovery of Plasmodium modulators by genome-wide analysis of circulating hemocytes in Anopheles gambiae. *Proc Natl Acad Sci* **106**: 21270.

Povelones M, Waterhouse RMR, Kafatos FCF, Christophides GGK. 2009. Leucine-rich repeat protein complex activates mosquito complement in defense against Plasmodium parasites. *Science* **324**: 258.

Ramirez JL, Garver LS, Brayner FA, Alves LC, Rodrigues J, Molina-Cruz A, Barillas-Mury C. 2014. The Role of Hemocytes in Anopheles gambiae Antiplasmodial Immunity. *J Innate Immun* **6**: 119–28.

Robert P, Escoufier Y. 1976. A unified tool for linear multivariate statistical methods: The RV-coefficient. *Appl Stat - J Roy St C* 25:8.

Rodrigues J, Brayner FA, Alves LC, Dixit R, Barillas-Mury C. 2010. Hemocyte differentiation mediates innate immune memory in Anopheles gambiae mosquitoes. *Science* **329**: 1353–5.

Smith RC, Eappen A, Radtke A, Jacobs-Lorena M. 2012. Regulation of anti-Plasmodium immunity by a LITAF-like transcription factor in the malaria vector Anopheles gambiae. *PLoS Pathog* **8**: e1002965.

Smith RC, Vega-Rodríguez J, Jacobs-Lorena M. 2014. The Plasmodium bottleneck : malaria parasite losses in the mosquito vector. *Mem Inst Oswaldo Cruz* **109**: 1–18.

Sinenko SA, Shim J, Banerjee U. Oxidative stress in the haematopoietic niche regulates the cellular immune response in Drosophila. *EMBO Rep* **13**: 83-9.

Tao D, Ubaida-Mohien C, Mathias D, King J, Pastrana-Mena R, Tripathi A, Goldowitz I, Graham D, Moss E, Marti M, et al. 2014. Sex-partitioning of the Plasmodium falciparum stage V gametocyte proteome provides insight into falciparum-specific cell biology. *Mol Cell proteomics* **13**:2705-24.

Upton LM, Povelones M, Christophides GK. 2014. Anopheles gambiae Blood Feeding Initiates an Anticipatory Defense Response to Plasmodium berghei. *J Innate Immun*

Vizcaíno JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Ríos D, Dianes J a, Sun Z, Farrah T, Bandeira N, et al. 2014. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat Biotechnol* **32**: 223–6.

Vlachou D, Schlegelmilch T, Christophides GK, Kafatos FC. Functional genomic analysis of midgut epithelial responses in Anopheles during Plasmodium invasion. *Curr Biol* **15**:1185-95.

Volz J, Müller HM, Zdanowicz A, Kafatos FC, Osta MA. 2006. A genetic module regulates the melanization response of Anopheles to Plasmodium. *Cell Microbiol* **8**:1392-405.

Wang L, Kounatidis I, Ligoxygakis P. 2014. Drosophila as a model to study the role of blood cells in inflammation, innate immunity and cancer. *Front Cell Infect Microbiol* **3**: 113.

Waterhouse RM, Povelones M, Christophides GK. 2010. Sequence-structure-function relations of the mosquito leucine-rich repeat immune proteins. *BMC Genomics* **11**:531.

Yassine H, Kamareddine L, Chamat S, Christophides GK, Osta MA. 2014. A Serine Protease Homolog Negatively Regulates TEP1 Consumption in Systemic Infections of the Malaria Vector Anopheles gambiae. *J Innate Immun* **6**:806-18.

Zettervall CJ, Anderl I, Williams MJ, Palmer R, Kurucz E, Ando I, Hultmark D. A directed screen for genes involved in Drosophila blood cell activation. *Proc Natl Acad Sci* **101**:14192-7.

Figure 1

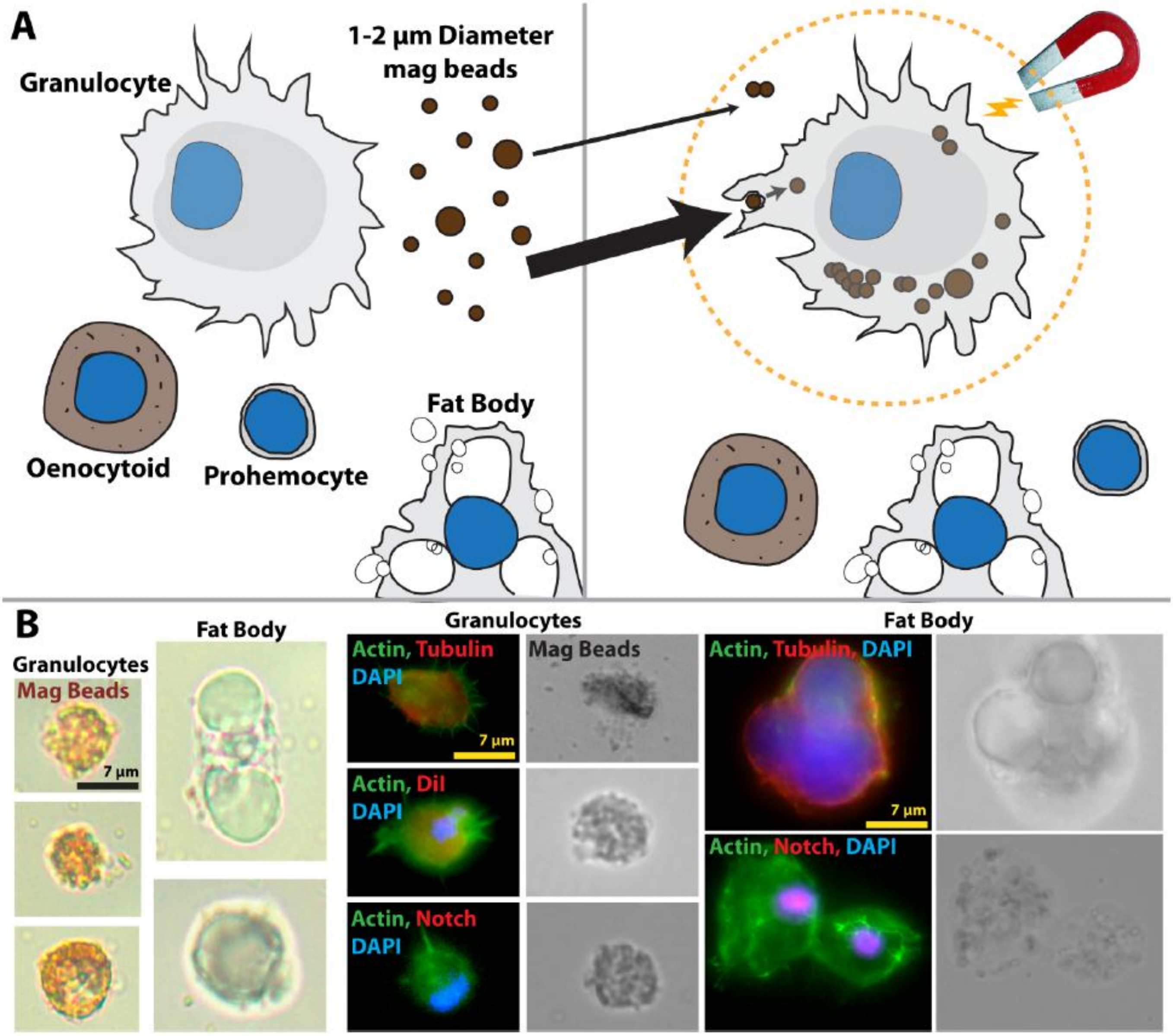

A
Granulocyte
1-2 µm Diameter
mag beads
Oenocytoid
Prohemocyte
Fat Body
B
Granulocytes
Mag Beads
7 µm
Fat Body
Granulocytes
Actin, Tubulin
DAPI
7 µm
Mag Beads
Actin, DiI
DAPI
Actin, Notch
DAPI
Fat Body
Actin, Tubulin, DAPI
7 µm
Actin, Notch, DAPI

