Oana Alina Tomescu

Eigenanalysis of dynamically important pathways in multiple omics cancer data.

Marshall Plan Report

Graz University of Technology
Institute for Knowledge Discovery
Bioinformatics Group

Harvard School of Public Health
Dana-Farber Cancer Institute
Supervisor: Aedin Culhane, PhD

Graz, November 2014

Abstract

The availability of multiple "omics" datasets from the same sample allows for a more complete understanding of pathway behavior in human diseases. However, pathway discovery is often based on flat gene lists which completely ignore the network topology of the pathways. Many methods use variations of Fisher's Exact Test to determine a data set's enrichment for a pathway while others consider only the ranks of the genes. More recent methods provide an approach to account for pathway structure but are computationally intensive and limited in their application. We propose an integrated pathway analysis approach where we combine feature (genes, proteins, CNV, etc.) scores from a multivariate analysis with an importance score for each feature in each pathway. These scores take into account the significance of each feature in the measured data sets as well as their topological importance within each pathway. We use two different measures for a feature's topological importance in a pathway and present results comparing enrichment in tumor and stroma microarray data from high grade serous ovarian cancer

Contents

1. Project Description

1.1. Background

Rapid advances in high throughput technologies have enabled quantification of multiple biological molecules at genome scale and reducing costs makes these accessible to most laboratories. Whilst many laboratories have applied genome wide gene expression profiling (transcriptomics), the network of biological pathways driving a phenotype is complex and challenging to unravel with data from a single omic screen. Increasingly, laboratories apply a multiple omics approach, generating data on multiple biological molecules, including mRNA, microRNA (miRNA), proteins, lipids, glycans, phosphoproteins and the epigenome.

The Cancer Genome Atlas (TCGA) has generated multiple molecular profiles per tumor, including gene expression on microarray platforms (Affymetrix GeneChip, Agilent microarrays), next generation RNA-sequencing (RNA-seq), profiles of protein expression, miRNA expression, exome and whole genome DNA sequencing to derive DNA mutations, copy number variation and loss of heterozygosity [36, 90]. Joint analysis of these data may provide unprecedented insights in the molecular mechanism and biological complexity of cancer, and lead to the discovery of new biomarkers of progression and response to therapy. However there are limited methods to perform integrated data analysis of multiple omics data and inferring driver biological pathways from individual data sets is challenging.

1.2. Methods

Dr. Culhane sugests the implementation and application of a dimension reduction approach, multiple co-inertia analysis (MCIA), for integrated analysis of several omics datasets. Based on a covariance optimisation criterion, MCIA enables the simultaneous projection of several datasets with matched cases into the same dimensional space. MCIA finds successive principal axes (eigenvectors) from individual principal component analyses that maximize a covariance function and it calculates the contribution of each individual to overall structure (i.e. to what extent each gene deviates or agrees with what the majority of genes support).

MCIA can be applied to meta-analysis of multiple omics datasets, where the number of features exceeds the number of observations. Datasets may have different numbers of features. MCIA reveals the features (eg genes) among multiple datasets that have highest correlation and variance. The output from MCIA is a matrix of eigenvectors for each data type which are projected and visualized in the same scale and represent the most variant features with the highest correlation. The simultaneous projection of multiple datasets on the same scale makes it possible to extract the union of concordant features across datasets, and thus is useful for omics data integration.

Through their work on integrative omics analysis, Tomescu et al. [125], show the relevance of integrative analysis techniques and highlight the need for multiple analysis methods. DI Tomescu's experience in the application and understanding of integrative analysis methods, especially the co-inertia analysis [33], shown in the study of *P. falciparum* (the parasite causing the most severe form of malaria), make her the most suitable candidate for the implementation of MCIA.

Pathway discovery is essential to progressing biomarker discovery. This is challenging when data from only one molecular level (eg gene expression) is available since many genes are regulated post-transcriptionally and gene expression may not correlate with protein activity [64]. Whilst multiple approaches that test which pathway enrichment of a given list of genes (or proteins) have been described, the simplest being a one tailed Fisher exact test or hypergeometric distribution, others account for the rank of genes

[123]. Traditional methods [123] for determining functional pathway enrichment treat pathways as a list of elements, while ignoring their inherent connectivity. Recent methods such as FunNet [102], PARADIGM [118, 129] and multi-level ontology analysis MONA [115] provide an approach to integrate multiple data types in the context of pathways but are computationally intensive or are limited in their application.

Dr. Culhane and DI Tomescu suggest the development of a network based pathway eigenanalysis approach which integrates multiple data types via MCIA, and collects and combines the most variant set of features (eg genes, proteins, miRNAs) from each platform and projects these onto a network of biological pathways such that the most variant pathway, which is most likely to perturb the network, can be extracted.

Using the network of pathways curated by the Reactome project [31] scores will be assigned to pathway elements based on their contribution to the information flow in the network. This flow-based approach [105] rewards both highly linked hubs and bottlenecks nodes which may have few connections but bridge different clusters within the network. The dynamical importance for a given pathway will then be quantified by correlating the scores from the network analysis with the MCIA principal component.

1.3. Specific aims

Aim 1: Develop an eigenanalysis approach based on MCIA to predict the dynamical importance of functional pathways and apply it as a new integrated analysis method to multiple omics ovarian cancer data.

Aim 2: Implementation of the eigenanalysis method as a platform independent Bioconductor/R package.

This is an innovative project that will develop a new computational approach to pathway discovery and it addresses current bottlenecks in the field of cancer research. We apply a simple linear algebra eigenanalysis approach that is scalable and potentially computationally efficient in pathway discovery across multiple high throughput omics data.

Aim 1

Aim 1 consists of the development of an eigenanalysis approach to predict the dynamical importance of Reactome pathways through integrated analysis of multiple omics data and it's application to the TCGA serous ovarian cancer data. The following steps are required:

- Extract eigenvectors of dynamical importance from Reactome pathways: i) Download flat files of the Reactome database [31], ii) build a network adjacency matrix using the Bioconductor *graph* package, iii) calculate the information flow associated with the network adjacency matrix which is defined as the largest eigenvalue of the adjacency matrix [105]. Steps ii) and iii) will be applied to extract eigenvalues associated with each pathway in Reactome. In addition, the dynamical importance of each pathway in the network will be calculated as the relative change in the largest eigenvalue of the network adjacency matrix upon its removal. For each pathway exclude its nodes, repeating steps ii) and iii), to discover the change in the entire network upon removal of the pathway. This provides an objective quantification of the relative importance of each pathway in the network.
- Extract eigenvectors of the most covariant features in the high grade serous ovarian cancer: Download TCGA ovarian data (microarray, RNA-Seq gene expression, proteomics, miRNA, etc) and apply MCIA to these datasets. The eigenvector associated with the first eigenvalue from each dataset of features will be concatenated to build a union of all features (gene, proteins, miRNAs).
- Score Reactome pathways: Multiply the Reactome pathway eigenvectors from the first step with the most variant ovarian cancer eigenvector in the second step to produce a single value which ranks the Reactome pathways by the amount of information flow in the network and by the variance of the omics data.

The above steps briefly outline the approach. Additionally, the method will be applied to synthetic or well-known data to score its performance compared to PARADIGM [118, 129], probably the most popular network-based pathway approach that can be applied to multiple omics data.

Aim 2

Aim 2 consists of the implementation of the eigenanalysis in a platform independent Bioconductor/R package. R package creation and submission should be straightforward. Dr. Culhane is experienced in creation of Bioconductor packages, having written three Bioconductor packages previously [33, 52, 117]. DI Tomescu is experienced with the script language R/Bioconductor as well as with Windows and UNIX based operating systems, making the platform independent package creation straightforward.

2. Introduction

The ultimate goal of science is the understanding of the world by discovering the underlying laws that govern it. This is, of course, highly complicated. Human kind has achieved major breakthroughs but the are still a lot of questions to be answered. For example: physicists can explain our every day world with Einstein's theory of relativity but if they have to explain the world of atoms, molecules, laser or superconductors 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 sometimes a particle. Or maybe it is neither a weave nor a particle, it is something that has just not been discovered yet. 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 types of data 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 of specific levels of abstraction.

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 kinases; different experimental platforms, such as Agilent and Affymetrix or multiple time points.

Integrative analysis is a rapid growing research field today. This is due to the unprecedented wealth of available data which is caused by technological improvements and, at the same time, dropping costs for experiments. While at the end of the year 2000, according to [46], there were only 1760 published articles on integrative analysis, in September 2014 the number exploded to 18500 publications.

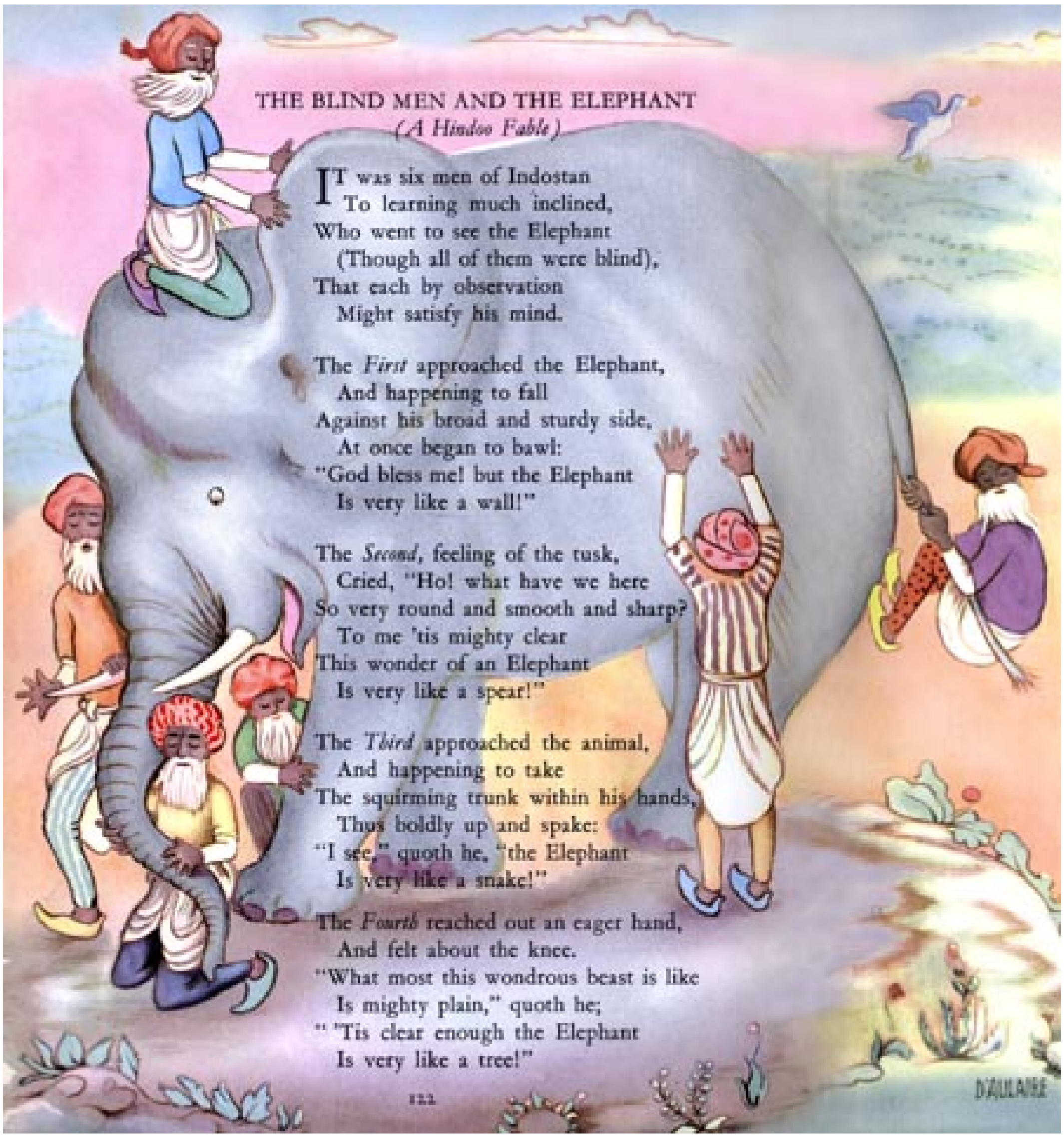
The goal of an integrative analysis is data discovery on one hand and data exploitation on the other hand. Some of the used methods provide also the opportunity of data visualization which promotes the overall understanding of the problem.

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 2.1). 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 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 from 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 may different levels, such as: DNA, RNA, protein and 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 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 same study. As mentioned above, one driving force of the development



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

of integrative data analysis is the increasing amount of available data sets. These sets would not be so abundant if the technology needed to generate them would not have developed so fast and the costs would not have dropped so quickly. Section 2.1 introduces the different types of data that are currently used in integrative analysis while section 2.2 provides an overview of various types of integrative analysis methods.

2.1. 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", while *omics* represents a variant of the suffix *oma* [140]. *Oma* originated from the Greek $\omega\mu\alpha$ which is a sequence composed of ω, a letter belonging to the word stem and $\mu\alpha$, a genuine Greek suffix used for abstract nouns.

The different types of omics data are presented in the following subsections. Genomics is the first data type that is introduced, accompanied by a short summary of the major discoveries of molecular biology that led to the vast research field as we know it today. The introduction is continued with transcriptomics, followed by proteomics.

2.1.1. Genomics

According to the World Helth Organisation (WHO) "Genomics is defined as the study of genes and their functions, and related techniques".

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 [81] that traits are passed down to the children and children's children in a predictable way through, what today are called, genes.

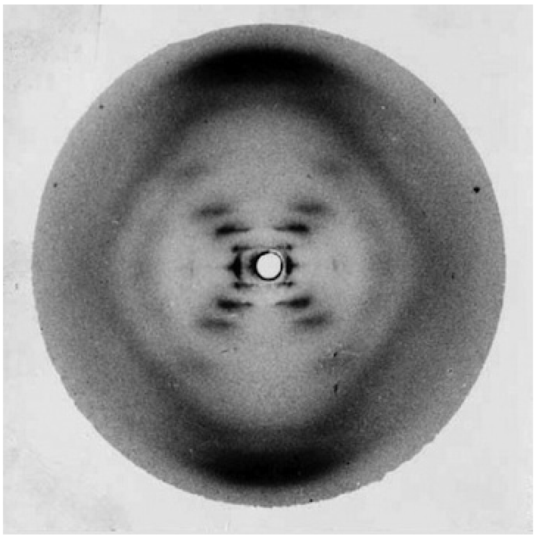


Figure 2.2.: Photograph 51 by Rosalind Franklin showing an X-ray image of the DNA. Figure published in [42]

The next important milestone was the discovery of DNA by Friedrich Miescher in 1869 [84]. 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 [42] together with the famous "photograph 51" (see Figure 2.2) in the same issue of Nature 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 [97], Watson and Crick proposed their 3D model for the DNA [139] 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

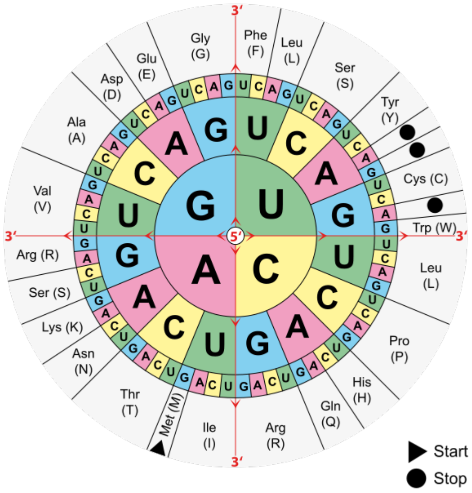


Figure 2.3.: Aminoacids table. The direction of reading for the genetic code (inner side) of the proteins (outer side) starts at 5′ and goes to 3′. Public domain figure.

that "the specific pairing we have postulated immediatly suggests a possible copying mechanism for the genetic material."

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 [92] that the genetic code (see Figure 2.3) for phenylalanine was UUU (three consecutive uracil bases). This was the stating point in 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. 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

Figure 2.4.: 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. Nature cover author: ; Science cover author: Ann Elliott Cutting

steps in which polymerase and dideoxynucleotides triphosphates (ddNTP) are added to the mixture. In each of the four parallel processes another 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 process yields sequences ending in the same base. In order to read the sequenced DNA piece one has to use electrophoresis. The method was published [114] in same year as Sanger et al. sequenced the bacteriophage Φ X174 [112] followed by the bacteriophage λ [113] in 1982.

Another key tool for molecular biology is the polymerase chain reaction (PCR). Developed in 1983 by Mullis Kary, the PCR [109] is used to *in vitro* amplify the DNA. 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 *Haemophilis influenza Rd.* [41] 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. 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 [72] (see Figure 2.4) as well as Venter et al. [132] 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 [1] 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 [139] 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 [139] to the human genome.

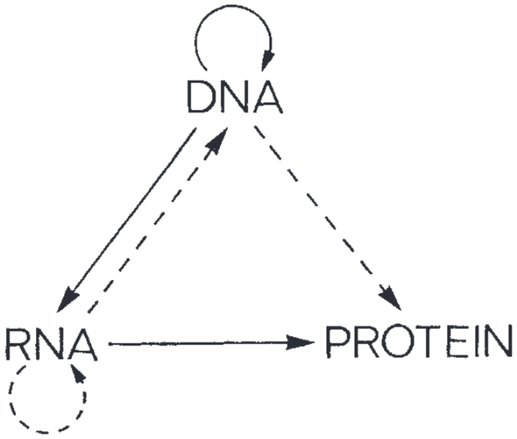


Figure 2.5.: Central dogma of molecular biology. Digram by Francis Crick as it was published in [30].

The Human Genome Project was announced to be finished in 2003. This was two and a half years before the planed end with a large part 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 [30].

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 to protein or back to DNA. The article also included a diagram (see Figure 2.5) 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 is known

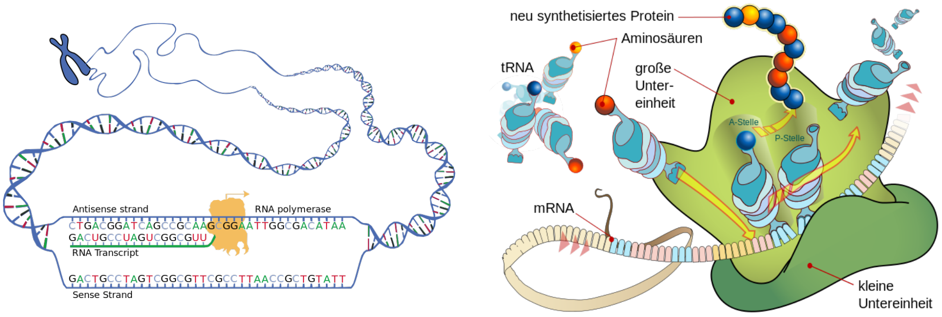


Figure 2.6.: The two main processes involved in gene regulation: transcription (left) and translation (right). Public domain graphics.

as: "DNA makes RNA, RNA makes protein" which emphasizes the two processes which govern the protein production in a biological system: transcription and translation.

The synthesis of RNA by using DNA as a template is called transcription [68]. Through this process, in which the DNA bases A,T,C,G are translated to A,U,C and G, three types of RNA are created: messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). In case of mRNA, the process of transcription is divided in two subprocesses: synthesis and processing. A graphical representation is shown in Figure 2.6.

The synthesis of proteins based on an mRNA template is called translation [68]. A protein is 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 2.6.

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 are not completely understood. The mechanism of gene regulation are so complex that they have 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 mile stones in the history of molecular biology led the research filed as we know it today and make it possible for us to access large amounts

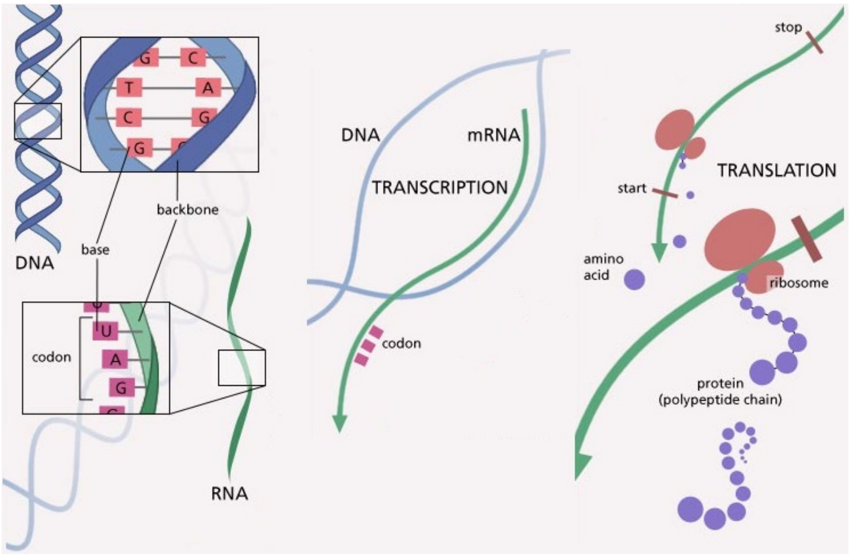


Figure 2.7.: Illustration of the Central dogma of molecular biology (adapted from [59]).

of information that we need to answer

2.1.2. Transcriptomics

Transcriptomics is the technology used to study the transcriptome which is defined by Velculescu et al. in [131] 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 [8], differential display [77] or dotblot analysis [75]. One drawback shared by all of the 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 [2] by using serial analysis of gene expression (SAGE). It represented one of the earliest application of the Sanger sequencing method [114] and was composed of two steps as described in Figure 2.8: "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 [130] 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 [89] 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. These steps are summarized in Figure 2.9.

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

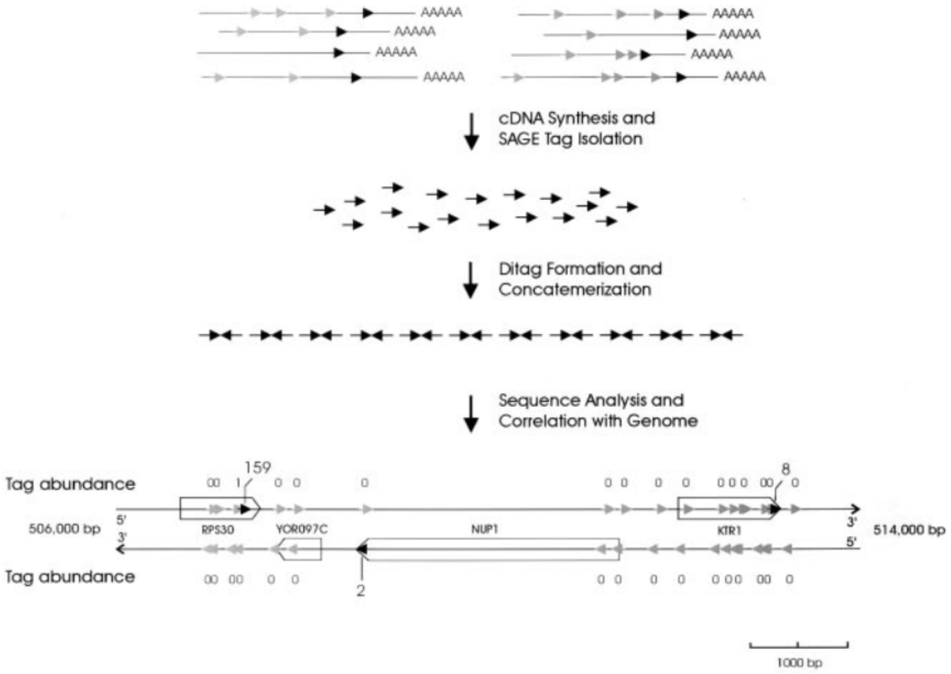


Figure 2.8.: Serial analysis of gene expression: method used for the caracterization of the first mammalian transcriptome.

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 also: 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 [18, 37], mutational analysis [53], drug discovery and development [34], cancer research [15, 25, 50, 126], microbial applications [35, 122].

Mcroarrays dominated the research community because they stand for high throughput technology at a very reasonable price [136]. However, there are limitations that have to be taken into account: cross-hybridization can lead to high background levels which will cause erroneous data [94]; 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 use of an already existing genome sequence [136].

The end of the microarray era started with the second-generation or next-generation sequencing (NGS) technology [119]. As emphasized by Wang and colleagues in [136], NGS based approaches directly determine the cDNA sequence in contrast to microarray based methods that use already existing genome information.

