TESIS DOCTORAL

BIOINFORMATICS TO INTEGRATE PROTEIN AND GENE INFORMATION IN A RELATIONAL CONTEXT, APPLICATION TO HUMAN PROTEOMIC AND TRANSCRIPTOMIC DATA

CONRAD FRIEDRICH DROSTE

DIRECTOR

DR. JAVIER DE LAS RIVAS SANZ

SALAMANCA, JULIO DE 2017

El **Dr. Javier De Las Rivas Sanz**, con D.N.I. 15949000H, Investigador Científico del Consejo Superior de Investigaciones Científicas (CSIC), director del grupo de Bioinformática y Genómica Funcional en el Instituto de Biología Molecular y Celular del Cáncer (CiC-IBMCC), y profesor del Programa de Doctorado y del Máster de Biología y Clínica del Cáncer de dicho Instituto y la Universidad de Salamanca (USAL).

CERTIFICA

Que ha dirigido esta Tesis Doctoral titulada "BIOINFORMATICS TO INTEGRATE PROTEIN AND GENE INFORMATION IN A RELATIONAL CONTEXT, APPLICATION TO HUMAN PROTEOMIC AND TRANSCRIPTOMIC DATA" realizada por **D. Conrad Friedrich Droste**, alumno del Programa de Doctorado de 2012/2013 de la Universidad de Salamanca.

y AUTORIZA

La presentación de la misma, considerando que reúne las condiciones de originalidad y contenidos requeridos para optar al grado de Doctor por la Universidad de Salamanca.

En Salamanca, a 7 de julio de 2017

Dr. Javier De Las Rivas Sanz

To my parents, Heike and Friedel and my loved ones.

I have to apologize to Pigena for all the time I missed to enjoy with her.

ACKNOWLEDGEMENTS & APPRECIATIONS

I am deeply grateful to my Ph.D. director Dr. Javier De Las Rivas for the opportunity to realize this work in his research group. His guidance, encouragement and support during these time were always appreciated and needed.

I owe a very important debt to Dr. Manuel Fuentes and Paula Díez which allowed me to support their proteomic research. Without their tremendous efforts to generate the proteomic data and interpret the results a big part of this Ph.D. project would have not been achieved.

My deepest heartfelt appreciations go to all the current and previous members of the Bioinformatics and Functional Genomics group of the Cancer Research Center Salamanca (CIC-IBMCC). I only can say thank you, for the countless support and encouragement during this Ph.D. time.

I also thank all members of the Cancer Research Center from Salamanca, without this precious scientific environment this research could not have been done.

I would also like to express my gratitude to the Junta de Castilla y León for their financial support during my predoctoral time.

TABLE OF CONTENTS

1 INTRODUCTION

The basic unit of a complex organism like Homo sapiens is the cell. Cells are complex biological systems. Cellular proliferation, differentiation and interactions between other cells and the environment require the production, assembly, operation and regulation of many thousands of components. The investigation how cells work on the molecular level with such high precision is one of the most challenging questions of modern molecular biology **(Zhu et al. 2007)**.

Thanks to technical advantages biological researchers have new opportunities to gain insights into these complex systems on different cellular levels: (1) Lower costs and reduced time to sequence the genome of cells and organisms show the information the DNA provides, like protein coding regions, open reading frames, regulatory elements, operons, mutations, deletions, duplications, etc. (2) Development of new techniques like RNA-Seq and technical improvements and cost reductions of highly frequented techniques like microarrays give researchers the opportunity to investigate the transcriptomic of biological systems. The transcriptomic shows, which genes are active and how active they are in a specific biological context, like tissues, development or disease states. (3) The next level is the proteomic of a cell, which is a rapidly developing area not only on the quality level - which describes inter alia the specific biological condition of proteins in a cell or cellular compartment -, but also on the quantity level – which describes inter alia how the expression level of proteins changes in a specific biological condition.

Having access to these large OMIC datasets is only the first step to gain biological knowledge. The second step is to interpret these data in a reasonable and reproducible way **(Alyass et al. 2015)** .The information has to be processed, analysed, interpreted, stored and shared with the biological community in order to make progress in science. Bioinformatics is the discipline which helps to face these challenges. The **Figure 1** shows the grow of entries in mayor sequence (GenBank, EMBL, KEGG GENES), protein (PIR, PRF and SWISS-PROT), 3D structure databases (PDB) and KO (KEGG ORTHOLOGY).

This project focuses on processing, analysis, integration and interpretation of proteomic and transcriptomic datasets – called Proteogenomics. Section 1.1 explains this term and its state of the art in detail. It also describes the development of a bioinformatic tool to integrate the datasets with biological network datasets – the Path2enet tool. Therefore section 1.2

describes the state of the art of biological network analysis and key-concepts of this bioinformatic method.

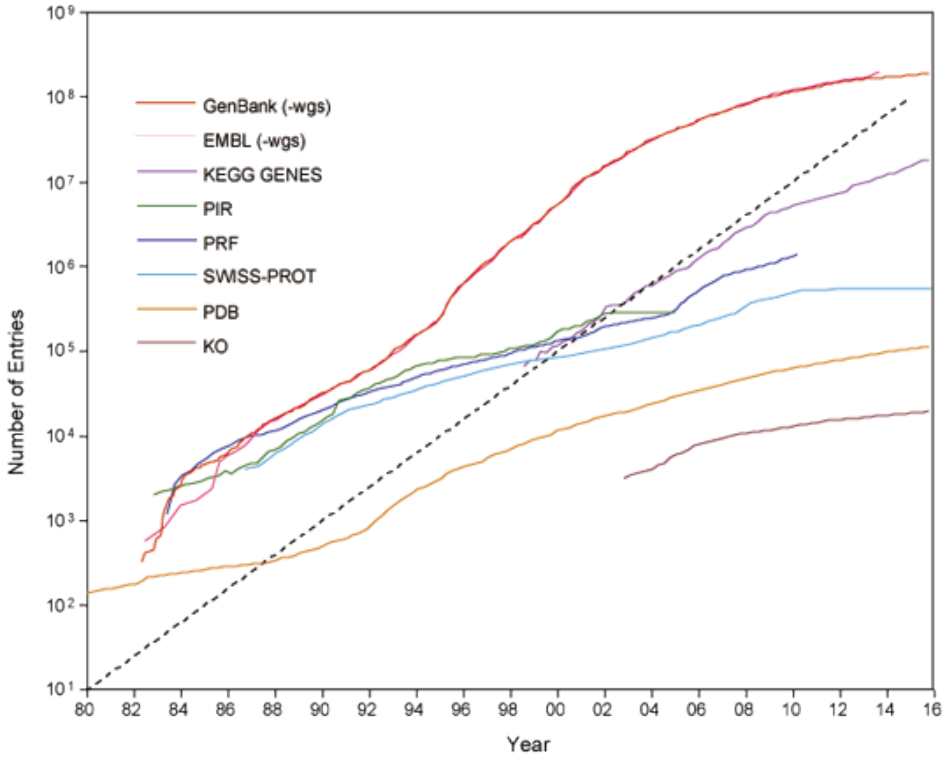


Figure 1: Growth of sequence and structure databse in the last 36 years. (1)

For the project datasets of B- and T-lymphocytes and leukaemia patients or cell lines were primarily used. So the excursus in section 1.3 gives a small overview of the immune system, the role of B- and T-lymphocytes and leukaemia.

The basic unit of a complex organism like Homo sapiens is the cell. Cells are complex biological systems. Cellular proliferation, differentiation and interactions between other cells and the environment require the production, assembly, operation and regulation of many thousands of components. The investigation how cells work on the molecular level with such high precision is one of the most challenging questions of modern molecular biology **(Zhu et al. 2007).**

Experiments regarding the interaction of cells have produced large amounts of various data, which are stored in many different databases. One objective of bioinformatics is the development of new tools for the analysis of these databases so that new scientific results can be generated out of the already existing data.

One tool for the analysis of data of several databases is the meta-analysis which combines different datasets to get new results. Important biological datasets are the interactome of different species, especially the human interactome. The interactome is the collective term for the whole set of molecular interactions in a species. The human interactome is based on *"around 25,000 protein-coding genes, around 1,000 metabolites and an undefined number of distinct proteins and functional RNA molecules"* **(Barabasi et al. 2011)**. In total, this sums up to at least more than 100,000 cellular components in the human interactome. These components are related to each other in different ways. The number of relations exceeds substantially the

number of components. The complexity of the human interactome makes its analysis difficult.

The objective of the analysis of the human interactome is to support researchers to understand the molecular mechanism of a cell. Understanding these mechanisms is imperative to identify key-players in the interactome. The detection of key-players in the human interactome is important because defects of key-players are often the reason of major diseases like cancer. Therefore, the key-players of an interactome are perfect drug targets. If drugs are able to re-transfer defected key-players into their initial non-defected state diseases could be cured.

Scientists of different disciplines, like cell biology, genetics, biochemistry, biophysics and bioinformatics put substantial efforts in investigating the key-players in the human interactome. Adequate bioinformatic tools to identify these key-players can support biological researchers substantially in analysing the human interactome.

The human interactome is very complex. Investigators normally reduce its complexity by focusing their investigations on one part of it. Proteins and their interactions with receptors, metabolites, hormones, transcription factors and especially other proteins are an important part in the human interactome.

Many researches pay particular attention to the investigation of protein-protein-interactions. They use modern experiments like the two-hybrid system **(Suter et al. 2008)** or Mass Spectrometry. These techniques are used to verify: **(i)** older, already published protein relations; **(ii)** protein relations predicted from model organism; and **(iii)** new protein-protein-interactions (**Rual et al. 2005)**. Protein-protein-interaction networks are interesting for further studies, because they represent the largest and most diverse datasets available **(Zhu et al. 2007).**

The importance of proteins and their interactions in organism explains why such huge data volumes and databases of proteins and protein-protein-interactions are available.

Bioinformatics can use this data for the creation of protein-protein-interaction-networks. The analysis of these networks with statistic methods enables to identify their key-players. But to find these key-player it is important to put the interactomes in a biological context. The improvements in transcriptomics and proteomics increased the number of biological datasets substantially.

These datasets are important to decide, which protein or gene is expressed in a specific tissue or development stage. The integration of network-specific information with the omic datasets is one of the major objectives of the bioinformatic tool "*Path2enet*" which was developed fow this project. It allows inter alia to select nodes which are expressed in biological context,

1.1 Status of proteogenomics methods and analysis

Proteomics is the comprehensive, integrative study of proteins and their biological functions. The goal of proteomics is often to produce a complete and quantitative map of the proteome of a species, including defining protein cellular localization, reconstructing protein interaction networks and complexes, and delineating signaling pathways and regulatory post-translational protein modifications **(Alexey and Nesvizhskii 2014)**. Bearing this in mind, the -omics integration allows the complete characterization of a cell. In fact, an increasing number of studies have been developed in this field in the last years. For instance, Yizhak and collaborators have introduced a new method for the integration of proteomic and metabolomics data with genomics called IOMA (Integrative Omics- Metabolic Analysis) in order to predict metabolic flux distributions **(Yizhak et al. 2010)**. Other researcher combine transciptomic and proteomic data in order to understand regulatory behavior of cells **(Haider and Pal 2013)**. On the transcriptomic level the RNA-seq method is increasing its influence, even so there are still challenges in processing the big amout of generated data, as well to target complex transcriptome to characterize rare RNA isoforms **(Wang et al. 2009).**

On proteomic level mass spectrometry appears as the promising technology for simultaneous identification of thousands of proteins in one single assay. It measures gene products at the translational level and allows a great coverage of the proteome **(Díez, Droste et al. 2015)**. However, several bottlenecks are also presented in this methodology and are quite diverse for sample preparation (eg. high salt concentrations, detergents, subcellular fractionaton…), protein digestion (i.e. highly dependent of specific proteases or combinations of them), getting enough accuracy and precision of measurements, and processing results with high-efficiency database search engines **(Feist and Hummon 2015; Aebersold and Mann 2003)**.

Recently, it has been developed a novel proteomics approach based on a technique which couples size exclusion chromatography (SEC) with microsphere-based affinity proteomics (MAP) 13–15. The **Figure 2** shows the benefits the proteomic approach can have for biological researcher.

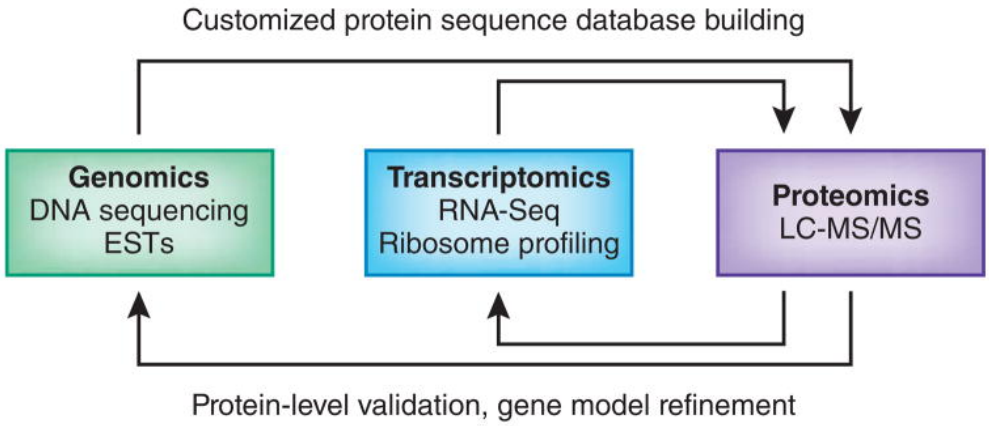


Figure 2: T*he proteogenomics concept combines various source of genomic (DNA sequencing, expressed sequence tags (ESTs)), transcriptomics (RNA-seq, microarrays) and proteomics (LC-MS/MS). Each data type can help to improve the quality and analysis of the other datasets* ***(Alexey and Nexvizhsii 2014)***

In this thesis we present a multi-dimensional characterization of the protein profile of a Burkitt's lymphoma B-cell line based on the analysis of evidence at the protein level by affinity proteomics (SEC-MAP) and MS/MS assays and its correlation with the number of gene

products identified by RNA-Seq and microarray data)

1.1.1 Gene Expression datasets

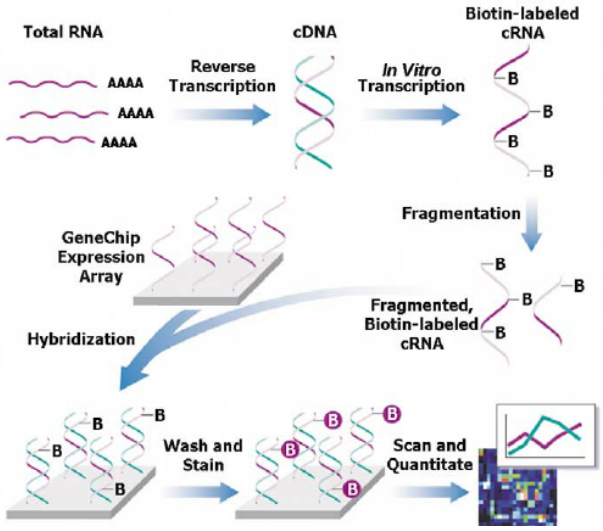
In the presented thesis I used three sources of transcriptomic expression datasets. In the following paragraph I will provide a short description of these data sources: 1.1.1.1 Expressed Sequence Tags [EST], 1.1.1.2 Microarray and 1.1.1.3 RNA-seq.

1.1.1.1 ESTs

Expressed Seqence Tags [ESTs] are the oldest sources of expression based transcriptomic data used in the thesis. “ESTs are short (usually approximately 300-500 base pairs), single-pass sequence reads from cDNA. They represent the genes expressed in a given tissue and/or at a given developmental stage (2).” They were used already in the data analysis of gene expression by tissue and delopmental stage in 1994. (Fields C., 1994). In this work we use the database as source for ESTs. Unigene is a database of the National Center for Biotechnology Information [NCBI] , a division of the National Library of Medicine [NLM] at the National Institutes of Health [NIH] of the United States of America . This database has high quality and will only take up non repetitive Expressed Sequence Tags [ESTs] which have at least 100 base pairs. Unigene has its focus on transcripts of protein-coding genes of the nuclear genome.

1.1.1.2 Microarray

High-density oligonucleotide microarrays, also known as expression microarrays, is one of high-throughput genomic techniques that has achieved a greater impact in biomedical research in the last two decades. Among the most popular platforms is the technology developed by *Affymetrix* (https://www.affymetrix.com/site/mainPage.affx)



This platform is based on hybridization of RNA from samples in the study with 25 nucleotides long complementary strands called as probes. Each probe generally represents only one gene and allows to detect its individual expression.

Probes location in *Affymetrix* 3’ ITV microarrays are restricted to 3’ regions because reverse transcription and amplification depends on poly(A) tails. As consequence, expression signal from this platform allows the measurement of gene expression, but it is not possible to obtain information about alternative splicing events.

This technology contributed tremendously to biomedical research in recent years. There are thousands of published studies describing relevant biological insight from expression microarrays in part due to the relative low cost and high reproducibility of the platform. Nowadays, 3' ITV technology is deprecated given the new generation of microarrays, e.g. based on random priming, and sequencing techniques, mainly Next Generation Sequencing. However, 3' ITV expression left a valuable legacy, a large number of publicly available datasets that are continually revisited **(Rustici et al. 2013)**. It is remarkable for example the number of large cancer datasets .

1.1.1.3 RNA-Seq

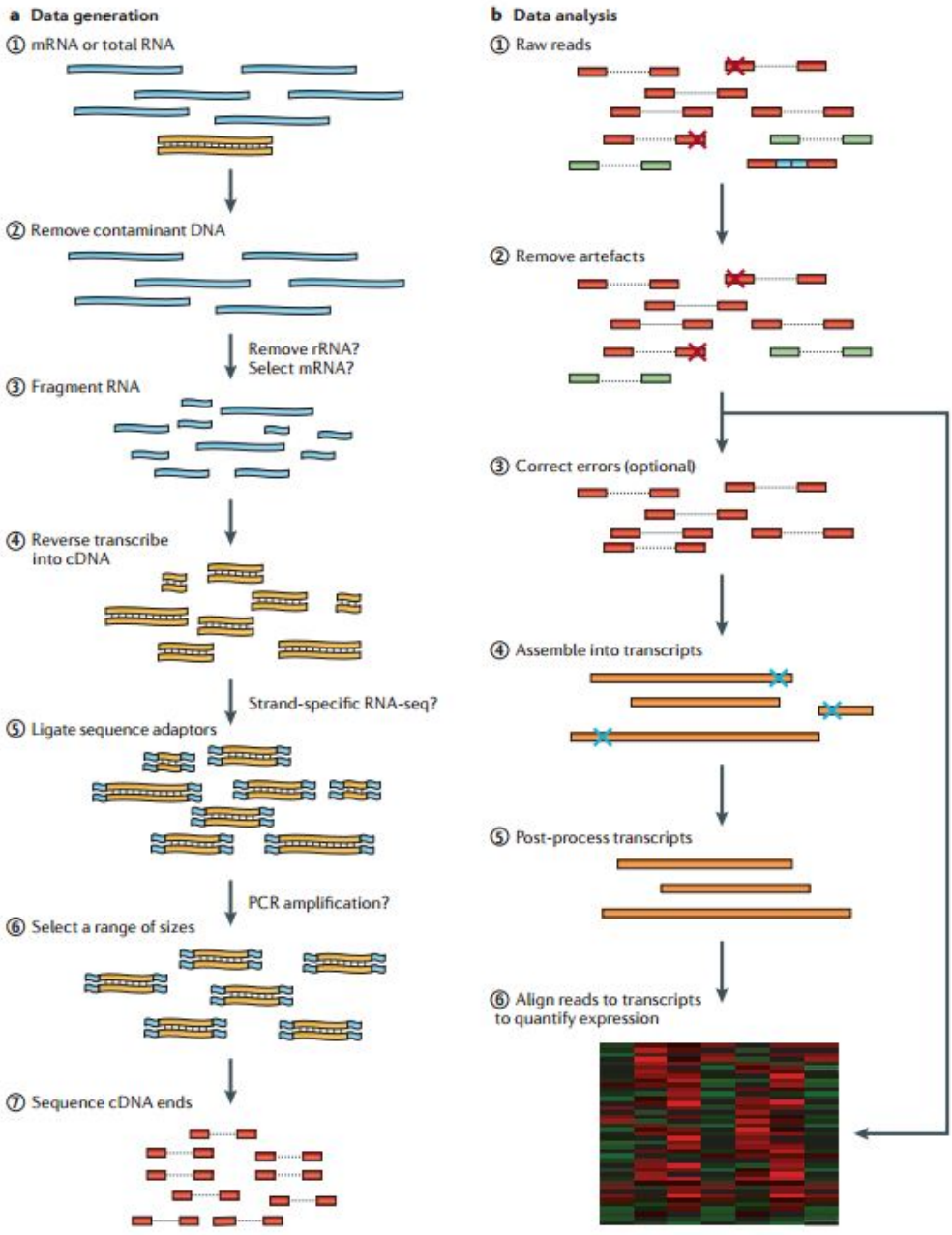


Figure 4 *Workflow of the RNA-Seq data collecting (a) and (b) computational prosessing of the datasets.* ***(Martin and Wang 2011)***

The newest transcriptomic dataset and most promising technique is the *"RNA-Seq"*. RNA sequencing is "an experimental protocol that uses nextgeneration sequencing technologies to sequence the RNA molecules within a biological sample in an effort to determine the primary sequence and relative abundance of each RNA" **(Martin and Wang 2011)**.RNA is converted to a library of cDNA fragments. To each of the cDNA fragments adaptors are subsequently added. Then high-throughput sequencing technology produce a short sequence from each cDNA strand. The alignment of these sequences to a reference genome or transcriptome divides them into three categories: (1) exonic reads, junction reads and poly(A)end-reads. Combining the three types a base-resolution expression profile of each gen is gained. The technical processing of the RNA-Seq datasets is explained in section 3.4.3. An overview of the process gives **Figure 4**.

There are several advantages: (1) researchers can use all DNA-Sequencing technologies; (2) amplification is not obligated, but does not have a limit of quantification either; (3) reads are annotated to reference transcripts or genome; (4) creation of a genome-scale transcription map, if no reference database is available; (5) obtaining transcriptional structure and/or level of expression at the same time for each gene. Therefor, RNA-Seq is good for species without genome information; (6) very low background signal and (7) requires less RNA sample **(Martin and Wang 2011).**

The RNA-Seq is still under development and it has some limitations: (1) the cDNA libraries used does not allow profiling all types of transcripts; (2) the fragmentation of large RNA molecules into (200-500bp) is necessary; (3) short reads can be identical (abundant RNA species or PCR artefacts); (3) improving of Bioinformatic methods to analyze the data; and (4) high percentage of transcripts surveyed is very cost intensive, but necessary to detect important genes. **(Wang et al. 2009)**

RNA-Seq is a promising technique which will provide new insights in the transcriptomic level of cells. Because it depends on sequencing techniques it will be even better if these techniques advance and are less cost intensive as they are today. Path2enet includes two a priori processed RNA-Seq datasets of homo sapiens. Both datasets contain FPKM per protein (Uniprot Identifier). The datasets are described in section 3.1.2

1.1.2 Proteomic datasets

In this paragraph I will give a short introduction to the two proteomic techniques used in this thesis. The paragraph 1.1.2.1 describes the LC-MS/MS proteomic data generation and 1.1.2.1. the 1.1.2.2 the newer antibody-based SEC-MAP technique

1.1.2.1 Mass Spectrometry LC-MS/MS

Mass spectrometry-based proteomics is an indispensable tool for molecular and cellular biology. The technique is useful to deal with the large-scale determination of gene and cellular function at protein level. For Systems biology it is important to study protein-protein interactions via affinity-based isolations on a small and proteome-wide scale. It helps to map the organelles of organisms **(Aebersold and Mann 2003**). The LC-MS/MS experiments can be devided in 5 steps. The typical proteomics experiment consists of five stages. (i) Isolations of proteins by biochemical fraction or affinity selectioin with a final step of one-dimensional gel electrophoresis (ii) proteins are enzymatically degraded to peptides (iii) separating the peptides via high-pressure liquid chromatography (LC) (iv) mass spectrum of the peptides eluting (normal mass spectrum – MS) and (v) generating a prioritized list of peptides for

fragmentation and a series of tandem mass spectrometric (MS/MS) experiments ensues.

In stage 1, the proteins to be analysed are isolated from cell lysate or tissues by biochemical fractionation or affinity selection. This often includes a final step of one-dimensional gel electrophoresis, and defines the 'sub-proteome' to be analysed. MS of whole proteins is less sensitive than peptide MS and the mass of the intact protein by itself is insufficient for identification. Therefore, proteins are degraded enzymatically to peptides in stage 2, usually by trypsin, leading to peptides with C-terminally protonated amino acids, providing an advantage in subsequent peptide sequencing. In stage 3, the peptides are separated by one or more steps of high-pressure liquid chromatography in very fine capillaries and eluted into an electrospray ion source where they are nebulized in small, highly charged droplets. After evaporation, multiply protonated peptides enter the mass spectrometer and, in stage 4, a mass spectrum of the peptides eluting at this time point is taken (MS1 spectrum, or 'normal mass spectrum'). The computer generates a prioritized list of these peptides for fragmentation and a series of tandem mass spectrometric or 'MS/MS' experiments ensues (stage 5). Figure 5(A) is an overiew of such shotgun LC-MS/MS approach. The next step is to match MS/MS spectra with theoretical spectra predicted for each peptide contained in a protein sequence database. Sequence tag-assisted database searching starts with extraction of short tags followed by database searching, in which the list of candidate peptides is restricted to those peptides only, that contain one of the extracted sequence tags, allowing for mutations in the sequences of candidate database peptides. Peptide sequence can also be extracted directly from the spectrum using d*e novo* sequencing (extracted sequences can then be searched in a protein sequence database to find the exact or a homologous peptideisolation of a given peptide ion, fragmentation by energetic collision with gas, and recording of the tandem or MS/MS **(Alexey and Nesvizhskii 2014)**.

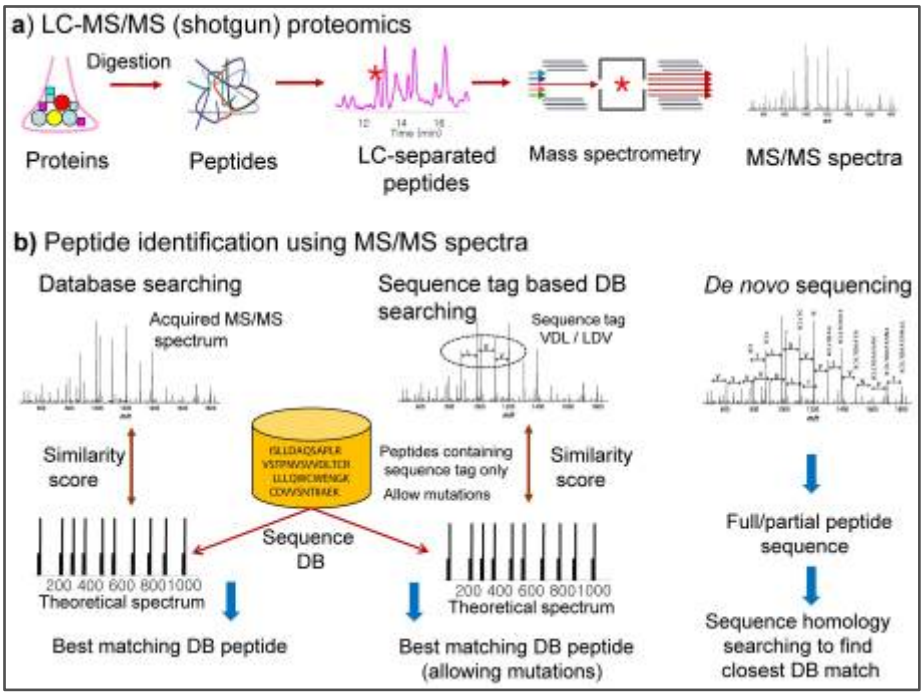


Figure 5: *Overview of a peptide and protein identification process with LC-MS/MS shotgun proteomics.*

1.1.2.2 Antibodies - SEC-MAP

Size Exclusion Chromatography Microsphere-based Affinity Proteomics Arrays (SEC-MAP) is an antibodybased proteomic approach to identify pre-selected proteins in the cell. SEC-MAP has been successfully employed to detect hundreds of proteins in a single sample and provide crucial information about protein size and subcellular localization. Since few amount of cells is required (~5-10 x 106 cells), the SEC-MAP approach could be helpful as a sensitive high-content tool for protein profiling in different cells (including lymphocytic ones) **(Kanderova 2016).** However, previous knowledge about the purpose of the study is required to select the antibodies included in the SEC-MAP array. First the Microsphere-based Affinity Proteomics Arrays (MAP) array and a differently color-coded microspheres hvae to be designed and selected to use this technique. These proteins have to be releveant for the study and are selected by the researcher. Additionally, a qualitative approach has been designed in order to identify which antibody/protein tandems are suitable for immunodetection and/or immunoprecipitation techniques. The **Figure 6** shows the design of a SEC-MAP proteomic approach. Because this type of proteomic data is more specific, it is often used in combination with LC-MS/MS and/or transcriptomic data.

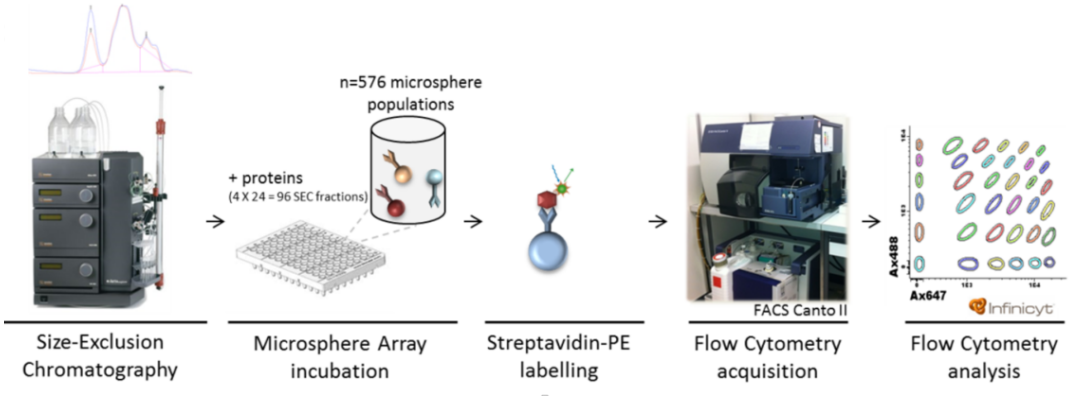


***Figure 6**: Overview of a SEC-MAP experimental design and analysis.*

1.2 Status of biological network analysis

1.2.1 Graph Theory

The grow of the OMIC datasets available explained in setion 1.1 has lead to a increase of knowledge how these elements interact with each other. These interactions are the basis of biological networks. Like mentioned above, these biological networks are very large and have to be interpreted statistically and visually in a form that researcher can gain insights in biological processes. **(Barabasi et al. 2011)** Therefore, the biological networks are interpreted as graphs. The entry (vertice) of a graph can be any cellular component like a gene protein or metabolite. The interaction/ relation between two vertices is called edge. The networks can be directed or undirected. **Figure 7** shows the differences. In the directed graph it is not possible to go from E to any other vertex in the network, because it is a dead end. In the undirected graph you can reach any other point starting from ventex E.

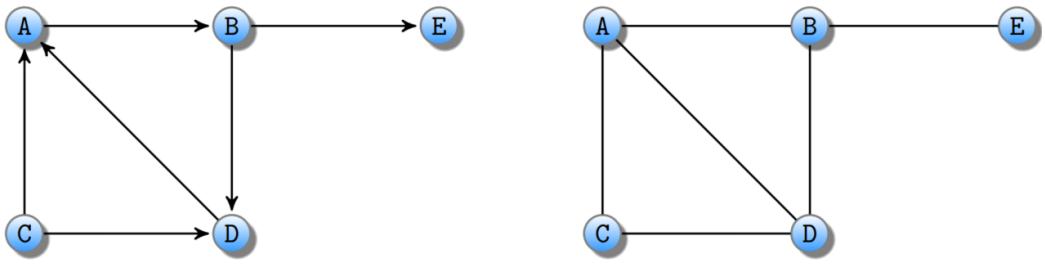


Figure 7 shows on the right side a directional graph and on the left side an unidirectional graph.

Metabolic networks, based on enzymes and compounds, are often represented as directed graphs. Portein-protein physical interaction networks are interpreted as undirected networks. To gain insights into large graphs, the following topological parameters are important:

shortest path of two different vertices in a network. The shortest path in right panel of **Figure 7** from C to E is C-D-B-E or C-A-B-E.
distance is the number of vertices to walk the *shortest path*. In our example from C to E it is 3 and from A to E it is 2
diameter in our example the shortest path with the most nodes to pass is from C to E. This length is called diameter of the network.
In section 3.4.1.5 important network paramets like ***degree***, ***betweennes***, ***clustering coefficient*** and ***eigenvector*** are explained.

Key-players in a network are *hubs* which can be devided in wo categories: (1) party hubs and (2) date hubs. Party hubs are nodes of a network which interact with many other nodes of the network at the time. Party hubs are often important inside a single module or subnetwork of the network. Date hubs are different, because they interact with different partners at different network locations and times. They connect two modules or subnetworks in a network. If you remove the date hubs of a network, the network looses connectivity. If you remove party hubs, the connectivity of the networks keeps nearly the same.

Statistically party hubs have a high level of degree and clustering coefficient. Date hubs normally do not need a high degree, but a very high betweenness.

1.2.2 Biological network datasets

The following sections describe the different type of network data souces. Section 1.2.1.1 explains the different kind of protein-protein-interaction datasets available. Sectrion 1.2.1.2

1.2.2.1 Protein-Protein Interaction networks

In network analysis it is important how the gene-protein / protein-protein interaction is proven. A very good example to demonstrate how databases can provide curated and non-curated data at the same time is the protein database “Universal Protein knowledgebase” [UniProtKB] . This database provides a comprehensive, high-quality and freely accessible resource of protein sequence and functional information to the scientific community. But the database has two kind of datasets: “UniProt Knowledgebase/Swiss-Prot” [UniProtKB/Swiss-Prot] and “UniProtKB/Translated EMBL Nucleotide Sequence Data Library“ [TrEMBL]. The highly accurate dataset Swiss-Prot has less entries but provides higher quality. TrEMBL provides more proteins but the quality is much lower.

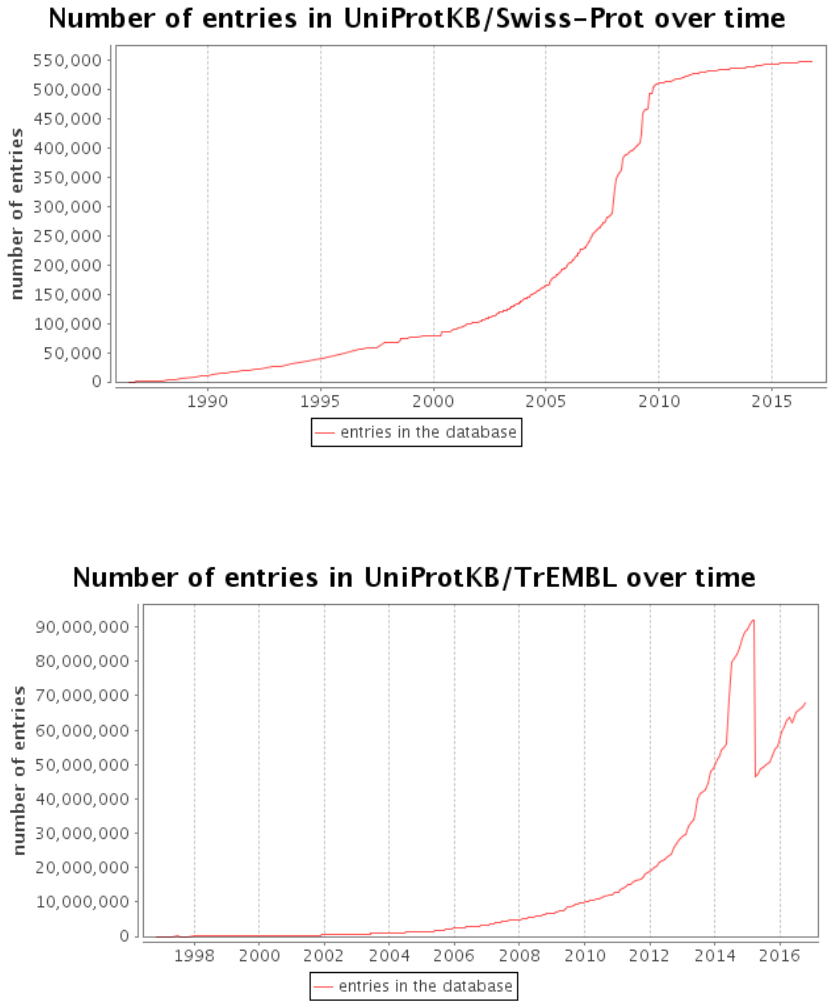


Figure 7: This charts show the grow of entries in the SwissProt and TrEMBL database of UniprotKB. The drop in 2015 in the TrEMBL database was the result of removing redundant entries

For an integrated network-based analysis scientists need really reliable data of interactions. There are two types of interaction databases. The first kind of databases are called Primary databases. The primary databases obtain their information directly from the scientific literature. Curators, experts in protein-protein-interactions, have to evaluate, interprete and compilate the results of published experiments in order to include an interaction or not. These experts follow common annotation protocols such as those set out in (Orchard et al., 2012) or (The UniProt Consortium 2014). The researchers, for their part, are asked to follow a series of recommendations when describing their experimental results (Orchard et al., 2007). There are several protein-interaction databases like ”Biomolecular interaction network database” [BIND] , ”Biological General Repository for Interaction Datasets” [BioGRID] , “Database of Interacting Proteins” [DIP] , “Human Protein Reference Database” [HPRD] , ”Interaction Database” [IntAct] und “Molecular Interaction Database” [MINT] .

The second kind of interaction database try to combine and evaluate the entries of several primary databases. Therefor, they are called Meta-databases. The bioinformatic web tool “Agile Protein Interaction DataAnalyzer”, which is used in this work, [APID] unifies these databases and provides additional information: **(i)** in which database can specific protein-protein-interactions be found; **(ii)** which kind of experiments have been made; and **(iii)** how

many experiments have been made to verify a protein-protein-interaction. Figre X shows the graphical representation of the "Notch Signaling Pathway" in APID. The integration of various databases in one datasets helps bioinformatics preparing an integrative network-analysis. Other important databases are iRefWeb **(Turner et al. 2010)**, Mentha **(Calderone et al. 2013)**, HINT **(Das and Yu 2012)**, GeneMania **(Warde-Farley et al. 2010)** and STRING **(von Mering et al. 2003)**.

Another source for protein-protein interactions are interactions based on the 3D structure of the inivolved proteins. This interactions are based on nuclear magnetic resonance spectroscopy **(Volkman et al. 1998)** , electron microscopy **(Kostyuchenko et al. 2003)** or X-ray crystallography **(Omar et al. 2016)**. The Protein Data Bank (PDB) **(Berman et al. 2000)** provides this information.

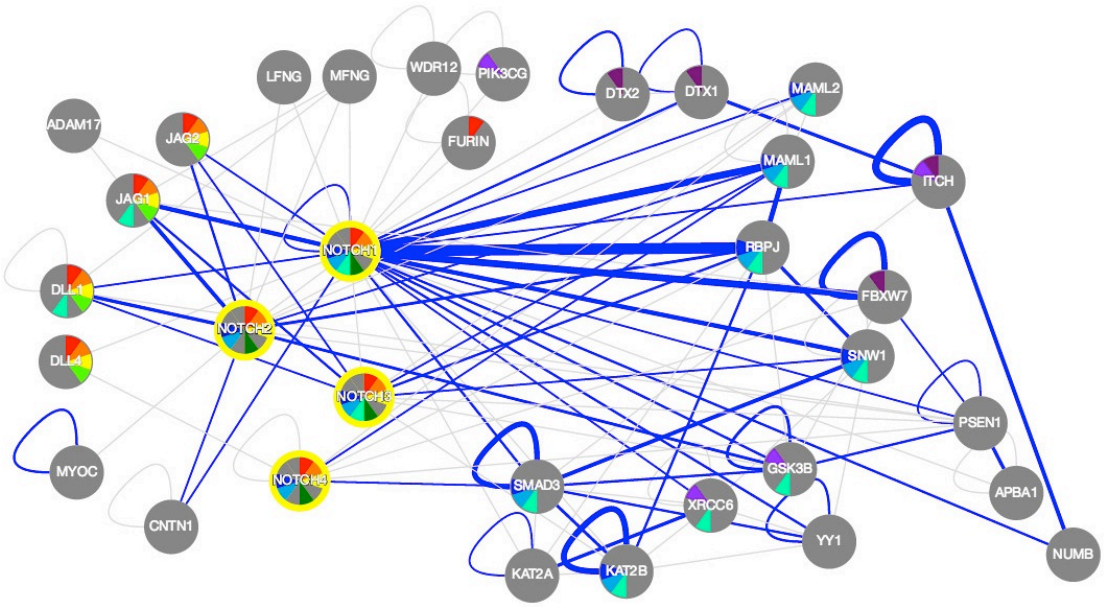


Figure 8 shows the graphical representation of the "Notch Signaling Pathway" generated with the APID-tool. It includes the interactions of the pathway, the number of experiments available to confirm the interaction (edges) and information of functional enrichment (color of nodes)

1.2.2.2 Biological Pathways

Another major source of highly curated interactions are pathway based databases. Important databaes are "KEGG", "BioCyc", "Pathway Commons" and "Reactome". These databases store data of important processes in the cell on (1) gene; (2) proteomic and (3) metabolic level.

Pathways can be categorized in metabolic pathways and non-metabolic pathways. Metabolic pathways can be considered as indirect protein-protein-interactions, and which can also be described as a network of enzyme-enzyme relations. Non-metabolic pathways can be described as generalized protein-protein-interaction networks. You can find **(i)** direct relations of proteins like phosphorylation; **(ii)** indirect interactions, like relations of transcription factors; and **(iii)** transcribed gene products via gene expressions.

The regulation of cellular pathways is essential in complex organism. Diseases like cancer are the result of deregulated pathways and cellular networks. This is the reason why the databases provide information of pathways in diseases, too. KEGG stores at the moment (May. 2017) 23

cancer annotated pathways like “Melanoma”, “Acute myeloid leukemia”, “Prostata Cancer”, “Colorectal cancer”, “Viral carcinogenesis” and “Small cell lung cancer”. In KEGG highlights oncogenes and tumor suppressor genes in its pathways. This makes it easier to understand their crucial role in cellular pathways and why their malfunction leads to cancer. **Figure 10** shows the graphical representation of the pancreatic cancer in KEGG with the oncogenes and tumor suppressor genes highlighted in red.

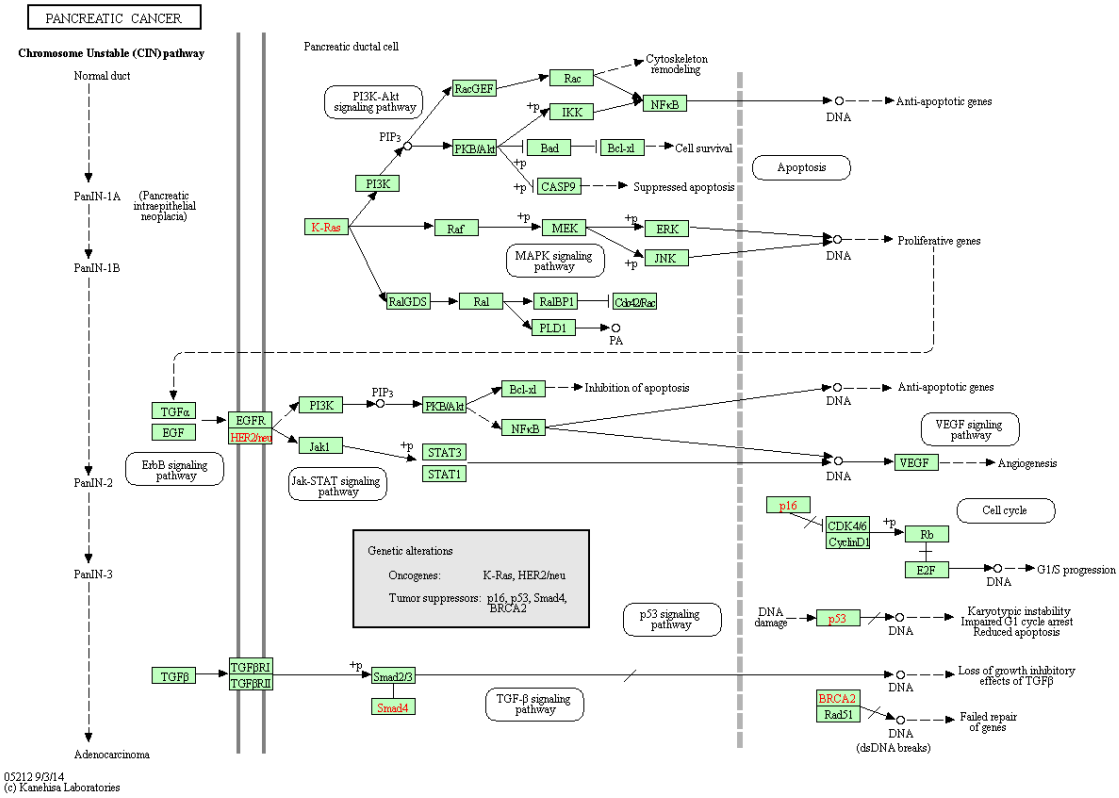


Figure 9 Graphical representation of the pancreatic cancer in KEGG. The oncogenes (KRAS, HER2/neu) and tumor suppressors (p16, p553, Smad4, BRCA2) are highlighted in red in the pathway.

1.2.3 Bioinformatic tools available

Existing libraries and packages to analyse and visualize the KEGG database in R and/or Bioconductor.

There are several “packages” available in R and/or Bioconductor which use data of the KEGG database in different ways. There are annotation packages like “KEGG.db”, which map for example “Enzyme Commission numbers” to “Gene”, using data provided by KEGG (3). The package “Gene2Pathway” predicts the mapping of a gene to a KEGG pathway based on its domain signature, using the “InterPro” database, which “offers predicted protein domain annotation for 19,000 of all 23,000 genes in the “IPI human” database” (4).

The R package "KEGGgraph" (5) is a tool to combine data of pathways provided by KEGG, like"KEGG Markup Language" [KGML] files, with graph methods provided by the R environment. It is able to parse KGML files, which can be downloaded from an ftp server of KEGG, and transform them into igraphs-objects. Igraphs-objects are the basis of analyzing and visualizing graphs in the R environment. These igraphs-objects can be merged to generate graphs from several KGML files. "KEGGgraph" can use non-metabolic and metabolic KGML files provided by KEGG (**Zhang and Wiemann 2009)**.

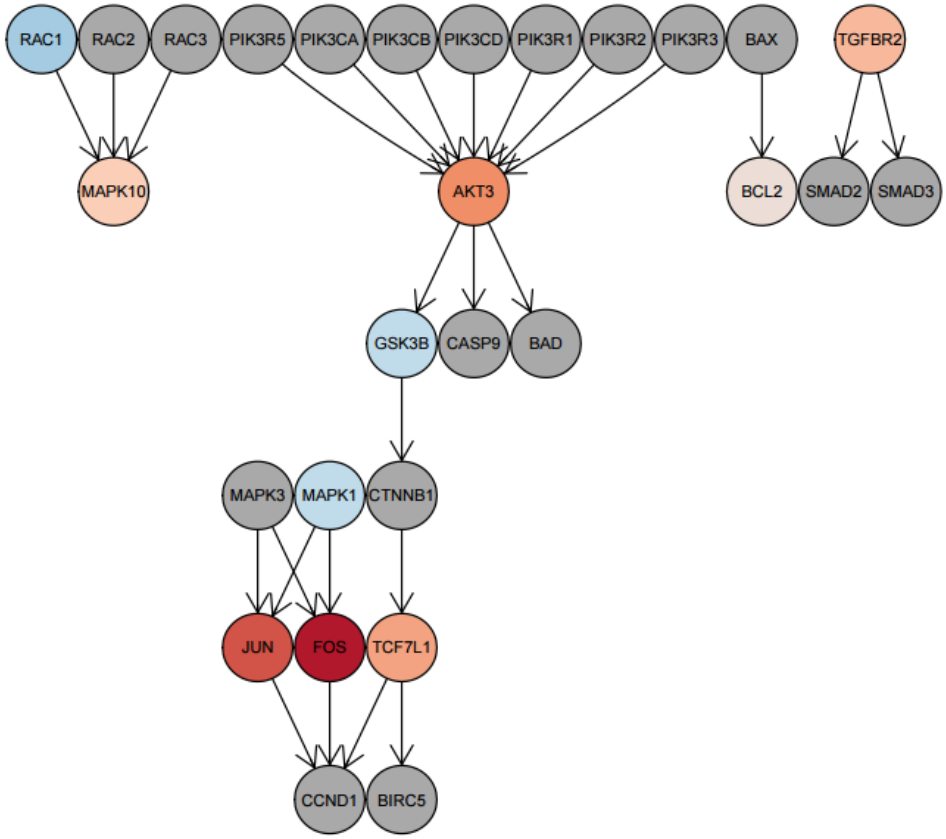


Figure 10 Representation of KEGGgraph of the colorectal cancer pathway and a differential expression microarray dataset.

A similar approach to transform biological pathways in networks enables the R package "PaxtoolsR". It contains a parser for the BioPAX OWL files and enables to search for networks in the Pathway Commons database. Pathway Commons databases include: BIND, BioGRID, CORUM, CTD, DIP, Drug Bank, HPRD, HumanCyc, IntAct, KEGG, MirTarBase, Panther, PhosphoSitePlus, Reactome, RECON, TRANSFAC. This tool provides interesting features to have access to a large source of datasets.

Another tool to use the BioPAX level 2 and 3 formats is the "rBiopaxParser". It allows to merge different graphs as well.

The package “Pathview” (6) uses the data “KEGGgraph” provides but enables a very KEGG oriented view of the pathways. **Figure 12** shows an example representation of “Cell Cycle” pathway which data is parsed by KEGGgraph and rendered by “Pathview”. The function allows to integrate processed protein or gene expression datasests.

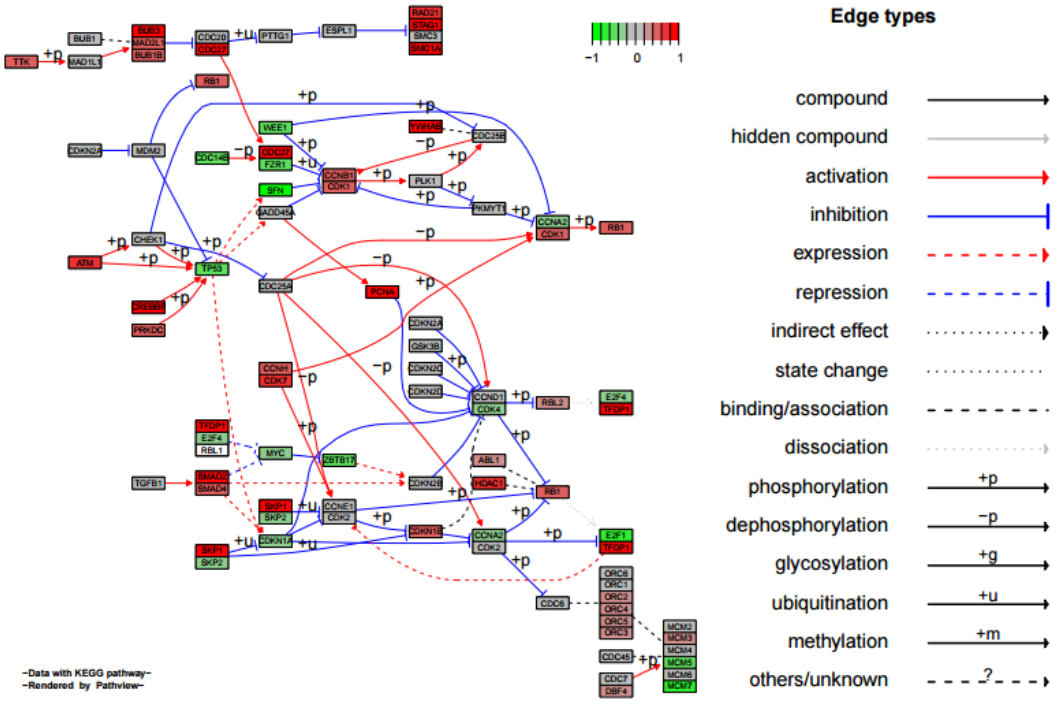


Figure 12 shows the “Cell Cycle” pathway representation of the Pathview package.

A similar approach to use the data of KEGG and generate a pathway-like representation like Pathview uses the KEGGlincs package. But this packges uses the Cytoscape

The “MetaboSignal” (7) R-package is designed to combine metabolic and signaling KEGG pathways into networks. The networks are useful to explore the topological relationship between genes (signaling- or enzymatic-genes) and metabolites. This allos to combine genetic information with metabolic data to build regulatory networks.

The package “BioNet” **(Beisser et al. 2010)** is an integrated network-analysis package. It provides data of the “Human Protein reference database” and combines it with gene-expression data-sets. It uses the “heinz” algorithm to find the maximal significantly deregulated set of interconnected genes in a cellular network. The “heinz”-algorithm combines close connection of maximum-scoring connected subnetworks and “prize-collecting Steiner trees” (PCST). It incorporates **(i)** edge weights; and **(ii)** analyzed modules of a predefined size to find deregulated subnetworks.

The package “pwOmics” is a tool to integrate pathway-abased datasets and time-series of omics data. It uses proteomic and transcriptomic data to calculate the expression difference of upstream transcription factors and the genes regulated by these transcription factors. The downstream analyisis searches for expression changes in the proteomic dataset of regulatory molecules. The package uses the pathway datasets, which the *rBiopaxParser* handles.

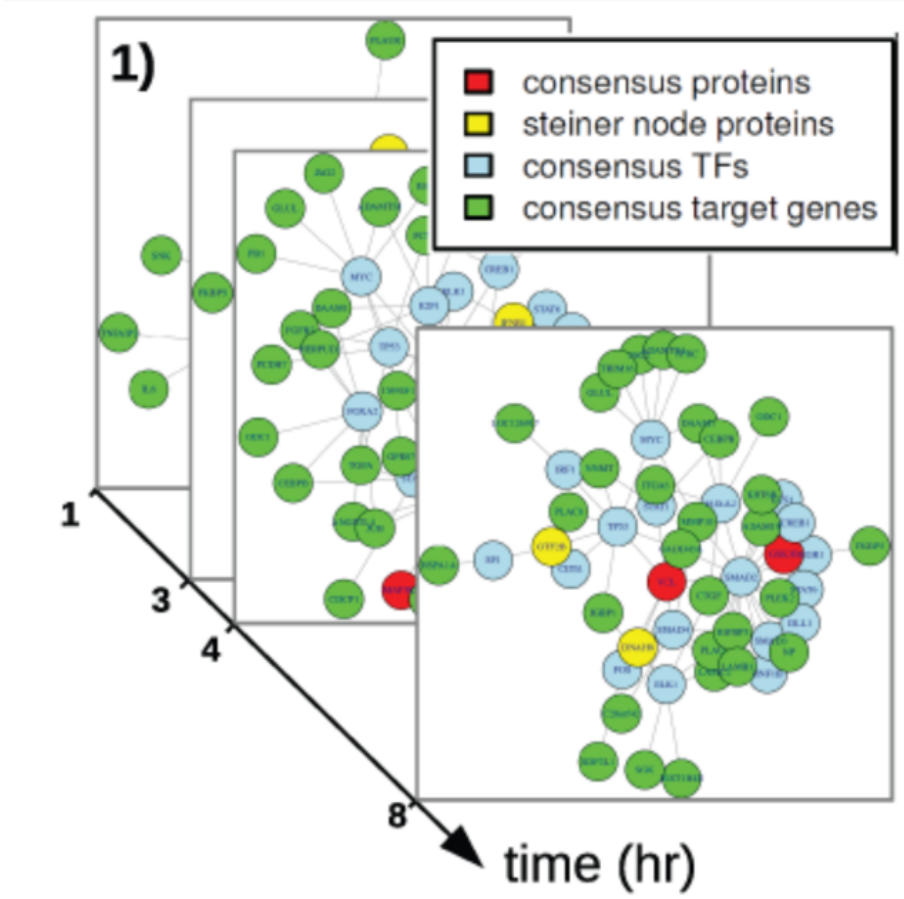


Figure 13: Representation of the downstream and upstream analysis of pwOmics.

The “graphite”-package **(Sales et al. 2012)** provides access to other pathway-databases as well. It incorporates data of: (1) “Biocarta”; (2) “KEGG”; (3) “NCI/Nature Pathway Interaction Database”; and (4) “Reactome”. The package includes the function “clipper” (8) which uses graphs decomposition theory and pathway topology analysis to identy the signal pathways which have the greatest association with a specific phenotype.

Similar topology-based pathway analysis tools are SPIA, DEGraph, TopologyGSA, TAPPA, PRS and PWEA. The tools have the approach to find the most significant paths or pathways in an expression dataset. To describe all in detail would expand the introduction and topology pathway analysis is not the key point of this thesis.

The R package “SubpathwayMiner” is a tool for pathway identification. It is written to annotate gene sets obtained by high-throughput experiment like microarrays to pathways stored in the KEGG database. The creation of sub-pathways can be modified by the user setting up a distance parameter. The distance parameter is the maximal distance between all enzymes in a metabolic pathway like “citrate cycle pathway”. The created cliques found in the metabolic pathways according the distance parameter are called “sub-pathways”.

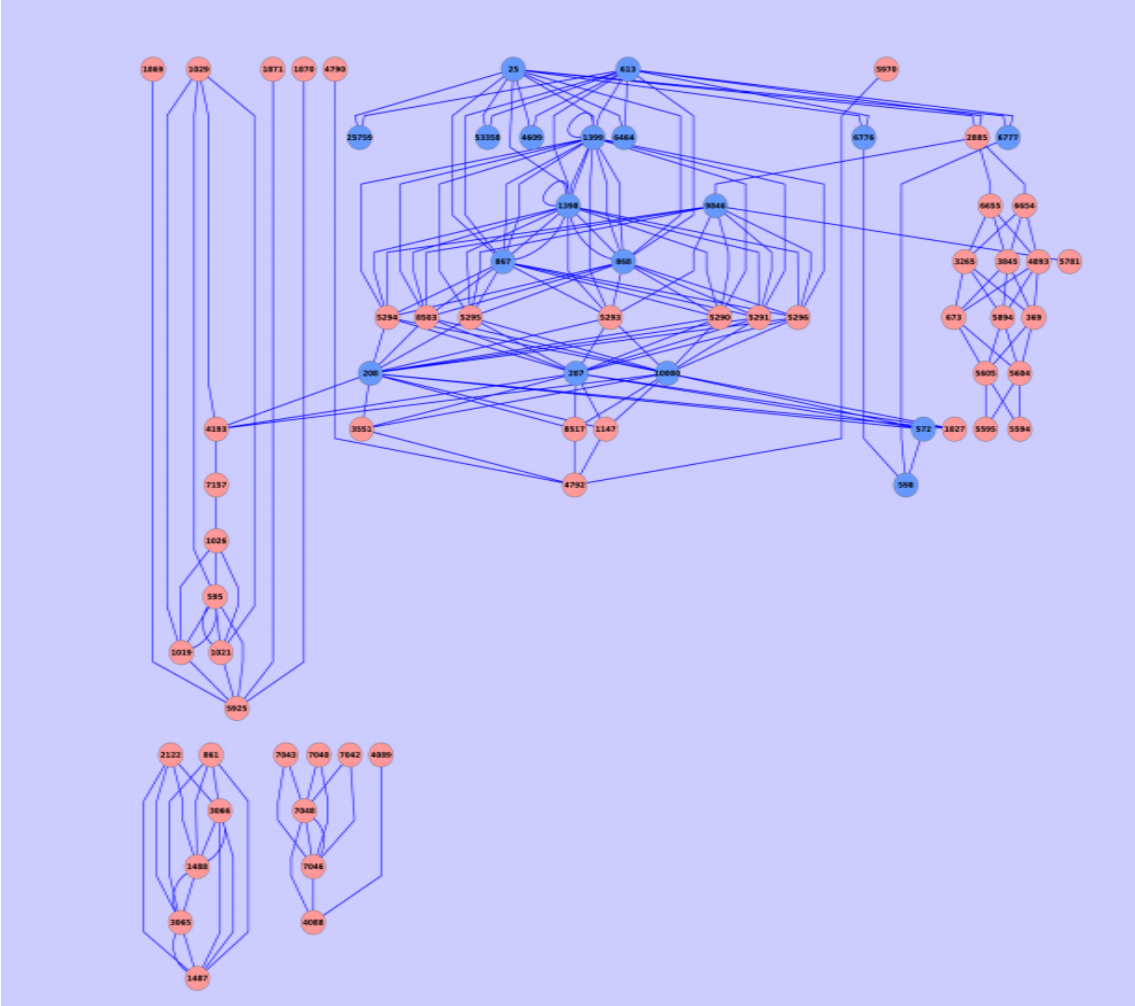


Figure 14 shows the representation of”Chronic myeloid leukemia” in clipper. The function highlighst the most significant path in blue.

“SubpathwayMiner” provides **(i)** functions such as “sub-pathway” annotation and identification of metabolic pathways based on either “enzyme commission” [EC] numbers, “KEGG Orthology identifiers” [KO] or entire pathways; **(ii)** functions to create statistically enriched “sub-pathway” out of data of differential gene expression; and **(iii)** functions to visualize metabolic pathways and “sub-pathways”. “SubpathwayMiner” supports most organism in KEGG and can be updated regularly **(Li et al. 2009)**.

Another type of network-based biology analysis provides the package FGNet (7). This tool enables to create functional gene networks derived from biological enrichment analyses of DAVID, GeneTerm Linker, gage and topGo.

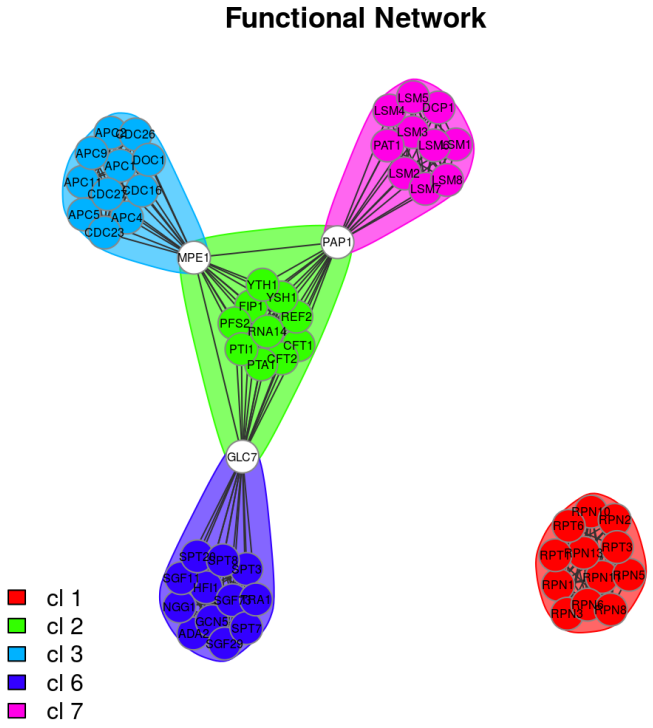


Figure 14 shows a functional gene network created with FGNet. White nodes are members of several annotation clusters and connect them.

The package “GOstats” enables a characterization of genes in a cellular network. This allows to analyze which type of genes are significant deregulated in a certain cellular subnetwork.

2 OBJECTIVES

2.1 Problem position and hypothesis

At present **most of the biomolecular studies**, specially on human samples, that are published in high-quality scientific journals (and reported in PubMed) are focused in **a detailed demostrative analysis of one or a small group of proteins**, with a frequent indifference or disregard of all the biomolecular partners that are not tested in the performed experiments, but are really working and interacting together with the protein or proteins studied. This inductive simplification is intrinsic to the scientific method but it is against the **current trend of omic global-scale data production**. We think that such frequent neglect or omission of the complex biological context could be repair with a more proper **application of bioinformatic and computational biology tools and methods to the biological studies**. In this way we present in this Thesis an important problem that affects current research in biological sciences and we suggest the application of better bioinformatic tools and methods as a clear approach to improve such controversial situation, proposing a step in this direction.

2.2 Objectives

The **general objective** of this **Doctoral Thesis** is the development and application of **bioinformatic** algorithms and methods to integrate, analyze and visualize various sources of genomic, proteomic and biomolecular information applied mainly to human data. The management and integrative analysis of the multiple and complex **data** currently available for proteins and genes at global level (i.e. at "omic" scale) is a challenge for biomedical studies and a clear scenario to apply and develop new bioinformatic methods and tools. Within this framework, the present **Doctoral Thesis** develops some specific approaches to integrate several layers of biological information and experimental data (such as: biological pathways data, protein-protein interaction data, experimental genomic and proteomic expression datasets) in order to enable the generation of biomolecular networks and facilitate a global overview and analysis of biomolecular processes in specific biological contexts. These methods are applied, as **specific cases-of-study**, to several experimental datasets: **(i)** human lymphocytes (B- and T-cells), **(ii)** transcriptomic and proteomic data obtained for some lymphoma human B-cell lines and **(iii)** proteomic data obtained for cells isolated from B-cell leukemia patients.

To be more precise, the following **four specific objectives** are proposed:

Objective 1. Design and development of an integrated **biological database** that combines **pathways, proteins and expression** information. This database will include three types of data: first, information about human biological pathways (obtained from **KEGG**) integrated with the corresponding protein information (obtained from **UniProt**); second, protein-protein physical interaction data (derived from **APID**); third, human gene expression data from different cell types, tissues and organs (obtained from several transcriptomic resources: Expression Sequences Tags [**ESTs**] data, mRNA microarrays data and RNA-Sequencing data).

Objective 2. Development of a **bioinformatic tool or application** which provides a translation of **biological pathways in protein networks** integrating several layers of information about the biomolecular nodes in a multiplex view. The tool uses the **integrated biological database** generated in the first objective, **transforming the pathways into networks** and enriching the networks with experimental protein interaction data and gene expression data. The tool also analyses the expression data to determine if a given gene/protein in a network (i.e., a node) is active (ON) or inactive (OFF) in a specific cellular context or sample type. In this way, we want to reduce the complexity of the networks and reveal the proteins that are active (expressed) under specific conditions. As a whole, the bioinformatic tool will be designed to analyze and visualize single or multiple pathways in form of ***pathway-expression-networks***. As a case of study, we applied this tool to the investigation of some specific human cell types: B- and T-lymphocytes.

Objective 3. Development and application of an **integrative analysis of proteomic datasets** (obtained by mass-spectrometry or by antibody specific identification) and **transcriptomic datasets** (obtained by gene expression mRNA microarrays and RNA-Sequencing) in a qualitative way. This **proteogenomic** approach wants to provide a global overview of which protein and genes are active in a cell or patient sample and which biological functions they perform. As a case of study, we applied this method to a dataset of Burkitt's lymphoma B-cells.

Objective 4. Development and application of a procedure to **integrate quantitative proteomic and phospoproteomic data** (obtained by mass-spectrometry) and calculate **relative protein expression levels**, as well as, mapping the identified protein profiles to specific biological processes and signaling pathways. As a case of study, we applied this method to a set of B-cell chronic lymphocytic leukemia (CLL) patients with different clinical states.

All the presented methods are performed by using **human samples** isolated from different tissues, development stages, cell lines or patients. All datasets are at present public available and in several cases they come from the experimental work done by our collaborators of the Lab11 of the CiC-IBMCC (USAL/CSIC).

3 MATERIAL AND METHODS

3.1 Data to build pathway-derived networks in an expression specific context

This section describes the biological databases and data that we used to generate our application (**Path2enet**) in order to build integrated biomolecular networks. These are the following: **(1)** Biological pathways and the derived information on genes/proteins, with the type of associations that they have in the pathways; **(2)** Protein-Protein interaction data; **(3)** Pre-processed gene expression data (derived from datasets of ESTs, microarrays and RNA-Seq); and **(4)** Transcriptomic expression data provided by the user.

3.1.1 Network specific data

We used two kinds of interaction data to generate the networks. A highly curated pathway database, *Kyoto Encyclopedia of Genes and Genomes* [KEGG]; and a meta-database of protein interactions which contains experimentally proven physical protein-protein interactions, called *Agile Protein Interaction DataAnalyzer* [APID].

3.1.1.1 Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG is a knowledge-base database of biological pathways for systematic analysis of gene functions, linking genomic information with higher order functional information **(Kanehisa and Goto 2000)**. The genomic information is stored in the KEGG GENES database and the higher order functional information is stored in the KEGG PATHWAY database. KEGG PATHWAY is a curated pathway database which represents a high quality resource for biomolecular associations between proteins in a well-defined and known biological context. KEGG PATHWAY provides manually drawn reference pathways on several categories. One category is "metabolic pathways" (like the *Citrate cycle* pathway) which can be described as networks of enzyme-enzyme relations. Another category is "non-metabolic pathways" (i.e. "Genetic Information Processing", "Environmental Information Processing", "Cellular Processes", etc) which in most cases can also be described as networks of protein-protein associations. The links or association between proteins in the pathways can be: **(i)** direct relations, in reactions like phosphorylation; **(ii)** semi-direct interactions, like relations via specific transcription

factors; and **(iii)** indirect interactions, like the ones of transcribed gene-products via common gene regulation.

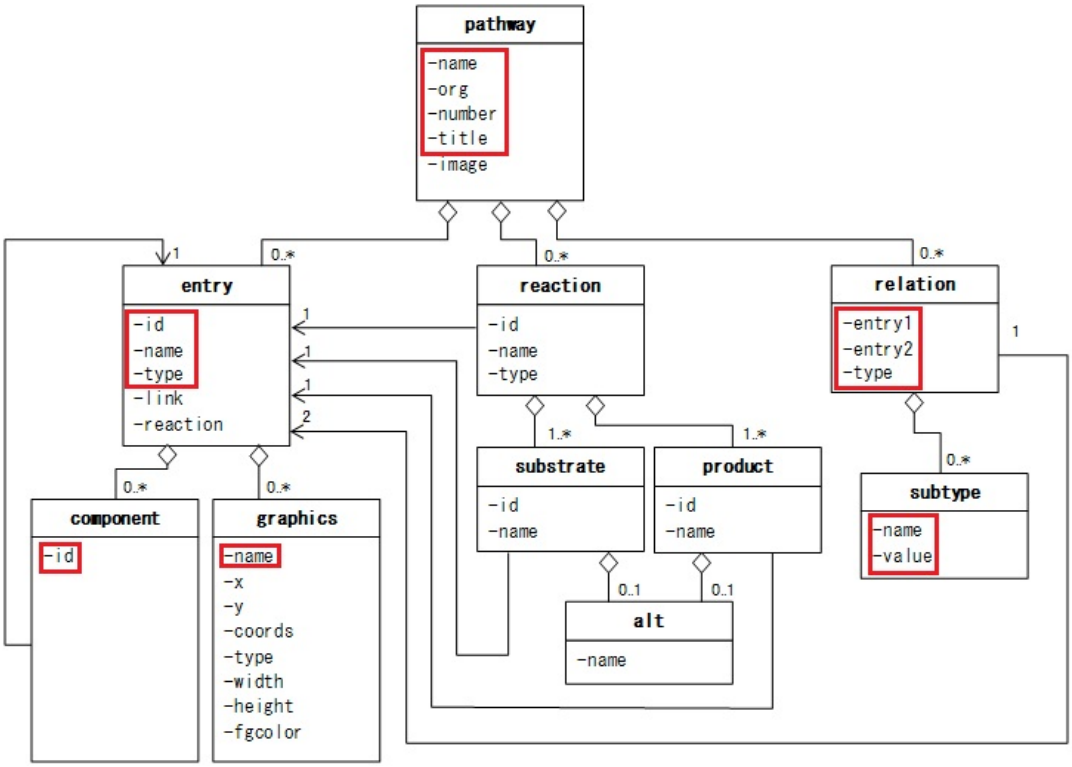


Figure 15: *Overview of KGML. The red boxes mark the data used by Kegg2MySQL*

KEGG stores its data in files using the "KEGG Markup Language" [KGML]. KGML files are written in the "Extensible Markup Language" [XML] (9). These files enable to exchange graphical pathways of KEGG with other databases or programs. KEGG also uses the BR file (KEGG BRITE HIERARCHY FILES that include all the categories of KEGG database) to describe the complete structure of the data. As indicated above, the KGML files are divided into two major categories: **(i)** metabolic pathways; and **(ii)** non-metabolic pathways. In this work we used the KGML files (stucture shown in **Figure 2**) and BR files of May 2016 to create the database for the pathway-based network (i.e. the one included in Path2enet tool). Both file types are public available on the KEGG PATHWAY website. In particular, in all our studies we used the KGML-files of *Homo sapiens* [hsa].

Both metabolic and non-metabolic pathways can be read as networks if we extract the information as nodes and edges. In this way, **nodes** in the pathways are proteins or gene-products (considered as central cellular biomolecules); and **edges** are different types of reactions or associations between two or more proteins. The pathways, apart from biochemical reactions, can also include specific information about protein-protein physical interactions.

The objective of bioinformatic application and tool that we are going to develop (i.e. Path2enet) is representing pathways as networks trying to transfer and incorporate all the information related to protein-protein associations into networks. Therefore, from the KGML files we extract all the information relevant to create protein networks, that in each case corresponds to: **(i)** which proteins interact; and **(ii)** which type of relations or associations

occur between each specific protein pair. As indicated above, **Figure 15** gives an overview of the structure of the KGML-files used, reflecting the type of data-architecture that Path2enet incorporates and uses.

3.1.1.2 Protein-Protein Interaction data (APID)

Agile Protein Interaction DataAnalyzer (APID) is an interactive bioinformatic web resource developed to integrate and analyze in a unified and comparative platform the main currently known information about physiscal protein-protein interactions demonstrated by specific small-scale or large-scale experimental methods **(Prieto and De Las Rivas 2006; Alonso-López 2016)**. This tool has received a major update in 2016. In this way, the whole dataset included in APID is focused on highly validaded **protein-proteni physical interactions** (PPIs). Moreover, APID provides quality levels of each interaction like the number of experiments, methods and publications.

APID includes various primary sources of protein-protein-interactions: DIP, IntAct, MINT, HPRD, BioGRID and BioPlex. It also includes interactions based on 3D-structures of complexes in the PDB and PDBsum data base. **Table 1** shows the number of interactions included in these databases in December 2016.

Table 1: *Number of human interactions and proteins included in several source databases.*

Database	Type of nodes	Type of interactions	Number of proteins	Number of interactions
IntAct	Proteins and other molecules	Experimental (ppis)	28,749	108,219
PDB	Proteins	Experimental (ppis)	30,089	1,297
HPRD	Proteins	Experimental (ppis)	30,047	41,740
DIP	Proteins	Experimental (ppis)	27,701	5,543
BioGRID	Proteins and other molecules	Experimental (pp-i & genetic-i)	56,907	107,010
BioPlex	Proteins	Experimental (ppis)	7,668	55,154
APID	Proteins	Experimental (ppis)	29,701	349,144

APID also integrates and provides information on the functional annotation of the proteins obtained from diferent databases: GO, Pfam, Interpro and Reactome.

As we focus our study on human pathways, in our tool Path2enet we included the data about the "human interactome" of APID (from December 2016) which was available on its website, http://cicblade.dep.usal.es.

3.1.2 Expression data

This section describes which experimental datasets the Path2enet tool uses to generate tissue and cell specific networks. On the one hand, the tool provides experimental datasets *a priori* processed and, on the other hand, it can use input of experimental datasets provided by the user. The tool includes 3 gene expression data sources: Expression Sequence Tags [EST], microarrays, and RNA-Seq.

3.1.2.1 Expression Sequence Tags (ESTs)

UNIGENE is a database of the National Center for Biotechnology Information (NCBI) (10), a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH) of the United States of America (11). This database has high quality and will only take up non repetitive Expressed Sequence Tags (ESTs) which have at least 100 base pairs (12). ESTs are short (usually approximately 300-500 base pairs), single-pass sequence reads from cDNA. They represent the genes expressed in a given tissue and/or at a given developmental stage. UNIGENE has its focus on transcripts of protein-coding genes of the nuclear genome (13).