Figure 2

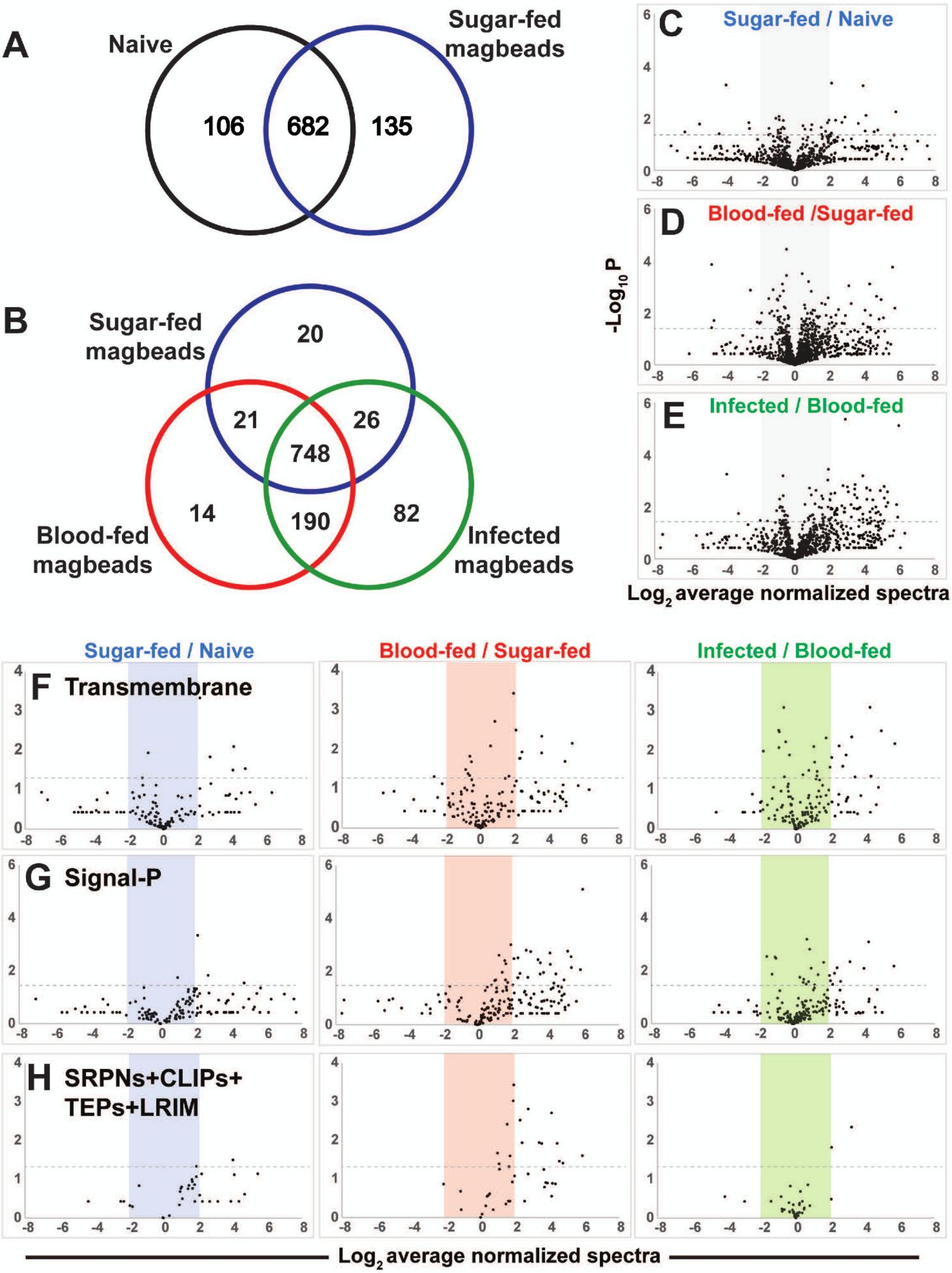
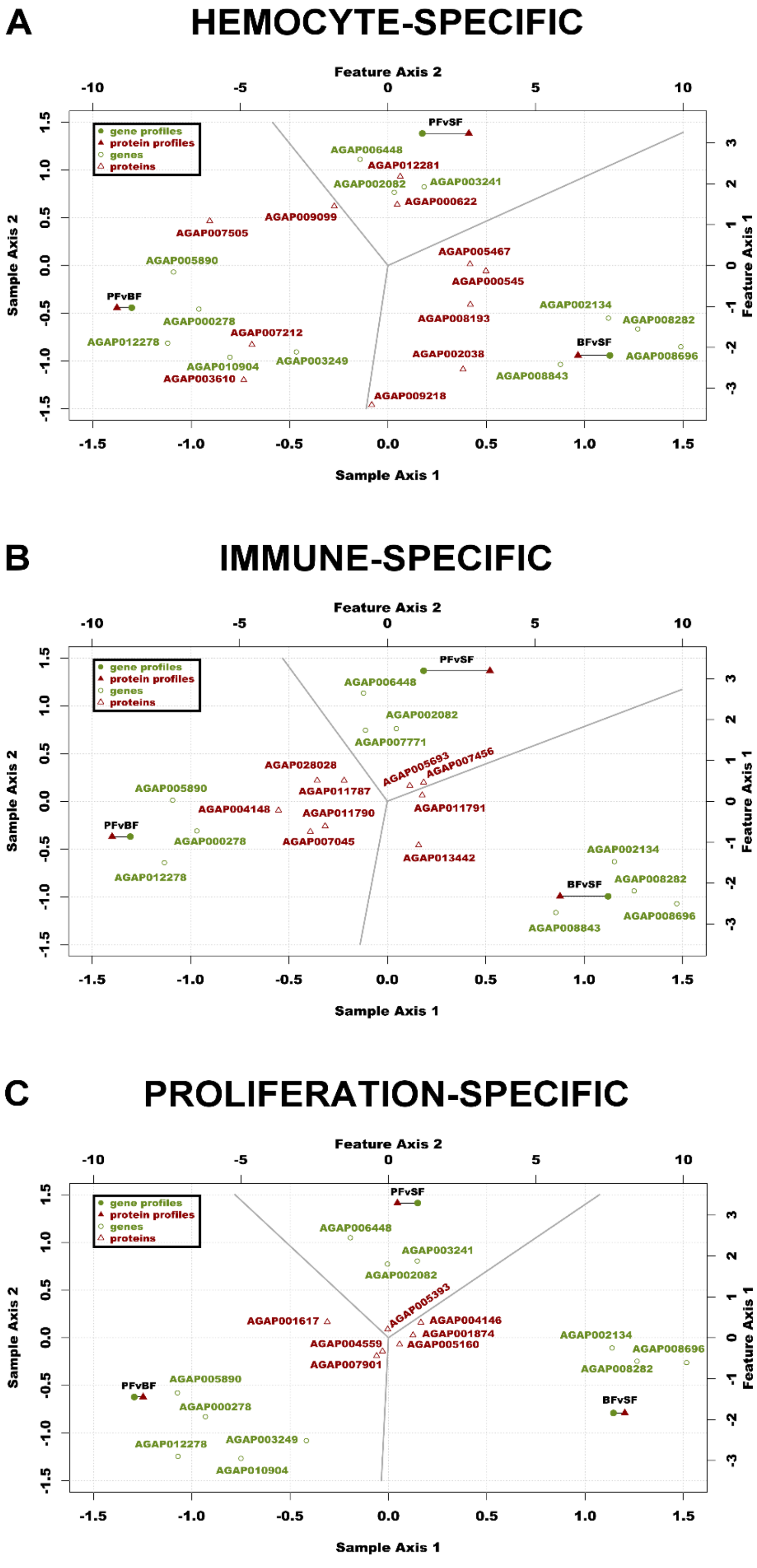

A
Naive
Sugar-fed magbeads
106
682
135
B
Sugar-fed magbeads
20
21
26
748
14
190
82
Blood-fed magbeads
Infected magbeads
C Sugar-fed / Naive
D Blood-fed /Sugar-fed
E Infected / Blood-fed
-Log10 P
Log2 average normalized spectra
Sugar-fed / Naive
Blood-fed / Sugar-fed
Infected / Blood-fed
F Transmembrane
G Signal-P
H SRPNs+CLIPs+TEPs+LRIM
-Log10 P
Log2 average normalized spectra
-8 -6 -4 -2 0 2 4 6 8
0 1 2 3 4
0 2 4 6