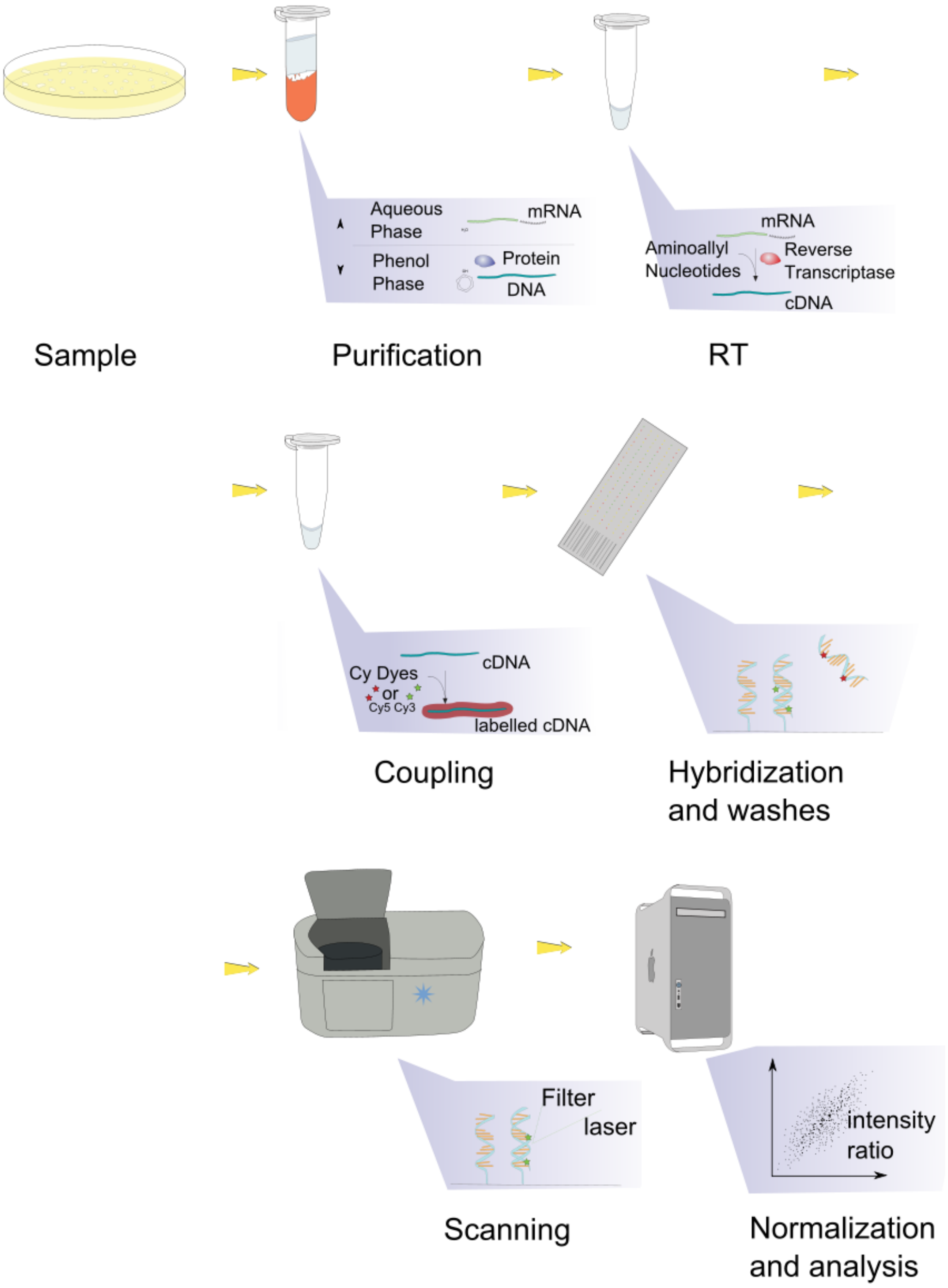


Figure 2.9.: Microarray work flow.Public domain figure.

Even if Sanger sequencing was used for cDNA [44] and expressed sequence tag (EST) sequencing [16] the low throughput, high costs and being non-quantitative led to the development of tag based methods such as SAGE [130], cap analysis of gene expression (CAGE) [70, 121] and massively parallel signature sequencing (MPSS) [17] which are high throughput and provide precise gene expression levels. Nevertheless, these technologies also suffer from limitations such us high costs, use of short read tags that can not be uniquely mapped to the genome as well as non-isoform specificity[136]. 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 in [47] describe these steps in detail and provide additional information on this technology.

An outstanding review [78] 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 [14, 23, 79] 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 [56, 78, 82, 100, 104, 120].

Based on NGS, a new method was developed for the identification and quantification of transcriptomes: RNA-seq (RNA-sequencing) [80, 88]. Generally speaking, the work flow of RNA-seq is composed of the following steps [136]: 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 Ilumina IG (formally known as Solexa) [14], Applied Biosystems SOLiD [23] and Roche 454 Life Sciences [79] are used, although Illumina IG seems to be the most used [128] platform.

Some of the most noteworthy advantages [23, 40, 136] 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 [40]: transcriptome profiling of non-model organisms [29, 133], model transcripts identification [106], study of RNA modification [13, 98] and quantification of allele-specific gene expression [108].

2.1.3. 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 mechanisms of translational and transcriptional regulation.

The term *proteomics* was defined [48] 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 [10, 138, 142]. Nevertheless, studies that deserved the name proteomics have been conducted since 1975, when the two dimensional gel, developed by O'Farrell [93], was used in studies in which mouse [67] and guinea pig [116] protein mappings were conducted. An example of a two dimensional gel of proteins from *Bacillus subtilis* can be seen in Figure 2.10.

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 [38] used for

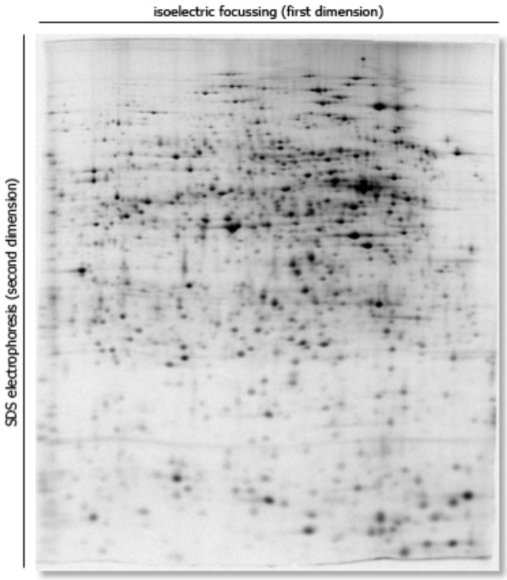

isoelectric focussing (first dimension)
SDS electrophoresis (second dimension)

Figure 2.10.: Example of a two dimensional gel. Figure released under GNU Free Documentation License.

the sequencing of proteins. Later, the group around Stephen Kent developed microsequencing techniques [4, 5, 6] 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 [9]. This breakthrough was achieved by the ability to quantify and identify proteins which was used in the study of protein interaction networks [24] and by revealing the protein composition of cellular organelles [11, 143]. Figure 2.11 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 [20]: identification of the coding gene, 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 [58] which will generate statistically significant peptides matches but also peptide quantification through isotope-labeling or label-free comparisons. As examples we mention here the chemically labeling [107] (iTRAQ) and stable isotope labeling with amino acids in culture (SILAC) [95].

Like 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 [27]. 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.

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 [96] and the biological/disease driven HPP [3].

Applications of proteomics include drug discovery such as crizotinib [127] which is successfully used in the treatment of lung cancer, biomarkers discovery for various diseases such as schizophrenia [71] or breast cancer [76] and comparative proteogenomics [51] with focus on improving gene prediction and identification of rare post-translational modifications.

Recently, two major studies of the human proteome were published [65, 141]. 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.

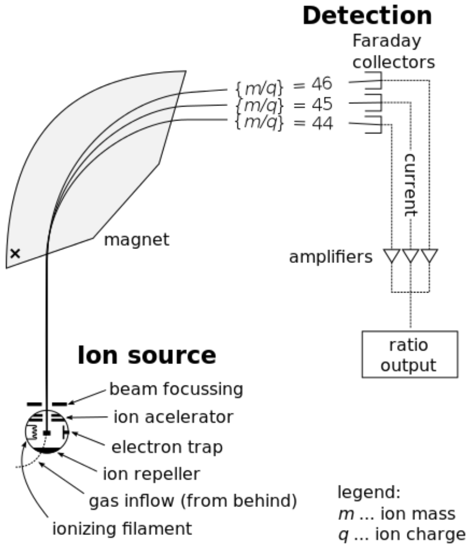


Figure 2.11.: Principals of a simple mass-spectrometer. Public domain figure.

2.1.4. Other Omics Data Sets

In the previous sections the most well studied omics data types were presented. Other omics data types include, but are not limited to metabolomics, glycomics, lipidomics and localizomics. An overview of the various omics data types is shown in Figure 2.12.

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 [63].

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.

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

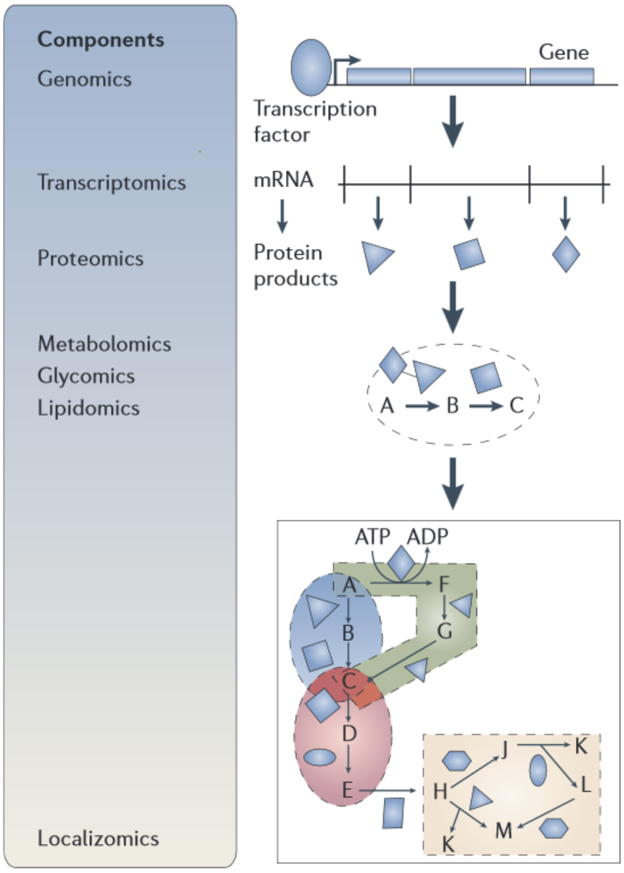


Figure 2.12.: Overview of omics data. Figure adapted from [63].

2.2. Integrative Data Analysis

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 [39], or they can be measured on the same omics level but on different organisms/platforms [7, 73].

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 [135], somatic mutations, copy number and gene expression data[85], chromatin maps and gene expression profiles [69], genotypic variation at DNA level and gene expression data [22], CHIP-seq and RNA-seq data [45], transcriptomics and proteomics data[28, 39, 74].

With transcriptomic and proteomic data, most analysis techniques are based on the direct correlation between transcripts and proteins. Cox and colleagues [28] present different approaches based on correlation and clustering. Other correlation-based studies have also been performed in [19, 21, 26, 49, 74, 87, 137]. Statistical methods based on correlations are presented in [66, 91]. The premise of a direct correlation between transcripts and proteins is not valid in eucaryotic organisms, due to post-transcriptional and post-translational regulation [39, 55]. Other approaches are based on network analysis [61, 62] and statistical methods such as analysis of variation, clustering and gene set enrichment [54, 124, 134].

Piruzian *et al.* [101] 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.* [99] 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 [54] 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 [134].

In [124], 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 [57, 63, 144] for omics data in general and in [55] for transcriptome and proteome data in particular.

Omics data alone are not enough when it comes to the fundamental questions of molecular biology. In order to exploit the available data one has to combine them in a meaningful way. Integrative data analysis addresses this need.

3. Methods and Results

This chapter summarizes the used methods and presents the obtained results. In section 3.1 Reactome, the data base which provides the network based pathway information and the R package *graphite* which was designed to facilitate the interaction with Reactome through the use of R are introduced.

3.1. The pathway data base Reactome

The focus of this section is the network analysis of Reactome. By employing the R package graphite, inspection of the Reactome pathways was performed. Additionally, three repositories that provide data on human protein/genetic/molecular interactions were investigated.

Within Reactome, we examined each pathway as well as the union of all pathways in one single network. STRING, a protein-protein interaction data base, proved to also (in addition to Reactome) possess the mathematical properties needed for our further analysis. Taking into account two different repositories would both enlarge the assessment power of the method and provide a comparison framework.

In order to integrate the information from the two repositories, the nodes of the networks have to be ranked. Two different measurement techniques (dynamical importance and betweenness) are investigated and applied to the largest strongly connected component of Reactome and STRING as well as to the 166 strongly connected Reactome pathways.

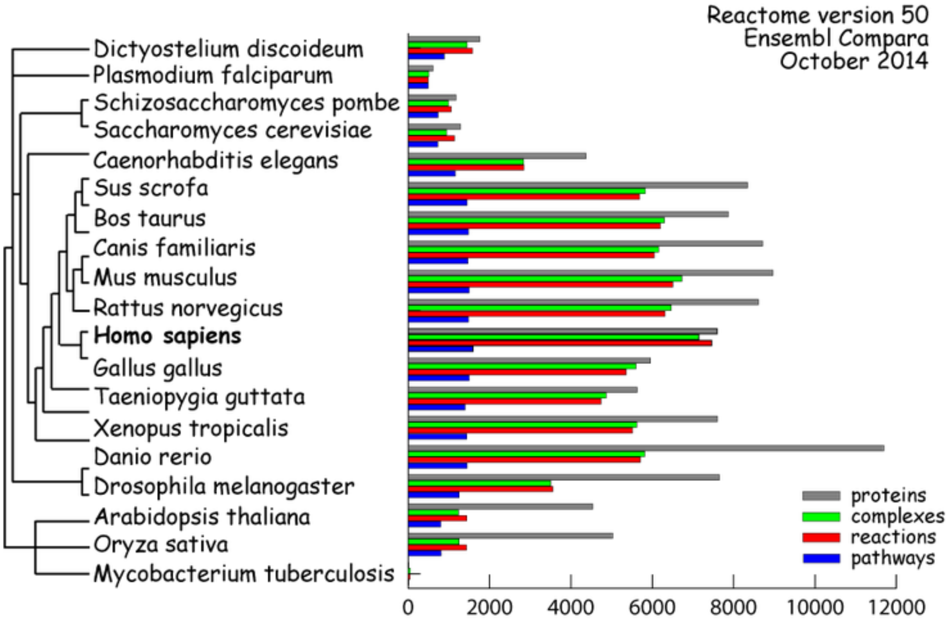


Figure 3.1.: Overview of the Reactome content: species, proteins, complexes, reactions and pathways. Figure adapted from [103].

3.1.1. Using the R package graphite to access pathways from Reactome

According to the Reactome homepage [103], Reactome [31] is a "free, open-source, curated and peer reviewed pathway database". It includes pathways for a number of organisms of interest such as *Homo sapiens, P. falciparum, Sacharomyces cerevisiae, Mus musculus*. Figure 3.1 shows the number of proteins, complexes, reactions and pathways for each species. A detailed overview of the species, proteins, complexes, reactions and pathways is included in the Appendix A.1. At the time point of the analysis, Reactome contained 1240 pathways for *Homo sapiens*.

The R package *graphite* [110] can be used to access Reactome pathways. The package includes pathways from different databases including Reactome (BioPax format), KEGG (KGML format), BioCarta (BioPax format), NCI (BioPax format) and SPIKE. For the BioPax format a pathway is identified through the xml tag `pathway`. Additionally, a number of rules are applied [111] to translate a Reactome pathway into a *graphite* pathway. The basic conversion rules are shown in Figure 3.2. Additional simplification rules are shown in A.2 and A.3. These rules are very important because the structure of the converted pathway is the basis of our further analysis.

An example of a Reactome pathway from *graphite* can be seen bellow:

```
1 > reactome[[3]]
2 "Abacavir metabolism" pathway from reactome
3 Number of nodes       = 4
4 Number of edges       = 3
5 Type of identifiers = native
6 Retrieved on          = 2014-04-02
7
8 >plot(graph.edgelist(as.matrix(reactome[[3]]@edges[,c
     (1,2)]), directed = T))
```

The graphical display of the pathway can be achieved by using the plotting function shown on line 8. The simplistic result of this call can be seen on Figure 3.3

At the time point of the analysis, *igraph* included 1240 pathways for *H. sapiens* while the most recent Reactome version (47) provides 1491 pathways.

```
> library("graphite")
> packageVersion("graphite")
[1] 1.8.1
> length(reactome)
[1] 1240
```

Email contact with Gabriele Sales (package maintainer) revealed that there are no scripts used for the translation of Reactome pathways into graphite. They used a combination of department intern tools that can not be shared. However, we were offered the possibility of individual translation of potentially not included (due to the new release of Reactome) pathways. It

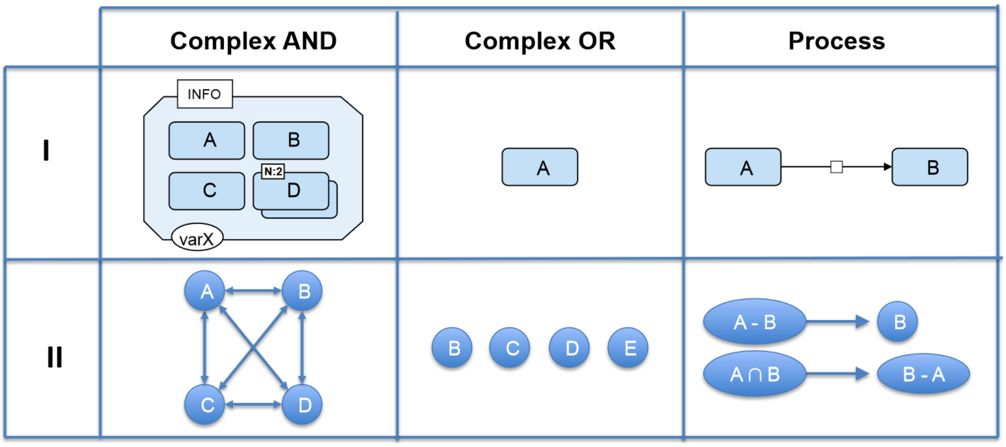


Figure 3.2.: Basic conversion rules from a Reactome pathway to a graphite network. Figure addapted from [111].

was also mentioned that an update will be available together with the next Bioconductor release.

3.1.2. Pathway Connectivity

In order to assess the dynamical importance of the pathways, the largest eigenvalue of the adjacency matrix has to be a real valued number. According to the Perron-Frobenius Theory, this can be ensured if the graph representing the pathway is strongly connected. A directed graph is called strongly connected if there is a path in each direction between each pair of vertices in the graph. A directed graph is weakly connected if replacing all of its directed edges with undirected edges produces a connected (undirected) graph. In other words, a directed weakly connected graph becomes strongly connected when directed edges are replaced by undirected edges.

Employing the R package *igraph* [32] we assessed that from the 1280 pathways 166 are strongly connected and 902 are weekly connected using the following script:

```
rm(list = ls())
```

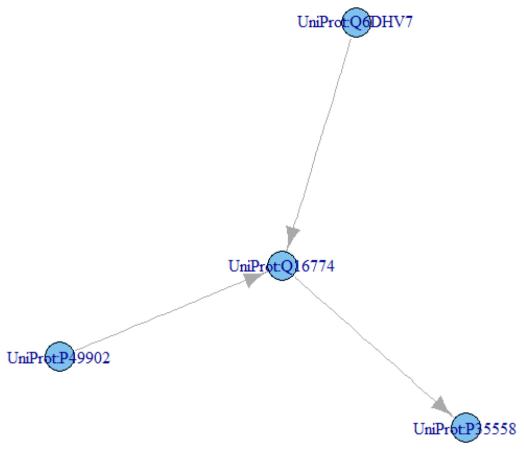


Figure 3.3.: Abacavir metabolism. Graphic output generated with the R package *igraph*.

```
library("graphite")
library("igraph")

isStronlyConnected = function(reactome.pathway){
  edges = as.matrix(reactome.pathway@edges[,c(1,2)])
  reactome.graph = graph.edgelist(edges, directed = TRUE)
  strongly.connected = is.connected(graph = reactome.graph,
  mode = "strong")
  return(strongly.connected)
}

pathways.sc = lapply(X = reactome, FUN = isStronlyConnected)
```

Statistical Summary of the Pathways in graphite.

The distribution of the number of nodes and edges in each pathway can be seen in Figure 3.4. Additionally, the summary statistics for the number of nodes and ages are shown:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.00    9.00   20.50   51.98   52.00 1874.00

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.0    20.0   127.5  1631.0   731.0 89160.0
```

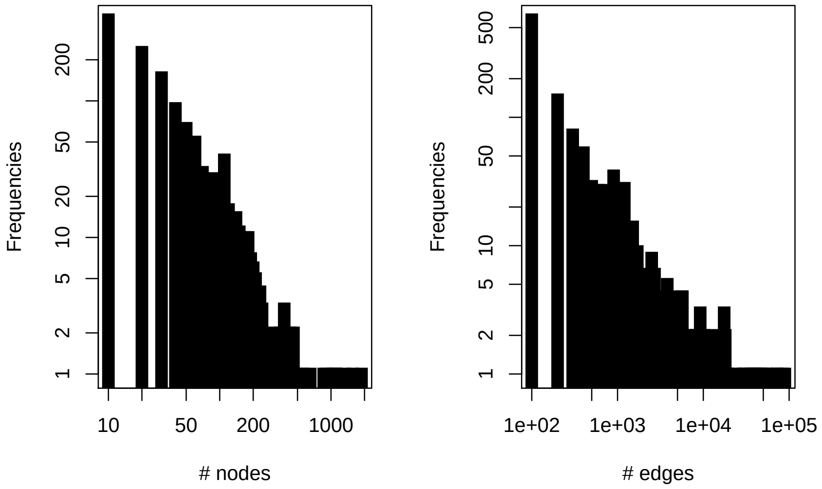


Figure 3.4.: Distribution of nodes and edges per pathway as provided by the *graphite* package.

Largest Strongly Connected Component

Since only 166 pathways were identified as strongly connected, we assess the size of the largest strongly connected component if all pathways are combined into one large pathway. According to our analysis, this component includes 4675 of the total 7166 vertices.

3.1.3. Inspection of Other Data Repositories for Interaction Networks Using ARepA.

ARepA is an acronym for Automated Repository Acquisition, and is designed as a command-line tool to easily fetch 'omics data from multiple heterogeneous repositories and process them in a standardized way. It's main features include gene ID standardization and file standardization (R readable, tab-delimited format). Currently, ARepA fetches human data from the following repositories: STRING, BioGRID, GEO, and IntAct.

BioGrid is an online interaction repository with data compiled through comprehensive curation efforts. It includes protein and genetic interactions.

IntAct provides a freely available, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available.

STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources: genomic context, high-throughput experiments, conserved coexpression, previous knowledge. STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable.

The downloaded data files was imported into R. In order to compare the networks from the different repositories, Table 3.1 was created:

	Reactome	STRING	IntAct	BioGrid
graph	directed	undirected	undirected	directed
vertices	7166	16893	4975	6116
edges	280086	324075	9689	18768
weak	6954	16867	4518	5769
strong	4675	16867	4518	1

Table 3.1.: Comparison of different repositories.

The largest strongly connected components can be found in STRING and Reactome. The interactions in STRING include known and predicted interactions. This probably explains the huge difference in the number of nodes of the largest strongly connected component.

3.1.4. Ranking methods

In order to be able to rank the nodes in the pathways we have to apply a certain measurement technique. In this scenario we chose two methods: the dynamical importance and the betweenness of nodes.

While the dynamical importance of a node focuses on the change in the largest eigenvalue of the corresponding adjacency matrix upon it's removal, the betweenness is based on the number of shortest paths passing through that node.

These two distinct techniques allow us to take into account different aspects of the pathways under study and compare the obtained results.

Dynamical Importance

The dynamical importance as it was used here was defined and described in [105]: For a network, I_k, the dynamical importance of node k is defined as the change in the largest eigenvalue λ of the corresponding adjacency matrix upon it's removal (3.1).

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

Additionally, Restrepo and colleagues ([105]) provide $\hat{I}_k$, an approximation for I_k, that decreases the computation time for large networks (3.2).

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

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.

According to the Perron-Frobenius theory [83], the dynamical importance of the nodes under study is a real value ($I_k \in \mathbf{R}$) iff the corresponding network is strongly connected.

The implementation of the dynamical importance in R uses the function *eigen* which computes the left and right eigenvectors of the adjacency matrix of the network and it is shown bellow:

```
dynamicalImportance = function(g){
if(!is.igraph(g))
stop("Not an igraph object")

ad.mat = as.matrix(get.adjacency(g))
ad.mat = ad.mat + matrix(data = 1/(dim(ad.mat)[1]),
   ncol = dim(ad.mat)[1],nrow = dim(ad.mat)[2])
r.ev = eigen(ad.mat)$vectors[,1]
l.ev = eigen(t(ad.mat))$vectors[,1]

tmp1 = abs(r.ev)*abs(l.ev)
tmp2 = t(abs(r.ev))%*%abs(l.ev)
dyn.imp = tmp1/as.numeric(tmp2)
names(dyn.imp) = V(g)$name
return(dyn.imp)
}
```

For very big graphs, like the merged Reactome graph, the computation time is very high. One possible way of decreasing it would be to use the power method (lines 27 and 28) to compute the left and right eigenvectors:

```
myPowerCalcEV = function(mat){
accuracy = 10
b = as.matrix(runif(dim(mat)[2]),nrow = dim(mat)[2], ncol = 1)
diff = rep(100,dim(mat)[2])
while( length(which(diff > 10^(-1*accuracy)))>1){
cat(".")
tmp = mat%*%b
b.new = (tmp)/(norm(tmp,"f"))
diff = abs(b-b.new)
b = b.new
}
return(b)
}

dynamicalImportancePM = function(g){
if(!is.igraph(g))
stop("Not an igraph object")

ad.mat = as.matrix(get.adjacency(g))
ad.mat = ad.mat + matrix(data = 1/(dim(ad.mat)[1]), ncol = dim(ad.mat)[1],nrow = dim(ad.mat)[2])
r.ev = myPowerCalcEV(ad.mat)
l.ev = myPowerCalcEV(t(ad.mat))

tmp1 = abs(r.ev)*abs(l.ev)
tmp2 = t(abs(r.ev))%*%abs(l.ev)
dyn.imp = tmp1/as.numeric(tmp2)
names(dyn.imp) = V(g)$name
return(dyn.imp)
}
```

Betweenness

The betweenness B_k of node k was defined by L. Freeman [43] as the number of shortest paths between any two nodes i and j going through node k divided by the total number of shortest paths between i and j.

$$B_k = \sum_{i,j \neq k} \frac{\#sp_{ikj}}{\#sp_{ij}} \tag{3.3}$$

The R package *igraph* [32] provides the functionality needed to compute the betweenness.

In summary, using the P package *igraph* and the self implemented function to compute the dynamical importance, the two network centrality measures can be computed as:

```
dyn.imp.reactome = dynamicalImportancePM(reactome.as.igraph)
bet.ness.reactome = betweenness(reactome.as.igraph)
```

3.1.5. Preliminary Results

The dynamical importance and betweenness of the largest strongly connected components from Reactome and STRING as well as the strongly connected pathways from Reactome were computed.

For the largest strongly connected component in Reactome and STRING, the dynamical importance of each node is plotted against the square root of the product between the in and out degree of that node. Figure 3.5 shows the results for Reactome while Figure 3.6 presents the results for STRING.

We notice that the dynamical importance displays a more linear dependence on the geometric mean of the in and out degrees of the nodes. In contrast, the betweenness shows a less linear dependence. It seems that the betweenness does not directly depend on the degrees of the nodes.

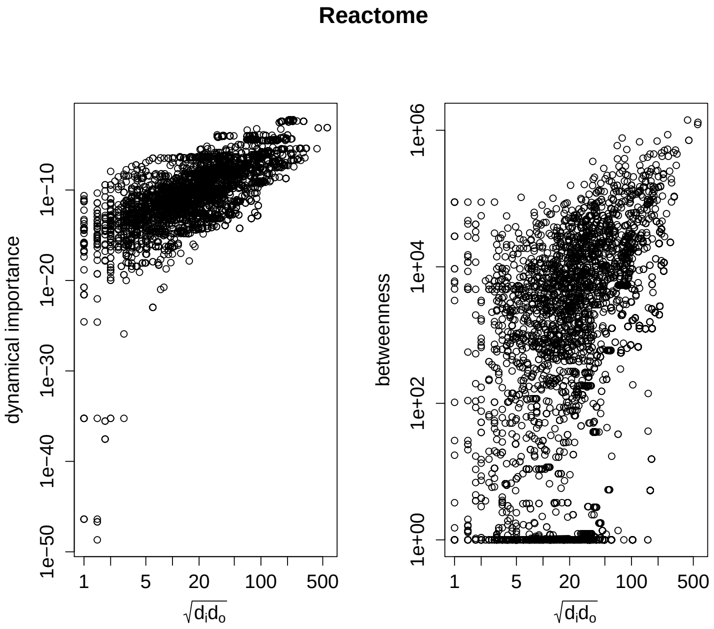


Figure 3.5.: Reactome: dynamical importance and betweenness.

Figure 3.7 shows the largest dynamical importance and the largest betweenness of each of the 166 strongly connected pathways in Reactome plotted against the number of nodes and edges in that pathway.

In Figures 3.5,3.7 and 3.6 the plotted betweenness is one higher than the computed one. The betweenness has high values and by adding one we do not change the conclusions drown from the figures. But adding one helps to visualize a betweenness equal to zero on a logarithmic axis.

In conclusion, there are nodes with betweenness equal to zero and there are pathways with the largest betweenness equal to zero. Such pathways are shown in Figure 3.8.

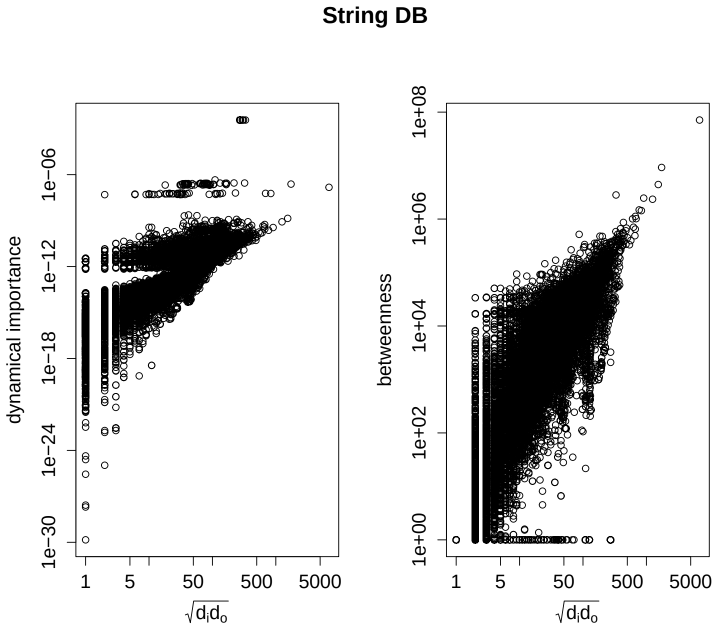


Figure 3.6.: STRING DB: dynamical importance and betweenness.

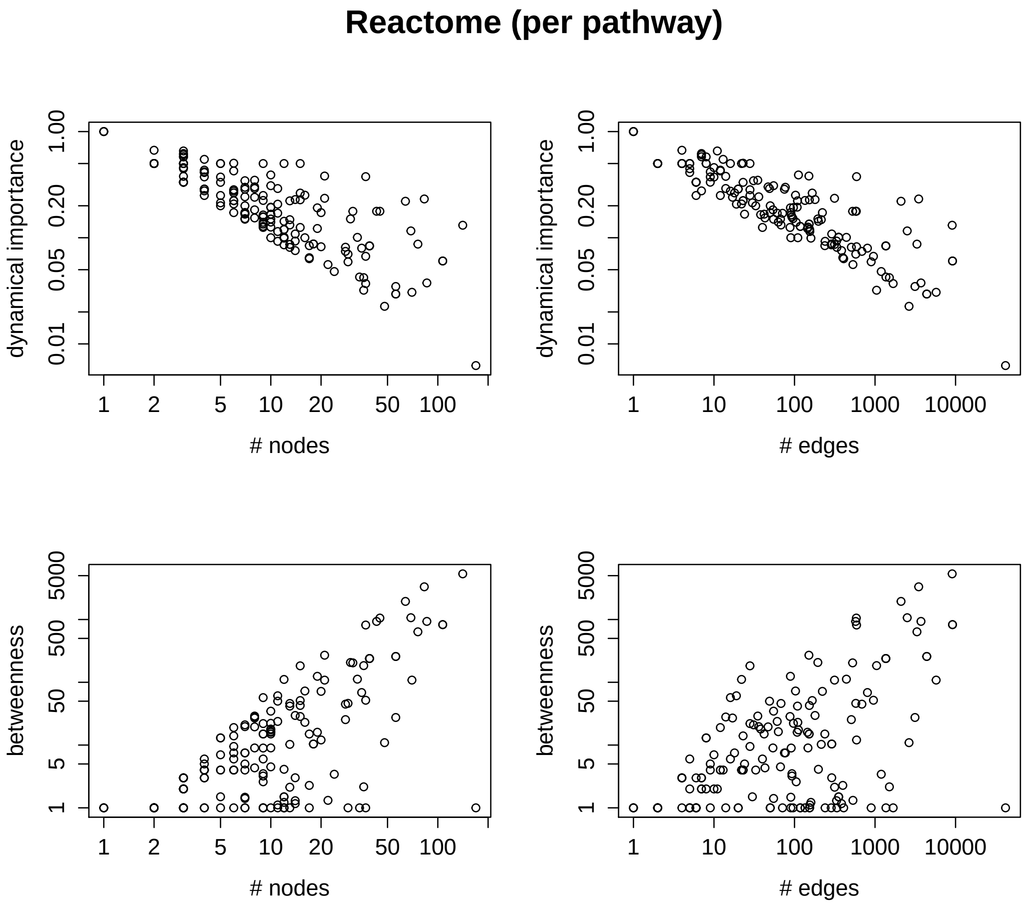


Figure 3.7.: The largest dynamical importance and betweenness of the strongly connected pathways in Reactome.

By analyzing the shown networks, the reason for the zero betweenness is the high connectivity of the networks. The shortest path never has to go through a node: Due to the high connectivity there are always alternatives that are shorter.

In such cases we will use the in and/or out degree of the nodes as a measure to rank them. This will ensure that the nodes are equally important in their pathway, but the larger a pathway of this kind is, the more important a node will be.

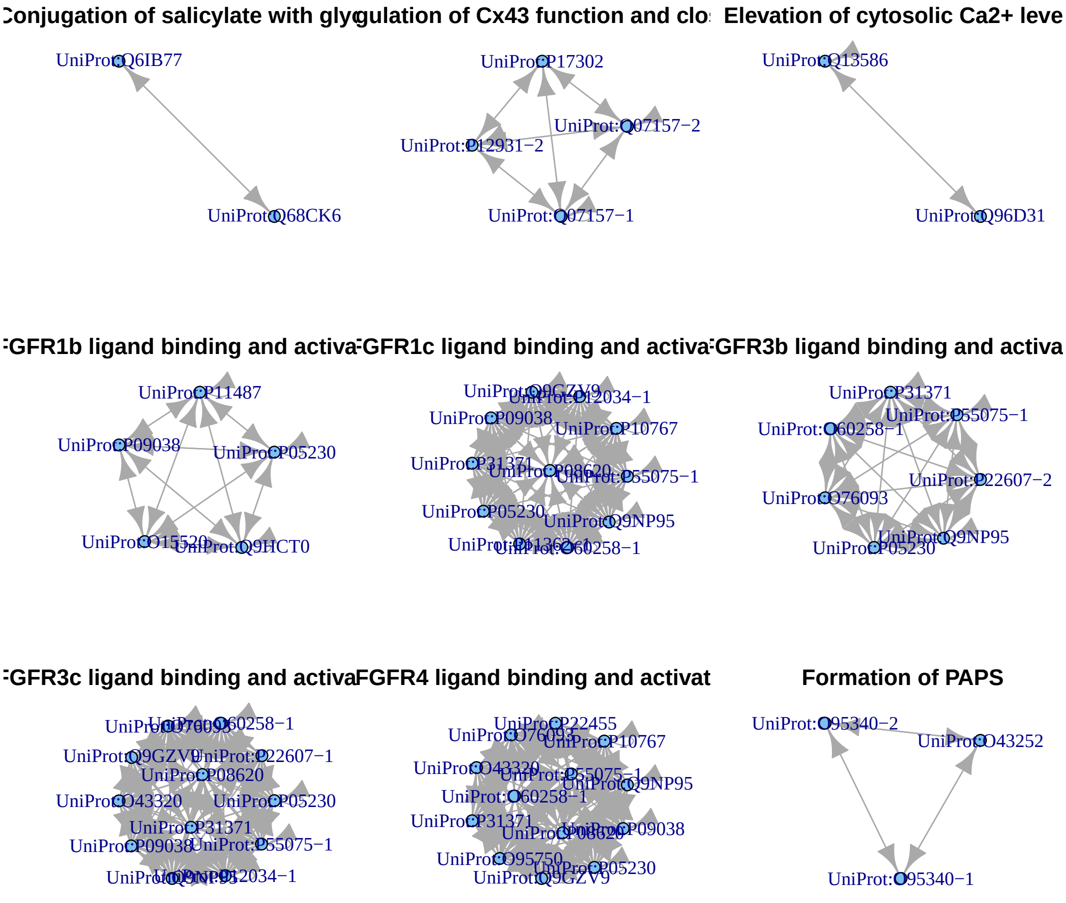


Figure 3.8.: The largest dynamical importance and betweenness of the strongly connected pathways in Reactome.

3.2. Multiple Co-Inertia Analysis and the TGF-beta receptor signaling in EMT Pathway

This section is mostly dedicated to the deeper understanding of the multiple co-inertia analysis (MCIA). Additionally, the scoring resulting from the betweenness and dynamical importance were compared on the TGF-beta receptor signaling in EMT (epithelial to mesenchymal transition) pathway.

3.2.1. Multiple Co-Inertia Analysis

MCIA is an integrative analysis method based on an ordination method like principal component analysis (PCA), correspondence analysis (CA) or non-symmetric correspondence analysis (NSCA) that transform the data into a comparable space.

Ordination methods

The mathematical analysis starts with the investigation of the statistical triplet (X, Q, D). X is the matrix containing the measurements of genes (rows) in different conditions (columns). We assume X to have n rows and q columns. Q $(p \times p)$ is an inner product used to measure distance between n points in $\mathbb{R}^p$. D $(n \times n)$ is a inner product used to measure relationships between p points in $\mathbf{R}^n$.

The purpose of the analysis determines Q and D. In case of the centered PCA applied on the covariance matrix of X, $Q = I_p$ the $p \times p$ identity matrix and $D = \frac{1}{n} I_n$ the $n \times n$ identity matrix. Additionally, X has to be normed by the columns mean: $X = [x_{ij} - m(x_j)]$, where $m(x_j)$ is the mean of the j^{th} column. If PCA is carried out on the correlation matrix of X than Q and D remain unchanged and $X = \left[\frac{x_{ij} - m(x_j)}{sd(x_j)}\right]$, where $sd(x_j)$ is the standard deviation of the j^{th} column of X.

In case of the NSCA, Q is a diagonal $p \times p$ matrix containing the row weights r_i. D is the $n \times n$ identity matrix showing equal weights for all columns. $X = \left[\frac{p_ij}{r_j} - c_j\right]$, where $p_{ij} = \frac{x_{ij}}{N}$, N is the sum over all entries in X. The row weights $r_i = \frac{x_{i+}}{N}$, where x_{i+} the row sum. The column weights $c_j = \frac{x_{+j}}{N}$, where x_{+j} the column sum.

In order to visualize the rows and columns of X in the same space, a duality plot is computed. For this, both rows and columns are projected into a smaller hyperspace following the steps:

- Diagonalization of the inertia operators $W_K Q = X^T DXQ$ and $W_K D = XQX^T D$

  $Q = E^T E \quad D = B^T B$ Cholesky decomposition

  $\Omega = BXT \Rightarrow \Omega^T \Omega = EX^T B^T BXE^T \Rightarrow \Omega^T \Omega = V \Lambda V^T$

  $\Omega = BXT \Rightarrow \Omega \Omega^T = BXE^T EX^T B^T \Rightarrow \Omega \Omega^T = U \Lambda U^T$

- Computation of principal axis and principal components

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

  $G = B^T U \quad K = B^{-1} U$ K are principal components

- Projection into the smaller hyperspace

  $L = XQA$ ... projection of the rows of X onto the principal axis

  $C = X^T DK$ ... projection of the columns of X onto the principal components

3.2.2. Multiple Co-Inertia Analysis - Mathematical Description

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. From all matrices X_k a matrix X is computed: $X = [\omega_1 X_1 | \omega_2 X_2 | ... | \omega_K, X_K]$, where ω_k is the inverse sum of the eigenvalues of X_K.

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

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

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.

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

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

First order solution

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

It can be shown that since v maximizes g it also maximizes:

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

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.

3.2.3. TGF-beta receptor signaling in EMT Pathway

In order to examine in more detail the effects of the ranking imposed by the two selected methods we apply them to the Reactome pathway *TGF-beta receptor signaling in EMT*. The structure of the network can be seen Figure 3.9. We notice that this is a strongly connected network with 15 nodes and 313 edges. We chose this pathway because it is a strongly connected, cancer related pathway with a small enough number of nodes so that it can be displayed 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,

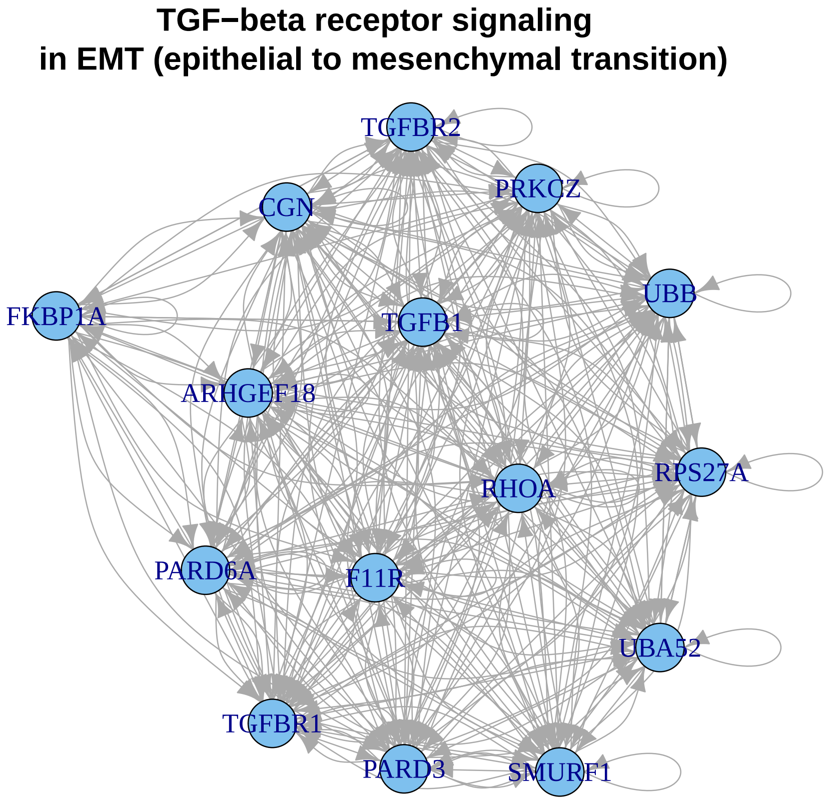


Figure 3.9.: The TGF-beta receptor signaling in EMT pathway

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."

The results of the ranking based on dynamical importance and betweenness can be seen in Tables 3.3 and 3.2.

We notice that the highest ranked gene is in both cases TGFBR1. Additionally, the dynamical importance ranks 7 genes as second: RHOA, PRKCZ, ARHGEF18, PARD3, PARD6A, CGN and F11R. The betweenness ranks only CGH second followed by the other 6 genes ranked second by the dynamical

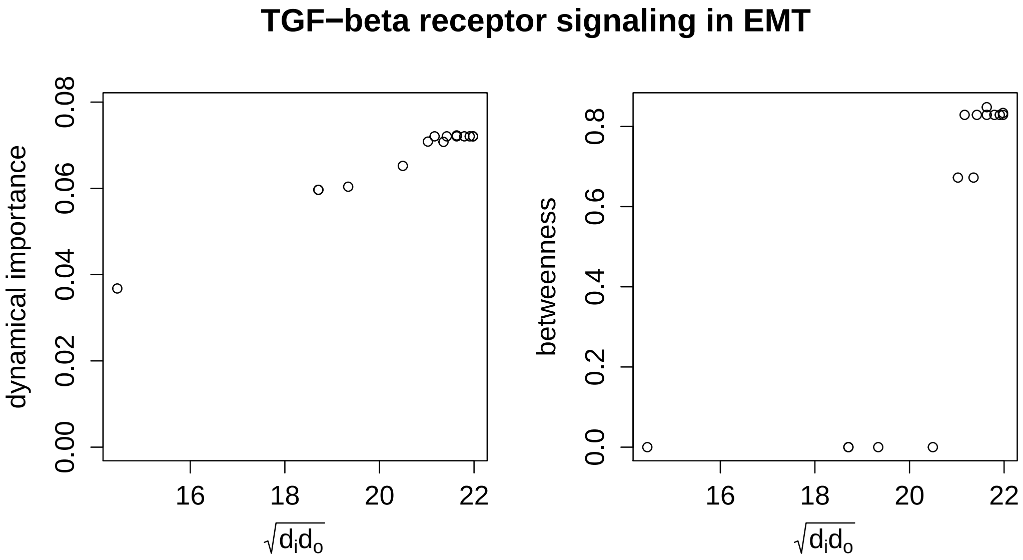


Figure 3.10.: Ranking of the nodes in the TGF-beta receptor signaling in EMT.

importance. Both methods rank TGFBR2 and TGFB1 next. These two genes are equally important based on their betweenness but not to their dynamical importance. The rest of the genes have a betweenness of zero but a slightly lower dynamical importance as the last ranked gene. The betweenness of zero is caused by the smaller number of in/out edges compared to the other genes.

The rankings derived from the betweenness and dynamical importance scores are very similar, especially for the high ranked genes. For the low ranked genes it seems that the betweenness tends to assign scores equal to zero while the dynamical importance still provides a viable ranking. This relationship can also be seen in the dependence of the ranking scores on the geometric mean of the in and out degrees in Figure3.10.

Rank	Gene	dyn. imp.
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 3.2.: Dynamical importance.

Rank	Gene	betweenness
1	TGFBR1	0.84809
2	CGN	0.83367
3	RHOA	0.82891
3	PRKCZ	0.82891
3	ARHGEF18	0.82891
3	PARD3	0.82891
3	PARD6A	0.82891
3	F11R	0.82891
4	TGFB1	0.6724
4	TGFBR2	0.6724
5	UBB	0
5	FKBP1A	0
5	RPS27A	0
5	UBA52	0
5	SMURF1	0

Table 3.3.: Betweenness.

3.3. Eigenanalysis of mRNA and RNAseq Omics Data

This section was dedicated to applying the multiple co-inertia analysis (MCIA) to The Cancer Genome Atlas (TCGA) data and downstream integrative pathway enrichment analysis based on the already derived dynamical importance and betweenness scores.

3.3.1. The Cancer Genome Atlas data

Gene expression of tumors from ovarian cancer patients were profiled using Affymetrix customized platform HG U133 plus 2.0 and RNA-sequencing on Illumina HiSeq platform. Data were downloaded from the NCI-TCGA data portal. The Affymetrix data was normalized and summarized by lowess. A pre-processing pipeline (RSEM) was applied to the Illumina RNA-sequencing data to determine the transcript expression levels. The alignment and gene expression quantification in RNAseq were obtained by MapSplice and RSEM. In our analysis missing values were replaced with a positive value far smaller than the lowest expression value in the dataset (10e-10) and then, the expression values were log10 transformed. 266 out of 489 patient samples were present across both datasets and included in the analysis. Only genes mapping to an official gene symbol were retained and duplicated genes were excluded. In the RNA sequencing dataset 20.135 genes were detected and those with more than 15 missing values were removed, yielding 12.042 and 15,840 gene expressions in Affymetrix and RNASeq respectively.

3.3.2. Multiple Co-inertia Analysis

A typical omics dataset is a matrix where the number of features exceeds the number of measurements (row and columns of the matrix, respectively). Prerequisite for MCIA is a set of tables where either features or measurements are matched and have equal weights. MCIA is performed in two

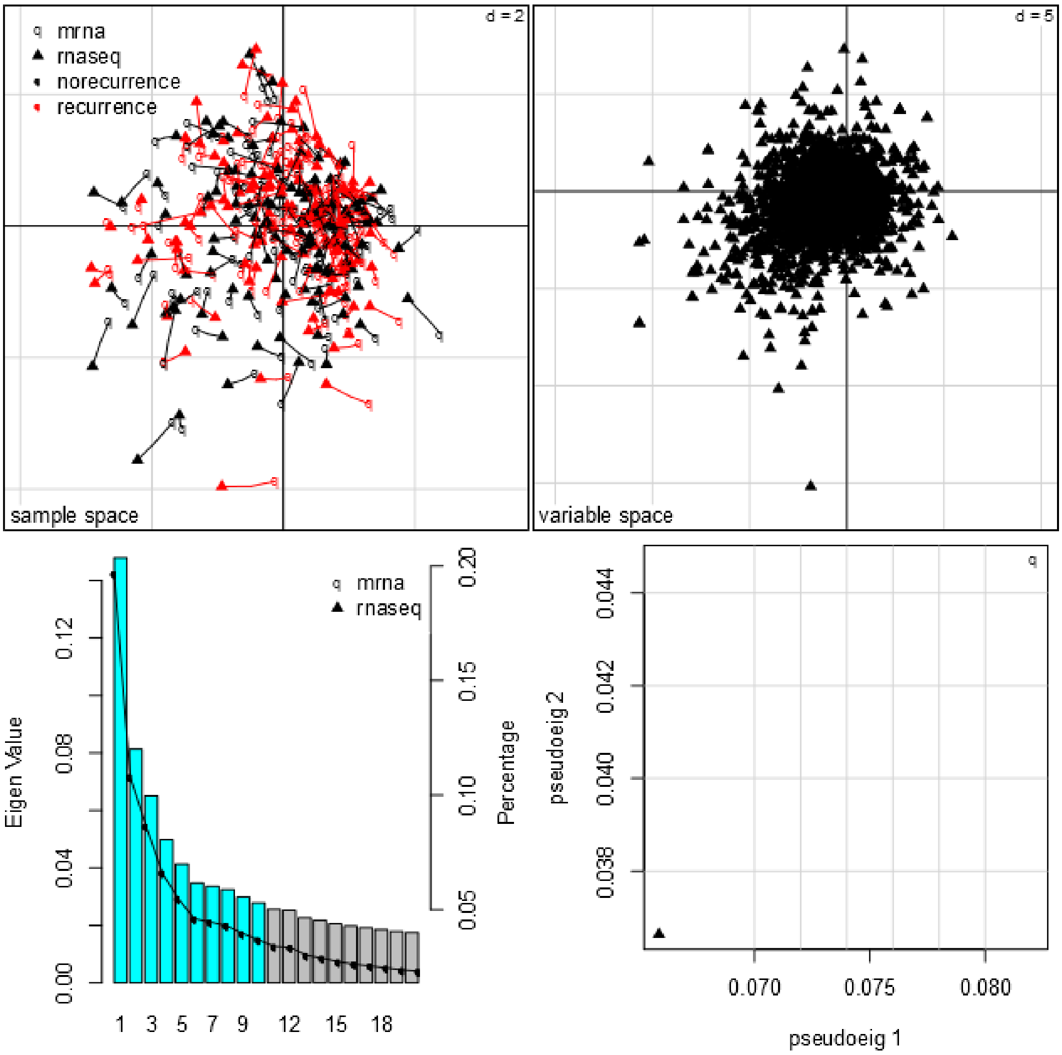


Figure 3.11.: MCIA result

steps to represent features or measurements as points along several axes. In the first step, a one table ordination method, such as PCA, COA or nonsymmetric correspondence analysis is applied on each dataset separately, which transforms data into the same comparable space.

The second step is MCIA. MCIA maximizes the correlation of each individual table with a consensus reference structure through synthetic analysis, which finds a set of reference axes sequentially. In order to find the solution for the first dimension, MCIA determines a single reference axis (referred to as common component v_1) and a set of auxiliary axes for each table ($u_{11} \ldots u_{1K}$) so as to maximize the sum of the co-variances between each of the auxiliary axes u_{1k} and the v_1. The first order solutions of u_{11} to u_{1K} and v_1 are given by the first principal component of the concatenated weighted matrix. The subsequent solutions are found with residual matrices from the calculation of the first order solution with the constraint that the rest order axes are orthogonal with the previous sets. These steps are repeated so that the desired number of axes (principal components, dimensions) are generated.

As a result, MCIA provides a simultaneous ordination of columns (measurements) and rows (features) of multiple tables within the same hyperspace, with features or measurements sharing similar trends will be closely projected.

3.3.3. Analysis Results

In this section, the results of MCIA and of the integrated pathway enrichment analysis will be shown.

MCIA

Figure 3.11 shows the results of MCIA. The upper left plot shows all samples projected onto the common MCIA space. One sample is represented by a circle (coordinates in the microarray space) connected to a triangle (coordinates in the RNAseq space). The color red represents the samples where there was a recurrence after chemotherapy while black samples did not show recurrence.

The upper right plot shows the variable space. Here, each gene in the two data set is shown. The triangles code the genes in the RNAseq data and the circles code genes in the microarray set. It looks like the RNAseq dominates the microarray data.

The lower left plot shows the eigenvalues of the concatenated matrix. We notice that the first and highest eigenvalue only accounts for 12 % of the variance. The lower right plot shows the pseudoeigenvalues of the microarray and RNAseq data.

Integrated pathway enrichment analysis

In order to perform the integrated pathway enrichment analysis the contribution of each gene to the computed MCIA axis has to be extracted. Additionally, the scores computed by the dynamical importance and/or betweenness are needed. These will account for the overall importance of each gene in the already annotated Reactome pathways.

Figure 3.12 shows histograms of the contribution of each microarray gene (upper left plot) and RNAseq gene (upper right plot) as well as the dynamical importance (lower left) and betweenness (lower right) scores.

All scores have been normalized and transformed so that they map to the same range and can be multiplied with each other. The normalization was done by dividing all scores through the maximum and adding one. In this way all scores are in the interval $[1, 2]$.

We notice that there is a difference in the histograms of the MCIA scores. The mRNA MCIA scores seem to be higher since more RNAseq features have low values as compared to the number of mRNA features with values in the same range. It also noteworthy that the betweenness scores are lower than the dynamical importance scores.

After the MCIA scores were combined with the dynamical importance and betweenness scores the top 5% genes were used for pathway enrichment analysis. The same analysis was also performed on only the MCIA top 5%

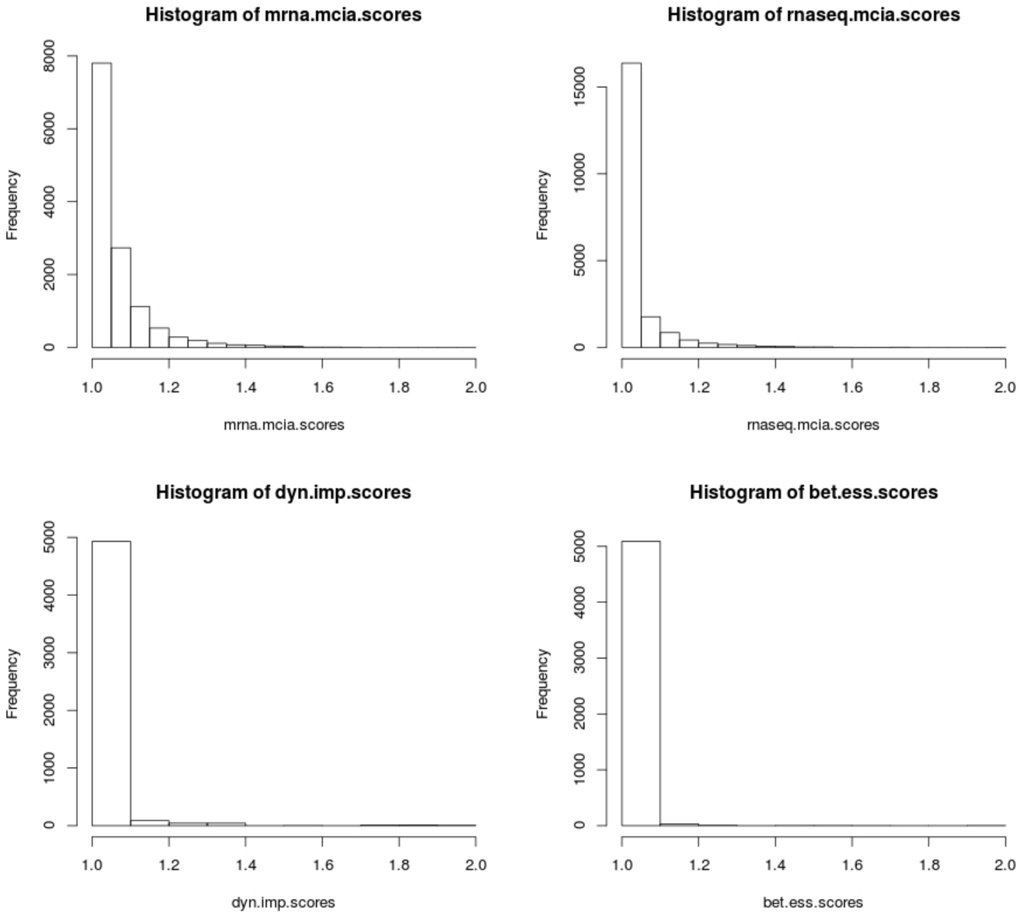


Figure 3.12.: Histograms of MCIA, dynamical importance and betweenness scores.

ranked genes. These results can be seen in Figure 3.13, 3.14 and 3.15. The colors of the bars code for the p value and a map of color-to-p value is shown on the right of each plot. The length of the bars show the number of genes in a pathway.

We notice that through the combination of the MCIA scores with the network scores, the number of enriched pathways increases. This means that through the multiplication with the dynamical importance and the betweenness scores genes that were not so important in a pathway according to MCIA become important and lead to the enrichment of that pathway.

Additionally, it is also noteworthy that the inclusion of the betweenness scores yields results that are closer to the results form the MCIA alone. This is not the case for the inclusion of the dynamical importance scores. This is due to the way the two network scores influence each gene. Form these results it seems that the betweenness scores and the MCIA rank the genes in a similar way. On the other hand, the dynamical importance seems to complement the MCIA results.

A detailed examination of the results of the MCIA scores alone (Figure 3.13) show different enriched pathways such as: signal tranduction, FGFR based pathways, G-alpha signaling pathways, various metabolic pathways and GPRC based pathways. A large amount of the pathways are based on the FGF family.According to Reactome, Signaling through FGFR is summarized as: "The 22 members of the fibroblast growth factor (FGF) family of growth factors mediate their cellular responses by binding to and activating the different isoforms encoded by the four receptor tyrosine kinases (RTKs) designated FGFR1, FGFR2, FGFR3 and FGFR4. These receptors are key regulators of several developmental processes in which cell fate and differentiation to various tissue lineages are determined. Unlike other growth factors, FGFs act in concert with heparin or heparan sulfate proteoglycan (HSPG) to activate FGFRs and to induce the pleiotropic responses that lead to the variety of cellular responses induced by this large family of growth factors. An alternative, FGF-independent, source of FGFR activation originates from the interaction with cell adhesion molecules, typically in the context of interactions on neural cell membranes and is crucial for neuronal survival and development. Upon ligand binding, receptor dimers are formed and their intrinsic tyrosine kinase is activate causing phosphorylation of multiple

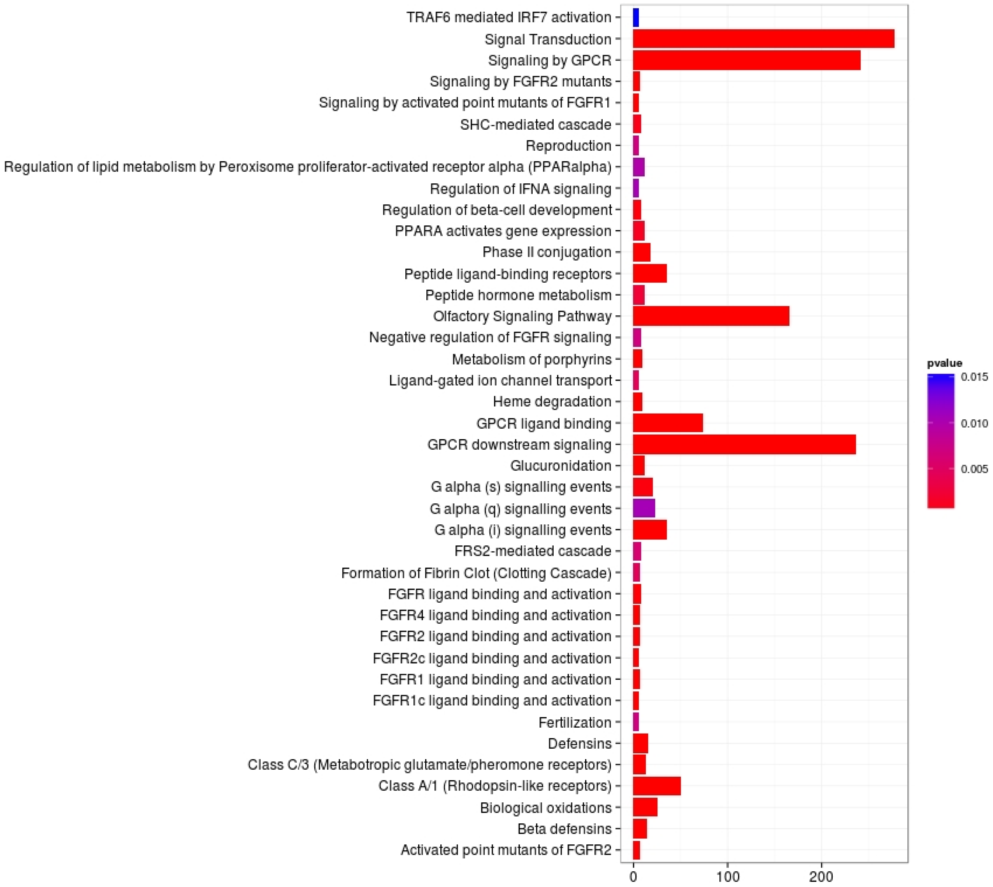


Figure 3.13.: Pathway analysis of the MCIA top 5% ranked genes.

tyrosine residues on the receptors. These then serve as docking sites for the recruitment of SH2 (src homology-2) or PTB (phosphotyrosine binding) domains of adaptors, docking proteins or signaling enzymes. Signaling complexes are assembled and recruited to the active receptors resulting in a cascade of phosphorylation events. This leads to stimulation of intracellular signaling pathways that control cell proliferation, cell differentiation, cell migration, cell survival and cell shape, depending on the cell type or stage of maturation."

Additionally, G alpha signaling events such as: G alpha (i) signaling events ("The classical signalling mechanism for G alpha (i) is inhibition of the cAMP dependent pathway through inhibition of adenylate cyclase. Decreased production of cAMP from ATP results in decreased activity of cAMP-dependent protein kinases."), G alpha (s) signaling events ("The general function of the G alpha (s) subunit (Gs) is to activate adenylate cyclase, which in turn produces cAMP, leading to the activation of cAMP-dependent protein kinases (often referred to collectively as Protein Kinase A). The signal from the ligand-stimulated GPCR is amplified because the receptor can activate several Gs heterotrimers before it is inactivated.") and G alpha (q) signaling events ("The classic signalling route for G alpha (q) is activation of phospholipase C beta thereby triggering phosphoinositide hydrolysis, calcium mobilization and protein kinase C activation. This provides a path to calcium-regulated kinases and phosphatases, GEFs, MAP kinase cassettes and other proteins that mediate cellular responses ranging from granule secretion, integrin activation, and aggregation in platelets. Gq participates in many other signalling events including direct interaction with RhoGEFs that stimulate RhoA activity and inhibition of PI3K. Both in vitro and in vivo, the G-protein Gq seems to be the predominant mediator of the activation of platelets.") were enriched.

The enriched pathways based on the MCIA and the betweenness scores also include signaling through FGFR, but in addition other pathways such as: signaling by EGFR, constitutive PI3K/AKT Signaling in cancer and extracellular matrix organization were enriched. Signaling by EGFR is described as: "The epidermal growth factor receptor (EGFR) is one member of the ERBB family of transmembrane glycoprotein tyrosine receptor kinases (RTK). Binding of EGFR to its ligands induces conformational change that unmasks the dimerization interface in the extracellular domain of EGFR, leading to

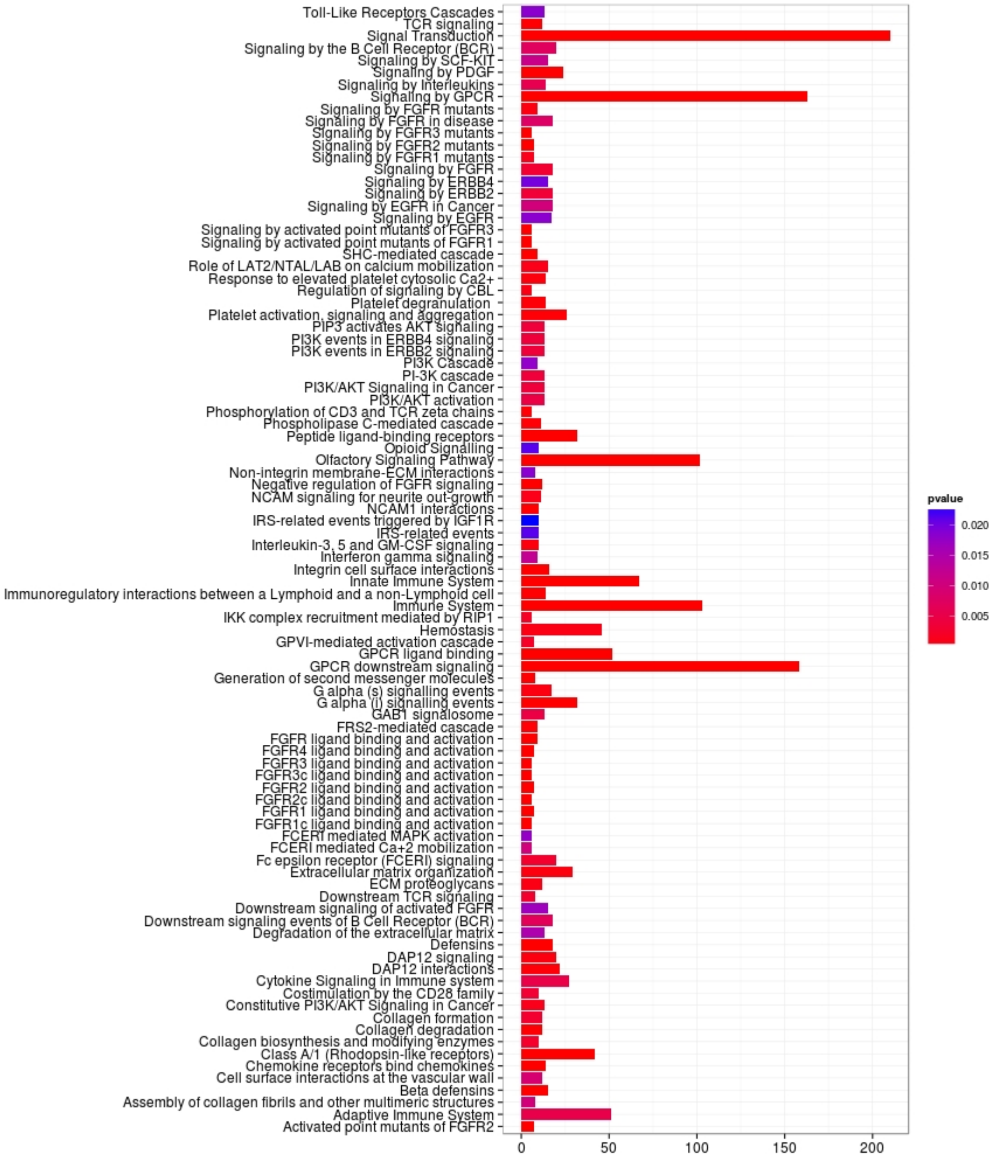


Figure 3.14.: Integrated pathway analysis of MCIA and betweenness scores

receptor homo- or heterodimerization at the cell surface. Dimerization of the extracellular regions of EGFR triggers additional conformational change of the cytoplasmic EGFR regions, enabling the kinase domains of two EGFR molecules to achieve the catalytically active conformation. Ligand activated EGFR dimers trans-autophosphorylate on tyrosine residues in the cytoplasmic tail of the receptor. Phosphorylated tyrosines serve as binding sites for the recruitment of signal transducers and activators of intracellular substrates, which then stimulate intracellular signal transduction cascades that are involved in regulating cellular proliferation, differentiation, and survival. Recruitment of complexes containing GRB2 and SOS1 to phosphorylated EGFR dimers either directly, through phosphotyrosine residues that serve as GRB2 docking sites, or indirectly, through SHC1 recruitment, promotes GDP to GTP exchange on RAS, resulting in the activation of RAF/MAP kinase cascade. Binding of complexes of GRB2 and GAB1 to phosphorylated EGFR dimers leads to formation of the active PI3K complex, conversion of PIP2 into PIP3, and activation of AKT signaling. Phospholipase C-gamma1 (PLCG1) can also be recruited directly, through EGFR phosphotyrosine residues that serve as PLCG1 docking sites, which leads to PLCG1 phosphorylation by EGFR and activation of DAG and IP3 signaling. EGFR signaling is downregulated by the action of ubiquitin ligase CBL. CBL binds directly to the phosphorylated EGFR dimer through the phosphotyrosine Y1045 in the C-tail of EGFR, and after CBL is phosphorylated by EGFR, it becomes active and ubiquitinates phosphorylated EGFR dimers, targeting them for degradation."

One of the most encouraging pathway is constitutive PI3K/AKT Signaling in cancer which is described in Reactome as: "Class IA PI3K is a heterodimer of a p85 regulatory subunit (encoded by PIK3R1, PIK3R2 or PIK3R3) and a p110 catalytic subunit (encoded by PIK3CA, PIK3CB or PIK3CD). In the absence of activating signals, the regulatory subunit stabilizes the catalytic subunit while inhibiting its activity. The complex becomes activated when extracellular signals stimulate the phosphorylation of the cytoplasmic domains of transmembrane receptors or receptor-associated proteins. The p85 regulatory subunit binds phosphorylated motifs of activator proteins, which induces a conformational change that relieves p85-mediated inhibition of the p110 catalytic subunit and enables PI3K to phosphorylate PIP2 to form PIP3. The phosphoinositide kinase activity of PI3K is opposed by the phosphoinositide phosphatase activity of PTEN. PIP3 acts as a messenger that recruits PDPK1 (PDK1) and AKT (AKT1, AKT2 or AKT3) to the plasma membrane. PDPK1 also possesses a low affinity for PIP2, so small amounts of PDPK1 are always present at the membrane. Binding of AKT to PIP3 induces a conformational change that enables TORC2 complex to phosphorylate AKT at a conserved serine residue (S473 in AKT1). Phosphorylation at the serine residue enables AKT to bind to PDPK1 and exposes a conserved threonine residue (T308) that is phosphorylated by PDPK1. AKT phosphorylated at both serine and threonine residues dissociates from the plasma membrane and acts as a serine/threonine kinase that phosphorylates a number of cytosolic and nuclear targets involved in regulation of cell metabolism, survival and gene expression.

Signaling by PI3K/AKT is frequently constitutively activated in cancer. This activation can be via gain-of-function mutations in PI3KCA (encoding catalytic subunit p110alpha), PIK3R1 (encoding regulatory subunit p85alpha) and AKT1. The PI3K/AKT pathway can also be constitutively activated by loss-of-function mutations in tumor suppressor genes such as PTEN.

Gain-of-function mutations activate PI3K signaling by diverse mechanisms. Mutations affecting the helical domain of PIK3CA and mutations affecting nSH2 and iSH2 domains of PIK3R1 impair inhibitory interactions between these two subunits while preserving their association. Mutations in the catalytic domain of PIK3CA enable the kinase to achieve an active conformation. PI3K complexes with gain-of-function mutations therefore produce PIP3 and activate downstream AKT in the absence of growth factors. While AKT1 gene copy number, expression level and phosphorylation are often increased in cancer, only one low frequency point mutation has been repeatedly reported in cancer and functionally studied. This mutation represents a substitution of a glutamic acid residue with lysine at position 17 of AKT1, and acts by enabling AKT1 to bind PIP2. PIP2-bound AKT1 is phosphorylated by TORC2 complex and by PDPK1 that is always present at the plasma membrane, due to low affinity for PIP2. Therefore, E17K substitution abrogates the need for PI3K in AKT1 activation.

Loss-of-function mutations affecting the phosphatase domain of PTEN are frequently found in sporadic cancers, as well as in PTEN hamartoma tumor syndromes (PHTS). PTEN can also be inactivated by gene deletion or epigenetic silencing, or indirectly by overexpression of microRNAs that target PTEN mRNA (Huse et al. 2009). Cells with deficient PTEN function have increased levels of PIP3, and therefore increased AKT activity.

Because of their clear involvement in human cancers, PI3K and AKT are targets of considerable interest in the development of small molecule inhibitors. Although none of the currently available inhibitors display preference for mutant variants of PIK3CA or AKT, several inhibitors targeting the wild-type kinases are undergoing clinical trials. These include dual PI3K/mTOR inhibitors, class I PI3K inhibitors, pan-PI3K inhibitors, and pan-AKT inhibitors. While none have yet been approved for clinical use, these agents show promise for future therapeutics. In addition, isoform-specific PI3K and AKT inhibitors are currently being developed, and may provide more specific treatments along with reduced side-effects." In addition to the FGFR and G alpha signaling events, when the MCIA scores are combined with the dynamical importance scores other promising pathways are found: beta-catenin independent WNT signaling which is described in Reactome as "Humans and mice have 19 identified WNT proteins that were originally classified as either 'canonical' or 'non-canonical' depending upon whether they were able to transform the mouse mammary epithelial cell line C57MG and to induce secondary axis formation in Xenopus. So-called canonical WNTs, including Wnt1, 3, 3a and 7, initiate signaling pathways that destabilize the destruction complex and allow beta-catenin to accumulate and translocate to the nucleus where it promotes transcription. Non-canonical WNTs, including Wnt 2, 4, 5a, 5b, 6, 7b, and Wnt11 activate beta-catenin-independent responses that regulate many aspects of morphogenesis and development, often by impinging on the cytoskeleton. Two of the main beta-catenin-independent pathways are the Planar Cell Polarity (PCP) pathway, which controls the establishment of polarity in the plane of a field of cells, and the WNT/Ca2+ pathway, which promotes the release of intracellular calcium and regulates numerous downstream effectors.", chemokine receptors bind chemokines "Chemokine receptors are cytokine receptors found on the surface of certain cells, which interact with a type of cytokine called a chemokine. Following interaction, these receptors trigger a flux of intracellular calcium which leads to chemotaxis. Chemokine receptors are divided into different families, CXC chemokine receptors, CC chemokine receptors, CX3C chemokine receptors and XC chemokine receptors that

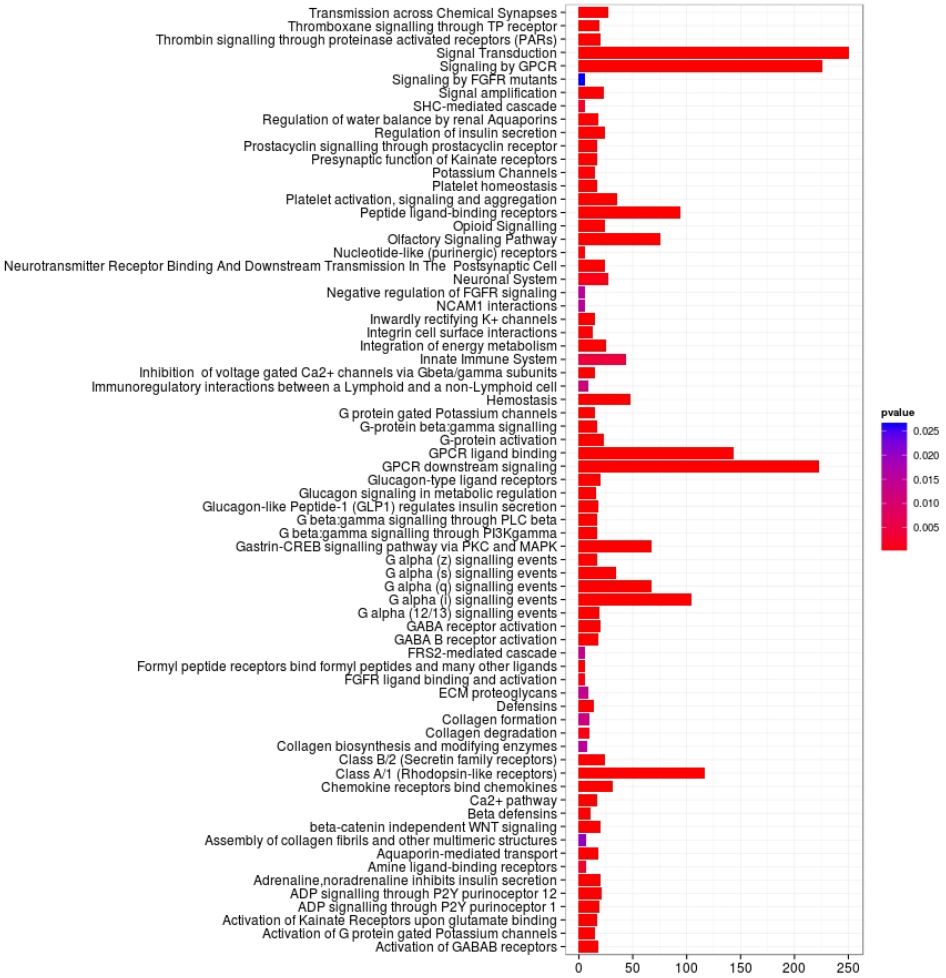


Figure 3.15.: Integrated pathway analysis of MCIA and dynamical importance scores

correspond to the 4 distinct subfamilies of chemokines they bind.", integrin cell surface interactions "The extracellular matrix (ECM) is a network of macro-molecules that underlies all epithelia and endothelia and that surrounds all connective tissue cells. This matrix provides the mechanical strength and also influences the behavior and differentiation state of cells in contact with it. The ECM are diverse in composition, but they generally comprise a mixture of fibrillar proteins, polysaccharides synthesized, secreted and organized by neighboring cells. Collagens, fibronectin, and laminins are the principal components involved in cell matrix interactions; other components, such as vitronectin, thrombospondin, and osteopontin, although less abundant, are also important adhesive molecules.

Integrins are the receptors that mediate cell adhesion to ECM. Integrins consists of one alpha and one beta subunit forming a noncovalently bound heterodimer. 18 alpha and 8 beta subunits have been identified in humans that combine to form 24 different receptors.