Our application Path2enet includes a priori processed EST files of *Homo sapiens* of May 2016. The dataset includes 87,916 observations of 52 variables (tissue or development specific). Section 4.1 describes how the tool generates user specific EST datasets .

3.1.2.2 Microarray expression datasets

3.1.2.2.1 Pre-processed datasets of Gene Expression Barcode 3.0

The *Gene Expression Barcode (version 3.0)* offers a dataset from high-density oligonucleotide microarrays that store 17,268 gene/protein entries detected in 195 tissues and cell lines. The Path2enet tool processes and integrates the tissue specific dataset based on the HGU133 Plus 2.0 (Human) microarray (abc-tis-gpl570-formatted_v3.csv) and the cell line specific dataset (abc-cell-gpl570-formatted_v3.csv) obtained from Gene Expression Barcode 3.0 website: http://barcode.luhs.org/index.php?page=transcriptome.

3.1.2.2.2 B- and T-lymphocytes datasets

To perform this analysis of the B- and T-Lymphocytes we downloaded and normalized an expression dataset that included 163 human samples. These samples were genome-wide expression microarrays of platform Human Genome U133 Plus 2.0 from *Affymetrix* (GEO reference: GPL570). The samples corresponded to naive B cells (CD19+), 32 microarrays; T cells (CD4+), 96 microarrays; and T cells (CD8+)

We include below the list of the 163 high-density oligonucleotides expression microarrays from human B-cells and T-cells used in this study (taken from Gene Expression Omnibus, GEO, database).

B cells (CD19+) 32 microarrays, samples ID in GEO (http://www.ncbi.nlm.nih.gov/geo/):

```
GSM595853 GSM595858 GSM595863 GSM595870 GSM609252 GSM609254
GSM609257 GSM609258 GSM609259 GSM609260 GSM609262 GSM609263
GSM609264 GSM609265 GSM609266 GSM609267 GSM629980 GSM629981
GSM705297 GSM705298 GSM705299 GSM705300 GSM705301 GSM746743
GSM746744 GSM746745 GSM746746 GSM746747 GSM746748 GSM746750
GSM746751 GSM746752
```

T cells (CD4+) 96 microarrays, samples ID in GEO (http://www.ncbi.nlm.nih.gov/geo/):

GSM251101 GSM251105 GSM251110 GSM251111 GSM251114 GSM251126
GSM251129 GSM251192 GSM251194 GSM304415 GSM304420 GSM304945
GSM304946 GSM322864 GSM345116 GSM345117 GSM345118 GSM345119
GSM345120 GSM345121 GSM345122 GSM345123 GSM345124 GSM345125
GSM345126 GSM345127 GSM345128 GSM345129 GSM345130 GSM345131
GSM345132 GSM345133 GSM345134 GSM345135 GSM345136 GSM345137
GSM345138 GSM345139 GSM345140 GSM345141 GSM345142 GSM345143
GSM345144 GSM345145 GSM345282 GSM345283 GSM345284 GSM364915
GSM364916 GSM371639 GSM371640 GSM371641 GSM371646 GSM371652
GSM372721 GSM372722 GSM372723 GSM372724 GSM372725 GSM372726
GSM372727 GSM372728 GSM372729 GSM372730 GSM403596 GSM413790
GSM413792 GSM413795 GSM413796 GSM413798 GSM413800 GSM413803
GSM413804 GSM472023 GSM548000 GSM548001 GSM595854 GSM595859
GSM595864 GSM595871 GSM642302 GSM642303 GSM642304 GSM705302
GSM705303 GSM705304 GSM705305 GSM705306 GSM788303 GSM788304
GSM788305 GSM788306 GSM788307 GSM788308 GSM788309 GSM788310

T cells (CD8+) 35 microarrays, samples ID in GEO (http://www.ncbi.nlm.nih.gov/geo/):

GSM159405 GSM159406 GSM198958 GSM371708 GSM371709 GSM371710
GSM371711 GSM371712 GSM372731 GSM372732 GSM372733 GSM372734
GSM372735 GSM372736 GSM372737 GSM372738 GSM372739 GSM372740
GSM372741 GSM403597 GSM595856 GSM595861 GSM595865 GSM595872
GSM705312 GSM705313 GSM705314 GSM705315 GSM705316 GSM826755
GSM826757 GSM826758 GSM826760 GSM826761 GSM826763

3.1.2.2.3 Ramos cell line dataset

To perform the gene expression profiling and analysis of the *Ramos* B-cells, we used a raw dataset of 3 mRNA samples corresponding to biological replicates of these cells. The details about these samples are described in section 3.2.1.1.

3.1.2.3 RNA-Seq dataset

Path2enet includes two a priori processed RNA-Seq datasets of *Homo sapiens*. Both datasets contain FPKM per protein (with Uniprot Identifiers).

The first RNA-Seq dataset is of the Human Body MAP 2.0 (ArrayExpress experiment E-MTAB-513 (14). It includes 16 tissues a three replicates. Path2enet uses the aligned BAM files that Ensemble provides (15) annotated to the GRCh37 (16).

The second RNA-Seq dataset is of The Human Protein Atlas (ArrayExpress experiment E-MTAB-2836 (17). This dataset includes 32 different tissues and different number of replicates. The data is annotated to ENSEMBL version 83 GRCh38. The Human Protein Atlas stores a preprocessed zip-file (rna_tissue.csv.zip (18) which the Path2enet tool uses.

3.2 Data used for the qualitative proteomics analyses

The following sections describe the dataset used in the two qualitative proteogenomic studies of the *Ramos* Burkitt's lymphoma-derived B-cell lines. Section 3.2.1 explains the global transcriptomic approaches based on microarray and LC-MS/MS datasets. Section 3.2.2 describes the global transcriptomic dataset of RNA-Seq and LC-MS/MS dataset, as well as the affinity antibody approach of 549 antibodies corresponding to 417 selected proteins of the *Ramos* cells.

3.2.1 Qualitative analysis of *Ramos* Burkitt's lymphoma-derived B-cell line

This section describes the transcriptomic and proteomic datasets used in the publication **"Integration of Proteomics and Transcriptomics Datasets for the Analysis of a Lymphoma B-Cell Line in the Context of the Chromosome-Centric Human Proteome Project" (Díez, Droste et. al. 2015)**. The experimental design and parameters to gain the proteomic datasets are written in detail in the publication. In this Thesis pre-processed proteomic datasets are used.

3.2.1.1 Microarray expression dataset

To perform the gene expression profiling and analysis of the *Ramos* B-cells, we used a raw dataset of three samples of mRNA from biological replicates of these cells hybridized on *Affymetrix* Human Gene ST 1.0 high-density oligonucleotide microarrays. The data are available at GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) database: series number GSE40168; samples GSM987747, GSM987748, GSM987749; platform GPL6244 ([HuGene-1_0-st]).

3.2.1.2 Mass Spectrometry derived proteomics dataset

The LC-MS/MS-dataset analysed in this study was produced by the proteomics group of the Cancer Research Centre (IBMCC, CSIC/USAL/IBSAL, 37007 Salamanca, Spain) leaded by Dr Manuel Fuentes. The raw dataset is uploaded and available on ProteomeXchange PXD001933 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD001933).

The analysis started with two raw proteomic peptide datasets (flat .txt files) corresponding to 4 different sample types (that were 4 different protein fractions isolated from the cells: cytoplasma, organelle, membrane, and nucleus) and for three experimental replicas. The two raw datasets for each sample corresponded to: **(i)** all proteins who are identified with at least 1 peptide; and **(ii)** all proteins who are identified with at least 2 peptides. In this way, we had a total of 2x4x3= 24 raw files with the following information: protein accession number of UNIPROT database (isoforms are included), protein name, percent sequence coverage (Coverage), predicted molecular weight (Predicted MW [Da]), total number of spectra assigned above the specified confidence level cutoff (PSMs), number of unique peptides assigned above the specified confidence level cutoff (# Unique Peptides), and number of total peptides assigned above the specified confidence level cutoff (# Peptides). The protein evidence (PE) information was included in the dataset processed with the neXtprot database search (release 2014-09-19).

3.2.2 Integrated analysis of MS/MS, RNA-Seq and Afiinity Proteomics of *Ramos* B-Cell data

This section describes the transcriptomic and proteomic datasets used in the submitted study **"Comprehensive combination of affinity proteomics, MS/MS and RNA-Sequencing datasets for the analysis of a lymphoma B-cell line in the context of the Chromosome-Centric Human Proteome Project" (Díez, Droste et. al. in preparation 2017)**. The experimental design and parameters to gain the proteomic datasets is written in detail in the publication. In this Thesis pre-processed proteomic datasets are used.

3.2.2.1 RNA-Seq dataset

The transcriptomic data corresponds to RNA-Seq for a *Ramos* B-cell line obtained with Illumina Genome. Analyzer IIx with paired layout (experiment SRX105534: http://www.ncbi.nlm.nih.gov/sra/SRX105534) was taken from the study SRP00931 (http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP009316) from SRA (Sequence Read Archive) database. The analysis of the gene expression in this *Ramos* B-cell line was done to calculate the values of FPKM (fragment per kilobase of exon per million fragments mapped) for each gene (see section 3.4.3 for the details of the data processing).

3.2.2.2 LC-MS/MS dataset

The LC-MS/MS-dataset analysed in this study was produced by the proteomics group of the Cancer Research Centre (IBMCC, CSIC/USAL/IBSAL, 37007 Salamanca, Spain) leaded by Dr Manuel Fuentes. The raw dataset is uploaded and available on ProteomeXchange. For the study presented in section 4.3.1 the data was uploadet under the ID PXD001933 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID= PXD001933)) **(Díez, Droste et. al. 2015)** for the study presented in 4.3.2 PXD003939 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD003939) **(Díez, Droste et. al. in preparation 2017)**.

Both datasets analysed are 12 MS/MS data files (corresponding to 3 replicas x 4 subcellular compartments) which contain the number of peptides found for each protein. These datasets were created with the neXtprot database search (release 2014-09-19 and release 2016-02). Each file included the following information: neXtProt-ID, protein name, percent sequence coverage (Coverage), predicted molecular weight (Predicted MW [Da]), total number of spectra assigned above the specified confidence level cutoff (PSMs), number of unique peptides assigned above the specified confidence level cutoff (# Unique Peptides), and number of total peptides assigned above the specified confidence level cutoff (# Peptides and protein evidence group, PE group).

3.2.2.3 Antibody dataset

The proteomics group of Dr. Manuel Fuentes also provided us with a dataset of Size Exclusion Chromatography Microsphere-based Affinity Proteomics Arrays [SEC-MAP] which included 549 antibodies corresponding to 417 distinct proteins considered as important or relevant for the *Ramos* B-cells studied.

3.3 Quantitative analysis of proteome and phosphoproteome of B-cell lymphocytosis of patients' samples

This section describes the datasets corresponding to the qualitative and quantitative proteome and phosphoproteome determination done for samples from 5 B-CLL patients. This work have been submitted for publication with the title **"Revealing Cell Signaling Pathways in Chronic Lymphocytic Leukemia Tumor B-cells by Integration of Global Proteome and Phosphoproteome Profiles"** (Díez P., Droste C. et. al, submitted 2017). The experimental design and parameters to gain the proteomic datasets are written in detail in the publication. This research was also done in colaboration with the proteomics group of the Cancer Research Centre leaded by Dr. Manuel Fuentes. The raw datasets are available via ProteomeXchange with identifier PXD005997. In this Thesis the datasets generated in this work are used.

3.4 Methodoly and technical environment

The following sections describe the methodologies and technical environment that we used for the development of Path2enet and for the integrative analyses of proteogenomic data. Section 3.4.1 describes the bioinformatic platform of Path2enet in detail and explains the network analysis the tool performs. Sections 3.4.2 and 3.4.3 focus on the bioinformatic methods and environment which we used to process the transcriptomic datasets (i.e. microarrays and RNA-Seq). In the results section we discuss how we used these datasets in detail in the Path2enet tool and in the proteogenomic studies.

3.4.1 Technical environment of the Path2enet tool

3.4.1.1 The general environment

At the beginning of the development of the *Path2enet tool* the *R-version "3.2.0"* was used. Later in the development process the newer *R-version "3.3.0"* has been used. The *Path2enet tool* is compatible with both *R*-versions. *R* was installed on the "*Unix-type"* operating systems *"Centos 6.5"* and *"Centos 7"*.

3.4.1.2 R and Bioconductor environment

R is a free software environment for statistical computing and graphics (R-cran, https://www.r-project.org/). The R environment is a suite of integrated of software for data manipulation, calculation and graphical display. R was developed by Ross Ihaka and Robert Gentleman, at the University of Auckland (New Zealand), being the first stable beta version relased in 2000. It is based on the programming language “S”, which is an easy-to-learn and a well-developed programming language, created by Rick Becker, John Chambers and Allan Wilks at Bell Laboratories (**https://cran.r-project.org/)**.

R is particularly suitable for analyzing large sets of data and for visualizing the results in a clearly arranged way. It is an open source software and has an open development environment. It is licensed under the “General Public License - Version 2” and therefore it is free of charge. This is the reason why many scientists use R for their statistical data analysis. The R environment is a dynamic environment because many people offer their written software solutions, called “packages” or “libraries” to the community. The R environment can be used on operation systems based on “UNIX”, “Windows” and “Mac”.

Bioconductor is an initiative for the collaborative creation of extensible software for computational biology and bioinformatics **(Gentleman, Carey et al. 2004).** The software is free of charge and provides packages for the analysis of microarrays, high throughput assays, sequence data and many packages for annotation. The project was created because many biological processes are of computational nature, and the use of computational and statistical models are a key way to understand them. The analysis of high-throughput large-scale biomolecular data (“omic data”) also needs strong computational and statistical power. The intention of Bioconductor is to improve the collaboration of scientists by creating an environment that meets “conceptual, computational and inferential challenges” **(Gentleman, Carey et al. 2004).** An open and shared programming environment is also a guarantee for a better reproducibility of the statistical analysis of complex data created in many biology experimental studies.

3.4.1.3 The basic R-libraries and R-packages of the Path2enet tool

The R-packages and libraries which are essential for the Path2enet tool are listed in **Table 2**. The packages XML, RMySQL and DBI are needed to generate the Kegg2MySQL-database. The packages BioIDMapper, Rcurl and org.Hs.eg.db are annotation packages. RGtk, gWidgets and gWidgetsRGtk2 are needed to visualize the networks. igraph is a package that provides many functions to analyze and visualize networks. The igraph-object is a standard to store networks and network attributes. The author transferred the data of the databases into igraph-object. This ensures that existing packages can use the created networks and analyse and visualize them.

Table 2: Packages and libraries of "R" and "Bioconductor" essential for the created R-functions

Package	Description	Version
XML	Collection of informatics tools to parse *"XML"* files. development stage Important to parse the *"KGML"* files provided by *"KEGG"*	XML_2.6-0 (19)
RMySQL	Packages to interact between *"R"* and *MySQL*	R*MySQL*_0.7-4 (20)
DBI		DBI_0.2-5 (21)
gWidgetsRGtk2	Collection of libraries in order to build graphical user interfaces. The libraries are important to visualize the created graphs.	gWidgetsRGtk2_0.0-69 (22)
gWidgets		gWidgets_0.0-41 (23)
RGtk2		RGtk2_2.12.18 (24)
Igraph	Functions that are needed to analyze and visualize graphs. Creates and uses the *igraph-objects*.	igraph_0.5.5 (25)
BioIDMapper	Libraries needed to translate *KEGGIDs* into *"gene symbol"* and *UniprotIDs*.	BioIDMapper_2.4 (26)
RCurl		Rcurl_1.4-3 (27)
org.Hs.eg.db		org.Hs.eg.db_2.3.6 (28)

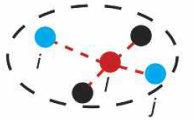
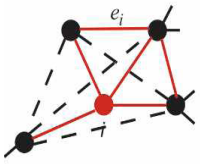
3.4.1.4 MySQL

MySQL software is very useful to store and access large content of data. It is a robust "Structured Query Language" [SQL] database server. It is an open source software, licensed under the terms of the "General Public License" (GPL) (GPL). Therefore, it can be used by many public users without paying license fees.

The data volume of the data included in APID, KEGG and UNIGENE is large. MySQL is designed to handle large datasets in a structures way. In this way, we integrated all the data from APID, KEGG and UNIGENE in a MySLQ database. The Path2enet tool is designed to access these databases instead of storing the raw original data. The applied version of the MySQL client is "MySQL-client-5.1" and the version of the installed MySQL-server version is "MySQL-serveR-5.1". The reference to the used MySQL syntax is "MySQL 5.1 Reference Manual"(MySQL5.1.Reference Manual).

3.4.1.5 Network parameters

Network parameters are the result of the statistical analysis of the topological characteristics of a given network. The Path2enet tool calculates the network parameters in order to identify key-players in large networks. Small networks may be analyzed by discussing a graphical view, but biological researchers have to analyze large networks statistically in order to extract the relevant information derived from the network graphs. For the analysis of networks the Path2enet tool treats the edges in the network as undirected. Four parameters are calculated for the network graphs: "Degree", "Betweenness", "Clustering Coefficient" and "Eigenvector". These parameters are explained in **Table 3**.

Parameter	Image	Formula	Description
Degree		k_i = number of links connected to node i	The number of links connected to one vertex is defined as its degree. (Zhu, Gerstein et al. 2007).
Betweenness		$b_l = \sum_{ij} p_{ij}(l)/p_{ij}$	Betweenness is the fraction of the shortest paths between all pairs of vertices that pass through one vertex or link. Betweenness estimates the traffic load through one node or link assuming that the information flows over a network primarily following the shortest available paths. **(Zhu, Gerstein et al. 2007**).
		p_{ij}: number of shortest paths between i and j **$p_{ij}(l)$:** numer of shortest paths between l and j going through node l	
Clustering Coefficient		$c_i = \frac{2e_i}{k_i(k_i-1)}$	The clustering coefficient of one vertex can be calculated as the number of links between the vertices within its neighborhood divided by the number of links that are possible between them. **(Zhu, Gerstein et al. 2007)**
		e_i: number of existing links (labeled in red) among the k_i nodes that connect to node i	

Eigenvector		$x_i = \frac{1}{\lambda}\sum_{j=1}^{n} A_{ij}x_j,$ x_i: eigenvector centrality of node i A: adjacency matrix with an edge between vertices i and j x_j: centrality of node j λ: is a constant (31)	Eigenvector centrality scores correspond to the values of the first eigenvector of the graph adjacency matrix; these scores may, in turn, be interpreted as arising from a reciprocal process in which the centrality of each actor is proportional to the sum of the centralities of those actors to whom he or she is connected. In general, vertices with high eigenvector centralities are those which are connected to many other vertices which are, in turn, connected to many others (and so on) (32)

Table 3: Explanations of parameters calculated by igraph (32).

3.4.2 Processing of the microarray datasets

The pre-processing, normalization and signal calculation of these data were done using the Bioconductor packages affy, frma, oligo29 and pd.hugene.1.0.st.v130. The table 4 explains the packages in detail.

Table 4: *The table describes the packages used for the microarray analysis*

Package	Description	Version
affy **(Gautier et al. 2004)**	Processing and analyzing of oligo nucleotide arrays.	affy_1.54.0 (33)
frma **(McCallet al. 2010)**	This package allows preprocessing and analyzing of singe microarrays and microarray batches. It contains the function barcode with the gene expression barcode algorithm.	frma_1.28.0 (34)
barcode **(McCall et al. 2011)**	Calculation of the ON/OFF state of a probe set in a microarray platform based on a calculated and approved expression level. Section 3.4.2.2 explains this method more in detail.	frma_1.28.0 (35)
hgu133afrmavecs	frma uses this vectors for microarrays of type hgu1331 and hgu133plus2. These vectors are necessary for the barcode algorithm. frmaTools version 1.19.3 and hgu133ahsentrezgcdf and hgu133plus2entrezgcdf version 19.0.0	hgu133afrmavecs_1.5.0 (36)
hgu133plus2frmavecs		hgu133plus2frmavecs_1.5.0 (37)
pd.hugene.1.0.st.v1 **(Carvalho 2015)**	I used these packages for the analysis of the microarray datasets of the *Ramos* cell line in the result section 3.3.	pd.hugene.1.0.st.v1_3.14.1 (38)
oligo **(Carvalho and Irizarry 2010)**	PD hugene: Platform Design Info for Affymetrix HuGene-1_0-st-v1 Oligo: Preprocessing of Oligonucleotide Microarray	oligo_1.40.1 (39)

3.4.3 Processing of the RNA-Seq datasets

We used following steps to proceed the RNA-Seq datasets:

1. Prefetch SRR387395 run from SRA database with SRA tools **(Leinonen R, et al. 2011)**

2. Convert SRA file to paired end fastq files with SRA tools
3. Trimming with Trimmomatic **(Bolger, Lohse and Usadel 2014).** Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170. performing the following:
    a. Remove adapters (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10)
    b. Remove leading low quality or N bases (below quality 3) (LEADING:3)
    c. Remove trailing low quality or N bases (below quality 3) (TRAILING:3)
    d. can the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15)
    e. Drop reads below the 36 bases long (MINLEN:36)
4. Align to EnsEMBL GRCh37 genome with STAR **(Dobin et. al. 2013)**
5. Creating, sorting and indexing BAM file with SAMtools **(Li et al. 2009)**
6. Obtaining FPKM for genes and isoforms with CuffLinks **(Trapnell et. al. 2010)**

4 RESULTS

The result section is divided in three major parts. The first part (4.1) describes the processing and preparing of the data to build the biomolecular networks with the Path2enet tool and how the tool accesses the generated databases and adapts the data sources to the needs of the user. The second part (4.2) explains how the Path2enet tool uses the information to create, analyse and visualize the biomolecular networks and its performance on an experimental dataset of human B- and T-cells. The third part (4.3) shows the results of the proteogenomic analysis of the Burkitt-Lymphoma *Ramos* B-cell line. The last section (4.4) shows the results of the quantitative proteomic analysis of B-cells isolated from five Chronic Lymphocytic Leukemia (CLL) patients.

4.1 Processing of biological data to build the biomolecular networks

The Path2enet tool uses network data and pre-processed transcriptomic data to generate the biomolecular networks. Section 4.1.1 explains how the identifier (ID) mapping of Path2enet works and which datasets uses for its ID mapping functions. Section 4.1.2 shows how the Path2enet processes (i) the protein-protein physical interaction data of APID and (ii) the pathway data of KEGG PATHWAY to build the MySQL-databases to store this information. It also explains how the user can access theses databases via R. Section 4.1.3 describes the pre-processed transcriptomic data included and how the users can explore and adapt these data to their needs.

4.1.1 ID mapping

A key step to achieve an adequate integration of biological data from different sources is the use of one common identifier to provide the mapping of all datasets. Unfortunately, this is not frequently the case because many databases use different identifiers to store its information. Therefore, in this work we used several tools for mapping identifiers [ID-Mapping] to translate/map the major identifier of every used database to one selected identifier, that was the *UniProt Accession* [UniProt ID] (UniProt, 2017). We have chosen this identifier, because in the case of proteins the UniProt database **(i)** provides high quality mapping tables and tools; **(ii)** it shows if a protein has been manually “reviewed” or not (indicated with a label entry in the UniProt Swissprot database); **(iii)** it is the key identifier used in APID and **(iv)** this identifier is the most highly used in proteomic research and proteomic resources.

The function *IDmappingFunction* of the Path2enet tool generates the mapping tables needed: **(i)** in the function KeggXML2SQLDatabase to generate the KeggMySQL database inside Path2enet, **(ii)** to create the preprocessed transcriptomic datasets; **(iii)** to integrate the experimental data to the networks and **(iv)** to build the biomolecular networks.

4.1.1.1 Generating ID mapping for KeggXML2SQLDatabase and EST dataset

Path2enet package generates an ID mapping table to map ENTREZ Gene IDs (identifiers) or KEGG IDs to UniProt IDs (i.e. UniProt *Accession* and UniProt *Entry Name*, which for example for KRAS are: "P01116" and "RASK_HUMAN", respectively). To map the identifiers the *IDmappingFunction* obtains and processes the ID mapping table that UniProt database provides (via its ftp server) (Uniprot Database). In this way we generate a table called *referenceTable*. This *referenceTable* is important because it links the identifiers KEGG PATHWAY uses [Kegg-ID] with the referred UniProt IDs. This step is necessary to generate the databases and to integrate the different data sources with common primary IDs. In this way, the final and unique key identifiers of the generated networks are the UniProt IDs, to allow a correct integration and comparison of different data.

In order to create the *referenceTable* for building the KeggMySQL database and the Unigene EST dataset for *Homo sapiens* the user only has to input "HUMAN" as organism and the directory to store the data files.

```
tmp <- tempdir()

referenceTable <- IDmappingFunction(directory=tmp, org="HUMAN")
```

For the specific organisms we used the same abbrevations as UniProt. UniProt provides ID mapping tables for ARATH, CAEEL, CHICK, DANRE, DICDI, DROME, ECOLI, HUMAN, MOUSE, RAT, SCHIPO and YEAST.

Path2enet tool at present only includes the ID mapping table for human. The last mapping included in the tool was generated in May 2016.

```
data(referenceTable)
```

UniprotID	GeneID	KEGG	UniProtKB.ID	UniGene
A0A183	448835	hsa:448835	LCE6A_HUMAN	Hs.62927
A0A5E8	10634	hsa:10634	A0A5E8_HUMAN	Hs.322852
A0AUZ9	151050	hsa:151050	KAL1L_HUMAN	Hs.282260
A0AUZ9	151050	hsa:151050	KAL1L_HUMAN	Hs.591638
A0AV02	84561	hsa:84561	S12A8_HUMAN	Hs.658514
A0AV05	340385	hsa:340385	A0AV05_HUMAN	Hs.521942
A0AV47	11073	hsa:11073	A0AV47_HUMAN	Hs.593379

Figure 16. *Small view of the data object referenceTable included in the Path2enet.*

The lines in grey boxes above show some of R code used to generate the *referenceTable* and **Figure 15** shows a brief view of such table presenting as an example the ID mapping for 6 human proteins mapped to 6 UniProt IDs and 6 KEGG IDs, but to 7 ENTREZ Gene IDs and 7 UniGene IDs. In this way, it shows that the mapping from gene to proteins, and vice versa, is not unique.

Table 5 Dimensions and number of unique identifiers per database identifier for Homo sapiens.

Identifier (given name in the *referenceTable*)	**Unique Entries**
UniProt Accession (UniProtID)	133,808
Entrez Gene IDs (GeneID)	19,204
Kegg IDs (KEGG)	19,126
UniProt Knowledgebase IDs (UniProtKB.ID)	13,3808
UniGene IDs (UniGene)	22,205
TOTAL dimension of the *referenceTable*	**157,356 x 5**

4.1.1.2 Generating ID mapping for the transcriptomic datasets

For the mapping of the RNA-Seq and microarray datasets to UniProt IDs, three other identifiers are necessary: ENSEMBL GeneID (Ensembl) and the identifier of the microarray probe sets of the Affymetrix platforms HGU133A and HGU133Plus2 (Affy). The Barcode algorithm supports these two platforms for human.

The annotation of microarray is complicated, because probe sets can be ambiguous. Ambiguous probe sets are misleading and the Path2enet tool shall remove them in the mapping process. Therefore, the function uses the chip definition file from the BRAINARRAY tool that maps the probe sets from the Affymetrix HGU133 Plus 2.0 array to ENSEMBL Gene IDs (CDF file named “HGU133Plus2.HS.ENSG”) and Affymetrix HGU133A (CDF file named “HGU133A_Hs_ENSG”) to ENSEMBL Gene IDs. Theses CDFs do not include ambiguous probes sets. In this work I use the 20th version of the BRAINARRAY tool (41).

In the next step the tool maps the ENSEMBL Gene IDs to UniProt Swissprot identifiers in order to include only reviewed proteins. The user has to download the UniProt_sprot.fasta.gz (41) that that UniProt provides on his ftp server.

```
IDmappingFunction(org = “HUMAN”, brainarray=”path/to/CDF/file”,
uniprot=”path/to/uniport_sprot.fasta.gz”)
```

The Path2enet tool includes both ID mapping tables for the HGU133A and HGU133Plus2 platforms: *referenceTableHGU133A* and *referenceTableHGU133Plus2*

```
data(referenceTableHGU133A)

data(referenceTableHGU133Plus2)
```

Table 6. Number of unique identifiers in: referenceTableHGU133A and referenceTableHGU133Plus2.

Identifier (given name in the *Tables*)	**referenceTableHGU133A**	**referenceTableHGU133Plus2**
UniProt Accession (UniProtID)	20,198	20,198
Entrez Gene IDs (GeneID)	19,062	19,062
Kegg IDs (KEGG)	18,556	18,556
UniProt Knowledgeb.ID (UniProtKB.ID)	22,114	22,114
Unigene IDs (UniGene)	20,198	20,198
ENSEMBL IDs (Ensembl)	20,579	20,579
Affymetrix IDs (Affy)	17,474	32,219
TOTAL dimensions of the tables	**491,489 x 7**	**663,000 x 7**

4.1.2 Setting up the network data included in Path2enet tool

Path2enet tool uses the KEGG PATHWAY and APID to generate its networks. This paragraph describes how Path2enet tool integrates these datasets in a MySQL-environment and how the user can access this data via R.

4.1.2.1 Setting up the MySQL database

Path2enet package builds and uses a MySQL relational database to store and later access the data and information needed from the pathways. It uses a relational database SQL because: **(i)** the tool should be able to handle large data volumes; and **(ii)** it should have a quick response to allow external access.

Therefore, it is necessary to setup in the computer a MySQL database or to have access to an already installed SQL database. Further information and instructions about how to install and configure a MySQL database can be found at https://www.mysql.com/. The user need to have the privileges to create and access databases on the MySQL server.

4.1.2.2 APID

Section 3.1.1.2 describes the APID protein interaction meta-database. The Path2enet tool uses this data to have access to protein-protein physical interactions which are experimentally proven. The function *Apid2Sql* allows to transfer the data of APID to an MySQL server. APID provides the datasets on its website (43). The user has to download the interactome and enter the file destination to the function.

```
Apid2Sql(user="USER", host="IP/localhost", driver="MySQL",password="PW",
file=”path/to/interactome”, database="PPI")
```

Path2enet package contains the PPI of human. This data can be easily used to create the database.

```
data(ppiHumanUniP)

Apid2Sql(user="USER", host="IP/localhost", driver="MySQL",password="PW",
ppi=ppiHumanUniP, database="PPI")
```

4.1.2.3 KEGG

One objective of Path2enet tool is to store data obtained by KGML files provided by KEGG in a MySQL database. This database will be the basis for creating networks of protein relations in the R environment. In order to allow easy usage of the database the data should be accessible from R environment via a user friendly R function. The function should also enable that only certain pathway names or gene names are entered to obtain the corresponding data and information stored in the MySQL database. All this part of the Path2enet tool is written with the R language following MySQL syntax.

We designed Path2enet to work independently from their source databases and from other tools. Therefore, it was necessary to write a parser for the KEGG KGML files. The parser collects the data of KGML files, maps the KEGG IDs to UniProt IDs, stores them into R dataframes, which are then used to generate an independent new "pathways SQL database" that we called: *Kegg2MySQL / KeggSQL* (i.e. *Kegg2MySQL* was called in the first version of the tool, but know for simplicity we renamed it *KeggSQL*). The datasets chosen in our study were non-metabolic and metabolic KGML files for *Homo sapiens*. We focused the data collection on extracting all protein-relations which are stored in the KGML files.

In order to use the advantages of a database relational environment the Path2enet tool transfers the data of KGML and BR files provided by KEGG into the created pathways SQL database. Section 4.1.2.3.1 describes the function *KeggXML2SqlDatabase* of the Path2enet tool which fulfils this purpose and to do so: **(i)** it downloads and collects the hierarchical data of the KEGG BR files; **(ii)** downloads and collects the KEGG KGML files; **(iii)** it applies the mapping of KEGG IDs and UniPro IDs to identify each pathway and each protein.

4.1.2.3.1 Structure of the data KEGG PATHWAY provides

As indicated above, the function which builds the pathways database inside the Path2enet tool is called *KeggXML2SqlDatabase*. The first step is to download the KEGG BR files (br8901.keg) from the website: http://www.kegg.jp/kegg-bin/get_htext?br08901.keg . The number in the name brXXXX.keg is the version of the data file and changes within KEGG updates. The function automatically updates the version number.

```
A<b>Metabolism</b>
B  Global and overview maps
C    01100  Metabolic pathways
C    01110  Biosynthesis of secondary metabolites
C    01120  Microbial metabolism in diverse environments
C    01130  Biosynthesis of antibiotics
C    01200  Carbon metabolism
C    01210  2-Oxocarboxylic acid metabolism
C    01212  Fatty acid metabolism
C    01230  Biosynthesis of amino acids
C    01220  Degradation of aromatic compounds
B  Carbohydrate metabolism
```

Figure 17: *Category Metabolism of the KeggBR file.*

The KEGG BR files contain the hierarchy of the database. It has three levels or categories: [A], [B], and [C]. **Figure 17** shows a small part of one of these files. The [A] level shows the major category of the pathways like: Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Organismal Systems, Human Diseases and Drug Development. The [B] level includes 58 entries, for example: the Carbohydrate metabolism, Transcription, Translation, Signal transduction, Cancer, etc. The pathways are properly included in level [C] that contains their identifiers and specific names (in total 511 entries).

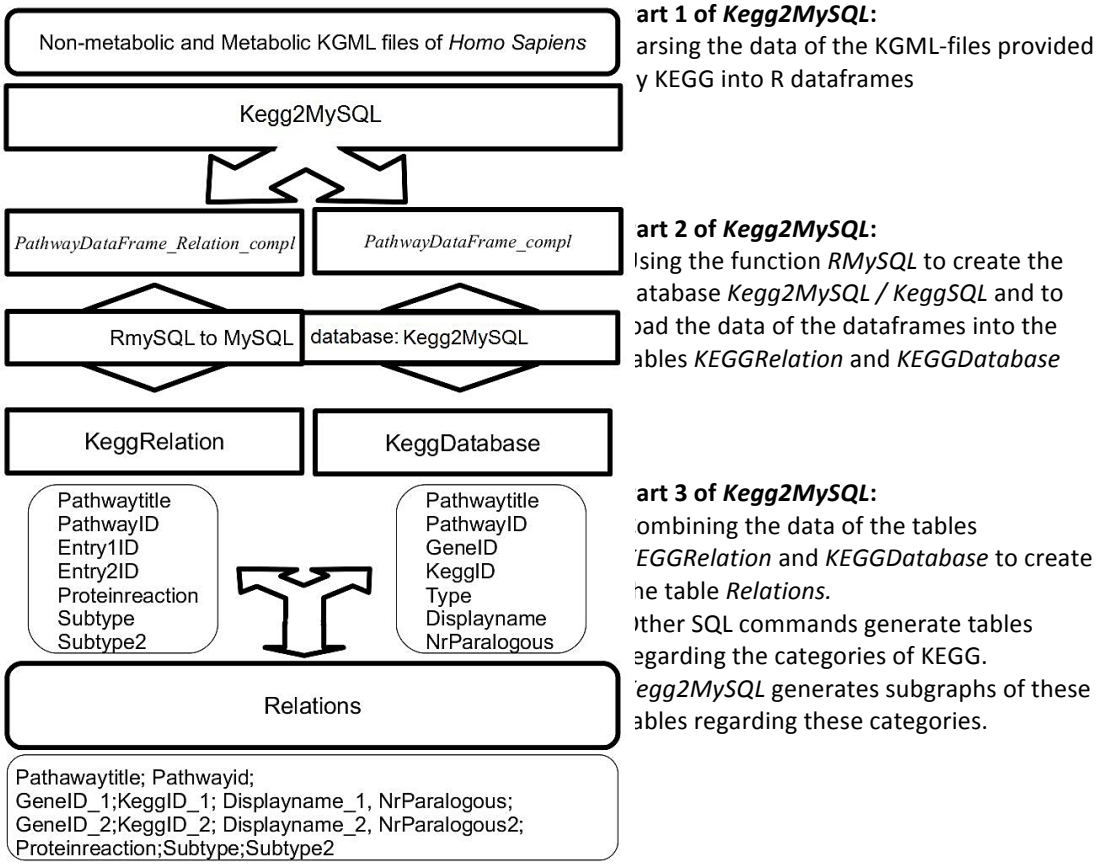


***Figure 18**. Overview of the creation of the KeggXML2SqlDatabase database. From KGML files until the table Relations. The function is divided in three parts, which are explained in the second column.*

The use of the KEGG identifier is needed to download a KGML file. To do so, Path2enet tool gets the information of the KEGG website. For example, the JAK-STAT signaling pathway has the ID: **hsa**04630, and can be downloaded using this URL: http://www.kegg.jp/kegg-bin/download?entry=**hsa**04630&format=kgml. The part highlighted in **bold** is the tri-letter-code that KEGG uses to differentiate organisms: **hsa** = *Homo sapiens*. Five numbers ID 04630 allow to identify the specific pathway and are also used to get KEGG BR file.

4.1.2.3.2 Generating KEGG database inside Path2enet

The function *KeggXML2SqlDatabase*, first, downloads the KEGG KGML files and stores the data in dataframes in R. Second, using the function *RMySQL* transfers these dataframes into a MySQL database called: *Kegg2MySQL / KeggSQL*. In this SQL database, initially it creates 2 tables: KEGGRelation and KEGGDatabase; that then are merged in only one major master table called *Relations*. This is the basis for all other tables created afterwards with specific queries.

- ***KeggXML2SqlDatabase* parser**

More in detail, ¡the first part of function *KeggXML2SqlDatabase* is a parser, which is a tool that collects data of a XML document and stores it in an object that can be used by other functions. It collects the data of the KGML files provided by KEGG and stores this data into 2 dataframes, called: "PathwayDataFrame_compl" and "PathwayDataFrame_Relation_compl".

The dataframe "PathwayDataFrame_compl" contains information of an entry section in a KGML file (see **Figure 16**). The entry section in a KGML file (for any file included in KEGG PATHWAYS) provides the following attributes: Pathwaytitle, PathwayID, KEGGID, EntryID, Type and Displayname. The attribute GeneID is the KEGGID (corresponding to a gene entry) mapped to the UniProtID with the ID mapping function. It is created in order to translate the gene KEGG identifiers into UniProt IDs. **Table 7** describes all the attributes providing examples.

Table 7: *Overview of the data stored in the dataframe "PathwayDataFrame_compl*

Attribute	Description	Example
Pathwaytitle	The title of the pathway in KEGG	"Notch Signaling Pathway"
PathwayID	The identifier of the pathway in KEGG	"path:hsa04330"
KEGGID	The identifier of an entry in KEGG	"hsa:6868"
GeneID	This column stores the UniProt IDs corresponding to the gene entries that KEGGID provides NA is used, if it is not possible to translate the KEGGID	"hsa:6868" → "ADM17_HUMAN" "ko:K04497" → "NA"
EntryID	This is the reference of an entry in the pathway	"17 "
Type	Explains what type of entry it is	"gene"
Displayname	A name presented by KEGG on its original image of a pathway.	"ADAM17"

The dataframe "PathwayDataFrame_Relation_compl" stores the following attributes: Pathwaytitle, PathwayID, Entry1ID, Entry2ID, Proteinreaction, Subtype and Subtype2. These attributes are important for explaining the relation between the entries of a pathway. **Table 8** describes the attributes and provides examples for each.

Table 8: *Overview of the data stored in the dataframe "PathwayDataFrame_Relation_compl"*

Attribute	Description	Example
Pathwaytitle	The title of the pathway in KEGG	"Chemokine signaling pathway"
PathwayID	The identifier of the pathway in KEGG	"path:hsa05146"
Entry1ID	The entry, with the reference number 53 in the KGML-file, has a relation with the entry with the reference number 56 of the pathway.	[entry1=]"53"
Entry2ID	See above "Entry1ID"	[entry2=]"56"
Proteinreaction	This attribute explains which kind of relation entry number 53 and 56 have.	"PPrel"
Subtype	This explains the relation more precisely.	"activation"
Subtype2	This column stores the type of the relation, too. If the type of relations is fully explained by Subtype the entry is marked as NA	"indirect effect"

- **Building of the *Kegg2MySQL* database**

On the basis of the two dataframes described, the second part of *KeggXML2SqlDatabase* creates the SQL database. To do so it uses *RMySQL*: **(i)** to access the data; **(ii)** to transfer the data from R to MySQL; and **(iii)** to generate or modify the database.

As indicated above, the *Kegg2MySQL / KeggSQL* database has 2 main tables: **(i)** *KEGGRelation* which stores the data of "PathwayDataFrame_Relation_compl"; and **(ii)** *KEGGDatabase* which stores the data of "PathwayDataFrame_compl". Other tables are created by SQL commands to fulfill the following objectives: **(i)** provide the data of the KGML files in a reasonable form; **(ii)** secure that large data volumes can be handled; **(iii)** divide the data into categories.

All tables in the database are indexed, which reduces the time needed to build the database structure and to process and interrogate the data afterwards. Biological databases can be quite large so that it is advantageous to use a program like MySQL, which is specialized in storing, processing and interrogating huge data volumes.

- **Structure of the *Kegg2MySQL* database**

The main data included in the *Kegg2MySQL* database are the non-metabolic and metabolic KGML files of *Homo sapiens*.

The table *KEGGDatabase* includes another column, which has its origin not in KEGG. The column is called "NrParalogous" and shows the number of entries, which share the same EntryID with a pathway in KEGG PATHWAYS. Entries which share the same EntryID with a KGML file are presented by the same Displayname in the pathways of KEGG. For example HRAS, KRAS and NRAS share the same EntryID(=35) with the pathway "Bladder cancer" in *Homo sapiens*. Therefore, in column "NrParalogous" of the table *KEGGDatabase*, these 3 proteins have value 3: "NrParalogous"(=3). In the graphical view of the KEGG pathway "Bladder cancer" from *Homo sapiens*, these 3 proteins (HRAS, KRAS and NRAS) are presented by the same Displayname(=RAS).(KEGG Pathway hsa).

A final table generated, called *Relations*, is a mixture of the tables *KEGGRelation* and *KEGGDatabase*. *KEGGDatabase* provides the information of the "protein/gene entries" of the KGML files and *KEGGRelation* provides the information of the "relations or links" of the KGML files (see **Table 9).**

Table 9: *Explanation of the attributes of the table Relations in the Kegg2MySQL database.*

Attributes	**Description**	**Example**
Pathwaytitle	The title of the pathway in KEGG	Acute myeloid leukemia
Pathwayid	The identifier of the pathway in KEGG	path:hsa05221
GeneID_1	The column stores the "gene symbol" of the first EntryID. NA is used, if it is not possible to translate the KEGGID	AKT1_HUMAN
*KEGG*ID_1	This is the KEGG-identifier of the first EntryID, which is related with the second EntryID.	hsa:207
Displayname_1	This is the placeholder of the entries stored in "EntryID_1" in graphical "KEGG" pathways.	AKT3, DKFZp434N0250, PKB-GAMMA, PKBG, PRKBG, RAC-P

NrParalogous	Number of how many entries share the first EntryID with the "Acute myeloid leukemia" pathway.	3
GeneID_2	This is the gene symbol of the entry that is related to the first entry.	CHUK_HUMAN
*KEGG*ID_2	KEGG-identifier of the second EntryID	hsa:1147
Displayname_2	Placeholders that could represent the second entry in a graphical KEGG pathway.	CHUK, IKBKA, IKK-alpha, IKK1, IKKA, NFKBIKA, TCF16
NrParalogous2	Number of how many entries shares the second EntryID with for example "Acute myeloid leukemia" pathway.	3
Proteinreaction	This attribute explains which kind of relation "AKT1" and "CHUK" have.	PPrel
Subtype	This explains the relation more precisely.	activation
Subtype2	This parameter describes the relation more in detail.	binding/association

Besides the hierarchy the KEGG BR file holds, Path2enet creates some specific meta-pathways. These are marked with prefix "Path2enet_*" generating also two tables that contain the information similar to *KeggDatabase* and *KeggRelations*.

These tables have stored the same attributes as the *Relations* table. The tables without the suffix "*MoreInfo" do not have the columns Pathwaytitle and PathwayID. This reduction is necessary because it reduces the repetition of interactions provided by different pathways.

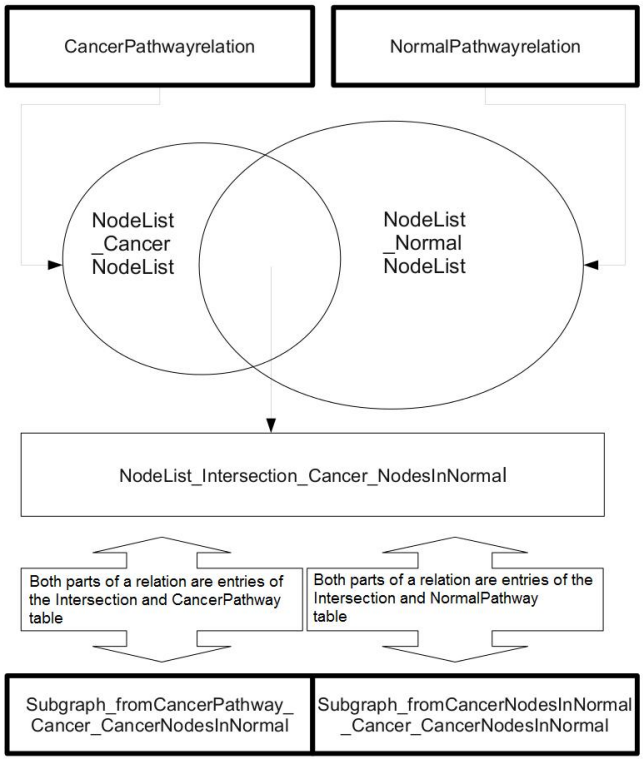


Figure 19: *Network subgraphs derived from CancerPathwayrelation and NormalPathwayrelation*

- **Comparative study of the networks derived from normal- or cancer-pathways**

Using Path2enet we generated two big network datasets reading-out all the human KEGG files

included in normal (i.e. not disease) pathways or in cancer associated pathways. These datasets were called: *NormalPathwayrelation* (which included 47,119 items) and *CancerPathwayrelation* (which included 2,535 items). We calculated the common nodes (i.e. the human proteins that are included in both datasets), generating the list of nodes in the intersection: NodeList_Intersection (see **Figure 19**). Finally, we create two networks (called "subgraphs" in our tool) that can be analysed and compared to find out which are the differences in the interaction patterns of the proteins and in the topological parameters of the network subgraphs. We called these networks:

– Subgraph_fromCancerPathway_Cancer_CancerNodesInNormal

– Subgraph_fromCancerNodesInNormal_Cancer_CancerNodesInNormal

4.1.2.4 Access to the database via R

The *searchFunction* of the Path2enet tool is written for users. It provides an easy way to interrogate information out of the *Kegg2MySQL* database. If the *Kegg2MySQL* database is hosted on a open server, it is possible to access information of this database without hosting the database on a local machine. The *searchFunction* depends on the *RMySQL* and the *DBI* packages. The *searchFunction* can be divided into three parts: **(i)** a connection to the database is established; **(ii)** a query to the database is created; and **(iii)** an output of the database is delivered to R. The parameters which can be modified by the user are explained in **Table 10**.

Table 10: *Overview of the parameters of the R searchFunction.*

Parameter	Explanation	Example
dbDriver	This parameter must have the name of the database environment in which the database is stored.	MySQL
dbUser	Name of a user, who has the right to access the database created with the script *KeggXML2SqlDatabase* .	Kegg2MySQL
dbHost	The user has to enter the ip address of the host of the database or localhost.	10.10.10.7
dbName	This input must have the name of the database that is created with *KeggXML2SqlDatabase*.	Kegg2MySQL
Password	The user password of the MySQL database has to be entered.	Kegg2MySQL
dbData	The table of the database, which contains the information the user wants to have.	Relations
GeneID	Information about one or more entries, which have the entered "gene symbol". Searches with more than one entry need the separation of the names by an "," or use a list of gene symbols.	"NRAS" "NRAS, KRAS"
KEGGID	Parameter similar to GeneID but used for KEGG identifiers	"hsa:4893 " "hsa:4893, hsa:3845"
PathwayID	Parameter for KEGG pathway identifier"	"path:hsa05221"
Pathwaytitle	Parameter for "pathway title".	"Acute myeloid leukemia"
Subtype	Parameter for Subtype.	"activation"
Proteinreaction	This parameter allows to filter the protein reaction of two entries.	"PPrel"

Parameter	Explanation	Example
selectGeneID	If this parameter is true, the output of the query has only the Uniprot-IDs names.	"FALSE" or "TRUE"
selectDisplayname	Similar to selectGeneID, but the output is the placeholder which KEGG assigns to the genes.	"FALSE" or "TRUE"
selectKEGGID	Similar to selectGeneID, but the output are KEGG identifiers	"FALSE" or "TRUE"
selectPathwaytitle	Similar to selectGeneID, but the output are KEGG identifiers	"FALSE" or "TRUE"
selectALL	If this parameter is set TRUE all information of the query will be used as output.	"FALSE" or "TRUE"
selectFREE	This parameter allows to select the columns of the output table individually.	"FALSE" or "GeneID, *KEGG*ID"
Relations	Parameter must be TRUE if the used table provides data of relations between entries, for example Relations.	"FALSE" or "TRUE"
noNA	This is a filter. If this parameter is TRUE, no entries that have a NA as GeneID or GeneID_1/2 will be presented as output.	"TRUE" or "FALSE"

The user can search with KEGG identifiers, pathway identifiers (like for example "hsa04330"), pathway titles or UniProt identifiers. We show below some ways of doing these searches.

```
kegglist <- c('hsa:11317' , 'hsa:5986' , 'hsa:51107' , 'hsa:55851' , 'hsa:5663' , 'hsa:5664' ,
'hsa:9541' , 'hsa:55534' , 'hsa:84441' , 'hsa:9794' , 'hsa:1487' , 'hsa:1488', 'hsa:1488',
'hsa:3065', 'hsa:3066' )

searchFunction(dbDriver="MySQL", dbUser="USERNAME", dbHost="IP/localhost",
dbName="KeggSQL", password="PW", dbData="Path2net_KeggDatabase", KeggID=kegglist,
selectALL='TRUE',  Local="FALSE")
```

```
genelist <- c( "4EBP1_HUMAN", "AAPK1_HUMAN" ,"AAPK2_HUMAN" ,"AKT1_HUMAN" ,
"AKT2_HUMAN" , "AKT3_HUMAN" , "AKTS1_HUMAN" ,"I4E1B_HUMAN" ,"IF4E2_HUMAN"
,"IF4E_HUMAN" , "MTOR_HUMAN",  "TSC2_HUMAN" , "P55G_HUMAN" ,"P85A_HUMAN")

signal.compl <- searchFunction(dbDriver="MySQL",  dbUser="USERNAME",
dbHost="IP/localhost", dbName="KeggSQL", password="PW",
dbData="Path2net_SignalPathwayrelation", print="FALSE" , GeneID=genelist ,
Relations="TRUE")

#Select only interactions

signal.geneID <- searchFunction(dbDriver="MySQL",  dbUser="USERNAME",
dbHost="IP/localhost", dbName="KeggSQL", password="PW",
dbData="Path2net_SignalPathwayrelation", print="FALSE" , GeneID=genelist ,
Relations="TRUE", selectGeneID="TRUE")
```

4.1.3 Processing the transcriptomic datasets

Data about proteins in APID and KEGG are tissue or cell-type unspecific. But most biologists conduct research on specific cell or tissue samples, because they are interested in information closely related to some specific biological context. To create specific protein-protein networks we made the assumption that each gene encoding for a protein in a protein-network has to be found as expressed in a specific tissue or cell type.

The Path2enet tool uses three types of *a priori* processed transcriptomic data to calculate such expression levels of specific proteins: **(i)** Expression Sequence Tags; **(ii)** Expression Microarrays; and **(iii)** RNA-Seq. This section describes the processing of the transcriptomic datasets and how the user can adapt the datasets for their needs.

4.1.3.1 EST data

```
data(estTissueUniP)
```

The UniGene database stores information on Expression Sequencence Tags (ESTs). The EST data give an overview about which genes are expressed in specific human tissues and cell types. If one specific gene does not have any EST in one specific tissue it indicates that it is not present in such tissue.

```
mkESTdb(dbDriver = "MySQL", dbUser = "root", dbHost = "localhost", password = NULL,
directory = tempdir(), dbData = "Unigene", ESTdataexits = FALSE, name = "Homo_sapiens",
species = "Hs.")
```

The implemented function *mkESTdb* downloads and generates the information in a new database in MySQL. The parameters to create the database are the same as for the *KeggXML2SqlDatabase* function**.** The name of the species has to fit the same that name the UniGene database and species input has (using the same abbreviation that the UniGene uses). Website ftp://ftp.ncbi.nih.gov/repository/UniGene/ shows the available datasets.

Table 11: *Short overview of the information provided by the "Unigene-MySQL-Database".*

Unigene	Tissue/Development-stage	ESTs
Hs.100043	Blood	2
Hs.100043	Brain	8
Hs.100043	embryonic_tissue	2
Hs.100043	Eye	17
Hs.100043	Heart	2
Hs.100043	Kidney	2
Hs.100043	Liver	1
Hs.100043	Lung	3
Hs.100043	lymph_node	6
Hs.100043	Mixed	13
Hs.100043	Mouth	1
Hs.100043	Muscle	4
Hs.100043	Pancreas	1
Hs.100043	Prostate	5
Hs.100043	Skin	0
Hs.100043	Spleen	6

Function *mkESTdb* generates the UniGene database. This R code generates an EST MySQL database for *Homo sapiens*. The database provides the information in a form, in which ESTs in different tissues and development-stages are mapped on genes. The information provided by the UniGene database is visualized in **Table 12**.

The function *createEstByTissue* uses this table and maps the UniGene identifiers to UniprotKB-IDs with the IDmapping function described in section 4.1.1. The user can also select which development stage are included in the dataset and if the ESTs are annotated to normal or cancer stage. The function *createEstByhas* checks the developmental stages and the normal/disease states are in UniGene:

```
stages <- createEstByTissue(dbDriver="MySQL", dbHost="localhost", dbUser="root",
password=NULL, dbData="Unigene", check="TRUE")
```

```
data(referenceTable)

estTissueUniP <- createEstByTissue(dbDriver="MySQL", dbHost="localhost", dbUser="root",
password=NULL, dbData="Unigene", cancer_source="normal",
developmental_stage="adult", referenceTable=referenceTable)

estTissueUniP <- unigene$est_by_uniprotidtissue
```

After selecting the right parameters the function generates the EST expression dataset. Path2enet package also includes an EST data object previously built with this function. This object includes ESTs annotated only to normal and adult in *Homo sapiens*, corresponding to 18,880 gene/protein entries detected in 51 human tissues.

4.1.3.2 Microarray Gene Expression Barcode

```
data(barcodeTissueUniP)
```

Path2enet also includes a large genome-wide expression resource derived from the analysis of hundreds of high-density oligonucleotide microarrays. This data was generated using the algorithm Barcode and can be found at Barcode website (45) **(McCall *et al.*, 2011; McCall *et al.*, 2014)**. The tissue specific dataset based on the HGU133 Plus 2.0 (Human) microarray (abc-tis-gpl570-formatted_v3.csv) and the cell line specific dataset (abc-cell-gpl570-formatted_v3.csv) are annotated to ENSEMBL Gene identifiers. The Path2enet tool maps these identifiers with its ID mapping function to UniProtKB IDs. This resource stores information about the expression of 17,268 gene/protein entries detected in 195 tissues and cell lines. The data is included in Path2enet package as a data object.

4.1.3.3 RNA-Seq data

```
data(rnaseqTissueUniP)
```

As mentioned in sections 3.1.2, Path2enet uses two RNA-Seq datasets: **(i)** from Human Body Map 2.0; **(ii)** and from Human Protein Atlas. Both are produced with Illumina sequencing platforms. Both datasets are included in Path2enet package and the user can access them

loading the R data object as indicated above. The processing of these datasets was different:

1. for **Human Body Map 2.0**: The tool takes the ftp://ftp.ensembl.org/pub/release-70/bam/homo_sapiens/genebuild/ *.bam archives, and uses cufflinks (on human genome GRCh37) to generate the FPKM values of the genes based on the ENSEMBL gene IDs. These gene IDs are then mapped to UniProt protein IDs with ID mapping function and the generated *referenceTable*. Section 3.4.3 describes in detail the workflow to generate FPKMs out of provided/generated *.bam-files. This data correspond to a expression set of 18,744 gene/protein entries in 16 human tissues.
2. for **Human Protein Atlas**: The tool takes the "rna_tisse.csv.zip" file (available at http://www.proteinatlas.org/about/download website). The gene IDs are then mapped to UniProt protein IDs with ID mapping function and the generated *referenceTable*. The tool stores the FPKM expression data of 19,078 gene/protein entries of 33 human tissues

4.2 *Path2enet* tool to generate, analyse and visualize the pathway-driven biomolecular networks

The following result section shows the capability of the Path2enet tool to build, analyse and visualize the pathway driven biological networks. It explains which method can be used to decide in which specific biological condition (ON) or (OFF) a node actually is. A case study on B- and T-lymphocytes shows its usability with some specific experimental datasets. Section 4.2.1 describes the general networks, analysis and visualization Path2enet provides. Section 4.2.2 shows the results of a case study of B- and T-lymphocytes.

4.2.1 Generating the biomolecular networks

The meta-analysis of APID, KEGG and transcriptomic datasets is only possible if the Path2enet-Tool can access all data-sets in a proper way. The Path2enet-Tool is able (i) to access MySQL-databases of APID, Unigene and the Kegg2MySQL-database which the tool has created; (ii) to integrate the pre-processed transcriptomic datasets of microarray and RNAseq; and (iii) to load user specific transcriptomic or proteomic datasets. The next step is to combine these databases so that as much information as possible can be interrogated.

The identifiers which generate a link between the three data sources are the UniProt identifiers and the UniProtKB ID. To be more independent of other packages the tool provides its own ID mapping function explained in paragraph 4.1.1. All datasets are mapped to UniProtKB ID.

The input is quite similar to the searchFunction but its purpose is to create and analyse protein-networks. Table X describes the input parameters. These protein-networks contain the information of all databases and data sources.

***Table 12**: Overview of the parameters of the R-script path2enet.*

Parameter	Explanation	Example
dbDriver, dbUser, dbHost, dbName, password, GeneID, KEGGID, Pathwaytitle, PathwayID, Subtype, Proteinreaction, noNA	These are the same parameters as in the R script *searchFunction.*	See **Table** 7.

Local	If this parameter is set to TRUE, only interactions who include both interaction partners in the "GeneID/KeggID" list of genes are used to generate the network. If it is set to FALSE any interaction including only one interaction partner in the "GeneID/KeggID" list of genes are used to generate the netwrok.	
dbData	The subset of pathways to look for interactions and compare the network using the subset of pathways of dbGlobal. The standard is "Relations" which includes all pathways.	
dbGlobal	The name of this parameter has to be taken from the *Kegg2MySQL* database. This table will be used to search for relations, who include only one entry, for example a gene, in the *dbGlobal* table. Therefore, *dbData* and *dbGlobal* can be similar.	dbData = NormalPathways dbGlobal = CancerPathways
"Directory"	In this directory the data frames with the results of the analysed graphs are stored, are exported as a csv-file. R must have the right to write into this folder.	"/tmp/"
dbPPI	The name of the protein-protein-interaction database the funcion should use.	
Barcode	This includes the pre-processed database of the Gene Expression Barcode database object explained in paragraph 4.3.1.2	
Unitissue	This includes the pre-processed database of the UniGene database object explained in paragraph 4.3.1.1	
Rnaseq	This includes the RNA-Seq database objects explained in paragraph 4.3.1.3	

The path2enet-function of the Path2enet tool combines and analyses the selected networks and transcriptomic datasets. The output of the function are two lists which: **(i)** store the created protein-networks as igraph-objects; and **(ii)** store the results of the analysis of these protein networks as "data frames". The created graphs are stored in the list under "*$Graphs" . The results of the analysed graphs are stored in the list "*$Analysed_Graphs". The created graphs and data frames are explained in the following paragraphs. A sample input and output for the “Notch Signaling Pathway” and its output is shown below.

```
graphs <- path2enet(dbDriver="MySQL", dbUser="root", dbHost="IP/localhost",
password="PW", dbName="KeggSQL", Pathwaytitle="Notch Signaling Pathway",
Local="FALSE", dbData="Path2net_SignalPathwayrelation", dbGlobal="Path2net_Normal",
dbPPI="PPI", unitissue=Unitissue, barcode=barcodeTissueUni,  rnaseq=rnaseqTissueUniP)
```

4.2.1.1 Graphs of the Path2enet tool

The Path2enet-Function generates three different graphs and compares them to each other.

Intension isto highlight shared interactions in the graphs. The generated graphs are (i) PPI: a graph created with interactions in the selected dbPPI protein-protein-interactions database; (ii) *PATHWLocal*: a graph created with interactions in the selected dbData dataset of the pathway database; and (iii) *PATHWGlobal*: a graph created with the interactions in the selected dbGlobal dataset of the pathway database. **Table 14** explains the graphs.

```
names(graphs)

[1] "Graphs"        "Analysed_Graphs"

names(graphs$Graphs)

[1] "PPI_PATHWglobal"       "PATHWlocal_PATHWglobal" "PPI_PATHWlocal"

names(graphs$Analysed_Graphs)

[1] "EdgeList_PATHWlocal_withGroups"              "EdgeList_withSubtype_withoutGroups"

[3] "Analysed_PATHWlocal_KeggIDs_withGroups"
"Analysed_PATHWlocal_UniprotKBIds_withGroups"

[5] "Analysed_PATHWlocal_UniprotKBIds_withoutGroups"  "Analysed_PPI"

[7] "Analysed_PATHWglobal_UniprotKBIds_withGroups"
"Analysed_PATHWglobal_UniprotKBIds_withoutGroups"Local="FALSE",
dbData="Path2net_SignalPathwayrelation", dbGlobal="Path2net_Normal",  dbPPI="PPI",
unitissue=Unitissue, barcode=barcodeTissueUni,  rnaseq=rnaseqTissueUniP)
```

***Table 13**: Explanation of the three protein-networks created with the R-script path2enet.*

Protein-network	Description
PPI	PPI is a graph created with stored protein-protein-interactions in APID.
PATHWLocal	It is the result of user search in table *dbData* of the *Kegg2MySQL* database. It selects the *dbData* itself. If the user searches with a list of gene symbols or KEGGIDs only relations with both parts on this list are in the graph (see **Figure 20**).
PATHWGlobal	It is the result of user search in table *dbGlobal* of the *Kegg2MySQL* database. It selects the *dbGlobal* itself. The difference between *PATHWLocal* and *PATHWGlobal* is that in the local only searches for the links in the query pathway but the global searches in the whole KEGG database.

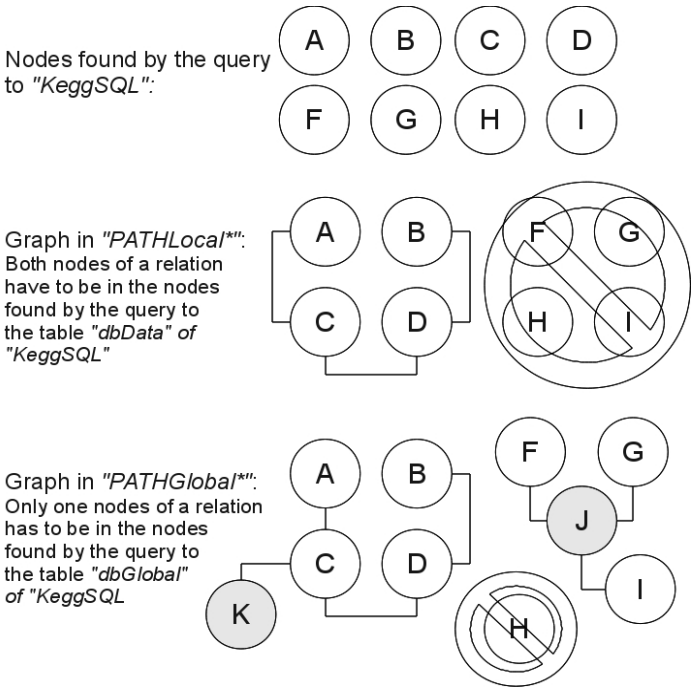


Figure 20: *Difference between PATHLocal* and PATHGlobal*. The white circles are nodes found by the query to the KeggSQL-database. The gray circles are nodes of a relation in dbGlobal with a node in the node list found by the query to KeggSQL, but are not part of the list.*

Table 15 describes to each other compared graphs. The Path2enet-Functions stores the networks in the sublists "PPI_PATHWGlobal", "PATHWLocal_PATHWGlobal" and "PPI_PATHWLocal". In every list there are two networks. The output are six graphs.

Table 14: *Explanation of three lists generated by Path2enet.*

PPI_PATHWglobal	*PPI2PATHWglobal:*Protein-protein-interactions found in the *APID* database are the *basic graph* and *PATHGlobal* is mapped on it.
	PATHWglobal2PPI: The *basic graph* is *PATHGlobal* and the graph *PPI* is mapped on it
PATHWLocal_PATHWGlobal	PATHWlocal2PATHWglobal: PATHWLocal is the basic graph and PATHWGlobal is mapped on it.
	PATHWglobal2PATHWlocal: PATHGlobal is the basic graph and PATHWLocal is mapped on it.
PPI_PATHWlocal	*PPI2PATHWlocal*: *PPI* is the *basic graph* and *PATHWLocal* is mapped on it.
	PATHWlocal2PPI: PATHWLocal is the basic graph and PPI is mapped on it.

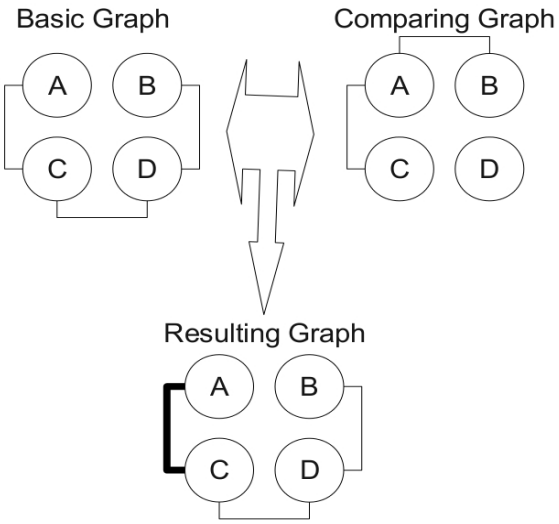


Figure 21: *Relation scheme of the Comparing Graph mapped on to the basic graph. Relations which both graphs have in common are bold.*

4.2.1.2 Attributes of the graphs in Path2enet

Path2enet creates *igraph*-objects which include the relations *Edge.attributes* and *Vertex.attributes* of graphs. *igraph* and other existing packages of *R* can be used to visualize and analyse these objects. The structure of an *igraph-object* in general is explained in the documentation of the *igraph-package*. Attributes of vertexes and edges are important to modify a graph in different ways. Vertex.attributes are demonstrated in **Table 16**.

Table 15: *Vertex.attributes* of *igraph-objects* created with *Path2enet.R* . The attribute "name" stores the names of the nodes. The other attributes store the transcriptomic information.

Name	**Blood**	**Brain**
embryonic_tissue	Eye	Heart
Kidney	Liver	Lung
lymph_node	Mixed	Mouth
Muscle	Pancreas	Prostate
Skin	Spleen	adrenal_gland
Bladder	Ear	Intestine
mammary_gland	Placenta	Stomach
Testis	Thymus	Thyroid
uncharacterized_tissue	pineal_gland	Vascular
Cervix	Bone	Ganglia
Tonsil	amniotic_fluid	Pharynx
adipose_tissue	Ovary	umbilical_cord
Uterus	salivary_gland	Peritoneum
bone_marrow	Nerve	pituitary_gland
connective_tissue	Larynx	spinal_cord
Epididymis	Esophagus	Parathyroid

Name	Blood	Brain
Bronchus		

The most important information stored in the generated graphs are the "vertex attributes" and the "edge attributes". The functions *list.vertex.attributes* and *list.edge.attributes* of the *igraph-package* in *R* show the vertex- and edge-attributes of an *igraph-object*. The "vertexes" are the nodes of a graph and the edges are the relation between these nodes. *Vertex.attributes* store information about a node of the graph. In the created graphs the nodes get their attributes from the *Unigene* database. *Path2enet* stores in which tissues or development-stages transcripts of a node, which represents a gene, are expressed. Another attribute of the node is its name.

The *Edge.attributes* store information about the relation between two nodes. This information is obtained by the *Kegg2MySQL*-database or by comparing the relations of two graphs with each other. **Table 17** describes the *Edge.attributes* stored in the graphs created with the *Path2enet.R* function.

Table 16: *The Edge.attributes stored in the igraph-objects. Subtype and Subtype2 are information obtained by the KeggSQL-database*.

Attributes	Description	Possibilities
Subtype	This describes the relation between two nodes of the graph. The information is obtained by the *Kegg2MySQL*-database.	"activation", "binding/association", "compound", "dephosphorylation", "dissociation", "expression", "indirect effect", "inhibition", "missing interaction", "NA", "phosphorylation", "repression", "state change", "ubiquination"
Subtype2	This attribute describes the relation of the two entries in detail. The information is obtained by the *Kegg2MySQL* database	"activation", "binding/association", "dephosphorylation", "dissociation", "expression", "indirect effect", "inhibition", "missing interaction", "NA", "phosphorylation", "ubiquination"
Intersection	The attribute *Intersection* is created by comparing two graphs with each other.	The output is: - *"TRUE",* if the relation exists in both graphs. - *"FALSE"*, if the relation exists only in one graph.

4.2.1.3 Visualization of graphs in Path2enet

For storing graphs the Path2enet tool generates *igraphs-objects* which can be used by network specific packages in R. These packages can visualize and analyse these igraphs-objects. This is very helpful because it enables other scientists easily to analyse the graphs generated by the Path2enet tool.

One example is the *igraph-package* which includes several functions to visualize and analyse *igraphs-objects*. Path2enet-Functions *graphTKplotterPATHW* and *graphTKplotterTissue* use the functions *plot.igraph* and *tkplotter* of the *igraph package* to visualize the *igraphs-objects* generated with *Path2enet*. The vertex and edge attributes of these *igraphs-objects* are explained in **Table 16** (vertex) and **Table 17** (edge). Both functions can use these attributes to create unique graphs, which can be tissue and development specific.

The graphs created with the functions *tkplotter* and *plot.igraph* are different:

The *plot.igraph* function allows visualizing graphs which include: **(i)** a title; **(ii)** a legend; and **(iii)** a subtitle. But the created graphs are static and nodes of bigger graphs can overlap them easily.

tkplotter generates dynamic graphs. The position of the nodes of these graphs can be modified by the user. Actually *tkplotter* does not allow to give a graph a title, a legend and a subtitle. This is very important if a network must be is presented to third parties.

To overcome certain limitations of the functions plot.igraph and tkplotter the Path2enet-Tool functions graphTKplotterPATHW and graphTKplotterTissue use them in parallel. GraphTKplotterPATHW can use the network layouts the user generates with the dynamic representation of the graphs in order to overcome overlapping and non-readable graph structures.

Examples of the visualized graphs are given in the following paragraphs: Paragraph 4.2.1.3.1 explains function graphTKplotterPATHW; Paragraph 4.2.1.3.2 explains the function graphTKplotterTissue.

4.2.1.3.1 The normal representation of graphs with *graphTKplotterPATHW*

The R-function graphTKplotterPATHW is the basic function for the visualization of the graphs created with the R-script Path2enet.

The attributes used to visualize the graph are the Edge.attributes stored in the .Several subtypes (activation, inhibition, expression, phosphorylation and other subtypes) are marked by a different color (blue, red, green, yellow and purple). Relations without a subtype are black. The relations which the basic graph has in common with the graph compared to are highlighted by bold edges. The nodes are colored depending on the results of the community function provided by the igraph-package.

The names of the nodes are UniprotIDs. Uniprot separates its entries of proteins in verified and unverified ones. Nodes with a suffix like “HUMAN”, e.g. “SUH_HUMAN” and “RBPJL_HUMAN”, are verified. The function removes the “*_HUMAN” postfix. They are entries of UniProt Knowledgebase/Swiss-Prot [UniProtKB/Swiss-Prot]. Proteins of UniProtKB/Swiss-Prot are highly verified. Nodes with names like “Q7Z6C” and “D3DRF” are members of UniProtKB/Translated EMBL Nucleotide Sequence Data Library [TrEMBL]. These proteins are not verified.

The function visualizes in Figure 8 and 9 the "Notch Signaling Pathway". The relations of its PPi graph are mapped on the basic graph PATHlocal.

The function graphTKplotterPATHW of the Path2enet-Tool uses plot.igraph of the igraph-package to draw **Figure 23** and uses tkplotter to draw **Figure 24**.

The graphs of the function graphTKplotterPATHW can provide essential information:

Nodes and their colors demonstrate inter alia: **(i)** which proteins are member of a graph; **(ii)** which proteins are verified or unverified with regard to the Uniprot database; and **(iii)** which nodes are members of a community. One example is the community of nodes “EP300”, “CBP”, “MAML1/2/3” and “KAT2A/B”. These proteins are Co-activators of „RBPJL” and “SUH”.

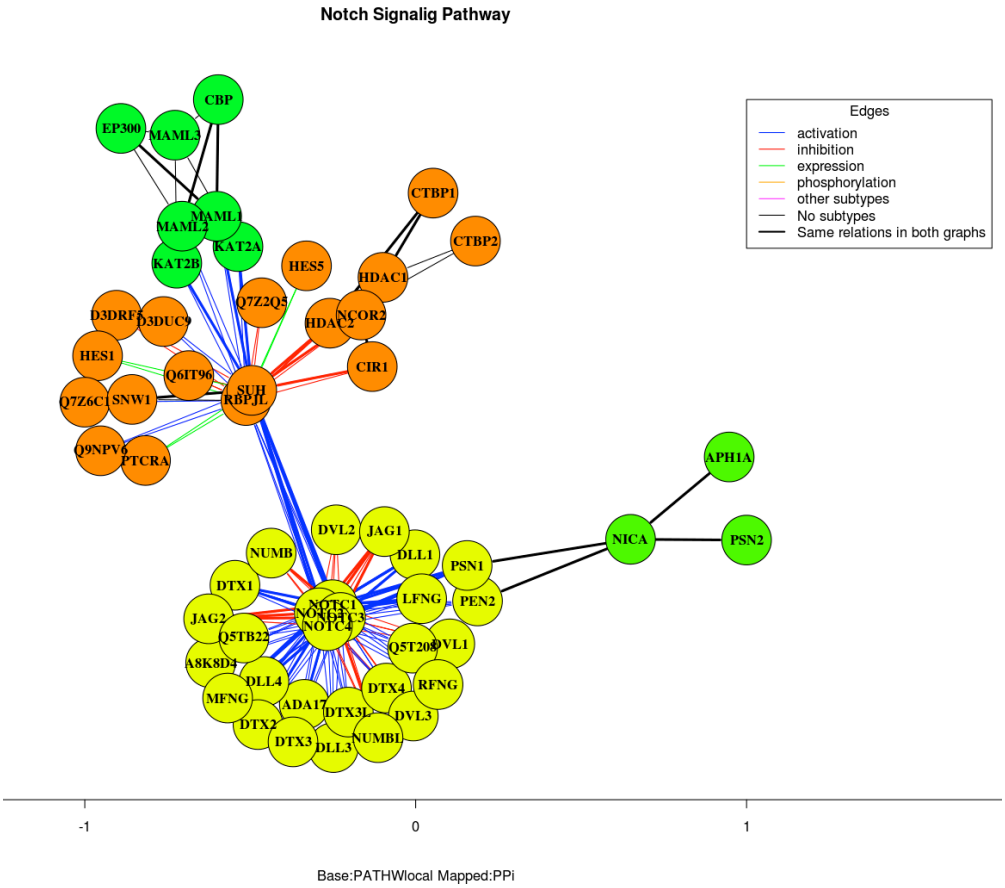


Figure 22: *PPI graph of the "Notch Signaling Pathway" mapped on PATHLocal.*

- *The bold edges are relations both graphs have in common.*
- *The color of the edges is explained in the legend.*
- *The color of the nodes follows the results of the calculation of the igraph function community*
- *Nodes close to each other and with the same color are members of a community.*
- *Graph plotted by the plot.igraph function of the igraph-package.*
- *Layout type: "fruchterman.reingold".*

The edges of the graph provide the following information: **(i)** which nodes interact exactly in the graph; **(ii)** which type of interaction have two nodes; **(iii)** which interactions have both graphs in common; and **(iv)** which interactions have only the basic graph.

