Proteomics as a multifaceted tool in medicine and environmental assessment

Jacob Kuruvilla

Division of Cell Biology, Department of Clinical and Experimental Medicine
Faculty of Medicine and Health Sciences, Linköping University, Sweden

Linköping 2017
During the course of the research underlying this thesis, Jacob Kuruvilla was enrolled in Forum Scientium, a multidisciplinary doctoral program at Linköping University, Sweden.
‘The real discoveries have always come from chaos. There are only patterns, patterns on top of patterns, patterns that affect patterns, patterns hidden by patterns and patterns within patterns. What we call chaos is just patterns we haven’t recognized. What we call random is just patterns we can’t decipher. What we can’t understand, we call it as nonsense and what we can’t read; we call gibberish. There is no free will and no variables.’

- Chuck Palahniuk
SUPERVISOR

Professor Susana Cristobal
Department of Clinical and Experimental Medicine
Linköping University

CO-SUPERVISORS

Professor Mats Lindahl
Department of Clinical and Experimental Medicine
Linköping University

Professor Mikael Sigvardsson
Department of Clinical and Experimental Medicine
Linköping University

FACULTY OPPONENT

Professor Jean Armengaud
Laboratory Innovative technologies for detection
CEA-Marcoule
France
ABSTRACT

Proteomics is evolving as a multi-faceted tool for addressing various biochemical and biomedical queries in the field of scientific research. This involves various stages, ranging from sample preparation to data analysis and biological interpretation. Sample preparation involves isolating proteins from the sample source, purifying and digesting them to initiate shotgun proteomics. Shotgun proteomics identifies proteins by bottom-up proteomic approaches where proteins are identified from the fragmentation spectra of their own peptides.

Paper I: deals with the simplification of functional characterization for nanoparticles intended for use in biomedicine. Proteomics was constructive in differentiating and semi-quantifying the surface of protein corona. This could be beneficial in predicting the interactions between nanoparticles and a biological entity like the cell or a receptor protein and provide initial valuable information related to targeting, uptake and safety.

Paper II: deals with understanding effects of TiO$_2$ nanoparticles on endothelial cells. A combinatorial approach, involving transcriptomics and proteomics was used to identify aberrations in the permeability and integrity of endothelial cells and tissues. Our study also investigated the correlation of size and how they motivated a differential cellular response. In case of intravenous entry for nanoparticles in targeted drug delivery systems, endothelial cells are the first barrier encountered by these drug carriers. This evaluation involving endothelial cell response could be very instrumental during the designing of NP based drug delivery systems.

Paper III: Pharmaceuticals and its metabolites could be very hazardous, especially if its disposal is not managed properly. Since water bodies are the ultimate sink, these chemicals could end up there, culminating in toxicity and other ‘mixture effects’ in combination with other factors. To evaluate the effects of the pharmaceutical, propranolol and climatic factors like low salinity conditions, a microcosm exposure was designed and shotgun proteomics helped understand its impact on mussel gills. In this study too, a combination of transcriptomics and proteomics unveiled molecular mechanisms altered in response to stressors, both individually and in combination.

Paper IV: An interplay of various factors like EBF1 and PAX5 determines B-cell lineage and commitment. This might have been materialized by direct and transient protein-protein interactions. A unique method called BioID helped screen relevant interactions in living cells by the application of a promiscuous biotin ligase enzyme capable of tagging proteins through biotinylation based on a proximity radius. Biotinylation of endogenous proteins enabled their selective isolation by exploiting the high affinity of biotin and streptavidin on streptavidin coated agarose beads, leading to their identification by mass spectrometry. The biotinylated proteins were potential candidate interactors of EBF1 and PAX5, which were later confirmed by sequencing techniques like ChIP-Seq, ATAC seq, and visualization techniques like proximity ligation assay (PLA).
LIST OF ORIGINAL PAPERS

This thesis is based on the following papers, referred by their roman numerals in the text.