Figure 3

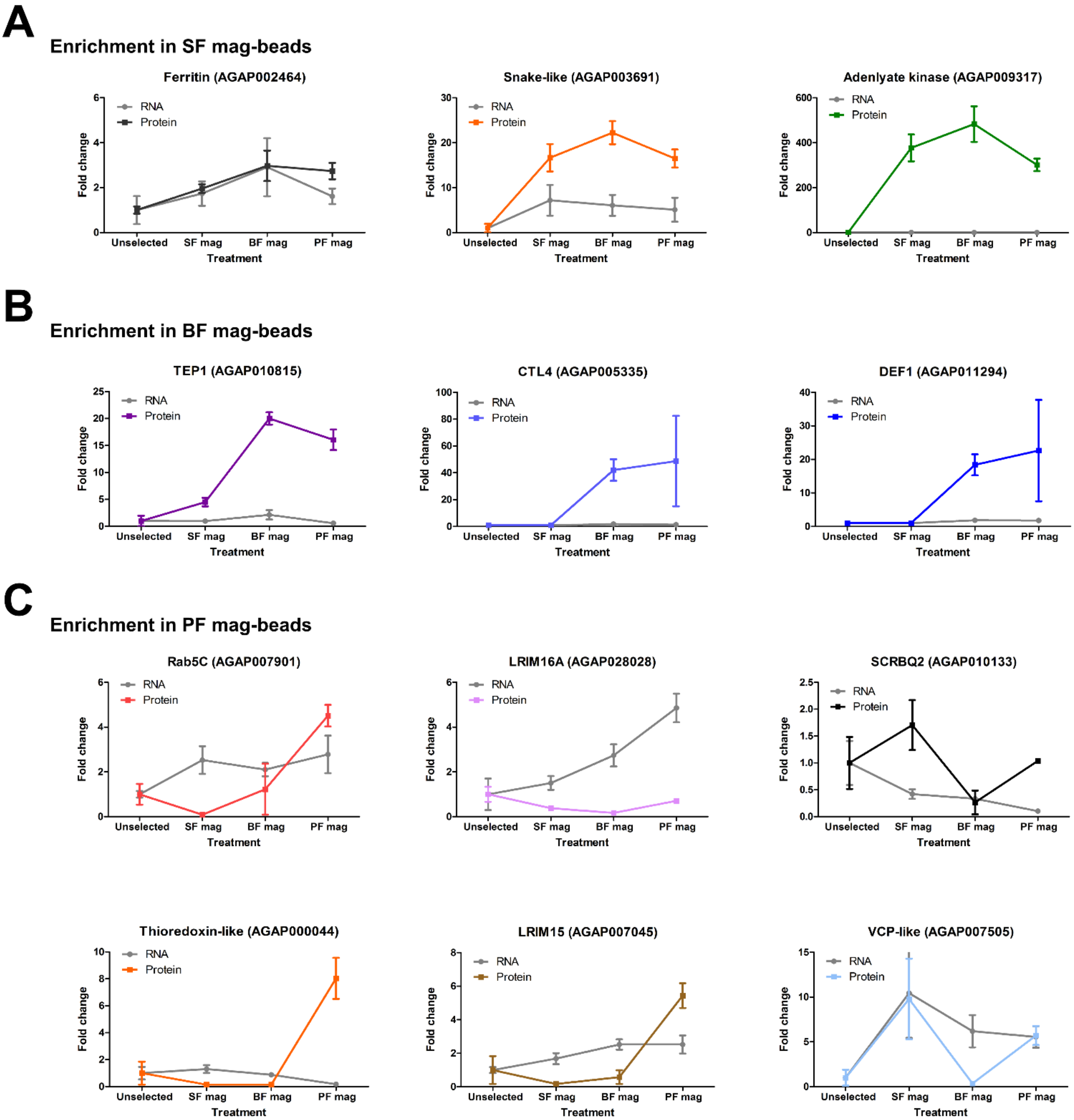


A HEMOCYTE-SPECIFIC

B IMMUNE-SPECIFIC

C PROLIFERATION-SPECIFIC

Figure 4



A Enrichment in SF mag-beads

B Enrichment in BF mag-beads

C Enrichment in PF mag-beads

Figure 5

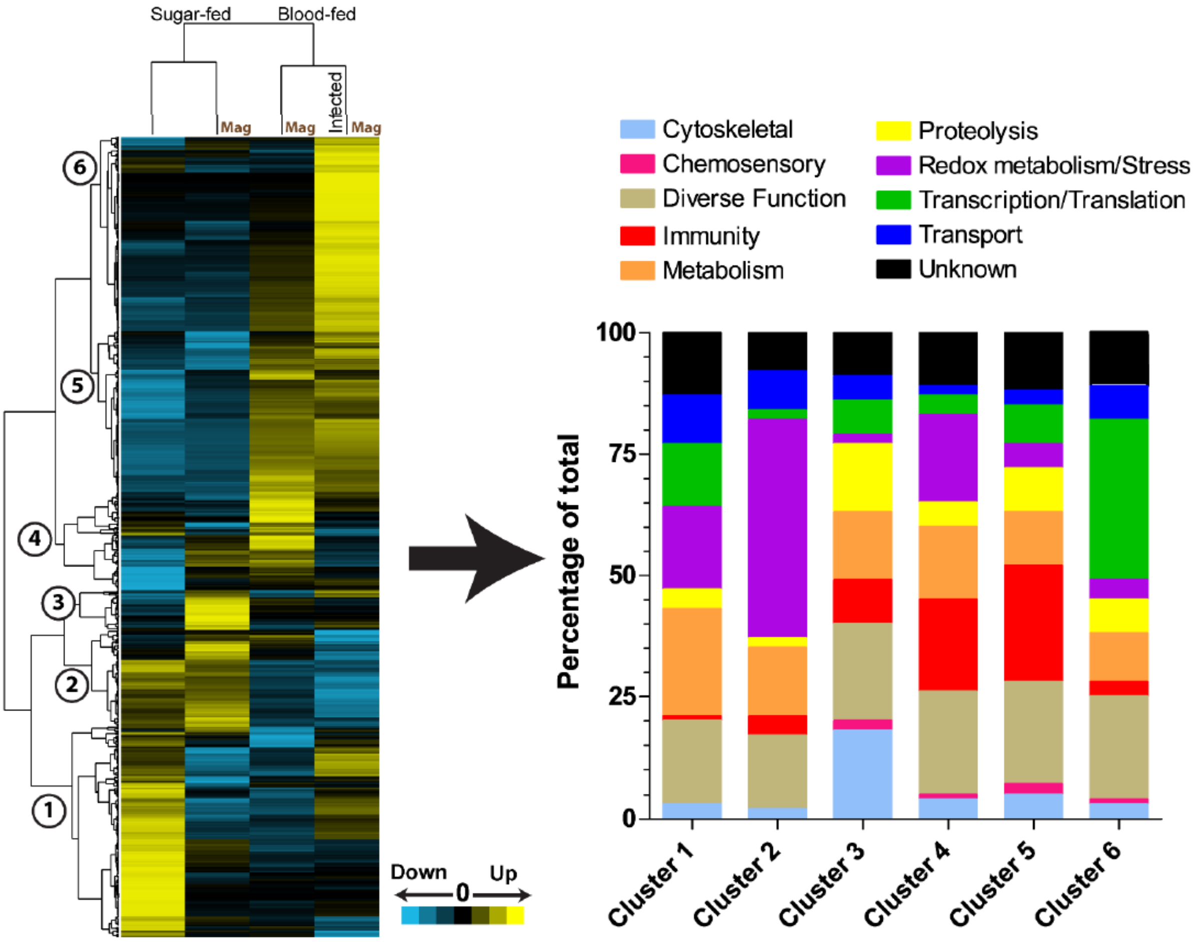
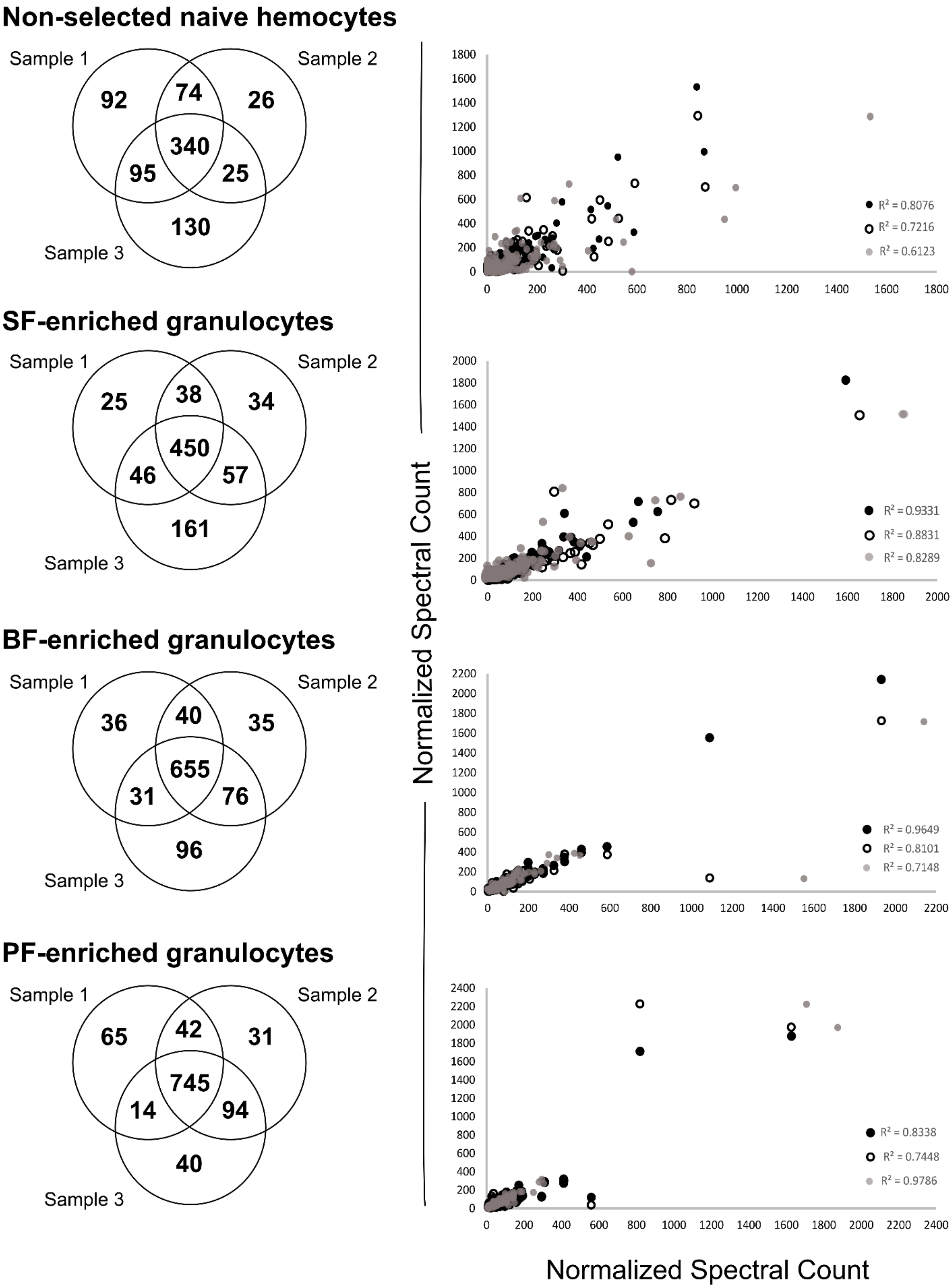
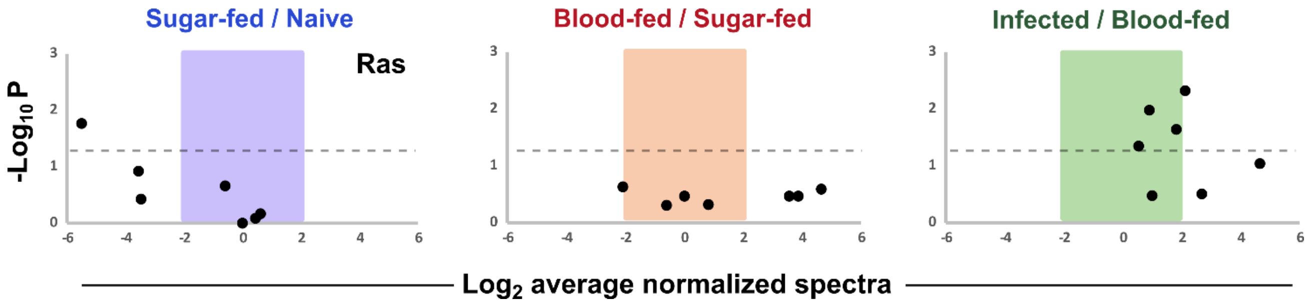

Sugar-fed
Blood-fed
Mag
Infected
Mag
Mag
6
5
4
3
2
1
Down
Up
0
Cytoskeletal
Chemosensory
Diverse Function
Immunity
Metabolism
Proteolysis
Redox metabolism/Stress
Transcription/Translation
Transport
Unknown
Percentage of total
100
75
50
25
0
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6

Figure S1



Non-selected naive hemocytes
Sample 1
Sample 2
Sample 3
92
74
26
340
95
25
130
SF-enriched granulocytes
Sample 1
Sample 2
Sample 3
25
38
34
450
46
57
161
BF-enriched granulocytes
Sample 1
Sample 2
Sample 3
36
40
35
655
31
76
96
PF-enriched granulocytes
Sample 1
Sample 2
Sample 3
65
42
31
745
14
94
40
Normalized Spectral Count
Normalized Spectral Count
0 200 400 600 800 1000 1200 1400 1600 1800
● R² = 0.8076
○ R² = 0.7216
● R² = 0.6123
0 200 400 600 800 1000 1200 1400 1600 1800 2000
● R² = 0.9331
○ R² = 0.8831
● R² = 0.8289
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200
● R² = 0.9649
○ R² = 0.8101
● R² = 0.7148
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400
● R² = 0.8338
○ R² = 0.7448
● R² = 0.9786

Figure S2


Sugar-fed / Naive
Blood-fed / Sugar-fed
Infected / Blood-fed
Ras
-Log$_{10}$ P
Log$_2$ average normalized spectra
3
2
1
0
-6
-4
-2
0
2
4
6

Dimension reduction techniques for the integrative analysis of multi-omics data

Chen Meng[1*], Oana A. Tomescu[2,3*], Gerhard G. Thallinger[2,3,4], Amin Moghaddas Gholami[5], Aedín C Culhane[6,7]

[1] Technische Universität München, Chair of Proteomics and Bioanalytics, Emil-Erlenmeyer-Forum 5, 85354 Freising, Germany

[2] Graz University of Technology, Institute for Knowledge Discovery, Bioinformatics, Petersgasse 14/V, 8010 Graz, Austria

[3] Omics Center Graz, Stiftingtalstrasse 24, 8010 Graz, Austria

[4] Austrian Centre of Industrial Biotechnology, Graz, Austria

[5] La Jolla Institute for Allergy and Immunology, 9420 Athena Circle, La Jolla, CA 92037, USA

[6] Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02115, USA

[7] Biostatistics, Harvard T.H. Chan School of Public Health, Huntington Ave, Boston, MA 02115, USA

* Joint first author

Corresponding author(s): Aedín C. Culhane, aedin@jimmy.harvard.edu, Dana-Farber Cancer Institute, 450 Brookline Ave, Boston, MA 02215, USA

Abstract

State-of-the-art next-generation sequencing, transcriptomics, proteomics, and other high-throughput "omics" technologies enable the efficient generation of very large experimental datasets. These data yield unprecedented views of molecular building blocks and the machinery of cells. Exploratory data analysis; such as clustering or dimension reduction, are an essential step in multivariate data analysis. There have been many recent developments in dimension reduction techniques; extensions of principal component analysis, that enable the simultaneous exploratory data analysis and integration of multiple high dimensional datasets. Such integrated data analysis provides an insight into the correlated structure across datasets, and may discover issues such as batch effects or outliers, and in addition to revealing known and potentially new biological knowledge. In this review, we explore dimension reduction techniques as one of the emerging approaches for data integration including meta-dimensional analyses, and how these can be applied to increase our understanding of the organization and dynamics of biological systems, in normal physiological function and disease.

Keywords

Multivariate analysis, Multi-omics data integration, Dimension reduction, integrative genomics, exploratory data analysis

Introduction

Genome scale molecular techniques including next generation sequencing and mass spectrometry, measure tens of thousands of mRNAs or proteins respectively, and are frequently applied to hundreds of biological samples. Exploratory data analysis (EDA) enables one to identify the major patterns in the data, including potential issues such as batch effects [1] and outliers and is one of the first steps in analysis of high-throughput molecular data [2]. Whilst EDA using hierarchical cluster analysis was widely applied to transcriptomics data [2], it has several limitations when studies have many samples with complex or diverse phenotypes. First it forces a hierarchical data structure on data. Second it assumes variables behave similarly over all samples and each object (variable or sample) is assigned to only one cluster. Cluster analysis, using cluster-of-cluster assignments applied to Cancer Genome Atlas Pan-cancer data of 3,527 specimens from 12 cancer type sources [3] was dominated by anatomical origin and failed to identify clusters of known biological pathways that are regulated in many cancers. While cluster analysis generally investigates pairwise distances or similarities between objects looking for fine relationships, dimension reduction or latent variable methods considers the global variance of the dataset and will thus highlight general gradients or patterns in the date [4]. Multi-table dimension reduction approaches provide a powerful and flexible approach for exploratory analysis of the correlated structure between multiple molecular data types. In this article, we will review multivariate extensions of principal component analysis (PCA) as they can be applied to the integration of multiple high dimensional datasets and EDA of the main characteristics, inter and intra-dataset correlations.