Storing all this information in one graph will support biological investigators to interpret protein-interactions.

Figures 22 and 23 include the above mentioned information. But although they represent the same igraph.object the differences are obvious: Figure 8 has a title, subtitle and a legend. The nodes overlap and it is difficult to separate nodes which are strongly related in the same community..

In Figure 9 the nodes do not overlap like in Figure 8. With function tkplotter it is possible to

separate nodes and communities of a graph manually. This allows to create graphs which biological researchers can better interpret. But the tkplotter of the igraph package is not able to include title, legend and subtitle in the graph..

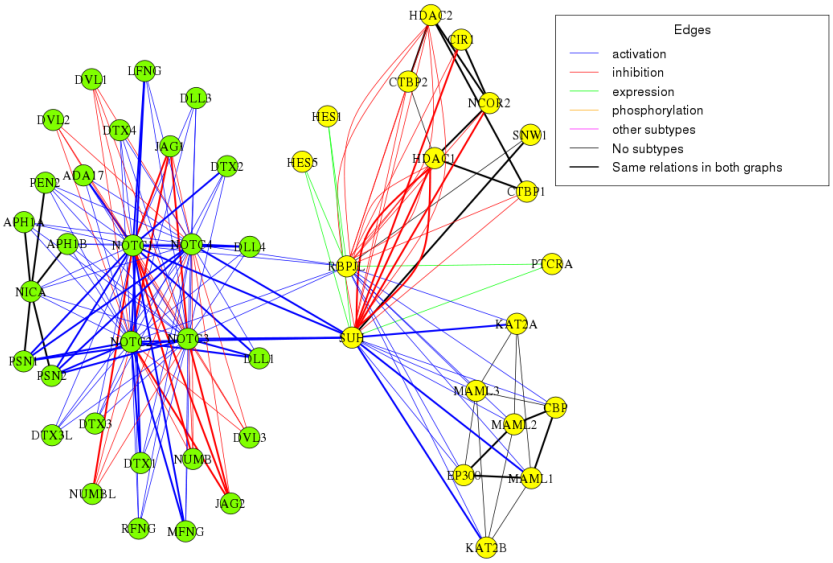


Figure 23: tkplotter graph using same data as in Figure X. The overlapping can be reduced individually.

4.2.1.3.2 Tissue specific representation of the graphs

Function graphTKplotterTissue integrates both types of attributes of igraphs-objects generated with Path2enet:

(i) Edge.attributes: The *graphTKplotterTissue* uses the Edge.attributes like the function *graphTKplotterPATHW* (4.2.1.3.1).

(ii) Vertex.attributes: The *graphTKplotterTissue* uses the Vertex.attributes of the igraphs-objects: **(i)** to mark all nodes yellow which have at least one “transcript” in the selected tissue; (ii a) to mark all nodes white which do not have any “transcript” in the selected tissue; or (ii b) to delete all nodes which do not have any “transcript” in the selected tissue; and **(iii)** to mark all nodes blue which have an unknown number of “transcript”.

In these examples I will use the UniGene EST transcripts to show the graphical representation of tissue-specific networks of the Path2enet-Tool. In the following sections we will talk about the combination of various transcriptomic levels and the case study with B- and T-cells.

Figure 24 demonstrates the functionality of the graphTKplotterTissue with the EST dataset of UniGene of liver. It shows clearly that this function enables to differentiate between “Notch/1/2/3/4” and “SUH”(“RBPJ”)/”RBPJL”. This is for example not possible with “Notch Signaling Pathway” provided by KEGG or protein-protein-interaction networks APID creates.

In **Figure 24** all nodes are deleted which do not have at least one transcript in liver according to *Unigene*. The graph has the following advantages: **(i)** it is tissue and development-stage specific; **(ii)** it provides a good overview of proteins which are probably interacting in liver in the *“*Notch Signaling Pathway*”*; and **(iii)** the number of nodes of the graph is reduced so that the graph is less complex.

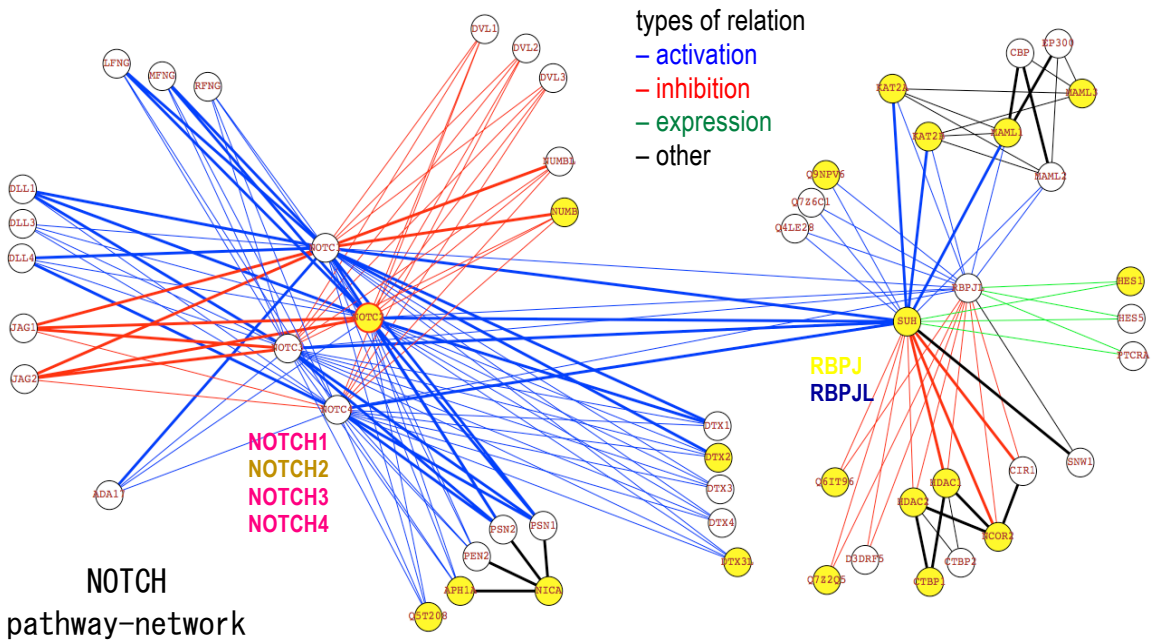


Figure 24: *graphTKplotterTissue: PATHWLocal as basic graph. On it the graph PPI is mapped.Selected tissue: liver [EST]]. All nodes that have one or more transcripts are marked as yellow.The nodes that do not have any transcript are colored white.The proteins "Notch2" and "RBPJ" have transcripts in the liver. The proteins "Notch1/3/4" and "RBPJL" do not have any transcript. The graph is the "Notch Signaling Pathway" of "KEGG" of homo sapiens.*

Such a tissue and development-stage specific graph is not possible in *KEGG*. But these graphs give biological researchers the option to choose the type of tissue which fits best regarding their research object. So they can easier guess which proteins are properly represented in a specific tissue or development-stage.

The next **Figure** shows, how the integration and comparison of PATHWLocal; PATHWGlobal and the transcriptomic dataset can give new insights into biological processes.

Comparing *PATHWLocal* with *PATHWGlobal* is useful for a biological researcher because: **(i)** he knows which proteins interact in the *"*Notch Signaling Pathway" in liver (*PATHWLocal* and *"*PATHWLocal Liver*")*; and **(ii)** he knows which other proteins interact with proteins of the *"*Notch Signaling Pathway*"* in *Kegg2MySQL* (*PATHWGlobal* and "PATHWGlobal Liver*"*).

For biological researchers it is easier to analyze *"*PATHWLocal Liver*"* and "PATHWGlobal Liver*"* than "PATHWLocal" and "PATHWGlobal" because the number of nodes and the complexity of these two graphs are reduced.

Comparing the images "PATHWGlobal Liver*"* with *"*PATHWLocal Liver*"* of **Figure 25** shows that the protein *"*RB*"* interacts with *"*HDAC/2*"* in "PATHWGlobal Liver*"* but it does not interact in *"*PATHWLocal Liver*"*. *KEGG* excludes *"*RB*"* in the *"*Notch Signaling Pathway*"*. The interaction in the *"*Notch Signaling Pathway*"* between *"*RB*"* and *"*HDAC/2*"* could be investigated, because in liver the three proteins have at least one transcript. To create a hypothesis like this is not possible by only using data of *KEGG*, *APID* or *Unigene*. But the combination of all three datasets allows creating such hypothesis.

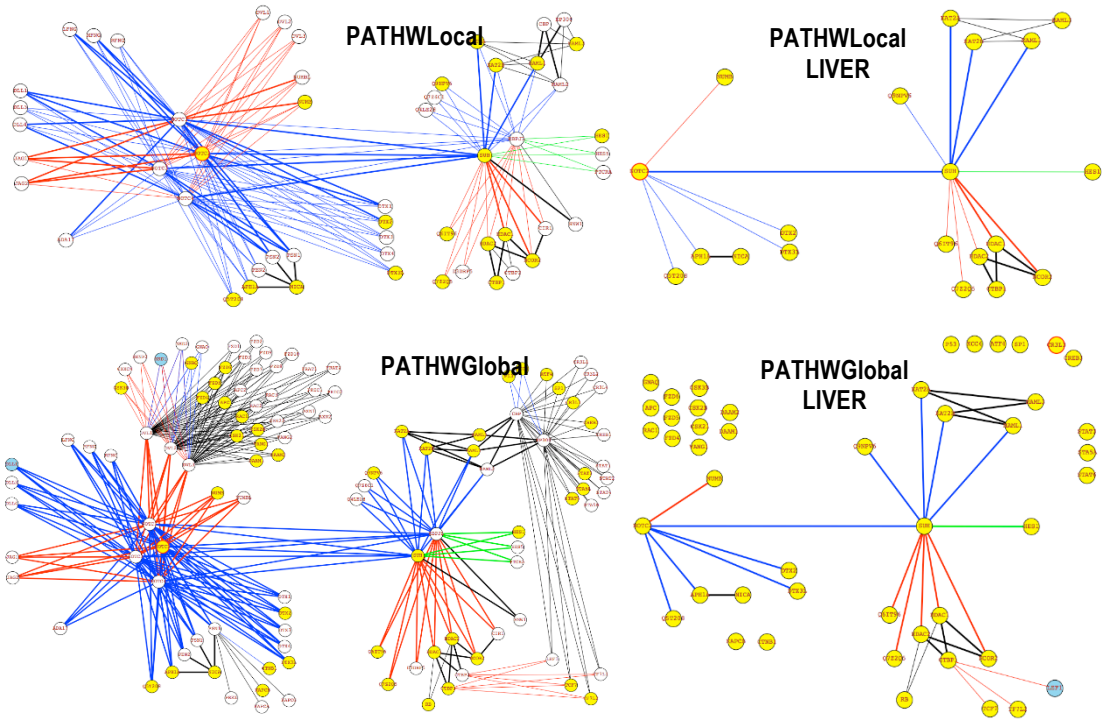


Figure 25: *Differences between PATHWLocal and PATHWGlobal of the "Notch Signaling Pathway". Tissue: liver Pathway:"Notch Signaling Pathway". PATHWGlobal shows relations of the nodes in "Notch Signaling Pathway" in the whole Kegg2MySQL database.*

4.2.1.4 Statistical analysis of graphs in Path2enet

Path2enet generates graphs and analyzes them statistically. It stores the statistical analysis in *R* matrices. These data frames are stored in a list called *Analysed_Graphs* and are saved as *".csv"*-files in the directory entered as a parameter of the function *path2enet* (*path2enet(Directory=" ")*). Paragraph 4.2.1.1.1 describes the general Analysed_Graphs and 4.2.1.1.2 shows the results of an analysis of a large cancer-pathways driven network.

4.2.1.4.1 General network analysis

The list *Analysed_Graphs* has 11 members which are matrices. The relations, which were found in the databases (*APID, KeggSQL)* regarding the input of the user, are memorized in the matrices starting with *"EdgeList*"*. In the matrices beginning with "Analysed*" the parameters *degree*, *betweenness*, *eigenvector* and *clustering_coefficient* of the created graphs (*PPI*, *PATHWLocal* and *PATHWGlobal*) are calculated, using functions of the *igraph-package*. **Table 18** describes in detail the matrices generated with *path2enet*.

Table 17: Explanation of Matrices created with Path2enet

Matrices	Description
EdgeList_PATHWlocal_withGroups	EdgeList of PATHWlocal with nodes combined to groups. The entries of the *"type"* group, which are separated by "|", like *"|hsa:1487|hsa:1488|hsa:3065|hsa:3066|"*, are treated like one node in the graph.

Matrices	Description
EdgeList_withSubtype_withoutGroups	EdgeList of PATHWlocal with individual nodes and the kind of the relation the nodes in the pathway have. The entries of the *"type"* group, which are normally separated by "|", like *"|hsa:1487|hsa:1488|hsa:3065|hsa:3066|"*, and treated like one node in the graph, are separated completely. They are treated like, "*hsa:1487"*, *"hsa:1488"*,*"hsa:3065"* and *"hsa:3066"*.
Analysed_PATHWlocal_KeggIDs_withGroups	The analysis of PATHWlocal with groups and KEGG identifiers.
Analysed_PATHWlocal_UniprotKBIds_withoutGroups	The analysis of PATHWlocal with UniprotKB identifiers and without groups.
Analysed_PPI	The analysis of PPI network with every node treated individually in the statistical analysis and UniprotKB identifiers.
Analysed_PATHWglobal_UniprotKBIds_withGroups	Results of the statistical analysis of the graph derived from the *dbData* table of the *Kegg2MySQL* database, but entries of the type *"group"* are separated and treated like individual nodes.
Analysed_PATHWglobal_UniprotKBIds_withoutGroups	The analysis of PATHWglobal with every node treated individually in the statistical analysis and UniprotKB identifiers.
Output	The output the function stores in the selected "Directory" are forteen .txt-files. The output is seperated in: **(i)** "EdgeList" of the graph; **(ii)** the statistical analysis with or withour "groups"; and **(iii)** pathway related dataset is saved with the KEGG identifiers and the UniprotKB identifiers.

4.2.1.4.2 Example of a large cancer-pathway network

Topological and graph analysis of a network is important. This section explains the analysis of the table *Path2enet_CancerPathway* of *Kegg2MySQL* database. The tool *path2enet* generates and analyses this network. **Table 18** is an excerpt of the first 30 entries resulting from the topological analysis. In total the graph has 377 nodes and 1,793 relations.

With the Path2enet tool *graphTKplotterPATHW* a biological researcher can analyze the graphical image of a network. He can search for important nodes in the graph and separate it in different clusters. This work is time-consuming and very subjective.

The analysis of network parameters (described in section 3.2.4) is faster and more objective than the analysis of a graphical network. **Table 18** for example shows that the proteins *AKT1/2/3* play an important role in *Path2enet_CancerPathway*, because they have the highest degree. Defects in the *threonine-protein kinase AKT1* [AKT1/2/3] are related to breast cancer, colorectal cancer and ovarian cancer (46). Defects in the *"Epidermal Growth factor receptor"*

[EGFR] are related to lung cancer (47). EGFR is 4th in **Table 18**.

***Table 18**. PATHWLocal graph: Analysis of top 35 entries corresponding to well identified proteins. Database used: CancerPathway of KeggSQL. Data sorting: Degree→ Eigenvector→ Betweenness first.*

"*KEGG*Translated"	KEGGID	Degree	Betweenness	Eigen-vector	Clustering-Coefficient
"AKT1_HUMAN"	"hsa:10000"	63	8099.95	1	0.09
"AKT2_HUMAN"	"hsa:207"	63	8099.95	1	0.09
"AKT3_HUMAN"	"hsa:208"	63	8099.95	1	0.09
"EGFR_HUMAN"	"hsa:1956"	53	2528.62	0.27	0.04
"IGF1R_HUMAN"	"hsa:3480"	49	2295	0.25	0.02
"PGFRA_HUMAN"	"hsa:5156"	49	2295	0.25	0.02
"PGFRB_HUMAN"	"hsa:5159"	49	2295	0.25	0.02
"FGFR1_HUMAN"	"hsa:2260"	43	1778.63	0.24	0.03
"PK3CA_HUMAN"	"hsa:5290"	42	3706.45	0.69	0.07
"PK3CB_HUMAN"	"hsa:5291"	42	3706.45	0.69	0.07
"PK3CD_HUMAN"	"hsa:5293"	42	3706.45	0.69	0.07
"PK3CG_HUMAN"	"hsa:5294"	42	3706.45	0.69	0.07
"P85A_HUMAN"	"hsa:23533"	40	2573.03	0.69	0.08
"P85B_HUMAN"	"hsa:5295"	40	2573.03	0.69	0.08
"PI3R5_HUMAN"	"hsa:5296"	40	2573.03	0.69	0.08
"RASK_HUMAN"	"hsa:3845"	37	5038.91	0.23	0.08
"RASH_HUMAN"	"hsa:3265"	35	2706.56	0.23	0.08
"RASN_HUMAN"	"hsa:4893"	35	2706.56	0.23	0.08
"MET_HUMAN"	"hsa:4233"	34	996	0.07	0
"NFKB1_HUMAN"	"hsa:4790"	33	5103.84	0.13	0.04
"TF65_HUMAN"	"hsa:5970"	33	5103.84	0.13	0.04
"GLI1_HUMAN"	"hsa:2735"	30	681.82	0	0.15
"GLI2_HUMAN"	"hsa:2736"	30	681.82	0	0.15
"PDPK1_HUMAN"	"hsa:5170"	27	3.77	0.61	0.44
"GSK3B_HUMAN"	"hsa:2932"	22	16232.11	0.19	0
"FZD1_HUMAN"	"hsa:11211"	22	861.83	0	0
"FZD2_HUMAN"	"hsa:7855"	22	861.83	0	0
"GRB2_HUMAN"	"hsa:2885"	22	665.44	0.1	0.11
"ERBB2_HUMAN"	"hsa:2064"	21	444.86	0.21	0.1
"CCND1_HUMAN"	"hsa:595"	19	4357.4	0.01	0.16
"HIF1A_HUMAN"	"hsa:2034"	18	2706.6	0.01	0
"MP2K1_HUMAN"	"hsa:5604"	17	3873.09	0.09	0
"MK01_HUMAN"	"hsa:5594"	17	3275.68	0.01	0.04
"FGFR2_HUMAN"	"hsa:2263"	17	300.91	0.19	0
"INSRR_HUMAN"	"hsa:3645"	17	300.91	0.19	0

"G1/S-specific cyclin-D1" [CCND1] is a regulatory component of the *"cyclin D1-CDK4- complex"* [DC] (Uniprot/P24385). It has a very high *betweenness* (4357.4) although it does not have a high degree (19). CCND1 connects different clusters in *CancerPathways* with each other. This demonstrates the regulatory function of CCND1. The gene of CCND1 may be an oncogene; in cases the gene is up-regulated the result is a directly altering progression through the cell cycle. Defects of the gene are a cause of multiple myeloma (Myeloma).

*"*3-phosphoinositide-dependent protein kinase" [PDPK1] phophorylates and activates PKB/AKT, PKA, PKC-zeta, RPS6KA1 and RPS6KB1(Uniprot/O15530) he protein has an average *degree* (27) and a low *betweenness* (3.77) but a high *eigenvector* (0.61) and *clustering coefficient* (0.44). Such a low *betweenness* demonstrates that this protein does not combine two or more clusters in the network like CCND1. The high *eigenvector* demonstrates that its interacting proteins are important in the network. PDPK1 interacts for example with AKT1/2/3, which are the most important proteins in the network. The *clustering coefficient* shows that the proteins which interact with PDPK1 are good connected with each other. In Uniprot/KB PDPK1 is not annotated to any disease. Biological researchers could investigate if a down- or up-regulation of this gene does not result in a major disease.

A full analysis of **Table** X would take a long time and much effort. But the example shows that the statistical analysis of large graphs is essential for biological researcher.

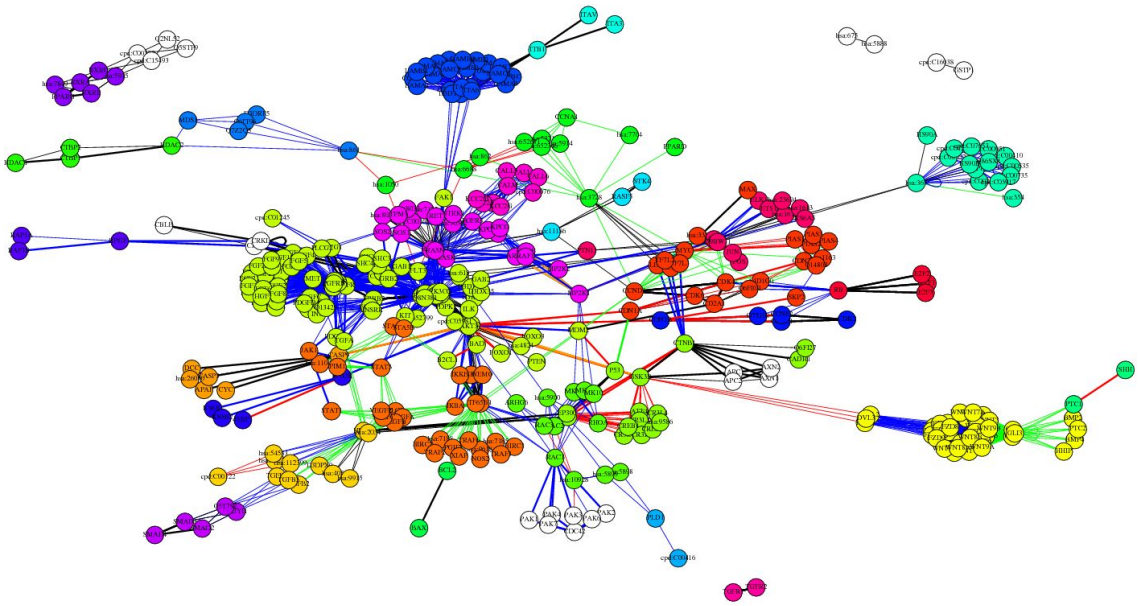


Figure 26*: graphTKplotterPATHW: Graph of table CancerPathways. Data basis: Table CancerPathway of Kegg2MySQL.*

Function: graphTKplotterPATHW of the Path2enet tool using the tkplotter of the igraph-package

Figure 26 shows the graph generated with the Path2enet tool's function *graphTKplotterPATHW*. It is based on the table *CancerPathways* of *Kegg2MySQL*-database. It demonstrates how difficult it can be to derive information out of large networks only with visual analysis. The statistical analysis is: **(i)** faster to accomplish; **(ii)** more objective; and **(iii)** better repeatable.

The graph also demonstrates that biological investigators can use the Path2enet tool to visualize and analyse large biological networks.

4.2.1.4.3 Combining statistical analysis and expression datasets

Using the gene expression data described above either from ESTs, Barcode and microarrays or RNA-Seq, Path2enet generates tissue or cell type specific networks. To do this, the user can select various vertex.attributes and set a filter value based on the type of expression attribute ("est", "barcode", "rnaseq""): **(i)** If a gene shows an expression value higher than the selected threshold, it is considered as ON-active; and **(ii)** if the value is lower, it is considered as OFF-inactive.

It is also possible to filter for various expression types at once and combine the results for an expression network of higher confidence. For each kind of a dataset, the user can select a specific threshold. The function generates a graph object only containing nodes, which have expression values above the selected threshold. The function graphParameters analysis this.

In this example (1) we fetched all liver specific attributes of a generated Notch Signaling Pathway; (2) set a threshold for the different types of expression dataset; (3) generated a graph which only contains nodes above the threshold; (4) analysed the graph; and (5) print it.

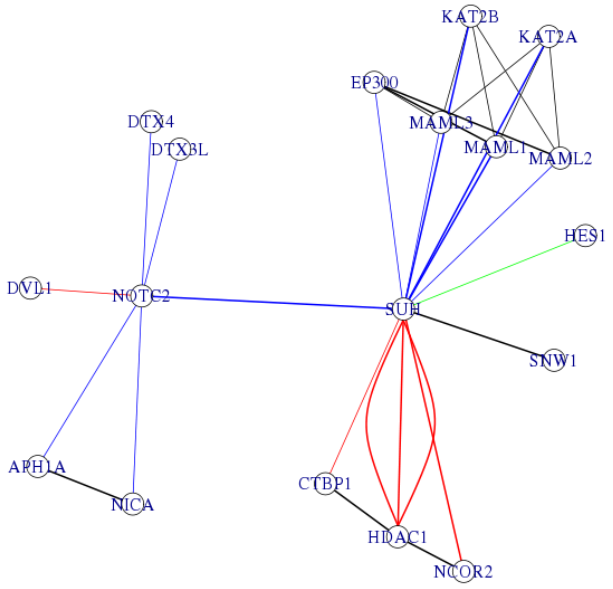


Figure 27: The genes of this network has an expression level in the samples studied higher value than the selected thresholds

4.2.2 Case Study B- and T-lymphocytes

In this section the Path2enet tool generates the pathway-driven biomolecular networks with the presented a priori-processed transcriptomic datasets and experimental datasets of B- and T-lymphocytes.

4.2.2.1 Including user specific datasets

Besides the pre-processed datasets, **Path2enet** enables to integrate experimental expression data as well. In the case study presented here, we use the algorithm to generate **specific pathway-expression-networks** of the **Notch signaling pathway** in three types of human cells: B cells (CD19+) and T cells (CD4+ and CD8+).

In this example we download 4 microarray data for each cell type and perform a comparison of the results for the B cells (CD19+) versus the T cells (CD4+ and CD8+). First, we download the CEL files and create the **AffyBatch** R objects:

In the next step, we use the expr2barcode function to perform an analisis with the Barcode algorithm and integrate the data into the Notch Signaling Pathway that was generated above, object “graph.NotchPPi”. We map the Affymetrix probesets to Uniprot IDs using referenceTableHGU133Plus. For further information about this mapping: “help(referenceTableHGU133Plus)”.

In the next step, we use the expr2barcode function to perform an analysis with Barcode algorithm and integrate the data into the Notch Signaling Pathway that was generated above: object “graph.NotchPPi”. We map the Affymetrix probesets to Uniprot IDs using referenceTableHGU133Plus. For further information about this mapping: “help(referenceTableHGU133Plus)”.

```
graphs <- path2enet(dbDriver="MySQL", dbUser="root", dbHost="IP/localhost",
password="PW", dbName="KeggSQL", Pathwaytitle="Notch Signaling Pathway",
Local="FALSE", dbData="Path2net_SignalPathwayrelation", dbGlobal="Path2net_Normal",
dbPPI="PPI", unitissue=Unitissue, barcode=barcodeTissueUni,  rnaseq=rnaseqTissueUniP)

graph.NotchPPi <- graphs$Graphs$PPI_PATHWlocal$PATHWlocal2PPI

NotchPPi.attributes <- list.vertex.attributes(graph.NotchPPi)

NotchPPi.attributes[grep("liver",NotchPPi.attributes)]

# "est_liver"      "barcode_liver"   "rnaseq_liver_bm" "rnaseq_liver_pa"

liver.attributes <- NotchPPi.attributes[grep("liver",NotchPPi.attributes)]

NotchPPi.liver <- graphParameters(graph.NotchPPi, vertex.attributes=liver.attributes,
barcode.value=0,est.value=0,rnaseq.value=0)

NotchPPi.liver.graph <- NotchPPi.liver$Reducedgraph

graphs.est.liver <- plot(NotchPPi.liver.graph, edge.color=E(NotchPPi.liver.graph)$color)
```

In the next step, we use the expr2barcode function to perform an analisis with the Barcode algorithm and integrate the data into the Notch Signaling Pathway that was generated above, object “graph.NotchPPi”. We map the Affymetrix probesets to Uniprot IDs using referenceTableHGU133Plus. For further information about this mapping: “help(referenceTableHGU133Plus)”.

```
data(referenceTableHGU133Plus)

expr2barcode.result.bcell_cd19_tcell_cd4_cd8 <-
expr2barcode(exprAffy=bcell_cd19_tcell_cd4_cd8,
phenotype="sample",referenceTable=referenceTableHGU133Plus,  igraph=graph.NotchPPi,
cutoff=1)
```

The generated object expr2barcode.result.bcell_cd19_tcell_cd4_cd8 includes four lists:

- *BarcodeGraph* includes the igraph object with the integrated results of the analisis.
- *BarcodeIndividual* includes the results of the barcode analisis of each microarray.
- *BarcodePheno* includes the mean of the barcode values of the phenotypes.

```
# Normal human T cells CD8+ from GSE14879
cd8 <- c("GSM371708", "GSM371709","GSM371710", "GSM371711")
dir.create(path="./TCELLCD8/")
for(cd8_file in cd8){
getGEOSuppFiles(GEO=cd8_file,makeDirectory=FALSE, baseDir="./TCELLCD8/")
}
files.dir <- dir("./TCELLCD8/")
lapply(paste("./TCELLCD8/",files.dir[grep("CEL.gz", files.dir)], sep=""), function(x) gunzip(x))
tcell_cd8 <- ReadAffy(celfile.path="./TCELLCD8/")
pData(tcell_cd8)[,1] <- "TCELLCD8+"
tcell_cd4_cd8 <- merge.AffyBatch(tcell_cd4,tcell_cd8)
# Normal human B cells from GSE24223
bcell_cd19 <- c("GSM595853", "GSM595858", "GSM595863", "GSM595870")
dir.create(path="./BCELLCD19/")
for(bcd19 in bcell_cd19){
getGEOSuppFiles(GEO=bcd19,makeDirectory=FALSE, baseDir="./BCELLCD19/")
}
files.dir <- dir("./BCELLCD19/")
lapply(paste("./BCELLCD19/",files.dir[grep("CEL.gz", files.dir)], sep=""), function(x)
gunzip(x))
bcell_cd19 <- ReadAffy(celfile.path="./BCELLCD19/")
pData(bcell_cd19)[,1] <- "BCELLCD19+"
```

- *AffyBatchUniprot* includes the expression set mapped to Uniprot-IDs.

4.2.2.2 Visualization in R

In the example above, we included 4 microarrays of each cell type. However, **Path2enet** package also includes a graph with a complete dataset corresponding to the analysis of the B cells (CD19+) and the T cells (CD4+ and CD8+), stored in the data object: “path2enet.BCELLCD19_TCELLCD8_TCELLCD4_NOTCHLOCAL_expr2barcode”. This dataset includes all 163 microarrays explained in section 3.1.2.2.2. The visualization and analysis are based on this dataset.

```
data(path2enet.BCELLCD19_TCELLCD8_TCELLCD4_NOTCHLOCAL_expr2barcode)

bcell_cd19_tcell_cd4_cd8_graph <-
path2enet.BCELLCD19_TCELLCD8_TCELLCD4_NOTCHLOCAL_expr2barcode$BarcodeGraph

bcell_cd19_barcode_graph  <- graphTKplotterTissue(bcell_cd19_tcell_cd4_cd8_graph,
tissue="barcode_Pheno_BCELL_CD19+", tkplot=TRUE, value=0.4)

#Shows the On/Off state based on the selected treshold.

head(bcell_cd19_barcode_graph$state_OnOff)

tcell_cd4_barcode_graph  <- graphTKplotterTissue(bcell_cd19_tcell_cd4_cd8_graph,
tissue="barcode_Pheno_TCELLCD4+", tkplot=TRUE, value=0.4)

tcell_cd8_barcode_graph  <- graphTKplotterTissue(bcell_cd19_tcell_cd4_cd8_graph,
tissue="barcode_Pheno_TCELLCD8+", tkplot=TRUE, value=0.4)
```

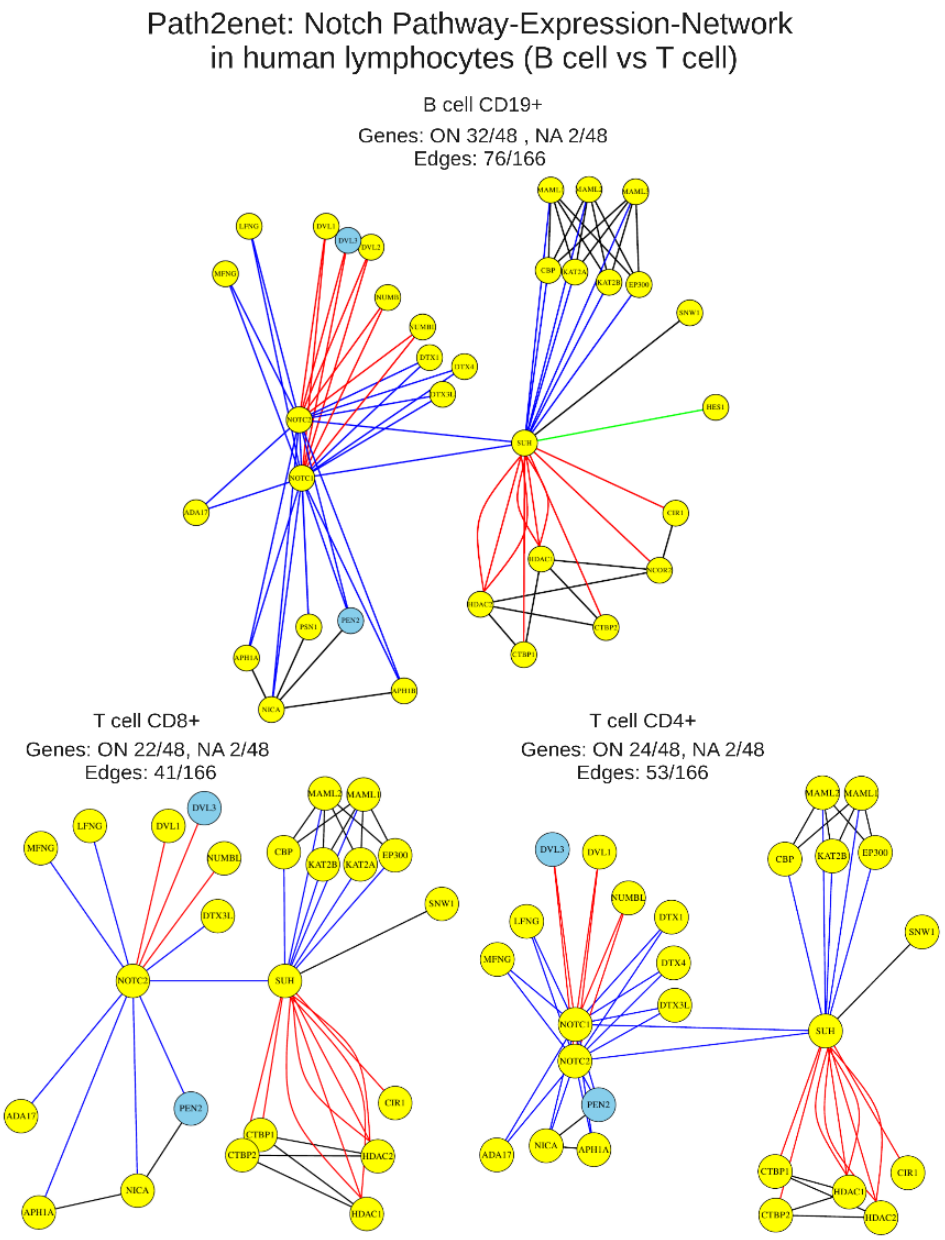


Figure 28 show the graphical representation of three generated graphs of Notch Signaling Pathway as Notch Pathway-Expression-Network in human lymphocytes (B cell vs T cell)

4.2.2.3 Defining the ON/OFF-status of nodes in biomolecular network

The Path2enet tool uses the **Gene Expression Barcode** algorithm to define if a gene is ON/OFF in a microarray dataset. Section 3.4.2 explains the algorithm in detail. **Figure 28** shows the results of the analysis. Using the treshold of 0.4 for the normalized expression, the B cell network expressed 34 of 48 of the NOTCH pathway proteins. In contrast, the T cell network expressed 22–24 of the NOTCH proteins. The discussion part of the Path2enet tool will explain this results.

***Table 19**: Results of the analysis performed for the NOTCH Signaling Pathway (including 48 genes) with Path2enet based on the gene expression Barcode algorithm for the datasets of B cell CD19+, T cell*

CD4+ and T cell CD8+. The datasets correspond to 163 samples of expression microarrays. The treshold to indicate if a gene was expressed (ON) or (OFF) is 0.4. Genes labeled with NA are not present in the expression platform and thus could not be annotated

Notch Signaling Pathway in human lymphocytes using *Path2enet* (expression level ON/OFF done with *Barcode*)							
Gene Symbols (KEGG)	Uniprot IDs	Bcell_CD19+		Tcell_CD4+		Tcell_CD8+	
ADAM 17	ADA17_HUMAN	0.802	ON	0.538	ON	0.621	ON
APH1A	APH1A_HUMAN	0.766	ON	0.568	ON	0.552	ON
APH1B	APH1B_HUMAN	0.594	ON	0.344	OFF	0.138	OFF
CREBBP	CBP_HUMAN	0.760	ON	0.622	ON	0.667	ON
CIR1	CIR1_HUMAN	0.969	ON	0.813	ON	0.517	ON
CTBP1	CTBP1_HUMAN	0.750	ON	0.852	ON	0.767	ON
CTBP2	CTBP2_HUMAN	0.445	ON	0.583	ON	0.578	ON
DLL1	DLL1_HUMAN	0.203	OFF	0.292	OFF	0.310	OFF
DLL3	DLL3_HUMAN	0.047	OFF	0.052	OFF	0.103	OFF
DLL4	DLL4_HUMAN	0.125	OFF	0.063	OFF	0.138	OFF
DTX1	DTX1_HUMAN	1.000	ON	0.417	ON	0.172	OFF
DTX2	DTX2_HUMAN	0.219	OFF	0.083	OFF	0.069	OFF
DTX3	DTX3_HUMAN	0.016	OFF	0.292	OFF	0.345	OFF
DTX3L	DTX3L_HUMAN	1.000	ON	0.990	ON	0.931	ON
DTX4	DTX4_HUMAN	1.000	ON	0.615	ON	0.207	OFF
DVL1	DVL1_HUMAN	1.000	ON	0.979	ON	1.000	ON
DVL2	DVL2_HUMAN	0.578	ON	0.266	OFF	0.259	OFF
DVL3	DVL3_HUMAN	–	NA	–	NA	–	NA
EP3000	EP300_HUMAN	0.906	ON	0.760	ON	0.810	ON
HDAC1	HDAC1_HUMAN	1.000	ON	1.000	ON	1.000	ON
HDAC2	HDAC2_HUMAN	0.656	ON	0.510	ON	0.517	ON
HES1	HES1_HUMAN	0.594	ON	0.198	OFF	0.103	OFF
HES5	HES5_HUMAN	0.000	OFF	0.031	OFF	0.034	OFF
JAG1	JAG1_HUMAN	0.227	OFF	0.224	OFF	0.069	OFF
JAG2	JAG2_HUMAN	0.031	OFF	0.286	OFF	0.172	OFF
KAT2A	KAT2A_HUMAN	0.750	ON	0.385	OFF	0.483	ON
KAT2B	KAT2B_HUMAN	1.000	ON	1.000	ON	0.828	ON
LFNG	LFNG_HUMAN	0.719	ON	0.510	ON	0.621	ON
MFNG	MFNG_HUMAN	0.891	ON	0.964	ON	0.966	ON
RFNG	RFNG_HUMAN	0.094	OFF	0.219	OFF	0.034	OFF
MAML1	MAML1_HUMAN	0.969	ON	1.000	ON	1.000	ON
MAML2	MAML2_HUMAN	0.719	ON	0.661	ON	0.810	ON
MAML3	MAML3_HUMAN	0.547	ON	0.281	OFF	0.069	OFF
NCOR2	NCOR2_HUMAN	0.431	ON	0.267	OFF	0.317	OFF
NCSTN	NICA_HUMAN	0.766	ON	0.479	ON	0.552	ON
NOTCH1	NOTC1_HUMAN	0.453	ON	0.427	ON	0.379	OFF
NOTCH2	NOTC2_HUMAN	0.914	ON	0.680	ON	0.664	ON
NOTCH3	NOTC3_HUMAN	0.031	OFF	0.073	OFF	0.069	OFF
NOTCH4	NOTC4_HUMAN	0.016	OFF	0.068	OFF	0.069	OFF
NUMB	NUMB_HUMAN	0.828	ON	0.341	OFF	0.371	OFF
NUMBL	NUMBL_HUMAN	0.516	ON	0.521	ON	0.569	ON
PSEN1	PSN1_HUMAN	0.581	ON	0.350	OFF	0.290	OFF
PSEN2	PSN2_HUMAN	0.281	OFF	0.104	OFF	0.046	OFF
PSENEN	PEN2_HUMAN	–	NA	–	NA	–	NA
PTCRA	PTCRA_HUMAN	0.146	OFF	0.323	OFF	0.253	OFF
SNW1	SNW1_HUMAN	0.984	ON	0.990	ON	0.914	ON
RBPJ	SUH_HUMAN	0.740	ON	0.691	ON	0.690	ON
RBPJL	RBPJL_HUMAN	0.000	OFF	0.063	OFF	0.103	OFF
		N proteins ON	32/48		24/48		22/48

4.3 Qualitative protogenomics analysis of lymphoma B-cell line *Ramos*

In the presented thesis two proteogenomic studies are included. Botch studies have the focus on the qualitative analysis of the transcriptomic and proteomic datasets of the lymphoma B-cell line *Ramos*. The first study (section 4.3.1) compares a genome-wide expression high-density oligonucleotide microarray transcriptomic dataset with a proteomic mass-spectrometry (LC MS/MS) dataset of four cell compartments (nucleus, cytoplasm, organelle, membrane). This study is a general approach how transcriptomic data and proteomic data complete each other to get deeper and better insights into the biological processes in the cell.

Section 4.3.2 describes a qualitative proteogenomic with the combination of RNA-Seq transcriptomics, MS/MS proteomics and antibody-based affinity proteomics (SEC-MAP). The SEC-MAP dataset is a focused approach on 413 lymphoma relevant proteins. For this study the participant researchers selected these proteins.

4.3.1 Global view of transcriptomic and proteomic data

The following paragraphs show the results of the global qualitative proteogenomic study of the lymphoma B-cell line *Ramos*. This research was done in cooperation with the Proteomics Unit. Cancer Research Centre (IBMCC, CSIC/USAL/IBSAL) directed by Dr.Manuel Fuentes. This group was involved in the design of the study, contributed the proteomic dataset and was involved in the discussion of the results and writing the manuscript of the resulting publication "**Integration of Proteomics and Transcriptomics Datasets for the Analysis of a Lymphoma B-Cell Line in the Context of the Chromosome-Centric Human Proteome Project**" **(Díez et al., 2015)**. The part of this work presented here corresponds to the processing, integration, analysis and visualization of the proteogenomic data in order to gain biological insights.

4.3.1.1 Processing the datasets

This paragraph describes the proteomic datasets the Proteomics Unit provided. The transcriptomic dataset of three microarrays of the *Ramos* cell line was identified and fully processed by our group. **Figure 30**. shows the general strategy of our study.

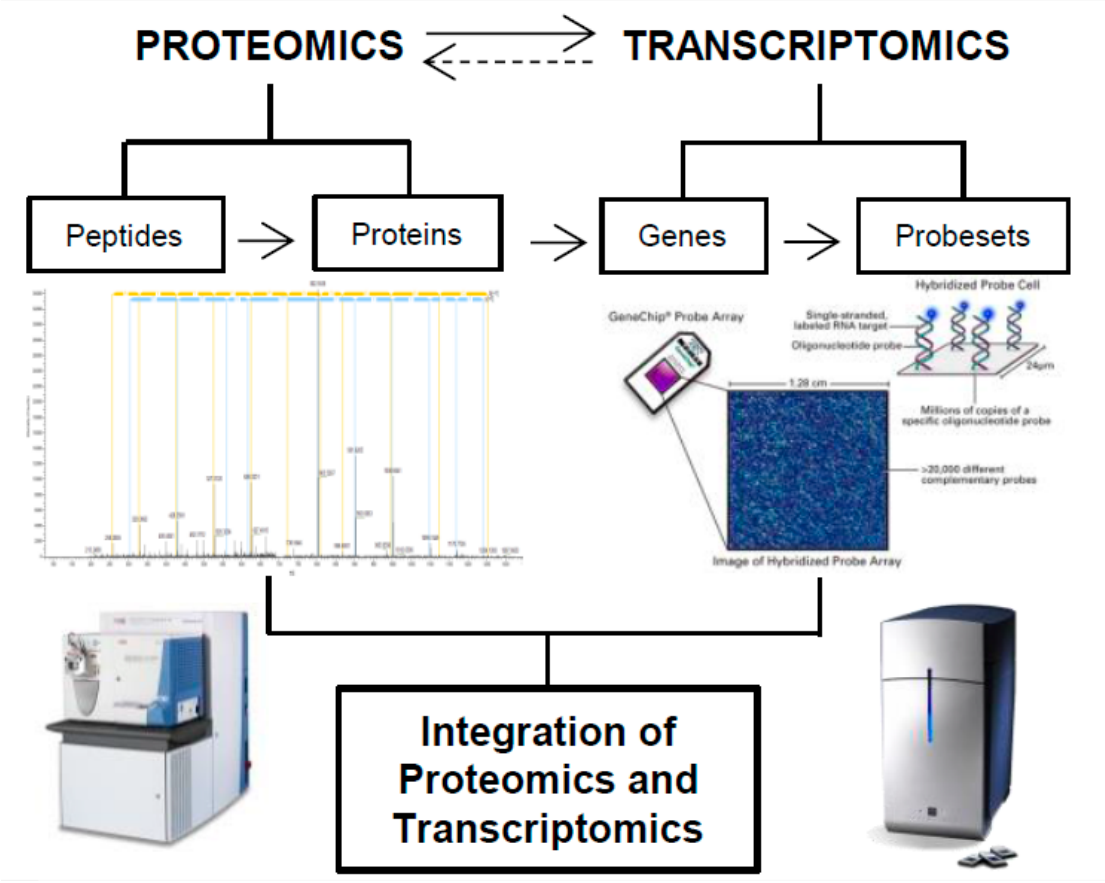


Figure 30: Integration of proteomics and transcriptomics workflow. Comparison of proteomics and transcriptomics data was made via mapping from peptides (obtained by an LC-MS/MS strategy) to DNA probes (Affymetrix Human Gene 1.0 platform).

4.3.1.1.1 Proteomic dataset

Original datasets provided:

The proteomic data was based on the number of proteins identified by one peptide (first dataset) or more then one peptide (second dataset). We analysed 4 different fragments cytoplasm, organelle, membrane, and nucleus. We had three experimental replicas. In total 2x4x3 txt files were integrated and compared. The datasets included following information: protein accession number of Uniprot database (isoforms are included), protein name, percent sequence coverage (Coverage), predicted molecular weight (Predicted MW [Da]), total number of spectra assigned above the specified confidence level cutoff (PSMs), number of unique peptides assigned above the specified confidence level cutoff (# Unique Peptides), and number of total peptides assigned above the specified confidence level cutoff (# Peptides). The protein evidence group (PE group) information was included in the dataset processed with the neXtprot database search (release 2014-09-19). In **Figure 31** the schema of generating the dataset is shown.

Processing of data:

This study is qualitative, so that the main information were **(i)** at least one proteotypic peptide protein is detected in the proteomic dataset; **(ii)** the protein is detected by more than one proteotypic peptide to have more confidence that the protein is in the dataset; and **(iii)** the cellular compartment in which a protein was found. This study did not focus isoform, so for the analysis this information is removed from the identifiers.

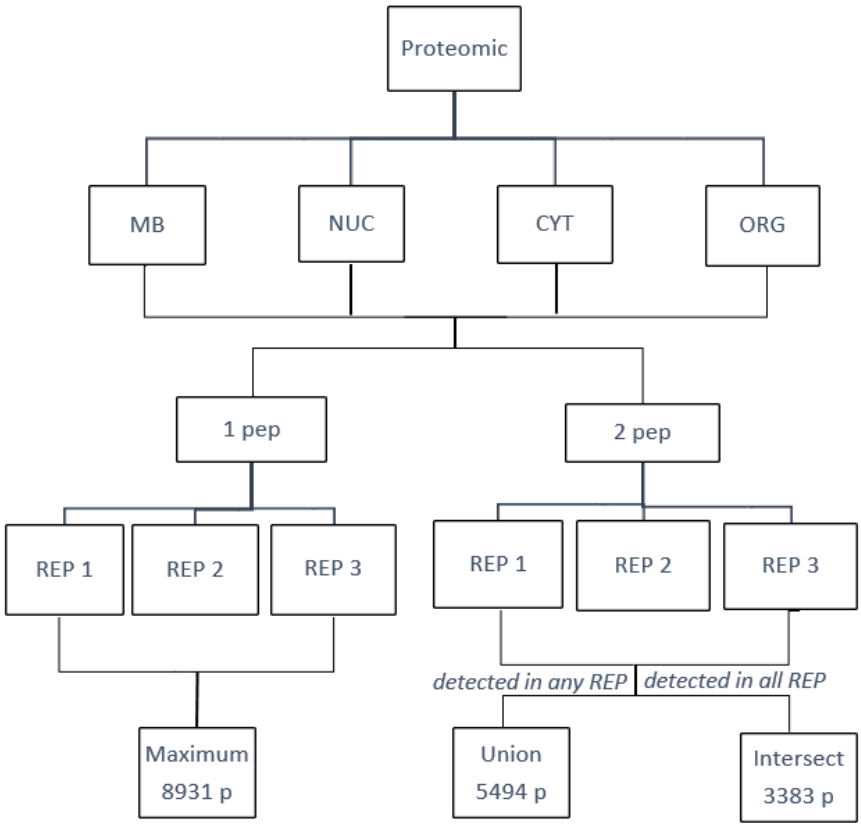


***Figure 31**: Structure of the original proteomic dataset and the generation of the Maximum, Union and Intersect datasets.*

The main steps to process the datasets are:

(1) Remove isoform information of the neXtProt identifier / Uniprot ACs
(2) Unification of the proteins of the subcellular fractions (i.e. cytoplasm, organelles, membranes, and nucleus) of each biological replicate
(3) Comparison of the three unified datasets of at least one peptide and at least two peptides
(4) Creation of datasets with different levels of coverage and confidence based on the results of step 3

The results are three mayor datasets: **(i)** *Intersection*: proteins with at least 2 peptides in all three replicates – this is the dataset has the highest confidence but lowest coverage; **(ii)** *Union*: proteins with at least 2 peptides in any replicate; and **(iii)** *Maximum*: proteins with at least 1 peptide in any replicate – this dataset has the highest coverage. **Table** X contains the number of proteins in each dataset and supplementary file X contains the proteins identified in each dataset.

Table 20*: The 3 different datasets generated out of the proteomic LC-MS/MS data.*

Dataset	Number of proteins (neXtProt)	Peptides	Confidence/Coverage
Intersection	3,383	2 in all 3 replicates	High / Low
Union	5,494	2 in 1 replicate	High / Medium
Maximum	8,931	1 in 1 replicate	Medium / High

4.3.1.1.2 Transcriptomic dataset

The transcriptomic dataset of the lymphoma B-cell line *Ramos* are *genome-wide expression high-density oligonucleotide microarray. These microarrays are downloaded of the Gene Expression Omnibus* (51) *[GEO] database [*GSE40168: GSM987747, GSM987748, GSM987749; platform GPL6244 ([HuGene-1_0-st])].

In this study we used the oligo and pd.hugene.1.0.st.v1 R-packages of Bioconductor to process the data. These packages are described in paragraph 3.4.2. The mayor steps are:

(1) Downloading and storage of the .CEL files
(2) Read in of the data (read.celfiles)
(3) Normalization of the three arrays with rma
(4) Getting the expression set matrix of the analysis
(5) Calculation of the mean of all three microarrays and storing it with the expression set matrix

In total 33,297 probsets are in the oligonucleotide microarray. To have a qualitative dataset of genes which are certainly expressed in the *Ramos* cell line, we set a threshold of the highest 25% of the expression dataset. In this study we could not use the barcode algorithm, because the function does not include a vector for the platform HuGene-1_0-st.

4.3.1.2 Integration and comparison of transcriptomic and proteomic data

In order to integrate the proteomic and transcriptomic datasets, both datasets needed one common identifier. The identifier used is Ensembl Gene ID, because **(i)** the neXtprot database provides a mapping table (nextprot_ensg.txt, release 2014-09-19) from neXtprot to Ensembl Gene ID; and **(ii)** The Brainarray Tool provides a mapping table (hugene10st_Hs_ENSG_mapping.txt; Version 18) from Affy Probe Sets to Enseml Gene ID, which has removed ambiguous Probe sets.

The next steps to integrate the datasets are, for
(i) the proteomic dataset:

(1) Mapping the neXtprot identifier of the *Intersect, Union and Maximum* to Ensembl Gene ID
(2) Find the Affy Probe Sets corresponding to the Ensembl Gene IDs
(3) Calculate the mean of the expression level of the Affy Probe Sets corresponding to an Ensembl Gene ID

(ii) the transcriptomic dataset, (complete dataset and highest 25 %):

(1) Mapping the Affy Probe sets to Ensembl Gene ID
(2) Calculate the mean of the Affy Probe Sets corresponding to one Ensembl Gene ID
(3) Mapping the Ensembl Gene ID to neXtprot ID
(4) Mapping the Ensembl Gene ID to Human Gene Symbol (to provide a better readable output)

The comparison of the transcriptomic and proteomic dataset are presented in **Table 22**. The datasets *Intersection*, *Union* and *Maximum* are explained in 4.3.1.1.1. They are mapped to 3433, 5540 and 8976 Ensembl Gene IDs which correspond to 4088, 6175 and 9494 Affy Probe Sets. The row **"Exclusive identifications" in Transcriptomics** means that these 1290 Ensembl Gene IDs were identified in the 25% of the highest expressed genes in the microarray platform, but not in the proteomic dataset. **"Exclusive identifications" in Proteomics** means that these proteins were identified in proteomics but not in the transcriptomic dataset. The corresponding Ensembl Gene IDs to the neXtProt identifier could not be mapped to Affy Probe Sets.

Table 21: [a] Proteins detected by at least 2 unique peptides in the MS proteomic experiments. [b] Proteins detected in all the 3 experimental biological replicates. [c] Proteins detected in any replicate. [d] Proteins detected by at least 1 unique peptide in the MS proteomic experiments. [e] Genes detected in the 25% higher expression quartile of the microarrays but not present in the MS data. [f] Proteins detected in the MS/MS data but not present in the expression microarrays.

	No. of Proteins (neXtProt)	**No. of Genes (Ensembl)**	
	Total neXtProt IDs	**ENSG IDs**	**Affymetrix probeset IDs**
Intersection [a, b]	3,383	3,433	4,088
Union [a, c]	5,494	5,540	6,175
Maximum [c, d]	8,931	8,976	9,494
"Exclusive identifications" in Transcriptomics [e]	-	**1,290**	-
"Exclusive identifications" in Proteomics [f]	**516**	-	-

The integration of proteomic and transcriptomic data makes it possible to compare the expression signal of the three proteomic datasets. **Figure 31** shows this distribution. Therefore the density plots of the log2 of the expression signal of the corresponding probesets and its boxplots were calculated. The expression of the whole expression signal corresponding to all 33,297 probesets included in the arrays (Purple) has the lowest expression mean. (Red) 9,494 probesets corresponding to 8,976 distinct human genes (ENSG IDs) and 8,391 proteins (neXtProt IDs), which showed at least one unique identifying peptide in the MS proteomic assays [maximum]. (Grey) 6,175 probesets corresponding to 5,540 distinct human genes (ENSG IDs) and 5,494 proteins (neXtProt IDs), which showed at least two identifying peptides in the proteomic assays [union]. (Green) 4,088 probesets corresponding to 3,433 distinct human genes (ENSG IDs) and 3,383 proteins (neXtProt IDs), which had at least two identifying peptides in proteomic assays and were found in all replicates of the proteomic isolations [intersection]. The figure shows: the more confident the proteomic signal is the higher is its expression.

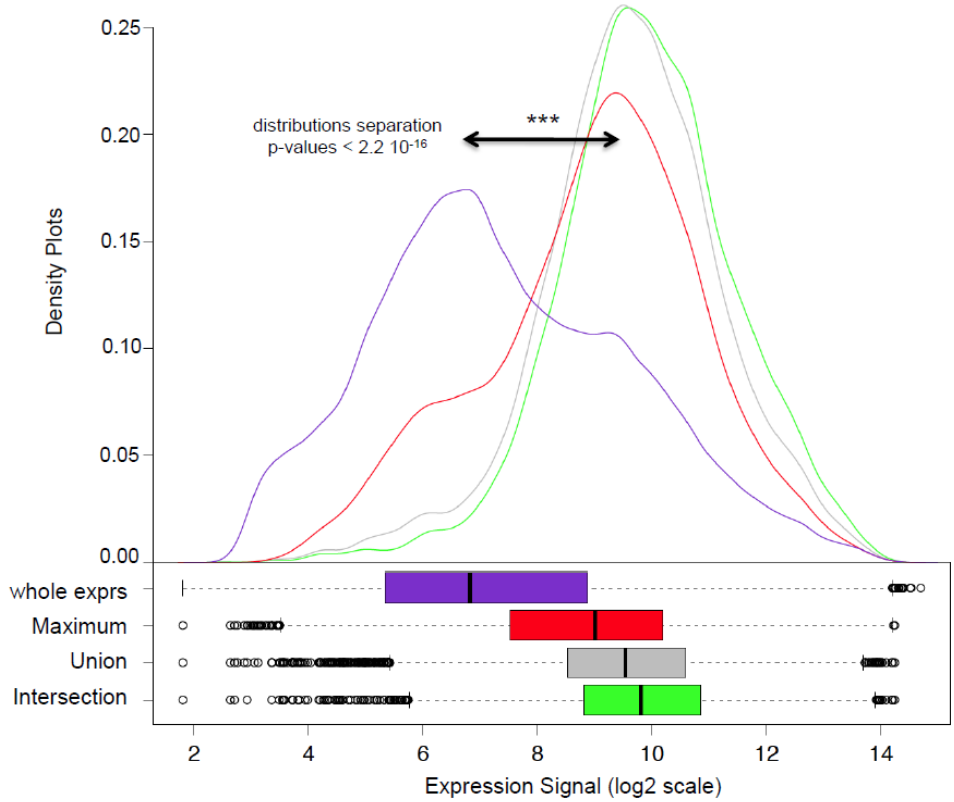


Figure 31: *Density plots and boxplots showing distributions of the whole gene expression signal versus the signal corresponding to genes detected in MS proteomic profiles for Ramos B-cells.*

4.3.1.3 Functional enrichment of proteomic and transcriptomic specific proteins/genes

This study includes a Functional Enrichment Analysis [FEA] off proteins identified in the proteomic Intersection datasets (**Table 23**), the 517 proteins exclusive in Proteomics (**Table 24**) and 1290 genes exclusive in Transcriptomics (**Table 25**). The FEA and clustering of the gene lists were done using the DAVID and GeneTerm-Linker tools. The main biological databases selected to find genes with annotated enriched terms were: **(i)** Gene Ontology (GO) using annotations spaces GOTERM_BP, GOTERM_CC and GOTERM_MF; **(ii)** the pathways database KEGG_PATHWAY; **(iii)** the INTERPRO and PFAM protein structural domain database; and **(iv)** the UNIGENE_EST and GNF_U133A_QUARTILE tissues-specific expression databases. To generate the functional clusters in DAVID we used classification stringency medium. All statistical analyses of data distributions, the comparisons and most of the mapping were done working in the R/Bioconductor environment.

The FEA of those 3,383 proteins detected in the *Intersection* dataset, which is coherent and steady revealed the expression of many essential proteins for general cell functions and house-keeping processes (e.g. anabolism and synthesis processes, together with catabolism and cellular respiration) as well as the activity and regulation of major biological macromolecules (DNA, RNA and proteins) involved in key maintenance processes like cell cycle, cell growth, cell proliferation, etc.

Table 22 *shows the results of the FEA analysis with the DAVID FAC tool of 1,290 exclusive identified in the highest expression quantile of the transcriptomic dataset. Each biological tearm enriched is shown in a row indicating the number of proteins assigned to such term and the corresponding number of proteins in the population to calculate*

the p-value enrichment using a hypergeometric test.

Annotation Database Categories	Biological Term enriched	N proteins in term	%	N proteins in Total	adj.p.value (Benjamini)	Genes IDs = 1290 g unique (total g mapped to functions = 904 g)	Fold Enrichment
	nucleus and nuclear genes						
SP_PIR_KEYWORDS	**nucleus**	**337**	37.28	893	**2.32E-24**	HMGN1, ITGB3BP, HMGN3, ZNF639, RP9, REST, CLI	**1.69**
GOTERM_CC_ALL	**GO:0005634~nucleus**	**375**	41.48	752	**3.06E-23**	HMGN1, ITGB3BP, HMGN3, ZNF639, RP9, REST, SGI	**1.56**
	transcription, gene expression, transcription factors						
SP_PIR_KEYWORDS	**transcription**	**196**	21.68	893	**4.41E-21**	ITGB3BP, ZNF639, REST, ZNF254, ZNF251, CTNNB1	**2.04**
SP_PIR_KEYWORDS	**transcription regulation**	**192**	21.24	893	**9.11E-21**	ITGB3BP, ZNF639, REST, ZNF254, CNOT6, ZXDA, ZN	**2.04**
GOTERM_BP_ALL	**GO:0010468~regulation of gene expression**	**244**	26.99	682	**1.43E-19**	ITGB3BP, ZNF639, REST, ZNF254, ZNF251, CTNNB1	**1.77**
GOTERM_BP_ALL	**GO:0045449~regulation of transcription**	**228**	25.22	682	**1.95E-19**	ITGB3BP, ZNF639, REST, ZNF254, ZNF251, CTNNB1	**1.81**
GOTERM_BP_ALL	**GO:0006350~transcription**	**195**	21.57	682	**9.98E-19**	ITGB3BP, ZNF639, REST, ZNF254, ZNF251, CTNNB1	**1.92**
GOTERM_BP_ALL	**GO:0010467~gene expression**	**223**	24.67	682	**1.23E-10**	ITGB3BP, MRPL42, ZNF639, RP9, REST, ZNF254, ZN	**1.54**
GOTERM_MF_ALL	**GO:0003676~nucleic acid binding**	**227**	25.11	701	**1.38E-09**	HMGN1, HMGN3, ZNF639, RP9, REST, ZNF254, ZNF2	**1.50**
GOTERM_MF_ALL	**GO:0008134~transcription factor binding**	**51**	5.64	701	**3.83E-05**	E2F1, ENY2, HMGN3, HTATIP2, E2F5, CTNNB1, CITE	**2.15**
GOTERM_MF_ALL	**GO:0003700~transcription factor activity**	**78**	8.63	701	**1.69E-04**	E2F1, ZNF83, E2F5, E2F7, NR6A1, ZNF639, ZXDA, CT	**1.73**
	zinc-finger like genes						
SP_PIR_KEYWORDS	**zinc-finger**	**179**	19.80	893	**1.92E-23**	MKRN1, ZNF639, RP9, REST, ZNF254, ZXDA, ZNF251	**2.24**
GOTERM_MF_ALL	**GO:0008270~zinc ion binding**	**202**	22.35	701	**4.34E-18**	ZNF639, RP9, REST, ZNF254, ZNF251, G2E3, ZNF77	**1.89**
UP_SEQ_FEATURE	**zinc finger region:C2H2-type 5**	**69**	7.63	893	**1.23E-10**	GTF3A, ZNF83, ZNF296, ZNF639, ZBTB39, ZNF675, F	**2.75**
UP_SEQ_FEATURE	**zinc finger region:C2H2-type 6**	**65**	7.19	893	**6.29E-11**	GTF3A, ZNF83, ZNF296, ZNF639, ZBTB39, ZNF675, F	**2.85**
INTERPRO	**IPR013087:Zinc finger, C2H2-type/integrase, DNA-binding**	**75**	8.30	784	**1.06E-10**	GTF3A, ZNF83, ZNF486, ZBTB34, ZNF675, REST, ZN	**2.57**
INTERPRO	**IPR015880:Zinc finger, C2H2-like**	**88**	9.73	784	**5.30E-11**	GTF3A, ZNF83, ZNF486, ZBTB34, ZNF296, ZMAT2, ZI	**2.35**
	apoptosis, cell death, cell cycle genes						
SP_PIR_KEYWORDS	**Apoptosis**	**45**	4.98	893	**1.01E-06**	ITGB3BP, E2F1, HTATIP2, CADM1, LITAF, NUAK2, GF	**2.54**
GOTERM_BP_ALL	**GO:0006915~apoptosis**	**57**	6.31	682	**9.43E-05**	E2F1, ITGB3BP, HTATIP2, CADM1, NUAK2, PMAIP1, S	**1.96**
GOTERM_BP_ALL	**GO:0012501~programmed cell death**	**57**	6.31	682	**1.47E-04**	E2F1, ITGB3BP, HTATIP2, CADM1, NUAK2, PMAIP1, S	**1.93**
SP_PIR_KEYWORDS	**cell cycle**	**43**	4.76	893	**8.40E-04**	CCNT2, E2F1, CKS1B, ING4, NEK2, ANAPC13, E2F7,	**2.01**
GOTERM_BP_ALL	**GO:0008219~cell death**	**61**	6.75	682	**0.001107626**	E2F1, ITGB3BP, HTATIP2, CADM1, NUAK2, PMAIP1, S	**1.76**
GOTERM_BP_ALL	**GO:0016265~death**	**61**	6.75	682	**0.001316266**	E2F1, ITGB3BP, HTATIP2, CADM1, NUAK2, PMAIP1, S	**1.74**
GOTERM_BP_ALL	**GO:0007049~cell cycle**	**63**	6.97	682	**0.002764063**	GAS2L3, E2F1, CCNT2, E2F7, OSGIN2, LATS1, CTNN	**1.68**
GOTERM_BP_ALL	**GO:0022402~cell cycle process**	**49**	5.42	682	**0.004480233**	GAS2L3, E2F1, BBS4, ING4, PPP2R3B, NEK2, ANAPC	**1.80**
GOTERM_BP_ALL	**GO:0006917~induction of apoptosis**	**30**	3.32	682	**0.032389471**	ITGB3BP, HTATIP2, CADM1, APH1A, PMAIP1, SERINC	**1.94**
GOTERM_BP_ALL	**GO:0012502~induction of programmed cell death**	**30**	3.32	682	**0.033012346**	ITGB3BP, HTATIP2, CADM1, APH1A, PMAIP1, SERINC	**1.93**
GOTERM_BP_ALL	**GO:0043065~positive regulation of apoptosis**	**37**	4.09	682	**0.034366558**	ITGB3BP, ING4, ING3, HTATIP2, CADM1, APH1A, PMA	**1.78**
	chromosome genes						
GOTERM_CC_ALL	**GO:0005694~chromosome**	**37**	4.09	752	**0.076807616**	HMGN1, ITGB3BP, HIST1H2AC, HMGN3, NEK2, REST	**1.70**
GOTERM_CC_ALL	**GO:0044427~chromosomal part**	**32**	3.54	752	**0.091331419**	HMGN1, ITGB3BP, HIST1H2AC, HMGN3, REST, ZNF3	**1.75**

The FEA analysis of the exlusively identified proteins in proteomics showed a gain of proteins related to the mitochondrial and ribosomal organelles, as well as cytoplasmic ones. This is a quite expected result because the genomic microarrays employed do not include probes for mitochondrial DNA. A loss of identifications associated to immunoglobulins (Ig) and major histocompatibility complex (MHC) proteins was also detected in the transcriptomic data probably due to Ig gene rearrangements and hypersomatic mutations of Ig genes that hamper the design of adequate array probes for these genes.

Table 23 *shows the results of the FEA analysis with the DAVID FAC tool of 516 proteins of the exclusive in proteomic identified proteins. Each biological tearm enriched is shown in a row indicating the number of proteins assigned to such term and the corresponding number of proteins in the population to calculate*

	ribosomal proteins & protein translation						
GOTERM_CC_ALL	**GO:0033279~ribosomal subunit**	**19**	4.60	333	**2.89E-08**	MRPS17, RPL17, MRPS24, RPL26, RPS15A, MRPS6,	**7.09**
SP_PIR_KEYWORDS	**ribosomal protein**	**21**	5.08	398	**4.49E-07**	MRPL53, MRPS17, RPL17, MRPS24, RPL26, RPS15A	**5.40**
SP_PIR_KEYWORDS	**ribonucleoprotein**	**25**	6.05	398	**5.57E-07**	MRPL53, MRPS17, RPL17, MRPS24, RPL26, RPS15A	**4.33**
GOTERM_CC_ALL	**GO:0005840~ribosome**	**22**	5.33	333	**3.26E-07**	MRPL53, MRPS17, RPL17, MRPS24, RPL26, RPS15A	**4.89**
GOTERM_CC_ALL	**GO:0030529~ribonucleoprotein complex**	**34**	8.23	333	**6.27E-07**	MRPS17, RPL17, RBM4, CWC15, RPS15A, RPL39, SF	**3.15**
SP_PIR_KEYWORDS	**ribosome**	**12**	2.91	398	**2.83E-05**	RPL17, RPS27, RPS28, RPL18A, RPL7, RPL13A, RPS	**7.94**
SP_PIR_KEYWORDS	**protein biosynthesis**	**18**	4.36	398	**3.10E-05**	RPL17, PDF, RPL26, RPS15A, RPS8, EIF4G1, EIF3C,	**4.63**
KEGG_PATHWAY	**hsa03010:Ribosome**	**13**	3.15	127	**8.65E-05**	RPL17, RPL26, RPS15A, RPL39, RPS8, RPS27, RPS2	**5.98**
GOTERM_BP_ALL	**GO:0006412~translation**	**26**	6.30	308	**1.07E-04**	RPL17, MRPS17, RPS15A, RPL13AP3, RPL39, EIF3C,	**3.60**
SP_PIR_KEYWORDS	**rna-binding**	**26**	6.30	398	**0.006213636**	RBM34, MRPS17, RBM4, RBM7, SRP19, YBX1, ZFP36	**2.33**
	immunoglobulin like and MHC proteins						
PFAM	**PF07654:Immunoglobulin C1-set domain**	**11**	2.66	345	**1.04E-04**	HLA-DQB1, HLA-DRB1, HLA-C, HLA-B, HLA-DMB, HLA	**9.77**
INTERPRO	**IPR003597:Immunoglobulin C1-set**	**12**	2.91	353	**1.22E-04**	HLA-DQB1, IGHG2, HLA-DRB1, HLA-C, HLA-B, HLA-DI	**8.33**
INTERPRO	**IPR003006:Immunoglobulin/major histocompatibility comp**	**13**	3.15	353	**1.02E-04**	HLA-DQB1, IGHG2, HLA-DRB1, ZNF841, HLA-C, HLA-I	**7.05**
GOTERM_CC_ALL	**GO:0042611~MHC protein complex**	**9**	2.18	333	**4.51E-04**	HLA-DQB1, HLA-DRB1, HLA-DRB4, HLA-C, HLA-B, HL	**7.54**
GOTERM_BP_ALL	**GO:0019882~antigen processing and presentation**	**11**	2.66	308	**0.006662349**	HLA-DQB1, HLA-DRB1, IFI30, HLA-C, HLA-B, HLA-DMI	**6.07**
GOTERM_MF_ALL	**GO:0032395~MHC class II receptor activity**	**6**	1.45	329	**0.007566434**	HLA-DQB1, HLA-DRB1, HLA-DRB4, HLA-C, HLA-B, HL	**14.53**
GOTERM_BP_ALL	**GO:0048002~antigen processing and presentation of pepti**	**7**	1.69	308	**0.010347178**	TAP2, IFI30, HLA-C, HLA-B, HLA-E, HLA-DMA, HLA-DR	**11.46**
KEGG_PATHWAY	**hsa04514:Cell adhesion molecules (CAMs)**	**10**	2.42	127	**0.074149415**	HLA-DQB1, HLA-DRB1, HLA-DRB4, HLA-C, HLA-B, HL	**3.03**
	mitochondrial proteins						
SP_PIR_KEYWORDS	**mitochondrion**	**37**	8.96	398	**0.001391843**	MRPS17, CHKB, PTPMT1, MRPL12, TOMM6, TOMM5	**2.15**
GOTERM_CC_ALL	**GO:0005739~mitochondrion**	**42**	10.17	333	**0.003133675**	MRPS17, CHKB, ALDOC, PTPMT1, APOBEC3F, MRPI	**1.85**
SP_PIR_KEYWORDS	**mitochondrion outer membrane**	**8**	1.94	398	**0.021233686**	CPT1B, MSTO1, TOMM6, TOMM5, PGAM5, CHKB, GI	**5.52**
GOTERM_CC_ALL	**GO:0031966~mitochondrial membrane**	**18**	4.36	333	**0.045084467**	CPT1B, ATP5J2, MSTO1, NDUFA9, CHKB, OTC, NDU	**2.18**
GOTERM_CC_ALL	**GO:0005740~mitochondrial envelope**	**18**	4.36	333	**0.076155089**	CPT1B, ATP5J2, MSTO1, NDUFA9, CHKB, OTC, NDU	**2.05**

The FEA analysis of exclusively identified genes in the highest quantile of the transcriptomic microarray dataset revealed identifications related to nuclear and DNA-binding proteins. This is probably due to the fact that isolating the nuclear fraction in the last step of the proteomics approach decreases the recovery for nuclear proteins. Additionally, around 300 of these transcripts exclusively identified in transcriptomics correspond to non-coding RNAs. Thus, it is obvious that their corresponding proteins have not been detected by the proteomics platform.

Table 24 *shows the results of the FEA analysis with the DAVID FAC tool of 3,383 proteins of the Intersection dataset obtained from Ramos B-cells. Each biological tearm enriched is shown in a row indicating the number of proteins assigned to such term and the corresponding number of proteins in the population to calculate the p-value enrichment using a hypergeometric test*

Annotation Database Categories	Biological Term enriched	N proteins in term	%	N proteins in Total	adj.p.value (Benjamini)	Proteins IDs (total number mapped = 2615 p)	Fold Enrichment
	general biological terms and functions						
GOTERM_BP_ALL	GO:0044237~cellular metabolic process	1420	54.45	2284	6.47E-54	ALDOA_HUMAN, NONO_HUMAN, PSPC1_HUMAN,	1.32
GOTERM_BP_ALL	GO:0019538~protein metabolic process	638	24.46	2284	3.22E-22	PCNP_HUMAN, RL4_HUMAN, RM19_HUMAN, SYR(	1.40
GOTERM_BP_ALL	GO:0010467~gene expression	628	24.08	2284	4.38E-13	SCML2_HUMAN, EXOS5_HUMAN, SNF8_HUMAN, N	1.29
GOTERM_BP_ALL	GO:0009056~catabolic process	306	11.73	2284	2.51E-13	ALDOA_HUMAN, HLTF_HUMAN, CATD_HUMAN, UI	1.51
GOTERM_BP_ALL	GO:0044248~cellular catabolic process	258	9.89	2284	1.15E-12	HLTF_HUMAN, CATD_HUMAN, UBAC1_HUMAN, D(	1.56
GOTERM_BP_ALL	GO:0044085~cellular component biogenesis	279	10.70	2284	3.14E-20	EXOS5_HUMAN, RL5_HUMAN, CATD_HUMAN, DK(	1.72
GOTERM_BP_ALL	GO:0070271~protein complex biogenesis	144	5.52	2284	1.01E-10	RANB9_HUMAN, MALT1_HUMAN, DECR_HUMAN,	1.76
GOTERM_BP_ALL	GO:0016070~RNA metabolic process	370	14.19	2284	4.79E-67	RU2A_HUMAN, EXOS5_HUMAN, RL5_HUMAN, NOI	2.44
GOTERM_BP_ALL	GO:0006396~RNA processing	255	9.78	2284	1.73E-61	RU2A_HUMAN, EXOS5_HUMAN, RL5_HUMAN, NOI	2.88
GOTERM_BP_FAT	GO:0008380~RNA splicing	141	5.41	2212	6.91E-36	RU2A_HUMAN, NONO_HUMAN, PR38A_HUMAN, H	3.04
GOTERM_BP_ALL	GO:0006259~DNA metabolic process	138	5.29	2284	8.26E-09	MCM5_HUMAN, ISG20_HUMAN, PNKP_HUMAN, M(	1.69
GOTERM_BP_ALL	GO:0034660~ncRNA metabolic process	119	4.56	2284	6.19E-33	RL5_HUMAN, EXOS5_HUMAN, WDR4_HUMAN, RR	3.20
GOTERM_BP_ALL	GO:0006399~tRNA metabolic process	62	2.38	2284	3.90E-17	THG1_HUMAN, WDR4_HUMAN, SYK_HUMAN, SYH	3.25
GOTERM_BP_ALL	GO:0015031~protein transport	257	9.85	2284	6.44E-32	SNF8_HUMAN, SRP09_HUMAN, RABE2_HUMAN, S	2.08
GOTERM_BP_ALL	GO:0046907~intracellular transport	240	9.20	2284	8.43E-36	AT2A2_HUMAN, SRP09_HUMAN, SCAM1_HUMAN,	2.26
	cell-specific biological terms and functions						
GNF_U133A_QUARTIL	PB-CD19+Bcells_3rd	1668	63.96	2271	9.21E-15	NONO_HUMAN, SAHH2_HUMAN, HNRPM_HUMAN,	1.11
CGAP_SAGE_QUARTI	1331:stem cell_null_3rd	557	21.36	2177	5.12E-130	ALDOA_HUMAN, EXOS5_HUMAN, ALDR_HUMAN, I	2.74
KEGG_PATHWAY	hsa04662:B cell receptor signaling pathway	26	1.00	1056	0.0430131	MALT1_HUMAN, CHP1_HUMAN, GRB2_HUMAN, NI	1.67

4.3.1.4 Identified missing proteins in the proteomic dataset

Missing proteins are proteins, which show evidence on the DNA level (open reading frames) and transcriptomic level (for example Microarrays, RNAseq, ESTs) that they exist, but were not detected in a proteomic expression set (MS/MS, antibodies). Therefore, we compared our proteomic datasets with the neXtProt database for missing proteins (release of 2015-04-28) in order to identify some in our dataset. In total we found 370 missing proteins in the Maximum dataset. In the datasets with lesser coverage but higher confidence we found in the Union dataset 32 proteins and 4 in the Intersection dataset. The missing proteins are also divided into 4 categories of protein existence [PE] (i.e., PE2 for experimental evidence at transcript level; PE3 for protein inferred from homology; PE4 for predicted protein; and PE5 for uncertain proteins). PE1 means they were detected on the proteomic level and of course excluded from the missing proteins.