I. Surface proteomics on nanoparticles: a step to simplify the rapid prototyping of nanoparticles

II. MultiOMICs reveals effects of ultra-small TiO₂ nanoparticles on endothelial cells

III. Shotgun proteomics to unravel marine mussel (Mytilus edulis) response to long-term exposure to low salinity and propranolol in a Baltic Sea microcosm

IV. EBF1 and PAX5 associate with SWI/SNF and Histone modifying complexes to modify the epigenetic landscape in early B-cell development
    *Authors contributed equally to the work
### ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS</td>
<td>Mass Spectrometry</td>
</tr>
<tr>
<td>LC</td>
<td>Liquid chromatography</td>
</tr>
<tr>
<td>BioID</td>
<td>Biotin identification system</td>
</tr>
<tr>
<td>NPs</td>
<td>Nanoparticles</td>
</tr>
<tr>
<td>PC</td>
<td>Protein Corona</td>
</tr>
<tr>
<td>PROP</td>
<td>Propanolol</td>
</tr>
<tr>
<td>TiO₂</td>
<td>Titanium dioxide</td>
</tr>
<tr>
<td>CMC</td>
<td>Critical micelle concentration</td>
</tr>
<tr>
<td>FASP</td>
<td>Filter aided sample preparation</td>
</tr>
<tr>
<td>MudPIT</td>
<td>Multi-dimensional protein identification technology</td>
</tr>
<tr>
<td>HPLC</td>
<td>High performance liquid chromatography</td>
</tr>
<tr>
<td>RP</td>
<td>Reversed phase</td>
</tr>
<tr>
<td>m/z</td>
<td>Mass / charge</td>
</tr>
<tr>
<td>ESI</td>
<td>Electrospray ionization</td>
</tr>
<tr>
<td>CID</td>
<td>Collision induced dissociation</td>
</tr>
<tr>
<td>PC</td>
<td>Protein corona</td>
</tr>
<tr>
<td>DTT</td>
<td>Dithiothreitol</td>
</tr>
<tr>
<td>IAA</td>
<td>Iodoacetamide</td>
</tr>
<tr>
<td>AJs</td>
<td>Adherens junction</td>
</tr>
<tr>
<td>TJs</td>
<td>Tight junctions</td>
</tr>
<tr>
<td>GJs</td>
<td>Gap junctions</td>
</tr>
<tr>
<td>TPP</td>
<td>Trans proteomic pipeline</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>EBF1</td>
<td>Early B-cell factor 1</td>
</tr>
<tr>
<td>PAX5</td>
<td>Paired box protein 5</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS

**INTRODUCTION** 1  
1. Philosophy of ‘Omics’ 2  
2. Proteomics 6  
   1. Why proteomics? 8  
   2. Designing proteomics experiments 8  
   3. Sample preparation 9  
   4. Mass spectrometry 14  
3. Proteomics in nanomedicine and risk assessment 17  
4. Proteomics in environmental assessment 19  
5. Proteomics, B-cells and cancer 23  

**SUMMARY** 27  
Paper I 28  
Paper II 31  
Paper III 36  
Paper IV 39  

**REFERENCES** 45  

**CONCLUSIONS AND FUTURE PERSPECTIVES** 57  

**POPULAR SCIENCE ABSTRACT** 61  

**ACKNOWLEDGEMENTS** 65
1. Philosophy of ‘Omics’

The suffix -ome stands for ‘all constituents considered’. In molecular biology, ‘omics’ attributes to a combination of technologies that can be used to measure some characteristic of a large family of molecules like genes, proteins, lipids or metabolites. It includes making sense of a pile of information coming from the measurement of a particular -ome and relating it to a biological function [1].

‘Omics – a suffix signifying the measurement of the entire complement of a given level of biological molecules and information that encompasses a variety of new technologies that can help explain both normal and abnormal cell pathways, networks, and processes via the simultaneous monitoring of thousands of molecular components.’

Maria V. Schneider and Sandra Orchard
Omics Technologies, Data and Bioinformatics Principles

Omics has prompted a paradigm shift from a study design which starts with a hypothesis to an explorative design where preliminary data could define problems and help suggest the hypothesis. Furthermore, these omics technologies need to be high throughput, capable of generating huge amounts of data; be it genes (genomics), RNA and gene expression (transcriptomics), proteins and its abundance (proteomics), lipids (lipidomics), or metabolites (metabolomics). This data should be able to facilitate a system level understanding of correlations and dependencies between molecular components [2]. In order to materialize this, there is a need of techniques and methods that can handle extremely complex biological samples with high throughput, along with high sensitivity and specificity [2]. Simultaneously, there is also a need of high end computational power to analyze and combine these omics approaches.