The integrin dimers can be broadly divided into three families consisting of the beta1, beta2/beta7, and beta3/alphaV integrins. beta1 associates with 12 alpha-subunits and can be further divided into RGD-, collagen-, or laminin binding and the related alpha4/alpha9 integrins that recognise both matrix and vascular ligands. beta2/beta7 integrins are restricted to leukocytes and mediate cell-cell rather than cell-matrix interactions, although some recognize fibrinogen. The beta3/alphaV family members are all RGD receptors and comprise aIIbb3, an important receptor on platelets, and the remaining b-subunits, which all associate with alphaV. It is the collagen receptors and leukocyte-specific integrins that contain alpha A-domains." and gastrin-CREB signaling via PKC and MAPK "Gastrin is a hormone whose main function is to stimulate secretion of hydrochloric acid by the gastric mucosa, which results in gastrin formation inhibition. This hormone also acts as a mitogenic factor for gastrointestinal epithelial cells. Gastrin has two biologically active peptide forms, G34 and G17.Gastrin gene expression is upregulated in both a number of pre-malignant conditions and in established cancer through a variety of mechanisms. Depending on the tissue where it is expressed and the level of expression, differential processing of the polypeptide product leads to the production of different biologically active peptides. In turn, acting through the classical gastrin cholecystokinin B receptor CCK-BR, its isoforms and alternative receptors, these peptides trigger signalling pathways which influence the expression of downstream genes that affect cell survival, angiogenesis and invasion”

3.4. Eigenanalysis of Tumor and Stroma Omics Data

The availability of multiple omics datasets from the same sample allows for a more complete understanding of pathway behavior in human diseases. However, pathway discovery is often based on flat gene lists which completely ignore the network topology of the pathways. Many methods use variations of Fisher's Exact Test to determine a data set's enrichment for a pathway while others consider only the ranks of the genes. More recent methods provide an approach to account for pathway structure but are computationally intensive and limited in their application. We propose an integrated pathway analysis approach where we combine feature (genes, proteins, CNV, etc.) scores from a multivariate analysis with an importance score for each feature in each pathway. These scores take into account the significance of each feature in the measured data sets as well as their topological importance within each pathway. We use two different measures for a feature's topological importance in a pathway and present results comparing enrichment in tumor and stroma microarray data from high grade serous ovarian cancer.

3.4.1. Motivation

The goal is to overcome the disadvantages of using flat gene lists for gene set enrichment analysis by accounting for the network topology of the pathways. The network topology plays a crucial role when computing the enrichment score of the toy pathway shown in Figure 3.16. The enrichment score should be higher when C and D are active as opposed to when F and G are active since the information flow through C and D is the largest in the network. This can be achieved by ranking the nodes according to their network centrality.

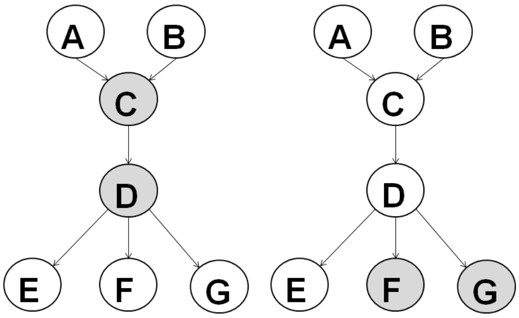


Figure 3.16.: Toy example: effect of network position of activated genes.

3.4.2. Data sets

Affymetrix U133 Plus 2.0 microarrays were used to analyse 38 tumor and stroma samples from high grade serous ovarian cancer [86]. For the pathway analysis we employed the curated pathway data base Reactome [31] which includes 1389 pathways and 6773 genes for homo sapiens.

3.4.3. Workflow

The work flow of the analysis is divided in two parts. In the first part (see Figure 3.18) the enrichment of the pathways is computed through the combination of the MCIA and the network gene scores for each pathway. In the second part the connectivity between the enriched pathways is computed through the combination of the MCIA scores with the network scores from the whole Reactome.

Pathway Scores

In order to compute the pathway enrichment first the MCIA has to be applied to the tumor and stroma data sets. The results of the analysis can be seen in Figure 3.17. In this figure, each sample is represented by a segment. Each end of a segment represents a data set. The blue circle represents the profile of the sample in the tumor data set and the brown triangle represents the profile of the sample in the stroma data set. The shorter the segment is,

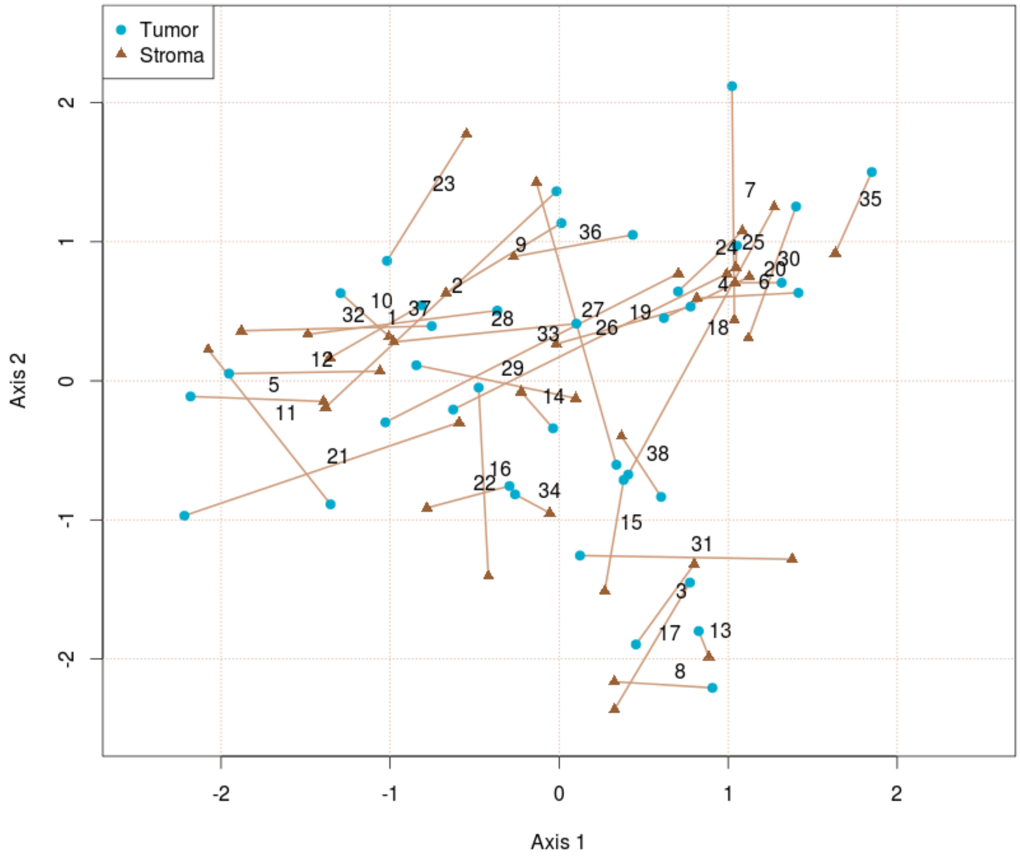


Figure 3.17.: MCIA result of the tumor and stroma data.

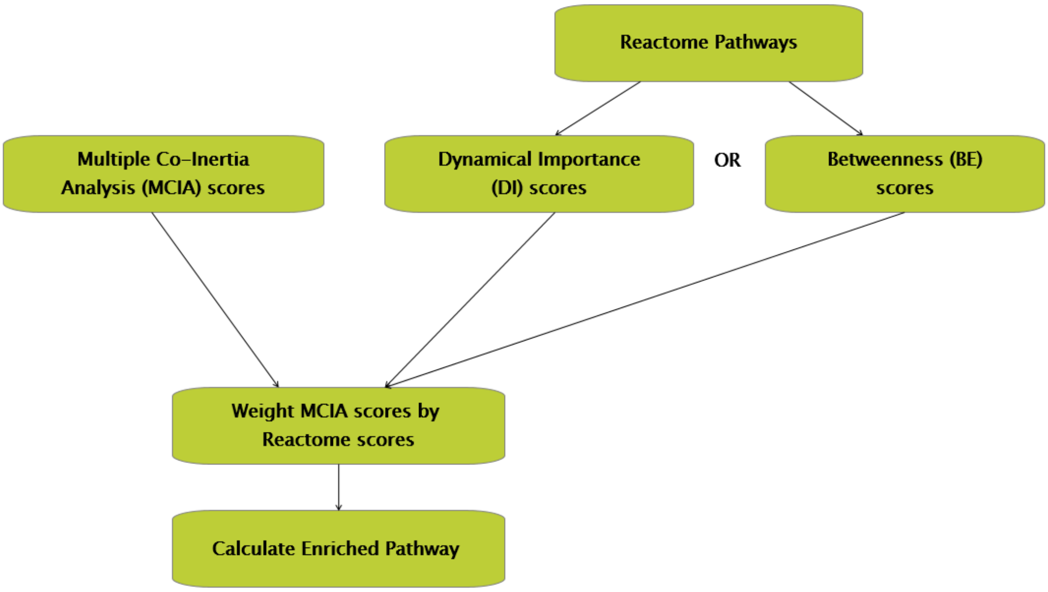

Reactome Pathways
Multiple Co-Inertia Analysis (MCIA) scores
Dynamical Importance (DI) scores
OR
Betweenness (BE) scores
Weight MCIA scores by Reactome scores
Calculate Enriched Pathway

Figure 3.18.: Workflow of the Eigenanlysis

the better agree the tumor and the stroma profile. If the profiles would be identical the length of the segment would be zero. We notice that there are sample with a high agreement between the tumor and the stroma profile (22,34,13) but there are also sample where the two profiles are very different (33,21,27).

Additionally, for each pathway in Reactome, the dynamical importance and the betweenness scores are computed. The R script written for this task can be found in the Appendix A.1. After all scores are computed these need to be combined. The combination of the scores was done in different ways: multiplication, correlation and spearman correlation. The R script used for the score combination was was included in the Appendix A.2.

A boxplot for the combined scores shows how these vary depending on the combination and normalization method as well as on the MCIA axis used for scoring. Each subplot of Figure 3.19 shows the influence of one from the ten MCIA axis used for the analysis. Additionally, in each subplot one can see the boxplots of (in this order):

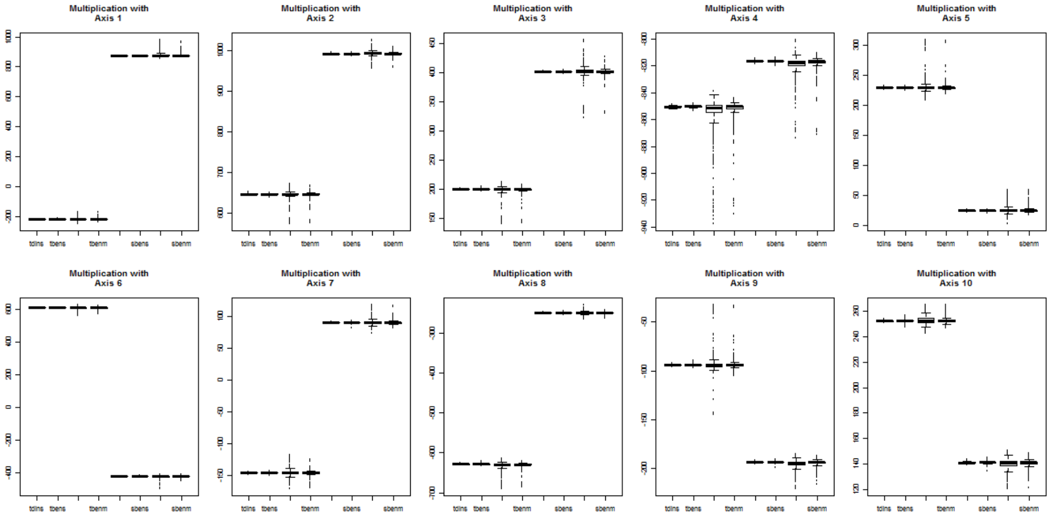


Figure 3.19.: Boxplots of pathway scores computed by multiplication of the normalized (through the sum or the maximum value) MCIA scores by the dynamical importance and the betweenness scores.

- tdins: tumor scores combined with the dynamical importance, normalized by the sum of the scores
- tbens: tumor scores combined with the betweenness, normalized by the sum of the scores
- tdinm: tumor scores combined with the dynamical importance, normalized by the maximum value of the scores
- tbenm: tumor scores combined with the betweenness, normalized by the maximum value of the scores
- sdins: stroma scores combined with the dynamical importance, normalized by the sum of the scores
- sbens: stroma scores combined with the betweenness, normalized by the sum of the scores
- sdinm: stroma scores combined with the dynamical importance, normalized by the maximum value of the scores
- sbenm: stroma scores combined with the betweenness, normalized by the maximum value of the scores

It is obvious that when the scores are combined through multiplication, there is a clear difference in the range of the pathway enrichment scores between the tumor and the stroma data. This difference can not be observed when the pathway enrichment is computed through pearson correlation or spearman correlation. The spearman correlation has the advantage of not being altered by outliers. Figures of pathways scores computed with these two methods are shown in the Appendix A.4 and A.5.

Pathway Connectivity

This analysis will not just deliver a flat list of enriched pathways, it will compute the connectivity of each pathway to the other enriched pathways and show them as a network. For this the connectivity between each pair of enriched pathways has to be computed. The end network of pathways will contain the top 10% connected pathways. The connectivity between two pathways is computed as follows:

First, all Reactome pathways are merged into a giant pathway. For this giant pathway, the dynamical importance and the betweenness of each gene is computed. Since the dynamical importance can be computed only for strongly connected pathways, we add to the adjacency matrix of the network one over the total number of genes. The next step is to combine through multiplication the rectome scores with the MCIA scores of the genes. The R script used for the computation of the network connectivity is included in the Appendix A.3.

In order to evaluate the network structure of the resulted enriched pathway, the assortativity of the network is examined. The assortativity [60] is defined as the correlation of the degrees of the directly linked nodes of a network. The value range between -1 and 1. Hong et al. showed that biological networks tend to have a negative assortativity while social networks tend to have a positive assortativity. In case of random networks the assortativity is around zero.

Based on different score combination methods as well as pathway selection method the assortativity of the network of enriched pathways was computed and is shown in Table 3.4. The different combination methods (see page 70) for the pathway enrichment score are shown in the columns. The rows of the table show the computation method for the connectivity of the enriched

	tdins	sdins	tbens	sbens	tdinm	sdinm	tbenm	sbenm
mult, amount	0.17	0.27	0.60	1.00	0.02	0.59	-0.12	0.589
mult, procent	0.23	0.27	0.67	1.00	0.02	0.64	-0.15	0.51
p.corr, amount	-0.07	0.40	-0.04	0.50	-0.07	0.40	-0.04	0.50
p.corr, procent	-0.08	0.46	-0.32	0.15	-0.08	0.46	-0.32	0.15
s.corr, amount	-0.18	-0.12	-0.02	-0.15	-0.18	-0.12	-0.02	-0.15
s.corr, procent	-0.18	-0.09	-0.38	-0.28	-0.18	-0.09	-0.38	-0.28

Table 3.4.: Table with the assortativity of differently computed networks of enriched pathways.

pathways. If the number of pathways in the end network is computed as the top 26 or the top 10% is indicated by the words *ammount* and *procent*.

It can be noticed that if the pathway enrichment is computed by multiplication, the assortativity is almost always positive and very high. This indicates that the computed network of pathways does not show a biological-like behavior. If the pathway connectivity is measured with the pearson or the spearman correlation the assortativity becomes negative. The most negative values (which indicate biological-like network structure) are reached when the pathway enrichment is computed from the betweenness and the normalization is by dividing through the sum.

For these normalization and combination methods the computed network of enriched pathways is shown. Figure 3.20 shows the resulted network for the tumor data with the pathway scores computed based on the betweenness. The corresponding network for the stroma data is shown in Figure 3.21. In the Appendix A.6 and A.7the resulted networks of tumor and stroma with pathway scores based on the dynamical importance are shown. For comparison reasons the resulted network for tumor and stroma, computed by score multiplication are shown in Appendix A.8 and A.9.

By examining Figures A.8 and A.9 we notice what the assortativity of a network expresses. As already mentioned, biological networks have a negative assortativity. This means that nodes with a high connectivity tend to connect to nodes with a low connectivity and vice versa. If a network has a positive assortativity highly connected nodes tend to be linked to other

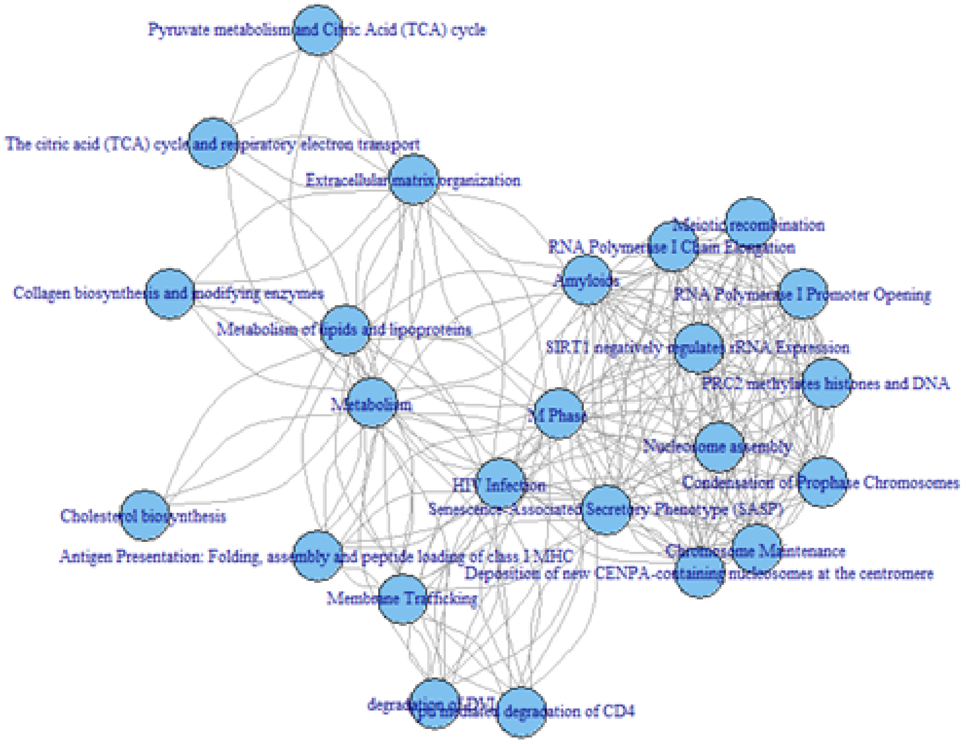

Pyruvate metabolism and Citric Acid (TCA) cycle
The citric acid (TCA) cycle and respiratory electron transport
Extracellular matrix organization
Meiotic recombination
RNA Polymerase I Chain Elongation
Amyloids
Collagen biosynthesis and modifying enzymes
RNA Polymerase I Promoter Opening
Metabolism of lipids and lipoproteins
SIRT1 negatively regulates rRNA Expression
PRC2 methylates histones and DNA
Metabolism
M Phase
Nucleosome assembly
Condensation of Prophase Chromosomes
HIV Infection
Senescence-Associated Secretory Phenotype (SASP)
Cholesterol biosynthesis
Chromosome Maintenance
Antigen Presentation: Folding, assembly and peptide loading of class I MHC
Deposition of new CENPA-containing nucleosomes at the centromere
Membrane Trafficking
degradation of DVL
Vpu mediated degradation of CD4

Figure 3.20.: Network of enriched pathways for the tumor data. Enrichment computed based on the betweenness

highly connected nodes and nodes with a low connectivity are linked to nodes with a similar connectivity.

A detailed inspection of the pathway network of the tumor data reveals that there are chromosome centered pathways like: deposition of new CENPA-containing nucleosomes at the centromere ("Eukaryotic centromeres are marked by a unique form of histone H3, designated CENPA in humans. In human cells newly synthesized CENPA is deposited in nucleosomes at the centromere during late telophase/early G1 phase of the cell cycle. Once deposited, nucleosomes containing CENPA remain stably associated with the centromere and are partitioned equally to daughter centromeres during S phase. A current model proposes that pre-existing CENPA at the centromere drives recruitment of new CENPA, however this has not been proved. The deposition process requires at least 3 complexes: the Mis18 complex, HJURP complex, and the RSF complex. HJURP binds newly synthesized CENPA-H4 tetramers before deposition and brings them to the centromere for deposition in new CENPA-containing nucleosomes. The exact mechanism of deposition remains unknown."), senescence associated secretory profile ("The culture medium of senescent cells in enriched in secreted proteins when compared with the culture medium of quiescent i.e. presenescent cells and these secreted proteins constitute the so-called senescence-associated secretory phenotype (SASP), also known as the senescence messaging secretome (SMS). SASP components include inflammatory and immune-modulatory cytokines (e.g. IL6 and IL8), growth factors (e.g. IGFBPs), shed cell surface molecules (e.g. TNF receptors) and survival factors. While the SASP exhibits a wide ranging profile, it is not significantly affected by the type of senescence trigger (oncogenic signalling, oxidative stress or DNA damage) or the cell type (epithelial vs. mesenchymal). However, as both oxidative stress and oncogenic signaling induce DNA damage, the persistent DNA damage may be a deciding SASP initiator. SASP components function in an autocrine manner, reinforcing the senescent phenotype, and in the paracrine manner, where they may promote epithelial-to-mesenchymal transition (EMT) and malignancy in the nearby premalignant or malignant cells. Interleukin-1-alpha (IL1A), a minor SASP component whose transcription is stimulated by the AP-1 (FOS:JUN) complex, can cause paracrine senescence through IL1 and inflammasome signaling. Here, transcriptional regulatory processes that mediate the SASP are annotated. DNA damage triggers ATM-mediated activation of TP53, resulting in the increased level of CDKN1A (p21). CDKN1A-mediated inhibition of CDK2 prevents phosphorylation and inactivation of the Cdh1:APC/C complex, allowing it to ubiquitinate and target for degradation EHMT1 and EHMT2 histone methyltransferases. As EHMT1 and EHMT2 methylate and silence the promoters of IL6 and IL8 genes, degradation of these methyltransferases relieves the inhibition of IL6 and IL8 transcription. In addition, oncogenic RAS signaling activates the CEBPB (C/EBP-beta) transcription factor, which binds promoters of IL6 and IL8 genes and stimulates their transcription. CEBPB also stimulates the transcription of CDKN2B (p15-INK4B), reinforcing the cell cycle arrest.

CEBPB transcription factor has three isoforms, due to three alternative translation start sites. The CEBPB-1 isoform (C/EBP-beta-1) seems to be exclusively involved in growth arrest and senescence, while the CEBPB-2 (C/EBP-beta-2) isoform may promote cellular proliferation (Atwood and Sealy 2010 and 2011). IL6 signaling stimulates the transcription of CEBPB, creating a positive feedback loop. NF-kappa-B transcription factor is also activated in senescence through IL1 signaling. NF-kappa-B binds IL6 and IL8 promoters and cooperates with CEBPB transcription factor in the induction of IL6 and IL8 transcription. Besides IL6 and IL8, their receptors are also upregulated in senescence and IL6 and IL8 may be master regulators of the SASP. IGFBP7 is also an SASP component that is upregulated in response to oncogenic RAS-RAF-MAPK signaling and oxidative stress, as its transcription is directly stimulated by the AP-1 (JUN:FOS) transcription factor. IGFBP7 negatively regulates RAS-RAF (BRAF)-MAPK signaling and is important for the establishment of senescence in melanocytes.") and PRC2 methylates histones and DNA ("Polycomb group proteins are responsible for the heritable repression of genes during development. Two major families of Polycomb complexes exist: Polycomb Repressive Complex 1 (PRC1) and Polycomb Repressive Complex 2 (PRC2). PRC1 and PRC2 each appear to comprise sets of distinct complexes that contain common core subunits and distinct accessory subunits. PRC2, through its component EZH2 or, in some complexes, EZH1 produces the initial molecular mark of repression, the trimethylation of lysine-27 of histone H3 (H3K27me3). How PRC2 is initially recruited to a locus remains unknown, however cytosine-guanine (CpG) motifs and transcripts have been suggested. Different mechanisms may be used at different loci. The trimethylated H3K27 produced by PRC2 is bound by the Polycomb subunit of PRC1. PRC1 ubiquitinates histone H2A and maintains repression.").

There are also metabolism centered pathways such as: pyruvate metabolism and citric acid cycle ("Pyruvate metabolism and the citric acid (TCA) cycle together link the processes of energy metabolism in a human cell with one another and with key biosynthetic reactions. Pyruvate, derived from the reversible oxidation of lactate or transamination of alanine, can be converted to acetyl CoA. Other sources of acetyl CoA include breakdown of free fatty acids and ketone bodies in the fasting state. Acetyl CoA can enter the citric acid cycle, a major source of reducing equivalents used to synthesize ATP, or enter biosynthetic pathways. In addition to its role in energy generation, the citric acid cycle is a source of carbon skeletons for amino acid metabolism and other biosynthetic processes. One such process included here is the interconversion of 2-hydroxyglutarate, probably derived from porphyrin and amino acid metabolism, and 2-oxoglutarate (alpha-ketoglutarate), a citric acid cycle intermediate.") and metabolism of lipids and lipoproteins ("Lipids are hydrophobic but otherwise chemically diverse molecules that play a wide variety of roles in human biology. They include ketone bodies, fatty acids, triacylglycerols, phospholipids and sphingolipids, eicosanoids, cholesterol, bile salts, steroid hormones, and fat-soluble vitamins. They function as a major source of energy (fatty acids, triacylglycerols, and ketone bodies), are major constituents of cell membranes (cholesterol and phospholipids), play a major role in their own digestion and uptake (bile salts), and participate in numerous signaling and regulatory processes (steroid hormones, eicosanoids, phosphatidylinositols, and sphingolipids). Because of their poor solubility in water, most lipids in extracellular spaces in the human body are found as complexes with specific carrier proteins. Regulation of the formation and movement of these lipoprotein complexes is a critical aspect of human lipid metabolism, and lipoprotein abnormalities are associated with major human disease processes including atherosclerosis and diabetes. The central steroid in human biology is cholesterol, obtained from animal fats consumed in the diet or synthesized de novo from acetyl-coenzyme A. (Vegetable fats contain various sterols but no cholesterol.) Cholesterol is an essential constituent of lipid bilayer membranes and is the starting point for the biosyntheses of bile acids and salts, steroid hormones, and vitamin D. Bile acids and salts are mostly synthesized in the liver. They are released into the intestine and function as detergents to solubilize dietary fats. Steroid hormones are mostly synthesized in the adrenal gland and gonads. They regulate energy metabolism and stress responses (glucocorticoids), salt balance (mineralocorticoids), and sexual development and function (androgens and estrogens). At the same time, chronically elevated cholesterol levels in the body are associated with the formation of atherosclerotic lesions and hence increased risk of heart attacks and strokes. The human body lacks a mechanism for degrading excess cholesterol, although an appreciable amount is lost daily in the form of bile salts and acids that escape recycling. Aspects of lipid metabolism currently annotated in Reactome include lipid digestion, mobilization, and transport; fatty acid, triacylglycerol, and ketone body metabolism; peroxisomal lipid metabolism; phospholipid and sphingolipid metabolism; cholesterol biosynthesis; bile acid and bile salt metabolism; and steroid hormone biosynthesis. ").

Additionally, pathways of the extracellular matrix are enriched: extracellular matrix organisation ("The extracellular matrix is a component of all mammalian tissues, a network consisting largely of the fibrous proteins collagen, elastin and associated-microfibrils, fibronectin and laminins embedded in a viscoelastic gel of anionic proteoglycan polymers. It performs many functions in addition to its structural role; as a major component of the cellular microenvironment it influences cell behaviours such as proliferation, adhesion and migration, and regulates cell differentiation and death. ECM composition is highly heterogeneous and dynamic, being constantly remodeled and modulated, largely by matrix metalloproteinases (MMPs) and growth factors that bind to the ECM influencing the synthesis, crosslinking and degradation of ECM components (Hynes 2009). ECM remodeling is involved in the regulation of cell differentiation processes such as the establishment and maintenance of stem cell niches, branching morphogenesis, angiogenesis, bone remodeling, and wound repair. Redundant mechanisms modulate the expression and function of ECM modifying enzymes. Abnormal ECM dynamics can lead to deregulated cell proliferation and invasion, failure of cell death, and loss of cell differentiation, resulting in congenital defects and pathological processes including tissue fibrosis and cancer.") and collagen biosynthesis and modifing enzymes ("The biosynthesis of collagen is a multistep process. Collagen propeptides are cotranslationally translocated into the ER lumen. Propeptides undergo a number of post-translational modifications. Proline and lysine residues may be hydroxylated by prolyl 3-, prolyl 4- and lysyl hydroxylases. 4-hydroxyproline is essential for intramolecular hydrogen bonding and stability of the triple helical collagenous domain. In fibril forming collagens approximately 50% of prolines are 4-hydroxylated; the extent of this and of 3-hydroxyproline and lysine hydroxylation varies between tissues and collagen types. Hydroxylysine molecules can form cross-links between collagen molecules in fibrils, and are sites for glycosyl- and galactosylation. Collagen peptides all have non-collagenous domains; collagens within the subclasses have common chain structures. These non-collagenous domains have regulatory functions; some are biologically active when cleaved from the main peptide chain. Fibrillar

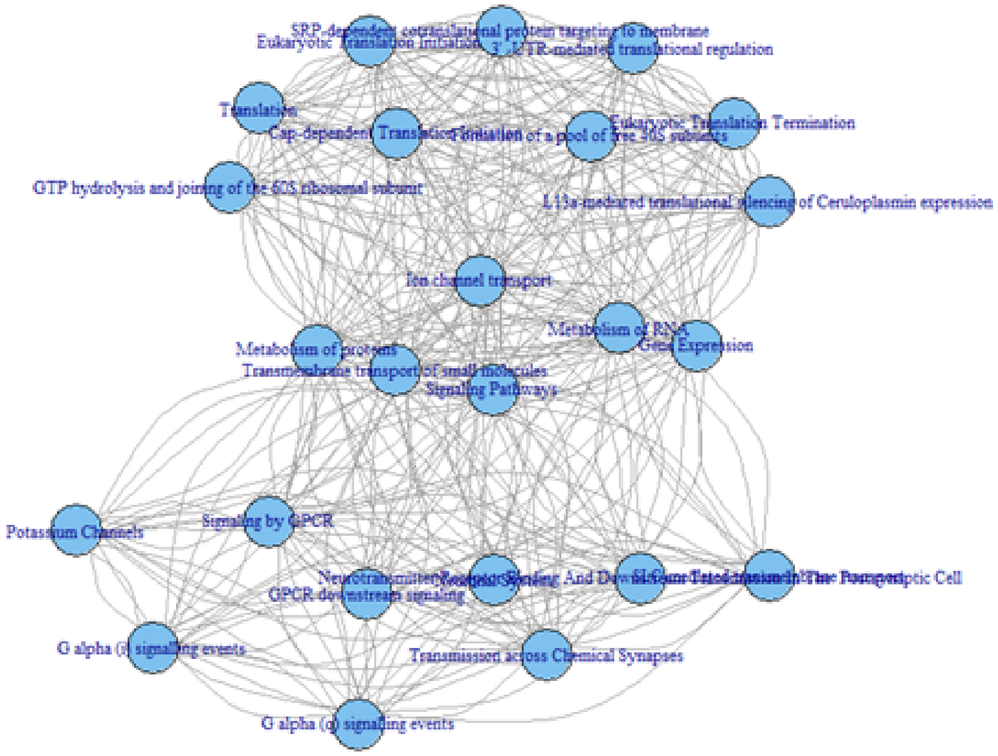

SRP-dependent cotranslational protein targeting to membrane
Eukaryotic Translation Initiation
3' -UTR-mediated translational regulation
Translation
Cap-dependent Translation Initiation
Formation of a pool of free 40S subunits
Eukaryotic Translation Termination
GTP hydrolysis and joining of the 60S ribosomal subunit
L13a-mediated translational silencing of Ceruloplasmin expression
Ion channel transport
Metabolism of RNA
Gene Expression
Metabolism of proteins
Transmembrane transport of small molecules
Signaling Pathways
Potassium Channels
Signaling by GPCR
Neurotransmitter Receptor Binding And Downstream
Postsynaptic Cell
GPCR downstream signaling
G alpha (i) signalling events
Transmission across Chemical Synapses
G alpha (q) signalling events

Figure 3.21.: Network of enriched pathways for the stroma data. Enrichment computed based on the betweenness

collagens all have a large triple helical domain (COL1) bordered by N and C terminal extensions, called the N and C propeptides, which are cleaved prior to formation of the collagen fibril. The C propeptide, also called the NC1 domain, is highly conserved. It directs chain association during intracellular assembly of the procollagen molecule from three collagen propeptide alpha chains. The N-propeptide has a short linker (NC2) connecting the main triple helix to a short minor one (COL2) and a globular N-terminal region NC3. NC3 domains are variable both in size and the domains they contain.")

A closer look at the pathway network enriched in stroma shows a number of transaltional pathways as well as signaling centered pathways. Interesting pathways are: signaling by GPCR (“G protein-coupled receptors (GPCRs; 7TM receptors; seven transmembrane domain receptors; heptahelical receptors; G protein-linked receptors [GPLR]) are the largest family of transmembrane receptors in humans, accounting for more than 1% of the protein-coding capacity of the human genome. All known GPCRs share a common architecture of seven membrane-spanning helices connected by intra- and extracellular loops. The extracellular loops contain two highly-conserved cysteine residues that form disulphide bonds to stabilize the structure of the receptor. They recognize diverse messengers such as light, odorants, small molecules, hormones and neurotransmitters. Most GPCRs act as guanine nucleotide exchange factors; activated by ligand binding, they promote GDP-GTP exchange on associated heterotrimeric guanine nucleotide-binding (G) proteins. There are two models for GPCR-G Protein interactions: 1) ligand-GPCR binding first, then binding to G Proteins; 2) "Pre-coupling" of GPCRs and G Proteins before ligand binding. These in turn activate effector enzymes or ion channels. GPCRs are involved in a range of physiological roles which include the visual sense, smell, behavioural regulation, functions of the autonomic nervous system and regulation of the immune system and inflammation.”), potassium channels (“Potassium channels are tetrameric ion channels that are widely distributed and are found in all cell types. Potassium channels control resting membrane potential in neurons, contribute to regulation of action potentials in cardiac muscle and help release of insulin form pancreatic beta cells. Broadly K+ channels are classified into voltage gated K+ channels, Hyperpolarization activated cyclic nucleotide gated K+ channels (HCN), Tandem pore domain K+ channels, Ca2+ activated K+ channels and inwardly rectifying K+ channels.”), L13a-mediated translational silencing of Ceruloplasmin expression (“While circularization of mRNA during translation initiation is thought to contribute to an increase in the efficiency of translation, it also appears to provide a mechanism for translational silencing. This might be achieved by bringing inhibitory 3′ UTR-binding proteins into a position in which they interfere either with the function of the translation initiation complex or with the assembly of the ribosome (Mazumder et al 2001). Translational silencing of Ceruloplasmin (Cp) occurs 16 hrs after its induction by INF-gamma. Although the mechanism by which silencing occurs has not yet been determined, this process is mediated by the L13a subunit of the 60s ribosome and thought to require circularization of the Cp mRNA. Between 14 and 16 hrs after INF gamma induction, the L13a subunit of the 60s ribosome is phosphorylated and released from the 60s subunit. Phosphorylated L13a then associates with the GAIT element in the 3′ UTR of the Cp mRNA inhibiting its translation.")

Different pathways were shown to be enriched in tumor and stroma. Some are already linked to cancer, the other require further investigation. All shown pathway descriptions were taken from Reactome [103].

The described method is currently being summarized into an R package. We are working on the platform independent implementation. We plan to publish the R package together with the manuscript which describes the eigenanalysis of dynamically important pathways in multiple omics cancer data.

Appendix A.

Methods and Results