Table 25. *Number of unidentified proteins found in the three different proteomic datasets and corresponding to four levels of protein existence (PE) classification.*

Dataset	PE2	PE3	PE4	PE5	TOTAL
Maximum	273	37	5	55	370
Union	18	3	-	11	32
Intersection	-	-	-	4	4

4.3.2 Focus on selected proteins in transcriptomic and proteomic data

The following paragraphs show the results of a qualitative proteogenomic study of the lymphoma B-cell line *Ramos* focused on selected proteins. The reason of the focus on proteins is the usage of affinity antibodies in the study. This research was done also in collaboration with the group of Dr. Manuel Fuentes resulting in the preparation of the following publication "**Comprehensive combination of affinity proteomics, MS/MS and RNA-Sequencing datasets for the analysis of a lymphoma B-cell line in the context of the Chromosome-Centric Human Proteome Project" (Díez et al. in preparation 2017)**. For this work we did the processing, integration, analysis and visualization of the proteogenomic data.

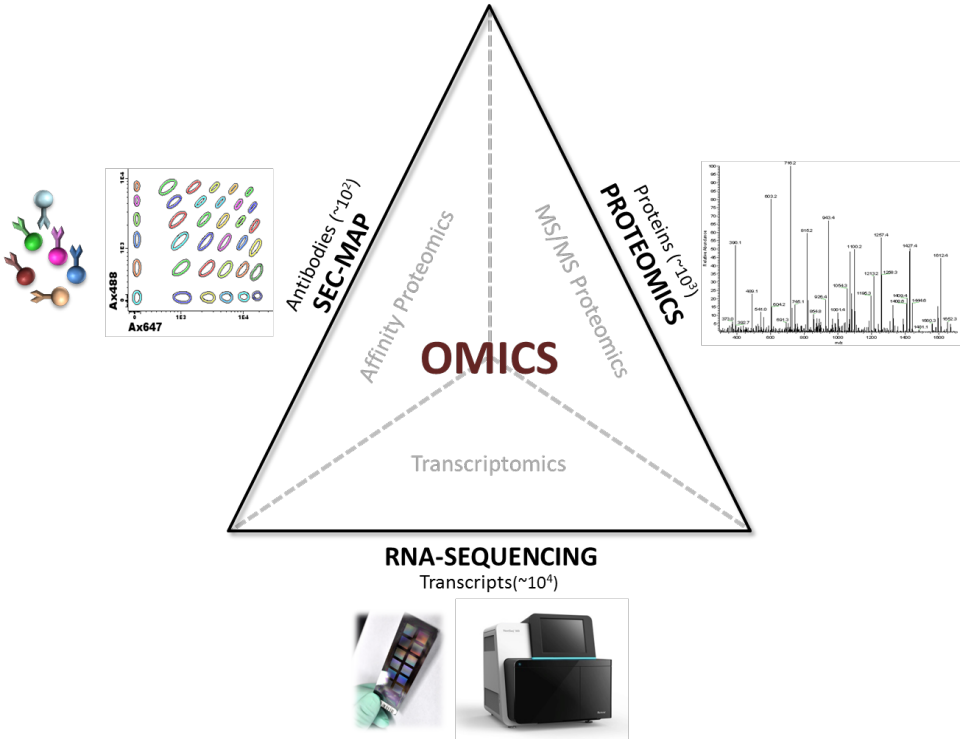


Figure 32: Overview of the different kind of transcriptomic and proteomic datasets to integrate them into an OMIC dataset to give a detailed characterization of the Ramos cell line.

In this study a multi-dimensional characterization of the protein profile of a Burkitt's lymphoma B-cell line based on the analysis of evidence at the protein level by affinity proteomics (Size Exclusion Chromatography- Microsphere-based Affinity Proteomics Arrays [SEC-MAP]) and MS/MS assays and its correlation with the number of gene products identified by RNA-Seq was done. **Figure 32** gives an overview of the characterization.

Section 4.3.2.1 describes the processing of the transcriptomic and proteomic datasets. The next paragraph explains the results of the integration and the proteogenomic analysis. Section 4.3.2.3 shows the results of the FEA analysis and the last paragraph 4.3.2.4 the missing proteins identified in this study.

4.3.2.1 Processing the datasets

In this study three kinds of biological datasets are used. The transcriptomic dataset is a global RNA-Seq dataset of the *Ramos* cell line. The proteomic datasets are a global Proteomics LC-MS/MS dataset of the same cell line and a 417 lymphoma relevant proteins focused screening with affinity antibodies [SEC-MAP]. **Figure 34** shows a schema of the processed datasets and

the objective to create an OMICS datasets of the B lymphoma cell line *Ramos*.

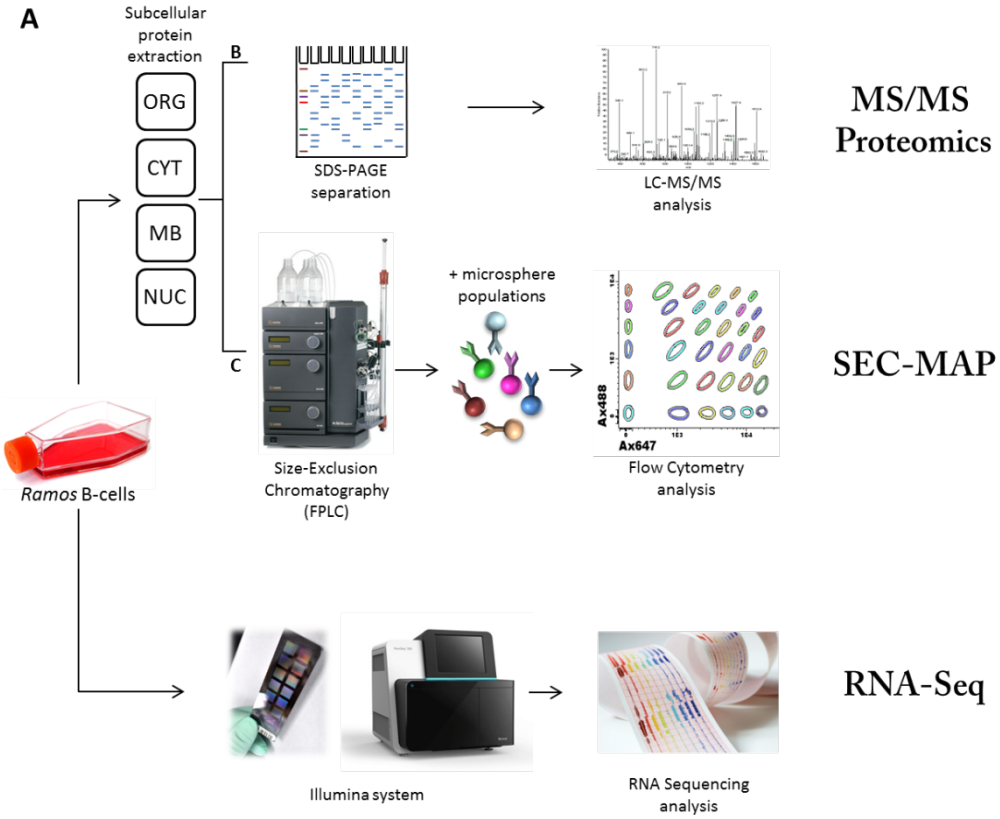


Figure 34: *Integration of global transcriptomic (RNA-Seq), global Proteomics (MS/MS) and on 417 proteins focused antibody dataset (SEC-MAP) to create a OMICS dataset of the B lymphoma Ramos cell line.*

4.3.2.1.1 Proteomic data

The Proteomics datasets contains a MS/MS-dataset similar to 4.3.1.1 and an affinity antibody dataset of 549 antibodies corresponding to 413 selected proteins.

LC-MS/MS dataset

The 12 MS/MS datasets (3 replicas x 4 subcellular compartments) contain the number of peptides found for each protein. First, all these datasets were unified in one and the number of peptides found per protein was summed up. In this way, we identified all the proteins containing at least one peptide in every of the MS/MS datasets (that we called in this study "complete mapping") and contains 5,707 proteins mapped to neXtProtIDs. The number of peptides in each replica and the number in each cellular compartment were also calculated and summed up. This provided seven subsets (1-3 replicas and CYT, MB, ORG, NUC). The genes associated with the proteins were mapped to chromosomes with the R-package BiomaRt.

SEC-MAP dataset

The SEC-MAP dataset contains the information of 549 antibodies corresponding to 417 proteins. In order to distinguish between detected and non-detected proteins, a Qualitative Antibody Score (QAS) System was established based on a signal peak detector tool. By QAS, it is possible to select the protein entities which have been detected by this approach based on the antibody reliability and discard the antibodies with poor or low performance. This score

was measured by an expert interpretation by the Proteomics Unit. Cancer Research Centre of the detected proteins on the depicted line plot. **Table** X explains the scoring scheme. The score defines if a protein based on the SEC-MAP experiment is in the *Ramos* cell line or not. The dataset is annotated to neXtProt-IDs

***Table 26:** . List of criteria for SEC-MAP evaluation. Microsphere array results were evaluated scoring the peaks detected in the SEC fractions. Peak identifications were positively and negatively scored according to the criteria shown in the table.*

Criterion	QAS value
1. Well defined peaks over background noise (≥140% difference between the minimum and the maximum value of the peak)	+1
2. The peak is observed in the expected subcellular fraction based on the literature (neXtProt database)	+1
3. Peak matching the expected molecular weight (MW) for the protein (+/-1 SEC fraction)	+1
4. When conditions 2 and 3 are both achieved	+1
5. Peaks present in fractions >19, as long as this localization does not match with the expectedly MW (i.e. the peak correspond with proteolytic forms)	-3
6. Two or more antibodies against the same target protein showing a highly similar elution profile*	+1

*This criterion must only be applied when the protein has already scored ≥2 points for the previous criteria. *SEC*, Size-Exclusion Chromatography

4.3.2.1.2 Transcriptomic data

The transcriptomic data corresponds to RNA-Seq for a *Ramos* B-cell line obtained with Illumina Genome Analyzer IIx with paired layout (experiment SRX105534: http://www.ncbi.nlm.nih.gov/sra/SRX105534) taken from the study SRP00931 (http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP009316). Section 3.4.3 explains the processing of these RNA-Seq files in order to gain *Fragments Per Kilobase Of Exon Per Million Fragments Mapped* [FPKM] per transcript.

In total 20,533 Ensembl Gene IDs are detected. The detected genes are mapped to neXtProt IDs with the ID maping table within the neXtProt database release 02-2016 (nextprot_ensg.txt) (52). 19,518 neXtProt IDs could be mapped within the RNA-Seq dataset. The value of 1 FPKM is the threshold to define if a gene/protein is present in the dataset or not. In the dataset 9,523 neXtProt IDs had FPKM>1.

4.3.2.2 Integration and comparison of transcriptomic and proteomic data

To create the proteogenomic dataset the transcriptomic RNA-Seq and Proteomics datasets MS/MS and SEC-MAP have the neXtprot ID as key identifier. Before all three datasets are combined, a global comparison of the MS/MS and RNA-Seq dataset is done.

To accomplish the -omics integration, the neXtProt IDs (corresponding to identified proteins by the MS/MS approach) were mapped into ENSG IDs (for the genes) identified by RNA-Seq approach. This was done using the ID maping table within the neXtProt database release 02-2016 (nextprot_ensg.txt) (52). In MS/MS proteomics we identified 5,707 proteins (neXtProts IDs) which refer to 5,982 genes (ENSG IDs). 5,672 out of 5,707 unique proteins of the MS/MS dataset could be also be found in the RNA-Seq dataset. RNA-Seq analysis allows the detection of 20,533 genes (ENSG IDs) corresponding to 19,518 neXtProt IDS.. When considering genes with an FPKM value ≥1, 5,157 out of these 5,672 mapped proteins were identified. **Table 2** gives an overview of the global comparison of the transcriptomic and proteomic datasets.

Thus, the combination of both datasets, once applied to the global molecular characterization of *Ramos* B-cell, displays a correlation of 91 % with high accuracy and great overlap within each analytical level.

Table 2. Integration of MS/MS proteomics and RNA-Seq transcriptomics [a].

	no. of genes (ENSG IDs)	no. of proteins (neXtProt IDs)
RNA-Seq[b]	20,533	19,518
RNA-Seq (with FPKM≥1)[c]	9,535	9,523
MS/MS[d]	5,982	5,707
Detection by MS/MS + RNA-Seq[e]	5,947	5,672
Detection by MS/MS + RNA-Seq (with FPKM≥1)[f]	5,165	5,157

[a]The table shows the number of gene and protein distinct IDs identified in the RNA-Seq transcriptomics and MS/MS proteomics assays, respectively: Row 1, RNA-Seq transcriptomics data; row 2, MS/MS proteomics data; row 3, MS/MS proteomics mapped to RNA-Seq transcriptomics; row 4, RNA-Seq transcriptomics mapped to neXtProt IDs; row 5, MS/MS proteomics mapped to RNA-Seq transcriptomics. The columns in the table correspond to **(i)** all the human neXtProt IDs detected; and **(ii)** the mapped Ensembl IDs. [b] genes detected in the RNA-Seq transcriptomics data independently of the FPKM value. [c] genes detected in the RNA-Seq transcriptomics data with FPKM≥1 mapped to neXtProt IDs. [d] proteins detected by at least 1 unique peptide in the MS/MS proteomics experiment. [e] detection by MS/MS proteomics and RNA-Seq transcriptomics. [f] detection by MS/MS proteomics and RNA-Seq transcriptomics presenting at least 1 unique peptide per protein and FPKM≥1 for MS/MS and RNA-Seq data, respectively.

The next step in the study was the integration of the SEC-MAP dataset with the selected 413 proteins. All three datasets are annotated to neXtprot IDs. **Figure 34** shows a density plots of the expression signal of $\log^2$ (FPKM+1) of the transcriptomic signal (19,518 ENSG) separated to the genes identified based on the dataset of the proteomic MS/MS (5,672 ENSG) and SEC-MAP (413 ENSG) approaches.

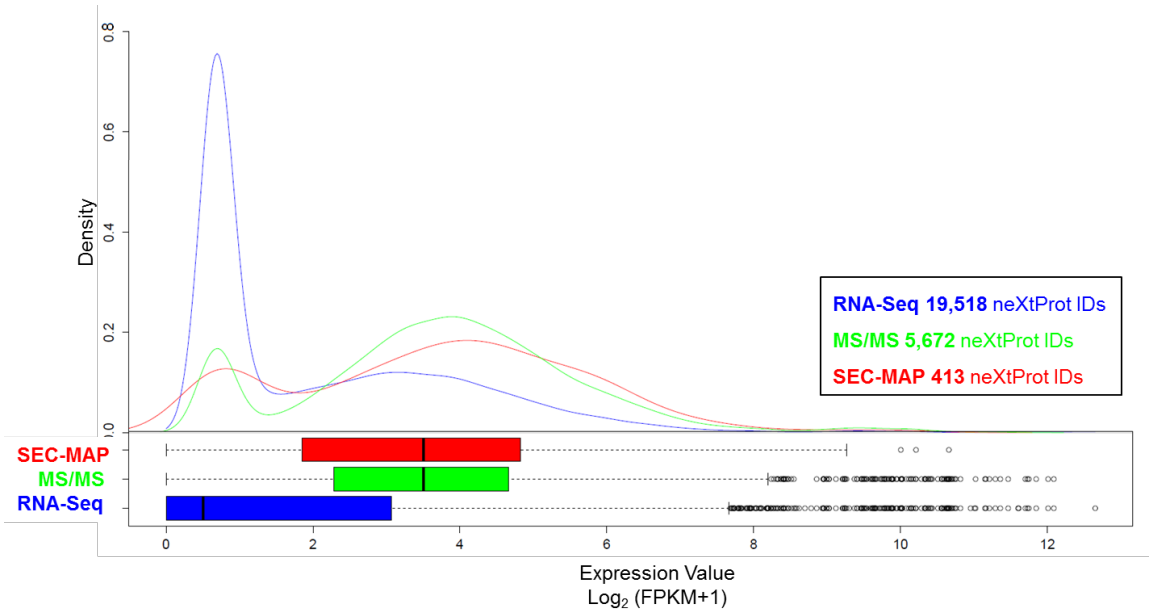


Figure 34: *Density plots and box plots of the distribution of the expression signal measured in log2 (FPKM+1) in the processed RNA- Seq dataset of Ramos B-cell line. The Figure compares the expression signal of all genes which could be mapped to 19,518 neXtProt IDs (RNA-Seq complete – 19,518 IDs, Blue) versus the signal of the genes corresponding to the proteins detected in the complete MS experiments (MS/MS-complete mapping – 5,672 neXtProt IDs, Green) and the proteins identified in the SEC-MAP experiment (SEC-MAP – 413 neXtProt IDS, Red).*

On the basis of the combination of these datasets, it seems that the highest reliable detection of proteins corresponds to proteins presenting at least 1 gene with FPKM ≥1; concerning MS/MS proteomics with at least ≥1 unique peptide for the target proteins; finally, proteins detected by SEC-MAP with a QAS value≥1. **Figure 35** shows a distribution of expression frequency of the 413 proteins based on these values.

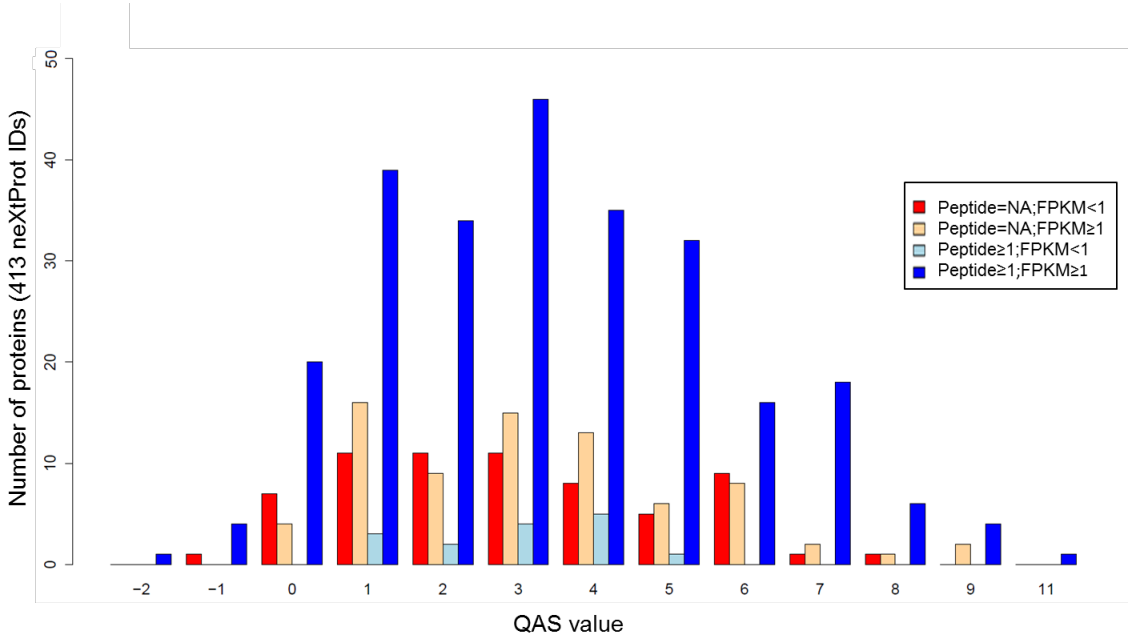


***Figure 35**: Comparison of the expression frequency of the 413 proteins present in the SEC-MAP array along the datasets of MS/MS, RNA-Seq, and SEC-MAP. A protein is considered as expressed in a dataset if its: (1) number of peptides ≥ 1; (2) log2 (FPKM+1) ≥1 [FPKM]; or (3) QAS value ≥ 1 [AB]. There are eight different stages: (1) Peptide=NA;FPKM<1 (Red) and AB<1: Protein non-detected in all datasets; (2) Peptide=NA; FPKM<1 (Red) and AB≥1: Protein only detected in SEC-MAP; (3) Peptide=NA; FPKM≥1(beige) and AB<1: Protein only detected in RNA-Seq; (4) Peptide=NA; FPKM≥1(beige) and AB≥1: Protein detected in RNA-Seq and SEC-MAP; (5) Peptide≥1; FPKM<1 (light blue) and AB<1: Protein only detected in MS/MS; (6) Peptide≥1; FPKM<1 (light blue) and AB≥1: Protein detected in MS/MS and SEC-MAP; (7) Peptide≥1; FPKM≥1 (dark blue) and AB<1: Protein detected in MS/MS and RNA-Seq; (8) Peptide≥1; FPKM≥1(dark blue) and AB≥1: Protein detected in all datasets.*

Once these conditions were established, the analysis revealed a 56% (231/413) overlapping between the 3 approaches (eg. CD79A, AIFM1, B2M, BAX, HRAS, PML), which are proteins related to B-cell receptor (BCR) crosslinking and linked intracellular signals, such as RAS activation pathways, and MYC interaction pathways, among others. Specifically, MS/MS proteomics identified 65.6% (271/413) proteins present in the SEC-MAP array; RNA-Seq identified 80.6% (333/413) proteins, and SEC-MAP 91.0% (376/413) proteins (**Figure 35**). It is also remarkable that 8 proteins (JUN, CD44, CALB2, IL3RA, CTBP2, SEPT5, CDC14A, and MAPRE3) were undetectable by the 3 approaches. The venn-diagramm in **Figure 36** visualises the overlap and differences of proteins in the datasets.

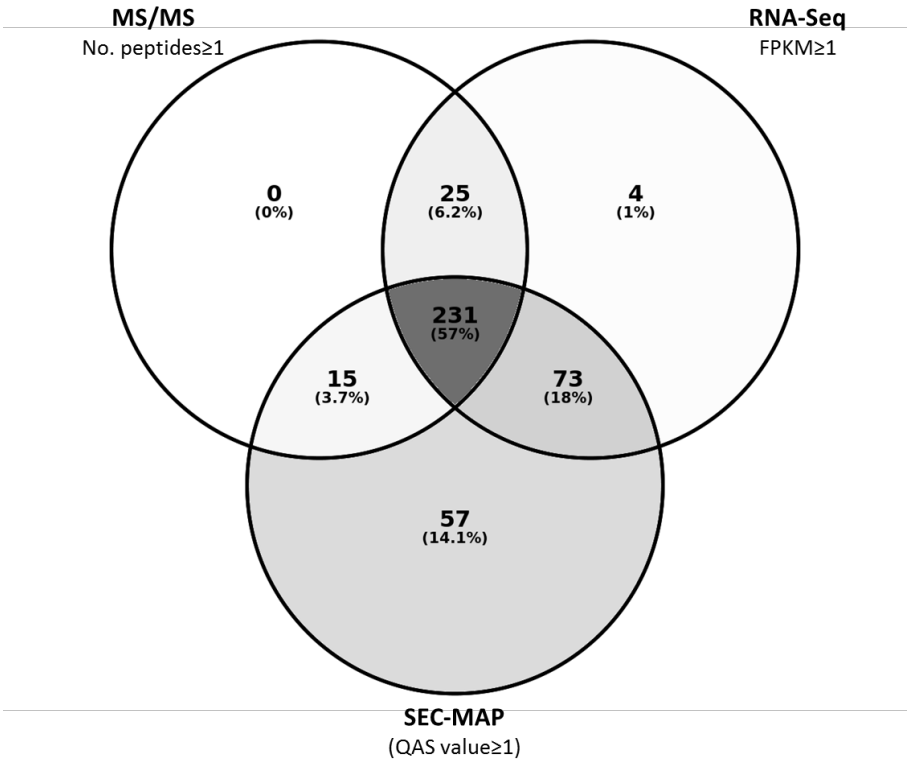


Figure 36: *Venn diagram of the expression of the 405 proteins out of the 413 included in the SEC-MAP array. A protein is considered as expressed in a dataset if its: (1) number of peptides≥1; or (2) FPKM≥1; or (3) the QAS value≥1 for MS/MS, RNA-Seq, and SEC-MAP approaches, respectively. A total of 231 proteins is detected in all three approaches. 8 proteins are non-detected in any of the mentioned approaches and they are not represented in the Venn diagram. SEC-MAP detected 376 proteins with a QAS value≥1; RNA-Seq detected 333 proteins with an FPKM≥1, and MS/MS proteomics detected 271 proteins with at least 1 peptide/protein.*

Also the overlapping between proteomics and transcriptomics was high, there is a significant variation when focusing on specific proteins. Thus, the integration of new –omics strategies (as SEC-MAP) implies the increasing of the knowledge about a specific cell or disease. In this study, it has been depicted that 13.8% proteins (57/413) were only detected by the SEC-MAP approach against the 1% (4/413) only detected by RNA-Seq and 6.1% (25/413) detected by RNA-Seq and MS/MS proteomics. This analysis also reveals that 3.6 % (15/413) proteins (with very low transcripts levels) are accurately detected by SEC-MAP and MS/MS, which could be due to the enrichment achieved by affinity reagents and subcellular fractionation. Of course, these identified proteins require a further validation due to lack of previous evidence at the transcript level since the RNA-Seq technique is only a snapshot of the transcriptomic level at one specific point.

4.3.2.3 Functional enrichment of the proteins

To identify the functions of the proteins/genes detected in the MS/MS and RNA-Seq dataset but not in SEC-MAP dataset and reverse two FEA are performed. Another FEA with all the proteins detected in all three datasets is done.

A FEA was done with the 57 proteins exclusively identified by the SEC-MAP array (Supporting Information S8) showing enrichment in functions related to membrane proteins, including receptors, plasma membranes, cell adhesion molecules and transmembrane proteins. These results suggest that immune-affinity allows the specific and selective identification of proteins which could be difficult to detect by other approaches due to several methodological challenges, such as isolation of subcellular compartment, suitable proteotypic enzymes, or relative low abundance.

On the other side, 29 antibodies (eg. CD22, AUKA, CASP3, TRAF1….) did not efficiently work or were totally ineffective (QAS value<1) for detecting proteins by SEC-MAP; even testing more than one antibody clone against the same protein. Most of these 29 proteins were accurately detected by both MS/MS proteomics and RNA-Seq. The FEA revealed that functions including lumen and nuclear proteins, phosphorylated intracellular signaling proteins, and GTPase regulation was undetectable by SEC-MAP. These results could be expected because highly modified proteins (including post-translational modifications, PTMs) will require highly specific antibodies; in addition, specific organelle/membrane isolation protocols will be required for detection of proteins exclusively located in these subcellular compartments.

4.3.2.4 Missing proteins

Like mentioned in point 4.3.1.4 researcher try to find Proteomics evidence of proteins with DNA or transcriptomic proof, which have not been described in a proteomic dataset. These proteins are not easy to detect by MS/MS approaches since their characteristics or abundance in samples is difficult to overcome. In this study, we have mapped our results into the neXtProt database (release 2016-02). **Table 27** shows the results separated by proteomic evidence PE

Table 27: *Number of unidentified proteins in MS/MS dataset separated by protein evidence (PE)*

Dataset	PE2	PE3	PE4	PE5	TOTAL
Complete Mapping	27	9	1	0	37

4.4 Proteome and phosphoproteome in CLL B-cells

The following paragraphs show the results of an overall quantitative proteome and phosphoproteome of primary cells from patients with B-CLL and CLL-like monoclonal B-cell lymphocytosis (MBL) and the phosphoproteins potentially involved in the baseline pathological B-cell behavior using a high-resolution massspectrometry-based approach. Five tumor B-cell proteomes and phosphoproteomes from different chronic lymphocytic leukemia (B-CLL) are studied. This research was also done in collaboration with Dr. Manuel Fuentes group preparing a publication entitled "**Revealing Cell Signaling Pathways in Chronic Lymphocytic Leukemia Tumor B-cells by Integration of Global Proteome and Phosphoproteome Profiles**" **(Díez et al. submitted 2017)**. In this work we did the processing, analysis and visualization of the proteogenomic data in order to gain biological insights.

Section 4.4.1 explains the steps to process the proteomic and phosphoproteomic datasets. The next section describes the qualitative study and 4.4.3 the quantitative analysis proceed in this study.

4.4.1 Processing the dataset

In this study it is not only important if a protein is expressed, it is also important on which level a protein is expressed in a patient. The study contains samples of 4 CLL patients and 1 CLL-like MBL patient. **Figure 37** gives an overview of the dataset and the differences in the B-CLL and MBL patients.

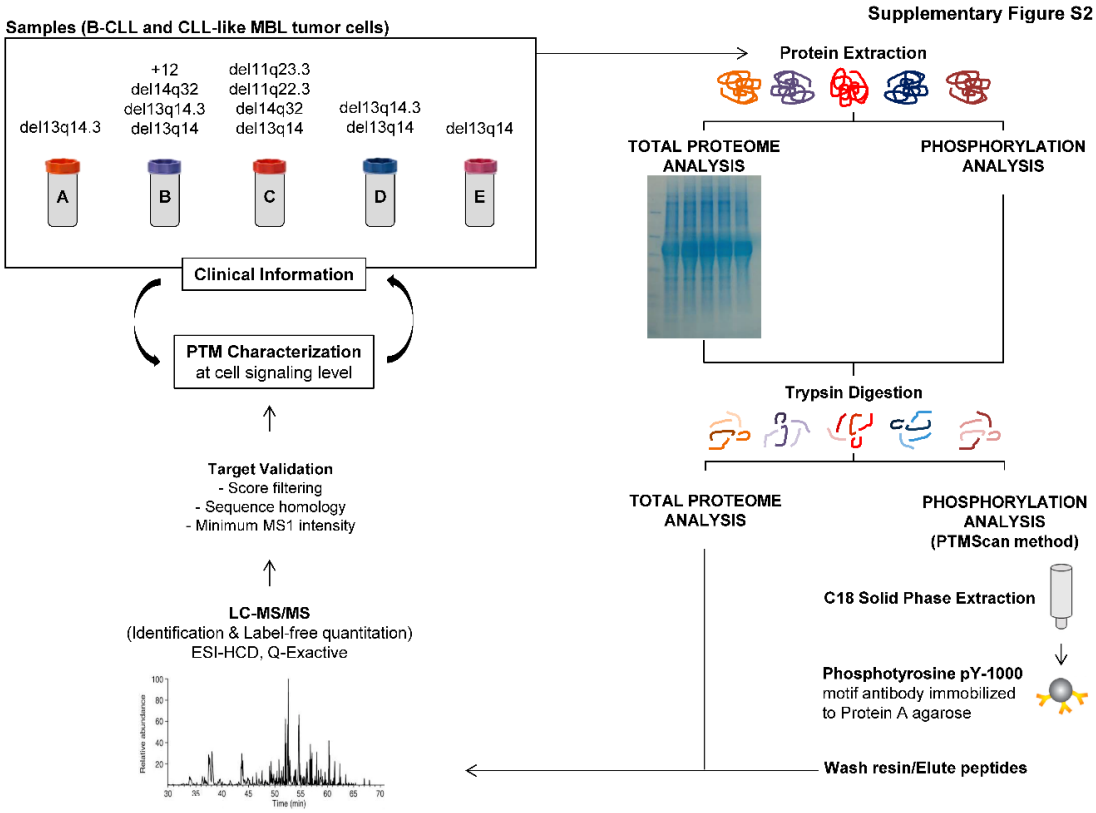


***Figure 37**: General overview of the procedure. CLL/MBL tumor B cell samples were processed to extract the proteins and the analysis of the proteome and the phosphoproteome were carried out in parallel. The PTMScan method (from Cell Signalling Technology) was performed by using a phosphotyrosine pY-1000 antibody. An LC-MS/MS approach was employed for the identification of the proteins and phosphoproteins and the results were integrated with the clinical data.*

Of each patient 2 independent replicas on proteome and phosphoproteome are analysed. The dataset received of the Proteomics Unit contained for each peptide an expression value. This raw dataset had to proceed quality measurements on peptide level:

(1) Removing all ambiguous peptides

(2) Removing all peptides if one replicate doubles the other replicate (0.5>Min/Max), the peptide is considered as untrustworthy and the value is replaced with NA

In total 13,504 unique unambiguous peptides corresponding to 2,970 proteins are in the proteome dataset. 594 unique unambiguous peptides corresponding to 327 proteins are in the phosphoproteome dataset.

(3) The mean of the peptides in the two replicas of each patient is calculated.
(4) If the mean of the replicates is below the threshold of 500,000 the value of this peptide is set to 0
(5) The mean of the peptides corresponding to a protein is accumulated

The processed dataset shows various protein identifiers (UniProt ID, UniProt-Swissprot, UniProt-Accession), the genomic identifier in UniProt, description of the protein, experimental evidence at protein level PE, and chromosomal location of the protein. The sequences of the combined peptides per protein are separated with ";" and its number of peptides is shown in the "Peptid-Count" column. The table shows the expression per replicate after the processing and the mean of the signal. The difference of these levels and the z-score are calculated as well. In this study FEA analysis for proteins who differentiate patients in the quality and/or quantitative way were done with the DAVID-tool explained in the method section 3.4.3

4.4.2 Qualitative analysis

The qualitative analysis part is similar to the section 4.3 but in this case we have one proteomic and one phosphoproteomic dataset of 4 CLL patients and 1 CLL-like MBL patient. Therefor we do not need to map the identifiers to any other identifier. The value if a protein is present (ON) or absent (OFF) in a patient is if the protein expression value is greater than 0. This means **(i)** at least the mean of one peptide in the two replicas has to pass the threshold of 500,000; and **(ii)** has to be a trustworthy peptide. In paragraph 4.4.2.1 we compare the patients to each other and in paragraph 4.4.2.2 we compare the proteome and phosphoproteome.

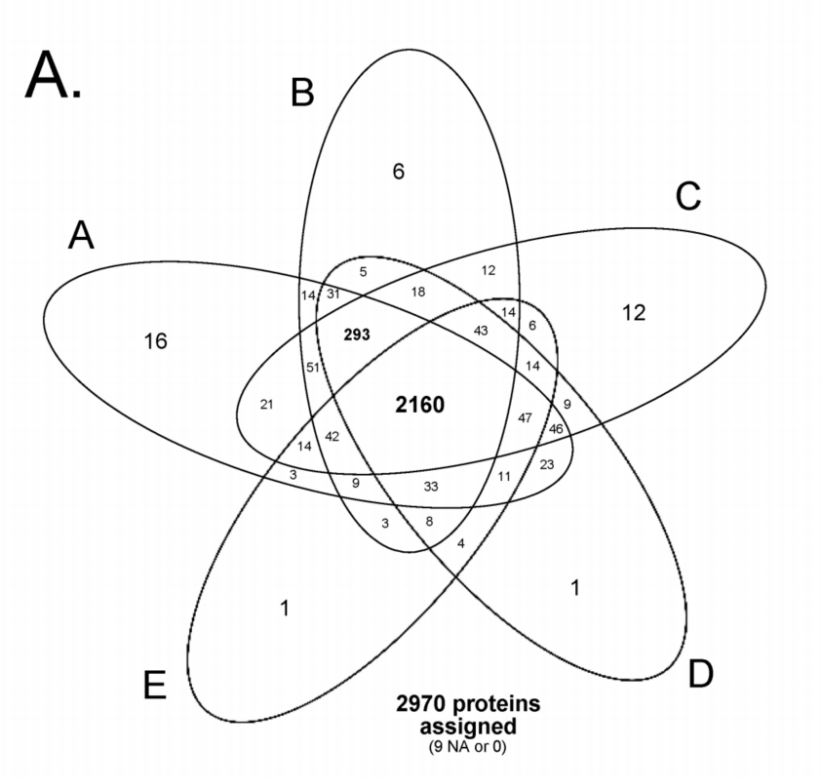


Figure 38: Venn-diagram representation of proteins identified in the CLL (A-D) and MBL tumor B-cell samples (E) analyzed. This Figure illustrates how many proteins were expressed in common or in only one or a subset of the 5 samples analyzed for the whole proteome dataset (2,979 proteins in total)

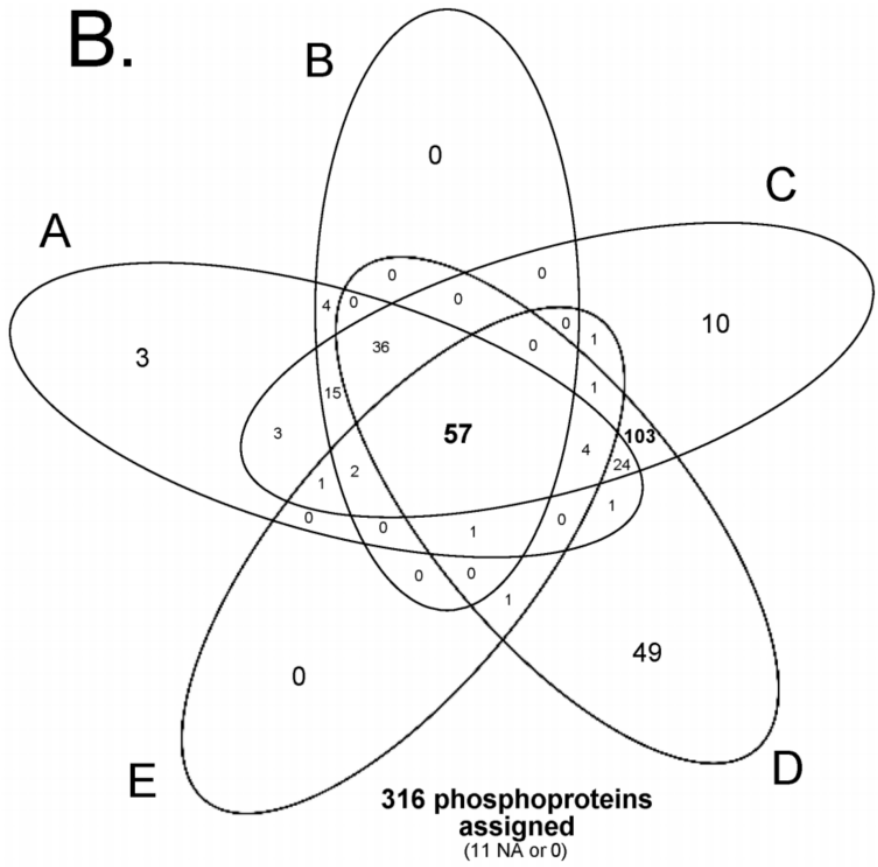


Figure 39: Venn-Diagram representation of phosphoproteins identified inthe CLL (A-D) and MBL tumor B-cell samples (E) analyzed. This Figure illustrates how many proteins were expressed in common or in only one or a subset of the 5 samples analyzed for the whole the phosphorylated protein dataset (327 proteins).

The qualitative analysis of the phosphoproteome in **Figure 39** shows as a result (**Figure 39B**), only 57 of 327 phosphoproteins (17%) identified were detected in common in the five CLL/MBL cell samples analyzed. Such phosphoproteins corresponded to proteins directly involved in protein phosphorylation (protein kinases such as PRKCB, LYN, SYK, ATM, BLK, JAK1, JAK2), signal transduction (e.g. KHDRBS1, STAM2, STAP1), and intracellular protein transport (NSF, CUL3). Thus, most phosphoproteins identified were present in only a subset of the samples or just in one sample. Briefly, 103/327 (32%) phosphoproteins were uniquely identified to be phosphorylated in samples C and D (e.g. BLK, CD19, PLCG2, and SCIMP); FEA showed they had specific roles in RNA binding, the spliceosome, at the same time they contained relevant interaction domains such as the SH3, SAP, KH, and LIM domains. In turn, sample A showed three uniquely phosphorylated proteins (HERC1, NUDT3, and PSMD9) and 10 and 49 proteins were restricted to sample C (DCP1B, GGA2, GBE1, IFIT5, JAK1, PPA2, SRRT, SRSF7, XRCC6, and ZNF24), and D (e.g. CBL, DOCK11, DOK2, EXOSC10, IKZF3, ILF3, LSP1, PDCL3, PTPN18, RIPK2, SASH3, SETD1A, and YBX1), respectively **(Diez, et al. submitted 2017)**.

4.4.3 Quantitative analysis

In this study besides the qualitative proteomic data the expression level (quantitative) data of the proteins is included. The main obstacles in the quantitative analysis are: **(i)** low number of patient datasets (2 replicas of 5 patients); and **(ii)** proteins can only compared in the different patients (horizontal) but not to each other in the same sample (vertical).

Of course this project cannot increase the number of samples in the analysis part but develop strategies to compare the proteins horizontal. In this thesis uses two analysis strategies: **(i)** percentage of the max value; and **(ii)** ranking method, 1 (highest) to 5 (lowest).

In the following paragraphs describe the results of the differential expression in the cancer patients of the protein in the proteome and phosphoproteome (4.4.3.1), combine the differential expression analysis with important DNA regions in CLL and MBL (4.4.3.2) and (4.4.3.3) the results of a FEA of important differential expressed proteins.

4.4.3.1 Differential expression of proteins in the samples of th CLL patients

All proteins identified in common across the CLL/MBL cell samples analyzed were ranked (score 1 to 5) based on their expression levels within the five studied samples (i.e. for each given protein, 1 was assigned to the sample presenting the highest expression value and 5 to that presenting the lowest one). **Figure 40A** (left panel) illustrates this ranking in which 1,880/2,978 (63%) proteins with the highest score of 1 belonged to sample A, whereas 1,704/2,434 (70%) proteins with the lowest score of 5 belonged to sample E. Interestingly, for each sample a different score predominated (i.e. 66% proteins of sample A had a score of 1;50% proteins of sample C scored 2; 45% proteins of sample B scored 3; 45% proteins of sample D scored 4; and 62% proteins of sample E scored 5). Interestingly, sample E showed a greater dispersion vs. samples A-D (**Figure 40A**, right panel).

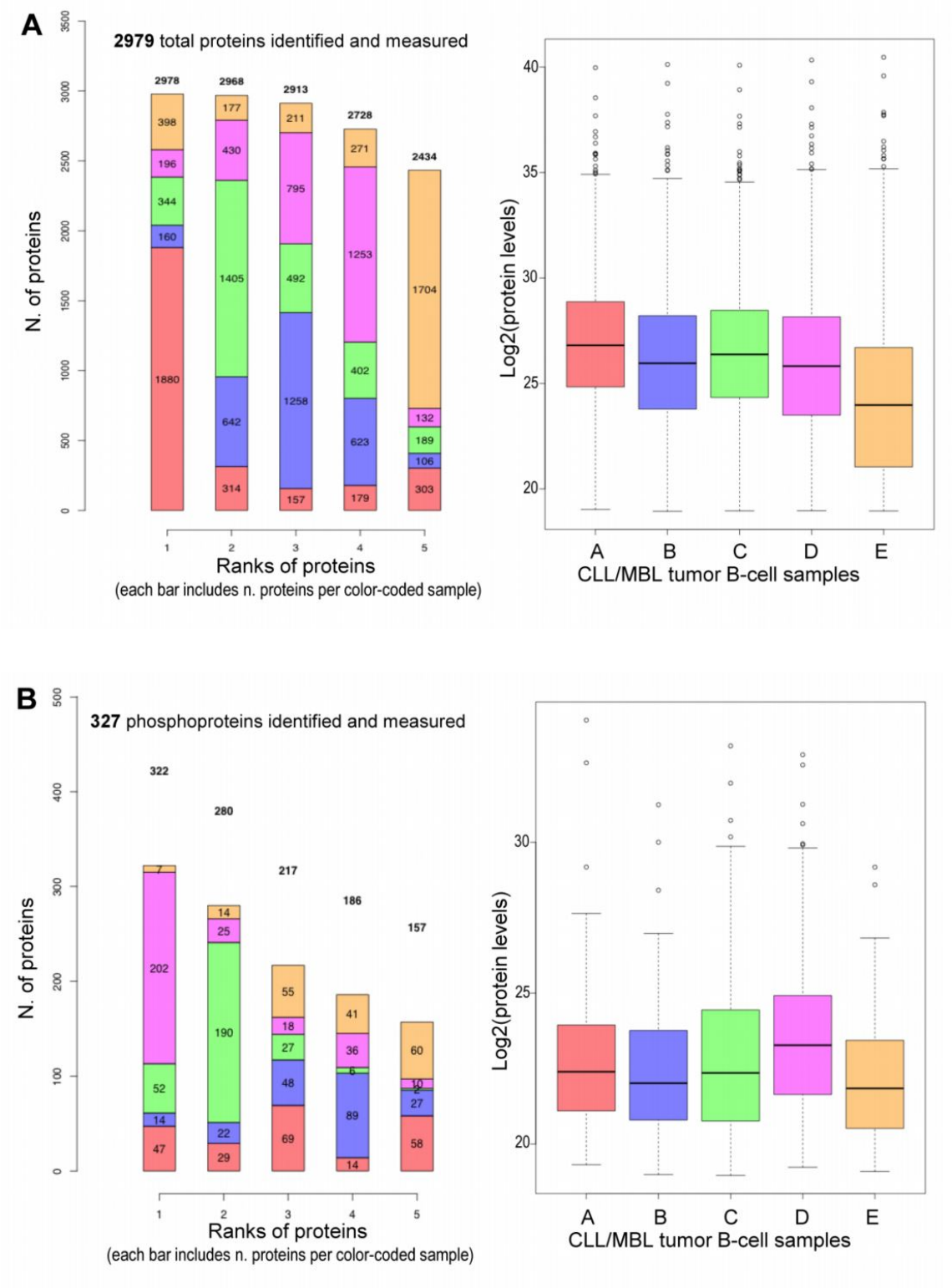


***Figure 40**: Quantitative proteomic expression data. On the left side in panels A) and B) the distribution of proteins found to be expressed at different levels per color-coded sample –sample A, red; sample B, blue; sample C, green; samples D, magenta; sample E, yellow - (ranking score in which 1 represents the highest levels and 5 the lowest levels per protein) is showing the number of ranks decreased because NA values did not get a rank value. In the right side, the distribution of the log2 expression values per sample (color-coded as in the left panels) is shown as box plots.*

Based on the rank position (1-5) per protein, detected for each sample (A-E), a hierarchical clustering algorithm was applied (**Figure 40** left panel) which revealed two main groups: group 1 included samples A and C and group 2 comprised samples B, D and E. The hierarchical clustering of the phosphoproteins showed different groups. Group one was patient A-B and the second group was C, D and E, in which the ranked based expression profile of patient D and E were closer to each other.

Supplementary Figure S4

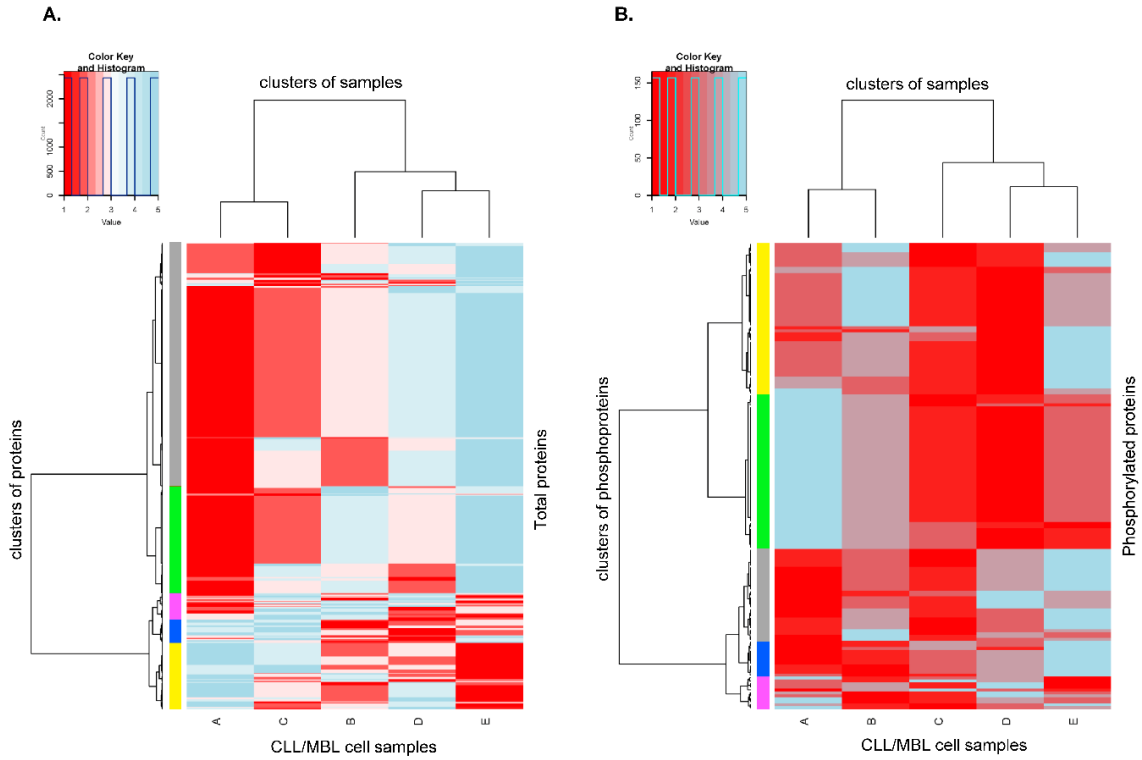


Figure 41: *shows the heatmaps based on the horizontal rank analysis of the expression levels of the proteins (A) and phosphoproteins (B). The maximum expression level per protein is assigned with a 1 and the lowest with a 5. The colors are gradually changed from 1 red to 3 white and to 0 blue.*

Besides the rank based measurements a percentage based heatmap and clustering analysis was done. The proteome proteins show two groups: group 1 included samples A and C and group 2 consisted of samples B, D and E. The proteins the clustering algorithm identified to differentiate the samples were analysed. In **Figure 41** a FEA of the grey labeled proteins (1,167) showed that such clustering was particularly based on proteins involved in proteasome activity (e.g. PSMD1, PSMD9, PSMA1), immune regulation (e.g. BCLF1, IgHM, CD2-P, CD5, CD47, CD48, CD53, CD79B, IGBP1, IGHC, IGHG1, B2M, ZAP70), HLA proteins (-A, -C, -DMB, -DRB1, -E), and the BCR signaling pathway (SYK, LSP1, MYCBP, BLNK, ATM). Another group of proteins (yellow-labeled) differentially expressed in the two clusters, consisted of 980 proteins, mostly related to HLA and antigen presentation (**Figure 42**, panel A). Other groups of proteins contributing to the clustering (clusters labeled as green, blue, and magenta) showed fewer differences across distinct samples analyzed. Among them, a group (green-labeled) of quite homogeneously expressed proteins within the two groups of samples corresponded to proteins directly related to protein synthesis and expression (**Figure 42**, panel A).

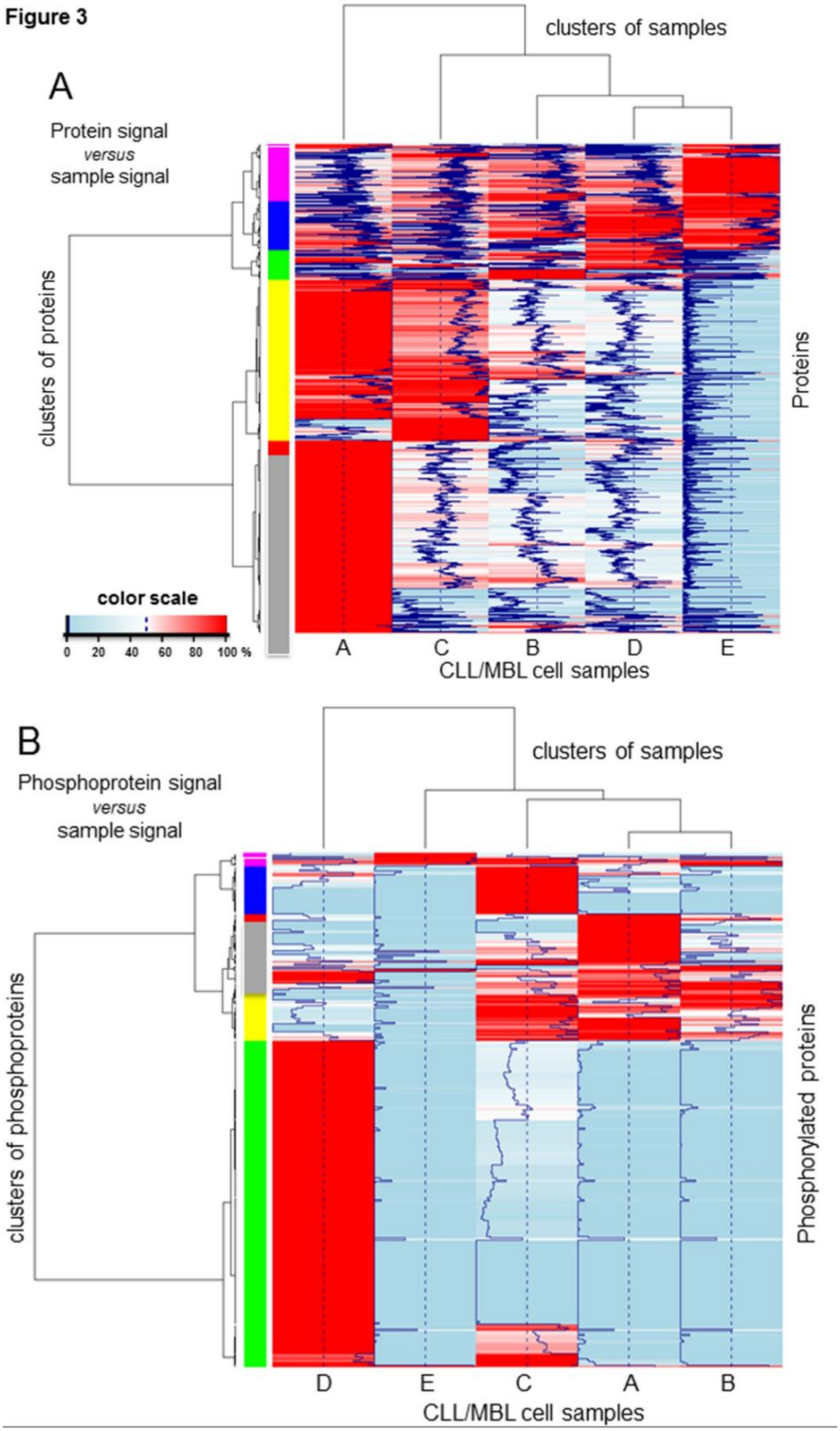
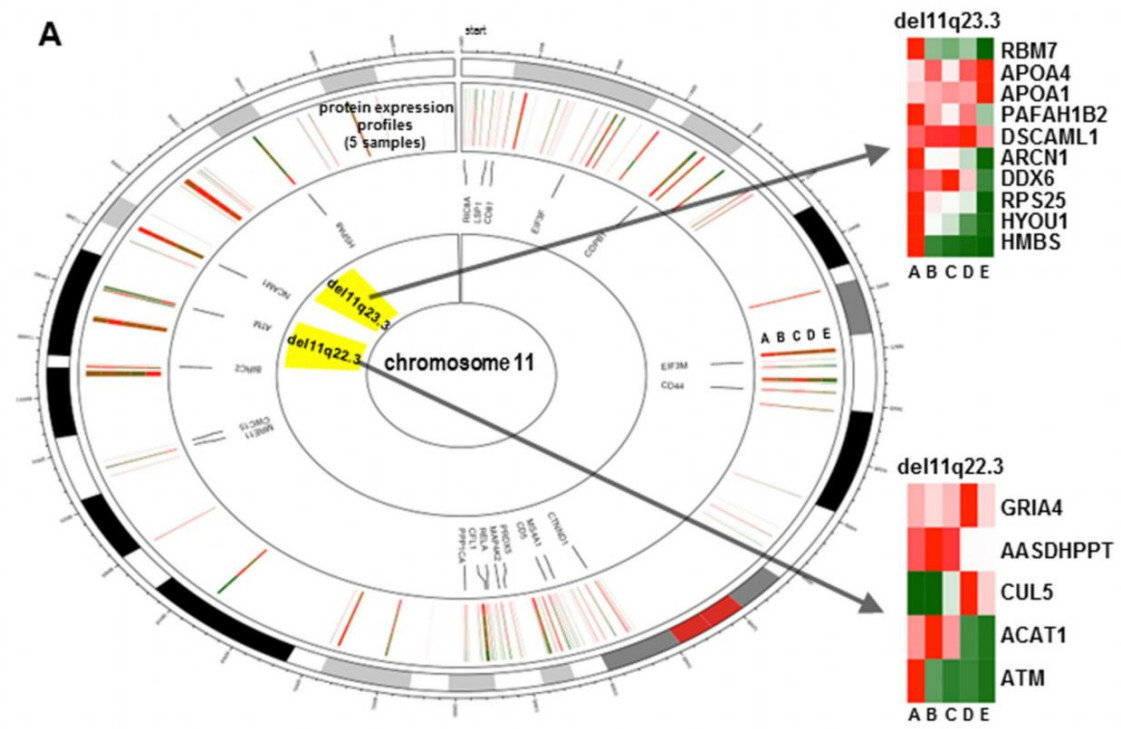


Figure 42: *Heatmaps based on the percentage comparison of the expression values of the proteins identified and quantitatively measured in CLL/MBL tumor B-cell samples. The maximum value is set to 100% and the other values are percentages of this value. Colors gradually change from 100% (red) to 50% (white) and 0% (blue). Distinct clusters of proteins are color-coded in the first column on the left of each heatmap. Panel A) shows the results of the whole proteome dataset and panel B) the phosphorylation dataset.*

As described above, significantly different phosphoproteome profiles were observed across the different samples analyzed once relative expression levels were considered (**Figure 42B**). Overall, two groups of samples (group 1: samples A to C showing UM-IGHV; group 2: samples D and E showing M-IGHV) were identified based on five main groups of proteins found to be differentially expressed between them. The most contributing group of proteins consisted of 201 proteins (green-labeled cluster); these included the phosphorylated LYN, BCLAF1, SYK, CD2AP, TP53BP1, BLK, LSP1, RAB10, CD19 (only phosphorylated in C and D), and BTK, among other proteins. The remaining clusters of phosphoproteins (labeled as yellow, grey, blue, and magenta in **Figure 42B**) contributed less to the clustering. No clear association was found between the tumor cell phosphoproteome and tumor cytogenetics, the IGHV mutational status and other features of the disease investigated.

4.4.3.2 Analysis and visualization of proteins in the deletions of chr 11 and 13

The percentage-based expression values of the proteome are also mapped to important chromosomal positions for CLL and MBL disease. This study actively searched for those proteins coded in the differentially altered (e.g. deleted) chromosomal regions, specific patterns were observed, particularly for del11q22.3, del11q23.3, and del13q14. Thus, lower expression levels of ATM and CUL5 were associated with del11q22.3 (sample C), decreased amounts of MLL, SCN4B, CD3D, ARCN1, and TREH were also found in sample C that carried del11q23.3, and low RB1 levels were associated with del13q14 (**Figure 43B**).



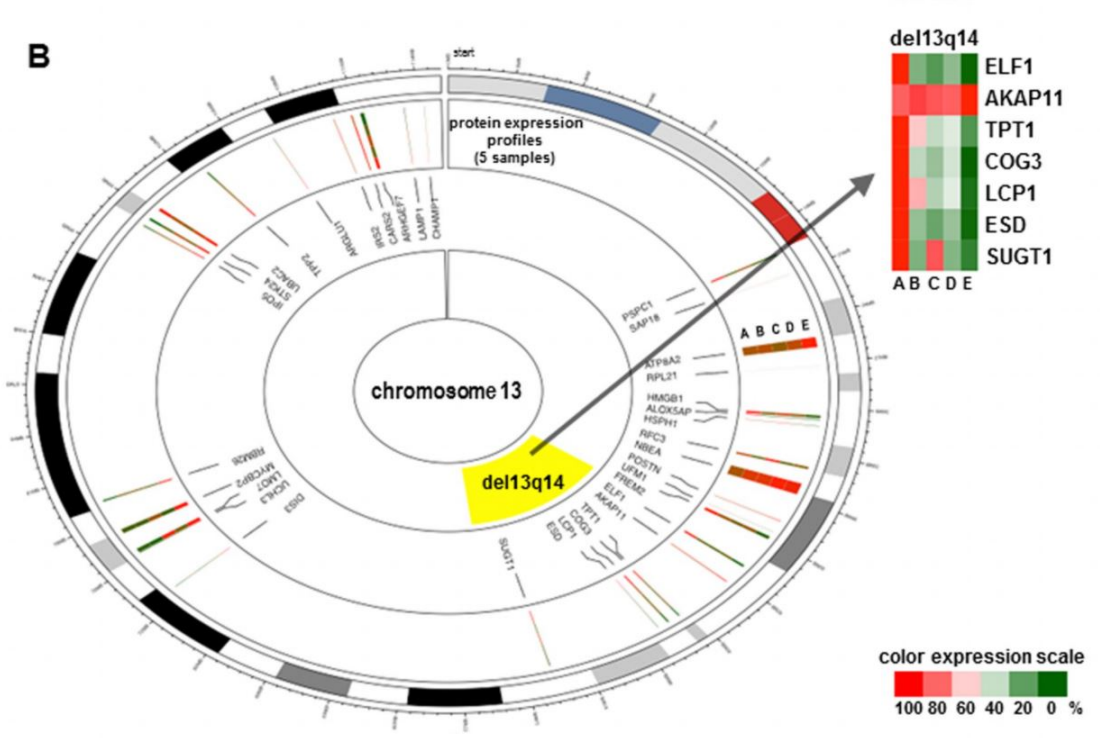


***Figure 43**: Heatmaps based on the percentage comparison of the expression values of the proteins identified and quantitatively measured in CLL/MBL tumor B-cell samples. The maximum value is set to 100% and the other values are percentages of this value. Colors gradually change from 100% (red) to 50% (white) and 0% (blue). Distinct clusters of proteins are color-coded in the first column on the left of each heatmap. Panel A) shows the results of the whole proteome dataset and panel B) the phosphorylation dataset.*

5 DISCUSSION

5.1 Path2enet

The *Results* chapters 4.1 and 4.2 above explained the **input**, **process** and **output** of the *Path2enet-Tool* functions. They also presented the graphic visualization, generated with the tool, including a comparative view of the **pathway graphs** produced by *KEGG*, and the **pathway-driven expression networks** produced [see section 4.2.1.]. On the basis of this detailed information, the main purpose of the *Discussion* chapter is to put these results into technical and biological perspective, and to contrast the tool against similar bioinformatic tools.

Section 5.1.1 discusses the technical part of the written software. Section 5.1.2 discusses the interaction and transcriptomic data used in Path2enet. Section 5.1.3 highlights the benefits of the Path2enet tool for biological researchers and 5.1.4 compares it to similar solutions in R. 5.1.5 discusses the results of the case study presented in the project.

5.1.1 Technical considerations

The written code in R is functional and can be used in the R-2.6.3 environment. The generated functions are new in R, although there are many packages working with data provided by KEGG. The functions are written user-friendly. The user only needs basic knowledge of the R language. In our test within a small user group the system operated stable. Tests with a larger number of users have to follow.

The following functions of the Path2enet tool were tested: **(i)** build the KeggSQL database out of KGML-files; **(ii)** establish connections between the R environment and the MySQL-databases; **(iii)** create protein-networks combining the data extracted from APID, KEGG and transcriptomic data of EST, microarray and RNA-Seq; **(iv)** visualize the protein-networks using igraph R package; and **(v)** analyse the networks topology and graph characteristics using also igraph.

The KeggSQL database fulfills its purposes: **(i)** it parses the KGML-files, collects all protein related data and stores it into a relational database; **(ii)** the created database has short access times, because the tables are all indexed; **(iii)** it can handle large data volumes; and **(iv)** the R functions can access the data either with the relational database hosted on a local machine or on a server.

The combination of the open source programs R and MySQL insures that most investigators can use the written functions free of charge.

5.1.2 Selection of data sources

In section 5.1.2.1 we discuss the interaction datasets the selection of the network data sources and in 5.1.2.2 the expression data sources we have selected for the Path2enet tool.

5.1.2.1 Interaction datasets

The Path2enet tool uses the *Agile Protein Interactomes DataServer* [APID] and *KEGG PATHWAY* as network datasets to generate its pathway-driven expression networks. The tool is also focused on human pathways and interactome, because it is the most relevant for our studies.

I have chosen APID because it focus on "the generation and delivery of unified compendiums of known and experimentaly proven protein-protein physical interactions" **(Alonso-López D. et al. 2016).** The datThe aset does not include predicted interactions like STRING, GeneMANIA or ConsensusPaTHDB. APID provides as well different quality levels of the interactions it hosts. These quality levels are based on number of experiments, methods and publications of each publication. The user can choose which confidence networks he wants to use for the Path2enet tool. I used the less confident but highest coverage network on default to include as many interactions as possible. Besides the quality of the dataset APID includes the interactomes of 448 organisms. To use Path2enet for a different organism, the user would have to adapt the ID mapping table of section 4.1.1.1 to map the KeggID of the organism he wants to investigate. But the Path2enet tool provides automatic annotation for mouse, rat, yeast and e.coli besides human. The dataset is also freely available.

The KEGG PATHWAY database is not complete free. Since May 2011 the KEGG FTP (53) resources are for paid subscribers only. But the KEGG database is freely available on the web and has a high popularity with over 500,000 visitors per month. KEGG PATHWAY includes 517 pathways and 515,449 references in total.The Path2enet tool uses the data via its webservice which will be free of charge in the future (54). But the online database of KEGG PATHWAY (http://www.kegg.jp/) provides KGML files for each biological pathway on its website. For example, in the case of the human NOTCH signaling pathway (KEGG ID reference: hsa04330) the KGML file can be downloaded freely as "hsa04330.xml". The link for this file is: http://www.kegg.jp/kegg-bin/download?entry=hsa04330&format=kgml. In this way, any specific pathway is accessible via its KGML file in the KEGG website and Path2enet R package provides functions to download these files and create a MySQL database derived from the KGMLs. Moreover, to facilitate the use of the pathway KMGL files within the application Path2enet, we also provided an SQL dump file ("Path2enet_KeggSQL.sql") generated with all the KMGL files of Homo sapiens (this datafile is provided at: http://bioinfow.dep.usal.es/path2enet/). This allows the creation of the necessary SQL database within the user's computer to query for specific pathways and to use the other functions of Path2enet. This database resource is not just a compendium of KMGL files from KEGG given that it provides some important added values: **(i)** it includes a mapping of all the gene and protein identifiers (IDs) from KEGG to the IDs of UniProtKB (used as the reference protein database in Path2enet); and **(ii)** it includes a relational SQL structure, based on the extracted data from the pathways, that allocates such information in two principal indexed tables: one describing the pair-wise links or relations between protein pairs, and another one describing the characteristics of each singular protein.

With respect to the use of other formats, other than XML and KGML, Path2enet can also use any database or resource provided in a "network structure" as an igraph object, because the tool includes functions to read and load in R igraph objects. The function *bar2graph* allows the integration of the expression datasets into igraph objects and the user can choose a indtifier for the dataset he wants to include like "rnaseq_". But the igraph object vertexces and the dataset have to be mapped to same identifier, in our case the UniprotKB identifier. For the use of other standard formats, such as SBML or BioPAX, there are already tools that address this scope. For example KEGGtranslator], an easy-to-use stand-alone application that can visualize and convert KGML formatted XML-files into multiple output formats. This tool supports a plethora of output formats, being able to increase the information in translated documents beyond the scope of the KGML document. KEGGtranslator converts KEGG files (KGML formatted XML-files) to SBML, BioPAX, SIF, SBGN, SBML-qual, GML, GraphML and LaTeX. Moreover, in Bioconductor (https://www.bioconductor.org/) there are packages to parse, modify and visualize BioPAX data, like rBiopaxParser or PaxtoolsR. At the moment, we are working on a workflow to use these packages to create SQL databases, similiar to the SQL described above, but using data from other pathway resources such as Reactome or Pathway Commons. This work is under development, but one of main problems in the use of these resources is not the use of standard formats, like BioPAX or SBML, but the accurate mapping to standard protein identifiers from UniProtKB.

APID and KEGG PATHWAYS are different in its structure and functional annotation. KEGG PATHWAY stores the interactions in categories (like cellular processes) and cellular functions (for example cell cycle), whereas APID stores its large interactome based on experimentally proven physical interactions as one interactome. Provides the user with the possibility to select one pathway of KEGG PATHWAY and look for interactions of the nodes (proteins) of the pathway in APID. Often the user will find new possible interactions in the PPI network, because the curation of KEGG PATHWAY is often slower then curatioin of experimental proven interactions of APID. Path2enet-generates by processing the KGML-files KEGG provides 50,448 unique interactions for human and the APID dataset of protein interactions for human contains 284,263 interactions. To identify a new interation in APID does not directly implicit that this interaction is involved in the pathway. Therefore, the researcher can use the transcriptomic data to evaluate if both proteins are expressed in the same tissue or development stage in the preprocessed datasets. If this is the case, he can check his own experimental expression datasets if both proteins are "ON". The last step is to design an experiment to prove the interaction in a specific cellular condition.

In section 5.1.3 the benefits for biological researchers by the integration of the pathway, protein-protein-physical interaction dataset and expression datasets is discussed in detail.

5.1.2.2 Expression datasets

Besides the pathway and protein-protein-interaction datasets, Path2enet integrates three kinds of transcriptomic datasets ESTs, Microarrays and RNA-Seq into the generated networks. The tool processes 4 datasources: **(i)** ESTs (expressed sequence tags) from the Unigene database that includes 18,880 gene/protein entries detected in 51 human tissues (http://www.ncbi.nlm.nih.gov/unigene); **(ii)** *Barcode* gene expression from high-density oligonucleotide microarrays that store 17,268 gene/protein entries detected in 195 tissues and cell lines; **(iii)** RNA-Seq data of the Human Body Map 2.0 that stores FPKM expression data of 18,744 gene/protein entries in 16 human tissues; and **(iv)** RNA-Seq data from the Human Protein Atlas which stores the FPKM expression data of 19,078 gene/protein entries of 33

human tissues (http://www.proteinatlas.org).

The Path2enet tool uses these preprocessed data to evaluate if a gene is active "*ON*" or "*OFF*" in a specific tissue or biological condition (for example normal or cancer). The interpretation of the Unigene dataset is simple, because the reads are already annotated to a tissue or biological condition. The user can use Path2enet tool to generate a EST dataset which he needs for his studies. The user can also select a higher threshold of reads to create a more confidential network.

A similar approach is possible with the RNA-Seq datasets. The expression unit *Fragments Per Kilobase Of Exon Per Million Fragments Mapped* [FPKM] shows very good which proteins are expressed. In section 4.3.2.2 demonstrated, that the proteomic dataset and the RNA-Seq dataset of *Ramos* B-cell displays a correlation of 91 % with high accuracy and great overlap within each analytical level. The threshold we used in this study was FPKM>=1. This was only one experiment, which is discussed later in the tesis, but at the moment I did not find a great scale study to confirm a solid FPKM threshold to decide the "ON" or "OFF" state of a gene. The user can also include his own experimental RNA-Seq datasets with the function bar2graph. The RNA-Seq data should have as unit FPKM and shall be annotated to the UniprotKB.

The processing of microarray datasets is more complicated because microarray studies are more designed to measure relative expression of one gene in various biological condition (differentially expression) than to define the status of a gene as expressed or unexpressed in a microarray experiment. To face this problem Zilliox et al. compared the expression values of expected expressed genes in a tissue or biological state (cancer) with those known not expected to be expressed in a large scale study **(Michael J. Zilliox, Rafael A. Irizarry, 2007)** (51). The reason why the method is called *gene expression barcode* is because the expressed genes are coded with 1 and the unexpressed genes with 0. Uncertain genes have a value between 1 and 0. The results of the last updated version of the *Gene Expression Barcode* dateset is available in the package annotated to UniportKB.

Besides the pre-processed dataset, the vectors of the expression thresholds for the plattforms GPL96 "[HG-U133A] Affymetrix Human Genome U133A Array" and GPL570 "[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array" are available via the *Barcode* algorithm of the R package frma. These vectors are annotated to Affy probe Sets of the microarray platforms. These Affy probe sets can be ambigious. In order to improve the *Barcode* algorithm Path2enet removes ambigious probesets by using the mapping table of the *Brainarray* tool. This tool only maps unambigious and good probesets to Ensembl Gene Identifier [ENSG ID]. Path2enet maps these ENSG ID to UniprotKB. The function **expr2barcode** enables this analysis. Unfortunately it is only available for these platforms due to the lack of trained vectors. In the global integration of proteomic and transcriptomic expression datasets in the *Ramos* B-Cell line we faced this problem (section 4.3.1.1.2). In this study we set the threshold to the highest 25% expressed genes, which would be a work around for other researchers with the same problem, but the experimental trained and proven *Barcode* algorithm is the choice of the author.

5.1.3 Benefits of the Path2enet tool for biological research

This paragraph summarizes which benefits the Path2enet tool provides to biological research. The benefits are interrogated in the categories input, process and output. Using these standard IT-categories allows analyzing the basics of the software developed. This summary will also be useful for the users in 2 directions: **(i)** to recapitulate the results and the benefits of

the Path2enet tool; and **(ii)** to compare it quickly with KEGG, APID and similar bioinformatic tools. It shows these benefits as well on the case study on B- and T-cells done.

Table 28 explains the functions of the Path2enet tool and shows how they generate benefits to the users. The reader can use this brief overview of the functions **(i)** as reference; and **(ii)** as a quick step-in into the functions. More information about each specific function is provided in the Results 4.3.1 and 4.3.2 chapter.

KeggXML2SqlDatabase: This function allows inter alia to process the *KGML*-files provided by *KEGG*, and to build a customized *MySQL* database called *Kegg2MySQL*. It stores the data in *R* data-frames and translates the *KeggIDs* to UniprotKB-ID. To enable a quick access to the data the tool indexes all tables of the database.

The tables of the *Kegg2MySQL*-database are divided in categories and *KeggXML2SqlDatabase* creates subpathways for these categories. The user does not need any knowledge of *SQL*-syntax, because he can access the *Kegg2MySQL*-database via *R* without using *SQL*-syntax. This insures that many scientists can create and use the *Kegg2MySQL*-database. If the user has a basic knowledge of *SQL* he can modify the *Kegg2MySQL*-database - via *R* or directly - and share it with other scientists.

One benefit of the *Kegg2MySQL*-database is that biological researchers get a quick access to data provided by *KEGG* via *R*. Biological researchers can investigate the relations of proteins by only using *R* instead of using the web service of *KEGG*. Using the web service is time-consuming if investigators want to investigate the relations of many proteins.

Apid2Sql: This function stores the database of APID on a local machine and gives them quick access via R as well. The user can use these dataset and search for physical-physical protein interactions available in this meta-database.