Introduction to Dimension Reduction Analysis

Dimension reduction techniques reduce data dimensionality. For example, given a dataset, **X,** one can represent each gene as a numerical vector, with *n* elements, where *n* is the number of samples. These vectors could be plotted as points in sample dimensional space ($\mathbb{R}^n$), if the number of dimensions is small (typically <3). The goal of dimension reduction techniques is to identify new vectors in this space that capture most of the variance or information in the dataset. These new vectors are plotted to provide a graphical representation of the variance of a dataset. Figure 1A shows an example of the above mentioned genes as vectors in sample space. Vice versa, the columns or samples of the matrix can be represented in gene space ($\mathbb{R}^p$). EDA using PCA plots provides a visual plot of the underlying structure of a dataset and allows visual judgment of the number of clusters.

Dimension reduction methods arose in the early part of the 20th century [5,6] and have continued to evolve, often independently in multiple fields, giving rise to myriad of associated terminology which may be confusing to beginners. PCA can be computed using different methods including eigen-analysis, latent variable analysis, factor analysis, singular value decomposition (SVD) [7] or linear regression [4]. These generate a set of vectors called principal components (PC), but depending on the scientific field and method used to compute PCA, the vectors could also be called principal axes, eigenvectors, latent variables or latent factors. The most widely used approach is SVD. SVD decomposes a matrix with n columns and p rows to three new matrices (Figure 1B); a matrix of n×q (PCs of samples), p×q (PCs of variables) and q×q diagonal matrix of singular values (square root of eigenvalues). The number of PCs is q, where the maximum q is either n or p whichever is lower. Each PC is uncorrelated (orthogonal), and has an associated eigenvalue, which indicates the amount of variance captured by each PC. The PCs from most dimension reduction approaches are ranked such that their associated eigenvalues are monotonically decreasing. Most analyses will plot and examine only the first few PCs since these explain the most variant trends in the dataset. In an experiment with little complexity, (e.g. replicates of the same cell line treated with one condition) the first component might explain most of the variance in the variables and the remaining axes may simply be attributed to noise from technical or biological sources. However a complex dataset (a set of heterogonous tumors) may require multiple principal components to capture most of the variance. When two PCs are visualized on a plot, variables and samples with higher variance will have higher weights or loadings on that component, so they will be further plotted from the origin. Objects projected in the same direction from the origin are associated.

There are many dimension reduction approaches related to PCA (Table 1) such as, correspondence analysis (COA, CA), non-symmetrical correspondence analysis (NSC), multidimensional scaling

(MDS) and principal co-ordinate analysis(PCoA). These may be computed by SVD, but differ in how the data is transformed prior to SVD [7–9]. Classical MDS is a SVD of a distance matrix similar to PCoA. Although designed for contingency tables of non-negative count data, CA and NSC, decompose a chi-squared matrix [10,11], but have been successfully applied to continuous data including gene expression and protein profiles [12,13]. As described by Fellenberg et al. gene and protein expression can be seen as an approximation of the number of corresponding molecules present in the cell during a certain measured condition [13]. Additionally, Greenacre [9] emphasized that the descriptive nature of CA and NSCA allows their application on data tables in general, not only count data. These two arguments support the suitability of CA and NSCA as analysis methods for 'omics data. While CA investigates symmetric associations between two variables, NSCA captures asymmetric relations between variables.

Non negative matrix factorization (NMF) is an approach adopted from signal processing where it was used to solve the blind source deconvolution problem and has been widely applied in clustering analysis, face recognition and text mining [14]. Similar to PCA and other decomposition approaches, NMF seeks to explain the principal sources of variance in the data using a small number of vectors, however unlike PCA, it forces positive or non-negative constraint on the resulting data matrices and secondly (similar to Independent Component Analysis [15]) does not require orthogonality or independence in the components. This allows NMF to identify overlapping patterns in components. NMF is available in the R package NMF [16]. NMF is also called self modeling curve resolution or positive matrix factorization.

Since each method performs a different data transformation before decomposition, each is optimized for specific data properties. PCA is popular but is designed for analysis of multi-normal distributed data. If data is strongly skewed or extreme outliers presented, the first few axes will only separate a few objects with extreme values instead of displaying main axes of variation. If data is unimodal or non-linear trends, one may see distortion or artifact in plots, in which the second axis is an arched function of the first axis. This is called horseshoe effect in PCA and is well described with illustration [4]. Both non-metric MDS and CA perform better than PCA in these cases [8,17]. Unlike PCA, CA can be applied to sparse count data with many zeros. Independent Component

Analysis (ICA) is a generalization of PCA that does not constrain the axes to be orthogonal and is available in the R package FastICA [18]. Spectral map analysis is related to CA, and performed comparable to CA , each outperforming PCA in identification of clusters of leukemia gene expression profiles [8]. Principal co-ordinate analysis is an SVD of a distance matrix, and thus is versatile to different data structures, for example it can be applied to a matrix of distances between binary data, and is frequently applied in the analysis of microbiome data [19].

Dimension reduction techniques could also be applied in combination with variable selection, which is attractive in analysis of omics data as it reduces the complexity when interpreting high dimensional data. Several recent extensions of PCA include variable selection, often via a regularization step or L-1 penalization (e.g. LASSO) (Shen and Huang 2008). Variables with negligible loadings are excluded, producing a sparse matrix with fewer variables. Sparse, regularized or penalized versions of PCA and related methods, have all been described [20–24].

Integrative Analysis of two datasets

A number of extensions to PCA have been described that enable simultaneous decomposition and integrative analysis of paired data matrices into the same space [20,25–28] (Table 2). These include generalized SVD (gSVD) [25], coinertia analysis [CIA 29,30]and sparse or penalized extensions to partial least square (PLS) and canonical correlation analysis (CCA) [20,26,31,32]. It is useful to categorize these methods broadly as either descriptive or predictive [28].

Predictive two-table methods define the variables of dependent datasets in terms of the explanatory or independent ones is the other dataset, and methods include CCA and constrained forms of PCA or CA such as redundancy analysis or constrained correspondence analysis [4], respectively. CCA searches for linear combinations of variables (eigenvectors) from a pair of matrices that are maximally correlated. In the case of CCA, the eigenvectors are called canonical variates and the correlations between these are the canonical correlations. Each of these predictive methods requires an inversion or correlation or covariance matrix [20,28,33], which cannot be applied when the number of variables does not exceed the sample size [31].

Given the high dimensionality of ‘omics data where p>n, application of these methods requires a regularization step, which may be accomplished by adding a ridge penalty, that is, adding a multiple of the identity matrix to the correlation matrix [20,26,34]. Penalized CCA [35], sparse CCA [36], CCA-l1, CCA-EN [26] and CCA-group sparse [37] have applied it to integrative analysis of two omics datasets. Witten et al., [20] provide an elegant comparison of various CCA extensions accompanied by a unified approach to compute both penalized CCA and sparse PCA. They use a fast and efficient implementation of regularized singular value decomposition and have implemented this in the R package penalized multivariate analysis PMA [20]. In addition, Witten and Tibshirani [38] extended the sparse CCA into a supervised framework. Supervised CCA selects variables from the two data that are not only highly correlated but also associated with a dependent variable. This method could be used to integrating two datasets and a quantitative phenotype, for example, selecting variables from both genomics and transcriptomics data and link them to drug sensitivity data. This may provide a similar solution to between group coinertia analysis [28] or spare PLS discriminant analysis [39] described below.

PLS, similar to principal component regression, is a dimension reduction coupled with a regression model, such that the eigenvectors have high covariance with a response variable [40]. Response variables can be univariate or multivariate but the latter is more challenging as it has to find the PCs that explains all of the response variables simultaneously [40]. To analyze the omics data where p>>n, similar to CCA, sparse PLS (sPLS) extensions have been described. PLS components can be calculated via kernel-PLS, iterative local regression algorithm such as nonlinear iterative PLS (NIPALS) or statistically inspired modification of PLS (SIMPLS) [40]. In a recent comparison, sPLS performed comparably to spare CCA [26]. Similar to supervised CCA, sPLS approach has also been extended to classification, sPLS discriminant analysis (sPLS-DA) is achieved by coding the response matrix Y with dummy variables and has been applied to classification and variable selection of the microarray and SNP data [39].

By contrast to predictive approaches, co-inertia analysis (CIA) is a descriptive non-constrained approach for coupling pairs of data matrices [27–29,41]. CIA can be seen as the PCA of the table of cross covariances between the variables of the two tables. CIA is performed in two steps: i) application of a dimension reduction technique such as PCA, CA or NSC to the initial datasets and ii) identification of orthogonal axes on which the projection of the datasets computed in i) are maximally covariant [30,41]. CIA does not require an inversion step, and can be applied to datasets to genomics data without regularization or penalization. Since CIA can be coupled with several dimension reduction approaches, including PCA, CA or even PCoA [28], it is flexible and can be applied to binary, categorical, discrete counts or continuous data. Co-inertia analysis is closely related to CCA [28].While CCA maximizes the correlation between eigenvectors which is known to be sensitive to detection of outliers, co-inertia maximizes the squared co-variance between eigenvectors. The relationship between CIA and Procrustes analysis [27] and CCA [28] have been well described, and a comparison of sCCA (with elastic net normalization), sPLS and coinertia is provided by Le Cao [26]. CIA and sPLS both maximize the covariance between eigenvectors and efficiently identify joint and individual variance in paired data. By contrast CCA-EN maximizes the correlation between eigenvectors and will discover effects present in both datasets, but may omit to discover strong individual effects [26]. Both sCCA and sPLS are sparse methods and variables selected by these methods are similar, whereas, CIA does not require penalization, and variables have a marginally different rank and some redundancy compared to the sparse methods [26].

Three-way and N-way PCA

Possibly the simplest multiple table or multi-block analysis is when k number of tables has the same rows and the same columns. These could be a longitudinal analysis of the same samples and same variables over time or a study of the same variables and samples in different locations, or samples from different sources. Analysis of such *variables x* samples x time data are called a 3-mode decomposition, triadic, cube or three-way table analysis, tensor decomposition, three-way PCA, three mode PCA, three-mode Factor Analysis (3MFA), Tucker-3 model, Tucker3, TUCKALS3 among others. There is a history of such analysis in ecology where counts of species and environment variables are measurements over different seasons [28,42–44], and also in psychology where different standardized tests are measured on study populations multiple times [45–48]. French statisticians developed STATIS "Structuration des TAbleaux a Trois Indices de la Statistique" (organization of three way tables in Statistics) of which X-STATIS or Partial Triadic Analysis, an analysis of K tables with the same samples and variables [49], is the simplest implementation. STATIS also includes COVSTATIS, which handles multiple covariance matrices collected on the same samples, DISTATIS, which handles multiple distance matrices collected on the same samples and generalizes metric multidimensional scaling to three way distance matrices, and Canonical-STATIS, which generalizes discriminant analysis and combines it with DISTATIS to analyze multi-table discriminant analysis problems among others [reviewed by 42], and these are available in the R package ade4. In psychometrics, Carroll and Chang's canonical decomposition (CANDECOMP) and Harshman's parallel factor analysis (PARAFAC) are the same model, proposed independently, collectively called the CP model, which is very similar to Tucker-3. Both methods are available in the R packages ThreeWay [50] and Principal Tensor Analysis on k modes , PTaK [51]. Tucker3 which can be considered a "complete" version of Partial Triadic Analysis [52]. Whist three-way PCA methods have a rich and lengthy history in other fields. CANDECOMP/PARAFAC (CP) and Tucker3 can be consider higher order generalizations of SVD and PCA [52,53]. The relationship between CP, Tucker3 and other tensor or higher decompositions are reviewed by Kolda and Bader [53]. Multi-linear subspace learning, multi-linear PCA or higher order PCA, tensor subspace analysis, tensor PCA or higher order (HO) SVD are extensions of SVD or PCA applied to multi-table, N-way, tensors or arrays of data [46,51,53–56] and have been applied in analysis of multi-omics data analysis to find the most variant variables or samples among multidimensional arrays or tensors [55,57,58]. However these tensors or N-way data decomposition may not find an optimal low rank approximation, or be orthogonally decomposable. Other tensor decompositions include INDSCAL, PARAFAC2, CANDELINC, DEDICOM, and PARATUCK2 as well as nonnegative variants of all of the above, many of these are available in the N-way Toolbox or R package threeWay [50,51,53].