```r
#### R script to compute the dynamical importance and the betweenness for each pathway in reactome

rm(list = ls())
source(paste(.path,"src\\dynamicalImportance.R",sep = ""))
library(graphite)
library(igraph)

rankGenesInPathway = function(pwy, rank.method){
  edges = pwy@edges; edges
  rank = rep(0,length(pwy@nodes))
  names(rank) = pwy@nodes

  if(dim(edges)[1] != 0)
  {
    help1 = unique(edges[which(edges$direction == "undirected"),c(1,2)]); help1
    help2 = unique(edges[which(edges$direction == "undirected"),c(2,1)])
    colnames(help2) = c("src","dest"); help2
    help3 = unique(edges[which(edges$direction == "directed"),c(1,2)]); help3

    all.edges = as.matrix(unique(rbind(help1,help2,help3))); all.edges
    pwy.graph = graph.edgelist(all.edges, directed = T)

    if(rank.method == "di")
    rank = dynamicalImportance(pwy.graph)
    if(rank.method == "be")
    rank = betweenness(pwy.graph)
  }
  return(rank)
}

load(file = paste(.res.dir.RData, "reactome_pathways_as_symbols_with_nodes.RData", sep = ""))

dyn.imp = lapply(X = reactome.pathways,FUN = rankGenesInPathway, rank.method = "di")
save(dyn.imp, file = paste(.res.dir.RData,"dyn_imp_per_path_with_nodes.RData",sep = ""))

bet.ness = lapply(X = reactome.pathways,FUN = rankGenesInPathway, rank.method = "be")
save(bet.ness, file = paste(.res.dir.RData,"bet_ness_per_path_with_nodes.RData",sep = ""))


load(file = paste(.res.dir.RData, "reactome_pathways_as_symbols_with_nodes_and_edges.RData", sep = ""))


dyn.imp = lapply(X = reactome.pathways,FUN = rankGenesInPathway, rank.method = "di")
save(dyn.imp, file = paste(.res.dir.RData,"dyn_imp_per_path_with_nodes_and_edges.RData",sep = ""))

```

```r
bet.ness = lapply(X = reactome.pathways,FUN = rankGenesInPathway, rank.method = "be")
save(bet.ness, file = paste(.res.dir.RData,"bet_ness_per_path_with_nodes_and_edges.RData",sep = ""))
```

Listing A.1: R script to compute the dynamical importance and the betweenness for each pathway in reactome

```r
##### R script to combine MCIA scores with the dynamical importance and the betweenness scores for each pathway in reactome

rm(list = ls())
source("set_paths_for_current_project.R")
load(file = paste(.res.dir.RData,"all_scores.RData", sep =""))

di.nm = apply(X = dyn.imp.scores,MARGIN = 2,FUN = function(col){ if(max(col) != 0) col/max(col)+1 else col})
be.nm = apply(X = bet.ness.scores,MARGIN = 2,FUN = function(col){if(max(col) != 0) col/max(col)+1 else col+1})
di.ns= apply(X = dyn.imp.scores,MARGIN = 2,FUN = function(col){if(sum(col) != 0)  col/sum(col)+1 else col})
be.ns = apply(X = bet.ness.scores,MARGIN = 2,FUN = function(col){if(sum(col) != 0) col/sum(col)+1 else col+1})

#range(di.nm);range(di.ns)
#range(be.nm);range(be.ns)

n.axis = dim(mcia.scores.stroma)[2]

pathway.scores.per.axes = list()

for(axis in 1:n.axis){
```

```
#axis = 1
mcia.axis = axis
#hist(mts)
#hist(dyn.imp.scores[,1])

mts = mcia.scores.tumor[,paste("Axis",mcia.axis,sep="")]
mss = mcia.scores.stroma[,paste("Axis",mcia.axis, sep="")]

range(mts)
range(mss)


# multiplication

tdi.ns = sweep(x = di.ns, MARGIN = 1,STATS = mts, FUN = "*")
sdi.ns = sweep(x = di.ns, MARGIN = 1,STATS = mss, FUN = "*")

tbe.ns = sweep(x = be.ns, MARGIN = 1,STATS = mts, FUN = "*")
sbe.ns = sweep(x = be.ns, MARGIN = 1,STATS = mss, FUN = "*")

tdi.nm = sweep(x = di.nm, MARGIN = 1,STATS = mts, FUN = "*")
sdi.nm = sweep(x = di.nm, MARGIN = 1,STATS = mss, FUN = "*")

tbe.nm = sweep(x = be.nm, MARGIN = 1,STATS = mts, FUN = "*")
sbe.nm = sweep(x = be.nm, MARGIN = 1,STATS = mss, FUN = "*")

# pearson correlation
```

```

tdi.ns.c = apply(X = di.ns,MARGIN = 2,FUN = cor,y = mts)
sdi.ns.c = apply(X = di.ns,MARGIN = 2,FUN = cor,y = mss)

tbe.ns.c = apply(X = be.ns,MARGIN = 2,FUN = cor,y = mts)
sbe.ns.c = apply(X = be.ns,MARGIN = 2,FUN = cor,y = mss)

tdi.nm.c = apply(X = di.nm,MARGIN = 2,FUN = cor,y = mts)
sdi.nm.c = apply(X = di.nm,MARGIN = 2,FUN = cor,y = mss)

tbe.nm.c = apply(X = be.nm,MARGIN = 2,FUN = cor,y = mts)
sbe.nm.c = apply(X = be.nm,MARGIN = 2,FUN = cor,y = mss)

# spearman correlation

tdi.ns.cs = apply(X = di.ns,MARGIN = 2,FUN = cor,y = mts, method = "spearman")
sdi.ns.cs = apply(X = di.ns,MARGIN = 2,FUN = cor,y = mss, method = "spearman")

tbe.ns.cs = apply(X = be.ns,MARGIN = 2,FUN = cor,y = mts, method = "spearman")
sbe.ns.cs = apply(X = be.ns,MARGIN = 2,FUN = cor,y = mss, method = "spearman")

tdi.nm.cs = apply(X = di.nm,MARGIN = 2,FUN = cor,y = mts, method = "spearman")
sdi.nm.cs = apply(X = di.nm,MARGIN = 2,FUN = cor,y = mss, method = "spearman")
```