Both functions allow the analysis of the relations of many proteins is interesting in particular after experiments with microarrays, RNA-Seq, LC-MS/MS expression dataset and high throughput essays. As already mentioned in section 1.2.4 *R* and *Bioconductor* provide many packages for the analysis of these experiments. The output of experiments, like for example with microarrays, is a data frame in *R* which stores a list of genes/proteins which are over-expressed or suppressed in a specific tissue, development-stage or cell condition. Investigators want to know how these genes/proteins are related with each other; especially which impact has the manipulation of one or more genes on cellular pathways. Investigators can also use the data of microarrays / RNA-Seq of older studies. The *Gene Expression Omnibus* [GEO] (51) for example provides inter alia many datasets of microarray experiments, which investigators can access for free.

searchFunction: The investigators can use the output of these experiments and search for the relations of the genes in the *Kegg2MySQL*-database and APID-database. They only have to use the *searchFunction* of the *Path2enet tool.* It gives the biological researchers the possibility to select the relations of genes/proteins in a specific pathway, category or in the whole *Kegg2MySQL/APID*-databases. This enables a comparative view of the relations of the genes/proteins. Biological researchers can for example first select the category *NormalPathways* and then select the category *CancerPathways* of the KeggSQL-database Path2enet generates. So biological researchers can compare entries of *KEGG* which are annotated to the *"Cancer"* category, with entries which are not annotated to any disease. This is appropriate if a biological investigator researches the impact of the manipulation of an oncogene or tumor suppressor gene. The comparison of proteins and their relations could be

very interesting. This could inspire scientists to create new hypothesis, which they could prove experimentally. Section 4.1.2 explains the whole process in detail.

searchParalogous: If the biological researcher has found an interesting gene/protein, he can use the function *searchParalogous* to search for entries in *KEGG*, which share the same *EntryID* in one or more pathways. This allows identifying paralogous of a certain gene. Such paralogous could be better annotated than the gene of interest. Therefore, the biological researcher gets additional information of the gene, which he can verify by other experiments or data mining.

The *R* package *KEGGgraph* also parses the *KGML*-files of *KEGG*; however, *KeggXML2SqlDatabase* uses all *KGML*-files of a specie and *KEGGgraph* only supports one-by-one *KGML*-file. Moreover, *KEGGgraph* does not support the search for relations in *KEGG* and does not generate a database with categories and subpathways.

The combination of the *MySQL*-database *Kegg2MySQL*, the *Path2enet searchFunction* and *SearchParalogous* functions opens many opportunities for biological researchers to investigate datasets of biological experiments like microarray experiments. Using *Path2enet* allows predicting new relations of proteins, which investigators could proof experimentally. In *R* no package exists which provides functions to analyse the relations of many genes/proteins in *KEGG*. As no *SQL*-syntax is needed to search for the relations in *KEGG* insures that many scientists can use the *Kegg2MySQL*-database for their investigations.

The datasets of *KEGG* are well annotated, verified and used by many biological researchers. These were the reasons to use these datasets to generate the *Kegg2MySQL*-database. Then we decided to enrich this database with other interesting biological datasets out of *APID* and transcriptomic datasets like ESTs, RNA-Seq and microarrays. *APID* stores the information of major protein-protein-interaction databases (see section 4.1.2). *Unigene* stores the number of *ESTs* of genes found in different tissues and development-stages (see section 4.1.3), *Gene Expression Barcode* microarray datasets of different tissues, development-stages and cell lines and RNA-Seq datasets of *Human Body Map 2.0* and the *Protein Atlas* of different tissues.

Path2enet: These three datasets are the starting point for the *Path2enet* function *Path2enet* which is explained in detail in sections 4.2.1.1, 4.2.1.2 and 4.2.2.1. One of its major benefits is the ability to generate meta-analysis of *APID*, *KEGG* and transcriptomic datasets. The result of the meta-analysis is a new type of protein network, called “pathway-driven expression networks”. The function generates a protein network out of the relations found in *Kegg2MySQL*-database and *APID*. Then it enriches the generated protein networks with three transcriptomic data sources of each protein of the generated networks. In the creation process of the new type of protein networks *Path2enet* translates the *KEGGIDs* in *UniprotIDs*. In total the new type of protein networks: **(i)** combine the information of all three datasets; **(ii)** are tissue and development-stage specific networks; **(iii)** are stored in *igraphs-objects*, which can be analysed and visualized by many network specific *R*-packages; **(iv)** have exact nodes and relations; **(v)** allow classifying the proteins in verified and unverified with regard to the *Uniprot* database; **(vi)** compare the relations provided by *KEGG* with protein-protein-interactions provided by *APID*; **(vii)** allow classifying the relations in experimentally verified or unverified; and **(viii)** show the type of relation which two nodes of the network have. The advantages of the protein-networks provided in the *Path2enet tool* are explained in detail in section 4.2.1.3.

expr2barcode: Besides the a priori processed transcriptomic data the expr2barcode function allows to integrate experimental microarray datasets of the user and fulfils an analysis with Gene Expression Barcode algorithm to evaluate if a protein is expressed “On” or not “OFF” in

the experiment. This information is integrated with the networks the path2enet-functoin generates. This function gives researchers the opportunity to analyse their expression dataset in a network specific-context. They can compare their biological expression dataset with the already processed data as well to find interesting changes in the biological conditions in their samples.

graphTKplotterPATHW and ***graphTKplotterTissue:*** A visualized protein-network is a good starting point for the analysis of protein relations. As graphical pathways of *KEGG* provide a brief overview of the relations of genes in a pathway they are very supportive as an entry. However, graphical pathways of *KEGG* have several disadvantages: **(i)** the nodes of the pathways are placeholders; **(ii)** users cannot combine pathways; **(iii)** the pathways are tissue and development-stage unspecific; **(iv)** it is difficult to compare its pathways with other databases; and **(v)** the pathways cannot be analyse d statistically. These disadvantages do not allow a deep analysis of the relations of genes. The *Path2enet tool* functions *graphTKplotterPATHW* and *graphTKplotterTissue* mitigate these disadvantages and visualize the protein networks *Path2enet* in many different ways. **Figure** 8 and 9 show tissue and development-stage unspecific protein-networks generated with the *Path2enet tool*. This type of visualization helps biological researchers to get a good overview of protein relations or complete pathways. The **Figure**s demonstrate: **(i)** which type of relation two proteins have; **(ii)** which protein relations are provided by APID and *KEGG*; and **(iii)** which protein relations are provided only by *APID* or *KEGG*. Relations which are stored in both databases have a high level of verification. Instead of using placeholders every node in these graphs is treated individually. This ensures biological researchers to know exactly which proteins are related or not. They do not have to check, which genes/proteins really interact with each other, like for example in *KEGG*.

Figure 26 is a tissue-specific protein network generated with the *Path2enet tool* showing that it is possible to predict a protein relation (*Notch2* and *RBPJ*) in a specific tissue (liver) using the *graphTKplotterTissue*. *KEGG* alone cannot make such predictions because it: **(i)** replaces *Notch1/2/3/3* with *Notch*; **(ii)** replaces *RBPJ/RBPJL* with *RBPJ*; **(iii)** does not include transcriptomic information; and **(iv)** does not compare relations with other databases. This shows that biological researchers can use the *Path2enet tool* to analyse pathways and protein relations in order to get more information on the relation than only using the web-services of *KEGG* or *APID*.

Figure 28 shows a tissue-specific protein network of the *"*Notch Signaling Pathway*"*. All nodes are reduced if they do not have at least one transcript in the selected tissue or development-stage. This graph is useful if an investigator tries to analyse large pathways or protein-networks in a specific condition. The elimination of nodes reduces the complexity of a graph. This helps to analyse the protein-network, because a graph with fewer nodes can be better visualized than graphs with many nodes and is more relevant regarding the biological condition to analyse.

Figure 26 shows that the *Path2enet tool* also enables to search for protein-interactions of the members of the *"*Notch Signaling Pathway*"* in the whole *Kegg2MySQL*-database. These "global" pathways are helpful to identify new protein-relations. One example might be that after a microarray experiment a biological researcher has a list of genes, which are over-expressed and suppressed. Firstly, he wants to know, how these proteins directly interact with each other. In this case he can visualize the *"PATHLocal*"*-network the *Path2enet tool* generates. He will notice that many proteins of his list do not interact directly with each other. Secondly, he can visualize the *"PATHGlobal*"*-network which is generated with the *Path2enet*

tool and compare it with the first graph. In some cases he will notice that two proteins are over-expressed, which are related to each other by another protein. This protein might be a protein-kinase, which plays a major part in cell regulation. Even a slight change in the activation-process can cause the over-expression of other proteins. This could explain that the protein-kinase was not noticed in the microarray experiment. The change of its expression-level was too low to be noticed. But the over-expression of a protein related to this gene/protein is a hint that this protein might play an important role in the cell type and/or pathway which biological researchers investigate. The investigators can validate this hypothesis experimentally or check datasets of older microarray experiments in the *GEO*.

The **Figure**s show in which ways the *Path2enet tool* can visualize the new type of protein-networks. They also demonstrate how biological researchers can use them to analyse their datasets. The analysis of graphical protein-networks is a good starting point to interpret the results of biological experiments, which produce datasets with genes/proteins including numerous parameters like microarray experiments. Like mentioned before the *Path2enet tool* enables to use these datasets to generate protein-networks and visualize them. However, very often the number of gene/proteins is too big to visualize the generated protein-network in a proper way.

Figure 27 demonstrates such a big graph with 377 nodes and 1.793 relations. The *Path2enet tool* visualizes in this **Figure** the table *CancerPathways* of *Kegg2MySQL*. Such a big graph is not the exception but the rule. This is the reason why the *Path2enet tool* function *Path2enet* allows analyzing the generated networks statistically. The statistical analysis is the only way to get interesting information out of protein-networks which are too big to be visualized in a descent way. **Table 19** shows the results of the statistical analysis displayed in **Figure 27**. It only includes basic network parameters like *degree*, *betweenness*, *eigenvector* and *clustering-coefficient*. But even these parameters enable to do a rough analysis. It shows that AKT1/2/3, EGFR (both high *degree*), CCND1 (high *betweenness*) and PDPK1 (high *eigenvector* and *clustering* coefficient) are important proteins in the protein-network derived from *CancerPathways*. The proteins also show that the parameters change dependent on the function of the protein in the network. These examples are explained in detail in section 4.2.2. Biological researchers can use the *Path2enet tool's* statistical analysis to analyse large protein networks derived from their experiments. Another benefit of a statistical analysis in comparison to the graphical analysis is its higher objectiveness. Biological researchers have to learn to interpret the network parameters, but they could use the large amounts of experimental datasets of the *GEO* to train themselves. They can start with the analysis of small networks and visualize it in a descent way. Then they can analyse the graphical and statistical output of *Path2enet*. Then they can try other packages of *R*, which are specialized in analyzing networks like *igraph*. This is possible because the *Path2enet tool* only generates data frames and *R*-objects, which are standard objects in *R* and can be used by nearly all other packages. The possibility to do all investigation in an environment like *R* is a large benefit for biological researchers.

Finally, **Table 28** summarizes the benefits of the Path2enet tool divided in the 3 categories input, process and output.

***Table 28**: Function-wise summary of the benefits of the Path2enet tool. Details of the functions and a more detailed discussion of the graphs is found in the Results section.*

Function	Input	Process	Output
KeggXML2SqlDatabase [4.1.2.1]	*KGML*-files and BR-file of *KEGG*	Downloading *KGML*- and *BR*-files Parsing the data of the *KGML*-files to store it in *R* data-frames Creation of *Kegg2MySQL* Indexing the tables Counting the entries with the same *EntryID* Translation from *KeggID* to *UniprotKB*	*KeggSQL* on a local machine
Apid2Sql [4.1.2.2]	*Txt.-file* of *APID*	Creation of *PPI* database Indexing the tables	*APID* on a local machine
searchFunction [4.1.2.4]	User-friendly entry of: *GeneID*, *KeggID*, *Pathwaytitle*, *PathwayID*, etc [see **Table**s 7 and 8]	Searching entries and relations in *Kegg2MySQL* and *APID*	Results of the search are provided as a list of *R* data-frames
searchParalogous [4.1.2.4]		Searching proteins sharing the same *EntryID* in *KEGG PATHWAYS*	A list of entries which share the same *EntryID* with *KEGG*
Path2enet [4.2.1]	Same as SearchFunction to create the network Transcriptomic datasets of EST, microarray (Gene Expression Barcode Database), RNA-Seq (Human Body Map 2.0, ProteinAtlas)	Meta-analysis of *APID*, *KEGG* and *Transcriptomic* datasets Calculation of network parameters of the generated graphs Translating *KeggID* to *UniprotKB*	Generated *igraph objects* including the information coming from the 3 source databases: *R* data-frames of the network parameters calculated.txt files including all nodes and edges of the network, the type of relation and the network parameters calculated
graphTKplotterPATHW [4.2.1.3.1]	igraph objects generated with path2enet	Visualization of the graphs including the network parameters calculated: edge.attributes, *community*, *betweenness, etc.* *Visualization of the hubs in a network*	Nodes with *UniprotIDs* Nodes in the same *community* have the same color Type of protein relations Comparing relations provided by different sources, like protein-protein-interactions of *APID* or

			KEGG xamples: **Figure**s X and X	
graphParameters [4.2.1.1.3]	*igraph objects path2enet / expr2barcode* generates including tissue or development-stage specific transcriptomic data or experimental data	Creation of confidence networks based on the *vertex.attributes* of the selected transcriptomic datasets and thresholds Calculation of network parameters of the generated graphs	Reduced networks with nodes showing an expression value above the threshold in all selected attributes Data frame with the results of the network analysis **Figure** X	
expr2barcode [4.2.2.1]	*igraph objects path2enet / AffyBatch* objects of microarrays supported by frma/barcode plattforms GPL96 "[HG-U133A] Affymetrix Human Genome U133A Array", GPL570 "[HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array" referenceTable for the platform	The AffyProbe-Ids of the AffyBatch-object are mapped to UniprotKB-Ids. Data frames calculating the mean barcode value per protein for individual probes and phenotypes Loading the results of the barcode analysis in the igraph object	Generated of enriched *igraph objects* including the information coming from the microarray dataset *R* data-frames of the barcode analysis	
graphTKplotterTissue [4.2,4.2.1.3.2, 4.2.2.2]	*igraph objects path2enet / expr2barcode* generates including tissue or development-stage specific transcriptomic data or experimental data	Visualization of the graphs regarding its *edge* and *vertex.attibtutes* Creation of tissue and development specific networks networks	Proteins which have ESTs in the selected tissue are marked by color Networks with nodes which have at Least one EST in the selected tissue Relations are visualized in the same way as they are visualized in the *graphTKplotterPATHW* Examples: **Figure**s X to X	

5.1.4 Differential characteristics of Path2enet compared to similar R-packages

Paragraph 1.2.4 describes several R packages which use protein / gene associations derived from pathways and protein-protein physical interactions. Several packages also integrate expression data with their networks. In this paragraph I try to focus on the major differences of these packages compared to Path2enet.

The package *KEGGgraph* uses KGML-files of metabolic and non-metabolic pathways and converts them into directed and undirected igraph-objects. The *KEGGgraph* parses the KGML-files and uses the *BiomaRt* package to annotate the KEGG identifiers to *Human Gene Symbols*. The package allows to merge igraph-objects into meta-pathways. Some of the mentioned packages like *Pathview* and *MetaboSignal* rely on its output. The package *rBiopaxParser* and *PaxtoolsR* enables to build networks out of BioPAX OWL files. KEGG offers this file format on its licenced ftp server. Therefore, there are at least 3 packages available to transform KEGG pathways into networks. KEGGgraph even enables to use the enzyme-compound associations of metabolic pathways, which Path2enet does not process at the moment. But the Path2enet-Tool does not only use the information to build one network out of one or more KGML files. The function *KeggXML2SqlDatabase* is designed to create a MySQL-database out of all KGML-files which KEGG PATHWAY provides on its free of charge available website. It also uses the KEGG BR-file to structure the database in the categories of KEGG and builds meta-pathways based on this structure. In order to assure this essential function and to be more independent from other packages, the Path2enet-Tool includes its own KGML-parser and ID-mapping function. This is important because the backside of *Bioconductor* is, that sometimes the packages are not updated or do not adapt to new KGML-file formats.

The benefits of the MySQL datasets of KEGG and APID are discussed in the previous section under the points ***KeggXML2SqlDatabase***, ***Apid2Sql, searchFunction*** and ***searchParalogous***. None of these benefits are covered in the mentioned packages.

One deficit of the Path2enet-Tool is its limitation of pathway data sources. *PaxtoolR* and *rBiopaxParser* can process BioPAX-files, which is designed to be the standard format to exchange pathway data or access PATHWAY COMMON. The workaround is to map the igraph-objects of *PaxtoolR* and *rBiopaxParser* to UniprotKB identifier and to integrate the expression datasets into these igraph-object (**bar2graph**). But then all benefits of using a database environment like mentioned above are gone and the comparison to APID or other pathways has to be done individually.

Besides the generated database of KEGG and APID, Path2enet includes preprocessed expression data of ESTs, microarrays and RNA-seq. The user can explore this data in a network-based context and compare it to his own experimental data. The annotation of all data to the UniprotKB allow an easy integration with proteomic expression data which is still missing as preprocessed data source. At this point no similar tissue and cell specific data package was uploaded to the Bioconductor Data Packages (https://www.bioconductor.org/packages/release/data/experiment/) or R Data Package (https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html).

The **path2enet** integrates pathway associations, protein-protein physical interactions and the transcriptomic data. The function allows the user to freely select the proteins/genes and the pathway environment to generate an individual designed network. One example is the selection of a list of genes/proteins to find the relations they share in signaling or cancer annotated pathways. The packages *rBiopaxParser* and *KEGGgraph* allow to generate a network out off one or more KGML- or BioPAX-files but they cannot build networks out of a list of genes. The most similar approach allows the Pathway Commons Graph Query (function graphPc) of the *rBiopaxParser*. But this function does not allow to select a pathway environment. These pathway environments are also important in Path2enet to create the "*PATHGlobal**" networks and compare them directly with the user specific network. The function also compare the user specific network with a high-quality meta protein-protein

physical interaction database APID. The flexibility to generate and compare the networks in different biological context is not available in the other packages.

One of the objectives of the Path2enet-Tool is to define the "ON" or "OFF" status of a node in a generated network based on its expression value. The tools like *KEGGgraph*, *PaxtoolsR*, *rBiopaxParser*, *Pathview*, *BioNet* and *pwOmics* are designed to include more differential expression data. Therefore, Path2enet is the only tool to process expression data of microarrays with the *barcode* algorithm and *Brainarray*-tool (**expr2barcode**). It is also the only tool to provide functions to reduce the number of nodes based on one or more expression values and analyse these graphs (**graphParameters**) or visualize them (***graphTKplotterTissue***). The tools *SubpathwayMiner* and *BioNet* use differential expression data to find deregulated subgraphs in the networks in order to reduce the complexity for the interpretation of big networks. But these are different statistical approaches.

5.1.5 Case Study T- and B-lymphocytes

We used Path2enet to analyse the NOTCH signaling pathway in B cells and T-lymphocytes. We showed, that the expression networks based on a large microarray data set of these samples are different for each cell type, modulating the original general view of the canonical pathway provided by KEGG. The observed differences are not random. They have a clear biological meaning. We showed that only 2 out of the 4 NOTCH paralog proteins (NOTCH1, 2, 3, 4) were expressed in B cells and T cells. Thus, a clear signal in all lymphocytes was observed for NOTCH2; while NOTCH1 was also detected in B cells CD19+ and in T cells CD4+. We also found, that key regulators like DTX1 and HES1, are strongly expressed in B cells and less expressed, or not present, in T cells. All these results demonstrate the value of the cell-type and context specific networks generated with *Path2enet* .

5.2 Qualitative proteogenomic analysis of the Ramos cell line

In this section, we will discuss the results of the two qualitative proteogenomic studies **(Paula Díez, Conrad Droste et. al 2015)** and **(Paula Díez, Conrad Droste et. al in process 2017)**. The major differences of the transcriptomic and proteomic data sets of both studies are described in section 5.2.1. In section 5.2.2. we will discuss the differences of the only global based proteogenomic approach and the proteogenomic study of pre-selected proteins. In last section we try to highlight the benefits of the integration of transcriptomic and proteomic data analysis in both studies.

5.2.1 Comparison of transcriptomic and proteomic data sets

In the studies we had two kind of transcriptomic datasets (microarray and RNA-seq) and proteomic datasets (LC-MS/MS and SEC-MAP). Both studies have in common, that the global proteomic data was based on LC-MS/MS. In section 5.2.1.1 we discuss the different methods, to select the genes and proteins of the global approaches. Section 5.2.1.2 explains the results of data integration.

5.2.1.1 Identification of expressed proteins and genes

The selection of expressed proteins in the global LC-MS/MS datasets was similar. The data of the 4-subcellular fractionations (CYT, MB, ORG, NUC) was combined, in order to generate a proteomic dataset of each biological replica. These datasets were compared in 4 different ways, to define if a protein is present in the Ramos cellline or not: (i) Intersection (2 peptides in

all 3 replicas) 3,383 proteins; (ii) Union (2 peptides in any replicate) 5,494; (iii) Complete Case (1 peptide in each 3 replica) 5,707 proteins; and (iv) Maximum (1 peptide in any replica) 8,931 proteins. To define a reasonable threshold, based on peptides, which has sufficient confidence and coverage at the same time, is difficult and needs an expert review. We performed as well a Functional Enrichment Analysis with the 3,383 proteins detectec in the intersection dataset. This analysis revealed the expression of many essential proteins, for general cell functions and house-keeping processes (e.g. anabolism and synthesis processes, together with catabolism and cellular respiration), as well as the activity and regulation of major biological macromolecules (DNA, RNA and proteins) involved in key maintenance processes, like cell cycle, cell growth, cell proliferation, etc. **(Paula Díez, Conrad Droste et. al 2015)**. This shows, that the intersection dataset is of high confidence.

The result of the *Maximum* dataset was mapped to the chromosomes. We identified about 30% of all protein-coding genes (~5,000-6,000) present in human – X, Y chromosomes and mitochondrial DNA are excluded. It also showed, that this LC-MS/MS approach is proteome-wide and unbiased. In this dataset many proteins interacting with the MYC proto-ontogene, which plays a major role in Burkitt lymphoma Ramos B-cell line, were identified (Actl6a, Bcl2, Chd8, Gtf2I, Mapk1, Max, Mlh1, Mycbp2, Mycbp, Nmi, Nyfc, Pfdn5, Ruvb, Sap130, Smad2, Smad3, Smarca4, Smarcb1, Taf9, Wdr5, Yyi, among others).

Although the Maximum dataset shows biological meaningful results, we decided in our second study, to use a dataset of higher confidence the Complete Case. This dataset is similar to the intersection dataset in the first study, but needs only 1 peptide to confirm the expression of a protein per replica.

The next proteomic-based experimental approach is the antibody-based SEC-MAP technique. One drawback of this technique is, that the reasearcher selects the antibodies corresponding to the proteins he wants to investigate. In this case we selected 549 antibodies corresponding to 417 distinct proteins considered as important or relevant for the Ramos B-cells studied. To evaluate the quality of antibodies we designed the QAS value described in section 4.3.2.1.1. An antibody has to detect its corresponding protein, and has at least a QAS value of one or higher. 405 of 417 were detected in the dataset. The selection of antibodies and the evaluation of the antibodies experimental data needs an expert supervisor. The interpretation of the QAS allows to compare the antibodies in an objective way.

To find an objective solution to select the genes, based on the expression values in the transcriptomic data is difficult. For microarray dataset the Barcode algorithm is available to define the “On” or “OFF” state of a gene in an expression datasets. This algorithm is described in paragraph 3.4.2 and 5.1.2.2. But this algorithm supports only the plattforms GPL96 [HG-U133A], Affymetrix Human Genome U133A Array, and GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. We did not find a dataset of Ramos B-Cells of these platforms. The platform *Affymetrix* Human Gene ST 1.0 high-density oligonucleotide microarrays of the three Ramos replicas is not supported. Therefore, we had to select a threshold, based on our large experience with microarray data. We selected the threshold of 25% after normalization, but before, we mapped the Affy Probe Sets with the Brainarray-Tool to Ensembl Gene ID. We selected a high threshold value, to be sure that (**i**) no background noise is included in the datasets; (**ii**) we have a high probability, that the genes are expressed; and (**iii**) we still have a good confidence to coverage rate. In total 33,297 Affy Probe Sets are included in the data sets and 8,976 probe sets in the highest 25 %.

The other experimental expression dataset is RNA-seq based. RNA-seq values are based on counts of mRNA-fragments, and not on the density of an expression signal, like microarrays (see paragraphs 1.1.1 and 3.4.3). Therefore it is easier to select an expression threshold, which defines the "ON" or "OFF" state of a gene. We used the threshold FPKM >=1. This threshold cuts the number of identified genes and proteins, to 5,157 genes (ENSG IDs) with 5,672 mapped proteins (neXtProt IDs). In total the figures are: identified 20,533 genes (ENSG IDs) corresponding to 19,518 neXtProt IDs. This threshold is not experimental proven, like the Barcode algorithm, but the number of 5,672 mapped proteins is very similar to the proteomic Complete Case and Union datasets.

5.2.1.2 Integration of proteomic and transcriptomic data

In order to integrate the proteomic and transcriptomic datasets, we mapped the proteomic neXtProt IDs to Ensembl IDs. We used this direction, because one protein (in most cases) corresponds to one gene, but one gene can correspond to several proteins (i.e. alternative splicing).

The intersection, union and maximum datasets of the first study, were mapped to 3,433, 5,540, and 8,976 genes (i.e. Ensembl IDs). These genes were mapped to Affy Probe Sets 5,494, 6,175 and 9,494. Only 516 out of 8,931 proteins of the maximum dataset could not be mapped to the transcriptomic platform. This is a 94% overlap of the expression data in the microarrays over the proteomic data. The integration and comparison of RNA-seq and LC-MS/MS also showed an overlap of identified proteins. 5,672 out of 5,707 unique proteins of the MS/MS dataset could be also be found in the RNA-seq dataset (99,3%). When considering genes with an FPKM value ≥1 exact, 5,157 out of the 5,672 mapped proteins were identified (91%) (Section 4.3.2.2). These results indicate, that proteomic and transcriptomic studies, once applied to the global molecular characterization of a cell type, display a high accuracy and overlap within each analytical level.

In **Figure 31** and **34** we compared the expression levels of the proteins detected in the expression dataset. The density plots of the expression signal (log2 scale) of microarray and RNA-seq data revealed, that the expression level of genes, identified with the LC-MS/MS or SEC-MAP method, is significant higher than the overall expression value of the dataset. The comparison of the expression values of the genes, corresponding to the maximum, union and intersection dataset, showed, that the more confident the proteomic datasource is, the higher is the expression value. The comparison of the expression signal of the genes, corresponding to the SEC-MAP or LC-MS/MS in the RNA-seq dataset, did not show any significant differences.

Besides the high overlap of the experimental output, technologies had also a complementary part. In fact, comparing the results of the microarray and LC-MS/MS approaches, each one enables to cover the failures in identification ("exclusive identifications") due to its technical limits, with respect to the other strategy (paragraph 4.3.1.2 and 4.3.1.3). Performing the FEA for the exclusively identified proteins (516 proteins) in proteomics we identified a gain of proteins related to the mitochondrial and ribosomal organelles, as well as cytoplasmic ones. This is a quite expected result because the genomic microarrays employed do not include probes for mitochondrial DNA. A loss of identifications associated to immunoglobulins (Ig) and major histocompatibility complex (MHC) proteins was also detected in the transcriptomic data probably due to Ig gene rearrangements and hypersomatic mutations of Ig genes that hamper the design of adequate array probes for these genes.

Regarding exclusively identified proteins in transcriptomics (1,290 protein), the functional enrichment analysis revealed identifications related to nuclear and DNA-binding proteins. This is probably due to the fact that isolating the nuclear fraction in the last step of the proteomics approach decreases the recovery for nuclear proteins. On LC-MS/MS level it has to be proven if the recovery of lost proteins in the nuclear fraction could be improved by an other experimental design. Additionally, around 300 of these transcripts exclusively identified in transcriptomics correspond to non-coding RNAs. Thus, it is obvious that their corresponding proteins have not been detected by the proteomics platform.

The RNA-seq and LC-MS/MS dataset was not compared in such detail, because the SEC-MAP proteomic dataset was available. The differences are discussed in the following paragraph.

5.2.2 Global approach vs. pre-selected proteins

Also the overlapping between proteomics and transcriptomics was high, there is a significant variation when focusing on specific proteins. Thus, the integration of new –omics strategies (as SEC-MAP) implies the increasing of the knowledge about a specific cell or disease. In section 4.3.2.2, it has been depicted that 13.8% proteins (57/413) were only detected by the SEC-MAP approach against the 1% (4/413) only detected by RNA-Seq and 6.1% (25/413) detected by RNA-Seq and MS/MS proteomics. This analysis also reveals that 3.6 % (15/413) proteins (with very low transcripts levels) are accurately detected by SEC-MAP and MS/MS. Of course, these identified proteins require a further validation due to lack of previous evidence at the transcript level since the RNA-Seq technique is only a snapshot of the transcriptomic level at one specific point and the number of RNA-Seq datasets was only one.

The FEA analysis of the 57 proteins exclusive detected in the SEC-MAP dataset showed [4.3.2.3] that these proteins are related to functions to membrane proteins, including receptors, plasma membranes, cell adhesion molecules and transmembrane proteins. These results suggest that immune-affinity allows the specific and selective identification of proteins which could be difficult to detect by other approaches. 29 antibodies (eg. CD22, AUKA, CASP3, TRAF1….) did not efficiently work or were totally ineffective (QAS value<1) for detecting proteins by SEC-MAP; even testing more than one antibody clone against the same protein. Most of these 29 proteins were accurately detected by both MS/MS proteomics and RNA-Seq. FEA revealed, that functions including lumen and nuclear proteins, phosphorylated intracellular signaling proteins, and GTPase regulation, was undetectable by SEC-MAP. These results could be expected, because highly modified proteins (including post-translational modifications, PTMs).

MS/MS proteomic, microarray and RNA-Seq techniques allow the identification of thousands of gene products in a simultaneous manner. Nevertheless, SEC-MAP appears as a promising approach, because it allows the accurate detection of low abundance proteins, providing protein evidence for each gene with low transcript expression level. In addition, SEC-MAP could provide information about subcellular localization, MW and distribution (monomer vs multimeric/protein complexes), which facilitates the description of tissue or cellular protein profiles. Since SEC-MAP is based on affinity reagents; the reliability of this approach might be limited, depending onthe quality of antibodies employed for the construction of the array. Having this in mind, antibody content could be translated into a highly cost and time-consuming approach.

In summary, the SEC-MAP performance is easily integrated with other "-omics" approaches, and a useful approach for protein profiling. Of course, several aspects could be improved, such as optimizing QAS system, getting quantitative data for establishing differential protein profiles, sample preparation, or PTMs information, among others. Nevertheless, it seems a promising methodology and easy to combine with RNA-Seq, microarray and LC-MS/MS datasets, being a high-quality information source.

5.2.3 Benefits of a protegenomic analysis

Omic- technologies have the characteristic of generating vast amounts of data, which are of high usefulness for increasing the cellular knowledge. But the are a great challenge, when dealing with them. However, the integration of bioinformatics with these –omics strategies seems a necessary and promising approach to increase the knowledge in the field.

Integrating complementary methodologies, allow a better characterization and profiling of cells and diseases. In this sense, we have combined three different approaches (MS/MS proteomics, RNA-Seq transcriptomics, and SEC-MAP - as an example of affinity proteomics), to provide a full characterization of protein profile for a pathological Burkitt lymphoma B-cell.

The addition of transcriptomic data in proteomic studies has proven beneficial in increasing the number of detected proteins, either by providing evidence at the transcript level itself, or by selection of a suitable sample for protein detection, based on transcript level expression. In the two studies we identified in total 439 missing proteins, and 69 proteins with higher confidence level (removing the proteins detected by the maximum dataset).

5.3 Proteomic analysis of B-cell lymphocytosis of patient samples

The following paragraphs discuss the results of an overall quantitative proteome and phosphoproteome of B-CLL and CLL-like monoclonal B-cell lymphocytosis (MBL) primary cells and the phosphoproteins potentially involved in the baseline pathological B-cell behavior using a high-resolution massspectrometry-based approach **(Paula Díez, Conrad Droste et. al, submitted 2017)**. In paragraph 5.3.1 I will discuss the processing of datasets and in 5.3.2 the results of quantitative and qualitative analysis

5.3.1 The proteomic data

The first step in this study was to create a realiable dataset of proteomic data, and to find methods to analyse the quantitative expression values. The quantitative proteome and phosphoroteome datasets are really small. We only have one patient per phenotype, and two biological replicas per proteomic dataset. Therefore we can only show, that our experimental and analytical steps to purify and compare the data, can create reliable results. In the first step we had to remove all ambiguous peptides, regarding the mapping to neXtProt-ID. If a peptid is not exact, you cannot know, which protein is detected and receives its expression signal. This is comparable to the Brainarray-Tool for the annotation of microarray probe sets data. In the other studies, I started the analysis after the mapping to UniprotKB was proceed and the ambiguous peptides were already removed. In this proteomic study it is necessary to start at the peptide level, because the second step is to remove unreliable nodes, based on the comparison of the replicas (0.5>Min/Max), and setting peptides under the threshold of 500,000 to zero (in detail 4.4.1). These steps remove peptides of high differences, and set low expression values to the same value. We did not remove the low expressed proteins, because

we wanted to see, that these peptides show at least a signal. These rules reduce the number of proteins in the proteome dataset to 2,970 (13,504 peptides). This is comparable to the intersection dataset created in 4.3.1. The number of phophoproteom is smaller (327 proteins; 594 peptides), because only kinasas are in the dataset. Of note, more than half of the phosphopeptides identified (329/594) were detected, 253 for the first time, Therefore, they were not previously reported in neXtProt database (release February 2017). This shows, that the phosphoproteome approach is much more sensible than the general LC-MS/MS approach.

5.3.2 Comparison of proteomic data of cancer patients

The qualitative analysis showed, that the proteome dataset is much more similar in all five patients (2,160/2,970 proteins; 73%) than the phosphoproteomic (57/327 proteins; 17%). The subset of patient group D and C in phophoproteomic is substantially higher (103/327 proteins; 31%) as compared to other proteins samples. Such phosphoproteins corresponded to proteins, directly involved in protein phosphorylation (protein kinases such as PRKCB, LYN, SYK, ATM, BLK, JAK1, JAK2), signal transduction (e.g. KHDRBS1, STAM2, STAP1), and intracellular protein transport (NSF, CUL3). Thus, most phosphoproteins identified were present in only a subset of the samples, or just in one sample. Briefly, 103/327 (32%) phosphoproteins were uniquely identified to be phosphorylated in samples C and D (e.g. BLK, CD19, PLCG2, and SCIMP). FEA showed that they hav specific roles in RNA binding, the spliceosome. At the same time they contained relevant interaction domains, such as the SH3, SAP, KH, and LIM domains. Interestingly, kinases found to be phosphorylated in common, across all samples included proteins that drive various signal transduction pathways involved in the pathogenesis of CLL such as BCR signaling, chemokine receptor signaling, and toll-like receptor (TLR) signaling pathways. In all samples analyzed, the chemokine receptor and TLR signaling pathways were under-represented (phosphorylation being restricted in common to the CXCR4, STAT3, and TLR1 proteins) vs the BCR signaling pathway. Thus, many BCR-related proteins were found to be phosphorylated in all or the majority of samples (e.g. LYN, SYK, PI3K, BTK, ZAP70, PLCG3, ERK, NFAT, CD19, PRKCB, NFKB, JNK and VAV), suggesting that BCR signaling plays a critical role in maintenance/survival of CLL/MBL tumor cells **(Paula Díez, Conrad Droste et. al, submitted 2017)**.

The differential expression analysis showed disadvantages. We did not have reference protein expression data, to normalize all protein levels in order to compare proteins in one patient to each other. Therefore it was only possible, to compare the same proteins in various samples to each other. We also could only compare the absolut values of the expression signal to each other. We solved this problem in a (i) ranking based approach and (ii) to set the highest expression value to 100% and to calculate the percentage of the other signals regarding this value. In the proteome dataset we showed, that 1,880/2,978 (63%) proteins with the highest score of 1 belonged to sample A, whereas 1,704/2,434 (70%) proteins with the lowest score of 5 belonged to sample E. Differential phosphoproteome profiles, based on a scoring classification, showed higher scores (values 1 and 2) for samples D and C, respectively, while the remaining rank values (3-5) were homogeneously distributed across samples A, B, and E. A further specification is not possible with this approach. We can show tendencies, but not statisticallyexact results. The heatmaps in **Figure 42** show as well a hierarchical clustering, based on the percentage expression value. This clustering groups A and C and B, D and E together in the proteome result. Group A and C show a strong expression for the same proteins in a large part of the dataset. In the FEA of these proteins (1,167), we showed that such clustering was particularly based on proteins involved in proteasome activity (e.g. PSMD1, PSMD9, PSMA1), immune regulation (e.g. BCLF1, IgHM, CD2-P, CD5, CD47, CD48, CD53,

CD79B, IGBP1, IGHC, IGHG1, B2M, ZAP70), HLA proteins (-A, -C, -DMB, -DRB1, -E), and the BCR signaling pathway (SYK, LSP1, MYCBP, BLNK, ATM). The proteins which are more present in the other samples (980 proteins) are mostly related to HLA and Antigen presentation. **(Paula Díez, Conrad Droste et. al, submitted 2017)**.

To better compare the expression dataset in the patients, we mapped the percentage based expression signals to important parts of the chromosome. **Figure 43** shows the differences of expression values in important chromosomal regions, like del11q22.3, del11q23.3, and del13q14. In this approach we can show, that known deletions in leukemia patients lead to non or really low expression in the patients. Examples are lower expression levels of ATM and CUL5 were associated with del11q22.3 (sample C), decreased amounts of MLL, SCN4B, CD3D, ARCN1, and TREH were also found in sample C that carried del11q23.3, and low RB1 levels were associated with del13q14.

The qualitative proteome and phosphoproteome analysis has to be improved in several points: (i) Larger proteomic and phosphoproteomic sample size; (ii) reference values to standardisize the expression signal; and (iii) experimental proven parameters to include or exclude peptides in the dataset.

6 CONCLUSIONS

The work presented in this **Doctoral Thesis** has been focused in the development and application of **bioinformatic** algorithms and methods to integrate, analyze and visualize various sources of **transcriptomic and proteomic human data**. After our studies and results we can formulate the following **conclusions**:

1. We have designed and developed a free and open bioinformatic application called **Path2enet** that allows to integrate human **pathways, proteins and expression** information placing the proteins in a relational context thanks to the transfer that the tool does to a **network graph environment**. The generation of **interactive biomolecular networks** allows to explore the links and associations of the proteins, as well as their centrality, and this is done in a cell- or tissue-specific way thanks to the integration of genome-wide expression data that allow to identify which specific proteins are ON or present in a given biological context.

2. We have designed and applied several bioinformatic strategies to allow an **integrative analysis of proteomic data** (obtained by mass-spectrometry or by other experimental proteomic techniques) and **transcriptomic data** (obtained by different genome-wide expression transcriptomic techniques). This integration and comparative analysis has been applied to specific cases-of-study on human data (mainly human B lymphocytes from lymphoma cells and from leukemia patients) revealing a very good agreement in the global biomolecular profiling of the genes and proteins detected by both "omic" approaches.

3. The **integrative analysis of proteomic** and **transcriptomic data** reveals that, despite a general good agreement, the **sensitivity and precision** of these modern technologies is not the same. In this way, we always observed a larger genome coverage in the transcriptomic data and also better reproducibility when comparing replicate samples. By contrast, proteomics and in particular phospho-proteomics, showed a much larger sensitivity to detect specific biomolecular forms –that are not detectable at expression level– and in this way it provides a better framework for the identification of specific biomarkers.

4. Our work demonstrates that the **application of proteogenomic studies** with a **robust bioinformatic support** provides an excellent research scenario to undertake better biomolecular studies integrating two essential layers of the "omic" large-scale data. This strategy is essential to study **complex diseases**, like cancer, and it should be supported by a good collaboration and coordination of multidisciplary research teams where **bioinformatics and computational biology expertise** is critical, as it has been shown in this work.

7 BIBLIOGRAPHIC REFERENCES

Aebersold R.; Mann M. 2003. **Mass Spectrometry-Based Proteomics**. *Nature*, 422(6928):198-207. doi: 10.1038/nature01511

Aibar S., Fontanillo C., Droste C. and De Las Rivas J. 2015. **Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering.** *Bioinformatics,* 31(10): 1686-8. doi: 10.1093/bioinformatics/btu864

Aibar S., Fontanillo C., Droste C., Roson B., Campos-Laborie F. J., Hernandez-Rivas J.M. and De Las Rivas, J. 2015. **Analyse Multiple Disease Subtypes And Buildassociated Gene Networks Using Genome-Wide Expression Profiles.** *BMC Genomics*, 16 Suppl. 5: S3. doi: 10.1186/1471-2164-16-S5-S3

Alexey I.,Nesvizhskii. 2014. **Proteogenomics: Concepts, Applications, and Computational Strategies.** *Nature Methods***,** 11(11): 1114–1125. doi: 10.1038/nmeth.3144

Alonso-López D., Gutiérrez M.A., Lopes K.P., Prieto C., Santamaría R., De Las Rivas J. 2016. **APID Interactomes: Providing Proteome-Based Interactomes with Controlled Quality for Multiple Species and Derived Networks.** Nucleic Acids Research, 44(W1):W529-35. doi: 10.1093/nar/gkw363

Alyass A., Turcotte M. and Meyre D. 2015. **From Big Data Analysis to Personalized Medicine for All: Challenges and Opportunities.** *BMC Medical Genomics*, 8:33. doi: 10.1186/s12920-015-0108-y

Barabasi A.L., Gulbahce N. and Loscalzo J. 2011. **Network Medicine: A Network-Based Approach To Human Disease**. *Nature Review Genetics,* 12(1): 56-68. doi: 10.1038/nrg2918

Beisser D., Brunkhorst S., Dandekar T., Klau G. W., Dittrich M.T. and Müller T. 2012. **Robustness And Accuracy Of Functional Modules In Integrated Network Analysis.** *Bioinformatics*, 28(14): 1887–94. doi: 10.1093/bioinformatics/bts265

Beisser D., Klau G.W., Dandekar T., Müller T. and Dittrich, M.T. 2010. **Bionet: An R-Package For The Functional Analysis Of Biological Networks.** *Bioinformatics*, *26*(8): 1129–30. doi: 10.1093/bioinformatics/btq089

Berman H.M., Battistuz T., Bhat T.N., Bluhm W.F., Bourne P.E., Burkhardt K., Feng Z. et al. 2000. **The Protein Data Bank.** *Acta Crystallographica Section D Biological Crystallography*, 58 (Pt 6 No 1): 899-907

Bittner L. and Bellman R. 1961. **Adaptive Control Processes. A Guided Tour. XVI + 255 S.** *ZAMM - Journal of Applied Mathematics and Mechanics,* 42(7-8): 364–365. doi: 10.1002/zamm.19620420718

Bolger, Lohse and Usadel. 2014. **Trimmomatic: a Flexible Trimmer for Illumina Sequence Data.** Bioinformatics, 30(15):2114-20. doi: 10.1093/bioinformatics/btu170

Calderone, A., Castagnoli L. and Cesareni G.2013. **Mentha: A Resource for Browsing Integrated Protein-Interaction Networks.** *Nature Methods,* 10(8): 690-1. doi: 10.1038/nmeth.2561

Carvalho B.S. 2015. **pd.hugene.1.0.st.v1: Platform Design Info for Affymetrix HuGene-1_0-st-v1.** *R package version 3.14.1*.

Carvalho, B.S. and Irizarry, R.A. 2010. **A Framework for Oligonucleotide Microarray Preprocessing Bioinformatics**. *Bioinformatics,* 26(19): 2363-2367. doi: 10.1093/bioinformatics/btq431

Charoentong P., Angelova M., Efremova M., Gallasch R., Hackl H., Galon J., and Trajanoski Z. 2012. **Bioinformatics for Cancer Immunology and Immunotherapy.** *Cancer Immunol Immunother,* 61(11): 1885-903. doi: 10.1007/s00262-012-1354-x

Consortium, T. 1000 G.P. 2012. **An Integrated Map Of Genetic Variation From 1.092 Human Genomes.** *Nature*, 491(7422): 56–65. doi: 10.1038/nature11632

D'Antonio M., Pendino V., Sinha S. and Ciccarelli F. D. 2012. **Network Of Cancer Genes (NCG 3.0): Integration And Analysis Of Genetic And Network Properties Of Cancer Genes.** *Nucleic Acids Research*, 40(Database issue): D978–83. doi: 10.1093/nar/gkr952

Dai M., Wang P., Boyd A.D., Kostov G., Athey B., Jones E.G., Bunney W.E. et al. 2005. **Evolving Gene/Transcript Definitions Significantly alter the Interpretation of Genechip Data.** *Nucleic Acids Research,* 33(20): e175. doi: 10.1093/nar/gni179

Das J. and Yu H**.** 2012.**HINT: High-Quality Protein Interactomes and their Applications in Understanding Human Disease**. BMC System Biology, 6:92. doi: 10.1186/1752-0509-6-92

Davis J.C., Furstenthal L., Desai A., Norris T., Sutaria S., Fleming E.and Ma P. 2009. **The Microeconomics of Personalized Medicine: Today's Challenge and Tomorrow's Promise.** *Nature Reviews Drug Discovery*, 8(4): 279–86. doi: 10.1038/nrd2825

Díez P., Droste C., Almeida J., González M., Orfao A., De Las Rivas J. and Fuentes M. 2017. **Revealing Cell Signaling Pathways in Chronic Lymphocytic Leukemia Tumor B-cells by Integration of Global Proteome and Phosphoproteome Profiles.** *Cancer Research,* Article submitted in June 2017, ref. CAN-17-1635

Díez P., Droste C., Bartolomé R., Lécrevisse Q., Dégano R.M., Alonso-López D., Ibarrola N., Góngora R., Corrales F.J., Orfao A., Lund-Johansen F., De Las Rivas J. and Fuentes M. 2017. **Comprehensive combination of affinity proteomics, MS/MS and RNA-Sequencing datasets for the analysis of a lymphoma B-cell line in the context of the Chromosome-Centric Human Proteome Project.** *Journal of Proteome Research,* Article written, in preparation to be submitted, 2017

Díez P., Droste C., Dégano R.M., González-Muñoz M., Ibarrola N., Pérez-Andrés M., Garin-Muga A., Segura V., Marko-Varga G., LaBaer J., Orfao A., Corrales F.J., De Las Rivas J., Fuentes M. 2015. **Integration of Proteomics and Transcriptomics Datasets for the Analysis of a Lymphoma B-Cell Line in the Context of the Chromosome-Centric Human Proteome Project.** *Journal of Proteome Research,*14(9): 3530-40. doi: 10.1021/acs.jproteome.5b00474

Dittrich M.T., Klau G.W., Rosenwald A., Dandekar T. and Müller T. 2008. **Identifying Functional Modules In Protein-Protein Interaction Networks: An Integrated Exact Approach.** *Bioinformatics*, 24(13): i223–31. doi: 10.1093/bioinformatics/btn161

Dobin A., Davis C. A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M. and Gingeras T.R. 2013. **STAR: Ultrafast Universal RNA-Seq Aligner.** *Bioinformatics,* 29: 15-21. doi: 10.1093/bioinformatics/bts635

Droste C. and De Las Rivas J. 2016. **Path2enet: Generation of Human Pathway-Derived Networks in an Expression Specific Context** *BMC Genomics,* 17(Suppl 8): 731. doi: 10.1186/s12864-016-3066-7

Druker B. J., Guilhot F., O'Brien S. G., Gathmann I., Kantarjian H., Gattermann N., Deininger M. W. N., et al. 2006. **Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia.** *New England Journal of Medicine*, 355(23): 2408–2417. doi: 10.1056/NEJMoa062867

Durinck S., Spellman P.T., Birnez E., Huber W. 2009. **Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor Package biomaRt.** *Nature Protocol*, 4(8): 1184-91. doi: 10.1038/nprot.2009.97

Eisenberg E., Levanon E.Y. 2013. **Human Housekeeping Genes, Revisited.** *Cell Press, Trends in Genetics,* 29(10): 569-574. doi: 10.1016/j.tig.2013.05.010

Eisenberg E., Levanon E.Y. 2014. **Corrigendum to: Human Housekeeping Genes, Revisited.** *Cell Press, Trends in Genetics,* 30*(3):* 119-120. doi: 10.1016/j.tig.2014.02.001

Feist P., Hummon, A.B. **Proteomic Challenges: Sample Preparation Techniques for Microgram-Quantity Protein Analysis from Biological Samples**. 2015. *International Journal of Molecular Science*, 16(2): 3537–3563. doi: 10.3390/ijms16023537.

Fields C. 1994. **Analysis of Gene Expression by Tissue and Developmental Stage.** *Current Opinion in Biotechnol*gy, 5(6):595-598. doi: 10.1016/0958-1669(94)90080-9

Finn R.D., Marshall M. and Batemann A. 2005. **iPfam: visualization of protein-protein-interactions in PDB at domain and amino acid resolutions.** *Bioinformatics,* 21*(*3): 410-412. doi: 10.1093/bioinformatics/bti011

Fontanillo C., Nogales-Cadenas R., Pascual-Montano A., and De las Rivas, J. 2011. **Functional Analysis Beyond Enrichment: Non-Redundant Reciprocal Linkage of Genes and Biological Terms.** *PloS One*, 6(9): e24289. doi: 10.1371/journal.pone.0024289

Gautier L., Cope L., Bolstad B.M. and Irizarry R.A. 2004. **Affy--Analysis of Affymetrix GeneChip Data at the Probe Level.** *Bioinformatics,* 20(3): 307–315. doi: 10.1093/bioinformatics/btg405

Gentleman R., Carey V.J., Bates D.M., Bolstad B., Dettling M., Dudoit S., Ellis B., Gautier L., Ge Y. et al. 2004. **Bioconductor: Open Software Development for Computational Biology and Bioinformatics.** *Genome Biology,* 5: R80. doi: 10.1186/gb-2004-5-10-r80

Goh K. I., Cusick M. E., Valle D., Childs B., Vidal M. and Barabási A.L. 2007. **The Human Disease Network.** *Proceedings of the National Academy of Science of the United States of America,* 104(21): 8685-8690. doi: 10.1073/pnas.0701361104

Haider S.; Pal R. 2013. **Integrated Analysis of Transcriptomic and Proteomic Data.** Current Genomics, 14(2): 91–110. doi: 10.2174/1389202911314020003

Harrison C. 2011. **Genetic Signatures Uncover New Uses.** *Nature Reviews. Drug Discovery*, 10(10): 732–3. doi: 10.1038/nrd3565

Hieronymus H., Lamb J., Ross K. N., Peng X. P., Clement C., Rodina A., Nieto M. et al. 2006. **Gene Expression Signature-Based Chemical Genomic Prediction Identifies a Novel Class Of HSP90 Pathway Modulators.** *Cancer Cell*, 10(4): 321–330. doi: 10.1016/j.ccr.2006.09.005

Horvatovich P., Végvári Á., Saul J., Park JG., Qiu J., Syring M., Pirrotte P., Petritis K., Tegeler T.J., Aziz M., Fuentes M., Diez P., Gonzalez-Gonzalez M., Ibarrola N., Droste C., De Las Rivas J., Gil C., Clemente F., Hernádez M.L., Corrales FJ., Nilsson CL., Berven F.S., Bischoff R., Fehniger T.E., LaBaer J., Marko-Varga G. 2015. **In Vitro Transcription/Translation System: A Versatile Tool in the Search for Missing Proteins.** *Journal of Proteome Research,* 14(9): 3441-51. doi: 10.1021/acs.jproteome.5b00486

Hughes T. R., Marton M. J., Jones A. R., Roberts C. J., Stoughton R., Armour C. D., Bennett H. A., et al. 2000. **Functional Discovery via a Compendium of Expression Profiles.** *Cell*, 102(1): 109–126. doi: 10.1016/S0092-8674(00)00015-5

International HapMap Consortium. 2003. **The International HapMap Project**. *Nature,* 426(6968): 789–796. doi: 10.1038/nature02168

Irizarry R.A., Hobbs B., Collin F., Beazer-Barclay Y.D., Antonellis K.J., Scherf U. and Speed T.P. (2003). **Exploration, normalization, and summaries of high density oligonucleotide array probe level data** Biostatistics, 4(2): 249-64. doi: 10.1093/biostatistics/4.2.249

Johannes M., Fröhlich H., Sültmann H. and Beissbarth T. 2011. **Pathclass: An R-Package for Integration of Pathway Knowledge into Support Vector Machines for Biomarker Discovery.** *Bioinformatics*, 27(10): 1442–1443. doi: 10.1093/bioinformatics/btr157

Kanderova V., Kuzilkova D., Stuchly J., Vaskova M., Brdicka T., Fiser K., Hrusak O. et al. 2016.T. **High-resolution Antibody Array Analysis of Childhood Acute Leukemia Cells**. Mol. Cell. Proteomics 15, 1246–1261.

Kanehisa, M. and Goto S. 2000. ***KEGG*: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Research,* 28(1): 27-30. doi:10.1093/nar/28.1.27

Kato K., Yamashita R., Matoba R., Monden M., Noguchi S., Takagi T., and Nakai K. 2005. **Cancer Gene Expression Database (CGED): A Database for Gene Expression Profiling with Accompanying Clinical Information of Human Cancer Tissues.** *Nucleic Acids Research,* 33(Database issue):D533–536. doi:10.1093/nar/gki117

Khatri P., Sirota M. and Butte A. J. 2012. **Ten years of pathway analysis: current approaches and outstanding challenges.** *PLoS Computational Biology*, 8(2): e1002375. doi: 10.1371/journal.pcbi.1002375

Kim M.S., Pinto S.M., Getnet D. Getnet D., Nirujogi R.S., Manda S.S., Chaerkady R., Madugundu A.K. et al. 2014. **A Draft Map of the Human Proteome.** *Nature*, 509(7502): 575-81. doi: 10.1038/nature13302.

Kostyuchenko V.A., Leiman P.G., Chipman .PR., Kanamaru S., van Raaij M.J., Arisaka F., Mesyanzhinov V.V., Rossmann M.G.2003**. Three-Dimensional Structure of Bacteriophage T4 Baseplate.** *Nature Structural Biology, 10*(9):688-93. doi: 10.1038/nsb970

Lamb J., Crawford E. D., Peck D., Modell J. W., Blat I. C., Wrobel M. J., Lerner J. et al. 2006. **The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes,and Disease.** *Science*, 313(5795): 1929–35. doi: 10.1126/science.1132939

Lee H.C., Lai K., Lorenc M. T., Imelfort M., Duran C. and Edwards D. 2012. **Bioinformatic tools and Databases for Analysis Of Next-Generation Sequence Data.** *Briefings in Functional Genomics*, 11(1): 12–24. doi: 10.1093/bfgp/elr037

Li C., Li X. et al. 2009. **SubpathwayMiner: A Software Package for Flexible Identification of Pathways.** *Nucleic Acids Research,* 37(19): e131. doi: 10.1093/nar/gkp667

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G. et al. 2009. **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics*, 25(16): 2078-2079. doi: 10.1093/bioinformatics/btp352.

Magrane M., UniProt Consortium. 2011. **UniProt Knowledgebase: A Hub of Integrated Protein Data.** *Database,* 2011: bar009. doi: 10.1093/database/bar009.

Martin J.A., Wang Z. **Next-generation transcriptome assembly.** Nature Reviews Genetics, 12(10): 671-82. doi: 10.1038/nrg3068.

Matthews L., Gopinath G., Gillespie M., Caudy M., Croft D., De Bono B., Garapati P. et al. 2009. **Reactome Knowledgebase of Human Biological Pathways and Processes.** *Nucleic Acids Research*, 37(Database issue): D619–22. doi: 10.1093/nar/gkn863

McCall M.N., Bolstad B.M. and Irizarry R.A. 2010. **Frozen Robust Multiarray Analysis (fRMA).** *Biostatistics*, 11(2): 242–253. doi: 10.1093/biostatistics/kxp059

McCall M.N., Uppal K., Jaffee H.A., Ziliox M.J., Irizzarry R.A. 2011. **The Gene Expression Barcode: Leveraging Public Data Repositories to Begin Cataloging the Human and Murine Transcriptomes.** *Nucleic Acids Research,* 39(Database issue): D1011-5. doi: 10.1093/nar/gkq1259.

Metzker M.L. 2010. **Sequencing Technologies - The Next Generation**. *Nature Reviews Genetics*, 11(1): 31–46. doi: 10.1038/nrg2626.

Moch H., Blank P. R., Dietel M., Elmberger G., Kerr K. M., Palacios J., Penault-Llorca F. et al. 2012. **Personalized Cancer Medicine and the Future of Pathology.** *Virchows Archiv : An International Journal Pf pathology*, 460(1): 3–8. doi: 10.1007/s00428-011-1179-6

Nueda M. J., Carbonell J., Medina I., Dopazo J. and Conesa A. 2010. **Serial Expression Analysis: A Web Tool for the Analysis of Serial Gene Expression Data**. *Nucleic Acids Research*, 38(Web Server issue): W239–45. doi: 10.1093/nar/gkq488

Pacini C., Iorio F., Gonçalves E., Iskar M., Klabunde T., Bork P., Saez-Rodríguez J. 2013. **DvD: An R/Cytoscape Pipeline for Drug Repurposing Using Public Repositories of Gene Expression Data.** *Bioinformatics,* 29(1): 132-4. doi: 10.1093/bioinformatics/bts656.

Prieto C. and De Las Rivas J. 2006. **APID: Agile Protein Interaction DataAnalyzer.** *Nucleic Acids Research,* 34(Web Server issue): W298-302. doi: 10.1093/nar/gkl128

Prieto C., Risueño A., Fontanillo C. and De Las Rivas J. 2008. **Human Gene Coexpression Landscape: Confident Network Derived from Tissue Transcriptomic Profiles.** *PLoS One,* 3(12): e3911. doi: 10.1371/journal.pone.0003911.

Risueño, A., Fontanillo, C., Dinger, ME., & De Las Rivas, J. 2010. **GATExplorer: Genomic And Transcriptomic Explorer; Mapping Expression Probes To Gene Loci, Transcripts, Exons And Ncrnas.** *BMC Bioinformatics,* 11:221. doi: 10.1186/1471-2105-11-221.

Roberts, L. 2001. **A History of the Human Genome Project**. *Science*, *291*(5507): 1195–1195. doi: 10.1126/science.291.5507.1195

Rodríguez, A. E., Hernández, J. Á., Benito, R., Gutiérrez, N. C., García, J. L., Hernández-Sánchez, M., Risueño, A. et al. 2012. **Molecular Characterization of Chronic Lymphocytic Leukemia Patients with a High Number of Losses in 13q14.** *PloS One*, 7(11): e48485. doi: 10.1371/journal.pone.0048485

Rual, J.F., Venkatesan K., , Hao T., Hirozane-Kishikawa T., Dricot A., Li N., Berriz G.F., Gibbons F.D. et al. 2005. **Towards a Proteome-Scale Map of The Human Protein-Protein Interaction Network.** *Nature,* 437(7062): 1173-1178. doi: 10.1038/nature04209

Rustici G., Kolesnikov N., Brandizi M., Burdett T., Dylag M., Emam I., Farne A. et al. 2013. **ArrayExpress Update--Trends in Database Growth and Links to Data Analysis Tools.** *Nucleic Acids Research*, 41(Database issue):D987-90. doi: 10.1093/nar/gks1174.

Sales, G., Calura, E., Cavalieri, D. and Romualdi, C. 2012. **Graphite - A Bioconductor Package to Convert Pathway Topology to Gene Network.** *BMC Bioinformatics*, 13(1): 20. doi: 10.1186/1471-2105-13-20

Sant M., Allemani C., Tereanu C., De Angelis R., Capocaccia R., Visser O., Marcos-Gragera R. et al. 2010. **Incidence Of Hematologic Malignancies in Europe by Morphologic Subtype: Results of the HAEMACARE Project.** *Blood*, 116(19): 3724–3734. doi: 10.1182/blood-2010-05-282632

Schadt E.E., Friend, S.H. and Shaywitz D.A. 2009. **A Network View of Disease and Compound Screening.** *Nature Reviews Drug Discovery*, 8(4): 286–95. doi: 10.1038/nrd2826

Shendure, J. and Ji, H. 2008. **Next-Generation DNA Sequencing.** *Nature Biotechnology*, 26(10): 1135–1145. doi: 10.1038/nbt1486.

Shiromizu T, Adachi J, Watanabe S, Murakami T, Kuga T, Muraoka S, Tomonaga T. 2013. **Identification Of Missing Proteins in the Nextprot Database and Unregistered Phosphopeptides in the Phosphositeplus Database as Part of the Chromosome-Centric Human Proteome Project.** *Journal of Proteome Research,* 12(6): 2414-21. doi: 10.1021/pr300825v.

Sierra-Sánchez Á., Garrido-Martín D., Lourido L., González-González M., Díez P., Ruiz-Romero C., Sjöber R., Droste C., De Las Rivas J., Nilsson P., Blanco F. and Fuentes M. 2017. **Screening and Validation of Novel Biomarkers in Osteoarticular Pathologies by Comprehensive Combination of Protein Array Technologies.** *Journal of Proteome Research,* 16(5): 1890-1899. doi: 10.1021/acs.jproteome.6b00980

Sirota, M., Dudley, J. T., Kim, J., Chiang, A. P., Morgan, A. A., Sweet-Cordero, A., Sage, J., et al. 2011. **Discovery And Preclinical Validation of Drug Indications Using Compendia of Public Gene Expression Data.** *Science Translational Medicine*, 3(96): 96ra77. doi: 10.1126/scitranslmed.3001318

Slawski M., Boulesteix A.-L. and Bernau C. 2009. **CMA: A Comprehensive Bioconductor Package for Supervised Classification with High Dimensional Data.** *BMC Bioinformatics,* 9: 439. doi: 10.1186/1471-2105-9-439.

Suter, B.S., Kittanakom S. and Stagljar I. 2008. **Two-Hybrid Technologies in Proteomics Research**. *Current Opinion in Biotechnology,* 19(4): 316-323. doi: 10.1016/j.copbio.2008.06.005

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., et al. 2011. **The STRING Database in 2011: Functional Interaction Networks of Proteins, Blobally Integrated and Scored.** *Nucleic Acids Research*, 39 (Database issue): D561–8. doi: 10.1093/nar/gkq973

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., et al. 2003. **PANTHER:A Library of Protein Families and Subfamilies Indexed by Function.** *Genome Research*, 13(9): 2129–41. doi: 10.1101/gr.772403

Trapnell C., Williams B.A., Pertea G., Mortazavi A., Kwan G., van Baren M.J.,Salzberg S.L. et al. 2010. **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nature Biotechnology,* 28(5): 511–515. doi: 10.1038/nbt.1621.

Trapnell, C., Hendrickson D.G., Sauvageau M., Goff L., Rinn J.L.,and Pachter L. 2013. **Differential Analysis of Gene Regulation at Transcript Resolution with RNA-Seq.** *Nature Biotechnology,* 31(1): 46-53. doi: 10.1038/nbt.2450

Turner B., Razick S., Turinsky A.L., Vlasblom J., Crowdy E.K., Cho E., Morrison K. et al. 2010. **iRefWeb: Interactive Analysis of Consolidated Protein Interaction Data and their Supporting Evidence.** *Database,* 2010*:* baq023. doi: 10.1093/database/baq023*.*

Tusher, V. G., Tibshirani, R. and Chu, G. 2001. **Significance Analysis of Microarrays Applied to the Ionizing Radiation Response.** *Proceedings of the National Academy of Sciences of the United States of America*, 98(9): 5116–5121. doi: 10.1073/pnas.091062498

Uhlén M., Fagerberg L., Hallström B.M., Lindskog C., Oksvold P., Mardinoglu A., Sivertsson Å. et al. 2015. **Proteomics. Tissue-Based Map of the Human Proteome.** *Science,* 347(6220): 1260419. doi: 10.1126/science.1260419.

Ulitsky I., Krishnamurthy A., Karp R.M. and Shamir R. 2010. **DEGAS: De Novo Discovery of Dysregulated Pathways in Human Diseases.** *PloS One*, 5(10): e13367. doi: 10.1371/journal.pone.0013367

Vidal M., Cusick M. E.and BarabásiA.-L. 2011. **Interactome Networks and Human Disease.** *Cell*, 144(6): 986–98. doi: 10.1016/j.cell.2011.02.016

von Mering C., Huynen M., Jaeggi D., Schmidt S., Bork P. and Snel B. 2003. **STRING: A database of Predicted Functional Associations between Proteins.** *Nucleic Acids Research,* 31(1): 258-261. doi: 10.1093/nar/gkg034

Wang Z., Gerstein M., and Snyder, M. 2009. **RNA-Seq: A Revolutionary Tool for Transcriptomics**. *Nature Reviews Genetics*, 10(1): 57–63. doi: 10.1038/nrg2484

Warde-Farley et al. 2010. **The GeneMANIA Prediction Server: Biological Network Integration for Gene Prioritization and Predicting Gene Function.** *Nucleic Acids Research,* 38(Web Server issue):W214-20. doi: 10.1093/nar/gkq537.

Wei G., Twomey D., Lamb J., Schlis K., Agarwal J., Stam R.W., Opferman J.T. et al. 2006. **Gene Expression-Based Chemical Genomics Identifies Rapamycin as a Modulator of MCL1 and Glucocorticoid Resistance.** *Cancer Cell*, 10(4): 331–342. doi: 10.1016/j.ccr.2006.09.006

Wilhelm M., Schlegl J., Hahne H., Gholami A.M., Lieberenz M., Savitski M.M., Ziegler E. et al. 2014. **Mass-Spectrometry-Based Draft of the Human Proteome.** *Nature,* 509(7502): 582–587. doi: 10.1038/nature13319.

Yang W., Soares J., Greninger P., Edelman E.J., Lightfoot H., Forbes S., Bindal N. et al. 2013. **Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells.** *Nucleic Acids Research*, 41(Database issue): D955-61. doi: 10.1093/nar/gks1111

Yizhak, K.; Benyamini, T.; Liebermeister, W.; Ruppin, E.; Shlomi, T**.** 2010. **Integrating Quantitative Proteomics And Metabolomics with a Genome-Scale Metabolic Network Model.** *Bioinformatics*, 26(12) i255–i260. doi: 10.1093/bioinformatics/btq183.

Zhang J.D. and Wiemann S. 2009. ***Kegg*graph: A Graph Approach to *KEGG* PATHWAY in R and Bioconductor.** *Bioinformatics,* 25(11): 1470-1471. doi: 10.1093/bioinformatics/btp167.

Zhang X., Lu X., Shi Q., Xu X.Q., Leung H.C.E., Harris L.N., Iglehart J.D. et al. 2006. **Recursive SVM Feature Selection and Sample Classification for Mass-Spectrometry and Microarray Data.** *BMC Bioinformatics*, 7(1): 197. doi: 10.1186/1471-2105-7-197

Zhu J., Sanborn J.Z., Benz S., Szeto C., Hsu F., Kuhn R.M., Karolchik D. et al. 2009. **The Ucsc Cancer Genomics Browser.** *Nature Methods*, 6(4): 239–240. doi: 10.1038/nmeth0409-239

Zhu X., Gerstein M. and Snyder M. 2007. **Getting Connected: Analysis and Principles of Biological Networks.** *Genes & Development,* 21(9): 1010-1024. doi: 10.1101/gad.1528707

Zhu Y., Shen X. and Pan W. 2009. **Network-Based Support Vector Machine for Classification of Microarray Samples.** *BMC Bioinformatics,* 10 Suppl.1: S21. doi: 10.1186/1471-2105-10-S1-S21.

8 LIST OF WEB ADRESSES

1. **Kanehisa.** [On line] http://www.kanehisa.jp/en/db_growth.html.

2. **Ncbi.nlm.nih Glance.** [On line] http://www.ncbi.nlm.nih.gov/About/glance/ourmission.html.

3. **KEGG.db - Package Bioconductor.** [On line] http://www.bioconductor.org/packages/2.8/data/annotation/html/KEGG.db.html.

4. **Gene2Pathway.** [On line] http://cran.r-project.org/web/packages/gene2pathway/vignettes/gene2pathway.pdf.

5. **Bioconductor Packages KEGG Graph.** [On line] https://www.bioconductor.org/packages/release/bioc/html/KEGGgraph.html.

6. **Bioconductor.** [On line] https://www.bioconductor.org/packages/release/bioc/html/pathview.html.

7. **Bioconductor Packages.** [On line] https://www.bioconductor.org/packages/release/bioc/html/MetaboSignal.html.

8. **Clipper: Gene Set Analysis Exploiting Pathway Topology. .** [On line] R package version 1.16.0., 2016. https://www.bioconductor.org/packages/release/bioc/html/clipper.html .

9. **Tim Bray Extensible Markup Language.** [On line] http://www.w3.org/TR/2006/REC-xml-20060816/#sec-origin-goals.

10. **ncbi.nlm.nih.** [On line] http://www.ncbi.nlm.nih.gov/ .

11. **ncbi glance.** [On line] http://www.ncbi.nlm.nih.gov/About/glance/ourmission.html,.

12. **Unigene.** [On line] http://www.ncbi.nlm.nih.gov/books/NBK21083/ .

13. **NCBI Handbook.** [On line] http://www.ncbi.nlm.nih.gov/books/NBK21083/ .

14. **Ebi.ac.arrayexpress.** [On line] https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-513/samples/ .

15. **ftp.ensambl.bam.** [On line] ftp://ftp.ensembl.org/pub/release-70/bam/homo_sapiens/genebuild/.

16. **ftp.ensembl.gtf.** [On line] ftp://ftp.ensembl.org/pub/release-70/gtf/homo_sapiens/.

17. **ebi.ac.arrayexpres-2.** [On line] http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2836/.

18. **proteinatlas.** [On line] http://www.proteinatlas.org/download/rna_tissue.csv.zip.

19. **Packages XML.** [On line] http://cran.r-project.org/web/packages/XML/index.html .

20. **Packages RMySQL.** [On line] http://cran.r-project.org/web/packages/RMySQL/ .

21. **Packages DBI.** [On line] http://cran.r-project.org/web/packages/DBI/index.html.

22. **Packages gWidgetsRGtk2.** [On line] http://cran.r-project.org/web/packages/gWidgetsRGtk2/index.html .

23. **Packages gWidgets.** [On line] http://cran.r-project.org/web/packages/gWidgets/index.html .

24. **Packages RGTk2.** [On line] http://cran.r-project.org/web/packages/RGtk2/index.html.

25. **Packages igraph.** [On line] http://cran.r-project.org/web/packages/igraph/index.html.

26. **Packages BioIDMapper.** [On line] http://cran.r-project.org/web/packages/BioIDMapper/index.html .

27. **Packages RCurl.** [On line] http://cran.r-project.org/web/packages/RCurl/index.html.

28. **Packages org.Hs.eg.db.** [On line] http://www.bioconductor.org/packages/2.6/data/annotation/html/org.Hs.eg.db.html .

29. **GPL.** [On line] http://www.gnu.org/licenses/gpl.html.

30. **MySQL5.1.Reference Manual.** [On line] http://dev.MySQL.com/doc/refman/5.1/en/index.html .

31. **The mathematics of networks.** [On line] http://www-personal.umich.edu/~mejn/papers/palgrave.pdf,.

32. **Network analysis and visualization.** [On line] http://cran.r-project.org/web/packages/igraph/igraph.pdf,.

33. **Package affy.** [On line] http://bioconductor.org/packages/release/bioc/html/affy.html.

34. **Package frma.** [On line] http://bioconductor.org/packages/release/bioc/html/frma.html.

35. **Package barcode.** [On line] http://bioconductor.org/packages/release/bioc/html/frma.html.

36. **Package hgu133afrmavecs.** [On line] http://bioconductor.org/packages/release/data/annotation/html/hgu133afrmavecs.html.

37. **Package hgu133plus2frmavecs.** [On line] http://bioconductor.org/packages/release/data/annotation/html/hgu133plus2frmavecs.html.

38. **Package pd.hugene.1.0.st.v1.** [On line] http://bioconductor.org/packages/release/data/annotation/html/pd.hugene.1.0.st.v1.html.

39. **Package oligo.** [On line] http://bioconductor.org/packages/release/bioc/html/oligo.html.

40. **Uniprot Database.** [On line] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/.

41. **BRAINARRAY tool.** [On line] http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/20.0.0/ensg.asp.

42. **Uniprot_sprot.fasta.gz.** [On line] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.fasta.gz.

43. **APID.** [On line] http://cicblade.dep.usal.es:8080/APID/init.action.

44. **KEGG Pathway hsa.** [On line] http://www.genome.jp/KEGG/pathway/hsa/hsa05219.html .

45. **Barcode.** [On line] http://barcode.luhs.org/index.php?page=transcriptome.

46. **Uniprot/P31749.** [On line] http://www.uniprot.org/uniprot/P31749.

47. **Uniprot/P00533.** [On line] http://www.uniprot.org/uniprot/P00533.

48. **Uniprot/P24385.** [On line] http://www.uniprot.org/uniprot/P24385, .

49. **Myeloma.** [On line] http://www.ncbi.nlm.nih.gov/omim/254500.

50. **Uniprot/O15530.** [On line] http://www.uniprot.org/uniprot/O15530.

51. **Ncbi.nlm.nih.gov.** [On line] http://www.ncbi.nlm.nih.gov/geo/.

52. **Nextprot.** [On line] ftp://ftp.nextprot.org/pub/current_release/mapping.

53. **Bioinformatics.** [On line] http://www.bioinformatics.jp/en/keggftp.html.

54. **Kegg.** [On line] http://www.kegg.jp/kegg/docs/plea.html.

55. **HUPO Proteomics Standards Initiative.** [On line] http://www.psidev.info/.

56. **Especificación PSI-MI MITAB 2.7.** [On line] https://code.google.com/archive/p/psimi/wikis/PsimiTab27Format.wiki.

57. **Especificación PSI-MI XML.** [On line] http://www.psidev.info/mif.

58. **Reducing proteome redundancy.** [On line] http://www.uniprot.org/help/proteome_redundancy.

59. **IntAct Molecular Interaction Database.** [On line] http://www.ebi.ac.uk/intact/.

60. **MINT, the Molecular INTeraction database.** [On line] http://mint.bio.uniroma2.it/.

61. **Human Protein Reference Database.** [On line] http://www.hprd.org/.

62. **Biological General Repository for Interaction Datasets.** [On line] https://thebiogrid.org/.

63. **BioPlex (biophysical interactions of ORFeome-based complexes).** [On line] http://wren.hms.harvard.edu/bioplex/.

64. **Database of Interacting Proteins.** [On line] http://dip.doe-mbi.ucla.edu/dip/.

65. **UniHI.** [On line] http://www.unihi.org/.

66. **Hippie.** [On line] http://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/.

67. **PINA2.** [On line] http://cbg.garvan.unsw.edu.au/pina.

68. **HitPredict.** [On line] http://hintdb.hgc.jp/htp.

69. **Human Proteome Organization.** [On line] https://www.hupo.org/.

70. **The International Molecular Exchange Consortium (IMEx).** [On line] http://www.imexconsortium.org/.

71. **iRefWeb.** [On line] http://wodaklab.org/iRefWeb/.

72. **CORUM: The comprehensive resource of mammalian protein complex.** [On line] http://mips.helmholtz-muenchen.de/corum/.

73. **The MIPS Mammalian Protein-Protein Interaction Database.** [On line] http://mips.helmholtz-muenchen.de/proj/ppi/.

74. **Interologous interaction database.** [On line] http://ophid.utoronto.ca/ophidv2.204/.

75. **RCSB Protein data bank.** [On line] http://www.rcsb.org/pdb/home/home.do.

76. **Gene ontology consortium.** [On line] http://www.geneontology.org/.

77. **KEGG: Kyoto Encyclopedia of Genes and Genomes.** [On line] http://www.genome.jp/kegg/.

78. **Reactome: a curated pathway database.** [On line] http://www.reactome.org/.

79. **Pfam database.** [On line] http://pfam.xfam.org/.

80. **InterPro: protein sequence analysis & classification.** [On line] https://www.ebi.ac.uk/interpro/.

81. **Cytoscape: Network Data Integration, Analysis, and Visualization in a Box.** [On line] http://www.cytoscape.org/.

82. **Apache Log4j 2.** [On line] http://logging.apache.org/log4j/2.x/.

83. **Librerías JDBC.** [On line] http://docs.oracle.com/javase/tutorial/jdbc/.

84. **Framework JAMI.** [On line] https://github.com/MICommunity/psi-jami.

85. **ORM Hibernate.** [On line] http://hibernate.org/.

86. **ORM Doctrine.** [On line] http://www.doctrine-project.org/.

87. **Sistema gestor de bases de datos MySQL.** [On line] https://www.mysql.com/.