Correlated Structure in multi-tables

In ‘omics studies we frequently need to study the relationships between multiple datasets or datasets with different variables (and matched samples) or datasets with different samples (and matched variables) at the same time. The cancer genome atlas generated miRNA and mRNA transcriptome (RNAseq, microarray), DNA copy number, DNA mutation, DNA methylation and proteomics molecular profiles on each tumor. The NCI60, CCLE projects have pharmacological compound profiles in addition to exome sequencing and transcriptomic profiles. Generalizations of these decomposition EDA methods to three or more data types are required, and a few have been applied to ‘omics data and the need for application of such methods is attracting more attention of the community [55,59,60]. These methods transform multiple omics data onto the same scale and project all data onto the same lower dimensional space, which facilitates the visualization, comparison and integration of data across studies (Table 3).

Simultaneous decomposition and integration of multiple matrices is more complex that an analysis of paired data. In addition to matrix preprocessing as previously described, each dataset may have different number of variables, different scale or different internal structure and thus have different variance. This might produce global scores that are dominated by one or a few datasets. Therefore, it is crucial to transform the datasets before decomposition. In the simplest k-table analysis, all matrices could be weighted to have equal weight. However it is more common to give greater weight to smaller or less redundant matrices (MFA), matrices that have more stable predictive information (PCovR) or those that share more information with other matrices (STATIS). This class of methods could be generally expressed as the following model:

$$\begin{aligned}
&\mathbf{X}_1 = \mathbf{F}\mathbf{Q}_1{}^{\mathrm{T}} + \varepsilon_1 \\
&\vdots \\
&\mathbf{X}_k = \mathbf{F}\mathbf{Q}_k{}^{\mathrm{T}} + \varepsilon_k \\
&\vdots \\
&\mathbf{X}_K = \mathbf{F}\mathbf{Q}_K{}^{\mathrm{T}} + \varepsilon_K
\end{aligned}$$

The ($\mathbf{X}_1$, …, $\mathbf{X}_k$, …, $\mathbf{X}_K$) represent *K* omics datasets. For the convenience in expression, we specify rows as the same set of samples and columns are different variables. **F** is the “global score” matrix that is common for all datasets. The columns in **F** correspond to the principal component in the analysis of single dataset by PCA and are also called axis, dimensions or latent variables. However, because the global score matrix integrate the information from multiple datasets, it is not optimized to model the structure in any single dataset, it seeks to the joint pattern defined by multiple datasets. $\mathbf{Q}_1$, …, $\mathbf{Q}_{k,\ \ldots,}$ $\mathbf{Q}_K$ are the loadings or coefficient matrices. A high positive value indicates a strong positive contribution of the corresponding variable to the “global score”.

Multiple co-inertia analysis (MCIA) is an extension of co-inertia analysis (CIA) to three or more datasets [59,61]. MCIA is based on a "covariance optimization criterion" that simultaneously projects several datasets such as gene expression and proteomics data into the same dimensional space, then transforms the diverse sets of variables in the data onto the same scale. Apart from the global score matrix, MCIA also derive a set of "block scores" using linear combination of original variables in each matrix, and the global score are then further defined as the linear combination of "block scores". Instead of maximizing the covariance between scores from two datasets as in CIA, MCIA maximize the following criterion $\sum_{k=1}^{K} \operatorname{cov}^2(\mathbf{f}_k, \mathbf{f})$ with the constraints that $||\mathbf{q}_k|| = ||\mathbf{t}|| = 1$. Therefore, the global score **f** represent the most concordant structure of multiple datasets. The calculation of MCIA could use an *ad hoc* extension of the NIPALS PCA algorithm to the multi-table scenario [62]. It is an iterative algorithm - after calculating the global scores and block loading for the first dimension, the residual matrices are calculated by removing the variance account for by the variables loading. This process is called “deflation”. For the higher order solution, the same process is applied to the residual matrices and re-iterated until the desired number of dimensions is derived. Therefore, the computational time of the algorithm relies on the number of dimensions. MCIA is implemented in R package omicade4 and has been applied to integrative analysis transcriptomic and proteomic datasets of the NCI-60 cell lines [59].

Generalized canonical correlation analysis (GCCA) [45] is closely related to MCIA, and is a generalization of CCA to multi-dataset analysis [63–65] which has also been applied to the omics data analysis [34,38]. Typically, MCIA and GCCA will produce similar results [for a more detailed comparison see 59]. GCCA maximizes the same criterion as MCIA but only constrains the unit the variance of loading vectors [62]. GCCA employs a different deflation strategy than MCIA, it calculates the residual matrices by removing the variance with respect to the block scores. Since block scores, in comparison with the global score, are better representation for each single dataset, GCCA is more likely to find common variables across datasets regardless of the different structures across datasets. Witten et al. applied sparse multiple CCA to analyze gene expression and CNV data from diffuse large B-cell lymphoma patients and successfully identified "*cis* interactions" that are both up-regulated in CNV and mRNA data [38]. When applied to ‘omics data where n<p, a variable selection step is often integrated with the GCCA approach. Tenenhaus et al., applied sparse GCCA to combine gene expression, comparative genomic hybridization, and a qualitative phenotype measured on a set of 53 children with glioma [34]. Sparse multiple CCA [38] and SGCCA [34] are available in the R packages the PMA and RGCCA respectively. Similarly a higher order implementation of spare PLS is described [56] [56]

Consensus PCA is closely related to GCCA and MCIA, but has had less exposure to the omics data community. Consensus PCA optimizes the same criterion with the other two and subject to the same constraints as MCIA [62]. The deflation step of CPCA relies on the global score. As a result, it only guarantees the orthogonal of global scores and tends to finds common patterns in the various datasets. This property makes it is more suitable for discovery of the joint pattern of multiple datasets, such as the joint clustering problem.

NMF has also been extended jointly factorize multiple matrices. This method is based on a different concept than the methods reviewed until now. In the joint NMF, the values in global score **F** and coefficient matrices (**Q**$_1$, …, **Q**$_K$) are non-negative and there is no explicit definition of block loading. An optimization algorithm is applied to minimize an objective function, typically the sum of square of errors, i.e. $f = \sum_{k=1}^{K} {\varepsilon_k}^2$. It can be considered a non-negative implementation of PARAFAC, although it has also been implemented using the Tucker model [14,66–68]. Zhang et al., 2012 apply joint NMF to a three-way analysis of DNA methylation, gene expression and miRNA expression data to identify modules in each of these regulatory layers that are associated with each other [69].

A unifying framework using the duality diagram

In the 1970s, French statisticians Cazes [70], Cailliez and Pages [71] developed a unifying framework, called the duality diagram, which provides elegant approach to formulate all dimension reduction methods in a similar way and is implemented in the R package ade4 [74]. Publications from Jean Thioulouse [28], Stephane Dray [30] De la Cruz and Holmes [72] and Escofier [73] present excellent and detailed mathematical review of the duality diagram framework. This framework is based on the statistical triplet (**X,Dp,Dn**) where **X** is a matrix with *n* rows (observations or samples) and *p* columns (variables or genes) (Figure 1C). The space ($\mathbb{R}^{p}$) contains n elements (samples), the space $\mathbb{R}^{n}$ contains *p* variables. **Dn** and **Dp** are diagonal matrices of *n*x*n* and *p*x*p.* The matrix **Dn** defines the columns weights and is used as an inner product of $\mathbb{R}^{n}$ to compute the relationships between variables. The matrix **Dp** is a *p*x*p* symmetric matrix, and is used as an inner product in $\mathbb{R}^{p}$ to quantify the distance between *n* samples. From a geometrical point of view, analyzing the statistical triplet (**X,Q,D**) can be formulated as either finding principal axes of a dataset containing n points in $\mathbb{R}^{p}$ or as finding the principal components of p points in $\mathbb{R}^{n}$. PCA in the original scale corresponds to **Dp** **=** I_{p} (identity. Euclidean metric) and **Dn** is a matrix of uniform row weights (1/*n*). CA can be formulated as a duality diagram by defining **Dn** and **Dp** as the marginal frequencies of the original matrix and by standardizing **X** so that it captures the departure from independence of the original data. Coinertia or decomposition of multiple matrices is a simple extension of this (Figure 1D). Further examples and mathematical details are available in [27,72,73]

Example Case Study

To demonstrate integration of multi-datasets using dimensions reduction, we applied MCIA to analyze mRNA, miRNA and proteomics data of a subset of cells lines in the NCI60 panel, including cell lines from melanoma (ME), leukemia (LE) and central nervous system (CNS) tumors. The graphical output from this analysis includes a plot of the sample space, variable space and data weighting space (Figures 2A-B and E).

The scree plots (Figure 2D) shows the eigenvalue of each global scores. The eigenvalues can be interpreted similarly to PCA, a higher eigenvalue equates to greater importance of a global score. Since the first two eigenvalues were significantly greater, we only visualized the cell lines and variables on the first two dimensions. The goal in EDA to visualize and explore data, rather than that prove a hypothesis, so a researcher may be subjective in their selection of the number of components. For example, one can select the “elbow points” in the plot where the slope of eigenvalues’ decreasing goes from “steep” to “flat”. Or simply, select eigenvector with eigenvalues larger than the average [75].

In the sample space (Figure 2A), samples from different datasets are distinguished by different shapes. The coordinates in data space for each cell line ($\mathbf{F}_k$ in Figure 2A) are connected by lines to the global scores (**F**). In this space, the distance between two points indicates the concordance between to samples. Short lines reflect higher concordance. Most cell lines have concordance information in each datasets (mRNA, miRNA, protein) as indicated by relatively short lines. In addition, the amount of correlated structure between the two transformed datasets can be measured with the RV coefficient [76,77]; a generalized Pearson correlation coefficient for matrices. It has values between 0 and 1, where a higher value indicates higher co-structure. Here, we observed a relative high RV coefficients across the three datasets, ranging from 0.78 to 0.84. Smilde et al. [77] recently proposed a modified RV coefficient to fix a bias of the RV coefficient towards small datasets [77].

This analysis (Figure 2) shows that cell lines originating from the same anatomical source of tissue are projected close to each other and converged into clusters. Specifically, leukemia cell lines are positively weighted on the right side of first PC, whereas the other two cell lines are on the negative end of PC1. PC2 separates the melanoma cell line and CNS cell lines. The melanoma cell line LOX-IMVI, which lacks the melanogenesis, is projected close to the origin further away from the melanoma cluster. We also observed that the proteomics profiles of leukemia cell lines SR are projected close to the melanoma cell lines. We examined within tumor type correlations to the SR cell line (Figure 2C) and observed that the SR proteomics data had higher correlation with melanoma cell lines compared to leukemia cell lines. Given that the mRNA and miRNA are closer to the other leukemia cell lines, it suggests that there was a technical error perhaps mislabeling of the proteomics data for this cell line (Figures 2A and C).