The then, new discipline of genomics in 1988 was validated by the Human genome project 10 years later [3]. This kicked off a plethora of genome sequencing projects of organisms but later revealed the hidden complexity. To understand this, emphasis was laid on the relationship between the sequence and function leading to the birth of functional genomics [4]. The simplistic view of gene being the ultimate map of an individual rested in the deterministic concept of the gene which was extended throughout the genome. Based on sequenced genomes, the upcoming field of transcriptomics and proteomics unveiled elusiveness in the gene concept.

Sanger sequencing have been out performed by the latest technologies in both being high throughput and cost effective because of which the new term, ‘next generation sequencing’ came into being [5]. Transcriptomics deals with the study of RNA, transcripts produced by the genome, inclusive of mRNAs, non-coding RNAs and small
RNAs, where it identifies the structure of genes, their splicing patterns and other post-transcriptional modifications [6]. Techniques like microarrays were replaced by sequence tag-based approaches like SAGE (serial analysis of gene expression) [7]. Later came the age of RNA-sequencing which revolutionized the way transcripts were analyzed, by providing mapping and quantifying capabilities [8]. In contrast to genomics, transcripts were dynamic as their levels may vary during a given time in a certain developmental stage or due to external conditions.

Proteomics started its journey with the 2-dimensional electrophoresis (DE) gel used to map proteins from *E. coli*. This had to halt till Edman degradation could map proteins that were displayed. Proteomics had its most important breakthroughs by the development of mass spectrometry (MS) which identified molecules according to their mass to charge ratio. High throughput approaches have been in focus with the development of gel-free fractionation techniques, merging chromatography, along with very highly sensitive, robust mass spectrometers. Unlike genome, the proteome could differ between samples, because of it being highly dynamic and constantly adapting to changes, the cell or system it was exposed to. It also involves the study of splicing events and post-translational modifications which could answer many biological questions. There are several facts that prove the complexity from the gene to the phenotype: (a) just about 2% of human gene directly code proteins [9] (b) nearly entire DNA of eukaryotic organisms are transcribed [10] (c) noncoding RNA transcripts in higher organisms is thought to have important regulatory role [11] (d) spatially and temporally regulated post-transcriptional and post-translational regulation is important in the control of various functions of gene products [12].

The boundaries between these ‘omics’ studies have been growing into each other; thereby including the fields of interactomics which is related to interactions between the genome and proteome; between proteins and even proteins and metabolites [13] (figure 1). Recently developed technologies in all the ‘omics’ fields have led to the accumulation of vast amounts of data which was limited by ways to analyze and extract biological information. But very recent developments in bioinformatics have managed great strides towards extracting insights from these data by analyzing the properties and contents of different omics data sets, ultimately developing workflows to integrate various data types [14].

Despite the accomplishments omics have achieved, much remains to be figured out both in the technological and computational level. The high-throughput technologies corresponding to all omics generally come simultaneously with the lack of sensitivity, precision and selectivity. The difficulty of the task is to keep similar but different objects separate and to accurately measure concentrations in parallel that vary by several orders of magnitude [2]. Low abundance signals might not be detected or could be masked against the background. The combination of poorly understood biological variation, technological biases, artifacts and the high costs that in practice,
limit the number of replicates, challenge available statistical methods, ultimately limiting accuracy and confidence of the results [2]. Keeping technical issues apart, the biological concept could also be under question especially when there is a need to go beyond the scope of study. The completeness of the entity (gene, transcript or protein) has always been a question.

Figure 1: The central dogma that traces biological information from the genome to the phenotype. Components depict the molecular content of the system, interactions specify links between different molecular components while functional state depicts an integrated approach related to the phenotype. Genes are transcribed to mRNA and translated to proteins. The proteins functionalize to catalase reactions to form metabolites [14].

A combined transcriptomics and proteomics approach could give information about the pattern of RNA expression and the associated protein expression patterns could be mapped according to the central dogma. In many cases, non-correspondence between RNA and proteins exists because of different half-lives and post transcription machinery [15]. There are other factors that should be kept in mind when we study
the transcript profile. Studies have been considering average data from a collection of cells rather than a single cell approach. A global transcriptomics and proteomics approach would be the ideal solution to answer the biological question, though a combination on the same sample would not be feasible as they involve destruction of the sample, making it unusable for the other set of analysis.
2. Proteomics