88. **Sistema gestor de bases de datos de grafos Neo4j.** [On line] https://neo4j.com/.

89. **HTML Kickstart .** [On line] http://www.99lime.com/elements/.

90. **Librería jQuery.** [On line] https://jquery.com/.

91. **Librería qTip2.** [On line] http://qtip2.com/.

92. **Paquete de iconos vectoriales Font Awesome.** [On line] http://fontawesome.io/.

93. **Librería Charts de Google Developers.** [On line] https://developers.google.com/chart/.

94. **Librería DataTables.** [On line] https://datatables.net/.

95. **Librería cytoscape.js.** [On line] http://js.cytoscape.org/.

96. **Librería arbor.js.** [On line] http://arborjs.org/.

97. **Extensión arbor para cytoscape.js.** [On line] https://github.com/cytoscape/cytoscape.js-arbor.

98. **Complex Portal.** [On line] https://www.ebi.ac.uk/intact/complex/.

99. **GeneTerm Linker.** [On line] http://gtlinker.cnb.csic.es/.

100. **GeneCodis.** [On line] http://genecodis.cnb.csic.es/.

101. **Enrichr.** [On line] http://amp.pharm.mssm.edu/Enrichr/.

102. **HuRI: The Human Reference Protein Interactome Mapping Project.** [On line] http://interactome.baderlab.org/.

103. **Interactome3D.** [On line] http://interactome3d.irbbarcelona.org/.

104. **Licencia Apache 2.0.** [On line] http://www.apache.org/licenses/LICENSE-2.0.

105. **Punto de acceso a la descarga de archivos en HPRD.** [On line] http://www.hprd.org/download.

106. **Punto de acceso a la descarga de archivos en BioGRID.** [On line] http://thebiogrid.org/download.php.

107. **Punto de acceso a la descarga de archivos en BioPlex.** [On line] http://wren.hms.harvard.edu/bioplex/downloadInteractions.php.

108. **Esquema XSD del formato XML usado en la base de datos UniProt.** [On line] http://www.uniprot.org/docs/uniprot.xsd.

109. **Librería Simple API for XML (SAX).** [On line] http://www.saxproject.org/.

110. **UniProt Proteomes.** [On line] http://www.uniprot.org/proteomes/.

111. **Servicios REST de UniProt.** [On line] http://www.uniprot.org/help/programmatic_access.

112. **Ontology Lookup Service (OLS).** [On line] http://www.ebi.ac.uk/ols/index.

113. **UniProt (Universal Protein Resource).** [On line] http://www.uniprot.org/.

114. **MatrixDB: The Extracellular Matrix Interaction Database.** [On line] http://matrixdb.univ-lyon1.fr/.

115. **PDBsum: Pictorial database of 3D structures in the Protein Data Bank.** [On line] http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html.

116. **AmiGO 2.** [On line] http://amigo.geneontology.org/amigo.

117. **IntAct FTP.** [On line] ftp://ftp.ebi.ac.uk/pub/databases/intact/current.

118. **NCBI Taxonomy FTP.** [On line] ftp://ftp.ncbi.nih.gov/pub/taxonomy/.

119. **OBO Foundry.** [On line] http://www.obofoundry.org/.

120. **Estándares HTML y CSS (W3C).** [On line] https://www.w3.org/standards/webdesign/htmlcss.

121. **Licencia MIT.** [On line] https://opensource.org/licenses/MIT.

122. **UniProt REST .** [On line] http://www.uniprot.org/help/programmatic_access.

123. **UniProt ID Mapping.** [On line] http://www.uniprot.org/mapping/.

124. **Mentha.** [On line] http://mentha.uniroma2.it/.

125. **STRING.** [On line] http://string-db.org/.

126. **HINT.** [On line] http://hint.yulab.org/.

127. **GeneMania.** [On line] http://genemania.org/.

128. **ncbi.nlm.nih.** [On line] http://www.ncbi.nlm.nih.gov.

9 LIST OF FIGURES

10 LIST OF TABLES

11 APPENDIX: SCIENTIFIC PUBLICATIONS

Scientific Publications of CONRAD F. DROSTE directly related with the work presented in this Thesis (they are enclosed after this Appendix).

Díez P., Droste C., Dégano R.M., González-Muñoz M., Ibarrola N., Pérez-Andrés M., Garin-Muga A., Segura V., Marko-Varga G., LaBaer J., Orfao A., Corrales F.J., De Las Rivas J. and Fuentes M. 2015. **Integration of Proteomics and Transcriptomics Data Sets for the Analysis of a Lymphoma B-Cell Line in the Context of the Chromosome-Centric Human Proteome Project.** *Journal of Proteome Research,* 14(9): 3530-40. doi: 10.1021/acs.jproteome.5b00474

Díez P., Droste C., Bartolomé R., Lécrevisse Q., Dégano R.M., Alonso-López D., Ibarrola N., Góngora R., Corrales F.J., Orfao A., Lund-Johansen F., De Las Rivas J. and Fuentes M. 2017. **Comprehensive combination of affinity proteomics, MS/MS and RNA-Sequencing datasets for the analysis of a lymphoma B-cell line in the context of the Chromosome-Centric Human Proteome Project.** *Journal of Proteome Research,* Article written, in preparation to be submitted, 2017.

Díez P., Droste C., Almeida J., González M., Orfao A., De Las Rivas J. and Fuentes M. 2017. **Revealing Cell Signaling Pathways in Chronic Lymphocytic Leukemia Tumor B-cells by Integration of Global Proteome and Phosphoproteome Profiles.** *Cancer Research,* Article submitted (ref. CAN-17-1635) in June 2017.

Droste C. and De Las Rivas J. 2016. **Path2enet: Generation of Human Pathway-Derived Networks in an Expression Specific Context** *BMC Genomics,* 17(Suppl 8): 731. doi: 10.1186/s12864-016-3066-7

Other Scientific Publications of CONRAD F. DROSTE that correspond to work done during his predoctoral period, but they are not directly related with the work presented in this Thesis.

Aibar S., Fontanillo C., Droste C. and De Las Rivas J. 2015. **Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering.** *Bioinformatics,* 31(10): 1686-8. doi: 10.1093/bioinformatics/btu864

Aibar S., Fontanillo C., Droste C., Roson B., Campos-Laborie F.J., Hernandey-Rivas J.M. and De Las Rivas J. 2015. **Analyse multiple disease subtypes and build associated gene networks using genome-wide expression profiles.** *BMC Genomics*, 16(Suppl 5): S3. doi: 10.1186/1471-2164-16-S5-S3

Horvatovich P., Végvári Á., Saul J., Park JG., Qiu J., Syring M., Pirrotte P., Petritis K., Tegeler T.J., Aziz M., Fuentes M., Diez P., Gonzalez-Gonzalez M., Ibarrola N., Droste C., De Las Rivas J., Gil C., Clemente F., Hernádez M.L., Corrales FJ., Nilsson CL., Berven F.S., Bischoff R., Fehniger T.E., LaBaer J., Marko-Varga G. 2015. **In Vitro Transcription/Translation System: A Versatile Tool in the Search for Missing Proteins.** *Journal of Proteome Research,* 14(9): 3441-51. doi: 10.1021/acs.jproteome.5b00486

The Author(s) *BMC Genomics* 2016, **17**(Suppl 8):731
DOI 10.1186/s12864-016-3066-7

BMC Genomics

RESEARCH Open Access

Path2enet: generation of human pathway-derived networks in an expression specific context

Conrad Droste and Javier De Las Rivas*

From 6th SolBio International Conference 2016 (SolBio-IC&W-2016)
Riviera Maya, Mexico. 22-26 April 2016

Abstract

Background: Biological pathways are subsets of the complex biomolecular wiring that occur in living cells. They are usually rationalized and depicted in cartoon maps or charts to show them in a friendly visible way. Despite these efforts to present biological pathways, the current progress of bioinformatics indicates that translation of pathways in networks can be a very useful approach to achieve a computer-based view of the complex processes and interactions that occurr in a living system.

Results: We have developed a bioinformatic tool called *Path2enet* that provides a translation of biological pathways in protein networks integrating several layers of information about the biomolecular nodes in a multiplex view. *Path2enet* is an R package that reads the relations and links between proteins stored in a comprehensive database of biological pathways, KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/), and integrates them with expression data from various resources and with data on protein-protein physical interactions. *Path2enet* tool uses the expression data to determine if a given protein in a network (i.e., a node) is active (ON) or inactive (OFF) in a specific cellular context or sample type. In this way, *Path2enet* reduces the complexity of the networks and reveals the proteins that are active (expressed) under specific conditions. As a proof of concept, this work presents a practical "case of use" generating the pathway-expression-networks corresponding to the NOTCH Signaling Pathway in human B- and T-lymphocytes. This case is produced by the analysis and integration in *Path2enet* of an experimental dataset of genome-wide expression microarrays produced with these cell types (i.e., B cells and T cells).

Conclusions: *Path2enet* is an open source and open access tool that allows the construction of pathway-expression-networks, reading and integrating the information from biological pathways, protein interactions and gene expression cell specific data. The development of this type of tools aims to provide a more integrative and global view of the links and associations that exist between the proteins working in specific cellular systems.

Keywords: Biological pathway, Protein network, Gene network, Network analysis, Transcriptomics, Expression, Gene coexpression, Bioinformatics, R package

* Correspondence: jrivas@usal.es
Bioinformatics and Functional Genomics Group, Cancer Research Center (CiC-IBMCC, CSIC/USAL/IBSAL), Consejo Superior de Investigaciones Cientificas (CSIC), Salamanca, Spain

Background

Large-scale "omic" experiments that capture the physical associations and links between genes, proteins and other molecular components within the cells are producing extensive data on biomolecular interactions which are stored in new generation databases and resources [1]. The human interactome, for example, is composed of around 20,000 protein-coding genes, around 1000 metabolites and a still undefined number of distinct proteins and functional RNA molecules [2]. In total, this sums up to more than 100,000 cellular components expected to form the complex machinery of human cells. These components are related to each other in different ways. The number of relations and functional associations substantially exceeds the number of components, making the interactome a large relational system difficult to depict and analyze. Despite this complexity, the nature of the cellular interactomes allows to render or transcribe them into biomolecular "networks" that can integrate different layers of information to generate comprehensive spaces, providing a better view of the cellular systems. Moreover, the "networks" can be analyzed with computers to explore and quantify the centrality and the weight of the different components, and to find clusters or modules of highly related elements. This is the framework that drived us to develop the bioinformatic application tool here presented, called *Path2enet*.

Path2enet is an R package that reads the relations and links between proteins stored in the major and highly curated pathways database KEGG (Kyoto Encyclopedia of Genes and Genomes) [3, 4] and integrates them with gene expression data from various resources as well as experimentally determined data from protein-protein physical interactions taken from APID [5, 6]. *Path2enet* tool uses the expression data to determine if a given node (protein) in a generated network is active (ON) or inactive (OFF) in a specific cellular context, cell type or condition. The transformation of pathways into comprehensive networks plus the mapping of active –i.e., expressed– nodes, can help researchers to integrate different levels of molecular information, placing it in a relational specific context. In addition, the integration of protein-protein interaction data within a pathway-network view can help to find relevant relations and critical nodes in the processes studied. As a practical example, we applied *Path2enet* tool to the analysis of the NOTCH Signaling Pathway in human lymphocytes in order to uncover the specific differences between B cells (CD19+) and T cells (CD4+ or CD8+).

Methods

Integration of pathways, molecular interactions and expression resources

Path2enet is an R package that uses and integrates several databases and resources to generate pathway-derived networks in an expression specific context. These resources are the following: (**A**) pathways data, the tool collects the pathways information from KEGG, taking the KGML-files and generating a MySQL database from such files (this data integration provides a set that contains 50,448 unique interactions for human) [3, 4]; (**B**) protein-protein interaction data, the tool also uses a dataset of human protein-protein physical interactions (PPIs) from the dataserver APID [6], which at the time of building the package contained 284,263 unique interactions of human proteins; and (**C**) gene expression data, the tool integrates four types of expression information. These are: (**C1**) ESTs (expressed sequence tags) from the Unigene database that includes 18,880 gene/protein entries detected in 51 human tissues (http://www.ncbi.nlm.nih.gov/unigene); (**C2**) *Barcode* gene expression from high-density oligonucleotide microarrays that store 17,268 gene/protein entries detected in 195 tissues and cell lines [7, 8]; (**C3**) RNA-Seq data of the Human Body Map 2.0 (ArrayExpress Experiment E-MTAB-513) that stores FPKM expression data of 18,744 gene/protein entries in 16 human tissues (these FPKMs –fragments per kilobase of exon per million reads– were calculated using *Cufflings 2.2.0* algorithm [9] and annotated to *Ensembl GRCh37* with the R-package *Biomart* [10]; and (**C4**) RNA-Seq data from the Human Protein Atlas which stores the FPKM expression data of 19,078 gene/protein entries of 33 human tissues (http://www.proteinatlas.org) [11].

Calculation of expression level to identify ON/OFF genes

Beside the pre-processed expression datasets provided in several of the integrated resources, *Path2enet* uses the gene expression *Barcode* algorithm with the R package *fRMA* [8] to evaluate if a gene is expressed (i.e., is ON, active and present) or not (i.e., such gene is OFF, not-active and therefore not expressed) in a studied set of samples. The user can also incorporate and apply in *Path2enet* his own expression ON/OFF thresholds, for example using experimental RNA-Seq data. However, the identification of such thresholds is not trivial and the *Barcode* algorithm is most efficient in this task.

ID mapping and data unification

For the ID mapping and integration, *Path2enet* uses *Brainarray* [12] or *Gatexplorer* [13] within R to annotate the probeset identifiers of the microarrays to *Ensembl* gene identifiers.

To achieve a correct unification of databases and resources, *Path2enet* uses as key identifiers (IDs) of the genes/proteins the entry IDs from *UniProtKB* database [14]. Therefore, the KEGG gene and *Ensembl* gene identifiers in the datasets are annotated to the *UniProt* entry IDs using the mapping tables that *UniProt* provides. *Path2enet* also uses the R package *RMySQL* [15] to build and to connect to the *MySQL* databases using

R programming. Finally, in order to build the networks, *Path2enet* uses the R package *igraph* [16], which is a tool that provides outputs that can be introduced in Cytoscape.

Selection of an experimental dataset to apply *Path2enet*

As a practical example, we applied *Path2enet* to analyze the NOTCH Signaling Pathway in human lymphocytes, detecting the way in which this pathway is expressed in these cells and also finding the specific differences in activated genes/proteins between "naive" B cells (B cells that have not been exposed to an antigen) and T cells. To perform this analysis we downloaded and normalized an expression dataset that included 163 human samples. These samples were genome-wide expression microarrays of platform Human Genome U133 Plus 2.0 from *Affymetrix* (GEO reference: GPL570). The samples corresponded to naive B cells (CD19+), 32 microarrays; T cells (CD4+), 96 microarrays; and T cells (CD8+), 35 microarrays. The specific. CEL files (i.e., the raw data) that correspond to these samples are indicated in Additional file 1, and are available in the Gene Expression Omnibus (GEO) database from NCBI.

Software availability and implementation

Path2enet has been developed in R (free software environment for statistical computing and graphics, https://www.r-project.org/). In this way, a full operative R package has been built and it is available at http://bioinfow.dep.usal.es/path2enet. The software will be uploaded to the R CRAN package repository (CRAN.R-project.org) once this article is published. An R vignette (enclosed as Additional file 2) is provided as a guided tutorial to facilitate the installation and use of the *Path2enet* package.

Results and discussion

Building networks and performing analysis with *Path2enet*

Path2enet is a bioinformatic application tool that integrates the information of pathways, protein-protein interactions and expression datasets (obtained with microarrays, RNA-Seq or ESTs) from different tissues and cell types. *Path2enet* uses these datasets to build a network view of biological pathways in an expression-specific context. The tool is capable of identifying the genes/proteins that are ON in specific samples applying the *Barcode* algorithm, and allows the use of specific experimental expression data to present focused views of the human pathways map as specific biomolecular networks.

In the networks built using *Path2enet*, the "nodes" correspond to the proteins included in the queried pathway plus the information about the active- or inactive-state of such proteins (derived from the expression data of the cell-types or the tissues studied in each case). The "edges" of the network correspond to the links or associations between the biomolecular entities (derived from the information included in the pathways). These links can be activation, inhibition, expression, phosphorylation, etc. In order to facilitate further analysis of the networks, the edges generated by *Path2enet* are taken as undirected.

Considering the coverage over the map of human pathways, *Path2enet* can generate two different types of networks. The first is the "local" network which strictly includes the nodes of the canonical pathway selected from KEGG. For example, in the case of the NOTCH Signaling Pathway (KEGG ID: hsa04330) (Fig. 1a) the "local" network retrieves the 48 genes/proteins that are included in this pathway for human (*Homo sapiens*). Thus, *Path2enet* generates a network where each node is a protein and the edges are colored according to the type of association reported in the pathway (Fig. 1b). The second type of network that *Path2enet* can build is the "global" pathway-network that includes all the "local" nodes and links from a given KEGG pathway, plus all the extra "external" nodes that such nodes can be linked to in other pathway charts (i.e., it provides the links to other nodes in any biological pathway of the whole human repertoire). In this way, *Path2enet* is not restricted to predefined pathways since it can create large networks blending multiple layers of biological information.

Once a network is built with *Path2enet*, calculations of the network topological parameters (such as degree, betweeness, clustering coefficient, eigenvector value, etc.) can be performed, because the tool generates *igraph* objects [16], that can be studied with graph analysis tools. In this way, *Path2enet* provides ways to identify hubs and clusters in the network.

Application of *Path2enet* to build the NOTCH pathway-network of B and T cells

In the case study presented in this article we used *Path2enet* to generate expression networks of the NOTCH signaling pathway in three types of human cells: B cells (CD 19+) and T cells (CD 4+ and CD8+) (Fig. 2). To achieve this, we used a sample dataset of microarray expression (indicated in Methods).

First, we needed to apply the gene expression *Barcode* analysis to the gene products present in the NOTCH pathway-network (Fig. 1b) to identify which nodes were active in these cell types. The quantitative results of these analyses are presented in Fig. 3. Using the treshold of 0.4 for the normalized expression, the B cell network expressed 34 of 48 of the NOTCH pathway proteins. In contrast, the T cell network expressed 22–24 of the NOTCH proteins. It was very interesting to show that in all lymphocytes DLL1/2/3 and JAG1/2 were absent (i.e., they were OFF). In fact, these proteins are ligands of the

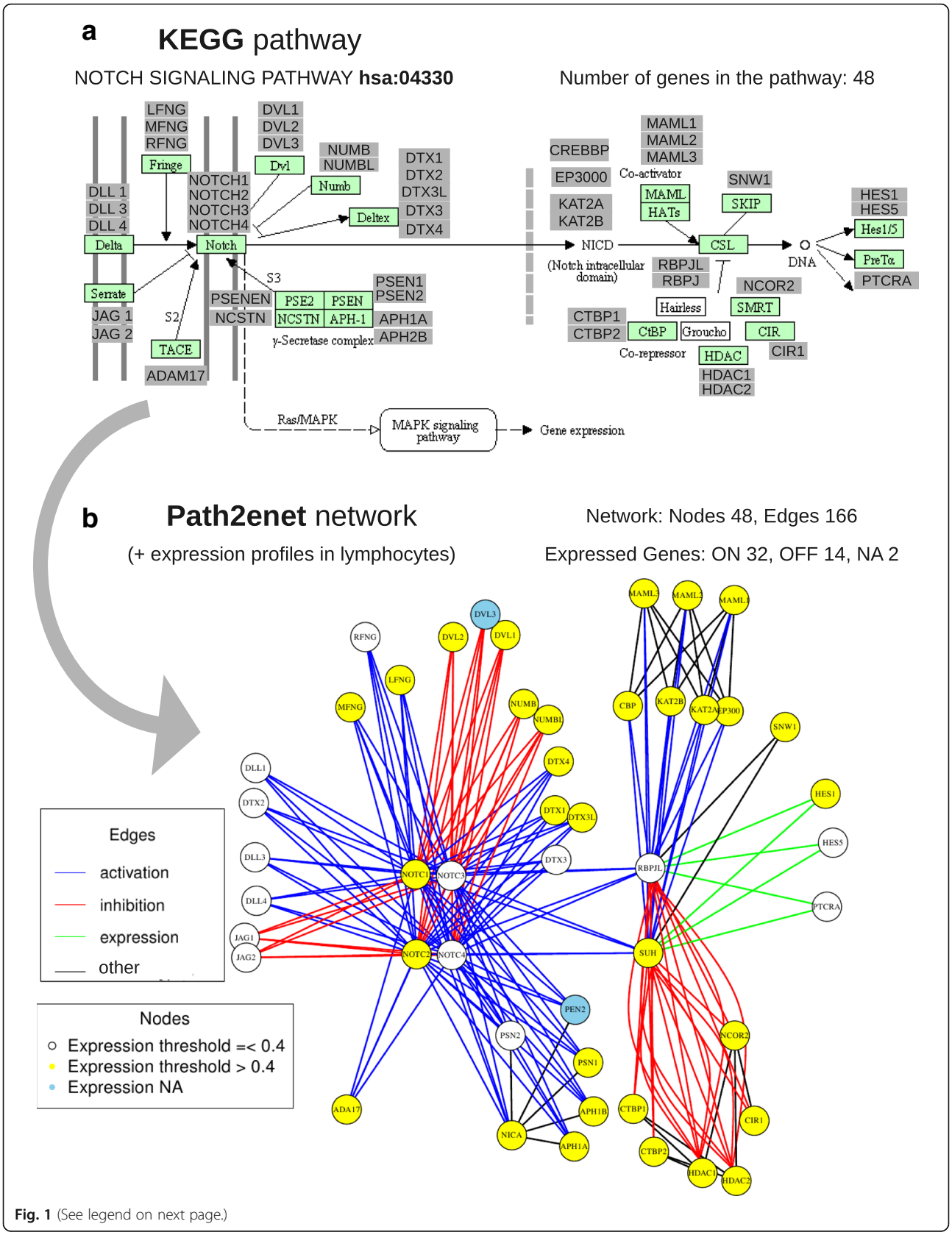


Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 Figure showing the NOTCH Signaling Pathway and the transformation into a network using *Path2enet*: (**a**) the canonical map included in the KEGG database (hsa:04330), showing all the distinct proteins that are included in each node of this pathway (total 48 proteins); (**b**) transfromation of the canonical pathway to a network done with *Path2enet*, that produces a new view incorporating all the information about the gene products that are expressed (i.e., active) based on the pathway and on the expression data sets of lymphocytes that are incorporated (163 samples: 32 B cells and 131 T cells). *Path2enet* uses the gene expression *Barcode* algorithm to evaluate if a gene is expressed (ON, yellow) in the network or not (OFF, white). The genes in blue correspond to NA (not assigned) since the tool could not assign them because they are not present in the expression platform used. The panels inside the figure indicate: "Edges" the characteristics of the links/relations that connect each node (blue, activation; red, inhibition; green, expression; black, other type of link); "Nodes" the expression level assigned to each node in the sample type studied that in this case were lymphocytes (white nodes, when their expression is below the threshold ≤0.4; yellow nodes, when their expression is above the threshold >0.4; blue nodes, when the expression level is not assigned)

NOTCH receptors of lymphocytes coming from the external cells that connect to them, therefore they should not be present in the lymphocytes. This is clearly shown in the quantitative analysis (Fig. 3), since all these genes were labeled OFF (not expressed) in B cells and in T cells.

We also observed that the only NOTCH paralogs detected in the lymphocytes were NOTCH2 and some NOTCH1. It is well known that NOTCH2 is preferentially expressed in mature naive B cells and interacts with DTX1, thus playing an important role in B cell development [17]. We also saw that the level of DTX1 in B cells was much higher (DTX1 = 1.00) than in T cells CD4+ (0.41) or CD8+ (0.17) (Fig. 3). This result is also in agreement with several studies that have shown that T cells are normally developed in absence of DTX1 [18].

Finally, another differential protein found expressed in B cells but not in T cells was the transcription factor HES1. The presence and role of this transcription factor in lymphocytes has been proven in several studies [19, 20]. In fact, it has been indicated that in T cells HES1 is dispensable beyond the beta selection checkpoint [21]. This explains our detection of HES1 in B cells CD19+ and its abscence in T cells CD4+ and CD8 + .

As a whole the data presented in Figs. 2 and 3 were very consistent with our current knowledge of the role of the NOTCH pathway in human B and T lymphocytes, enhancing the value of generating well defined "pathway-expression-networks" for specific cell types which is the scope of *Path2enet*.

Path2enet tool for pathways: usability and formats

KEGG pathways database (http://www.kegg.jp/) provides KGML files for each biological pathway on its website. For example, in the case of the human NOTCH signaling pathway (KEGG ID reference: hsa04330) the KGML file can be downloaded freely as "hsa04330.xml". The link for this file is: http://www.kegg.jp/kegg-bin/download?entry=hsa04330&format=kgml. In this way, any specific pathway is accessible via its KGML file in the KEGG website and *Path2enet* R package provides functions to download these files and create a MySQL database derived from the KGMLs (as explained in the R vignette included with *Path2enet*). Moreover, to facilitate the use of the pathway KMGL files within the application *Path2enet*, we also provided an SQL dump file ("Path2enet_KeggSQL.sql") generated with all the KMGL files of *Homo sapiens* (this datafile is provided at: http://bioinfow.dep.usal.es/path2enet/). This allows the creation of the necessary SQL database within the user's computer to query for specific pathways and to use the other functions of *Path2enet*. This database resource is not just a compendium of KMGL files from KEGG given that it provides some important added values: (i) it includes a mapping of all the gene and protein identifiers (IDs) from KEGG to the IDs of *UniProtKB* (used as the reference protein database in *Path2enet*); (ii) it includes a relational SQL structure, based on the extracted data from the pathways, that allocates such information in two principal indexed tables: one describing the pair-wise links or relations between protein pairs, and another one describing the characteristics of each singular protein.

With respect to the use of other formats, other than XML and KGML, *Path2enet* can also use any database or resource provided in a "network structure" as an *igraph* object, because the tool includes functions to read and load in R *igraph* objects. For the use of other standard formats, such as SBML or BioPAX, there are already tools that address this scope. For example *KEGGtranslator* [22], an easy-to-use stand-alone application that can visualize and convert KGML formatted XML-files into multiple output formats. This tool supports a plethora of output formats, being able to increase the information in translated documents beyond the scope of the KGML document. *KEGGtranslator* converts KEGG files (KGML formatted XML-files) to SBML, BioPAX, SIF, SBGN, SBML-qual, GML, GraphML and LaTeX. Moreover, in *Bioconductor* (https://www.bioconductor.org/) there are packages to parse, modify and visualize BioPAX data, like *rBiopaxParser* [23] or *PaxtoolsR* [24]. At the moment, we are working on a workflow to use these packages to create SQL databases, similiar to the SQL described above, but using data from other pathway resources such as Reactome or Pathway Commons. This work is under development, but one of main problems in the use of these resources is not the use of standard formats, like BioPAX or SBML, but the accurate mapping to standard protein identifiers from UniProtKB.

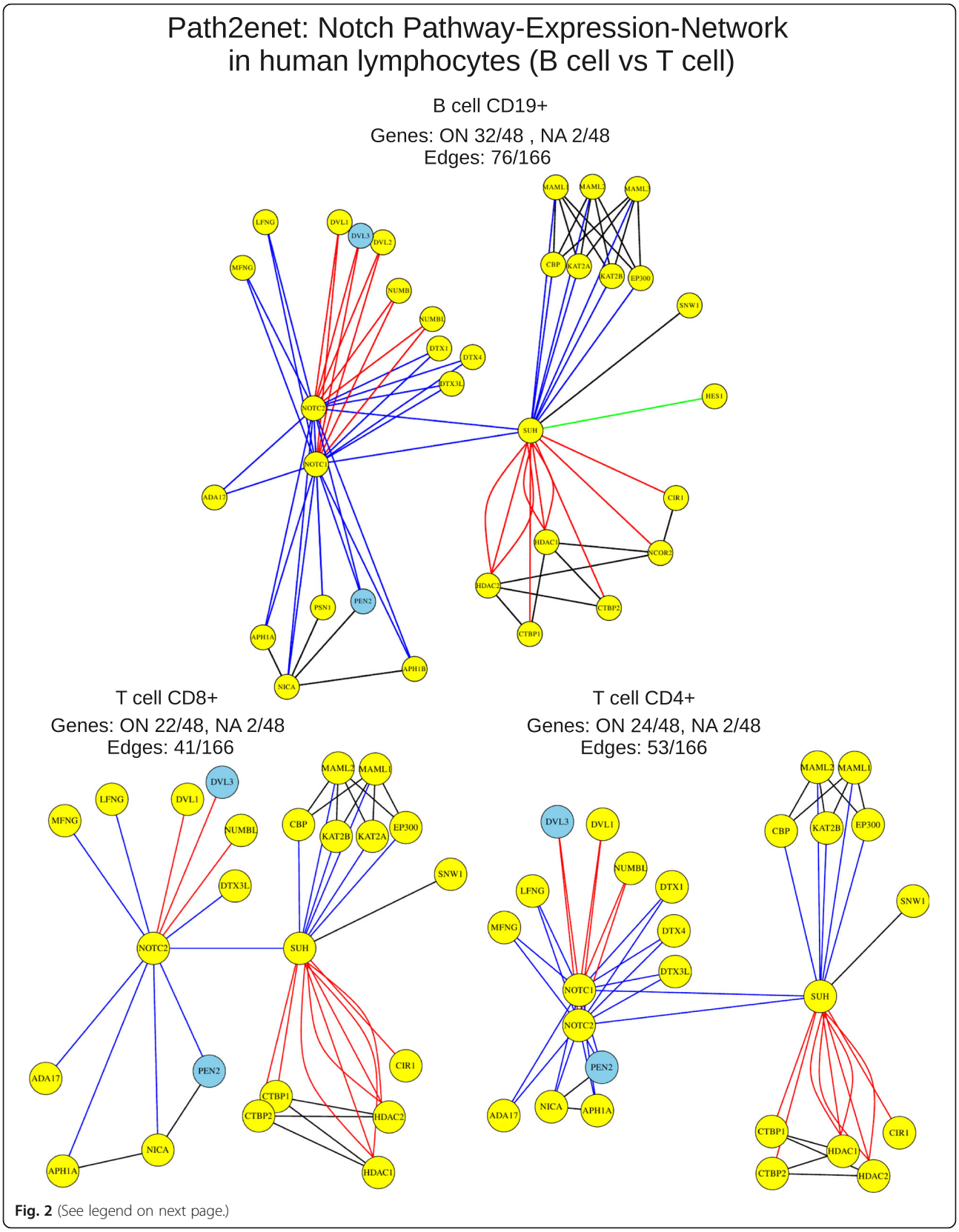


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Pathway-expression-networks produced for B cell CD19+ (top), T cell CD8+ (bottom left) and T cell CD4+ (bottom right) based on the transformation of the NOTCH Signaling Pathway done with *Path2enet*. The tool removed all nodes below the expression treshhold of 0.4. The expression network produced for B cells reveals 32 active nodes (ON) plus 2 NAs out of 48 proteins and 76 edges (out of 166 maximum). The expression network for T cells reveals 22–24 active nodes (ON) plus 2 NAs out of 48 proteins and 41–53 edges (out of 166 maximum)

Conclusions

Path2enet produces pathway-expression-networks reading and integrating high quality pathway data, protein interaction data and expression cell specific data. The development of this type of tools can be very useful to achieve a more integrative and global view of the links and association between the proteins working in specific cellular systems. The tool is not restricted to predefined pathways since it can create large networks blending multiple layers of biological information. Moreover, the tool can use either pre-processed expression data from selected repositories or experimental expression data from RNA-Seq or microarrays.

In this study we applied *Path2enet* to the analysis of the NOTCH signaling pathway in B cells and T cells. We showed that the expression networks based on a large microarray data set of these samples are different for each cell type, modulating the original general view of the canonical pathway provided by KEGG. Moreover, the observed differences have clear biological meaning,

Notch Signaling Pathway in human lymphocytes using *Path2enet* (expression level ON/OFF done with *Barcode*)

Gene Symbols (KEGG)	Uniprot IDs	Bcell_CD19+		Tcell_CD4+		Tcell_CD8+	
ADAM 17	ADA17_HUMAN	0.802	ON	0.538	ON	0.621	ON
APH1A	APH1A_HUMAN	0.766	ON	0.568	ON	0.552	ON
APH1B	APH1B_HUMAN	0.594	ON	0.344	OFF	0.138	OFF
CREBBP	CBP_HUMAN	0.760	ON	0.622	ON	0.667	ON
CIR1	CIR1_HUMAN	0.969	ON	0.813	ON	0.517	ON
CTBP1	CTBP1_HUMAN	0.750	ON	0.852	ON	0.767	ON
CTBP2	CTBP2_HUMAN	0.445	ON	0.583	ON	0.578	ON
DLL1	DLL1_HUMAN	0.203	OFF	0.292	OFF	0.310	OFF
DLL3	DLL3_HUMAN	0.047	OFF	0.052	OFF	0.103	OFF
DLL4	DLL4_HUMAN	0.125	OFF	0.063	OFF	0.138	OFF
DTX1	DTX1_HUMAN	1.000	ON	0.417	ON	0.172	OFF
DTX2	DTX2_HUMAN	0.219	OFF	0.083	OFF	0.069	OFF
DTX3	DTX3_HUMAN	0.016	OFF	0.292	OFF	0.345	OFF
DTX3L	DTX3L_HUMAN	1.000	ON	0.990	ON	0.931	ON
DTX4	DTX4_HUMAN	1.000	ON	0.615	ON	0.207	OFF
DVL1	DVL1_HUMAN	1.000	ON	0.979	ON	1.000	ON
DVL2	DVL2_HUMAN	0.578	ON	0.266	OFF	0.259	OFF
DVL3	DVL3_HUMAN	–	NA	–	NA	–	NA
EP3000	EP300_HUMAN	0.906	ON	0.760	ON	0.810	ON
HDAC1	HDAC1_HUMAN	1.000	ON	1.000	ON	1.000	ON
HDAC2	HDAC2_HUMAN	0.656	ON	0.510	ON	0.517	ON
HES1	HES1_HUMAN	0.594	ON	0.198	OFF	0.103	OFF
HES5	HES5_HUMAN	0.000	OFF	0.031	OFF	0.034	OFF
JAG1	JAG1_HUMAN	0.227	OFF	0.224	OFF	0.069	OFF
JAG2	JAG2_HUMAN	0.031	OFF	0.286	OFF	0.172	OFF
KAT2A	KAT2A_HUMAN	0.750	ON	0.385	OFF	0.483	ON
KAT2B	KAT2B_HUMAN	1.000	ON	1.000	ON	0.828	ON
LFNG	LFNG_HUMAN	0.719	ON	0.510	ON	0.621	ON
MFNG	MFNG_HUMAN	0.891	ON	0.964	ON	0.966	ON
RFNG	RFNG_HUMAN	0.094	OFF	0.219	OFF	0.034	OFF
MAML1	MAML1_HUMAN	0.969	ON	1.000	ON	1.000	ON
MAML2	MAML2_HUMAN	0.719	ON	0.661	ON	0.810	ON
MAML3	MAML3_HUMAN	0.547	ON	0.281	OFF	0.069	OFF
NCOR2	NCOR2_HUMAN	0.431	ON	0.267	OFF	0.317	OFF
NCSTN	NICA_HUMAN	0.766	ON	0.479	ON	0.552	ON
NOTCH1	NOTC1_HUMAN	0.453	ON	0.427	ON	0.379	OFF
NOTCH2	NOTC2_HUMAN	0.914	ON	0.680	ON	0.664	ON
NOTCH3	NOTC3_HUMAN	0.031	OFF	0.073	OFF	0.069	OFF
NOTCH4	NOTC4_HUMAN	0.016	OFF	0.068	OFF	0.069	OFF
NUMB	NUMB_HUMAN	0.828	ON	0.341	OFF	0.371	OFF
NUMBL	NUMBL_HUMAN	0.516	ON	0.521	ON	0.569	ON
PSEN1	PSN1_HUMAN	0.581	ON	0.350	OFF	0.290	OFF
PSEN2	PSN2_HUMAN	0.281	OFF	0.104	OFF	0.046	OFF
PSENEN	PEN2_HUMAN	–	NA	–	NA	–	NA
PTCRA	PTCRA_HUMAN	0.146	OFF	0.323	OFF	0.253	OFF
SNW1	SNW1_HUMAN	0.984	ON	0.990	ON	0.914	ON
RBPJ	SUH_HUMAN	0.740	ON	0.691	ON	0.690	ON
RBPJL	RBPJL_HUMAN	0.000	OFF	0.063	OFF	0.103	OFF
		N proteins ON	32/48		24/48		22/48

Fig. 3 Results of the analysis performed for the NOTCH Signaling Pathway (including 48 genes) with *Path2enet* based on the gene expression *Barcode* algorithm for the data sets of B cell CD19+, T cell CD4+ and T cell CD8+. The data sets correspond to 163 samples of expression microarrays. The treshold to indicate if a gene was expressed (ON) or (OFF) is 0.4. Genes labeled with NA are not present in the expression platform and thus could not be annotated

as demonstrated, for example, when only 2 out of the 4 NOTCH paralog proteins (NOTCH1, 2, 3, 4) were expressed in B cells and T cells. Thus, a clear signal in all lymphocytes was observed for NOTCH2; while NOTCH1 was also detected in B cells CD19+ and in T cells CD4+. We also found that key regulators like DTX1 and HES1 are strongly expressed in B cells and less expressed, or not present, in T cells. All these results give support to the the value of the networks that *Path2enet* generates that are cell-type and context specific. In conclusion, users have the possibility to combine several pathways and include protein-protein interaction data to find key players in a specific biological context either for normal or for pathological samples.

Additional files

Additional file 1: List of 163 high-density oligonucleotides expression microarrays from human B-cells and T-cells used in this study (taken from Gene Expression Omnibus, GEO, database). (DOCX 125 kb)

Additional file 2: R vignette provided as a guided tutorial to facilitate the installation and use of the *Path2enet* package. (HTML 9198 kb)

Acknowledgements
We acknowledge the funding provided to Dr. J. De Las Rivas group by the Local Government, "Junta de Castilla y Leon" (JCyL, Valladolid, Spain, grant number BIO/SA08/14); and by the Spanish Government, "Ministerio de Economia y Competitividad" (MINECO) with grants of the ISCiii co-funded by FEDER (grant references PI12/00624 and PI15/00328). We also acknowledge a PhD research grant to Conrad Droste ("Ayudas a la Contratación de Personal Investigador") provided by the JCyL with the support of the "Fondo Social Europeo" (FSE).

Declarations
About this supplement
This article has been published as part of *BMC Genomics* Volume 17 Supplement 8: Selected articles from the Sixth International Conference of the Iberoamerican Society for Bioinformatics on Bioinformatics and Computational Biology for Innovative Genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-17-supplement-8.

Funding
The publication costs for this article were funded by the research grant PI12/00624, from the *Instituto de Salud Carlos III* (ISCiii) co-funded by the *Fondo Europeo de Desarrollo Regional* (FEDER).

Availability of data and materials
The data and materials supporting the results of this article, incuding the R package *Path2enet* and all the Additional files, are available at: http://bioinfow.dep.usal.es/path2enet/. In particular, the SQL file "Path2enet_KeggSQL.sql" is available at such URL.

Authors' contributions
CD developed and documented the R package including the integration of all the databases and resources that this tool uses. He also carried out the data collection for several analyses, trials and comparisons using the package. JDLR designed the study, coordinated the trials along the software developed, supervised the data analysis and wrote the manuscript. CD also helped to write the manuscript. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable

Ethics approval and consent to participate
Not applicable. Our work only uses human data from open public databases and it does not include any personal information.

Published: 25 October 2016

References
1. Aranda B, Blankenburg H, Kerrien S, Brinkman FS, Ceol A, Chautard E, et al. PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nat Methods. 2011;8:528–9.
2. Yang X, Coulombe-Huntington J, Kang S, Sheynkman GM, Hao T, Richardson A, et al. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. Cell. 2016;164:805–17.
3. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
4. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–62.
5. Prieto C, De Las Rivas J. APID: Agile Protein Interaction DataAnalyzer. Nucleic Acids Res. 2006;34:W298–302.
6. Alonso-López D, Gutiérrez MA, Lopes KP, Prieto C, Santamaria R, De Las Rivas J. APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. Nucleic Acids Res. 2016;44:W529–35.
7. McCall MN, Uppal K, Jaffee HA, Ziliox MJ, Irizzarry RA. The Gene Expression Barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. Nucleic Acids Res. 2011;39:D1011–5.
8. McCall MN, Jaffee HA, Zelisko SJ, Sinha N, Hooiveld G, Irizzarry RA, Ziliox MJ. The Gene Expression Barcode 3.0: improved data processing and mining tools. Nucleic Acids Res. 2014;42:D938–43.
9. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2012;31:46–53.
10. Durinck S, Spellman PT, Birnez E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4:1184–91.
11. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015;347:1260419.
12. de Leeuw WC, Rauwerda H, Jonker MJ, Breit TM. Salvaging Affymetrix probes after probe-level re-annotation. BMC Res Notes. 2008;1:66.
13. Risueño A, Fontanillo C, Dinger ME, De Las Rivas J. GATExplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs. BMC Bioinformatics. 2010;11:221.
14. Magrane M, UniProt Consortium. UniProt Knowledgebase: a hub of integrated protein data. Database. 2011;2011:bar009.
15. James DA, Debroy S. RMySQL: R interface to the MySQL database. R package version 0.9-3. 2012. http://CRAN.R-project.org/package=RMySQL.
16. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal, Complex Systems. 2006;1695:1–9.
17. Saito T, Chiba S, Ichikawa M, Kunisato A, Asai T, Shimizu K, et al. Notch2 is preferentially expressed in mature B cells and indispensable for marginal zone B lineage development. Immunity. 2003;18:675–85.
18. Lehar SM, Bevan MJ. T cells develop normally in the absence of both Deltex1 and Deltex2. Mol Cell Biol. 2006;26:7358–71.
19. Maillard I, Koch U, Dumortier A, Shestova O, Xu L, Sai H, et al. Canonical notch signaling is dispensable for the maintenance of adult hematopoietic stem cells. Cell Stem Cell. 2008;2:356–66.
20. Yu X, Alder JK, Chun JH, Friedman AD, Heimfeld S, Cheng L, Civin CI. HES1 inhibits cycling of hematopoietic progenitor cells via DNA binding. Stem Cells. 2006;24(4):876–88.
21. Wendorff AA, Koch U, Wunderlich FT, Wirth S, Dubey C, Brüning JC, et al. Hes1 is a critical but context-dependent mediator of canonical Notch signaling in lymphocyte development and transformation. Immunity. 2010;33:671–84.

22. Wrzodek C, Dräger A, Zell A. KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats. Bioinformatics. 2011;27:2314–5.
23. Kramer F, Bayerlova M, Klemm F, Bleckmann A, Beissbarth T. rBiopaxParser - an R package to parse, modify and visualize BioPAX data. Bioinformatics. 2013;29:520–2.
24. Luna A, Babur Ö, Aksoy BA, Demir E, Sander C. PaxtoolsR: pathway analysis in R using Pathway Commons. Bioinformatics. 2016;32:1262–4.

Journal of proteome research

Article

pubs.acs.org/jpr

Integration of Proteomics and Transcriptomics Data Sets for the Analysis of a Lymphoma B-Cell Line in the Context of the Chromosome-Centric Human Proteome Project

Paula Díez,[†,‡] Conrad Droste,[§] Rosa M. Dégano,[‡] María González-Muñoz,[†] Nieves Ibarrola,[‡] Martín Pérez-Andrés,[†] Alba Garin-Muga,[∥] Víctor Segura,[∥] Gyorgy Marko-Varga,[⊥] Joshua LaBaer,[#] Alberto Orfao,[†] Fernando J. Corrales,[∥] Javier De Las Rivas,[*,§] and Manuel Fuentes[*,†,‡]

[†]Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

[‡]Proteomics Unit. Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

[§]Bioinformatics and Functional Genomics Research Group, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

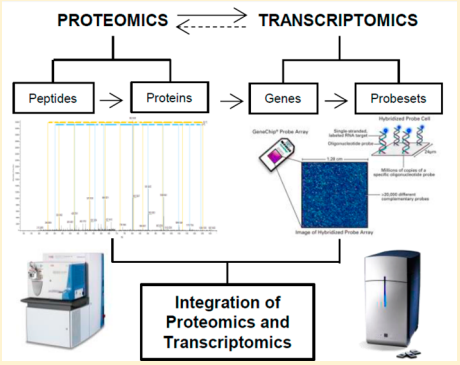
[∥]Division of Hepatology and Gene Therapy, Proteomics and Bioinformatics Unit, Centre for Applied Medical Research (CIMA), University of Navarra, 31008 Pamplona, Spain

[⊥]Clinical Protein Science and Imaging, Biomedical Centre, Department of Biomedical Engineering, Lund University, BMC D13, 221 84 Lund, Sweden

[#]Biodesign Institute, Arizona State University, 1001 South McAllister Avenue, Tempe, Arizona 85287, United States

Supporting Information

ABSTRACT: A comprehensive study of the molecular active landscape of human cells can be undertaken to integrate two different but complementary perspectives: transcriptomics, and proteomics. After the genome era, proteomics has emerged as a powerful tool to simultaneously identify and characterize the compendium of thousands of different proteins active in a cell. Thus, the Chromosome-centric Human Proteome Project (C-HPP) is promoting a full characterization of the human proteome combining high-throughput proteomics with the data derived from genome-wide expression profiling of protein-coding genes. Here we present a full proteomic profiling of a human lymphoma B-cell line (*Ramos*) performed using a nanoUPLC-LTQ-Orbitrap Velos proteomic platform, combined to an in-depth transcriptomic profiling of the same cell type. Data are available via ProteomeXchange with identifier PXD001933. Integration of the proteomic and transcriptomic data sets revealed a 94% overlap in the proteins identified by both -omics approaches. Moreover, functional enrichment analysis of the proteomic profiles showed an enrichment of several functions directly related to the biological and morphological characteristics of B-cells. In turn, about 30% of all protein-coding genes present in the whole human genome were identified as being expressed by the *Ramos* cells (stable average of 30% genes along all the chromosomes), revealing the size of the protein expression-set present in one specific human cell type. Additionally, the identification of missing proteins in our data sets has been reported, highlighting the power of the approach. Also, a comparison between neXtProt and UniProt database searches has been performed. In summary, our transcriptomic and proteomic experimental profiling provided a high coverage report of the expressed proteome from a human lymphoma B-cell type with a clear insight into the biological processes that characterized these cells. In this way, we demonstrated the usefulness of combining -omics for a comprehensive characterization of specific biological systems.



KEYWORDS: *C-HPP, lymphoma B-cell line, protein expression profile, transcriptomics, subcellular fractionation*

INTRODUCTION

Upon successful completion of the Human Genome Project in 2003—approximately 20 055 protein-coding genes have been reported according to neXtProt version of September 19, 2014—a major challenge remains as regards the understanding of how gene expression levels relate to the regulatory behavior

Special Issue: The Chromosome-Centric Human Proteome Project 2015

Received: May 28, 2015
Published: July 28, 2015

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. 2015, 14, 3530−3540

of cells.[1,2] After the genomics era, the scientific community realized that DNA coding sequences are not by themselves sufficient to provide an overview of cellular biological processes.[3] The genome remains nearly constant throughout the lifetime of a cell and there are no significant changes regarding the cell type once it achieves its differentiated specific state. However, both the transcriptome and the proteome are much more dynamic and they can vary with the functional state of the cell or in response to intra- and extra-cellular environmental signals. Henceforth, studying changes in mRNA and protein levels can provide a clear and accurate readout of the cell state.[1] Proteins are usually the final regulatory and effector molecules of cells coded by genes, and proteomics allows a comprehensive and integrative study of all proteins in a cellular system.[4] The main goal of proteomics is often to generate a complete and quantitative map of proteins, including cellular localization of proteins, identification of protein complexes, protein isoforms, and post-translational protein modifications (PTM). In fact, proteomes are characterized by large protein-abundance differences, cell-type, time-dependent expression patterns, and PTMs, all of which carry biological information that is not commonly accessible by genomics or transcriptomics data.

In turn, transcriptomics methodologies are the most used to determine the active expression of predicted protein-coding genes. In this respect, RNA deep sequencing technology has emerged as a promising strategy that provides a whole transcriptome shotgun approach to quantifying detailed genome-wide expression.[5,6] Previous to the development of high-throughput RNA and DNA sequencing technologies, microarray technologies have been considered powerful large-scale methods that have been applied for gene expression profiling of multiple samples in many different biological studies. Nevertheless, the detection and quantification of expressed mRNAs do not unequivocally determine the presence of the corresponding translated proteins. Regulatory mechanisms, such as PTMs or silencing processes, can result in an imbalance between the transcribed and the translated portions, as well as the half-live differences between transcripts and proteins.[3,7−11]

Proteomics measures proteins directly, providing information about the active genes at the translational level, and can be used as verification of gene expression.[3,12] Major advances which have occurred in proteomics over the last years have allowed detection and validation of putative genes, together with the added benefit for genome annotation.[13,14] However, many challenges still remain in these approaches due to proteome complexity.[15] Among others, these are related to (i) the sample preparation procedures, because to get maximum coverage of the proteome it is necessary to use multiple sample fractionation methods, either through protein extraction protocols or processing techniques; (ii) the peptide separation optimized via the usage of combinations of different proteases (e.g., trypsin, Lys-C, proteinase K, etc.) to increase peptide recovery leading to increase protein sequence coverage; (iii) the precision and accuracy of mass spectrometry (MS) measurements, because to reduce the occurrence of false positive hits it is recommended to use instruments with high resolution, high sensitivity, fast scanning speed, and high mass accuracy (ppm) such as the Fourier Transform Ion Cyclotron Resonance (FT-ICR) and the Linear Trap Quadrupole-Orbitrap (LTQ-Orbitrap) (also affected by a different ionization efficiency of different peptides); (iv) the data processing methods and database search engines (i.e., Mascot, Sequest, OMSSA), as well as the scoring and validation parameters (false discovery rate − FDR, precursor mass tolerance), which determine the robustness of the results about the identified peptides and proteins[9,13−13].

All these challenges are still open in proteomics; therefore, the integration of transcriptomics and proteomics data, working with adequate bioinformatics strategies, can offer new insights in the field[1,16] and provide reliable information about how genes and proteins are regulated and integrated at the molecular, cellular, and organismal levels to control a set of biological responses.[17]

The Human Proteome Organization (HUPO) has coordinated the efforts of the international community promoting several initiatives to describe the human proteome through a well-planned working scheme. The project is partially organized according to a chromosome-based strategy (C-HPP) where scientific groups from different nationalities agreed to characterize the proteome of a specific chromosome.[18] All 24 chromosomes plus the mitochondrial DNA have been already assigned to many teams from 21 different countries. The vast heterogeneity, wide dynamic range, and different ionization efficiencies of peptides are causing a restriction in detection and quantification capacities of large-scale proteomics studies. Hence, C-HPP groups are now integrating transcriptomics and proteomics data sets in order to better guide the genome-wide proteomics analysis.[19] Specifically, the Spanish team of the Human Proteome Project (SpHPP) addresses the protein mapping of chromosome 16, and that it is the reason why a lymphoma cell line has been used in this research, because a great number of proteins from B-cells are encoded in chromosome 16.[2,20] The development of studies integrating proteomics and transcriptomics may lead to the full characterization of chromosomes and also to determine the relationship between the transcripts and their products (i.e., proteins). Additionally, the usage of shotgun MS approaches generates huge amounts of data providing lots of information in which it might be possible to detect missing proteins. Also, the C-HPP initiative is expected to improve the knowledge about diseases and their biology contributing to the Biology and Disease (B/D)-HPP initiative.

Here, we present a workflow integrating transcriptomics and proteomics data sets about a specific biological sample. As a model, we selected the *Ramos* human B-cell lymphoma cell line, which is well-characterized at the gene expression level. In addition, the proposed approach has reported the identification of missing proteins that correspond to certain subset of protein-coding genes well-annotated in the human genome that have not been yet detected by any proteomic MS-based experimental approach. Additionally, a further comparison between neXtProt and UniProt database searches has been accomplished. On balance, all these points may be of great interest for the scientific community and, specifically, for the C-HPP consortium.

MATERIALS AND METHODS

Reagents

Protease inhibitor cocktail, 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), Tween 20, tris (2-carboxyethyl) phosphine hydrochloride (TCEP), phenylmethanesulfonyl fluoride (PMSF), digitonin, octylphenoxypolyethoxyethanol (IGEPAL), RPMI-1640 media, potassium ferrocyanide, sodium

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. 2015, 14, 3530−3540

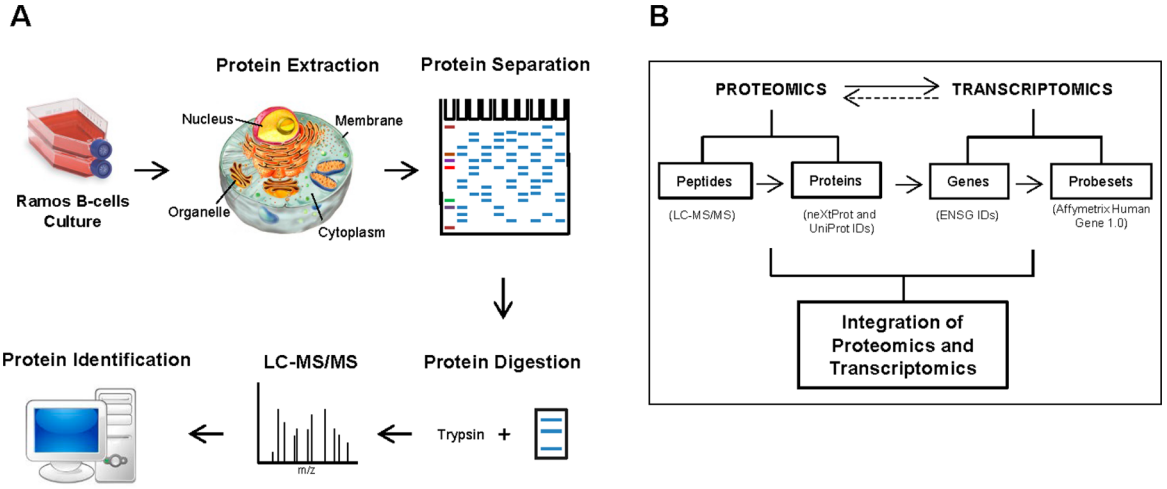


Figure 1. Schematic representation of the strategies followed for the integration of proteomics and transcriptomics analysis. (A) Overview of the experimental workflow. Subcellular protein extraction was performed from *Ramos* B-cells. After protein separation in an SDS-PAGE gradient gel (4−20%), proteins were digested with trypsin. The digests were analyzed using an nUPLC-LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher Scientific). SEQUEST and MASCOT database search algorithms were used for protein identification. (B) Integration of proteomics and transcriptomics workflow. Comparison of proteomics and transcriptomics data was made via mapping from peptides (obtained by an LC-MS/MS strategy) to DNA probes (Affymetrix Human Gene 1.0 platform).

thiosulfate, dithiothreitol (DTT), iodoacetamide (IAA), formic acid (FA), and acetonitrile (ACN) were purchased from Sigma (St. Louis/MO, U.S.A.). Heat-inactivated fetal bovine serum (FBS), L-glutamine, penicillin, and streptomycin were obtained from Gibco (Scotland, U.K.). n-Dodecyl-β-D-maltopyranoside was purchased from Affymetrix (Maumee, OH, U.S.A.), Coomassie Brilliant Blue was from Merck (Kenilworth, NJ, U.S.A.), and trypsin was from Promega (Madison, WI, U.S.A.). Monoclonal antihuman antibodies conjugated with phycoerythrin (PE): CD20 (clone L27), CD22 (clone HIB22), CD11a (clone HI111), CD45 (clone HI30), CD19 (clone HIB19), and CD3 (clone SK7) were from BD Biosciences (Pharmingen, San Diego, CA, U.S.A.); CD79b (clone CB3−1), and IgG1 mouse isotypic control antibody was from BioLegend (San Diego, CA, U.S.A.). Polyclonal antihuman antibodies conjugated with PE: kappa-Ig and lambda-Ig light chains were purchased from Cytognos SL (Salamanca, Spain).

The hypotonic lysis buffer used contained 30 mM HEPES pH = 8, 20% (v/v) glycerol, 15 mM KCl, 1 mM EDTA, 2 mM $MgCl_2$, 1 mM PMSF, 1 mM TCEP, and 1% (v/v) protease inhibitor cocktail. The hypertonic lysis buffer contains the same components as the hypotonic lysis buffer, except glycerol.

Cell Culture

The *Ramos* Burkitt's lymphoma-derived B-cell line (ATCC CRL 1596) was cultured at 37 °C in a humidified CO_2 incubator (5% CO_2) in complete RPMI media (RPMI-1640 medium supplemented with 10% (v/v) FBS, 200 mM L-glutamine, 10 000U/mL penicillin, and 10 000 μg/mL streptomycin).

Cell Harvest and Cell Lysis Methods

Cellular proteins were harvested by washing the cells (30×10^6 cells/experiment) twice with PBS, and cells were pelleted for 5 min at 1000*g*. The lysis buffer was then added at a volume equal to 5 times that of the cell pellet; all steps were performed at 4 °C (Figure 1A). The cell lysis methods A and B (Supplementary Figure 1) were performed in triplicate.

Method A. The hypotonic lysis buffer supplemented with 0.015% (w/v) digitonin was added to pelleted cells. After 30 min of rotation, the sample was centrifuged at 1500*g* for 5 min, and the cytoplasmic proteins (CYT) were collected in the supernatant. The remaining fractions were processed stepwise in an identical manner. For organelle proteins (ORG), the hypotonic lysis buffer with 0.5% (v/v) Tween 20 was used; for membrane proteins (MB), the hypotonic lysis buffer supplemented with 0.5% (v/v) IGEPAL detergent was employed, and lastly, nuclear proteins (NUC) were extracted by adding the hypertonic lysis buffer supplemented with 1% (w/v) n-dodecyl-β-D-maltopyranoside after 10 min incubation. Two washing steps were performed with nonsupplemented hypotonic lysis buffer between the distinct fractionation steps.

Method B. The CYT and ORG fractions were obtained as described in Method A. The NUC fraction was extracted with hypertonic lysis buffer supplemented with 140 mM NaCl. Lastly, the MB fraction was obtained after incubating for 5 min with hypotonic lysis buffer plus 1% (w/v) n-dodecyl-β-D-maltopyranoside.

Protein Quantification and SDS-PAGE Separation

After protein quantification by the Lowry-DC-Protein Assay as recommended by the manufacturer (Bio-Rad Laboratories, CA, U.S.A.), each sample was separated in a 4−20% gradient SDS-PAGE gel under reducing conditions. The same amount of protein (15 μg) was run for each fraction (CYT, ORG, NUC, MB). After electrophoresis, gels were stained in a solution of 0.5% (w/v) Coomassie Brilliant Blue. Gels were stored at 4 °C in an aqueous solution containing 1% (v/v) acetic acid, until analysis.

In-Gel Digestion and LC-MS/MS Analysis

Each gel lane was cut into five fragments and digested with trypsin following the method of Shevchenko et al.[21] with slight modifications. Briefly, gel pieces were destained with 15 mM potassium ferrocyanide and 50 mM sodium thiosulfate. Protein reduction and alkylation were performed with 10 mM DTT at 56 °C for 45 min, and with 55 mM IAA at room temperature for 30 min, respectively. Proteins were digested with trypsin (6.25 ng/mL) at 37 °C for 18 h. The peptide solution was acidified with FA and desalted by using C18-Stage-Tips

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. 2015, 14, 3530−3540

columns.[22] The samples were partially dried and stored at −20 °C until they were analyzed by LC-MS/MS.

A nanoUPLC system (nanoAcquity, Waters Corp., Milford, MA, U.S.A.) coupled to a LTQ-Velos-Orbitrap mass spectrometer (Thermo Fisher Scientific, San Jose,CA, U.S.A.) via a nanoelectrospray ion source (NanoSpray flex, Proxeon, Thermo) was used for reversed-phase LC-MS/MS analysis. Peptides were dissolved in 0.5% FA/3% ACN and loaded onto a trapping column (nanoACQUITY UPLC 2G-V/M Trap Symmetry 5 μm particle size, 180 μm × 20 mm C18 column, Waters Corp., Milford,MA, U.S.A.). Peptides were separated on a nanoACQUITY UPLC BEH 1.7 μm, 130 Å, 75 μm × 250 mm C18 column (Waters Corp., Milford,MA, U.S.A.) with a linear gradient from 7% to 35% solvent B (ACN/0.1% FA) at a flow rate of 250 nL/min over 120 min.

The nUPLC- LTQ-Orbitrap Velos was operated in the positive ion mode by applying a data-dependent automatic switch between survey MS scan and tandem mass spectra (MS/MS) acquisition. Survey scans were acquired in the mass range of m/z 400 to 1600 with a 60 000 resolution at m/z 400 with lock mass option enabled for the 445.120025 ion.[23]

The 20 most intense peaks having ≥2 charge state and above the 500 intensity threshold were selected in the ion trap for fragmentation by collision-induced dissociation with 35% normalized energy, 10 ms activation time, q = 0.25, ± 2 m/z precursor isolation width and wideband activation. Maximum injection time was 1000 and 50 ms for survey and MS/MS scans, respectively. AGC was 1×10^6 for MS and 5×10^3 for MS/MS scans. Dynamic exclusion was enabled for 90 s.

Database Search

Raw data were translated to mascot general file (mgf) format and searched against the neXtProt database (release September 19, 2014) using the target-decoy strategy with an in-house MASCOT Server v. 2.3 (Matrix Science, London, U.K.). Decoy database was created using the peptide pseudoreversed method, and separate searches were performed for target and decoy databases. Search parameters were set as follows: carbamidomethylation of cysteine as a fixed modification, oxidation of methionine and acetylation of the protein n-terminys as variable ones, precursor and fragment mass tolerance were set to 10 ppm and 0.8 Da, respectively, and fully tryptic digestion with up to two missed cleavages. FDR at PSM level (psmFDR) and protein level (protFDR) were calculated using MAYU.[24] Using C-HPP guidelines, protein identifications were obtained using the criteria psmFDR < 1% and protFDR < 1%. Lastly, protein inference was performed using the PAnalyzer algorithm,[25] and nonconclusive protein groups were discarded.

For the search against UniProt database, the MASCOT[26] and SEQUEST HT[27] algorithms were used to search for the acquired MS/MS spectra, using Thermo Scientific Proteome Discoverer software (v. 1.4.1.14) against a custom database of all human reviewed sequences downloaded from the UniProt database (February, 2014) and common contaminant sequences (e.g., human keratins, trypsin, BSA). Search parameters were the same as for the search against neXtProt database. Peptides having MASCOT ion scores of <20 were not considered for analysis. A 1% FDR using Percolator[28] was employed for peptide validation as well as for PSM level.

Supporting Information 1 contains raw data about the results obtained for representative samples.

Transcriptomic Analysis

To perform the gene expression profiling and analysis of the *Ramos* B-cells, we used a raw data set of three samples of mRNA from biological replicates of these cells hybridized on Affymetrix Human Gene ST 1.0 high-density oligonucleotide microarrays. The data are available at GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) database: series number GSE40168; samples GSM987747, GSM987748, GSM987749; platform GPL6244 ([HuGene-1_0-st]). The preprocessing, normalization, and signal calculation of these data were done using the Bioconductor packages oligo[29] and pd.hugene.1.0.st.v1.[30]

Integration of Transcriptomics and Proteomics Data Sets

In order to map the neXtProt IDs of the proteins identified in the LC-MS/MS proteomic assays to the corresponding genes and probesets detected by the gene expression transcriptomics assays, an ID-mapping procedure was run in two steps (Figure 1B): (1st) from protein neXtProt IDs to gene Ensembl IDs, using the mapping to genes provided by neXtProt database (release September 19, 2014); and (2nd) from the gene Ensembl IDs to gene Affymetrix probesets IDs, using the Brainarray tool[31] with the mapping table hugene10st_Hs_ENSG_mapping 18.0.0 corresponding to the arrays used. Before this ID-mapping, we unified all the neXtProt IDs obtained from the different isolated subcellular fractions (i.e., cytoplasm, organelles, membranes, and nucleus). Following these unification and mapping procedures, the neXtProt IDs and UniProt ACs were used to create different data sets with different levels of coverage and confidence based on the combination of the results obtained with three replicates (i.e., all proteomic experiments consisted of three independent biological replicates of the *Ramos* B-cells). In this way, the following data sets were generated (Table 1): (i) an *intersection* data set, including those proteins which were systematically identified in the three replicated experiments, with at least 2 proteotypic peptides at protFDR < 0.01; (ii) an *union* data set, including all proteins identified in any of the replicated experiments with at least 2 proteotypic peptides and protFDR < 0.01, and (iii) a *maximum* data set, including all proteins identified in any of the replicated experiments with at least 1 proteotypic peptide and protFDR < 0.01. These three data sets are included in each other, the third one being the one with the largest coverage (i.e., *maximum* data set) and therefore the one that provides the largest list of proteins identified.

To map proteins to genes in chromosomes, the Biomart[32] managed with R and Bioconductor tools were used. A brief R-script is provided as Supporting Information (Supporting Information 2) to show the details of the mapping protocol and the comparison of proteomic and transcriptomic data. Functional enrichment analysis (FEA) and clustering of the gene lists were done using the DAVID[33] and GeneTerm-Linker[34] tools. The main biological databases selected to find genes with annotated enriched terms were the following: (i) Gene Ontology (GO) using annotations spaces GOTERM_BP, GOTERM_CC and GOTERM_MF; (ii) the pathways database KEGG_PATHWAY; (iii) the INTERPRO and PFAM protein structural domain database; and (iv) the UNIGENE_EST and GNF_U133A_QUARTILE tissues-specific expression databases. To generate the functional clusters in DAVID, we used classification stringency *medium*.[33] All statistical analyses of data distributions, the comparisons, and most of the mapping were done working in the R/Bioconductor environment.[35]

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. 2015, 14, 3530−3540

Table 1. Number of Proteins and Genes Included in the Datasets Produced in the Analyses of *Ramos* B-Cells[a]

	no. of proteins (neXtProt)	no. of genes (Ensembl)	
	total neXtProt IDs	ENSG IDs	Affymetrix probeset IDs
intersection[b,c]	3383	3433	4088
union[b,d]	5494	5540	6175
maximum[d,e]	8931	8976	9494
"exclusive identifications" in transcriptomics[f]	-	1290	-
"exclusive identifications" in proteomics[g]	516	-	-

[a]The table indicates the number of protein and gene distinct IDs found in the proteomic and genomic assays, respectively: Row 1, *intersection* dataset; row 2, *union* dataset; row 3, *maximum* dataset (these datasets are defined in Materials and Methods). Row 4 ["Exclusive identifications" in Transcriptomics] includes the proteins that were not detected in proteomics but detected in the genomic data in the 25% highest expression quartile of the Affymetrix Human Gene ST 1.0 microarrays (calculating the expression signal average for the 3 arrays). Row 5 ["Exclusive identifications" in Proteomics] includes the genes that were not detectable by the genomic platform (i.e. genes not present in the microarray) but were detected by the proteomic approach. The columns in the table correspond to (i) all the human neXtProt IDs detected, (ii) the mapped Ensembl IDs, and (iii) the mapped Affymetrix probeset IDs. [b]Proteins detected by at least 2 unique peptides in the MS proteomic experiments. [c]Proteins detected in all the 3 experimental biological replicates. [d]Proteins detected in any replicate. [e]Proteins detected by at least 1 unique peptide in the MS proteomic experiments. [f]Genes detected in the 25% higher expression quartile of the microarrays but not present in the MS data. [g]Proteins detected in the MS/MS data but not present in the expression microarrays.

The same analysis strategy was performed for mapping UniProt IDs (using UniProtKB database release February 2014).

Immunophenotypic Analysis

Surface membrane expression of some proteins was validated by flow cytometry using a direct immunofluorescence technique with antihuman phycoerythrin (PE)-conjugated antibodies for the following proteins: CD3, CD11a, CD19, CD20, CD22, CD45, CD79b, kappa-Ig, and lambda-Ig light chains, together with a PE-conjugated isotype control antibody. All the antibodies were purchased from BDBiosciences (San José, CA, U.S.A.). Briefly, *Ramos* B-cells (0.5×10^6) were washed in 0.5% BSA in PBS and incubated with the antibodies for 15 min at room temperature, washed, and acquired on a FACSCanto II flow cytometer (BD Biosciences, San José, CA, U.S.A.) using the FACSDiva software (version 6.1, BD Biosciences). For data analysis, the Infinicyt software (Cytognos SL, Salamanca, Spain) was used.

Statistical Methods

For all continuous variables, mean values and their standard deviation (SD) were calculated. To evaluate the statistical significance of differences observed between groups, the two independent sample Student's *t* test was used for continuous variables displaying a normal distribution (SPSS software 18.0 package; SPSS, Inc., Chicago, IL). Statistical significance was set at a *P* value of <0.01.

RESULTS AND DISCUSSION

Subcellular Fractionation, Key To Increase Proteome Coverage

Mass-spectrometry techniques can increase the coverage of the proteome by using subcellular fractionation strategies for protein extraction.[20] This approach enables the in-depth analysis of biomolecules by reducing the sample complexity through isolation of different subcellular fractions. Here, we performed sequential extraction of proteins (from three biological replicates) by using a combination of different detergents specific for protein profiling from distinct subcellular

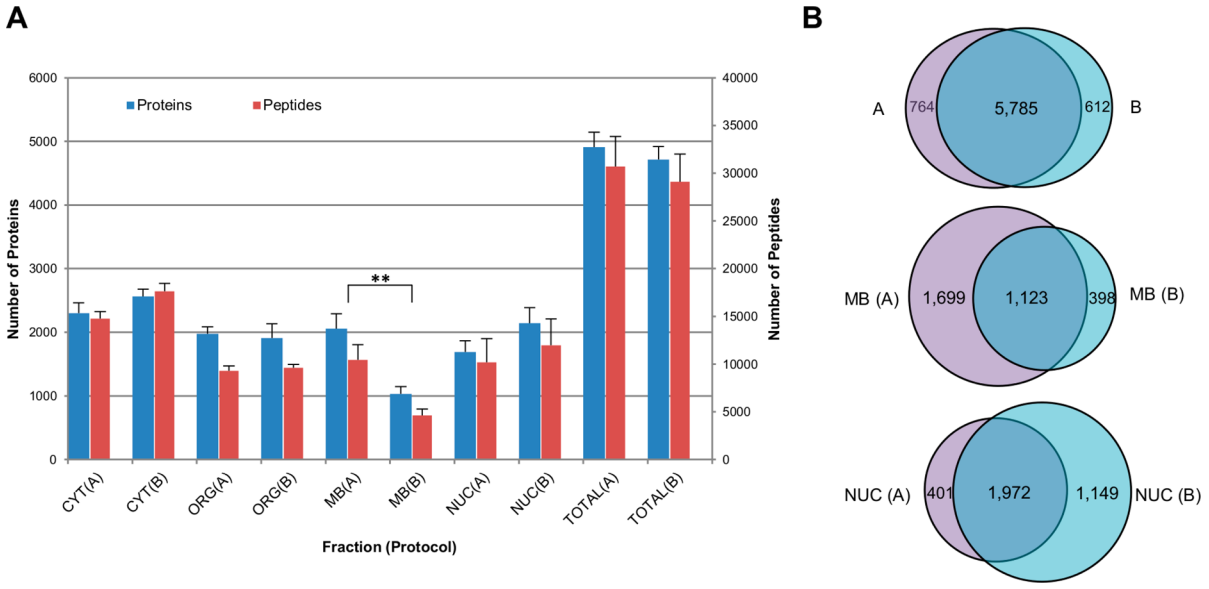


Figure 2. Comparison of protein extraction methods. (A) Total number of proteins and peptides (without duplicates) identified by LC-MS/MS assays for each subcellular fraction (CYT, cytoplasmic; ORG, organelle; MB, membrane; NUC, nuclear; and total protein extract, TOTAL) and protein extraction method (A or B). Each value was calculated as an average using the three replicates. (B) Robustness of LC-MS/MS assays for protein extraction methods A and B. Union of proteins from the three replicates (without duplicates) were considered for this comparison. Upper panel: comparison of total protein extracts. Middle and lower panels: comparisons of the membrane and nuclear proteins, respectively. ** $p < 0.01$.

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. 2015, 14, 3530−3540

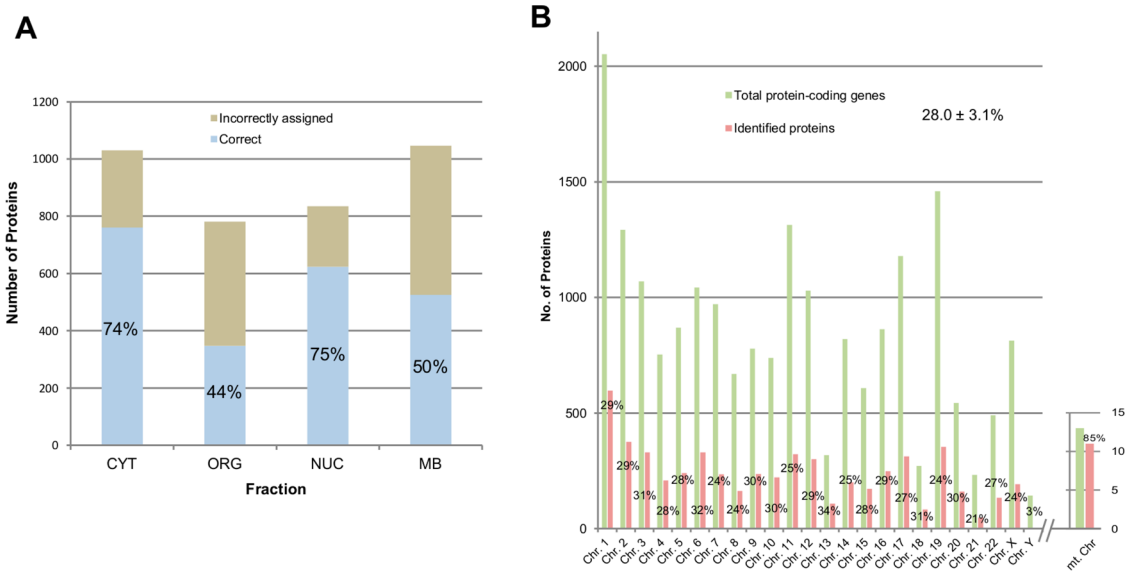


Figure 3. Subcellular protein localization. (A) Considering protein extraction method A and proteins identified in common in all three replicates (*intersection*), we validated our protein identifications for each subcellular fraction comparing with the subcellular location information given in the protein database. Correctly assigned proteins (in light blue) are the proteins whose subcellular location matches both in our MS/MS results and in the literature. Incorrectly assigned (in light brown) proteins are referred to not matching. (B) Chromosome mapping of proteins identified by at least one unique peptide (*maximum*). Green bars show the total protein-coding genes identified per chromosome (data from Ensembl release 78), red bars show the number of proteins identified by our MS/MS strategy. Corresponding percentages are noted within each bar.

compartments of *Ramos* lymphoma B-cells. The efficiency of solubilization and the maintenance of protein structure are directly dependent on the detergent choice, salt concentration, and pH. Thus, two similar approaches (methods A and B) were tested in parallel to isolate four different subcellular fractions of proteins (cytoplasmic, CYT; organelle, ORG; nuclear, NUC; and membrane, MB). These methods only differ in the extraction of membrane and nuclear proteins, which can be isolated by using IGEPAL and n-dodecyl-β-D-maltopyranoside (for method A) or NaCl and n-dodecyl-β-D-maltopyranoside (for method B), respectively. However, in both methods, the cytoplasmic and organelle fractions were similarly isolated using digitonin and Tween 20 (both are nonionic detergents), respectively. Digitonin effectively water-solubilizes membrane proteins, but not the nuclear ones which remain structurally intact. Thus, cytosolic proteins can be recovered. Second, we used Tween 20, which does not affect protein activity, to extract organelle-related proteins.

Supplementary Figure 2A displays the total amount of protein extracted from each subcellular fraction. By SDS-PAGE analysis (Supplementary Figure 2B), a homogeneous distribution of proteins independently of the molecular weight is observed.