```

  tbe.nm.cs = apply(X = be.nm,MARGIN = 2,FUN = cor,y = mts, method = "spearman")
  sbe.nm.cs = apply(X = be.nm,MARGIN = 2,FUN = cor,y = mss, method = "spearman")

  overall.pathway.scores = data.frame(tdins = colSums(tdi.ns),
  tbens = colSums(tbe.ns),
  tdinm = colSums(tdi.nm),
  tbenm = colSums(tbe.nm),
  sdins = colSums(sdi.ns),
  sbens = colSums(sbe.ns),
  sdinm = colSums(sdi.nm),
  sbenm = colSums(sbe.nm),
  tdinsc = tdi.ns.c,
  tbensc = tbe.ns.c,
  tdinmc = tdi.nm.c,
  tbenmc = tbe.nm.c,
  sdinsc = sdi.ns.c,
  sbensc = sbe.ns.c,
  sdinmc = sdi.nm.c,
  sbenmc = sbe.nm.c,
  tdinscs = tdi.ns.cs,
  tbenscs = tbe.ns.cs,
  tdinmcs = tdi.nm.cs,
  tbenmcs = tbe.nm.cs,
  sdinscs = sdi.ns.cs,
  sbenscs = sbe.ns.cs,
  sdinmcs = sdi.nm.cs,
  sbenmcs = sbe.nm.cs)
  rownames(overall.pathway.scores) = colnames(dyn.imp.scores)

  pathway.scores.per.axes[[mcia.axis]] = overall.pathway.scores
}
```

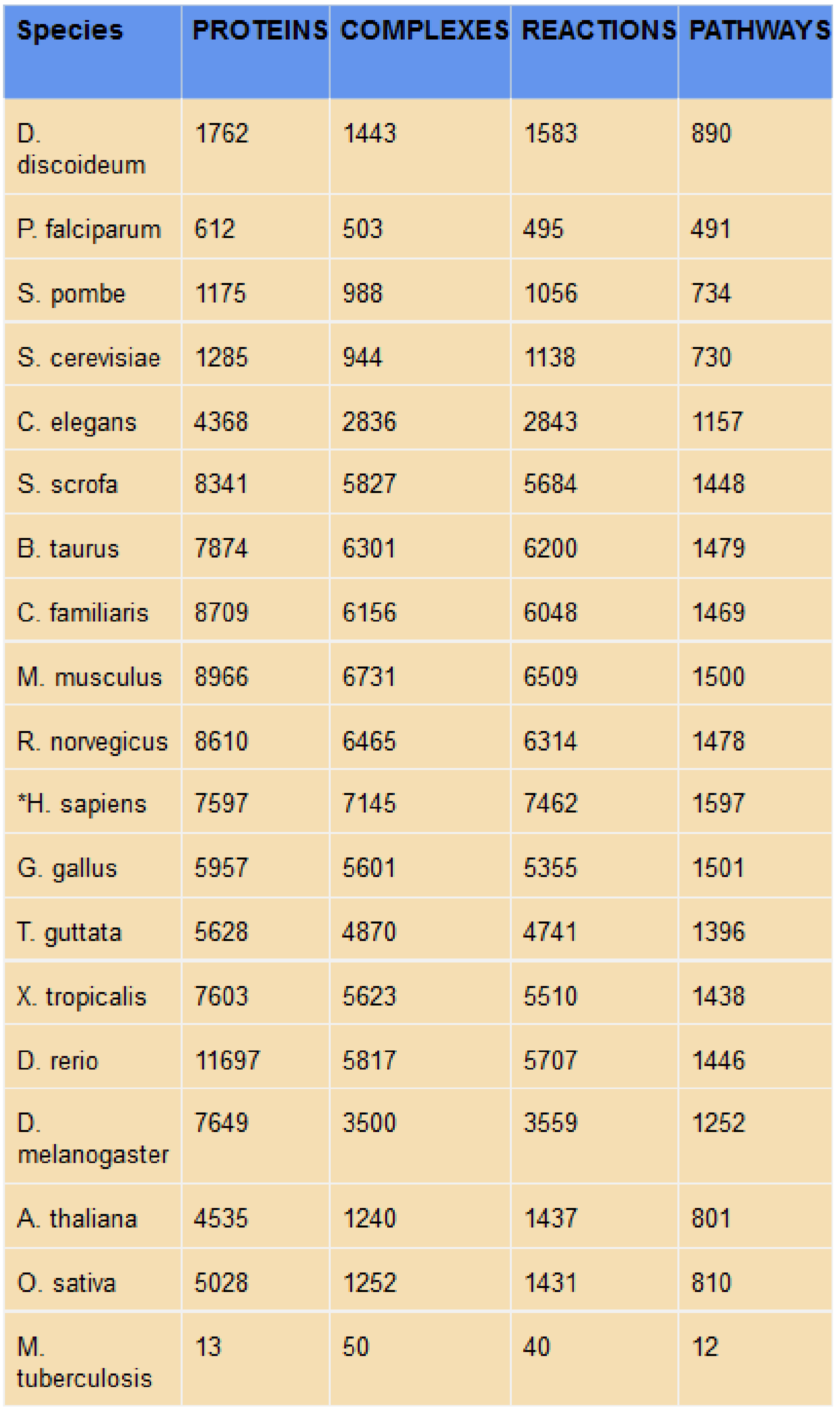
```r

save(pathway.scores.per.axes, file = paste(.res.dir.RData, "pathway_scores_for_all_mcia_axes.RData", sep = ""))
```

Listing A.2: R script to combine MCIA scores with the dynamical importance and the betweenness scores for each pathway in reactome

```r
# get resulted pathways
rm(list = ls())
source("set_paths_for_current_project.R")

load(paste(.res.dir.RData,"all_scores_for_net_connet_calc.RData", sep = ""))
load(file = paste(.res.dir.RData, "pathway_scores_for_all_mcia_axes.RData", sep = ""))

res.paths.a1 = pathway.scores.per.axes[[1]]
str(res.paths.a1)

use.procent = T; p = 0.01
use.amount = T; n.paths = 26

selected.pathways.procent = list()
selected.pathways.amount = list()

# select pathways
for(i in 1:dim(res.paths.a1)[2]){
  curr.scores = res.paths.a1[,i]
  names(curr.scores) = rownames(res.paths.a1)
  if(use.procent){
    qs = quantile(x = curr.scores, probs = c(p, 1-p), na.rm = T); qs
    sel.paths = curr.scores[which(curr.scores < qs[1]| curr.scores > qs[2])]; sel.paths; length(sel.paths)
    selected.pathways.procent[[i]] = names(sel.paths)
    names(selected.pathways.procent)[i] = colnames(
      res.paths.a1)[i]
  }
  if(use.amount){
    nn1 = round(n.paths/2,0); nn1
    nn3 = dim(res.paths.a1)[1]; nn3
    nn2 = nn3 - nn1+1; nn2

    help = sort(x = curr.scores, decreasing = F)
    sel.paths = help[c(1:nn1,nn2:nn3)]; sel.paths; length(sel.paths)

    selected.pathways.amount[[i]] = names(sel.paths)
    names(selected.pathways.amount)[i] = colnames(res.paths.a1)[i]
  }
}

save(selected.pathways.amount, selected.pathways.procent, file = paste(.res.dir.RData, "selected_pathways.RData", sep = ""))


##### compute connectivity between pathways
rm(list = ls())

computeConnectivity = function(ps, mcia.scores, reactome.scores){

  p1 = ps[1]; #cat("p1 = ", p1, "\n");
  p2 = ps[2]; #cat("p2 = ", p2, "\n");

  con = rep(0,dim(mcia.scores)[2])

  names(con) = colnames(mcia.scores)
  if(p1 == p2)
    return(con)

  rp1 = reactome.pathways[[as.character(p1)]]; #rp1
  rp2 = reactome.pathways[[as.character(p2)]]; #rp2

  nodes1 = rp1@nodes; nodes1
  nodes2 = rp2@nodes; nodes2

  mg = intersect(rownames(mcia.scores), names(
     reactome.scores))
  cn = intersect(intersect(nodes1,nodes2),mg); cn

  #cat("cn: ", cn, "\n")

  if(length(cn)>0){
    help.stats = matrix(reactome.scores[cn], ncol =
       1, nrow = length(cn));
    help.x = matrix(mcia.scores[cn,], nrow = length(
       cn), ncol = dim(mcia.scores)[2])
    colnames(help.x) = colnames(mcia.scores)
    con = colSums(sweep(x = help.x, MARGIN = 1, STATS
        = help.stats,FUN = "*")); #con
  }
  return(con)
}

# prepare result for visualisation

computeCytoscapeRes = function(path.con, axis, cyto.
   file){

  cytoscape.res = data.frame(SRC = "from", INT = "pp"
     , DEST = "to", stringsAsFactors =  F)

  for(i in 1:length(path.con)){
    ps.list[i]
    if(path.con[[i]][axis]!=0){
      help = c(ps.list[[i]]$x, "pp",ps.list[[i]]$y);
         help
    cytoscape.res = rbind(cytoscape.res, help)
  }
}

cytoscape.res = cytoscape.res[-1,]
path.con.graph = graph.edgelist(el = as.matrix(cytoscape.res[,c(1,3)]), directed = FALSE)

png(file = paste(.res.dir.figures,strsplit(cyto.file, split = "_")[[1]][1],".png", sep = ""), width = 1200, height = 700)
plot(path.con.graph, main = strsplit(cyto.file, split = "_")[[1]][1])
dev.off()

degree.path.con.graph = degree(graph = path.con.graph, mode = "total")
if(sd(degree.path.con.graph) != 0){
  igraph.assortativity = assortativity.degree(graph = path.con.graph,directed = T)
  cat("igraph.assortativity",igraph.assortativity,"\n")

  my.assortativity = cor(degree.path.con.graph[cytoscape.res[,1]],
  degree.path.con.graph[cytoscape.res[,3]], method = "pearson")
  cat("my.assortativity = ", my.assortativity, "\n")

}
else{
  cat("All nodes have the same degree: ", unique(degree.path.con.graph ), "\n")
  my.assortativity = NA
}

```