MCIA projects all variables into the same space. The variable spaces visualized the ($\mathbf{Q}_1$, …, $\mathbf{Q}_K$) in Figure 2B. Variables and samples projected in the same direction are associated. This allows one to select the variables most strongly associated with specific samples from each dataset for subsequent analysis. In our previous study [59], we already shown that the genes and proteins highly weighted on the melanoma side (positive end of second dimension) are enriched with melanogenesis functions; and genes/proteins highly weighted on the protein side is highly enriched in T-cell or immune related function. Therefore, in this example, we examined miRNA data, and selected the miRNAs with most extreme weights on the first two dimensions. Two miRNAs, miR142 and miR223 with highly weighted on the positive end of first dimension (leukemia) are commonly expressed in leukemia [78–82]. The miR142 plays an essential role in T-lymphocyte development. The miR-223 is regulated by the Notch and NF-kB signaling pathways in T-cell acute lymphoblastic leukemia [83].

The microRNA with most association to CNS cell lines was miR-409. It has been reported that this miRNA promotes the epithelial-to-mesenchymal transition (EMT) in prostate cancer [84]. Correspondingly, CNS cell lines show relative stronger mesenchymal phenotypes in NCI-60 which could relate to the high expression of this miRNA. On the positive end of the second dimension, we found miR-509, miR513 and miR506 strongly associated with melanoma cell lines, which are reported to initiate melanocyte transformation and promoting melanoma growth [85]

Challenges in integrative data analysis

Whilst biostatisticians and computational biologists are developing new extensions of dimension reduction analysis (described above) and other methods that can be used for exploratory data analysis and integration of molecular data, few gold standard or canonical test datasets exist and therefore it is often difficult to compare the performance of different methods. The community needs to define a set of test datasets for this purpose.

Whilst several dimension reduction approaches have been applied to molecular data, little consideration is often given to the underlying data structure. For example PCA is frequently applied to count data with many zeros, when CA is more appropriate. Equally there have been no systematic studies of the impact of the effects or potential loss of information when penalized methods, such as sparse CCA are used for data integration. These methods have the potential to reduce available data for downstream gene set analysis and integrative clustering analysis.

Most visualization approaches were designed for datasets with fewer variables, and visualization and interpretation of plots with thousands of variables can be complex. Within R, new packages such as ggord provide tools to plot higher quality ordination plots using ggplot. The R packages Plotly and ggplot2 let you create and share beautiful, interactive plots online. Both ggord and plotly are available on github.

Finally interpretation of long lists of biological variables (genes, proteins, miRNAs) remains a challenge and often one needs to search dispersed data sources to annotate these variables. The Bioconductor annotation project greatly facilitates quick and easy access to these within R [86,87]. One way to gain more insight into this list is to perform a gene set enrichment or pathway analysis (GSEA). An attractive feature of decomposition methods is that variable annotation can be projected into the same space to determine a score for Gene Ontology like biological processes, molecular functions and cellular compartments but also on pathways from databases like Reactome [12,13,59]. In multi-table decomposition, simultaneous analysis of multi-omics datasets will produce ranked lists of variables that are the most co-variant, variables that are highly associated with a sample or condition, and variables that are grouped together. These are on the same scale and can be concatenated to increase the power of gene set or pathway analysis [59]. After performing such an integrative GSEA it would also be informative to visualize, for example, the enriched pathway as a network by emphasizing the variables that contributed to its enrichment and the omics levels on which these were measured. An example is shown in [59 See Figure 3].

Conclusion

A rich resource of methods exist for dimension reduction and many methods have been developed in parallel by multiple fields. In this review we provide an overview of dimension reduction techniques that are both well-known (PCA) and also those which may not be widely used for the analysis of multi-omics data (CA, NSCA). We review methods for single table, two-table, 3-way and multi table analysis. Whilst numerous approaches for generalized decomposition of multi-table data or tensors have been described in other fields, few have been applied to the study of multi-omics data and few comparisons of these methods exist. There still is a significant challenge in extracting biologically and clinically actionable results from multi-omics data, however the field can leverage the rich resource of methods that other disciplines have developed.

Key Points

- There are many dimension reduction methods; extensions of principal component analysis, that can be applied to exploratory data analysis of a single data set, or integrated analysis of a pair or multiple datasets. In addition to exploratory analysis, these can be extended to cluster, supervised and discriminant analysis.
- The goal of dimension reduction is to map data onto a new set of vectors so that the variance (or information) in the data is explained by fewer vectors
- Multi-dataset methods multiple co-inertia (MCIA), multi-factor (MFA) or canonical correlations analysis (CCA) identify correlated structure between datasets with matched observations (samples). Each dataset may have different variables (genes, proteins, miRNA, mutations, drug response, etc).
- MCIA, CCA and related methods provide a visualization of consensus and incongruence in and between datasets, enabling discovery of potential outliers, batch effects or technical errors.

- Multi-dataset methods transform diverse variables from each dataset onto the same space and scale, facilitating integrative variable selection, gene set analysis, pathway and network analyses.

Biographical notes.

The authors are researches in computational and systems biology and are actively engaged in methods development for integrative analysis of multi-‘omics data. They have developed software in R/Bioconductor for dimension reduction and exploratory data analysis. Their packages include made4 (AC), Omicade4 (CM, AC, AMG), iBBiG (AC), mogsa (CM, AC, AMG).

Mr Chen Meng is a senior graduate student currently completing his thesis entitled “Application of multivariate methods to the integrative analysis of omics data” at Technische Universität München, Germany.

Ms Oana Tomescu is a senior graduate student who is will soon defend her thesis "Integrative Analysis of Omics Data. Enhancement of Existing Methods and Development of a Novel Gene Set Enrichment Approach" at the Graz University of Technology, Austria.

Dr. Amin Moghaddas Gholami is a bioinformatics scientist in the Division of Vaccine Discovery at La Jolla Institute for Allergy and Immunology. He is working on big data integrative and comparative analysis of multi-omics data. He was previously a bioinformatics group leader at Technische Universität München, where he supervised Chen Meng.

Dr. Gerhard Thallinger, is a principal investigator at Graz University of Technology, Austria. His research interests include analysis of next generation sequencing, microbiome, and lipidomics data. He supervises Oana Tomescu’s graduate studies.

Dr. Aedín Culhane is a research scientist in Biostatistics and Computational Biology at Dana-Farber Cancer Institute and Harvard Chan School of Public Health. She develops and applies methods for integrative analysis of ‘omics data in cancer. She co-supervises Chen and Oana on several projects.

Funding

Funding for this work was provided by DFCI BCB Research Scientist Developmental Funds, National Cancer Institute at the National Institutes of Health [grant numbers 2P50 CA101942-11, 1U19 AI111224-01, 1U19 AI109755-01] and Department of Defense BCRP [award number BC140682P1]. Views and opinions of, and endorsements by the author(s) do not reflect those of the US Army or the Department of Defense

R Supplement

R code to re-generate all figures in this article is available as a supplementary file.

References

1. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. 2010; 11:733–739
2. Brazma A, Culhane AC. Algorithms for gene expression analysis. 2005;
3. Hoadley KA, Yau C, Wolf DM, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 2014; 158:929–944
4. Legendre P, Legendre L. Numerical Ecology. 1998;
5. Pearson K. On lines and planes of closest fit to systems of points in space. Philos. Mag. 1901; 2:559–572
6. Hotelling H. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 1933; 24:417–441
7. Wall ME, Rechtsteiner A, Rocha LM. Singular Value Decomposition and Principal Component Analysis. Pract. Approach Microarray Data Anal. 2003; 91–109
8. Wouters L, Göhlmann HW, Bijnens L, et al. Graphical exploration of gene expression data: a comparative study of three multivariate methods. Biometrics 2003; 59:1131–1139
9. Greenacre M. Correspondence Analysis in Practice, Second Edition. 2007;
10. Beh EJ, Lombardo R. Correspondence Analysis: Theory, Practice and New Strategies. 2014;
11. . Theory and Applications of Correspondence Analysis. 1984;
12. Fagan A, Culhane AC, Higgins DG. A multivariate analysis approach to the integration of proteomic and gene expression data. Proteomics 2007; 7:2162–2171
13. Fellenberg K, Hauser NC, Brors B, et al. Correspondence analysis applied to microarray data. Proc. Natl. Acad. Sci. U. S. A. 2001; 98:10781–10786
14. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature 1999; 401:788–791
15. Comon P. Independent Component Analysis, a New Concept? Signal Process 1994; 36:287–314
16. Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics 2010; 11:367
17. Fasham MJR. A Comparison of Nonmetric Multidimensional Scaling, Principal Components and Reciprocal Averaging for the Ordination of Simulated Coenoclines, and Coenoplanes. Ecology 1977; 58:551–561
18. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw. Off. J. Int. Neural Netw. Soc. 2000; 13:411–430
19. Goodrich JK, Di Rienzi SC, Poole AC, et al. Conducting a Microbiome Study. Cell 2014; 158:250–262
20. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 2009; 10:515–534
21. Lee S, Epstein MP, Duncan R, et al. Sparse principal component analysis for identifying ancestry-informative markers in genome-wide association studies. Genet. Epidemiol. 2012; 36:293–302
22. Sill M, Saadati M, Benner A. Applying stability selection to consistently estimate sparse principal components in high-dimensional molecular data. Bioinforma. Oxf. Engl. 2015;

23. Zhao J. Efficient model selection for mixtures of probabilistic PCA via hierarchical BIC. IEEE Trans. Cybern. 2014; 44:1871–1883
24. Abdi H, Williams LJ, Valentin D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. Wiley Interdiscip. Rev. Comput. Stat. 2013; 5:149–179
25. Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc. Natl. Acad. Sci. U. S. A. 2003; 100:3351–3356
26. Lê Cao K-A, Martin PGP, Robert-Granié C, et al. Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinformatics 2009; 10:34
27. Dray S, Chessel D, Thioulouse J. Co-inertia analysis and the linking of the ecological data tables. Ecology 2003; 84:3078–3089
28. Thioulouse J. Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. Ann. Appl. Stat. 2011; 5:2300–2325
29. Culhane AC, Perrière G, Higgins DG. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics 2003; 4:59
30. Dray S. Analysing a pair of tables: coinertia analysis and duality diagrams. Vis. Verbalization Data 2014; 289–300
31. Braak CJFT. Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. Ecology 1986; 67:1167–1179
32. Hotelling H. Relations Between Two Sets of Variates. Biometrika 1936; 28:321–377
33. Hong S, Chen X, Jin L, et al. Canonical correlation analysis for RNA-seq co-expression networks. Nucleic Acids Res. 2013; 41:e95
34. Tenenhaus A, Philippe C, Guillemot V, et al. Variable selection for generalized canonical correlation analysis. Biostat. Oxf. Engl. 2014; 15:569–583
35. Waaijenborg S, Zwinderman AH. Sparse canonical correlation analysis for identifying, connecting and completing gene-expression networks. BMC Bioinformatics 2009; 10:315
36. Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. Stat. Appl. Genet. Mol. Biol. 2009; 8:Article 1
37. Lin D, Zhang J, Li J, et al. Group sparse canonical correlation analysis for genomic data integration. BMC Bioinformatics 2013; 14:245
38. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat. Appl. Genet. Mol. Biol. 2009; 8:Article28
39. Cao K-AL, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics 2011; 12:253
40. Boulesteix A-L, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief. Bioinform. 2007; 8:32–44
41. Dolédec S, Chessel D. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshw. Biol. 1994; 31:277–294
42. Abdi H, Williams LJ, Valentin D, et al. STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. Wiley Interdiscip. Rev. Comput. Stat. 2012; 4:124–167
43. Bénasséni J, Dosse MB. Analyzing multiset data by the Power STATIS-ACT method. Adv. Data Anal. Classif. 2011; 6:49–65