The MS/MS data analysis of the four subcellular fractions described revealed a high number of identifications at the protein and peptide levels. In these assays, a very tight precursor mass tolerance (0.8 Da) and an FDR lower than 1% for both PSM, peptide and protein, were considered to reduce the chance of false positive identifications. The results obtained suggest that the differences in the number of identified proteins, between the protein extraction methods A and B, correlate with the detergent selected/used (Figure 2A). Specifically, the difference between the number of membrane proteins recovered with IGEPAL (method A) and n-dodecyl-β-D-maltopyranoside (method B) is statistically significant (p-value of 0.007). In case of method A, IGEPAL solubilizes membrane proteins, whereas it is not strong enough to lyse the nuclear membrane, which allows this subcellular compartment to remain intact for the effects of the n-dodecyl-β-D-maltopyranoside detergent (for extraction of nuclear proteins). In turn, in method B, the membrane fraction was isolated in the last step by using n-dodecyl-β-D-maltopyranoside.

Regarding nuclear proteins, we detected that usage of high salt concentrations (method B) was more efficient than n-dodecyl-β-D-maltopyranoside (method A) for nuclear protein extraction. In summary, these preliminary analyses indicate a slight increase in proteome coverage with method A versus method B. Therefore, the differences between both methods provide an explanation to the variations observed in protein recovery. In fact, with method A, a better recovery of membrane proteins was achieved, and we were particularly interested in the membrane fraction because it is the cellular compartment where more missing proteins are estimated to be located in this specific cell line.

For a more in-depth comparison between the two protein extraction methods, the robustness through the 3 replicates was analyzed with an overall overlap of 81%, 35%, and 56% between methods A and B for total, membrane, and nuclear proteins, respectively (Figure 2B). Of note, again method A identified a greater number of proteins than method B (up to 4 times). Therefore, on the basis of these preliminary analyses, we selected method A for further transcriptomics−proteomics comparisons owing to the higher proteome coverage not only at the total protein level but also at the membrane fraction of proteins.

Characterization of the Proteome of *Ramos* B-cells

As described above, *Ramos* B-cells derive from a Burkitt lymphoma carrying the *MYC* gene rearrangements (i.e., t(8;14)[36]), and they are often used as a model for proteomics of B lymphocytes. *MYC* gene was first discovered in Burkitt lymphoma patients as a proto-oncogene whose activation leads to the induction of cellular proliferation. Although our

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. 2015, 14, 3530−3540

proteomics strategy has not detected the Myc protein—mainly due to the fact that this protein is located on the nucleus, as it is further explained in the section entitled Exclusive Identifications from Comparison of both Proteomics and Transcriptomics)—a large number of proteins interacting with Myc have been identified (Actl6a, Bcl2, Chd8, Gtf2I, Mapk1, Max, Mlh1, Mycbp2, Mycbp, Nmi, Nyfc, Pfdn5, Ruvb, Sap130, Smad2, Smad3, Smarca4, Smarcb1, Taf9, Wdr5, Yyi, among others). Thus, this -omic platform allows the characterization of cell signaling pathways by identifying the components of the interactions leading to changes in cellular responses.

In order to evaluate the robustness of the MS/MS data set and its reliability for integration with transcriptomics data sets, the presence of cross-contamination between the subcellular fractions obtained with method A was determined according to the protein database. Thus, the percentage of the proteins identified in each subcellular fraction that were correctly isolated by the protein extraction method is presented in Figure 3A. In total, 74−75% was the proportion found for the fraction of proteins assigned by the protein database to cytoplasm (CYT) and nucleus (NUC). The lowest percentages for correct protein location corresponded to the organelle (ORG) and the membrane protein (MB) fractions. Both fractions were sequentially and consecutively extracted, and the nature of their protein mixture is highly related, hindering their correct localization. In addition, the literature-based protein database assignment can be ambiguous because many times there is more than one unique location for each protein, and additionally, there are membrane-associated organelle proteins that could be included in both fractions. Because our main goal was to profile the *Ramos* cell proteome, all proteins identified were grouped into a unique data set independently of their correct or incorrect location. In this regard, it is important to keep in mind that subcellular fractionation was specifically used to increase protein recovery, not for the independent analysis of the fractions.

On the basis of chromosome mapping (Figure 3B) of the proteins identified, overall we identified about 30% of all protein-coding genes (~5000−6000) present in the human genome (without considering the X and Y chromosomes, and mitochondrial DNA). This confirms that this approach is proteome-wide and unbiased, proving that selecting for a specific cell type is a useful approach for total proteome coverage. Interestingly, the proteins identified also covered about an 85% of the total protein-coding genes of the mitochondrial DNA, using the LC-MS/MS approach here employed. This good coverage of the mitochondrial proteins indicates a great improvement with respect to other proteomic procedures and platforms (e.g., GFP-tagging analysis[37]). This is mainly due to the development of proteomic strategies with high sensitivity and accuracy in combination with the usage of subcellular fractionation approaches that allow the enrichment of the organelle proteomes such as the mitochondrial one.[38]

Other studies have recently published proteomic data related to B-cells lymphomas by using SILAC approaches. Mann's lab has developed a super-SILAC to classify large B-cell lymphomas subtypes by their protein profiles.[39] In turn, Rüetschi and colleagues studied the same disease by a SILAC-based quantitative approach.[40] A comparison of our results revealed an overlap with 1696 and 2141 proteins identified in Mann's and Rüetschi's studies, respectively, constituting 54% and 60% of proteins in common (data not shown). The differences in identification might be due to the absence of subcellular fractionation steps in these studies that improves the recovery and protein identification. Even so, B-cell specific proteins have been detected by both approaches, demonstrating the feasibility of proteomics strategies to characterize and model the disease.

Integration of Proteomics and Transcriptomics Data Sets

The characterized MS/MS proteomic data was compared to transcriptomic measurements performed on *Ramos* B-cells (3 biological replicates) with genome-wide expression high-density oligonucleotide microarrays (Figure 4). To accomplish the proteomics-transcriptomics integration, the neXtProt IDs (corresponding to identified proteins by the LC-MS/MS approach) were mapped into the Ensembl IDs (for the genes), and these were mapped into the Affymetrix probesets IDs that identify the gene-specific DNA oligo probes which are present in the microarrays (Supporting Information 3,4 and 5). Such comparison was addressed from different ways termed *intersection*, *union*, and *maximum*, as described above in the Materials and Methods section. On the basis of this strategy, the following goals were pursued: (i) to identify as many proteins as possible using this approach present in *Ramos* B-cells (*maximum* and *union*), and (ii) to characterize those proteins that are unequivocally identified with this proteomic approach (*intersection*). Briefly, 3383, 5494, and 8931 known human proteins (i.e., neXtProt IDs) were found within the *intersection*, *union*, and *maximum* data sets respectively (Table 1). These data sets corresponded to 3433, 5540, and 8976

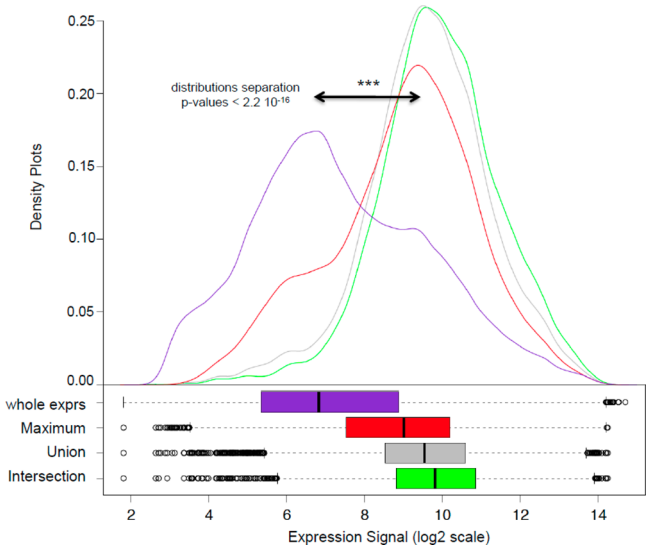


Figure 4. Density plots and boxplots showing the distributions of the whole gene expression signal versus the signal corresponding to the genes detected in the MS proteomic profiles for *Ramos* B-cells. The expression was measured with *Affymetrix* high-density oligonucleotide expression arrays type Human Gene 1.0. (Purple) Whole expression signal corresponding to all the 33 297 probesets included in the arrays [whole exprs]. (Red) 9494 probesets corresponding to the 8976 distinct human genes (ENSG IDs) and 8391 proteins (neXtProt IDs), which showed at least one unique identifying peptide in the MS proteomic assays [*maximum*]. (Gray) 6175 probesets corresponding to the 5540 distinct human genes (ENSG IDs) and 5494 proteins (neXtProt IDs), which showed at least two identifying peptides in the proteomic assays [*union*]. (Green) 4088 probesets corresponding to the 3433 distinct human genes (ENSG IDs) and 3383 proteins (neXtProt IDs), which had at least two identifying peptides in the proteomic assays and were found in all the replicates of the proteomic isolations [*intersection*].

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. 2015, 14, 3530−3540

genes (i.e., Ensembl IDs). These results indicate that proteomics and transcriptomics studies, once applied to the global molecular characterization of a cell type, display a high accuracy and overlap within each analytical level. Such overlap can be estimated considering the proportion of proteins that are detected by proteomics and transcriptomics, because—as we explain below—there were only 516 proteins exclusive of proteomics (i.e., not detected by the transcriptomic platform) out of a total of 8931, and this corresponds to a 94% of overlap of the expression data over the proteomic data.

FEA of those 3383 proteins detected in a coherent and steady manner (*intersection* data set) by the proteomics assays (Supporting Information 6) revealed the expression of many essential proteins for general cell functions and house-keeping processes (e.g., anabolism and synthesis processes, together with catabolism and cellular respiration) as well as the activity and regulation of major biological macromolecules (DNA, RNA, and proteins) involved in key maintenance processes like cell cycle, cell growth, cell proliferation, and so forth.

On the other side, mapping of proteins to tissue-type and cell-type databases provided a clear enrichment in B-cell specific genes and proteins that was in agreement with the character and properties of a lymphocytic cell type. For example, regarding B-cell receptor (BCR) cross-linking, many intracellular signaling cascades appended to be activated leading to regulation of gene expression. Moreover, synthesis initiation proteins (eIF3a/h proteins), proteins for cellular adhesion and costimulatory signaling (CD11a), accessory signal transduction components (such as CD20, CD19, CD79a or CD79b), protein tyrosine kinases (Fyn, Lyn, Syk and Btk), and proteins related to the activation of Ras signaling pathway (N-, K-, and H-Ras) were identified. In turn, proteins belonging to the B-cell receptor signaling pathways have been widely detected including those related to BCR internalization (SHP-1, CD19, Bam32), cytoskeletal rearrangements and integrin activation (Bam32, PLCγ2, DAG), transcription (Blnk, Grb2, SHP-1, Erk1/2, Raf), proteasomal degradation (PKC, Carma1, Bcl10, MALT1, IKK, NF-κB) and growth arrest and apoptosis (Akt, FoxO), among others pathways that follow BCR activation.

Exclusive Identifications from Comparison of both Proteomics and Transcriptomics

As mentioned before, it is notorious that the overlap observed between both the proteomic and the transcriptomic methodologies applied to the same cell type, although these two experimental technologies had also a complementary part. In fact, comparing the results of both approaches, each one enables to cover the failures in identification ("exclusive identifications") due to its technical limits with respect to the other strategy.

Performing the FEA for the exclusively identified proteins in proteomics (Supporting Information 7), we identified a gain of proteins related to the mitochondrial and ribosomal organelles, as well as cytoplasmic ones. This is a quite expected result because the genomic microarrays employed do not include probes for mitochondrial DNA. In addition, we could infer that our subcellular extraction method in the proteomic procedure was effective on isolating organelle proteins, as it is shown for mitochondria. A loss of identifications associated with immunoglobulins (Ig) and major histocompatibility complex (MHC) proteins was also detected in the transcriptomic data probably due to Ig gene rearrangements and hypersomatic mutations of Ig genes that hamper the design of adequate array probes for these genes.

Regarding exclusively identified proteins in transcriptomics, the functional enrichment analysis (Supporting Information 7) revealed identifications related to nuclear and DNA-binding proteins. This is probably due to the fact that isolating the nuclear fraction in the last step of the proteomics approach decreases the recovery for nuclear proteins. Upon comparing the method chosen for this study (method A) with method B (Figure 2B), it seems clear that extracting the nuclear fraction in a previous step and with another detergent improves isolation of these proteins. Therefore, it could be appropriate to previously establish the protein fraction of interest to perform the best protein extraction method for such fraction. In our case, we select method A as the overall most suitable approach combining all fractions and also because we were interested in enriching for membrane proteins since these are the most commonly lost in many proteome-wide studies. Additionally, around 300 of these transcripts exclusively identified in transcriptomics correspond to noncoding RNAs. Thus, it is obvious that their corresponding proteins have not been detected by the proteomics platform.

Missing Proteins

Since thousands of human proteins have not been detected yet (as it is noted in the last release of missing proteins from the neXtProt database), the exploration of proteome data sets has become an indispensable exercise that may be accomplished to reduce the number of existing missing proteins. With this purpose, we have performed the mapping of our results into the neXtProt database for missing proteins (current release of April 28, 2015) obtaining the results shown in Table 2. Specifically, the searching has been carried out across the three data sets generated (*maximum, union*, and *intersection*) identifying up to 370 missing proteins from our *maximum* data set (containing 8931 neXtProt identifications). As the number of neXtProt IDs decreases from one data set to another (from *maximum* to *union* and *intersection*), the number of identified missing proteins decreases as expected (from 370 to 32 and 4, respectively). However, this reduction in number involves an increase in quality because these missing proteins have been identified in 3 different biological replicates and with, at least, 2 peptides per protein. In Supporting Information 8 is shown the information related to missing proteins identified for each data set and the related data (PE group, chromosome location).

Table 2. Missing Proteins Identified Across the Three Datasets Obtained after Searching against the neXtProt Database[a]

	PE group				
data set	PE2	PE3	PE4	PE5	total
maximum	273	37	5	55	370
union	18	3	-	11	32
intersection	-	-	-	4	4

[a]Proteins from each dataset (*maximum, union*, and *intersection*) were mapped into the neXtProt missing proteins database (April 28, 2015)—used as reference of missing proteins—to identify missing proteins. The missing proteins have been classified accordingly to their protein existence (PE) level (i.e., PE2 for experimental evidence at transcript level; PE3 for protein inferred from homology; PE4 for predicted protein; and PE5 for uncertain proteins).

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. 2015, 14, 3530−3540

Comparative Analysis of the Results Obtained with neXtProt and UniProt Database Searches

In a further analysis, the effect of database search on identification was evaluated. With this purpose, we additionally performed a search against the UniProt database and compared the results (Supporting Information 9, 10, 11, 12) with those obtained with neXtProt. In general, neXtProt database search generated a greater number of identified proteins what determined, consequently, an increased number of missing protein identifications compared to UniProt (Supporting Information 13 and 14). The integration of information from different databases (Swiss-Prot, Ensembl, Human Protein Atlas, PeptideAtlas....) makes possible a better characterization of the proteomes and justifies this increase in identifications (more spectrum and peptide information in the database leads to a possible increasing in identifications). In addition, this supposes an improvement in the characterization of the specific functions of identified proteins. In Supporting Information 15 are reported the FEA of proteins exclusively identified in proteomics and transcriptomics for searches performed against neXtProt and UniProt, respectively, reporting an enriched annotation of the functions for neXtProt search.

Immunophenotypic Characterization of the *Ramos* B-cells

In order to give support and provide some external validation of the characterization of the cell type studied in our proteomic and transcriptomic studies, we use an independent cell-oriented platform to characterize the immunophenotype of the *Ramos* B-cells. With this purpose, several proteins were selected and screened by multiparametric flow cytometry (FCM) (Supplementary Figure 3) showing high expression levels of the B-cell associated antigen—CD19, CD20, CD22 and CD45—as well as the coreceptor of the B-cell receptor, CD79b. Expression of the lambda-Ig light chain was high as well, confirming the available data for the *Ramos* cell line (ATCC CRL 1596, www.atcc.org). Additionally, CD11a, an integrin involved in cellular adhesion and costimulatory signaling was evaluated showing a dim expression. As negative controls for FCM assays, T-cell marker CD3 and kappa-Ig light chain were also tested.

PERSPECTIVES AND CONCLUSIONS

Integration of proteomics and transcriptomics technologies has emerged as a potent strategy for the mapping of the biomolecules that define a cell type or a cellular state. The searches derived from both -omics methodologies are complementary, and once performed in conjunction one with the other, they allow a mutual validation and increase the total coverage of genome-wide protein-coding genes identified. With this goal in mind, the C-HPP initiative has promoted an effective combination of proteomics data into a genomic framework to achieve a full map of the human proteome that will provide a better understanding the studied biological systems.

Based on deep characterization of the proteome of the *Ramos* lymphoma B-cells by LC-MS/MS, a total of up to 6000 proteins and ~30 000 different peptides were identified. This data resulted in about 30% coverage of all human gene-coded proteins per chromosome found in the proteome of these lymphoma B-cells (such percentage being as high as 85% for the mitochondrial DNA). Integration of this proteomics data set with transcriptomics data sets (derived from high-density oligonucleotide microarrays technology) showed an 82% overlap between both technologies. Despite this, several gaps were identified with each of the two technologies. Regarding proteomics, special attention must be given to the protein extraction method that could affect correct extraction of proteins from specific subcellular compartments; moreover, proteins expressed at very low concentrations and in highly complex multimers could also have problems in being detected. For this reason, extensive fractionation of the proteome of interest might contribute to improve the resolution of proteomics via detection of thousands of proteins simultaneously, improving current knowledge about the functional and biological cellular processes via higher coverage of the proteins involved in these mechanisms.

Finally, the characterization of PTMs within the lymphoma proteome would be of great interest to model the disease. In this sense, further studies will be performed to determine the PTMs that may influence the normal response behavior of these cells.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00474.

Additional data as noted in the text (ZIP)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: jrivas@usal.es. Phone: +34 923294819. Fax: +34923294743.
*E-mail: mfuentes@usal.es. Phone: +34 923294811. Fax: +34 923294743.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We gratefully acknowledge financial support from the Carlos III Health Institute of Spain (ISCIII, FIS PI11/02114, FIS PI14/01538, and FIS PI12/00624), Fondos FEDER (EU) and Junta Castilla-León SA198A12-2. The Proteomics Unit belongs to ProteoRed, PRB2-ISCIII, supported by grant PT13/0001. P.D. and C.D. are supported by a JCYL-EDU/346/2013 Ph.D. scholarship. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium[41] via the PRIDE partner repository with the dataset identifier PXD001933. We thank Peter Horvatovich for providing us the list of missing proteins (release 2015-04-28).

ABBREVIATIONS

ACN, acetonitrile; BCR, B-cell receptor; C-HPP, Chromosome-human Proteome Project; CYT, cytoplasmic proteins; DTT, dithiothreitol; FA, formic acid; FBS, fetal bovine serum; FCM, flow cytometry; FDR, False Discovery Rate; FEA, Functional Enrichment Analysis; FT-ICR, Fourier Transform-Ion Cyclotron Resonance; HEPES, 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid; HUPO, Human Proteome Organization; IAA, iodoacetamide; Ig, immunoglobulin; IGEPAL, octylphenoxypolyethoxyethanol; LTQ, Linear Trap Quadrupole; MB, membrane proteins; MHC, major histocompatibility complex; MS, mass spectrometry; NUC, nuclear proteins; ORG, organelle proteins; protFDR, FDR at protein level; psmFDR, FDR at PSM level; PMSF, phenylmethane-

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. 2015, 14, 3530−3540

sulfonyl fluoride; PTM, post-translational modification; TCEP, tris (2-carboxyethyl) phosphine hydrochloride

REFERENCES

(1) Muñoz, J.; Heck, A. J. R. From the human genome to the human proteome. *Angew. Chem., Int. Ed.* **2014**, *53*, 10864−10866.
(2) Segura, V.; Medina-Aunon, J. A.; Mora, M. I.; Martínez-Bartolomé, S.; Abian, J.; Aloria, K.; Antúnez, O.; Arizmendi, J. M.; Azkargorta, M.; Barceló-Batllori, S.; et al. Surfing transcriptomic landscapes. A step beyond the annotation of chromosome 16 proteome. *J. Proteome Res.* **2014**, *13*, 158−172.
(3) Ansong, C.; Purvine, S. O.; Adkins, J. N.; Lipton, M. S.; Smith, R. D. Proteogenomics: Needs and roles to be filled by proteomics in genome annotation. *Briefings Funct. Genomics Proteomics* **2008**, *7*, 50−62.
(4) Jensen, O. N. Interpreting the protein language using proteomics. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 391−403.
(5) Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57−63.
(6) Toung, J. M.; Morley, M.; Li, M.; Cheung, V. G. RNA-sequence analysis of human B-cells. *Genome Res.* **2011**, *21*, 991−998.
(7) Clamp, M.; Fry, B.; Kamal, M.; Xie, X.; Cuff, J.; Lin, M. F.; Kellis, M.; Lindblad-Toh, K.; Lander, E. S. Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 19428−19433.
(8) Woo, S.; Cha, S. W.; Merrihew, G.; He, Y.; Castellana, N.; Guest, C.; Maccoss, M.; Bafna, V. Proteogenomic database construction driven from large scale RNA-seq data. *J. Proteome Res.* **2014**, *13*, 21−28.
(9) Renuse, S.; Chaerkady, R.; Pandey, A. Proteogenomics. *Proteomics* **2011**, *11*, 620−630.
(10) Nilsson, C. L.; Berven, F.; Selheim, F.; Liu, H.; Moskal, J. R.; Kroes, R. A.; Sulman, E. P.; Conrad, C. A.; Lang, F. F.; Andrén, P. E.; et al. Chromosome 19 annotations with disease speciation: A first report from the global research consortium. *J. Proteome Res.* **2013**, *12*, 135−150.
(11) Schwanhäusser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. Global quantification of mammalian gene expression control. *Nature* **2011**, *473*, 337−342.
(12) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **2014**, *11*, 1114−1125.
(13) Castellana, N.; Bafna, V. Proteogenomics to discover the full coding content of genomes: A computational perspective. *J. Proteomics* **2010**, *73*, 2124−2135.
(14) Krug, K.; Nahnsen, S.; Macek, B. Mass spectrometry at the interface of proteomics and genomics. *Mol. BioSyst.* **2011**, *7*, 284−291.
(15) Jacob, F.; Goldstein, D. R.; Fink, D.; Heinzelmann-Schwarz, V. Proteogenomic studies in epithelial ovarian cancer: established knowledge and future needs. *Biomarkers Med.* **2009**, *3*, 743−756.
(16) Haider, S.; Pal, R. Integrated analysis of transcriptomic and proteomic data. *Curr. Genomics* **2013**, *14*, 91−110.
(17) Cox, B.; Kislinger, T.; Emili, A. Integrating gene and protein expression data: Pattern analysis and profile mining. *Methods* **2005**, *35*, 303−314.
(18) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.et al. The human proteome project: current state and future direction. *Mol. Cell. Proteomics* **2011**, *10*, 10.1074/mcp.M111.009993.
(19) Paik, Y. K.; Omenn, G. S.; Uhlen, M.; Hanash, S.; Marko-Varga, G.; Aebersold, R.; Bairoch, A.; Yamamoto, T.; Legrain, P.; Lee, H. J.; et al. Standard guidelines for the chromosome-centric human proteome project. *J. Proteome Res.* **2012**, *11*, 2005−2013.
(20) Segura, V.; Medina-Aunon, J. A.; Guruceaga, E.; Gharbi, S. I.; Gonzälez-Tejedo, C.; San Chez Del Pino, M. M.; Canals, F.; Fuentes, M.; Casal, J. I.; Martínez-Bartolomé, S.; Elortza, F.; Mato, J. M.; Arizmendi, J. M.; Abian, J.; Oliveira, E.; Gil, C.; Vivanco, F.; Blanco, F.; Albar, J. P.; Corrales, F. J.; et al. Spanish human proteome project: Dissection of chromosome 16. *J. Proteome Res.* **2013**, *12*, 112−122.
(21) Shevchenko, A.; Tomas, H.; Havlis, J.; Olsen, J. V.; Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **2006**, *1*, 2856−2860.
(22) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2007**, *2*, 1896−1906.
(23) Olsen, J. V.; de Godoy, L. M. F.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **2005**, *4*, 2010−2021.
(24) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8*, 2405−2417.
(25) Prieto, G.; Aloria, K.; Osinalde, N.; Fullaondo, A.; Arizmendi, J. M.; Matthiesen, R. PAnalyzer: a software tool for protein inference in shotgun proteomics. *BMC Bioinf.* **2012**, *13*, 288.
(26) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20*, 3551−3567.
(27) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.
(28) Brosch, M.; Yu, L.; Hubbard, T.; Choudhary, J. Accurate and sensitive peptide identification with mascot percolator. *J. Proteome Res.* **2009**, *8*, 3176−3181.
(29) Carvalho, B. S.; Irizarry, R. A. A framework for oligonucleotide microarray preprocessing. *Bioinformatics* **2010**, *26*, 2363−2367.
(30) Carvalho, B. Platform design info for Affymetrix HuGene-1_0-st-v1. R package version 3.8.0.
(31) Dai, M.; Wang, P.; Boyd, A. D.; Kostov, G.; Athey, B.; Jones, E. G.; Bunney, W. E.; Myers, R. M.; Speed, T. P.; Akil, H.; Watson, S. J.; Meng, F. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **2005**, *33*, E175.
(32) Durinck, S.; Spellman, P. T.; Birney, E.; Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **2009**, *4*, 1184−1191.
(33) Dennis, G.; Sherman, B. T.; Hosack, D. A.; Yang, J.; Gao, W.; Lane, H. C.; Lempicki, R. A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **2003**, *4*, P3.
(34) Fontanillo, C.; Nogales-Cadenas, R.; Pascual-Montano, A.; de Las Rivas, J. Functional analysis beyond enrichment: Non-redundant reciprocal linkage of genes and biological terms. *PLoS One* **2011**, *6*, e24289.
(35) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80.
(36) Williams, S. C.; Winter, G. Cloning and sequencing of human immunoglobulin V lambda gene segments. *Eur. J. Immunol.* **1993**, *23*, 1456−1461.
(37) Gabaldón, T.; Huynen, M. A. Shaping the mitochondrial proteome. *Biochim. Biophys. Acta, Bioenerg.* **2004**, *1659*, 212−220.
(38) Huber, L. A.; Pfaller, K.; Vietor, I. Organelle proteomics: Implications for subcellular fractionation in proteomics. *Circ. Res.* **2003**, *92*, 962−968.
(39) Deeb, S. J.; D'Souza, R. C. J.; Cox, J.; Schmidt-Supprian, M.; Mann, M. Super-SILAC Allows Classification of Diffuse Large B-cell Lymphoma Subtypes by Their Protein Expression Profiles. *Mol. Cell. Proteomics* **2012**, *11*, 77−89.
(40) Coiffier, B.; Lepage, E.; Briere, J.; Herbrecht, R.; Tilly, H.; Bouabdallah, R.; Morel, P.; Van Den Neste, E.; Salles, G.; Gaulard, P. CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *N. Engl. J. Med.* **2002**, *346*, 235−242.

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. 2015, 14, 3530−3540

(41) Vizcaíno, J.; Deutsch, E.; Wang, R.; Csordas, F.; Reisinger, F.; et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32*, 223−226.

DOI: 10.1021/acs.jproteome.5b00474
J. Proteome Res. 2015, 14, 3530−3540

Comprehensive combination of affinity proteomics, MS/MS and RNA-Sequencing datasets for the analysis of a lymphoma B-cell line in the context of the Chromosome-Centric Human Proteome Project

Paula Díez[1,2], Conrad Droste[3], Raquel Bartolomé[1], Quentin Lécrevisse[1], Rosa M. Dégano[2], Diego Alonso-López[4], Nieves Ibarrola[2], Rafael Góngora[1], Fernando Corrales[5], Alberto Orfao[1], Fridtjof Lund-Johansen[6], Javier De Las Rivas[3] and Manuel Fuentes[1,2]

[1] Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

[2] Proteomics Unit. Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

[3] Bioinformatics and Functional Genomics Research Group, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

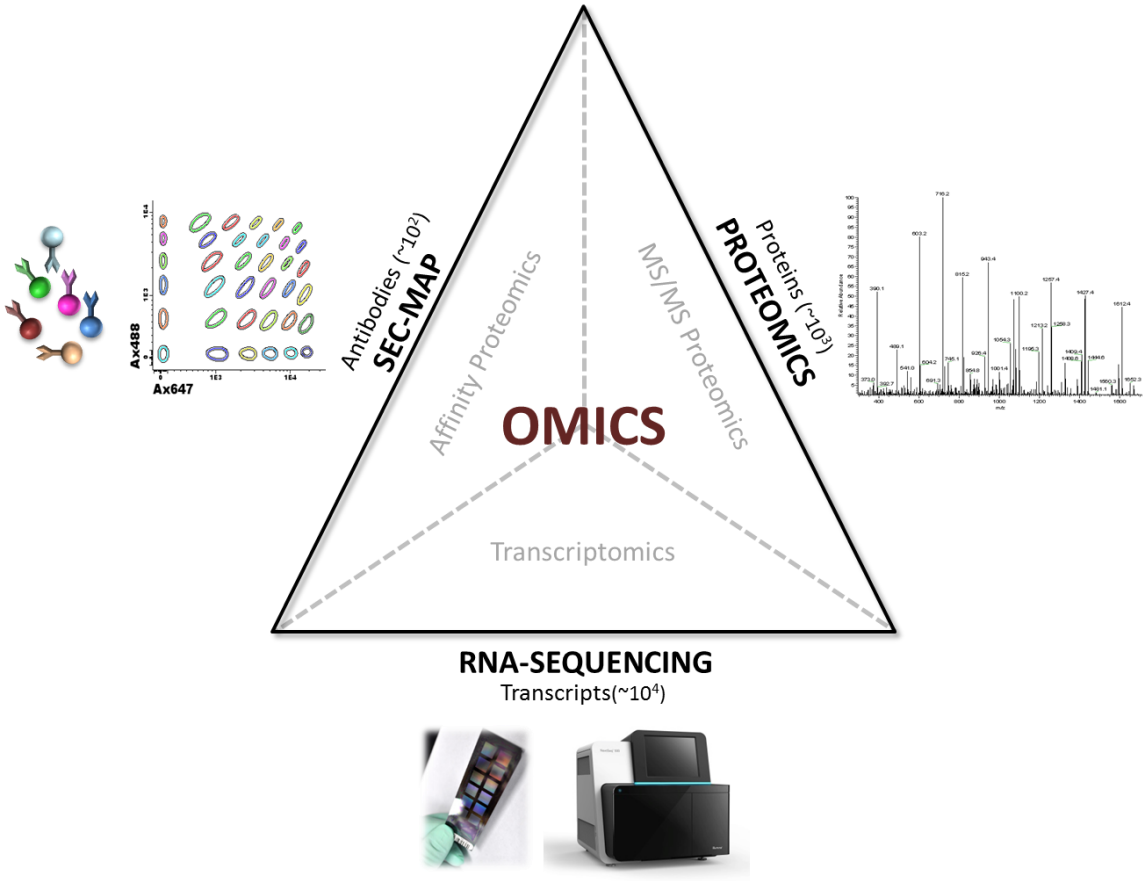
[4] Bioinformatics Unit, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

[5] Division of Hepatology and Gene Therapy, Proteomics and Bioinformatics Unit, Centre for Applied Medical Research (CIMA), University of Navarra , 31008 Pamplona, Spain.

[6] Protein Array Group, Department of Immunology, University of Oslo, 0372 Oslo, Norway

Topic: Combined analyses with mass spectrometry, RNA Sequencing, and polyclonal or monoclonal antibodies for targeted studies of tissue samples for orthogonal confirmation of tissue and organelle level expression.

Table of Content



Abstract

One of the major purposes of the Human Proteome Project (HPP) is the characterization of the proteins encoded by the genome using several complementary -omic techniques, such as proteomics, antibody-based proteomic profiling, transcriptomics and others, to achieve better coverage and identification of the still many *missing proteins* of the human proteome.

In the present study, a comprehensive combination of MS/MS proteomics, RNA-Seq transcriptomics, and antibody-based affinity proteomics (SEC-MAP) have been integrated to study a lymphoma B-cell line. Following a 4-subcellular fractionation, a systematic MS/MS proteomic analysis was performed achieving the identification of 5,672 unique proteins (neXtProt release 2016-02). A parallel analysis of the same B-cells using RNA deep sequencing provided the identification of 19,518 expressed genes and 5,707 protein coding genes (mapped to neXtProt). The proteomics/transcriptomics integration resulted in a 91% overlapping. Additionally, 37 missing proteins were identified in the MS/MS dataset (MS/MS data are available via ProteomeXchange with identifier PXD003939).

Moreover, when focusing on the 413 lymphoma relevant proteins screened by the SEC-MAP approach, a 55.9% overlapping was identified by the three mentioned techniques.

Finally, we have developed a bioinformatics pipeline to combine high-content proteomics data (MS/MS and affinity reagents) and transcriptomics as a useful integrative approach for full screening of protein profiles in the context of the chromosome-centric HPP.

Keywords: affinity-based proteomics, Human Proteome Project, microsphere, missing proteins, MS/MS, protein profiling, RNA-Sequencing, SEC-MAP, transcriptomics.

INTRODUCTION

In 2010, the Human Proteome Project (HPP) was launched with the overall aim of characterizing all the proteins encoded in the human genome and describing their role in biology and disease. A strategic plan was defined based on three technological pillars: mass spectrometry (MS), antibodies (affinity reagents), and the knowledge databases (KB) [1].

Moreover, the HPP has organized two transversal strategies on top of the pillars: chromosome-centric HPP (C-HPP) and Biology/Disease-HPP (B/D-HPP). The identification of at least one representative peptide for each protein-coding gene has been the main goal of C-HPP [2].

The integration of –omic technologies such as genomics, transcriptomics, proteomics and metabolomics offers a high-dimensional overview of the molecules present in a cell and, consequently, a view of the molecular entities in a particular cellular moment [3,4]. In fact, Genomics information offers a static and constant view of a cell state; meanwhile, Transcriptomics and Proteomics reflect a cellular dynamics in response to signals. Then, the integration of all these -omics provides an accurate readout of the cell which is crucial for understanding the cellular behavior and/or the biological environment [5,6].

Bearing this in mind, the -omics integration allows the complete characterization of a cell. In fact, an increasing number of studies have been developed in this field in the last years. For instance, Yizhak and collaborators have introduced a new method for the integration of proteomic and metabolomics data with genomics called IOMA (Integrative Omics- Metabolic Analysis) in order to predict metabolic flux distributions [7]. Also, Haider and colleagues focused their research on joining transcriptomic and proteomic data for the understanding of the regulatory behavior of cells since 8 different approaches [8]. Additionally, our recent study about profiling B-cells was performed using proteomic and genomics datasets, achieving a 94% overlap between both –omics approaches [9].

Most of these integration studies are supported by transcriptomics and proteomics methodologies, which have several features. Concerning transcriptomics, RNA-Sequencing (RNA-Seq) appears as the technology of choice for genome-wide differential gene expression studies. Compared to microarrays, RNA-Seq determines

RNA expression levels in a more accurately way. However, a great drawback of this technique is the necessity of efficient methods to process the big amount of generated data. Another big challenge is related to the targeting of complex transcriptomes to characterize rare RNA isoforms [10].

When talking about proteomics, mass spectrometry appears as the promising technology for simultaneous identification of thousands of proteins in one single assay. It measures gene products at the translational level and allows a great coverage of the proteome. However, several bottlenecks are also presented in this methodology and are quite diverse for sample preparation (eg. high salt concentrations, detergents, subcellular fractionaton…), protein digestion (i.e. highly dependent of specific proteases or combinations of them), getting enough accuracy and precision of measurements, and processing results with high-efficiency database search engines [11,12].

Recently, it has been developed a novel proteomics approach based on a technique which couples size exclusion chromatography (SEC) with microsphere-based affinity proteomics (MAP) [13–15]. SEC-MAP has been successfully employed to detect hundreds of proteins in a single sample and provide crucial information about protein size and subcellular localization. Since few amount of cells is required (~5-10 x 10^6 cells), the SEC-MAP approach could be helpful as a sensitive high-content tool for protein profiling in different cells (including lymphocytic ones) [16]. However, previous knowledge about the purpose of the study is required to select the antibodies included in the SEC-MAP array.

With this purpose, an in-house-assembled MAP array has been designed comprising a set of 576 populations of differently color-coded microspheres, each one carrying an antibody against a single human protein. These proteins are well-known to be relevant in several hematological pathologies. Additionally, a qualitative approach has been designed in order to identify which antibody/protein tandems are suitable for immunodetection and/or immunoprecipitation techniques. Furthermore, this approach has been integrated with transcriptomics and MS/MS datasets to describe protein profiles and ensures specificity of the affinity reagents.

Here, we present a multi-dimensional characterization of the protein profile of a Burkitt's lymphoma B-cell line based on the analysis of evidence at the protein level by

affinity proteomics (SEC-MAP) and MS/MS assays and its correlation with the number of gene products identified by RNA-Seq.

MATERIALS AND METHODS

Reagents

n-Dodecyl-β-D-maltopyranoside was purchased from Affymetrix (Maumee/OH, USA). Coomassie Brilliant Blue from Merck (Kenilworth/NJ, USA). Trypsin from Promega (Madison/WI, USA). Protease inhibitor cocktail; 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES); Tween 20; tris (2-carboxyethyl) phosphine hydrochloride (TCEP); phenylmethanesulfonyl fluoride (PMSF); digitonin; RPMI-1640 media; potassium ferrocyanide; sodium thiosulfate; dithiothreitol (DTT); iodoacetamide (IAA); formic acid (FA); acetonitrile (ACN); and bovine serum albumin (BSA) were purchased from Sigma (St. Louis/MO, USA). Heat-inactivated fetal bovine serum (FBS); L-glutamine; penicillin; and streptomycin were obtained from Gibco® (Scotland, UK). Biotin-PEO_4-NHS and Blocker Casein® were purchased from Thermo Fisher Scientific (Waltham/ MA, USA). Pure Goat IgG, mouse gamma globulin, and phycoerythrin (PE)-conjugated streptavidin were purchased from Jackson ImmunoResearch (West Grove/PA, USA).

The hypotonic lysis buffer used contained 30 mM HEPES pH=8, 20% (v/v) glycerol, 15 mM KCl, 1 mM EDTA, 2 mM $MgCl_2$, 1 mM PMSF, 1 mM TCEP, 1% (v/v) protease inhibitor cocktail. The hypertonic lysis buffer contains the same components as the hypotonic lysis buffer, except glycerol.

Methods

A general overview of the workflow followed is shown in Figure 1.

Cell Culture, Harvesting, and Lysis

The *Ramos* Burkitt's lymphoma-derived B-cell line (ATCC CRL 1596) was cultured at 37°C in a humidified CO_2 incubator (5% CO_2) in complete RPMI media (RPMI-1640 medium supplemented with 10% (v/v) FBS, 200 mM L-glutamine, 10,000 U/mL penicillin, and 10,000 μg/mL streptomycin).

A total of 30 x 10^6 cells were used for protein harvesting. Before starting the subcellular procedure, cells were washed twice with PBS and pelleted for 5 min at 1,000 g. Then, four sub sequentially steps were performed adding different lysis buffers (at a volume equal to 5 times that of the cell pellet). Briefly, the hypotonic lysis buffer supplemented with 0.015% (w/v) digitonin was added to pelleted cells. Then, the sample was rotated during 30 min and centrifuged at 1,500 g for 5 min to collect the cytoplasmic proteins (CYT) in the supernatant. The remaining fractions were processed stepwise in an identical manner. For organelle proteins (ORG), the hypotonic lysis buffer was supplemented with 0.5% (v/v) Tween 20; for the nuclear fraction (NUC), the hypertonic lysis buffer supplemented with 140 mM NaCl was used; and the membrane fraction (MB) was obtained after incubating for 5 min with the hypotonic lysis buffer plus 1% (w/v) n-Dodecyl-β-D-maltopyranoside. All steps were performed in triplicate and at 4˚C, and all buffers were supplemented with protease inhibitors (1%).

Protein Quantification and SDS-PAGE Separation

The Lowry-DC-Protein Assay was used for protein quantification following the manufacturer's recommendations (Bio-Rad Laboratories, CA, USA). Then, 15 μg of each protein sample (CYT, ORG, NUC, MB) were separated on a 4-20% gradient SDS-PAGE gel under reducing conditions (Figure 1A, 1B). After electrophoresis, 0.5% (w/v) Coomassie Brilliant Blue was employed to stain the gels and these were stored at 4°C in an aqueous solution containing 1% (v/v) acetic acid, until analysis.

Proteomics Analysis

Protein Digestion and LC-MS/MS Analysis

Each gel lane (containing 15 μg) was cut into 5 fragments and digested with trypsin following the method of Shevchenko et al. [17] with slight modifications. Briefly, gel pieces were destained with 15 mM potassium ferrocyanide and 50 mM sodium thiosulfate. Protein reduction and alkylation were performed with 10 mM DTT at 56ºC for 45 min, and with 55 mM IAA at room temperature (RT) for 30 min, respectively. Proteins were digested with trypsin (6.25 ng/mL) at 37ºC for 18 h. The peptide solution was acidified with FA and desalted by using C18-Stage-Tips columns [18]. The samples were partially dried and stored at -20ºC until analyzed by LC-MS/MS.

A nanoUPLC system (nanoAcquity, Waters Corp., Milford/MA, USA) coupled to an LTQ-Velos-Orbitrap mass spectrometer (Thermo Fisher Scientific, San Jose/CA, USA) via a nanoelectrospray ion source (NanoSpray flex, Proxeon, Thermo) was used for reversed-phase LC-MS/MS analysis. Peptides were dissolved in 0.5% FA/3% ACN and loaded onto a trapping column (nanoACQUITY UPLC 2G-V/M Trap Symmetry 5 µm particle size, 180 µm × 20 mm C18 column, Waters Corp., Milford/MA, USA). Peptides were separated on a nanoACQUITY UPLC BEH 1.7 µm, 130 Å, 75 µm × 250 mm C18 column (Waters Corp., Milford/MA, USA) with a linear gradient from 7% to 35% solvent B (ACN/0.1% FA) at a flow rate of 250 nL/min over 120 minutes.

The nUPLC- LTQ-Orbitrap Velos was operated in the positive ion mode by applying a data-dependent automatic switch between survey MS scan and tandem mass spectra (MS/MS) acquisition. Survey scans were acquired in the mass range of m/z 400 to 1600 with a 60,000 resolution at m/z 400 with lock mass option enabled for the 445.120025 ion [19]. The 20 most intense peaks having ≥2 charge state and above the 500 intensity threshold were selected in the ion trap for fragmentation by collision-induced dissociation with 35% normalized energy, 10 ms activation time, q = 0.25, ±2 m/z precursor isolation width and wideband activation. Maximum injection time was 1,000 ms and 50 ms for survey and MS/MS scans, respectively. AGC was 1×10^6 for MS and 5×10^3 for MS/MS scans. Dynamic exclusion was enabled for 90 s.

Database Search

Raw data were translated to mascot general file (mgf) format and searched against the neXtProt database (release 2016-02) using a target-decoy strategy. Peak lists (mgf file) obtained from MS/MS spectra were identified using Comet version 2015.01 rev. 2 [20]. The search was conducted using SearchGUI version 1.30.1 [21].

Protein identification was conducted against a concatenated target/decoy [22] version of the human complement of the neXtProt release 2016-02 (41,992 sequences). The decoy sequences were created by reversing the target sequences in SearchGUI. The identification settings were as follows: trypsin with a maximum of 2 missed cleavages; 10.0 ppm as MS1 and 0.5 Da as MS2 tolerances; fixed modifications: carbamidomethylation of cysteine (+57.021464 Da), and variable modifications: acetylation of protein n-terminus (+42.010565 Da) and oxidation of methionine (+15.994915 Da).

Peptides and proteins were inferred from the spectrum identification results using PeptideShaker version 0.41.1 [23]. Peptide Spectrum Matches (PSMs), peptides and proteins were validated at a 1.0% False Discovery Rate (FDR) estimated using the decoy hit distribution. At protein/peptide/PSM levels, the FDR (%), true positives, and false positives values for the datasets were: CYT1 (0.99, 2899, 29/0.98, 10474, 104/1.0, 17627, 178); CYT2 (1.0, 3183, 32/ 1.0, 12897, 130/ 1.0, 20147, 203); CYT3 (0.99, 2802, 28/ 0,99, 10857, 109/ 1.0, 5666, 57); MB1 (0.93, 1166, 11/ 0.97, 2651, 26/ 0.99, 1303, 13); MB2 (0.97, 1122, 11/ 0.97, 1937, 19/ 0.99, 1607, 16); MB3 (1.0, 893, 9/ 0.98, 1523, 15/ 1.0, 1291, 13); ORG1 (0.97, 2543, 25/ 0.99, 7503, 75/ 0.99, 10305, 103); ORG2 (0.99, 2294, 23/ 0.99, 6919, 69/ 0.98, 4933, 49); ORG3 (0.96, 2465, 24/ 0.98, 7045, 70/ 1.0, 10507, 106); NUC1 (0.99, 2498, 25/ 0.98, 7145, 71/ 0.97, 915, 9); NUC2 (0.96, 2577, 25/ 1.0, 8938, 90/ 1.0, 4469, 45); NUC3 (1.0, 2872, 29/ 0.99, 10937, 109/ 0.99, 4585, 46). We here acknowledge that the protein-level FDR is an estimate based on several imperfect assumptions and not all proteins surviving the threshold are "confidently identified".

The mass spectrometry data along with the identification results have been deposited to the ProteomeXchange Consortium [24] via the PRIDE partner repository [25] with the dataset identifier PXD003939 and 10.6019/PXD003939. During the review process, the data can be accessed with the following credentials upon login to the PRIDE website (http://www.ebi.ac.uk/pride/archive/login): Username: reviewer85106@ebi.ac.uk, Password: 0fVloZfQ.

Quantitative Analysis of MS/MS Datasets

Raw data was analyzed with the MaxQuant Suite v.1.5.3.30 [26], measuring its expression level through the Label-Free Quantification method MaxLFQ [27]. Quality control analysis was executed with the PTXQC package v. 0.80.1 [28] in R v.3.2.4 [29]. The ulterior analysis was performed with the Perseus framework v.1.5.3.2. Total proteins and exclusive proteins were also determined for each subcellular compartment (ORG, NUC, MB, CYT).

RNA-Sequencing Transcriptomics

The transcriptomic data corresponds to RNA-Seq for a *Ramos* B-cell line obtained with Illumina Genome Analyzer IIx with paired layout (experiment SRX105534:

http://www.ncbi.nlm.nih.gov/sra/SRX105534) taken from the study SRP00931 (http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP009316) [30] from SRA (Sequence Read Archive) database. The analysis of the gene expression in this *Ramos* B-cell line was done to calculate the values of FPKM (fragment per kilobase of exon per million fragments mapped) for each gene. To achieve this, the following steps were performed: (i) Fetching SRR387395 dataset from SRA database with SRA tools [31]; (ii) Conversion of the SRA file to paired-end fastq files with SRA tools; (iii) Trimming of the data with Trimmomatic [32]; (iv) Alignment of the reads to ENSEMBL GRCh37 genome with the program STAR [33]; (v) Processing of the alignment out with SAMtools [34] to generate a binary sequence alignment map (BAM); (vi) Using CuffLinks on the BAM files to calculate the FPKM value for each gene [35]; (vii) Mapping of ENSG_IDs to neXtProt IDs using the ID mapping table within the neXtProt database release 2016-02 (ftp://ftp.nextprot.org/pub/current_release/mapping/). After all these analyses a total of 19,518 neXtProt IDs could be mapped within the RNA-Seq dataset, out of this 9,523 neXtProt IDs had FPKM>1.

Microsphere-based Affinity Proteomics (MAP) Arrays

Size Exclusion Chromatography (SEC)

Protein samples obtained after the subcellular procedure were labeled with 1 mg/mL biotin-PEO_4-NHS for 30 min at 4ºC and sequentially filtered through a 0.22 µm Ultra-free filter (Millipore, Billerica MA, USA) by centrifugation at 12,000 g for 5 min at 4°C, and loaded onto a Superdex 200 10/300 column (GE Healthcare Life Sciences, Uppsala, Sweden). Fractionation was performed using a Äkta FPLC system (GE Healthcare Biosciences, Pittsburgh, PA, USA) at 4°C; in this step, PBS containing 0.05% Tween 20 was used as running buffer. Per each subcellular fraction, 24 size-exclusion chromatography (SEC) fractions of 0.5 mL were collected (Figure 1A, 1C). These fractions were aliquoted and immediately frozen at -80°C, until analysis.

Microsphere-based Affinity Proteomics (MAP) Arrays

Different populations of fluorescent microspheres were generated using maleimide derivatives of fluorescent dyes, as described elsewhere [15,36]. For the fluorescent labeling of the distinct microsphere populations, 4 different dyes at different concentrations and ratios were used: 6 distinct fluorescent levels of both Alexa Fluor 488 (Ax488) and Alexa Fluor 647 (Ax647), and 4 distinct concentrations of both Pacific Orange (PacO)

and Pacific Blue (PacB). This led to a total of 576 distinctively color-coded microsphere populations. Then, maleimide protein-G (Fitzgerald Industries, Acton, MA, USA) was immobilized on the microspheres surface and different capture antibodies (Supporting Information S1) were attached to 549 out of 576 color-coded microsphere populations (the remaining 27 microsphere populations were empty, without any antibody, and used as controls). Finally, all different color-coded microsphere populations were mixed together to form a microsphere suspension array, and they were frozen stored (-80°C) in small (one test) aliquots until used.

Immunoprecipitation of Biotin-Labeled Proteins with MAP Array

Size-fractionated subcellular lysates were thawed and placed into V-bottom 96-well plates (Steriling®, Newport, UK) at identical protein concentrations. Then, aliquots of premixed microsphere arrays were thawed, pelleted and washed twice with PBS containing 1% Tween 20 (PBS-T). Prior to the incubation of the microsphere array with the lysates, microspheres were resuspended in PBS with 1% casein supplemented with 40 µg/mL of pure goat IgG and mouse gamma globulin to block unspecific binding (30 min at RT with continuous mild rotation). Afterward, 10 µL of the blocked microsphere array mixture were pipetted per well. PBS-T was added to fill up the wells to a final volume of 180 µL; then, the plate was sealed and incubated overnight at 4°C in the dark under continuous rotation. The next morning, the microspheres were pelleted, washed in PBS-T and labeled with 15 µL of phycoerythrin (PE)-conjugated streptavidin (2 µg/mL dissolved in PBS with 1% BSA) for 20 min at RT on the shaker. Afterward, the labeled microspheres were washed twice and measured in a flow cytometer.

Flow Cytometric Detection and Analysis

An 8-color FACSCanto II cytometer (BD Biosciences, San Jose, CA, USA) equipped with a high throughput automated sampler (HTS) was used for automated sample acquisition. Gating of the distinct color-coded microsphere populations was performed manually using the INFINICYT™ software (v. 1.7, Cytognos S.L., Salamanca, Spain) (Figure 2). Firstly, the debris and microsphere aggregates were excluded based on the light dispersion characteristics (FSC-A/SSC-A dot-plot for debris, FSC-A/FSC-H for microsphere aggregates. Figure 2A). An Ax488-Ax647 dot-plot (Figure 2B) was used to gate the Alexa microsphere populations (n=36). Afterward, a PacB-PacO dot-plot (Figure 2C) was used for the gating of Pacific microsphere populations (16 Pacific microsphere populations per each Alexa microsphere population). In total, 576

microsphere populations were identified (36 Alexa microsphere populations x 16 Pacific microsphere populations). Each microsphere population was correlated with its specific antibody, as well as its corresponding PE-median fluorescence intensity (MFI) value obtained, which directly relates to the amount of captured protein per color-coded microsphere population (Figure 2D). This information was exported to a text format database file. Data for individual color-coded microsphere populations corresponding to the same protein analyzed in the different SEC-fractions and subcellular compartments were aligned in the database (Supporting Information S2) and visualized in a graphical format. Data were formatted to generate line plots for each antibody/microsphere population included in the microsphere suspension array.

Integration of Transcriptomics, Proteomics, and SEC-MAP Datasets

The 12 MS/MS datasets (3 replicas x 4 subcellular compartments) contain the number of peptides found for each protein. First, all these datasets were unified in one and the number of peptides found per protein was summed up. In this way, we identified all the proteins containing at least one peptide in any of the MS/MS datasets (that we called "complete mapping") and which contains 5,672 proteins. The number of peptides in each replica and the number in each cellular compartment were also calculated and summed up. This provided seven subsets (1-3 replicas and CYT, MB, ORG, NUC). The genes associated with the proteins were mapped to chromosomes with the R-package BiomaRt [37]. A brief R-script is provided as Supporting Information (Supporting Information 3) to show the details of the mapping protocol. All these data are included in Supporting Information S4.

The neXtProt IDs were used as key identifiers to merge the datasets of RNA-Seq, MS/MS, and SEC-MAP. The number of proteins explored by the three techniques is 413 (this number is determined by the SEQ-MAP technique). These proteins were used to find the corresponding subsets in the RNA-Seq and MS/MS data. The results are integrated and presented in Supporting Information S5. When a protein had several gene IDs in the RNA-Seq data, we select the corresponding gene with the highest FPKM value. A protein was considered as "expressed" or present in a dataset according to the following thresholds: (1) number of peptides≥1; (2) FPKM≥1; or (3) QAS value≥1.

Functional enrichment analysis (FEA) was done using DAVID and GeneTerm-Linker tools [38,39] in order to detect any biological functions that may be enriched in some set or

subset of proteins. The main biological databases selected to find genes with annotated enriched terms were: (i) Gene Ontology (GO) using annotations spaces GO_BP, GO_CC and GO_MF; (ii) the pathways database KEGG_PATHWAY; and (iii) the INTERPRO protein structural domain database. To generate the functional clusters in DAVID, we used classification stringency *high*.

Statistical Methods

The statistical analysis and integration of the datasets were performed with the R/Bioconductor environment. The graphical representation of the Venn diagram was done with the web tool Venny 2.0 [40]. Statistical significance was set at a *P* value of <0.01. For proteomics quantification, differentially expressed proteins were calculated using a multi-test analysis (ANOVA analysis, permutation-based FDR, FDR < 1 %, 500 randomizations).

RESULTS

Evaluation of Subcellular Fractionation Strategy on Proteome Coverage

With the aim of increasing the coverage of the *Ramos* B-cell proteome, protein extraction was performed following a strategy which allows the fractionation in 4 subcellular compartments (CYT, MB, ORG, NUC). In Supporting Figure 1, the correlation for all possible combinations of the 12 MS/MS datasets (3 replicas x 4 subcellular compartments) is depicted showing the good correlation (Pearson coefficient > 0.854) between subcellular compartments and, therefore, the high reproducibility of the MS/MS results. In the PCA plot (Supporting Figure 2) it is also shown the similarity between replicas.

Additionally, we evaluated the differential expression between subcellular compartments of proteins detected by MS/MS approach using MaxQuant quantification. These datasets showed that most of the proteins (98%, 5,565/5,672) have been identified according to the reported subcellular localization (data not shown). In this sense, 34% (1,904/5,672) proteins were differentially detected in specific subcellular compartments with a q-value<0.05.

A Qualitative Antibody Score (QAS) System for Qualitative Evaluation of the Performance of Affinity Proteomics Tools

As it was previously described, SEC-MAP approach is a one-step high-throughput affinity methodology; however, the complexity and the high amount of generated data is considered as a challenge. In this study, we designed and developed a bioinformatics approach to gain qualitative information about the protein entities detected by this high-throughput affinity methodology.

Initially, the *Ramos* B-cell dataset was generated after flow cytometry analysis considering the PE intensity values of each color-coded microsphere population (as described in Materials and methods section) in each of 24 fractions (molecular weight (MW) based) obtained from SEC analysis and for each of the 4 subcellular compartments (CYT, ORG, NUC, MB). Thus, a complex multidimensional data matrix (576 x 24 x 4=55,296 data points) is created in the analysis. This multidimensional data contain info for each color-coded microsphere population where an MAP array - constituted by 576 color-coded microsphere populations containing 549 antibodies against 413 human proteins - is incubated with each MW fraction of the 24 obtained from SEC resulting from each one of the 4 processed subcellular compartments. Then, it is possible to obtain info from each protein based on the affinity recognition given by the antibody about the subcellular localization and the multimeric status (monomer *vs* multimeric/protein complex).

As a consequence, expert interpretation of the detected proteins on the depicted line plot (Figure 2D, as example) remains the most time-consuming part of the analysis. Thus, in order to distinguish between detected and non-detected proteins, a Qualitative Antibody Score (QAS) System was established based on a signal peak detector tool (similar to the ones described for the elution profiles in chromatography). By QAS, it is possible to select the protein entities which have been detected by this approach based on the antibody reliability and discard the antibodies with poor or low performance (due to cross-reactivity, low affinity/binding capacity, and epitope recognition, among others).

According to the peak detector tool, a peak was set considering a subset of SEC fractions, specifically five consecutive fractions comprising two fractions with increasing intensity values, a maximum, and two fractions with decreasing PE-intensity values. For the scoring of antibodies present in the array (QAS), the following criteria

(Table 1) were set: 1) the signal increment between the minimum and the maximum value of the peak should be ≥140% was positively scored; 2) the match between the obtained SEC elution profile and the predicted one based on the MW of the target protein (reported in neXtProt release 2016-02) was positively scored; 3) the match between detected subcellular localization and the predicted one based on the subcellular distribution -as reported in the neXtProt database - were positively scored; 4) achieving both criteria 2 and 3 is positively scored; 5) by contrast, the presence of proteolytic forms across the four latest SEC fractions (20-24) and lower the expected MW scored negatively; and 6) an extra positive score was given when similar peaks were detected with ≥2 different protein-specific antibody clones presented in the MAP array.

Considering criterion 2, in order to determine the approximate MW across every single SEC fraction, eight commercial protein standards (varying from 669 kDa to 6.5 kDa) were run (in the same conditions) on an FPLC system (Gel filtration HMW and LMW Calibration Kits, GE Healthcare Life Sciences). The elution profile of each standard protein was used to calculate the distribution coefficient (Kav), as described by Irvine et al [41], and a linear regression model was generated to estimate the observed MW of the proteins detected as a reactivity peak in each data set (Supporting Information S6). Then, the coincidence with the expected (calculated from the reported MW in the neXtProt database) and observed SEC fraction was evaluated, with a tolerance of ±1 SEC-fraction (~20% the MW) (Supporting Figure 3).

Based on the above-described QAS System, 89.8% color-coded microsphere populations (493/549) presented a QAS value ≥1 in the *Ramos* dataset and, therefore, were considered as microsphere populations containing an effective antibody for the target protein (as an acceptable antibody performance in the assay). For functional analysis, the antibody-microsphere populations with a QAS value < 1 were removed from the analysis as they were lacking in specificity for the detection of the target protein (Supporting Information S5).

Additionally, considering protein size profiles, we have been able to discriminate interacting *vs* non-interacting proteins and proteolytically degraded proteins in the samples from those with an uncleaved pattern. Moreover, some antibodies allowed the detection of the protein in more than one state (i.e. monomeric and multimeric, monomeric, and proteolytic). Of note, around 31.0% (170/549) proteins present

monomeric forms, whereas multimeric forms were detected for the majority of these proteins (52.8%, 290/549). In turn, some antibodies (25.7%, 141/549) were also associated with elution peaks corresponding to proteolytic forms, below the reported MW (Supporting Figure 3).

Moreover, as *Ramos* B-cells samples for SEC-MAP approach were prepared likewise for MS/MS, we evaluated the identification level within subcellular compartments between both proteomics approaches. After the SEC-MAP analysis (see below), we selected the differentially expressed proteins (92/413 proteins present in the SEC-MAP array) from the MS/MS datasets and compared their subcellular localization information (Supporting Information S4) with the SEC-MAP corresponding one (Supporting Information S5). The results show that NUC and ORG localizations are equally identified by both techniques (76 % and 83 %, respectively), whereas for CYT and MB the likeness is lower (38 % and 44 %, respectively).

Further analysis of these SEC-MAP datasets could reveal particular protein profiles of Burkitt's lymphoma B-cell line and could help to establish differential protein profiles between pathological cells and their healthy counterpart.

Combined Analysis of MS/MS, RNA-Seq and Affinity Proteomics (SEC-MAP) Datasets for a Lymphoma B-cell Line

The characterized MS/MS and SEC-MAP proteomics data sets were compared with transcriptomics measurements performed on *Ramos* B-cells with RNA-Seq (Figure 3). To accomplish the -omics integration, the neXtProt IDs (corresponding to identified proteins by the MS/MS approach) were mapped into ENSG IDs (for the genes) identified by RNA-Seq approach. Then, 5,672 unique proteins (neXtProt IDs) were identified in the 3 technical replicates of *Ramos* B-cell datasets (Supporting Information S4). In the same way, RNA-Seq analysis allows the detection of 20,533 genes (ENSG IDs) corresponding to 19,518 neXtProt IDS. In MS/MS proteomics we identified 5,707 proteins (neXtProts IDs) which refer to 5,982 genes (ENSG IDs). When considering genes with an FPKM value ≥1, 5,157 out of these 5,672 mapped proteins were identified. Thus, the combination of both datasets, once applied to the global molecular characterization of *Ramos* B-cell, displays a correlation of 91 % with high accuracy and great overlap within each analytical level. These mappings are depicted in Table 2 and Supporting Information S7.

Next, the integration of RNA-Seq, MS/MS proteomics, and SEC-MAP was performed focusing on the set of 413 proteins present in the SEC-MAP array (Figure 4). The content of SEC-MAP was selected according to previously reported knowledge about B-cells and lymphomas. Supporting Information S5 stores all the information concerning the results from the three methodological approaches. On the basis of the combination of these datasets, it seems that the highest reliable detection of proteins corresponds to proteins presenting at least 1 gene with FPKM ≥1; concerning MS/MS proteomics with at least ≥1 unique peptide for the target proteins; finally, proteins detected by SEC-MAP with a QAS value≥1. Once these conditions were established, the analysis revealed a 56% (231/413) overlapping between the 3 approaches (eg. CD79A, AIFM1, B2M, BAX, HRAS, PML), which are proteins related to B-cell receptor (BCR) crosslinking and linked intracellular signals, such as RAS activation pathways, and MYC interaction pathways, among others.

Specifically, MS/MS proteomics identified 65.6% (271/413) proteins present in the SEC-MAP array; RNA-Seq identified 80.6% (333/413) proteins, and SEC-MAP 91.0% (376/413) proteins (Figure 5). It is also remarkable that 8 proteins (JUN, CD44, CALB2, IL3RA, CTBP2, SEPT5, CDC14A, and MAPRE3) were undetectable by the 3 approaches.

An FEA of the 231 proteins detected by the three approaches (Supporting Information S8) depicted enrichment in leukocyte activity and regulatory proteins, among others. These results were expected in agreement with the character and properties of the lymphocytic cell type and array content.

Bridging the Gap between RNA-Seq, MS/MS Proteomics, and SEC-MAP

Bearing in mind the integration of these approaches, one of the pursued goals is the detection of low abundant transcript expression values (FKPM < 1) by MS/MS and SEC-MAP. This analysis reveals that 3.6 % (15/413) proteins (with very low transcripts levels) are accurately detected by SEC-MAP and MS/MS, which could be due to the enrichment achieved by affinity reagents and subcellular fractionation. Of course, these identified proteins require a further validation due to lack of previous evidence at the transcript level since the RNA-Seq technique is only a snapshot of the transcriptomic level at one specific point.

As it has been previously described, although the overlapping between proteomics and transcriptomics was high, there is a significant variation when focusing on specific proteins. Thus, the integration of new –omics strategies (as SEC-MAP) implies the increasing of the knowledge about a specific cell or disease. In this study, it has been depicted that 13.8% proteins (57/413) were only detected by the SEC-MAP approach against the 1% (4/413) only detected by RNA-Seq and 6.1% (25/413) detected by RNA-Seq and MS/MS proteomics.

A FEA was done with the 57 proteins exclusively identified by the SEC-MAP array (Supporting Information S8) showing enrichment in functions related to membrane proteins, including receptors, plasma membranes, cell adhesion molecules and transmembrane proteins. These results suggest that immune-affinity allows the specific and selective identification of proteins which could be difficult to detect by other approaches due to several methodological challenges, such as isolation of subcellular compartment, suitable proteotypic enzymes, or relative low abundance.

On the other side, 29 antibodies (eg. CD22, AUKA, CASP3, TRAF1….) did not efficiently work or were totally ineffective (QAS value<1) for detecting proteins by SEC-MAP; even testing more than one antibody clone against the same protein. Most of these 29 proteins were accurately detected by both MS/MS proteomics and RNA-Seq. The FEA (Supporting Information S8) revealed that functions including lumen and nuclear proteins, phosphorylated intracellular signaling proteins, and GTPase regulation was undetectable by SEC-MAP. These results could be expected because highly modified proteins (including post-translational modifications, PTMs) will require highly specific antibodies; in addition, specific organelle/membrane isolation protocols will be required for detection of proteins exclusively located in these subcellular compartments. Besides that, most of the antibodies contained in this SEC-MAP array have been validated for cell surface staining of viable cells (which required antibodies that have been developed against accessible epitopes of proteins in the native state, instead of denaturing conditions).

SEC-MAP Approach as a Strategy for Increasing Protein Evidence Coverage

Although MS/MS proteomics and RNA-Seq techniques allow the identification of thousands of gene products in a simultaneous manner, SEC-MAP appears as a promising approach because it allows the accurate detection of low abundance proteins,

providing protein evidence for each gene with low transcript expression level. In addition, SEC-MAP could provide information about subcellular localization, MW and distribution (monomer vs multimeric/protein complexes) which facilitates the description of tissue or cellular protein profiles. Since SEC-MAP is based on affinity reagents; hence, the reliability of this approach might be limited on the quality of the antibodies employed for the construction of the array. Having this in mind, antibody content could be translated into a highly cost and time-consuming approach.

In Figure 6, several representative SEC-MAP results are displayed in graphic format. Typically, SEC-MAP datasets are represented by the PE intensity (for each color-coded microsphere coupled with a specific antibody for a target protein) in each of 24 fractions (MW based) obtained from SEC analysis and for each of the 4 subcellular compartments (CYT, MB, ORG, NUC) (as it described in Materials and Methods section).

Figure 6A depicts profiles for proteins detected by the three approaches combined in this study. Generally, detected peaks correspond to expected SEC fractions, whereas other times reflect the localization of multimeric or degraded protein forms. For instance, peaks in SEC fraction #20 for HRAS, in SEC fraction #15 for AIFM1, and in SEC fractions #22-23 for B2M perfectly matched to the reported information in neXtProt. However, more peaks were identified for these proteins confirming the presence of multimeric or proteolytic forms. Complementary, profiles for the same protein (eg. BAX in Figure 6A) but from different antibody clones could be represented, allowing the confirmation of protein detection by direct comparison across several profiles.

Besides, Figure 6B shows a few representative proteins which have been only detected by SEC-MAP approach and not by MS/MS and RNA-Seq, being examples of the reliable detection of proteins in complex biological samples according to molecular size and subcellular localization.

On the other hand, in Figure 6C is depicted an example for ineffective detection of CD22. In this case, the profile pattern was totally different and is lacking any correlation with reported information for the protein; being considered a proof which directly reflect the inefficiency of the antibody to recognize the protein. By using the QAS, this qualitative observation was corroborated (QAS value = 0).

In addition, SEC-MAP allows the identification of immunoglobulins (Ig) and major histocompatibility complex (MHC) associated proteins which could not be easy to detect by other approaches due to Ig gene rearrangements and hypersomatic mutations.

In summary, the SEC-MAP performance is easily integrated with other "-omics" approaches and a useful approach for protein profiling in the context of C-HPP. Of course, several aspects could be improved such as optimizing QAS system, getting quantitative data for establishing differential protein profiles, sample preparation, or PTMs information, among others. Nonetheless, it seems a promising methodology and easy to combine with RNA-Seq and LC-MS/MS datasets, being a high-quality information source.

Missing Proteins

Since thousands of human proteins have not been detected yet (as recently reported by neXtProt release 2016-02), called missing proteins; then, the exploration of the proteome datasets is an indispensable exercise that may be accomplished to reduce the number of missing proteins. These proteins are not easy to detect by MS/MS approaches since their characteristics or abundance in samples is difficult to overcome. In this study, we have mapped our results into the neXtProt database (release 2016-02) identifying 37 *missing proteins* (27 belonging to protein existence (P.E.) level 2 (experimental evidence at transcript level), 9 in P.E.=3 (proteins inferred from homology), and 1 in P.E.=4 (predicted proteins). 19 out of the 37 missing proteins were identified with ≥2 peptides). *Missing proteins* belonging to chromosomes 16 present special interest since our HPP Consortium is focused on these chromosomes. Thus, we highlight EIF3CL and NOMO3 proteins from chromosome 16. These results are collected in Supporting Information S4.

CONCLUSIONS

Omic- technologies have the characteristic of generating vast amounts of data which are of high usefulness for increasing the cellular knowledge but constitute a great challenge when dealing with them. However, the integration of bioinformatics with these –omics strategies seems a necessary and promising approach to increase the knowledge in the field.

Integrating complementary methodologies allow a better characterization and profiling of cells and diseases. In this sense, we have combined three different approaches (MS/MS proteomics, RNA-Seq transcriptomics, and SEC-MAP as an example of affinity proteomics) to provide a full characterization of protein profile for a pathological Burkitt lymphoma B-cell line in the context of the C-HPP.

The addition of transcriptomics data in proteomics studies has proven beneficial in increasing the number of detected proteins either by providing evidence at the transcript level itself or by the selection of a suitable sample for protein detection based on transcript level expression.

Reports about the integration of proteomics and transcriptomics datasets are becoming more frequent; however, including other approaches (eg. affinity proteomics) is still a challenge according to mentioned above. In this study, the SEC-MAP approach has demonstrated to be useful, reproducible and accurate high-content proteomics tool for integration with other -omics datasets and for deciphering differential proteomics profiles in pathological situations within the framework of the Human Proteome Project.

ASSOCIATED CONTENT

Supporting Information S1. List of antibodies included in the microsphere array (SEC-MAP). A total of 549 antibodies were included in the microsphere array (SEC-MAP) out of 576 microsphere populations (576-549 = 27 empty microspheres). The table includes information about the ID for each antibody; the gene symbol, neXtProt ID and protein name for each protein targeted by the antibody; and the clone, isotype, source, and host of each antibody included in the array. The ID includes information about the gene symbol, molecular weight (value in brackets) and subcellular localization (value in parentheses. C, cytoplasm; O, organelle; M, membrane; N, nucleus).

Supporting Information S2. SEC-MAP results. The total of 549 antibodies included in the array and related information are depicted in the table. This file contains information about the SEC-MAP ID (name given to the antibody in the microsphere array); the gene symbol; the neXtProt ID corresponding to the target protein; the results concerning the 24-SEC fractions of each subcellular compartment (i.e. cytoplasm, organelle, nucleus, membrane) in which "1" means presence of peak comprising two fractions with increasing PE-intensity values, a maximum, and two fractions with decreasing PE-intensity values and"0" means absence of peak; number of total identified peaks per subcellular compartment; information about the localization of the target protein based on the literature; points given according to the QAS criteria; and total QAS value. *SEC-MAP*, Size-Exclusion Chromatography- Microsphere-based Affinity Proteomics; *PE*, phycoerythrin; *QAS*, Qualitative Antibody Score.

Supporting Information S3. R-script for integration proteomics and transcriptomics data.

Supporting Information S4. MS/MS results. The table stores the experimental data of the 5,672 proteins with at least one peptide in the MS/MS experimental replicas (complete mapping) and the number of FPKM found in the processed RNA-Seq data set SRX105534 for each corresponding gene. The identifiers of these 5,672 proteins are "neXtProt IDs", the corresponding 5,946 "Ensembl IDs" for the genes and their "HGNC Symbol", "Gene Symbol", "Gene Name" and "TSS ID" (Tissue Source Site identifier). The MS/MS experimental data is separated in different subsets. The first column ("# Validated peptides, complete mapping") unifies all experimental data sets and sums up each peptide found in the three replicas (complete mapping). The columns

"NUC, ORG, CYT, and MB" store the number of validated peptides found in the subcellular compartments nucleus (NUC), organelle (ORG), cytoplasm (CYT), and membrane (MB). The "PE" value shows the protein existence group to which each protein belongs. The column "MW (kDa)" contains the molecular weight of the protein. The column "Chromosome" stores the number of the chromosome where the gene is located. Concerning RNA-Sequencing, the column "FPKM" contains the FPKM value of each gene corresponding to each protein and the column "$\log_2$ (FPKM+1)" shows its mathematical transformation. The column "FPKM status" shows if the FPKM data of the gene is good and trustworthy. Also, missing proteins are collected in this Supporting Information.

Supporting Information S5. The table contains the information of the 413 proteins selected in the SEC-MAP experiments. Their identifiers are neXtProt IDs, their corresponding Ensembl IDs, Gene Symbol and Gene Name. The table shows the number of validated peptides found in the MS/MS experiments regarding the unification of the results of the three experimental replicas (complete mapping) and the number of peptides found in each subcellular compartment (NUC, ORG, CYT, MB). It also contains the FPKM values found in the processed RNA-Seq data set SRX105534 for each corresponding gene. Furthermore, it stores the QAS value (up to 5 for each protein) which indicate the quality of the antibodies in the SEC-MAP for each protein. The "Highest QAS value" column shows the highest QAS value for each protein if more than 1 antibody is present. The final column in the table shows in which experimental data set a protein is expressed ("+" means detection, "-" means non-detection). The second sheet shows the distribution of the expression of the proteins in regard to its expression in each experimental data set. It is the numerical data table for Figure 5. It shows as well the Venn diagram depicted in Figure 5. The third sheet in the file shows a data subset containing the proteins which have at least one corresponding gene with FPKM≥1 (333 ENSG). The fourth data sheet contains the list of unique neXtProt IDs (413 proteins).

Supporting Information S6. Linear regression model used to calculate the (approximated) molecular weight (MW) of monomeric proteins and protein complexes based on their elution profile in Size Exclusion Chromatography (SEC) fractions. **A.** List of protein standards used for calibration purposes, their characteristics and the experimental elution volumes. **B.** The Log_{10} value of the MW for each standard was

plotted against the distribution coefficient (K_{av}) calculated from the indicated formula, where V_e is the elution volume for each protein standard, V_0 is the void volume (8 mL) as determined by Blue Dextran, and V_c is the geometrical column volume (24 mL) from specifications of the column used.

Supporting Information S7. RNA-Seq transcriptomics and MS/MS proteomics mapping. The table includes the neXtProt ID, ENSG ID, Gene Symbol, TSS ID, FPKM, and FPKM status.

Supporting Information S8. Functional enrichment analysis of the set of proteins from *Ramos* B-cells exclusively identified in the MS/MS proteomics and RNA-Seq analyses (SEC-MAP-/MSMS+/RNA-Seq+, 29 proteins) (1st sheet); in the SEC-MAP analysis (SEC-MAP+/MSMS-/RNA-Seq-, 57 proteins) (2nd sheet); and in the 3 strategies (SEC-MAP+/MSMS+/RNA-Seq+, 231 proteins) (3rd sheet) done using DAVID FAC tool. Proteins were mapped to GO, KEGG, SwissProt, InterPro and Pfam databases. Each biological term enriched is shown in a row indicating the number of proteins assigned to such term and the corresponding number of proteins in the population to calculate the p-value enrichment using a hypergeometric test.

Supporting Figure 1. Multi-scatter plot depicting the relationship between the analysis results of the different subcellular samples (cytoplasm, CYT; membrane, MB; nucleus, NUC; organelle, ORG) and their replicas (1-3). The present values (in blue) correspond to the Pearson correlation coefficients.

Supporting Figure 2. PCA plot depicting the relationship between the analysis results of the different subcellular compartments (CYT, MB, ORG, NUC) via the two first principal components, PC1 and PC2. Each subcellular compartment is represented by three replicas. (+, red) CYT, cytoplasm, (●, dark blue) MB, membrane, (◆, light blue) NUC, nucleus and (■, green) ORG, organelle.

Supporting Figure 3. Molecular weight (MW) distribution. The $\log_{10}$ MW from expected results (considering the literature) was represented against the $\log_{10}$ observed MW in the SEC-MAP analysis. The dataset was classified according to their monomer/multimeric condition and the proteolysis state.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: jrivas@usal.es. Phone: +34 923294819. Fax: +34923294743

*E-mail: mfuentes@usal.es. Phone: +34 923294811. Fax: +34923294743

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We gratefully acknowledge financial support from the Spanish Health Institute Carlos III (ISCIII) for the grants: FIS PI11/02114, FIS PI114/01538, FIS PI12/00624, and FIS PI15/00328. We also acknowledge Fondos FEDER (EU) and Junta Castilla-León (grant SA198A12-2). The Proteomics Unit belongs to ProteoRed, PRB2-ISCIII, supported by grant PT13/0001. P.D. and C.D. are supported by a JCYL-EDU/346/2013 Ph.D. scholarship. We thank Rodrigo García Valiente for his support in the MaxQuant software.