In the mid 1970’s, a laboratory in Colorado reported a technique by which proteins could be resolved on a two-dimensional polyacrylamide gel and was described as ‘the sky on a starry night’. Just as Vincent van Gogh’s ‘starry night’ which depicts the view from the window of his asylum, just before sunrise at Saint-Remy-de-Provence. Van Gough was an intense depicter of natural phenomenon, knowing or unknowing could incorporate turbulence or chaos in his paintings especially when he was most psychotic. The cover page depicts the amount of order and chaos when proteins are cleaved to peptides or when protonated peptides are fragmented. Chaos exists in every deterministic nonlinear system where very small perturbations end in bigger diverging outcomes also called ‘butterfly effect’. In layman’s terms, a chaotic condition is normally meant to be uncontrollable and unpredictable because of many degrees of freedom that has to be dealt with simultaneously [16]. The task would be to locate and isolate the critical points against the background. A chaotic system is characterized by nonlinear dynamics where the change of output is not proportional to the change of input making it unpredictable in contrast to linear systems [17]. This has great pertinence in biological systems and different -omics where a large number of epistatic, epigenetic and external factors like environmental factors and stressors act in cohesion concluding in differentiation, development and death.

Life is the mode of action of proteins.

-Friedrich Engels

Proteins are life giving molecules, determining the function, structure, regulation and fate. They are made of 20 amino acid base pairs or the protein building blocks, and when attached together forms polypeptide chains. These chains attain a three-dimensional structure to form a protein with a structural or functional role. Proteomics, according to the classical definition is the large-scale characterization of the entire protein complement of a cell, tissue or organism; but presently, this omics study has taken up a more global and integrated perspective including post-translational modifications, protein-protein interactions and protein turnovers [13]. One of the most interesting feature of proteomics is its chaotic nature. The delicate variations in the ecosystem is translated to the phenotype which is reflected in the genotype only through proteomic plasticity.
Figure 2: In top-down proteomics, intact proteins are fractionated by either liquid or gas phase chromatography and introduced into a mass analyzer where it is fragmented and later identified. Meanwhile the bottom-up approach deals with digesting one or a mixture of proteins into peptides. The mixture of peptides is fractionated and analyzed by either peptide mass fingerprinting or tandem MS/MS.

There are two fundamental strategies employed in proteomics based research namely: top-down and bottom-up approaches (figure 2). In the top-down approach, intact proteins are analyzed. This decreases the complexity and preserves all information related to the status of the intact protein, including PTMs and sequence variations arising from mutations, truncations, and alternative splicing events but, as the physiochemical diversity of intact proteins is much greater than that of peptides, large-scale separation of intact proteins is challenging and, thus, traditional top-down studies have primarily focused on the analysis of a single or small number of proteins, typically obtained via affinity purification [18] [19] [20]. Meanwhile, bottom-up proteomics deals with proteins digested to peptides and information obtained through MS analysis of the resulting peptides is used to infer the identity, quantity, and modification status of the proteins present in the sample [21]. Bottom-up techniques leads to a mixture of peptides produced by proteolysis which makes it difficult to determine to which protein isof orm the peptide belongs to; but this has proven to be more robust and high throughput for identification and quantification [21]. Thus they are rightfully called bottom-up shotgun approach or multidimensional protein identification technology (MudPIT) [22]. Bottom-up proteomics involves order to chaos (cover page) where tissues are broken down to proteins which are even broken down to peptides and then ionized to molecules.
2.1 Why proteomics?

According to the universal code, genes are transcribed into RNA and then translated into proteins. Due to various genetic predispositions and environmental factors, the processes ending up in functional proteins are being modified at many levels. While genomics involves the study of 30,000 genes, transcriptomics deals with many more number of transcripts and proteomics, a million different proteins. Genomics does have an inventory of all genes of many organisms, mainly because of the fact that they are stretches of deoxy ribonucleic acid polymers made of four bases. This polymer acts as the ‘blueprint of life’, out of which proteins or the ‘building blocks’ are fabricated. So, the complexity at the root of life is daunting. Fortunately, state of art technologies along with very many tools capable of identifying and quantifying proteins enables us to measure the functional expression of the genes involved. This allows gene-transcript-protein-disease correlation and to study the phenotypic expression changes involved. The genome is normally static except when a mutation occurs, while the proteome is highly dynamic changing in response to both external and internal stimuli. Plants and animals are the most efficient machines that have existed on the face of earth. This is because, the information for coding all proteins is present in the genes but not all are expressed concurrently.

The real power of MS based proteomics is revealed when they are used in conjuncture with the latest molecular biology techniques classified under genomics, transcriptomics and site directed mutagenesis. One should go beyond the limits of engaging in an orgy of sequencing proteins and make use of all disciplines of science to answer the biological question. The ultimate goal of proteomics is deciphering protein structure and function, how they do it, with whom they interact and how they contribute to various cellular and molecular processes.