44. Lavit C, Escoufier Y, Sabatier R, et al. The ACT (STATIS Method). Comput Stat Data Anal 1994; 18:97–119
45. Carroll JD. Generalization of canonical correlation analysis to three or more sets of variables. Proc. Am. Psychol. Assoc. 1968; 227–228
46. Kroonenberg PM, De Leeuw J. Principal component analysis of three-mode data by means of alternating least squares algorithms. Psychometrika 1980; 45:69–97
47. Timmerman ME, Kiers HA. Three-mode principal components analysis: choosing the numbers of components and sensitivity to local optima. Br. J. Math. Stat. Psychol. 2000; 53 (Pt 1):1–16
48. Tucker LR. Some mathematical notes on three-mode factor analysis. Psychometrika 1966; 31:279–311
49. Thioulouse J, Chessel D. Les analyses multitableaux en ecologie factorielle. I : de la typologie d'etat a la typologie de fonctionnement par l'analyse triadique. Acta Oecologica Oecologia Gen. 1987; 8:463–480
50. Ferraro MAD, Kiers HAL, Giordani P. ThreeWay: Three-way component analysis. 2014;
51. Leibovici DG. Spatio-temporal multiway decomposition using principal tensor analysis on k-modes: the R package PTAk. J. Stat. Softw. 2010; 34:1–34
52. Kroonenberg PM. The analysis of multiple tables in factorial ecology. III.-three-mode principle component analyses: &#39;Analyse triadique complète&#39; Acta Oecologica 1989; 10:245–256
53. Kolda T, Bader B. Tensor Decompositions and Applications. SIAM Rev. 2009; 51:455–500
54. Henrion R. N-way principal component analysis theory, algorithms and applications. Chemom. Intell. Lab. Syst. 1994; 25:1–23
55. Ponnapalli SP, Saunders MA, Van Loan CF, et al. A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. PloS One 2011; 6:e28072
56. Zhao Q, Caiafa CF, M DP, et al. Multilinear Subspace Regression: An Orthogonal Tensor Decomposition Approach. Multilinear Subspace Regres. Orthogonal Tensor Decompos. Approach 2011; 1269–1277
57. Li W, Liu C-C, Zhang T, et al. Integrative analysis of many weighted co-expression networks using tensor computation. PLoS Comput. Biol. 2011; 7:e1001106
58. Omberg L, Golub GH, Alter O. A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. Proc. Natl. Acad. Sci. U. S. A. 2007; 104:18371–18376
59. Meng C, Kuster B, Culhane AC, et al. A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics 2014; 15:162
60. De Tayrac M, Lê S, Aubry M, et al. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. BMC Genomics 2009; 10:32
61. Chessel D, Hanafi M. Analyses de la co-inertie de $K$ nuages de points. Rev. Stat. Appliquée 1996; 44:35–60
62. Hanafi M, Kohler A, Qannari E-M. Connections between multiple co-inertia analysis and consensus principal component analysis. Chemom. Intell. Lab. Syst. 2011; 106:37–40
63. Takane Y, Hwang H, Abdi H. Regularized Multiple-Set Canonical Correlation Analysis. Psychometrika 2008; 73:753–775

64. Tenenhaus A, Tenenhaus M. Regularized Generalized Canonical Correlation Analysis. Psychometrika 2011; 76:257–284

65. Van de Velden M. On generalized canonical correlation analysis. 2011; Session IPS042:758

66. Kim H, Park H, Eldén L. Non-negative tensor factorization based on alternating large-scale non-negativity-constrained least squares. Bioinforma. Bioeng. 2007 BIBE 2007 Proc. 7th IEEE Int. Conf. On 2007; 1147–1151

67. Mørup M, Hansen LK, Arnfred SM. Algorithms for sparse nonnegative Tucker decompositions. Neural Comput. 2008; 20:2112–2131

68. Wang H-Q, Zheng C-H, Zhao X-M. jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. Bioinforma. Oxf. Engl. 2015; 31:572–580

69. Zhang S, Liu C-C, Li W, et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res. 2012; 40:9379–9391

70. Cazes P. Application de l'analyse des donn´ees au traitement de problemes geologiques. 1970;

71. Cailliez F, Pagès JP. Introduction à l'analyse des données. 1976;

72. De la Cruz O, Holmes S. THE DUALITY DIAGRAM IN DATA ANALYSIS: EXAMPLES OF MODERN APPLICATIONS. Ann. Appl. Stat. 2011; 5:2266–2277

73. Escoufier Y. The Duality Diagram: A Means for Better Practical Applications. Develoments Numer. Ecol. 1987; 139–156

74. Dray S, Dufour A. {The ade4 Package: Implementing the Duality Diagram for Ecologists}. J. Stat. Softw. 2007; 22:

75. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2010; 2:433–459

76. Holmes S. Using the Bootstrap and the RV Coefficient in the Multivariate Context. Proc. Conf. Data Anal. Learn. Symb. Numer. Knowl. 1989; 119–131

77. Smilde AK, Kiers H a. L, Bijlsma S, et al. Matrix correlations for high-dimensional data: the modified RV-coefficient. Bioinforma. Oxf. Engl. 2009; 25:401–405

78. Chiaretti S, Messina M, Tavolaro S, et al. Gene expression profiling identifies a subset of adult T-cell acute lymphoblastic leukemia with myeloid-like gene features and over-expression of miR-223. Haematologica 2010; 95:1114–1121

79. Dahlhaus M, Roolf C, Ruck S, et al. Expression and prognostic significance of hsa-miR-142-3p in acute leukemias. Neoplasma 2013; 60:432–438

80. Eyholzer M, Schmid S, Schardt JA, et al. Complexity of miR-223 regulation by CEBPA in human AML. Leuk. Res. 2010; 34:672–676

81. Lv M, Zhang X, Jia H, et al. An oncogenic role of miR-142-3p in human T-cell acute lymphoblastic leukemia (T-ALL) by targeting glucocorticoid receptor-α and cAMP/PKA pathways. Leukemia 2012; 26:769–777

82. Pulikkan JA, Dengler V, Peramangalam PS, et al. Cell-cycle regulator E2F1 and microRNA-223 comprise an autoregulatory negative feedback loop in acute myeloid leukemia. Blood 2010; 115:1768–1778

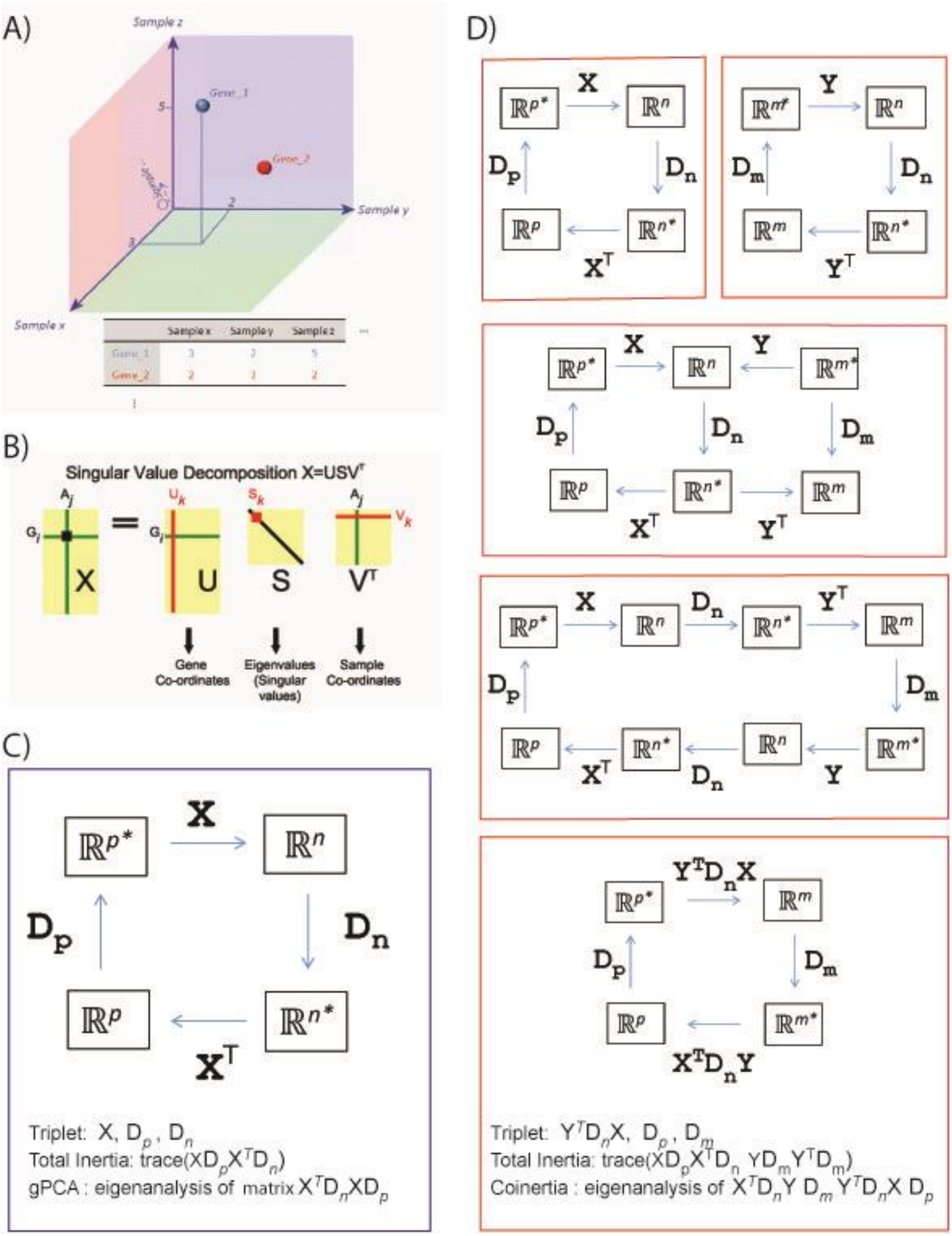
83. Kumar V, Palermo R, Talora C, et al. Notch and NF-kB signaling pathways regulate miR-223/FBXW7 axis in T-cell acute lymphoblastic leukemia. Leukemia 2014; 28:2324–2335

84. Josson S, Gururajan M, Hu P, et al. miR-409-3p/-5p promotes tumorigenesis, epithelial-to-mesenchymal transition, and bone metastasis of human prostate cancer. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 2014; 20:4636–4646
85. Streicher KL, Zhu W, Lehmann KP, et al. A novel oncogenic role for the miRNA-506-514 cluster in initiating melanocyte transformation and promoting melanoma growth. Oncogene 2012; 31:1558–1570
86. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods 2015; 12:115–121
87. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway based data integration and visualization. Bioinformatics 2013; btt285