ABBREVIATIONS

ACN, acetonitrile; Ax488, Alexa-488; Ax647, Alexa-647; BCR, B-cell receptor; B/D-HPP, biology/disease-Human Proteome Project; BSA, bovine serum albumin; C-HPP, Chromosome-centric Human Proteome Project; CYT, cytoplasm; DTT, dithiothreitol; ENSG ID, Ensembl ID; FA, formic acid; FBS, fetal bovine serum; FDR, False Discovery Rate; FEA, functional enrichment analysis; FPKM, fragment per kilobase of exon per million; GO, Gene Ontology; HEPES, 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid; HPP, Human Proteome Project; HTS, high-throughput automatic sampler; IAA, iodoacetamide; Ig, immunoglobulin; Kav, distribution coefficient; MAP, microsphere-based affinity proteomics; MB, membrane; MFI, median fluorescence intensity; MHC, major histocompatibility complex; MS, mass spectrometry; MW, molecular weight; NUC, nucleus; ORG, organelle; PacB, Pacific

Blue; PacO, Pacific Orange; PE, phycoerythrin; P.E., protein existence; PMSF, phenylmethanesulfonyl fluoride; PSM, Peptide Spectrum Matches; PTM, post-translational modification; QAS, Qualitative Antibody Score; RNA-Seq, RNA-Sequencing; RT, room temperature; SEC, size-exclusion chromatography; SEC-MAP, size-exclusion chromatography-microsphere-based affinity proteomics; TCEP, tris (2-carboxyethyl) phosphine hydrochloride.

REFERENCES

(1) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; et al. The human proteome project: current state and future direction. *Mol. Cell. Proteomics* **2011**, *10*, M111.009993.

(2) Aebersold, R.; Bader, G. D.; Edwards, A. M.; Van Eyk, J. E.; Kussmann, M.; Qin, J.; Omenn, G. S. The biology/disease-driven human proteome project (B/D-HPP): Enabling protein research for the life sciences community. *J. Proteome Res.* **2013**, *12*, 23–27.

(3) Ritchie, M. D.; Holzinger, E. R.; Li, R.; Pendergrass, S. A.; Kim, D. Methods of integrating data to uncover genotype–phenotype interactions. *Nat. Rev. Genet.* **2015**, *16*, 85–97.

(4) Hegde, P. S.; White, I. R.; Debouck, C. Interplay of transcriptomics and proteomics. *Curr. Opin. Biotechnol.* **2003**, *14*, 647–651.

(5) Segura, V.; Medina-Aunon, J. A.; Mora, M. I.; Martínez-Bartolomé, S.; Abian, J.; Aloria, K.; Antúnez, O.; Arizmendi, J. M.; Azkargorta, M.; Barceló-Batllori, S.; et al. Surfing transcriptomic landscapes. A step beyond the annotation of chromosome 16 proteome. *J. Proteome Res.* **2014**, *13*, 158–172.

(6) Nilsson, C. L.; Berven, F.; Selheim, F.; Liu, H.; Moskal, J. R.; Kroes, R. A.; Sulman, E. P.; Conrad, C. A.; Lang, F. F.; Andrén, P. E.; et al. Chromosome 19 annotations with disease speciation: A first report from the global research consortium. *J. Proteome Res.* **2013**, *12*, 135–150.

(7) Yizhak, K.; Benyamini, T.; Liebermeister, W.; Ruppin, E.; Shlomi, T. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* **2010**, *26*, i255–i260.

(8) Haider, S.; Pal, R. Integrated analysis of transcriptomic and proteomic data. *Curr. Genomics* **2013**, *14*, 91–110.

(9) Diez, P.; Droste, C.; Degano, R. M.; Gonzalez-Munoz, M.; Ibarrola, N.; Perez-Andres, M.; Garin-Muga, A.; Segura, V.; Marko-Varga, G.; LaBaer, J.; et al. Integration of Proteomics and Transcriptomics Data Sets for the Analysis of a Lymphoma B-Cell Line in the Context of the Chromosome-Centric Human Proteome Project. *J. Proteome Res.* **2015**, *14*, 3530–3540.

(10) Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63.

(11) Feist, P.; Hummon, A. Proteomic Challenges: Sample Preparation Techniques for Microgram-Quantity Protein Analysis from Biological Samples. *Int. J. Mol. Sci.* **2015**, *16*, 3537–3563.

(12) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422*, 198–207.

(13) Stuchlý, J.; Kanderová, V.; Fišer, K.; Cerná, D.; Holm, A.; Wu, W.; Hrušák, O.; Lund-Johansen, F.; Kalina, T. An automated analysis of highly complex flow cytometry-based proteomic data. *Cytometry. A* **2012**, *81*, 120–129.

(14) Holm, A.; Wu, W.; Lund-Johansen, F. Antibody array analysis of labelled proteomes: How should we control specificity? *N. Biotechnol.* **2012**, *29*, 578–585.

(15) Wu, W.; Slåstad, H.; de la Rosa Carrillo, D.; Frey, T.; Tjønnfjord, G.; Boretti, E.; Aasheim, H.-C.; Horejsi, V.; Lund-Johansen, F. Antibody array analysis with label-based detection and resolution of protein size. *Mol. Cell. Proteomics* **2009**, *8*, 245–257.

(16) Kanderova, V.; Kuzilkova, D.; Stuchly, J.; Vaskova, M.; Brdicka, T.; Fiser, K.; Hrusak, O.; Lund-Johansen, F.; Kalina, T. High-resolution Antibody Array

Analysis of Childhood Acute Leukemia Cells. *Mol. Cell. Proteomics* **2016**, *15*, 1246–1261.

(17) Shevchenko, A.; Tomas, H.; Havlis, J.; Olsen, J. V; Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **2006**, *1*, 2856–2860.

(18) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2007**, *2*, 1896–1906.

(19) Olsen, J. V; de Godoy, L. M. F.; Li, G.; Macek, B.; Mortensen, P.; Pesch, R.; Makarov, A.; Lange, O.; Horning, S.; Mann, M. Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol. Cell. Proteomics* **2005**, *4*, 2010–2021.

(20) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13*, 22–24.

(21) Vaudel, M.; Barsnes, H.; Berven, F. S.; Sickmann, A.; Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **2011**, *11*, 996–999.

(22) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* **2010**, *604*, 55–71.

(23) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **2015**, *33*, 22–24.

(24) Vizcaíno, J.; Deutsch, E.; Wang, R. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32*, 223–226.

(25) Martens, L.; Hermjakob, H.; Jones, P.; Adamsk, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. PRIDE: The proteomics identifications database. *Proteomics* **2005**, *5*, 3537–3545.

(26) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates,

individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372.

(27) Cox, J.; Hein, M. Y.; Luber, C. a; Paron, I. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **2014**, *13*, 2513–2526.

(28) Bielow, C.; Mastrobuoni, G.; Kempa, S. Proteomics Quality Control : Quality Control Software for MaxQuant Results. *J. Proteome Res.* **2015**, *15*, 777-787.

(29) R Development Core Team, R. *R:* A Language and Environment for Statistical Computing. **2011**.

(30) Schmitz, R.; Young, R. M.; Ceribelli, M.; Jhavar, S.; Xiao, W.; Zhang, M.; Wright, G.; Shaffer, A. L.; Hodson, D. J.; Buras, E.; et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. *Nature* **2012**, *490*, 116–120.

(31) Leinonen, R.; Sugawara, H.; Shumway, M. The sequence read archive. *Nucleic Acids Res.* **2011**, *39*.

(32) Bolger, A. M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120.

(33) Dobin, A.; Davis, C. A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T. R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21.

(34) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079.

(35) Trapnell, C.; Williams, B. a; Pertea, G.; Mortazavi, A.; Kwan, G.; van Baren, M. J.; Salzberg, S. L.; Wold, B. J.; Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **2010**, *28*, 511–515.

(36) Slaastad, H.; Wu, W.; Goullart, L.; Kanderova, V.; Tjønnfjord, G.; Stuchly, J.; Kalina, T.; Holm, A.; Lund-Johansen, F. Multiplexed immuno-precipitation with

1725 commercially available antibodies to cellular proteins. *Proteomics* **2011**, *11*, 4578–4582.

(37) Durinck, S.; Spellman, P. T.; Birney, E.; Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **2009**, *4*, 1184–1191.

(38) Dennis, G.; Sherman, B. T.; Hosack, D. A.; Yang, J.; Gao, W.; Lane, H. C.; Lempicki, R. A. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **2003**, *4*, P3.

(39) Fontanillo, C.; Nogales-Cadenas, R.; Pascual-Montano, A.; de Las Rivas, J. Functional analysis beyond enrichment: Non-redundant reciprocal linkage of genes and biological terms. *PLoS One* **2011**, *6*, e24289.

(40) Oliveros, J. C. VENNY. An interactive tool for comparing lists with Venn Diagrams. http://bioinfogp.cnnb.csic.es/tools/venny/index.html.

(41) Irvine, G. B. Determination of molecular size by size-exclusion chromatography (gel filtration). *Curr. Protoc. Cell Biol.* **2001**, *Chapter 5*, Unit 5.5.

FIGURE LEGENDS

Figure 1. A general overview of the protein profiling analysis of Ramos B-cells by three different approaches (MS/MS proteomics, SEC-MAP, and RNA-Sequencing). **A)** A visual description of steps followed for analyzing *Ramos* B-cells by applying MS/MS proteomics, SEC-MAP or RNA-Seq. **B and C)** Detailed description of the followed steps for MS/MS proteomics and SEC-MAP, respectively. Protein samples (ORG, CYT, MB, and NUC) were loaded onto an SDS-PAGE gel and separated for trypsinization before LC-MS/MS analysis in an Orbitrap Q-Velos. Peptide Shaker software was used for the database searching. Each protein sample (ORG, CYT, MB, and NUC) was sub-fractionated in 24 fractions by size-exclusion chromatography. Then, they were incubated with the microsphere array (576 microsphere populations) and the binding was detected by streptavidin-PE. Data analysis was done using INFINICYT™ software. *ORG*, organelle proteins; *CYT,* cytoplasmic proteins; *MB*, membrane proteins; *NUC*, nuclear proteins; *SEC-MAP,* Size-Exclusion Chromatography – Microsphere-based Affinity Proteomics; *RNA-Seq*, RNA-Sequencing; *PE*, phycoerythrin.

Figure 2. Flow cytometry gating for SEC-MAP analysis. Events of interest were selected, after debris removing, in an FSC-H/FSC-A dot-plot **(A)**. Then, microsphere populations were identified and classified in an AX488/Ax647 dot-plot **(B)** followed by a sub-classification in a PacB/PacO dot-plot **(C)**. In total, 576 microsphere populations were identified (36 alexas populations x 16 pacific populations) and their PE median fluorescence intensity values were used for the analysis **(D)**.

Figure 3. Density plots and box plots of the distribution of the expression signal measured in $\log_2$ (FPKM+1) in the processed RNA- Seq data set SRX105534 of *Ramos* B-cell line. The figure compares the expression signal of all genes which could be mapped to 19,518 neXtProt IDs (RNA-Seq complete – 19,518 IDs, Blue) versus the signal of the genes corresponding to the proteins detected in the complete MS experiments (MS/MS-complete mapping – 5,672 neXtProt IDs, Green) and the proteins identified in the SEC-MAP experiment (SEC-MAP – 413 neXtProt IDS, Red).

Figure 4. Comparison of the expression frequency of the 413 proteins present in the SEC-MAP array along the data sets of MS/MS, RNA-Seq, and SEC-MAP. A protein is considered as expressed in a dataset if its: (1) number of peptides ≥ 1; (2) $\log_2$

(FPKM+1) ≥1 [FPKM]; or (3) QAS value ≥ 1 [AB]. There are eight different stages: (1) Peptide=NA;FPKM<1 (Red) and AB<1: Protein non-detected in all data sets; (2) Peptide=NA; FPKM<1 (Red) and AB≥1: Protein only detected in SEC-MAP; (3) Peptide=NA; FPKM≥1(beige) and AB<1: Protein only detected in RNA-Seq; (4) Peptide=NA; FPKM≥1(beige) and AB≥1: Protein detected in RNA-Seq and SEC-MAP; (5) Peptide≥1; FPKM<1 (light blue) and AB<1: Protein only detected in MS/MS; (6) Peptide≥1; FPKM<1 (light blue) and AB≥1: Protein detected in MS/MS and SEC-MAP; (7) Peptide≥1; FPKM≥1 (dark blue) and AB<1: Protein detected in MS/MS and RNA-Seq; (8) Peptide≥1; FPKM≥1(dark blue) and AB≥1: Protein detected in all data sets.

Figure 5. Venn diagram of the expression of the 405 proteins out of the 413 included in the SEC-MAP array. A protein is considered as expressed in a dataset if its: (1) number of peptides≥1; or (2) FPKM≥1; or (3) the QAS value≥1 for MS/MS, RNA-Seq, and SEC-MAP approaches, respectively. A total of 231 proteins is detected in all three approaches. 8 proteins are non-detected in any of the mentioned approaches and they are not represented in the Venn diagram. SEC-MAP detected 376 proteins with a QAS value≥1; RNA-Seq detected 333 proteins with an FPKM≥1, and MS/MS proteomics detected 271 proteins with at least 1 peptide/protein.

Figure 6. Plots showing some examples of proteins included in the SEC-MAP array. Each graphic depicts the PE-expression profile for each indicated target protein. Each colored line corresponds to each of the 4 subcellular compartments (blue for cytoplasm, green for membrane, red for organelle, and black for nucleus). Asterisks indicate the position of the detected peak (according to the rules set in Materials and methods section). A) Examples of antibody profiles against proteins detected in the three strategies (MS/MS, RNA-Seq, SEC-MAP). BAX graphic lines are referred to three different clones included against the same BAX protein. B) Examples of antibody profiles against proteins detected only by SEC-MAP strategy. C) Example of an antibody profile against a protein non-detected by SEC-MAP but detected by MS/MS and RNA-Seq. *MFI*, median fluorescence intensity; *SEC*, size-exclusion chromatography.

Table 1. List of criteria for SEC-MAP evaluation. Microsphere array results were evaluated scoring the peaks detected in the SEC fractions. Peak identifications were positively and negatively scored according to the criteria shown in the table.

Criterion	QAS value
1. Well defined peaks over background noise (≥140% difference between the minimum and the maximum value of the peak)	+1
2. The peak is observed in the expected subcellular fraction based on the literature (neXtProt database)	+1
3. Peak matching the expected molecular weight (MW) for the protein (+/-1 SEC fraction)	+1
4. When conditions 2 and 3 are both achieved	+1
5. Peaks present in fractions >19, as long as this localization does not match with the expectedly MW (i.e. the peak correspond with proteolytic forms)	-3
6. Two or more antibodies against the same target protein showing a highly similar elution profile*	+1

* This criterion must only be applied when the protein has already scored ≥2 points for the previous criteria.

SEC, Size-Exclusion Chromatography

Table 2. Integration of MS/MS proteomics and RNA-Seq transcriptomics [a].

	no. of genes (ENSG IDs)	no. of proteins (neXtProt IDs)
RNA-Seq[b]	20,533	19,518
RNA-Seq (with FPKM≥1)[c]	9,535	9,523
MS/MS[d]	5,982	5,707
Detection by MS/MS + RNA-Seq[e]	5,947	5,672
Detection by MS/MS + RNA-Seq (with FPKM≥1)[f]	5,165	5,157

[a] The table shows the number of gene and protein distinct IDs identified in the RNA-Seq transcriptomics and MS/MS proteomics assays, respectively: Row 1, RNA-Seq transcriptomics data; row 2, MS/MS proteomics data; row 3, MS/MS proteomics mapped to RNA-Seq transcriptomics; row 4, RNA-Seq transcriptomics mapped to neXtProt IDs; row 5, MS/MS proteomics mapped to RNA-Seq transcriptomics. The columns in the table correspond to (i) all the human neXtProt IDs detected; (ii) the mapped Ensembl IDs. [b] genes detected in the RNA-Seq transcriptomics data independently of the FPKM value. [c] genes detected in the RNA-Seq transcriptomics data with FPKM≥1 mapped to neXtProt IDs. [d] proteins detected by at least 1 unique peptide in the MS/MS proteomics experiment. [e] detection by MS/MS proteomics and RNA-Seq transcriptomics. [f] detection by MS/MS proteomics and RNA-Seq transcriptomics presenting at least 1 unique peptide per protein and FPKM≥1 for MS/MS and RNA-Seq data, respectively.

FIGURES

Figure 1

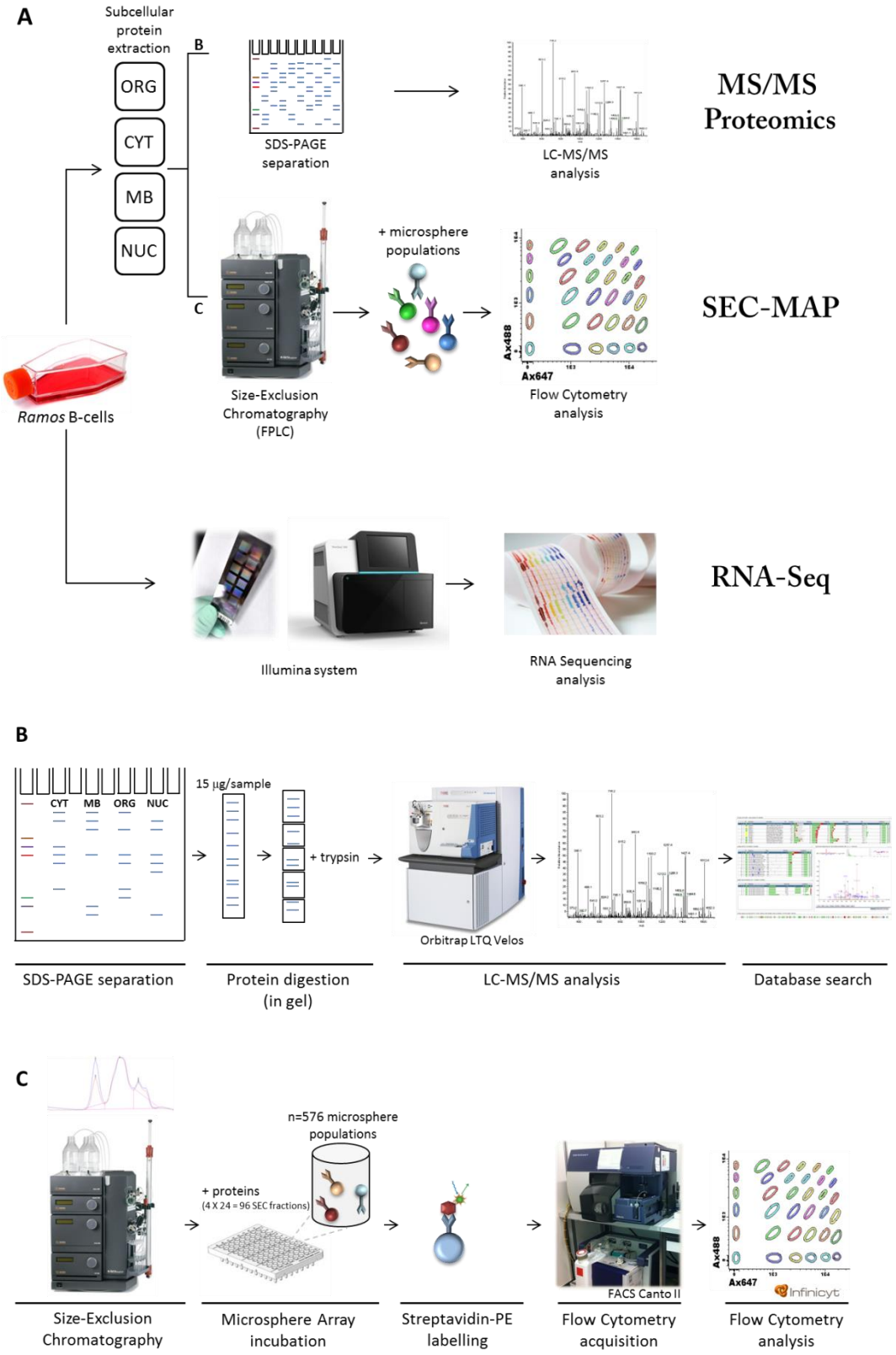


Figure 2

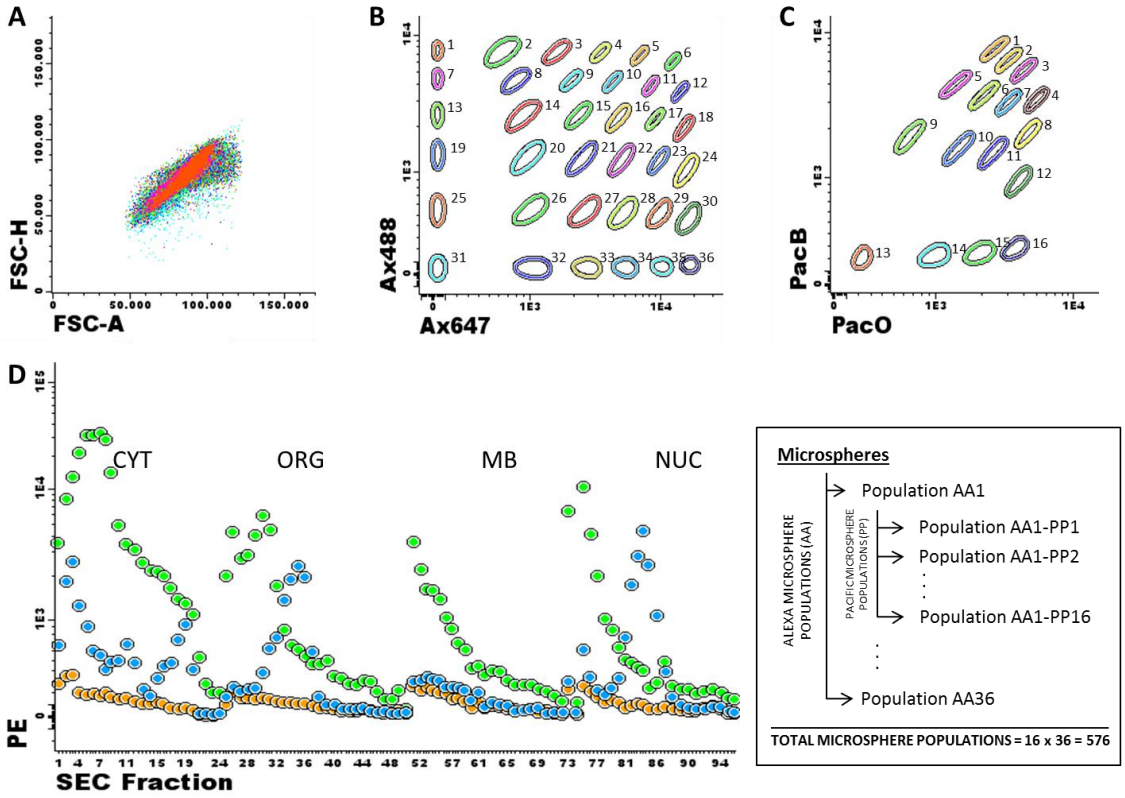

A
FSC-H
FSC-A
B
Ax488
Ax647
C
PacB
PacO
D
CYT
ORG
MB
NUC
PE
SEC Fraction
Microspheres
Population AA1
ALEXA MICROSPHERE POPULATIONS (AA)
PACIFIC MICROSPHERE POPULATIONS (PP)
Population AA1-PP1
Population AA1-PP2
Population AA1-PP16
Population AA36
TOTAL MICROSPHERE POPULATIONS = 16 x 36 = 576

Figure 3

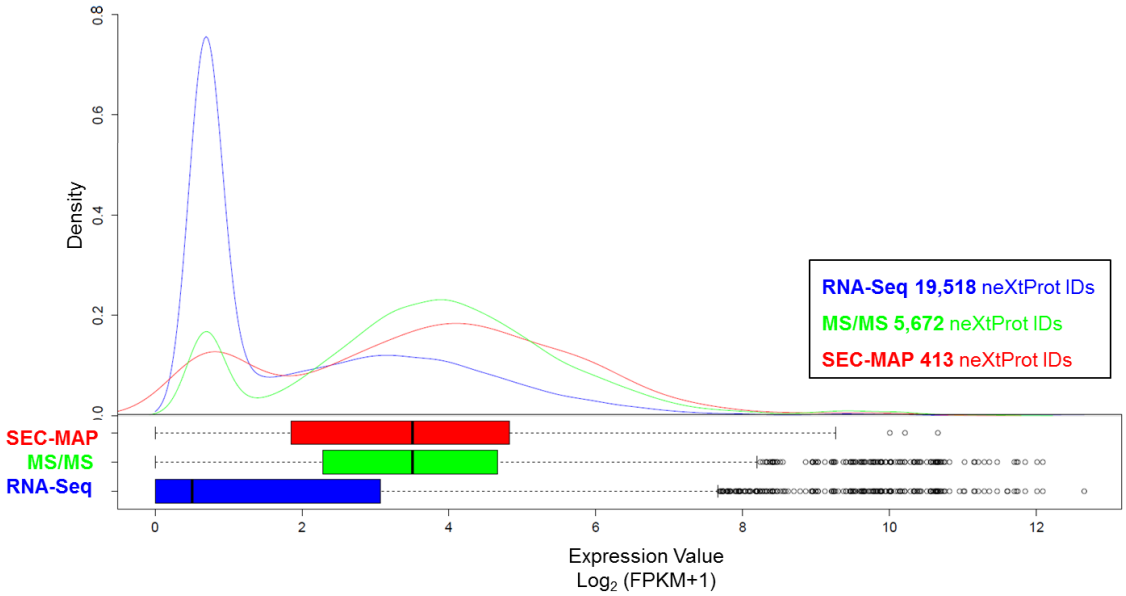

Density
0.0
0.2
0.4
0.6
0.8
RNA-Seq 19,518 neXtProt IDs
MS/MS 5,672 neXtProt IDs
SEC-MAP 413 neXtProt IDs
SEC-MAP
MS/MS
RNA-Seq
0
2
4
6
8
10
12
Expression Value
Log2 (FPKM+1)

Figure 4

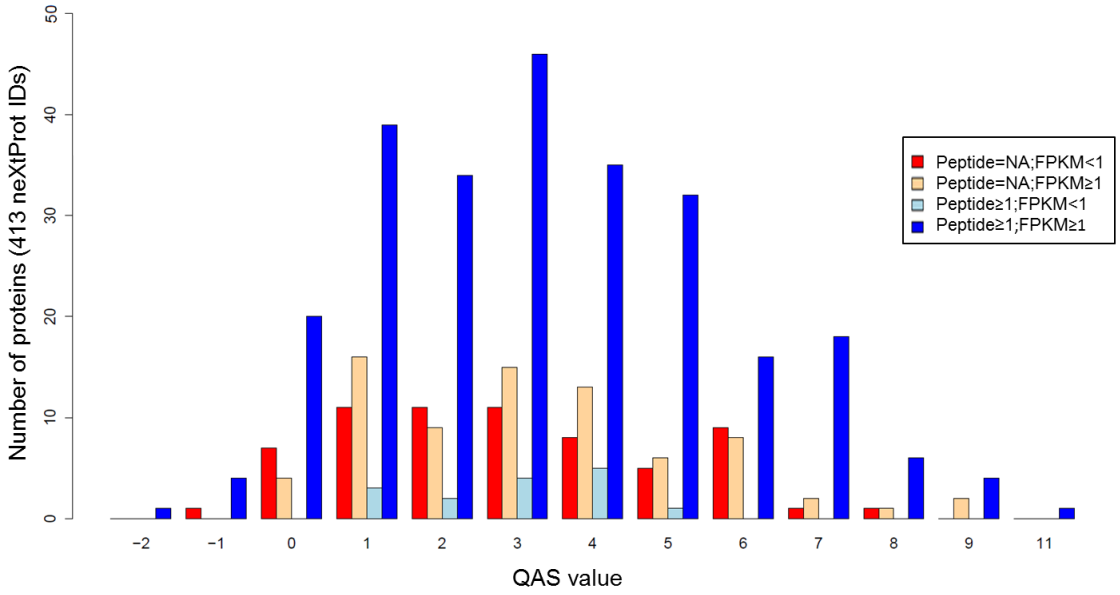

Number of proteins (413 neXtProt IDs)
0
10
20
30
40
50
−2
−1
0
1
2
3
4
5
6
7
8
9
11
QAS value
Peptide=NA;FPKM<1
Peptide=NA;FPKM≥1
Peptide≥1;FPKM<1
Peptide≥1;FPKM≥1

Figure 5

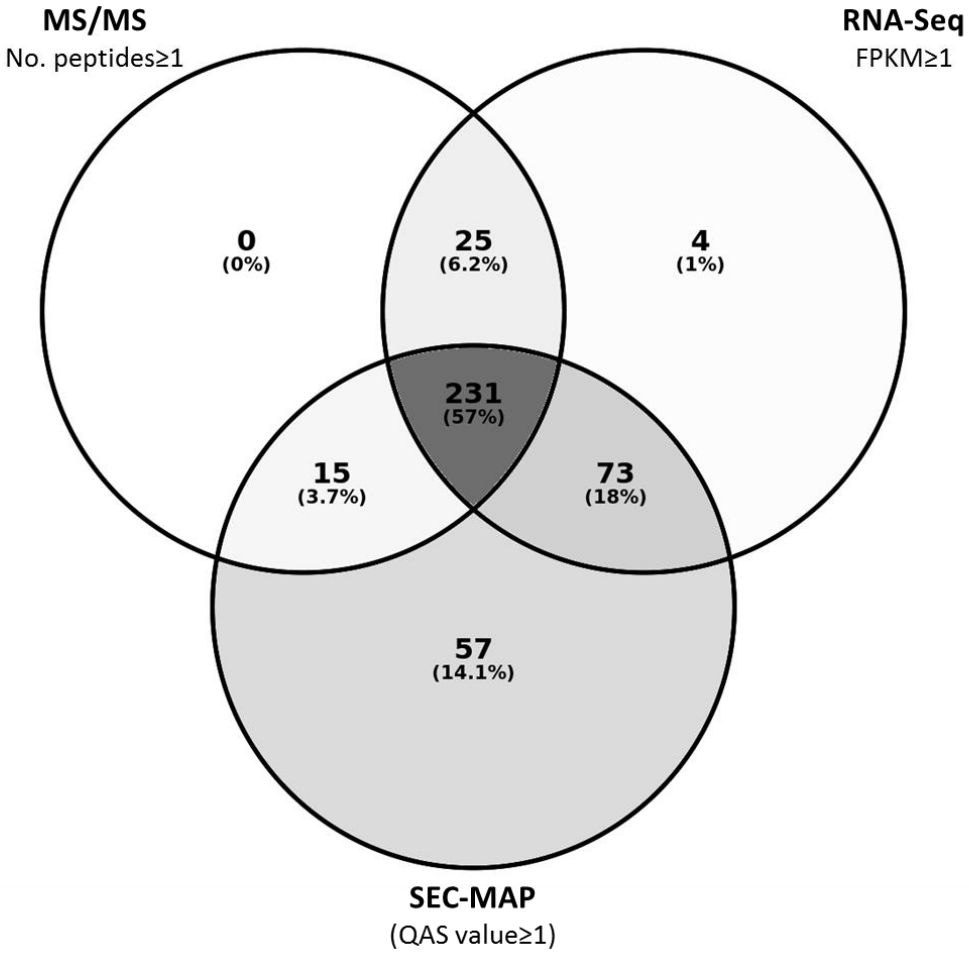
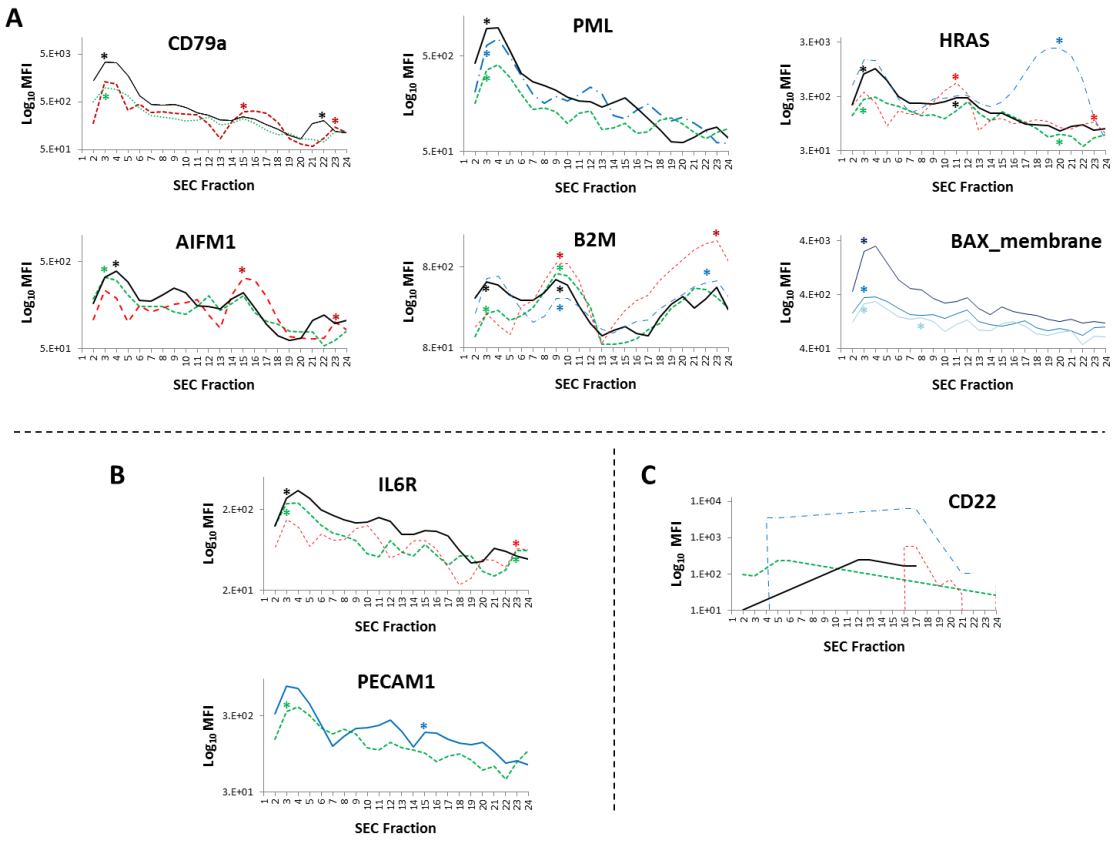

MS/MS
No. peptides≥1
RNA-Seq
FPKM≥1
0
(0%)
25
(6.2%)
4
(1%)
231
(57%)
15
(3.7%)
73
(18%)
57
(14.1%)
SEC-MAP
(QAS value≥1)

Figure 6


A
CD79a
PML
HRAS
AIFM1
B2M
BAX_membrane
Log10 MFI
SEC Fraction
B
IL6R
PECAM1
C
CD22

Revealing Cell Signaling Pathways in Chronic Lymphocytic Leukemia Tumor B-cells by Integration of Global Proteome and Phosphoproteome Profiles

Running title: Quantitative Proteome of Chronic Lymphocytic Leukemia B-cells

Paula Díez[1,2], Conrad Droste[3], Julia Almeida[1], Marcos González[4],

Alberto Orfao[1], Javier De Las Rivas[3*] and Manuel Fuentes[1,2*]

[1]Department of Medicine and Cytometry General Service-Nucleus. Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

[2] Proteomics Unit. Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

[3] Bioinformatics and Functional Genomics Research Group. Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain

[4] Service of Hematology. University Hospital of Salamanca (IBMCC/IBSAL), 37007 Salamanca, Spain

* Corresponding authors:

Manuel Fuentes, Ph.D., Department of Medicine and General Cytometry Service-Nucleus, Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain.
E-mail: mfuentes@usal.es; phone: +34 923294811; fax: +34 923294743

Javier De Las Rivas, Ph. D., Bioinformatics and Functional Genomics Research Group. Cancer Research Centre (IBMCC/CSIC/USAL/IBSAL), 37007 Salamanca, Spain.
E-mail: jrivas@usal.es; phone: +34 923294819; fax: +34 923294743

Financial support. We gratefully acknowledge financial support from the Spanish Health Institute Carlos III (ISCIII) for the grants FIS PI12/00905 and FIS PI14/01538. We also acknowledge Fondos FEDER (EU), Junta Castilla-León (grant BIO/SA07/15) and Fundación Solórzano (FS23/2015). The Proteomics Unit belongs to ProteoRed, PRB2-ISCIII, supported by grant PT13/0001, of the PE I+D+I 2013-2016, funded by ISCIII and FEDER. P.D. and C.D. are supported by a JCYL-EDU/346/2013 Ph.D. scholarship. The mass spectrometry proteomics

data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD005997.

Conflict of interest. The authors declare no conflict of interest.

Keywords: phosphoproteome; B-cell chronic lymphocytic leukemia (B-CLL); BCR signaling; cellular network; post-translational modification.

Word count: 4,746

Total number of figures/tables: 6

References: 49

ABSTRACT

B-cell chronic lymphocytic leukemia (B-CLL) is a blood cancer with highly heterogeneous genomic alterations and altered signaling pathways. To date, no studies have investigated the phosphoproteome of unstimulated B-CLL tumor cells by using an immunoaffinity enrichment method. Here, we report the overall (quantitative) proteome and phosphoproteome of B-CLL and CLL-like monoclonal B-cell lymphocytosis (MBL) primary cells and the phosphoproteins potentially involved in the baseline pathological B-cell behavior using a high-resolution mass spectrometry-based approach. Overall, 2,970 proteins and 327 phosphoproteins were quantified across the five studied primary tumor samples (ProteomeXchange; PXD005997) including 329 phosphopeptides reported here for the first time. Although the five tumor proteomes shared a significant overlap (73%), the phosphoproteomes varied significantly highlighting the importance of B-CLL phosphosignature investigations. Despite such heterogeneity, tumor B-cells from different CLL and MBL patients also displayed common phosphoproteins such those involved in BCR signaling, cell-cell interactions (particularly with other immune cells), and the NF-kB/STAT3 pathway. Notably, these phosphoprotein profiles were independent of the cytogenetic alterations and/or IGHV mutational status of tumor cells. Despite the great similarity observed between CLL and MBL cells, both groups showed different levels of phosphorylated ARIH2 and PTPN11 suggesting a potential role for these two proteins involved in apoptosis-mediated cell death in disease progression. In summary, our results provide new insights into the global phosphoproteome of unstimulated B-CLL and MBL cells and the immune signaling pathways involved in the development and progression of this tumor, also offering a comparative overview of the proteome and phosphoproteome of these tumor cells.

INTRODUCTION

B-cell chronic lymphocytic leukemia (B-CLL) is the most common human blood cancer in the Western world which is characterized by high genomic heterogeneity in the absence of a common genetic lesion (1)'(2). Despite the genome of the CLL B-cells has been investigated in great detail(3–6), their phosphoprotein profiles and the signaling pathways involved still remain to be fully understood. This is critical for the understanding of the malignant transformation of CLL cells since most human cancers are guided by deregulation/alteration of protein networks, including protein kinase signaling pathways (7,8); in turn, such altered protein networks might be used as target molecules for innovative therapies (9,10). Besides, genetic alterations and DNA damage also trigger high phosphorylation levels of specific substrates that affect cell signaling pathways by altering protein localization, protein-protein interactions, protein stability, and enzymatic activity (9,11). Thus, to better understand the pathogenic mechanisms at the tumor cell level, simultaneous analysis of the protein and phosphoprotein profiles of CLL B-cells might be crucial for better mapping the altered signaling pathways leading to inappropriate expansion and survival of tumor cells and thereby, to better understand, classify and treat the disease (12).

Currently, it is well established that the B-cell receptor (BCR) signaling pathway plays a critical role in the development of CLL leading to altered patterns of phosphorylation of specific proteins related to cell proliferation, differentiation, and survival (13–15). Thus, inhibition of BCR signaling pathways (e.g. via usage of BTK and PI3K inhibitors, among other targeted therapies) has become a validated treatment strategy across a variety of lymphoid malignancies[5]. However, in order to fully understand how proteins involved in BCR signaling interact and mediate a crosstalk with other functional pathways that regulate cytokine signaling, microtubule dynamics, cell-cell interactions, microenvironment signaling, and cell survival, demands a phosphoproteomic analytical approach (16). As an example, the LYN and SYK phosphoproteins trigger a signal transduction cascade that involves other kinases, adapter molecules, and second messengers (17,18); LYN phosphorylates activation motifs at different phosphosites on the alpha and beta chains of the immunoglobulin molecules, the SYK tyrosine kinase, together with several phosphatases, which generates an inhibitory effect on signal transduction (19). Consequently, B-cell antigen recognition, differentiation, and maturation pathways are critically

related to the phosphorylation levels of these and other signaling molecules, and their dynamics inside the cell.

Overall characterization of the phosphoproteomic profile of a given cell/cell population might be challenging, mainly due to the complexity of the distinct pathways involved, the low levels of stoichiometry and the high dynamic range of phosphopeptides (12,20). In this regard, phosphosite-enrichment strategies are highly attractive and valuable, particularly those based on the use of antibodies against the targeted phosphosites/phosphopeptides (21). Thus, mass spectrometry (MS)-centric phosphoproteomics has emerged as a golden tool for global post-translational (PTM) analysis of cellular phosphoproteomes due to its high sensitivity and the ability to simultaneously detect thousands of (single) amino acid modifications per assay (20). Moreover, the rise of label-free MS strategies – i.e. without isotopic labeling – allows quantitative comparisons between the relative amounts of the targeted phosphoproteins/phosphopeptides avoiding additional costs and tedious experimental procedures (22). To date, as stated by Thurgood et al. (23), there are only two documented phosphoproteomic studies in CLL (24,25), both performed by using immobilized metal affinity chromatography (IMAC) and focused on the CXCL12/CXCR4 signaling of CXCL12 stimulated B-CLL cells. The use of IMAC has been broadly used for enrichment and sequencing of phosphopeptides; however, the selectivity of most of the metal ions is limited when working with complex samples (26). Thus, the implementation of immunoaffinity enrichment methods (eg. PTMScan® strategy) (27) could satisfactory address the selective phosphoprofiling of complex mixtures revealing the involvement of specific proteins in cell signaling pathways.

In the present study, we investigated for the first time the whole phosphoproteome of CLL and monoclonal B-cell lymphocytosis (MBL) primary tumor B cells using an immunoaffinity enrichment protocol coupled with a high-resolution MS-based approach. Overall, we identified 13,504 peptides from a total of 2,970 quantified proteins and 594 phosphopeptides from 327 phosphoproteins, of which 329 corresponded to new phosphopeptides reported here for the first time.

MATERIALS AND METHODS

Patients and samples. Peripheral blood (PB) samples were obtained from 4 CLL patients and 1 CLL-like MBL patient (Table 1) after informed consent was given by each individual according to the guidelines of the local ethics committees of the University Hospital of Salamanca, the National DNA Bank-ISCIII, and the Declaration of Helsinki of 1975, as revised in 2008. The stratification of the CLL and CLL-like MBL patients was performed following the "Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia"[1] and the 2016 revision of the World Human Organization (WHO)[2]. In each case, fresh PB-derived tumor B-cells were purified from the PB mononuclear cells (PBMCs) by positive selection using the magnetic associated cell sorting (MACS) (Miltenyi, Bergisch Gladbach, Germany) at the Cytometry Service of the University of Salamanca (NUCLEUS, Salamanca, Spain) (28). MACS-sorted tumor B-cells were immediately stored frozen in liquid nitrogen to minimize the protein degradation at the National DNA Bank-ISCIII (Salamanca, Spain), until analyzed as described below. An aliquot of the MACS-sorted tumor B-cells were $CD5^+/CD19^+$ stained and analyzed by flow cytometry to determine the sorting purity (>99%) of the clonal tumor population.

Cytogenetic analysis were performed on MACS-purified and fixed tumor B-cells using fluorescence in situ hybridization (FISH) (29) and the probesets from Vysis (Abbot Molecular, Des Plained, IL). B-cell clonality (patterns of rearrangement of the IGHV genes) was determined following the guidelines described by the EuroClonality Consortium (30). Sequences were considered as somatically mutated if they contained 2% deviation from the germline sequence.

Protein extraction and quantification. Freshly-frozen B-cells were thawed, pelleted by centrifugation at 200 *g* (5 min at 4ºC) and washed three times with PBS. After draining off the total PBS volume without disturbing the cell pellet, 1 mL (per 1.25 x 10^8 cells) of a lysis buffer - 9 M urea, 1 mM activated sodium orthovanadate, 2.5 mM sodium pyrophosphate, 1 mM β-glycerol phosphate, 20 mM HEPES pH 8.0; all from Sigma, St. Louis, MO - was added to the cell pellet, followed by sonication on ice (3 bursts for 30 sec each) according to the manufacturer instructions. Both phosphatase inhibitors (i.e. sodium orthovanadate and pyrophosphate) were added to preserve the phosphoprotein stability. Then, samples were centrifuged at 20,000 *g* for 15 min and the supernatant containing the proteins was collected

150 (27) and stored at -80ºC until used. Protein concentration was determined by the Bradford
151 assay using the Coomassie Plus Protein Assay Reagent (Thermo, Waltham, MA).

152 **SDS-PAGE separation for whole proteome analysis.** Each protein sample (15 μg) was
153 separated on a 4-20% gradient SDS-PAGE gel under reduction conditions. After
154 electrophoresis, gels were stained in a Coomassie solution (Supplementary Figure S1A). Then,
155 each gel line was cut into gel pieces which were destained with 15 mM potassium ferrocyanide
156 and 50 mM sodium thiosulfate (both from Sigma).

157 **Protein digestion.** Proteins (in solution and from gel pieces for phosphorylation and total
158 proteome assays, respectively) were reduced in 4.5 mM DTT (Sigma) for 30 min at 45ºC and
159 alkylated with 100 mM iodoacetamide (Sigma) for 15 min at room temperature in the darkness.
160 Afterward, protein digestion was performed overnight in 10 μg/mL trypsin (Promega, Madison,
161 WI).

162 The peptide solution was acidified with 1% trifluoroacetic acid (TFA, Sigma) and desalted
163 with 360-mg SEP-PAK Classic C18 columns (Waters Corp., Milford, MA, USA). Peptide elution
164 was performed with 40% acetonitrile (Sigma) in 0.1% TFA, and peptides were dried under
165 vacuum conditions.

166 **PTMScan® method.** Lyophilized peptides were mixed with PTMScan® Direct reagent bead
167 slurries (phosphotyrosine pY-1000 motif antibody; Cell Signalling Technology, Danvers, MA,
168 USA) for 2.5 h at 4ºC (27) for immunoaffinity purification of phosphopeptides. Afterward, beads
169 were pelleted and washed, and the peptides were eluted from the beads with 0.15% TFA.
170 Enriched peptides were purified on C18-Stage-Tips columns (Thermo) and stored at -20ºC until
171 analyzed by LC-MS/MS. See Supplementary Methods for further details on LC-MS/MS analysis,
172 database search, western blot validation, data processing, functional enrichment analysis;
173 Supplementary Figure S2 displays a schematic illustration of the workflow of the whole
174 procedure.

RESULTS

Quantitative global proteome profile of CLL and CLL-like MBL cells.

The overall protein content of tumor B cells was analyzed for each of the 5 samples included in the study with 2 technical replicates using LC-MS/MS (Supplementary Table S1). A total of 13,504 unique peptides were identified corresponding to 2,970 unique proteins. From these proteins, 56 (2%) corresponded to proteins without evidence at the protein level: 47 with evidence at the transcript level (PE=2), 7 inferred by homology (PE=3), and 2 were predicted proteins (PE=4), according to the neXtProt database (February 2017 release) and following the HUPO guidelines (www.hupo.org).

Once we compared the proteins identified per sample (Supplementary Figure S3A), a high degree of correlation and similarity was observed both for intra-sample (5) and inter-sample (40) comparisons. Thus, the global distribution of the proteins identified in all samples analyzed (Figure 1A) showed a high degree of overlap among tumoral B-cells from all 5 samples investigated (2,160/2,970 proteins; 73%). Functional pathway analysis (Supplementary Table S2) for the proteins expressed in CLL/MBL tumor cells assigned them to 154 clusters associated to general cellular functions (e.g. metabolic and catabolic processes, cell cycle, cell-cell adhesion, and RNA transport, among others) and specific B-cell associated functional roles such as antigen processing and presentation, immunity, functions related to the complement system and the MHC class II molecules, and BCR signaling. From those 2,160 proteins expressed in common by tumor CLL cells from all 5 patients, 220 (10%) were found to be phosphorylated (Supplementary Table S1); among other proteins, these included BTK, PRKCB, STAT1, and SYK, which are key proteins in BCR-associated cell signaling. Additionally, proteins involved in the regulation of the cell cycle (e.g. ACIN1, ATM, STAT3, and WDR1), cell proliferation (e.g. EIF3L, LYN, SRRT), metabolism (e.g. GOT2, PPP1CA, PYGB), and the regulation of Ca^{2+} flux (e.g. ANXA6, CALM1) - up to a total of 20 FEA functional clusters that also were related to immune cells, signaling pathways linked to B and T cells, and the proteasome (Supplementary Table S2) - were also identified to be present in a phosphorylated form in the tumor cells.

In contrast, phosphorylated proteins whose expression was restricted to a single sample were restricted to a small number of functional pathways, including DYNLL1 and NBN for sample A, RFWD3 for sample B, PIK3R4 and ATRN for sample C, and ARIH2 for sample E.

The proteome profile across distinct B-CLL and CLL-like MBL tumor B cell samples.

All proteins identified in common across the CLL/MBL cell samples analyzed were ranked (score 1 to 5) based on their expression levels within the five studied samples (i.e. for each given protein, 1 was assigned to the sample presenting the highest expression value and 5 to that presenting the lowest one). Figure 2A (left panel) illustrates this ranking in which 1,880/2,978 (63%) proteins with the highest score of 1 belonged to sample A, whereas 1,704/2,434 (70%) proteins with the lowest score of 5 belonged to sample E. Interestingly, for each sample a different score predominated (i.e. 66% proteins of sample A had a score of 1; 50% proteins of sample C scored 2; 45% proteins of sample B scored 3; 45% proteins of sample D scored 4; and 62% proteins of sample E scored 5). Interestingly, sample E showed a greater dispersion vs. samples A-D (Figure 2A, right panel). Based on the rank position (1-5) per protein, detected for each sample (A-E), a hierarchical clustering algorithm was applied (Supplementary Figure S4A) which revealed two main groups: group 1 included samples A and C and group 2 comprised samples B, D and E.

CLL/MBL tumor cell protein profiles and other CLL/MBL cellular features.

As stated above, based on the overall proteome of CLL/MBL cells, two main groups of samples were observed (Figure 3A, Supplementary Table S3): group 1 included samples A and C and group 2 consisted of samples B, D and E. Such clustering was particularly based on proteins involved in proteasome activity (e.g. PSMD1, PSMD9, PSMA1), immune regulation (e.g. BCLF1, IgHM, CD2-P, CD5, CD47, CD48, CD53, CD79B, IGBP1, IGHC, IGHG1, B2M, ZAP70), HLA proteins (-A, -C, -DMB, -DRB1, -E), and the BCR signaling pathway (SYK, LSP1, MYCBP, BLNK, ATM) (1,167 grey labeled proteins in Supplementary Figure S4, panel A).

Another group of proteins (yellow-labeled) differentially expressed in the two clusters, consisted of 980 proteins, mostly related to HLA and antigen presentation (Supplementary Figure S4, panel A). Other groups of proteins contributing to the clustering (clusters labeled as

green, blue, and magenta) showed fewer differences across distinct samples analyzed. Among them, a group (green-labeled) of quite homogeneously expressed proteins within the two groups of samples corresponded to proteins directly related to protein synthesis and expression (Supplementary Figure S4, panel A).

In contrast, no clear association was found between tumor cell cytogenetics and IGHV mutational status, and the overall tumor cell protein expression profile. Despite this, when we actively searched for those proteins coded in the differentially altered (e.g. deleted) chromosomal regions, specific patterns were observed, particularly for del11q22.3, del11q23.3, and del13q14 (Supplementary Table S4). Thus, lower expression levels of ATM and CUL5 were associated with del11q22.3 (sample C), decreased amounts of MLL, SCN4B, CD3D, ARCN1, and TREH were also found in sample C that carried del11q23.3, and low RB1 levels were associated with del13q14 (Figure 4). Similarly, the expression levels of CD20, CD5, CD45, CD23, CD79B, CD49D, CD200 identified by the MS approach correlated well with the immunophenotypic profile found at the single cell level by flow cytometry (data not shown).

Phosphoproteome profile across distinct CLL and CLL-like MBL tumor B cell samples

The phosphoproteome profile of CLL/MBL tumor cells was evaluated after applying an affinity enrichment method for selective and specific detection of phosphotyrosine motifs (P-Tyr-1000; PTM) present in the overall protein fraction of each individual CLL and MBL cell sample. In addition to this PTM selection, phosphorylation of threonine (Thr) and serine (Ser), as well as oxidation of methionine (Met), were also identified. Overall, 594 phosphopeptides corresponding to a total of 327 phosphoproteins were identified in the five samples analyzed (Supplementary Table S5). Of note, more than half of the phosphopeptides identified (329/594) were detected for the first time and therefore, they had not been previously reported in the neXtProt database (February 2017 release). Interestingly also, 69 of 327 (21%) phosphorylated proteins were only identified via the enrichment method since they had not been detected in the (quantitative) global proteome dataset, these phosphoproteins included one protein with only experimental evidence at the transcriptional level (PE=2) and one protein inferred from homology (PE=3).

Overall, a high correlation (Supplementary Figure S3B) was observed between replicates of the same sample (r^2>0.99); in contrast, the correlation observed between different samples

was much lower (e.g. r^2 of between 0.63 and 1.00), as could be expected because of the higher phosphoproteome dynamism across different samples. As a result (Figure 1B), only 57 of 327 phosphoproteins (17%) identified were detected in common in the five CLL/MBL cell samples analyzed. Such phosphoproteins (Supplementary Table S5) corresponded to proteins directly involved in protein phosphorylation (protein kinases such as PRKCB, LYN, SYK, ATM, BLK, JAK1, JAK2), signal transduction (e.g. KHDRBS1, STAM2, STAP1), and intracellular protein transport (NSF, CUL3) (Supplementary Table S2). Thus, most phosphoproteins identified were present in only a subset of the samples or just in one sample. Briefly, 103/327 (32%) phosphoproteins were uniquely identified to be phosphorylated in samples C and D (e.g. BLK, CD19, PLCG2, and SCIMP); functional analysis showed they had specific roles in RNA binding, the spliceosome, at the same time they contained relevant interaction domains such as the SH3, SAP, KH, and LIM domains (Supplementary Table S2). In turn, sample A showed three uniquely phosphorylated proteins (HERC1, NUDT3, and PSMD9) and 10 and 49 proteins were restricted to sample C (DCP1B, GGA2, GBE1, IFIT5, JAK1, PPA2, SRRT, SRSF7, XRCC6, and ZNF24), and D (e.g. CBL, DOCK11, DOK2, EXOSC10, IKZF3, ILF3, LSP1, PDCL3, PTPN18, RIPK2, SASH3, SETD1A, and YBX1), respectively (Supplementary Table S5).

Differential phosphoproteome profiles based on a scoring classification (Figure 2B, left panel), showed higher scores (values 1 and 2) for samples D and C, respectively, while the remaining rank values (3-5) were homogeneously distributed across samples A, B, and E (Supplementary Figure S4B).

CLL/MBL tumor cell phosphoproteome profiles and other CLL/MBL cellular features

As described above, significantly different phosphoproteome profiles were observed across the different samples analyzed once relative expression levels were considered (Figure 3B). Overall, two groups of samples (group 1: samples A to C showing UM-IGHV; group 2: samples D and E showing M-IGHV) were identified based on five main groups of proteins found to be differentially expressed between them (Supplementary Table S3). The most contributing group of proteins consisted of 201 proteins (green-labeled cluster); these included the phosphorylated LYN, BCLAF1, SYK, CD2AP, TP53BP1, BLK, LSP1, RAB10, CD19 (only phosphorylated in C and D), and BTK, among other proteins. The remaining clusters of phosphoproteins (labeled as

289 yellow, grey, blue, and magenta in Figure 3B) contributed less to the clustering. No clear
290 association was found between the tumor cell phosphoproteome and tumor cytogenetics, the
291 IGHV mutational status and other features of the disease investigated.

DISCUSSION

B-CLL cells have been extensively characterized on genomic, transcriptomic, and immunophenotypic grounds, several of the genetic alterations and phenotypic markers identified proving to be of great diagnostic and/or prognostic utility (3,4,31,32). Nevertheless, a global overview of the phosphoproteome of unstimulated CLL cells has not been reported so far in detail. Only two studies (24,25) have addressed the phosphoprofiling of CLL, but both stimulating the cells and using the IMAC technology. Here, we investigate the application of a new approach based on immunoaffinity purification to improve the phosphopeptide recovery in a highly accurate manner. Enhancing high-throughput analysis of the tumor cell proteome by using high-resolution MS/MS approaches coupled to immunoaffinity enrichment processes might, therefore, contribute to getting a deeper insight into the disease and the potentially proteins and protein networks targetable by existing novel drugs (21,33). In turn, the analysis of the CLL proteome has been addressed by using LC-MS/MS strategies, particularly those based on stable isotope labelling as iTRAQ (34,35). However, a comprehensive comparison of both proteome and phosphoproteome data sets is lacking in CLL for revealing the potential role of the phosphoproteins.

Here, we investigated for the first time the overall quantitative phosphoproteome of (highly purified and unstimulated) CLL and MBL tumor B cells directly obtained from primary patient samples by using an immunoaffinity approach for phosphopeptide enrichment (PTMScan® method). For this purpose, those peptides (and therefore their corresponding proteins) which had phosphorylated tyrosine (Tyr), serine (Ser), and/or threonine (Thr) residues were purified, and screened, and their potential involvement in larger protein networks and signaling pathways altered in CLL/MBL cells were investigated to better understand the relevance of the phosphoproteome in the activation of tumor cell–associated signaling pathways applying a quantitative proteomics profiling approach. Moreover, the phosphoproteome dataset was compared to the proteome one in order to get insights into the involvement of such post-translational modifications in the development of the CLL. Overall, our results showed a highly similar and overlapping proteome profile for all five samples analyzed, in the absence of major differences between the 4 CLL and the CLL-like MBL tumor cell samples analyzed. As could be expected, those proteins expressed in common in CLL/MBL tumor B cells corresponded to

proteins involved in general cell functions as well as in specific B-cell activities. Despite this, multiple proteins showed a pattern of expression restricted to a subset of the samples or even a single tumor cell sample. Thus, sample A was the only one showing expression of DYNL1, a protein that has been reported to be able to neutralize the anti-apoptotic activity of BCL2 and favor expansion of the leukemic B cells (36,37). Similarly, expression of RFWD3 (a positive regulator of the stability of p53) was restricted to sample B which corresponded to the only sample showing deletion of chromosome 13q14 including deletion of the RB1 gene, in addition to del14q32, and trisomy +12.

For sample C, twelve proteins, including PIK3R4, were found to be differentially expressed vs. all other cases. Kristensen *et al.* (38) described PIK3R4 (key to initiate the autophagy process) as a potential prognostic biomarker in CLL, since high expression levels of this protein were associated with more aggressive disease. In turn, the CLL-like MBL sample E differentially expressed ARIH2 (also known as Triad1) which acts as a tumor suppressor protein. Briefly, ARIH2 is blocked by PTPN11, a protein that was also detected in the MBL proteome dataset but at very low expression levels (up to 20-fold less than in the CLL samples). Interestingly, PTPN11 was also found to be phosphorylated (pPTPN11; Tyr62, Tyr546, and Tyr584 phosphosites) in all CLL samples but not in the MBL cells analyzed. Altogether, these findings suggest that in CLL samples A-D, pPTPN11 could block the expression of ARIH2 (this protein was not detected in A-D samples) and, therefore, its suppressive leukemic effect, while expression of ARIH2 is not blocked and remains high in MBL, blocking CLL development and progression in the later MBL (sample E) case (39). Despite only one MBL sample was investigated, this sample was associated with the absence of expression of 293 proteins found in common in the other CLL tumor B-cell samples. Among others, these proteins included oncogenic proteins such as VAV1 (proto-oncogene), BCL2 (apoptosis regulator, phosphorylated on Thr464), ZAP70 (tyrosine protein kinase used as CLL prognostic marker), CD53, and CD20.

In-depth analysis of tumor cell, phosphoproteomics revealed only a small percentage of phosphorylated proteins in common to all samples, despite the overall similar global proteome of the CLL/MBL tumor cells analyzed. Because of the relevant role of protein phosphorylation in cell signaling, this might shed light on the understanding of the heterogeneity of the disease particularly in terms of genetic origin and progression. Interestingly, samples C and D (both

corresponded to stage C/IV CLL cases) shared around 50% of the phosphoproteins identified which might be considered as an asset of the critical PTM in this disease.

Interestingly, kinases found to be phosphorylated in common across all samples included proteins that drive various signal transduction pathways involved in the pathogenesis of CLL such as BCR signaling, chemokine receptor signaling, and toll-like receptor (TLR) signaling pathways (40–42). Of note, in all samples analyzed the chemokine receptor and TLR signaling pathways were under-represented (phosphorylation being restricted in common to the CXCR4, STAT3, and TLR1 proteins) vs the BCR signaling pathway. Thus, many BCR-related proteins were found to be phosphorylated in all or the majority of samples (e.g. LYN, SYK, PI3K, BTK, ZAP70, PLCG3, ERK, NFAT, CD19, PRKCB, NFKB, JNK and VAV) (43–48), suggesting that BCR signaling plays a critical role in maintenance/survival of CLL/MBL tumor cells (Figure 5). In line with previous observations (49), several proteins involved in the NF-κB and STAT3 signaling pathway related to CLL cell activation, proliferation, survival, adhesion and homing were also found to be increased in common in CLL samples (samples A to D) at significantly higher levels than in CLL-like MBL cells (i.e. sample E), suggesting a potential role for this specific signalling pathway in the malignant transformation of the disease.

Overall, our results provide a first integrated MS-based map of the unstimulated CLL global phosphoproteome, and show that CLL/MBL tumor B cells from different patients display a high degree of overlap in their global pattern of protein expression and the activation in common of the BCR signaling pathway; in contrast they display a highly diverse phosphoproteome that might contribute to both understand the common ontogenic mechanisms of the disease and to explain the heterogeneous clinical behaviour of the disease. Further studies in larger patient series and normal B-cell populations are required to confirm these findings, grouping the patients by their cytogenetic alterations, IGHV mutational status and/or Binet/Rai status.

ACKNOWLEDGEMENTS

The authors would like to thank all the clinicians and technicians in the Cytometry and Cell Purification Services of the University of Salamanca, the Spanish National DNA Bank (Banco Nacional de DNA Carlos III, University of Salamanca) and the Genomic Unit of Cancer Research Centre (IBMCC, USAL-CSIC) for their support in the data collection for the preparation of this manuscript. We are also grateful to Matt Stokes for his support during the study and to Ignacio Criado for the supports with the flow cytometry analysis.

Conflict of interest.

The authors declare no conflict of interest.

REFERENCES

1. Hallek M, Cheson BD, Catovsky D, Caligaris-Cappio F, Dighiero G, Dohner H, et al. Guidelines for the diagnosis and treatment of chronic lymphocytic leukemia: a report from the International Workshop on Chronic Lymphocytic Leukemia updating the National Cancer Institute-Working Group 1996 guidelines. Blood 2008;111:5446–56.

2. Swerdlow SH, Campo E, Pileri SA, Harris NL, Stein H, Siebert R, et al. The 2016 revision of the World Health Organization (WHO) classification of lymphoid neoplasms. Blood 2016;127:2375-2391.

3. Puente XS, Beà S, Valdés-Mas R, Villamor N, Gutiérrez-Abril J, Martín-Subero JI, et al. Non-coding recurrent mutations in chronic lymphocytic leukaemia. Nature 2015; 526:519–24.

4. Puente XS, Pinyol M, Quesada V, Conde L, Ordonez GR, Villamor N, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature 2011;475:101–5.

5. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, Lawrence MS, et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell 2013;152:714–26.

6. Schuh A, Becq J, Humphray S, Alexa A, Burns A, Clifford R, et al. Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. Blood 2012;120:4191–6.

7. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70.

8. Lemmon MA, Schlessinger J. Cell signaling by receptor tyrosine kinases. Cell 2010; 141:1117–34.

9. Casado P, Alcolea MP, Iorio F, Rodriguez-Prados JC, Vanhaesebroeck B, Saez-Rodriguez J, et al. Phosphoproteomics data classify hematological cancer cell lines according to tumor type and sensitivity to kinase inhibitors. Genome Biol 2013;14:R37.

10. Hallek M. Chronic lymphocytic leukemia: 2015 Update on diagnosis, risk stratification,

and treatment. Am J Hematol. 2015;90:446–60.

11. Stokes MP, Silva JC, Jia X, Lee KA, Polakiewicz RD, Com MJ. Quantitative profiling of DNA damage and apoptotic pathways in UV damaged cells using PTMScan direct. Int J Mol Sci. 2013;14:286–307.

12. Rush J, Moritz A, Lee KA, Guo A, Goss VL, Spek EJ, et al. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. Nat. Biotechnol. 2005;23:94–101.

13. Stevenson FK, Krysov S, Davies AJ, Steele AJ, Packham G. B-cell receptor signaling in chronic lymphocytic leukemia. Blood 2011; 118:4313–20.

14. Hoellenriegel J, Meadows SA, Sivina M, Wierda WG, Kantarjian H, Keating MJ, et al. The phosphoinositide 3???-kinase delta inhibitor, CAL-101, inhibits B-cell receptor signaling and chemokine networks in chronic lymphocytic leukemia. Blood 2011;118:3603–12.

15. Caligaris-Cappio F, Ghia P. Novel insights in chronic lymphocytic leukemia: Are we getting closer to understanding the pathogenesis of the disease? J. Clin. Oncol. 2008;26: 4497–503.

16. Purroy N, Carabia J, Abrisqueta P, Egia L, Aguiló M, Carpio C, et al. Inhibition of BCR signaling using the Syk inhibitor TAK-659 prevents stroma-mediated signaling in chronic lymphocytic leukemia cells. Oncotarget 2017; 8:742-56.

17. Takata M, Sabe H, Hata A, Inazu T, Homma Y, Nukada T, et al. Tyrosine kinases Lyn and Syk regulate B cell receptor-coupled Ca2+ mobilization through distinct pathways. EMBO J 1994;13:1341–9.

18. Mócsai A, Ruland J, Tybulewicz VLJ. The SYK tyrosine kinase: a crucial player in diverse biological functions. Nat Rev Immunol 2010;10:387–402.

19. Wiestner A. Emerging role of kinase-targeted strategies in chronic lymphocytic leukemia. Blood 2012; 120:4684–91.

20. Lawrence RT, Searle BC, Llovet A, Villen J. Plug-and-play analysis of the human

phosphoproteome by targeted high-resolution mass spectrometry. Nat Methods 2016;13:431–4.

21. Fleuren EDG, Zhang L, Wu J, Daly RJ. The kinome “at large” in cancer. Nat Rev Cancer 2016;16:83–98.

22. Riley NM, Coon JJ. Phosphoproteomics in the Age of Rapid and Deep Proteome Profiling. Anal Chem. 2016;88:74–94.

23. Thurgood LA, Chataway TK, Lower KM, Kuss BJ. From genome to proteome: Looking beyond DNA and RNA in chronic lymphocytic leukemia. J Proteomics 2017;155:73–84.

24. O’Hayre M, Salanga CL, Dorrestein PC, Handel TM. Phosphoproteomic analysis of chemokine signaling networks. Methods Enzymol. 2009;460:331–46.

25. O’Hayre M, Salanga CL, Kipps TJ, Messmer D, Dorrestein PC, Handel TM. Elucidating the CXCL12/CXCR4 signaling network in chronic lymphocytic leukemia through phosphoproteomics analysis. PLoS One 2010;5:e11716.

26. Thingholm TE, Larsen MR. Phosphopeptide Enrichment by Immobilized Metal Affinity Chromatography. Phospho-Proteomics Methods Protoc 2016; 1355:123–33.

27. Stokes MP, Farnsworth CL, Moritz A, Silva JC, Jia X, Lee K A, et al. PTMScan Direct: Identification and Quantification of Peptides from Critical Signaling Proteins by Immunoaffinity Enrichment Coupled with LC-MS/MS. Mol Cell Proteomics. 2012;11:187–201.

28. Nieto WG, Almeida J, Teodosio C, Abbasi F, Allgood SD, Connors F, et al. Commentary: Comparison of current flow cytometry methods for monoclonal B cell lymphocytosis detection. Cytom Part B Clin Cytom 2010;78B:S4–9.

29. Henriques A, Rodríguez-Caballero A, Nieto WG, Langerak AW, Criado I, Lécrevisse Q, et al. Combined Patterns of IGHV Repertoire and Cytogenetic/Molecular Alterations in Monoclonal B Lymphocytosis versus Chronic Lymphocytic Leukemia. PLoS One 2013;8:e67751.

30. Evans PAS, Pott C, Groenen PJTA, Salles G, Davi F, Berger F, et al. Significantly improved PCR-based clonality testing in B-cell malignancies by use of multiple immunoglobulin gene targets. Report of the BIOMED-2 Concerted Action BHM4-CT98-3936. Leukemia 2007;21:207–14.

31. Calin GA, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik SE, et al. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N Engl J Med. 2005;353:1793–801.

32. Chiorazzi N. Implications of new prognostic markers in chronic lymphocytic leukemia. Hematology Am Soc Hematol Educ Program 2012;2012:76–87.

33. Lim YP. Mining the tumor phosphoproteome for cancer markers. Clin. Cancer Res. 2005; 11:3163–9.

34. Alsagaby SA, Khanna S, Hart KW, Pratt G, Fegan C, Pepper C, et al. Proteomics-based strategies to identify proteins relevant to chronic lymphocytic leukemia. J Proteome Res. 2014;13:5051–62.

35. Eagle GL, Zhuang J, Jenkins RE, Till KJ, Jithesh P V., Lin K, et al. Total Proteome Analysis Identifies Migration Defects as a Major Pathogenetic Factor in Immunoglobulin Heavy Chain Variable Region (IGHV)-unmutated Chronic Lymphocytic Leukemia. Mol Cell Proteomics 2015;14:933–45.

36. Wong DM, Li L, Jurado S, King A, Bamford R, Wall M, et al. The Transcription Factor ASCIZ and Its Target DYNLL1 Are Essential for the Development and Expansion of MYC-Driven B Cell Lymphoma. Cell Rep. 2016;14:1488–99.

37. Renault TT, Chipuk JE. Getting away with murder: how does the BCL-2 family of proteins kill with immunity? Ann N Y Acad Sci. 2013;1285:59–79.

38. Kristensen L, Kristensen T, Abildgaard N, Thomassen M, Frederiksen M, Mourits-Andersen T, et al. High expression of PI3K core complex genes is associated with poor prognosis in chronic lymphocytic leukemia. Leuk Res. 2015;39:555–60.

39. Sakamoto KM, Grant S, Saleiro D, Crispino JD, Hijiya N, Giles F, et al. Targeting novel

492 signaling pathways for resistant acute myeloid leukemia. Mol. Genet. Metab. 2015;
493 114:397–402.

494 40. Muggen AF, Singh SP, Hendriks RW, Langerak AW. Targeting Signaling Pathways in
495 Chronic Lymphocytic Leukemia. Curr Cancer Drug Targets 2016;16:669–88.

496 41. Ganghammer S, Gutjahr J, Hutterer E, Krenn PW, Pucher S, Zelle-Rieser C, et al.
497 Combined CXCR3/CXCR4 measurements are of high prognostic value in chronic
498 lymphocytic leukemia due to negative co-operativity of the receptors. Haematologica
499 2016;101:e99-102.

500 42. Chatzouli M, Ntoufa S, Papakonstantinou N, Chartomatsidou E, Anagnostopoulos A,
501 Kollia P, et al. Heterogeneous functional effects of concomitant B cell receptor and {TLR}
502 stimulation in chronic lymphocytic leukemia with mutated versus unmutated Ig genes. J
503 Immunol 2014;192:4518–24.

504 43. Davids MS, Brown JR. Targeting the B Cell Receptor Pathway in Chronic Lymphocytic
505 Leukemia. Leuk Lymphoma 2012;8194:1–23.

506 44. Amrein PC, Attar EC, Takvorian T, Hochberg EP, Ballen KK, Leahy KM, et al. Phase II
507 study of dasatinib in relapsed or refractory chronic lymphocytic leukemia. Clin Cancer
508 Res 2011;17:2977–86.

509 45. Friedberg JW, Sharman J, Sweetenham J, Johnston PB, Vose JM, Lacasce A, et al.
510 Inhibition of Syk with fostamatinib disodium has significant clinical activity in non-Hodgkin
511 lymphoma and chronic lymphocytic leukemia. Blood 2010;115:2578–85.

512 46. Lannutti BJ, Meadows SA, Herman SE, Kashishian A, Steiner B, Johnson AJ, et al. CAL-
513 101, a p110delta selective phosphatidylinositol-3-kinase inhibitor for the treatment of B-
514 cell malignancies, inhibits PI3K signaling and cellular viability. Blood 2011;117:591–4.

515 47. Herman SE, Gordon AL, Hertlein E, Ramanunni A, Zhang X, Jaglowski S, et al. Bruton
516 tyrosine kinase represents a promising therapeutic target for treatment of chronic
517 lymphocytic leukemia and is effectively targeted by PCI-32765. Blood 2011;117:6287–
518 96.

519 48. Shinohara H, Kurosaki T. Comprehending the complex connection between PKCbeta,
520 TAK1, and IKK in BCR signaling. Immunol. Rev. 2009; 232:300–18.

521 49. Bruno S, Ledda B, Tenca C, Ravera S, Orengo AM, Mazzarello AN, et al. Metformin
522 inhibits cell cycle progression of B-cell chronic lymphocytic leukemia cells. Oncotarget
523 2015;6:22624–40.

FIGURE LEGENDS

Figure 1. Diagram representation of proteins (A) and phosphoproteins (B) identified in the CLL (A-D) and MBL tumor B-cell samples (E) analyzed. This figure illustrates how many proteins were expressed in common or in only one or a subset of the 5 samples analyzed for the whole proteome dataset (A, 2,979 proteins in total) or the phosphorylated protein dataset (B, 327 proteins).

Figure 2. Quantitative proteomic expression data. On the left side in panels **A)** and **B)** the distribution of proteins found to be expressed at different levels per color-coded sample – sample A, red; sample B, blue; sample C, green; samples D, magenta; sample E, yellow - (ranking score in which 1 represents the highest levels and 5 the lowest levels per protein) is showing the number of ranks decreased because NA values did not get a rank value. In the right side, the distribution of the log2 expression values per sample (color-coded as in the left panels) is shown as box plots.

Figure 3. Heatmaps based on the percentage comparison of the expression values of the proteins identified and quantitatively measured in CLL/MBL tumor B-cell samples. The maximum value is set to 100% and the other values are percentages of this value. Colors gradually change from 100% (red) to 50% (white) and 0% (blue). Distinct clusters of proteins are color-coded in the first column on the left of each heatmap. Panel **A)** shows the results of the whole proteome data set and panel **B)** the phosphorylation dataset.

Figure 4. Combined graphical visualization of the circular graphical representation of chromosome 11 (A) and chromosome 13 (B) proteome profile and the (percentage-based) heatmap for the proteins involved in deletions of both chromosomes. Each graphic has four layers. The outer layer shows the cytoband of the chromosome and the number of nucleotides in the chromosome as positional information. The following layer below shows the results of the heat maps mapped to the chromosomal position of the genes. The third layer shows the gene names and where their chromosomal position. The last layer shows the area of the del11q23.3, del11q22.3, and del13q14, respectively. The (percentage-based) protein expression values identified for those proteins that are coded by genes located in the deleted

552 chromosomal regions are pointed with large arrows and shown in red to green color-coded
553 heatmaps.

554 **Figure 5. B-cell receptor (BCR) signaling.** Proteins and phosphoproteins involved in BCR
555 signaling are depicted; proteins and phosphoproteins (flagged with a yellow dot) expressed in
556 samples A+B, C+D, and E are colored in green, red, and blue, respectively.