```
  write.table(cytoscape.res, file = paste(.res.dir.RData,cyto.file, sep = ""),
  quote = T, row.names = FALSE, col.names = T, sep = "\t")
  return(my.assortativity)
}
```

Listing A.3: R script to compute connectivity of enriched pathways



Species	PROTEINS	COMPLEXES	REACTIONS	PATHWAYS
D. discoideum	1762	1443	1583	890
P. falciparum	612	503	495	491
S. pombe	1175	988	1056	734
S. cerevisiae	1285	944	1138	730
C. elegans	4368	2836	2843	1157
S. scrofa	8341	5827	5684	1448
B. taurus	7874	6301	6200	1479
C. familiaris	8709	6156	6048	1469
M. musculus	8966	6731	6509	1500
R. norvegicus	8610	6465	6314	1478
*H. sapiens	7597	7145	7462	1597
G. gallus	5957	5601	5355	1501
T. guttata	5628	4870	4741	1396
X. tropicalis	7603	5623	5510	1438
D. rerio	11697	5817	5707	1446
D. melanogaster	7649	3500	3559	1252
A. thaliana	4535	1240	1437	801
O. sativa	5028	1252	1431	810
M. tuberculosis	13	50	40	12

Figure A.1.: Overview of the Reactome content: species, proteins, complexes, reactions and pathways. Figure adapted from [103].

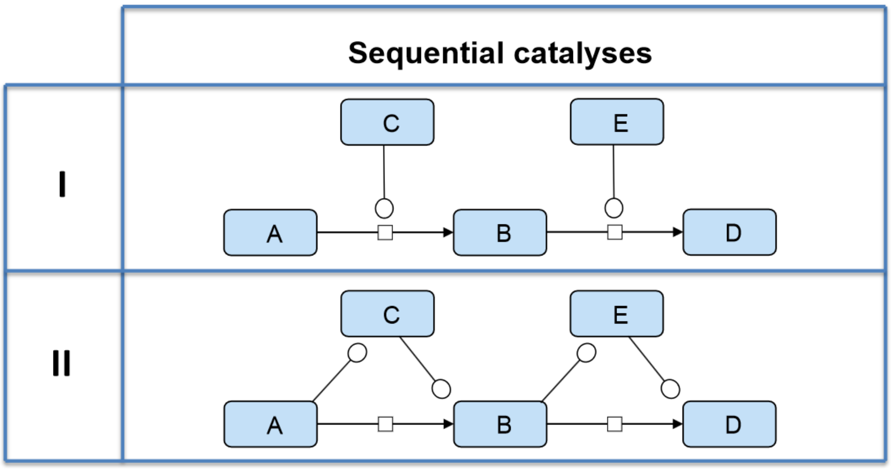


Figure A.2.: Example of simplification rules from a Reactome pathway to a graphite network: simplifying sequentially catalyzed processes. Figure addapted from [111].

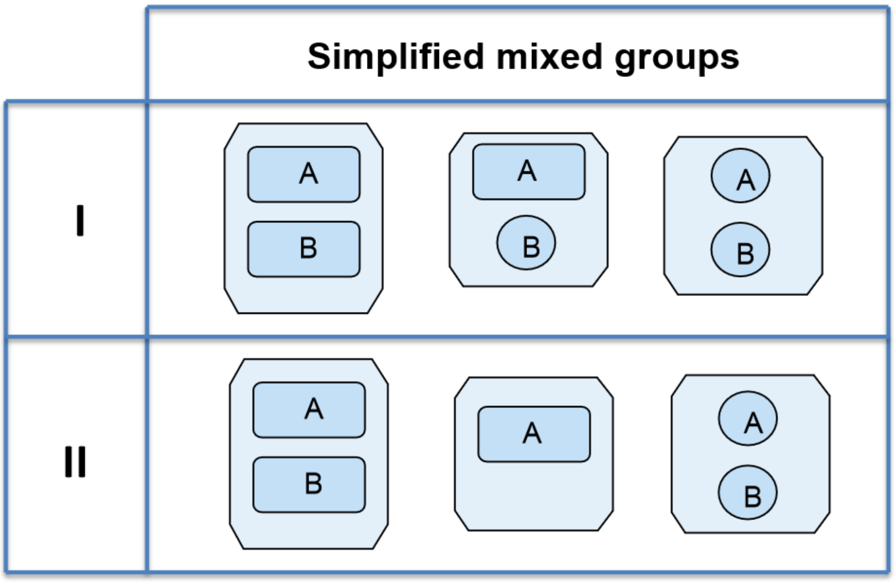


Figure A.3.: Example of simplification rules from a Reactome pathway to a graphite network: simplifying mixed groups. Figure addapted from [111].

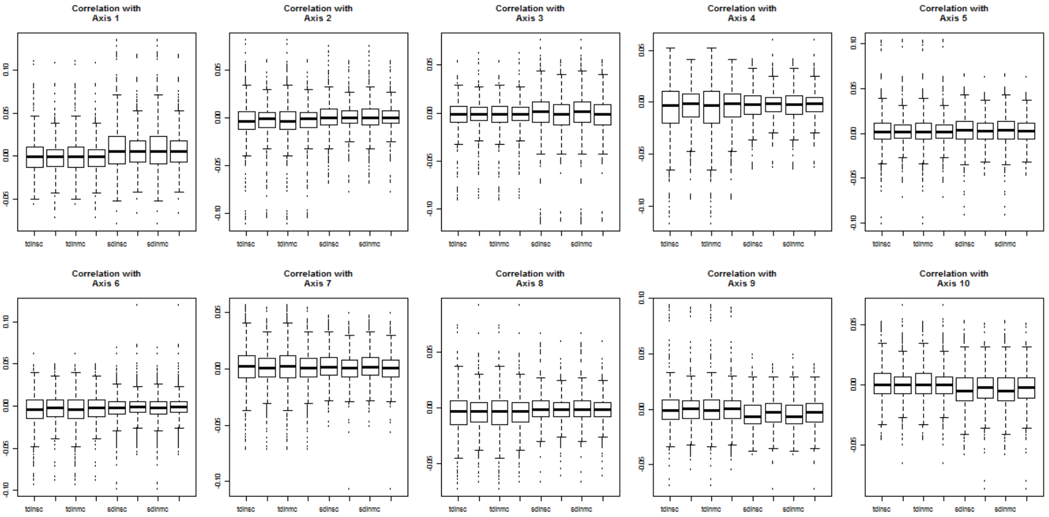


Figure A.4.: Boxplots of pathway scores computed by the pearson correlation of the normalized (through the sum or the maximum value)

MCIA scores and the dynamical importance and the betweenness scores.

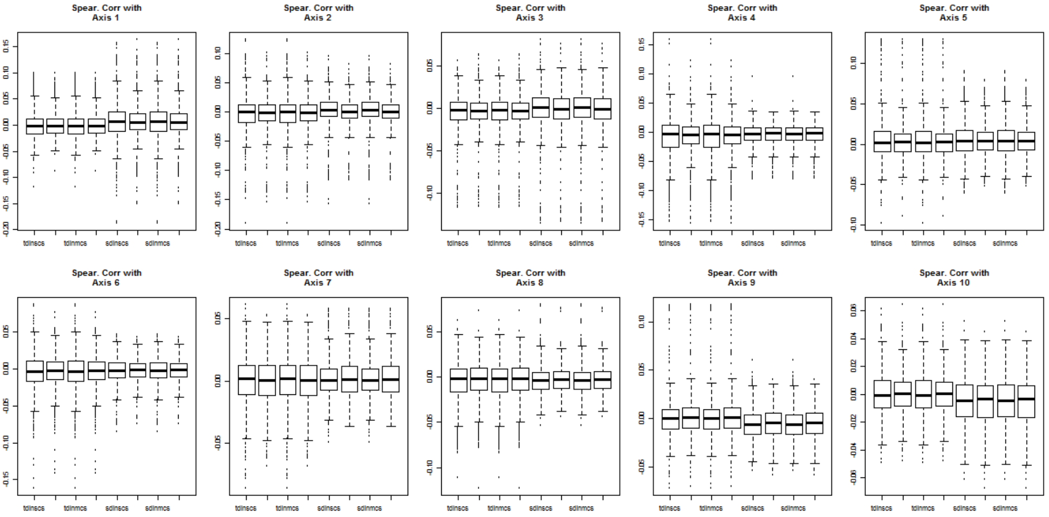


Figure A.5.: Boxplots of pathway scores computed by the pspearman correlation of the normalized (through the sum or the maximum value)

MCIA scores and the dynamical importance and the betweenness scores.

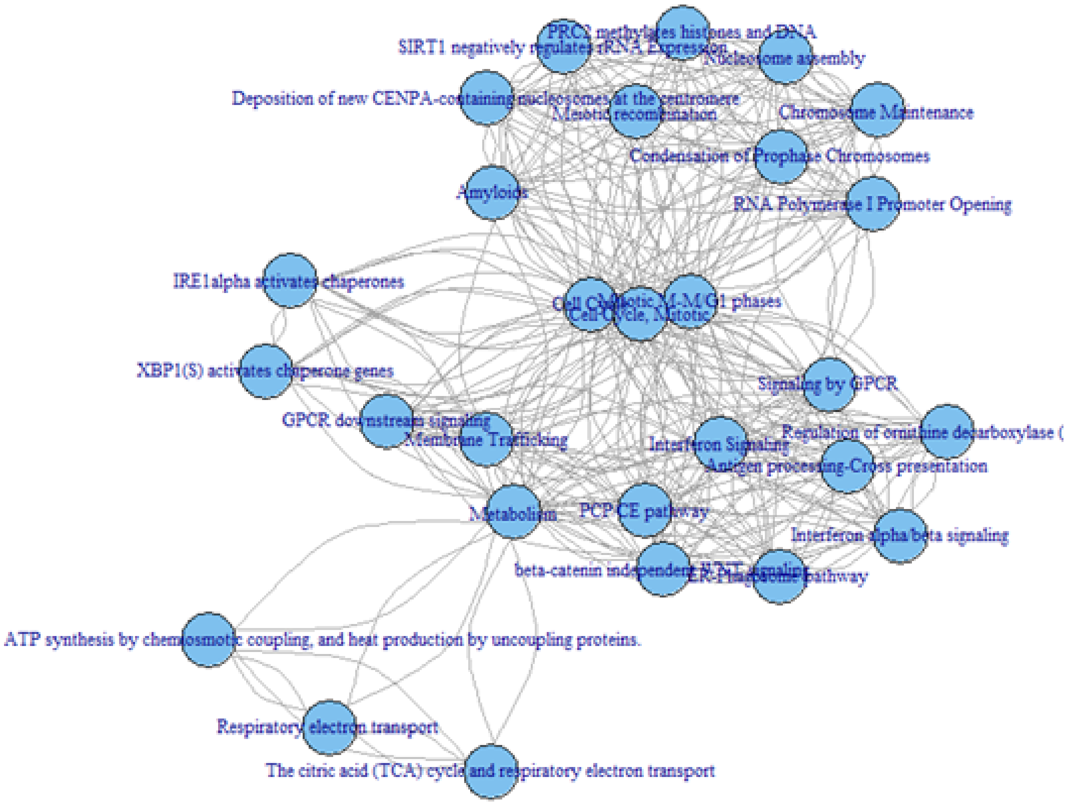

PRC2 methylates histones and DNA
SIRT1 negatively regulates rRNA Expression
Nucleosome assembly
Deposition of new CENPA-containing nucleosomes at the centromere
Meiotic recombination
Chromosome Maintenance
Condensation of Prophase Chromosomes
Amyloids
RNA Polymerase I Promoter Opening
IRE1alpha activates chaperones
Cell Cycle
Mitotic M-M/G1 phases
Cell Cycle, Mitotic
XBP1(S) activates chaperone genes
Signaling by GPCR
GPCR downstream signaling
Membrane Trafficking
Regulation of ornithine decarboxylase (
Interferon Signaling
Antigen processing-Cross presentation
Metabolism
PCP/CE pathway
Interferon alpha/beta signaling
beta-catenin independent WNT signaling
ER-Phagosome pathway
ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.
Respiratory electron transport
The citric acid (TCA) cycle and respiratory electron transport

Figure A.6.: Network of enriched pathways for the tumor data. Enrichment computed based on the dynamical importance.

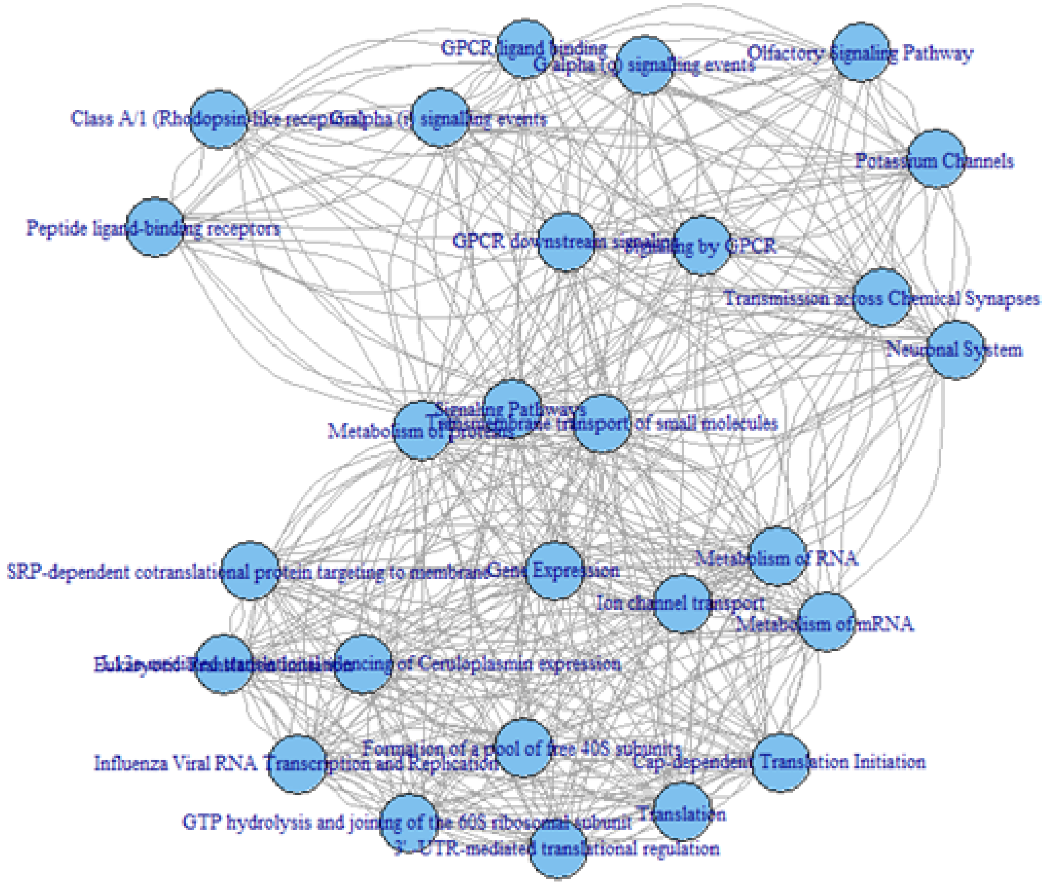

GPCR ligand binding
G alpha (q) signalling events
Olfactory Signaling Pathway
Class A/1 (Rhodopsin-like recep
Galpha (i) signalling events
Potassium Channels
Peptide ligand-binding receptors
GPCR downstream signaling
Signaling by GPCR
Transmission across Chemical Synapses
Neuronal System
Signaling Pathways
Metabolism of proteins
Transmembrane transport of small molecules
SRP-dependent cotranslational protein targeting to membran
Gene Expression
Metabolism of RNA
Ion channel transport
Metabolism of mRNA
Eukaryotic Translation Elongation
L13a-mediated translational silencing of Ceruloplasmin expression
Formation of a pool of free 40S subunits
Influenza Viral RNA Transcription and Replication
Cap-dependent Translation Initiation
GTP hydrolysis and joining of the 60S ribosomal subunit
Translation
3'-UTR-mediated translational regulation

Figure A.7.: Network of enriched pathways for the stroma data. Enrichment computed based on the dynamical importance.

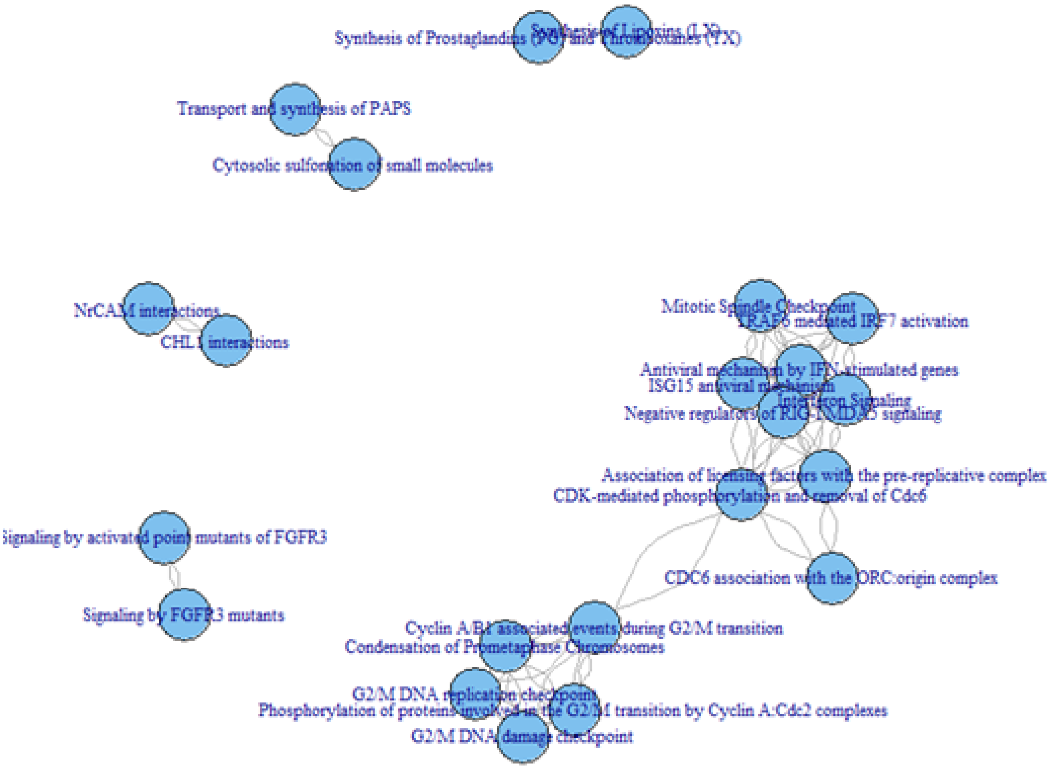

Synthesis of Lipoxins (LX)
Synthesis of Prostaglandins (PG) and Thromboxanes (TX)
Transport and synthesis of PAPS
Cytosolic sulfonation of small molecules
NrCAM interactions
CHL1 interactions
Signaling by activated point mutants of FGFR3
Signaling by FGFR3 mutants
Mitotic Spindle Checkpoint
TRAF6 mediated IRF7 activation
Antiviral mechanism by IFN-stimulated genes
ISG15 antiviral mechanism
Interferon Signaling
Negative regulators of RIG-I/MDA5 signaling
Association of licensing factors with the pre-replicative complex
CDK-mediated phosphorylation and removal of Cdc6
CDC6 association with the ORC:origin complex
Cyclin A/B1 associated events during G2/M transition
Condensation of Prometaphase Chromosomes
G2/M DNA replication checkpoint
Phosphorylation of proteins involved in the G2/M transition by Cyclin A:Cdc2 complexes
G2/M DNA damage checkpoint

Figure A.8.: Network of enriched pathways for the tumor data. Enrichment computed by multiplication.

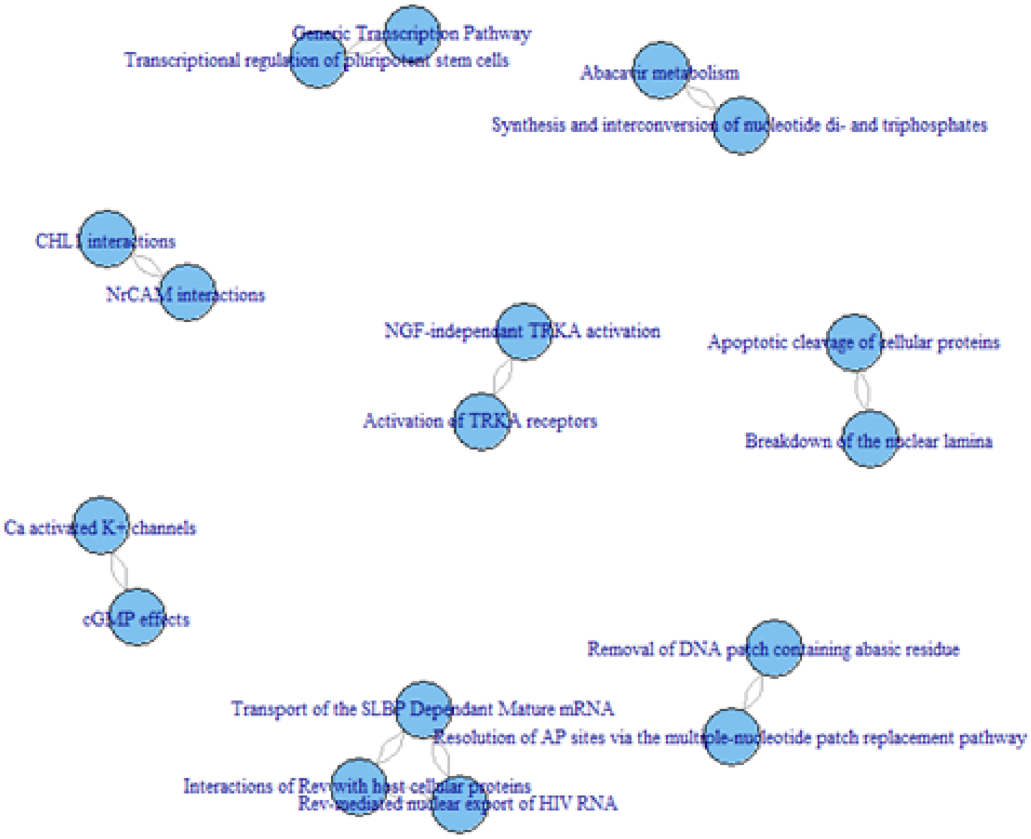

Generic Transcription Pathway
Transcriptional regulation of pluripotent stem cells
Abacavir metabolism
Synthesis and interconversion of nucleotide di- and triphosphates
CHL1 interactions
NrCAM interactions
NGF-independant TRKA activation
Apoptotic cleavage of cellular proteins
Activation of TRKA receptors
Breakdown of the nuclear lamina
Ca activated K+ channels
cGMP effects
Removal of DNA patch containing abasic residue
Transport of the SLBP Dependant Mature mRNA
Resolution of AP sites via the multiple-nucleotide patch replacement pathway
Interactions of Rev with host cellular proteins
Rev-mediated nuclear export of HIV RNA

Figure A.9.: Network of enriched pathways for the stroma data. Enrichment computed by multiplication

Bibliography

[1] Adams, M. D., Celniker, S. E., Holt, R. A., et al. (2000). The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–2195.

[2] Adams, M. D., Kelley, J. M., Gocayne, J. D., et al. (1991). Complementary dna sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656.

[3] Aebersold, R., Bader, G. D., Edwards, A. M., et al. (2013). The Biology/Disease-driven Human Proteome Project (B/D-HPP): Enabling Protein Research for the Life Sciences Community. *Journal of proteome research*, 12:23–7.

[4] Aebersold, R. H., Leavittt, J., Saavedra, R. A., Hood, L. E., and Kent, S. B. H. (1987). 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*, 84(October):6970–6974.

[5] Aebersold, R. H., Pipes, G., Hood, L. E., and Kent, S. B. (1988). N-terminal and internal sequence determination of microgram amounts of proteins separated by isoelectric focusing in immobilized pH gradients. *Electrophoresis*, 9:520–30.

[6] Aebersold, R. H., Teplow, B., Hood, L. E., and Kent, B. H. (1986). Electroblotting onto Activated Glass. *The Journal of Biological Chemistry*, 261(9):4229–38.

[7] Alter, O., Brown, P. O., and Botstein, D. (2003). Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci USA*, 100:3351–3356.

[8] Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977). 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*, 74(12):5350–5354.

[9] Andersen, J. S. and Mann, M. (2000). Functional genomics by mass spectrometry. *FEBS Letters*, 480:25–31.

[10] Anderson, N. G. and Anderson, N. L. (2000). Twenty years of two-dimensional electrophoresis: Past, present and future. *Electophoresis*, 17:443–453.

[11] Au, C. E., Bell, A. W., Gilchrist, A., et al. (2007). Organellar proteomics to create the cell map. *Current Opinion in Cell Biology*, 19:376–85.

[12] Author (2008). *Oxford English Dictionary*. Pblisher.

[13] Bahn, J. H., Lee, J.-H., Li, G., et al. (2012). Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Research*, 22:142–50.

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

[15] Benoit, G. R., Tong, J.-H., Balajthy, Z., and Lanotte, M. (2001). Exploring (novel) gene expression during retinoid-induced maturation and cell death of acute promyelocytic leukemia. In *Seminars in hematology*, volume 38, pages 71–85. Elsevier.

[16] Boguski, M. S., Tolstoshev, C. M., Bassett Jr, D. E., et al. (1994). Gene discovery in dbest. *Science*, 265(5181):1993–1994.

[17] Brenner, S., Johnson, M., Bridgham, J., et al. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology*, 18(June):630–4.

[18] Brown, P. O. and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature genetics*, 21(1 Suppl):33–7.

[19] Cagney, G., Park, S., Chung, C., et al. (2005). Human tissue profiling with multidimensional protein identification technology. *J Proteome Res*, 4:1757–1767.

[20] Ceciliani, F., Eckersall, D., Burchmore, R., and Lecchi, C. (2014). Proteomics in veterinary medicine: applications and trends in disease pathogenesis and diagnostics. *Veterinary pathology*, 51(2):351–62.

[21] Chen, Y.-r., Juan, H.-f., Huang, H.-c., et al. (2006). Quantitative proteomic and genomic profiling reveals metastasis-related protein expressio patterns in gastric cancer cells research articles. *J Proteome Res*, 5:2727–2742.

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

[23] Cloonan, N., Forrest, A. R. R., Kolle, G., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods*, 5(7):613–9.

[24] Collins, S. R., Kemmeren, P., Zhao, X.-C., et al. (2007). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Molecular & cellular proteomics : MCP*, 6:439–50.

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

[26] Corbin, R. W., Paliy, O., Yang, F., et al. (2003). Toward a protein profile of Escherichia coli: comparison to its transcription profile. *Proc Natl Acad Sci USA*, 100(16):9232–9237.

[27] Corthals, G. L., Wasinger, V. C., Hochstrasser, D. F., and Sanchez, J.-C. (2010). Review The dynamic range of protein expression : A challenge for proteomic research Proteomics and 2-DE. *Electrophoresis*, 21:1104–15.

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

[29] Crawford, J. E., Guelbeogo, W. M., Sanou, A., et al. (2010). De novo Transcriptome Sequencing in Anopheles funestus Using Illumina RNA-seq Technology. *PloS One*, 5(12):e14202.

[30] Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, 227:561–3.

[31] Croft, D., O'Kelly, G.and Wu, G., Haw, R., et al. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*, 39:691–697.

[32] Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, Complex Systems:1695.

[33] Culhane, A., Thioulouse, J., Perriere, G., and Higgins, D. (2005). MADE4: an R package for multivariate analysis of gene expression data. *Bioinforma Oxf Engl*, 21:2789–2790.

[34] Debouck, C. and Goodfellow, P. N. (1999). DNA microarrays in drug discovery and development. *Nature Genetics*, 21(1 Suppl):48–50.

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

[36] Dong, H., Luo, L., Hong, S., et al. (2010). Integrated analysis of mutations, miRNA and mRNA expression in glioblastoma. *BMC Syst Biol*, 4:163.

[37] Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nature Genetics*, 21(1 Suppl):10–4.

[38] Edman, P. (1950). Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chemica Scandinavica*, 4:283–93.

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

[40] Finotello, F. and Di Camillo, B. (2014). Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in Functional Genomics*, page elu035.

[41] Fleischmann, R. D., Adams, M. D., White, O., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (New York, N.Y.)*, 269(5223):496–512.

[42] Franklin, R. E. and Gosling, R. G. (1953). Molecular configuration in sodium thymonucleate. *Nature*, 171:740–1.

[43] Freeman, L. C. (1978). Centrality in Social Networks Conceptual Clarification. *Social Networks*, 1:215–239.

[44] Gerhard, D. S., Wagner, L., Feingold, E. A., et al. (2004). The status, quality, and expansion of the nih full-length cdna project: the mammalian gene collection (mgc). *Genome research*, 14(10B):2121–2127.

[45] Gerstein, M. B., Kundaje, A., Hariharan, M., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, 489:91–100.

[46] Google (2014). Search: "integrative analysis". http://scholar.google.at/. [Online; accessed 15-September-2014].

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

[48] Graves, P. R. and Haystead, T. a. J. (2002). Molecular Biologist's Guide to Proteomics. *Microbiology and Molecular Biology Reviews*, 66(1):39–63.

[49] Griffin, T. J. (2002). Complementary Profiling of Gene Expression at the Transcriptome and Proteome Levels in Saccharomyces cerevisiae. *Mol Cell Proteomics*, 1:323–333.

[50] Grouse, L. H., Munson, P. J., and Nelson, P. S. (2001). Sequence databases and microarrays as tools for identifying prostate cancer biomarkers. *Urology*, 57(4):154–159.

[51] Gupta, N., Benhamida, J., Bhargava, V., et al. (2008). Comparative proteogenomics: Combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Research*, 18:1133–42.

[52] Gusenleitner, D., Howe, E., Bentink, S., Quackenbush, J., and Culhane, A. (2012). iBBiG: Iterative Binary Bi-clustering of Gene Sets. *Bioinforma Oxf Engl*, 28:2484–2492.

[53] Hacia, J. G. (1999). Resequencing and mutational analysis using oligonucleotide microarrays. *Nature Genetics*, 21(1 Suppl):42–47.

[54] Hahne, H., Mäder, U., Otto, A., et al. (2010). A comprehensive proteomics and transcriptomics analysis of Bacillus subtilis salt stress adaptation. *J Bacteriol*, 192:870–882.

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

[56] Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nature reviews. Genetics*, 11:476–86.

[57] Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models-a review. *BioSystems*, 96:86–103.

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

[59] History, N. (2014). Deciphering the genetic code. http://history.nih.gov/exhibits/nirenberg/glossary.htm.

[60] Hong, S., Chen, X., Jin, L., and Xiong, M. (2013). Canonical correlation analysis for rna-seq co-expression networks. *Nucleic acids research*, 41(8):e95–e95.

[61] Hwang, D., Rust, A. G., Ramsey, S., et al. (2005a). A data integration methodology for systems biology. *Proc Natl Acad Sci USA*, 102:17296–17301.

[62] Hwang, D., Smith, J. J., Leslie, D. M., et al. (2005b). A data integration methodology for systems biology: Experimental verification. *Proc Natl Acad Sci USA*, 102:17302–17307.

[63] Joyce, A. R. and Palsson, B. O. (2006). The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol*, 7:198–210.

[64] Khatri, P., Sirota, M., and Butte, A. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput Biol*, 8:e1002375.

[65] Kim, M.-S., Pinto, S. M., Getnet, D., et al. (2014). A draft map of the human proteome. *Nature*, 509(7502):575–81.

[66] Kislinger, T., Cox, B., Kannan, A., et al. (2006). Global Survey of Organ and Organelle Protein Expression in Mouse: Combined Proteomic and Transcriptomic Profiling. *Cell*, 125:173–186.

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

[68] Knippers, R. (2006). *Molekulare genetik*. Georg Thieme Verlag.

[69] Kockmann, T., Gerstung, M., Schlumpf, T., et al. (2013). The BET protein FSH functionally interacts with ASH1 to orchestrate global gene activity in Drosophila. *Genome Biol*, 14:R18.

[70] Kodzius, R., Kojima, M., Nishiyori, H., et al. (2006). CAGE: cap analysis of gene expression. *Nature methods*, 3(3):211–22.

[71] Lakhan, S. E. (2006). Schizophrenia proteomics: biomarkers on the path to laboratory medicine? *Diagnostic Pathology*, 1:11.

[72] Lander, E. S., Linton, L. M., Birren, B., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[73] Lazar, C., Taminau, J., Meganck, S., et al. (2013). Geneshift: A nonparametric approach for integrating microarray gene expression data based on the inner product as a distance measure between the distributions of genes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 10(2):383–392.

[74] Le Roch, K. G., Johnson, J. R., Florens, L., et al. (2004). Global analysis of transcript and protein levels across the plasmodium falciparum life cycle. *Genome Res*, 14:2308–2318.

[75] Lennon, G. G. and Lehrach, H. (1991). Hybridization analyses of arrayed cdna libraries. *Trends in genetics*, 7(10):314–317.

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

[77] Liang, P. and Pardee, A. B. (1992). Differential display of eukaryotic messenger rna by means of the polymerase chain reaction. *Science*, 257(5072):967–971.

[78] Mardis, E. R. (2013). Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, 6:287–303.

[79] Margulies, M., Egholm, M., Altman, W. E., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80.

[80] Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–17.

[81] Mendel, G. (1866). Versuche über pflanzen-hybriden. In Verein, N., editor, *Verhandlungen des Naturforschenden Vereines in Brünn*. Naturforschender Verein, Brünn.

[82] Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature reviews. Genetics*, 11:31–46.

[83] Meyer, C. D. (2001). *Perron-Frobenius Theory*, chapter Matrix Analysis and Applied Linear Algebra. SIAM.

[84] Miescher, F. (1871). Über die chemische Zusammensetzung der Eiterzellen. *Hoppe-Seyers Medizinisch-Chemische Untersuchungen*, 4(4):441–60.

[85] Mo, Q., Wang, S., Seshan, V. E., et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci USA*, 110:4245–4250.

[86] Mok, S. C., Bonome, T., Vathipadiekal, V., et al. (2009). A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: microfibril-associated glycoprotein 2. *Cancer cell*, 16(6):521–532.

[87] Mootha, V. K., Bunkenborg, J., Olsen, J. V., et al. (2003). Integrated Analysis of Protein Composition, Tissue Diversity, and Gene Regulation in Mouse Mitochondria. *Cell*, 115:629–640.

[88] Mortazavi, A., Williams, B. a., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–8.

[89] National Institutes of Health: National Human Genome Research Institute (2014). Talking glossary of genetic terms. http://www.genome.gov/glossary/.

[90] Network, T. C. G. A. R. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474:609–615.

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

[92] Nirenberg, M. and Matthaei, H. (1961). The Dependendence of Cell-free Protein Synthesis in E. Coli upon Naturally Occurring or synthetic polyriconucleotides. *PNAS*, 47:1588–1602.

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

[94] Okoniewski, M. J. and Miller, C. J. (2006). Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 7:276–90.

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

[96] Paik, Y.-K., Jeong, S.-K., Omenn, G. S., et al. (2012). The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. *Nature biotechnology*, 30(3):221–3.

[97] Pauling, L. and Corby, R. B. (1953). Structure of the Nucleic Acids. *Nature*, 171:364.

[98] Peng, Z., Cheng, Y., Tan, B. C.-M., et al. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature biotechnology*, 30(3):253–60.

[99] Perco, P., Mühlberger, I., Mayer, G., et al. (2010). Linking transcriptomics and proteomic data on the level of protein interaction networks. *BMC Syst Biol*, 31:1780–1789.

[100] Pettersson, E., Lundeberg, J., and Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics*, 93:105–11.

[101] Piruzian, E., Bruskin, S., Ishkin, A., et al. (2010). Integrated network analysis of transcriptomic and proteomic data in psoriasis. *BMC Syst Biol*, 4:41–53.

[102] Prifti, E., Zucker, J., Clement, K., and Henegar, C. (2008). FunNet: an integrative tool for exploring transcriptional interactions. *Bioinforma Oxf Engl*, 24:2636–2638.

[103] Reactome (2014). Reactome.

[104] Reis-Filho, J. S. (2009). Next-generation sequencing. *Breast Cancer Research*, 11(Suppl 3):S12.

[105] Restrepo, J., Ott, E., and Hunt, B. (2006). Characterizing the dynamical importance of network nodes and links. *Phys Rev Lett*, 97:094102.

[106] Roberts, A., Pimentel, H., Trapnell, C., and Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics (Oxford, England)*, 27(17):2325–9.

[107] Ross, P. L., Huang, Y. N., Marchese, J. N., et al. (2004). Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. *Molecular & cellular proteomics : MCP*, 3:1154–69.

[108] Rozowsky, J., Abyzov, A., Wang, J., et al. (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology*, 7:522.

[109] Saiki, R., Scharf, S., Fred, F., et al. (1985). enzymatic Amplification of β-Globulin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. *Science*, 230:1350–4.

[110] Sales, G., Calura, E., and Romualdi, C. (2014a). *graphite: GRAPH Interaction from pathway Topological Environment*. R package version 1.10.1.

[111] Sales, G., Calura, E., and Romualdi, C. (2014b). Graphite Rules: How graphite solves network complexity. .

[112] Sanger, F., Air, G. M., Barrell, B. G., et al. (1977a). Nucleotide Sequence of Bacteriophage Ψ X174 DNA. *Nature*, 265(5596):687–95.

[113] Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F., and Petersen, G. B. (1982). Nucleotide Sequence of Bacteriophage λ DNA. *J. Mol. Biol.*, 162(4):729–73.

[114] Sanger, F., S, N., and R, C. A. (1977b). DNA sequencing with chain-terminating inhibitors. *PNAS*, 74:5364–7.

[115] Sass, S., Buettner, F., Mueller, N., and Theis, F. (2013). A modular framework for gene set analysis integrating multilevel omics data. *Nucleic Acids Res.* in press.

[116] Scheele, G. A. (1975). Two-Dimensional Gel Analysis of Soluble Proteins. Chracterization of Guinea Pig Exocrine Pancreatic Proteins. *The Journal of Biological Chemistry*, 250(14):5375–85.

[117] Schröder, M., Gusenleitner, D., Quackenbush, J., Culhane, A., and Haibe-Kains, B. (2013). RamiGO: an R/Bioconductor package providing an AmiGO visualize interface. *Bioinforma Oxf Engl*, 29:666–668.

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

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

[120] Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–45.

[121] Shiraki, T., Kondo, S., Katayama, S., et al. (2003). 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*, 100(26):15776–81.

[122] Soini, H. and Musser, J. M. (2001). Molecular Diagnosis of Mycobacteria. *Clinical Chemistry*, 47(5):809–814.

[123] Subramanian, A., Tamayo, P., Mootha, V., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 102:15545–15550.

[124] Takemasa, I., Kittaka, N., Hitora, T., et al. (2012). Potential biological insights revealed by an integrated assessment of proteomic and transcriptomic data in human colorectal cancer. *Int J Oncol*, 40:551–559.

[125] Tomescu, O., Mattanovich, D., and Thallinger, G. (2014). Integrative omics analysis. A study based on *P. falciparum* mRNA and protein data. *BMC System Biology*. in press.

[126] Triche, T. J., Schofield, D., and Buckley, J. (2000). Dna microarrays in pediatric cancer. *Cancer journal (Sudbury, Mass.)*, 7(1):2–15.

[127] Vaidyanathan, G. (2012). Redefining clinical trials: the age of personalized medicine. *Cell*, 148(6):1079–80.

[128] Van Verk, M. C., Hickman, R., Pieterse, C. M. J., and Van Wees, S. C. M. (2013). RNA-Seq: revelation of the messengers. *Trends in Plant Science*, 18(4):175–9.

[129] Vaske, C., Benz, S., Sanborn, J., et al. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinforma Oxf Engl*, 26:237–245.

[130] Velculescu, V. E., Zhang, L., Vogelstein, B., Kinzler, K. W., et al. (1995). Serial analysis of gene expression. *Science*, 270(5235):484–487.

[131] Velculescu, V. E., Zhang, L., Zhou, W., et al. (1997). Characterization of the Yeast Transcriptome. *Cell*, 88:243–251.

[132] Venter, J. C., Adams, M. D., Myers, E. W., et al. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–51.

[133] Vera, J. C., Wheat, C. W., Fescemyer, H. W., et al. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, 17:1636–47.

[134] Verhoef, S., Ballerstedt, H., Volkers, R. J. M., de Winde, J. H., and Ruijssenaars, H. J. (2010). Comparative transcriptomics and proteomics of p-hydroxybenzoate producing Pseudomonas putida S12: novel responses and implications for strain improvement. *Appl Microbiol Biotechnol*, 87:679–690.

[135] Wang, K.-S. and Liu, X. (2013). Integrative Analysis of Genome-wide Expression and Methylation Data. *J Biom Biostat*, 4:4–6.

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

[137] Washburn, M. P., Koller, A., Oshiro, G., et al. (2003). Protein pathway and complex clustering of correlated mrna and protein expression analyses in saccharomyces cerevisiae. *Proc Natl Acad Sci USA*, 100:3107–3112.

[138] Wasinger, V. C., Cordwell, S. J., Cerpa-poljak, A., et al. (1995). Progress with gene-product mapping of the Mollicutes : Mycoplasma genitalium. *Electophoresis*, 16:1090–1094.

[139] Watson, J. D. and Crick, F. H. (1953). A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–8.

[140] Wikipedia (2014). Omics — wikipedia, the free encyclopedia. http://en.wikipedia.org. [Online; accessed 15-September-2014].

[141] Wilhelm, M., Schlegl, J., Hahne, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, 509(7502):582–7.

[142] Wilkins, M. R., Sanchez, J.-C., Gooley, A. a., et al. (1996). Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It. *Biotechnology and Genetic Engineering Reviews*, 13(1):19–50.

[143] Yates, J. R., Gilchrist, A., Howell, K. E., and Bergeron, J. J. M. (2005). Proteomics of organelles and large cellular structures. *Nature reviews. Molecular cell biology*, 6:702–14.

[144] Zhang, W., Li, F., and Nie, L. (2010). Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology*, 156:287–301.