Figure Legends

Figure 1: The goal of dimension reduction techniques is to A) identify new vectors in this space that capture most of the variance or information in the dataset; for example genes as vectors in sample space ($\mathbb{R}^{n}$). These new vectors can be found using B) SVD which decomposes a matrix with n columns and p rows to three new matrices of n×q (PCs of samples), p×q (PCs of variables) and a q×q (eigenvalues). C) The Duality Diagram framework is based on the statistical triplet (**X,Dp,Dn**) where **X** is a matrix with *n* rows (observations) and *p* columns (variables). **Dn** and **Dp** are diagonal matrices of *n*x*n* and *p*x*p.* The matrix **Dn** defines the columns weights and is used as an inner product of ^{n} to compute the relationships between variables. It is called a "duality diagram" the dual operators $\mathbf{X}^{T}\mathbf{D}_{N}\mathbf{D}_{P}$ and $\mathbf{X}\mathbf{D}_{P}\mathbf{X}^{T}\mathbf{D}_{n}$ share the same spectrum [28,30]. A generalized PCA (gPCA) is a decomposition (SVD) of $\mathbf{X}^{T}\mathbf{DnDp}$. A simple PCA is performed when D_n the matrix of uniform row weights and $\mathbf{D}_{\mathrm{p}}$ is the identity (Euclidean metric). The this framework easily extends to multi-table analysis; D) a CIA [27] can be performed on two matrices, with X and Y which share common observations (n). D_p and D_m are metrics on $\mathbb{R}^{p}$, $\mathbb{R}^{m}$ respectively. A gPCA of these triplets would be decomposition of $\mathbf{X}^{T}\mathbf{D}_{n}\mathbf{X}\mathbf{D}_{p}$ and $\mathbf{Y}^{T}\mathbf{D}_{n}\mathbf{Y}\mathbf{D}_{m}$**.** When $\mathbb{R}^{n}$ and $\mathbb{R}^{n*}$ define the same space (ie the observations in **X,Y** are matched), these two diagrams can be merged. CIA is the eigenanalysis of the crossed diagram. If the columns of both tables are centered, then the total inertia of each table is simply a sum of variances and CIA of **X** and **Y** is in this case a sum of squared covariances trace($\mathbf{X}\mathbf{D}_{p}\mathbf{X}^{T}\mathbf{D}_{n}\mathbf{Y}\mathbf{D}_{m}\mathbf{Y}^{T}\mathbf{D}_{n}$). Further details are given in Thioulouse [28], De la Cruz and Holmes [72] and Dray [30]

Figure 2 Results of analysis of a MCIA of mRNA, miRNA and proteomics molecular profiles melanoma (ME), Leukaemia (LE) and central nervous system (CNS) NCI60 cell lines, showing plot of the first two components in sample space (A), variable space (B), a scree plot of the eigenvalues (D) and a plot of data weighting space (E). There appears to be a technical issue with the LE.SR proteomics data and C) shows the correlation coefficients of SR with other cell lines.



A)
Sample z
Gene_1
Gene_2
Sample y
Sample x
Sample x Sample y Sample z
Gene_1 3 2 5
Gene_2 2 3 2
B)
Singular Value Decomposition X=USV^T
A_j G_i U_k S_k A_j V_k
X = U S V^T
Gene Co-ordinates
Eigenvalues (Singular values)
Sample Co-ordinates
C)
$\mathbb{R}^{p*}$ X $\mathbb{R}^n$ D_n $\mathbb{R}^{n*}$ X^T $\mathbb{R}^p$ D_p
Triplet: X, D_p, D_n
Total Inertia: trace($XD_pX^TD_n$)
gPCA : eigenanalysis of matrix $X^TD_nXD_p$
D)
$\mathbb{R}^{p*}$ X $\mathbb{R}^n$ D_n $\mathbb{R}^{n*}$ X^T $\mathbb{R}^p$ D_p
$\mathbb{R}^{m*}$ Y $\mathbb{R}^n$ D_n $\mathbb{R}^{n*}$ Y^T $\mathbb{R}^m$ D_m
$\mathbb{R}^{p*}$ X $\mathbb{R}^n$ Y $\mathbb{R}^{m*}$ D_p D_n D_m $\mathbb{R}^p$ X^T $\mathbb{R}^{n*}$ Y^T $\mathbb{R}^m$
$\mathbb{R}^{p*}$ X $\mathbb{R}^n$ D_n $\mathbb{R}^{n*}$ Y^T $\mathbb{R}^m$ D_m $\mathbb{R}^{m*}$ Y $\mathbb{R}^n$ D_n $\mathbb{R}^{n*}$ X^T $\mathbb{R}^p$ D_p
$\mathbb{R}^{p*}$ Y^TD_nX $\mathbb{R}^m$ D_m $\mathbb{R}^{m*}$ X^TD_nY $\mathbb{R}^p$ D_p
Triplet: Y^TD_nX, D_p, D_m
Total Inertia: trace($XD_pX^TD_n$ $YD_mY^TD_m$)
Coinertia : eigenanalysis of X^TD_nY $D_mY^TD_nX$ D_p

Figure1: Dimension reduction techniques

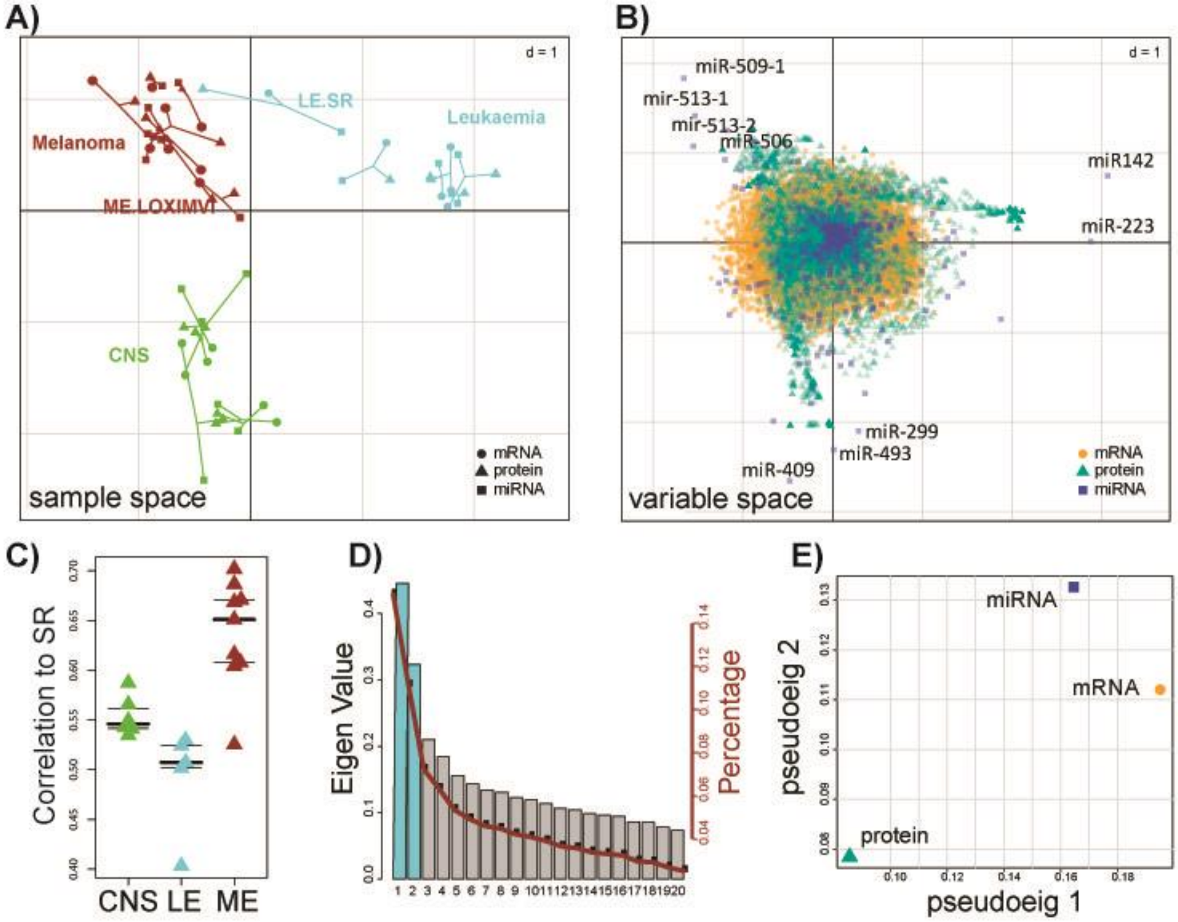


Figure 2: MCIA of mRNA, miRNA and proteomics molecular profiles of melanoma (ME), leukaemia (LE) and central nervous system (CNS) NCI60 cell lines

Table 1: Dimension reduction method for 1 dataset

Method	**Description**	**Name of function {R package}**
PCA	Principal component analysis	Prcomp {stats}, princomp {stats}, dudi.pca {ade4}, pca {vegan}, PCA {FactoMineR}, principal {psych}
CA, COA	Correspondence analysis	ca{ca}, CA{FactoMineR}, dudi.coa{ade4}
NSC	Non symmetric correspondence Analysis	dudi.nsc{ade4}

PCoA , MDS	Principal Co-ordinate Analysis/Multiple dimensional scaling	cmdscale{stats}
		dudi.pco{ade4}
		pcoa {ape}

NMF	Non-negative matrix factorization	nmf {nmf}
nmMDS	nonmetric multidimensional scaling	metaMDS {vegan}
sPCA, nsPCA, pPCA	Sparse PCA, non-negative sparse PCA, penalized PCA. (PCA with feature selection)	SPC {PMA}, spca {mixOmics}, nsprcomp {nsprcomp}, PMD {PMA}

NIPALS PCA	Non linear iterative partial least squares analysis
	(PCA on data with missing values)
		nipals {ade4}
		pca {pcaMethods}[1]
		nipals {mixOmics}

pPCA, bPCA	Probabilistic PCA, Bayesian PCA,	pca {pcaMethods}[1]
MCA	Multiple correspondence analysis	dudi.acm {ade4}, mca {MASS}
ICA,	Independent component analysis	fastICA {FastICA}
sIPCA	Sparse independent PCA (combines sPCA and ICA).	Ipca {mixOmics}
plots	Graphical resources	R packages including scatterplot3d, ggord[2], ggbiplot[3] , plotly[4]

[1] available in Bioconductor

[2] on github: devtools::install_github('fawda123/ggord')

[3] on github: devtools::install_github("ggbiplot", "vqv")

[4] on github: devtools::install_github("ropensci/plotly")

Table 2: Dimension reduction method for pairs of datasets

Method	Description	Feature Selection	R Function {package}
CCA	Canonical Correlation Analysis. Limited to n>p	No	cc(cca)
rCCA	regularized Canonical Correlation	Yes	rcc(cca)
sCCA	sparse CCA	Yes	cca(pma)
pCCA	Penalized CCA	Yes	spCCA(spCCA) supervised version

sPLS
pPLS
	sparse PLS
	penalized PLS
		Yes	spls(spls)
			spls(mixOmics)
			ppls(ppls)

sPLS-DA	sparse PLS-Discriminant Analysis	Yes	splsda(mixOmics)
cPCA	consensus PCA	No	cpca(mogsa)
CIA	coinertia analysis	No	coinertia (ade4) cia (made4)

Table 3: Dimension reduction method for multiple (>2) datasets

Method	**Description**	**Feature Selection**	**R Function {package}**
MCIA	Multiple coinertia analysis	No	mcia(omicade4)
gCCA	Generalized CCA	No	regCCA(dmt)
rGCCA	regularized Generalized CCA.	Yes	regCCA(dmt)+param
sGCCA	sparse generalized canonical correlation analysis.	Yes	sgcca(rgcca)