2.2 Designing proteomics experiments

Designing experiments in proteomics are very crucial, especially because a bad design would lead to an exceptional rate of false discoveries. Proteome characterization deeply depends on the dynamic range and detection sensitivity of experimental designs and results are generally reported as a collection of identifications with no information on the fraction of the proteome that was missed; thus they are difficult to evaluate and could be potentially misleading [23]. Therefore, the experimental design is critical for the success of a proteomics experiment and a good design must handle the complexity and the very wide range of protein abundances in question [23].

The first step towards setting up a proteomics experiment is to understand what the inputs and outputs to the process are, what their range is, and what input parameters one can control and secondly, in order to be able to perform a sound statistical
analysis after the experimentation is carried out, it is important to understand the process that has generated the data to falsify the research hypothesis [24]. Figure 3 depicts an overview of various steps involved in designing a proteomics workflow and relate considerations at the initial building of the hypothesis to selecting the right technology, pipeline and experimental design.

2.3 Sample preparation

Sample preparation is one of the most vital steps in proteomics as the reproducibility and quality impacts all downstream processing. It is vital for the generation of accurate and informative data. This has become the bottleneck in many proteomic experiments since advancements in downstream processes like fractionation and analysis soar high. Various factors of utmost importance included in sample preparation are sample harvesting, storage, handling, transportation, extraction and fractionation.

Protein extraction and purification is much more complicated compared to DNA or RNA because of their complexity. Proteins are heterogeneous with respect to their size, structure, pH, hydrophobicity, charge etc. [25]. Some of them are present in abundance while others in very less amounts. In addition, post-translational modifications, degradation, and high affinity for non-peptide entities makes proteomics even more challenging. Since proteins are very diverse in their physical and chemical properties, it is virtually impossible to devise a single strategy for extraction and preparation.
The colossal challenge during preparation arises from the fact that the proteome is a collection of 30 to 80% gene products expressed at both low levels (10 – 100 copies per cell) to high levels (10 000 – 100 000 copies per cell) and therefore they represent a dynamic range of at least 6 orders of magnitude in living cells, and about 10 to 12 orders of magnitude in body fluids [25]. The top ten most abundant plasma proteins account for approximately 90% of the total protein content in plasma, while all other proteins are present in the mere 10%, spanning more than 10 orders of magnitude in terms of concentration [26]. This is referred to as ‘masking effect’ created from abundant proteins by which peptides from low abundant proteins are hidden [27].

Currently, several fractionation methods for protein depletion are available but these approaches require relatively high amounts of starting material to ensure efficient enrichment of low-abundant proteins; otherwise, high- and medium-abundant proteins would not fully saturate their ligands and ultimately the elution would have the same profile as initial sample [28] [29].

Although any single protein preparation technology is unlikely to be ideal for all samples, the factors that should be taken into consideration are:

1. Efficient extraction and simplification of sample
2. Reduction of the dynamic range of protein concentrations
3. Removal of impurities or other interfering substances like nucleic acids, lipids, particulates, salts etc.
4. Compatible with downstream processing
5. Avoid any kind of unwanted modification of target proteins
6. Cost

**Cell lysis and solubilization**

The first step involves harvesting samples whether it is the protein corona attached on the surface of nanoparticles, gill tissues from aquatic mussels, endothelial cells or B-cells. The main aim of this step is to get rid of impurities and make the proteins in the sample accessible. After harvesting, the samples are broken down or lysed to extract the proteins. For this purpose, we have employed a chemical lysis extraction buffer in combination with some mechanical stimulus capable of breaking apart the cell and cell structures allowing the buffer to easily solubilize the proteins. The choice of buffer composition and mechanical stimulus depended on the sample source, sample size, composition and downstream process.

Proteins are polymers made of 20 amino acids. Each of these amino acids possess a basic structure with an amino group, carboxyl group, an alfa carbon and a R group or
the side chain which differs with different amino acids. Based on the nature of R groups, amino acids could be polar or non-polar. These amino acids are joined together by covalent bonds called peptide bonds formed between the carboxyl group of one amino acid to the amino group of another forming the primary structure. The primary structure fold and combine to form the secondary structures like an alfa helix or a beta sheet with the help of hydrogen bonds. When secondary structures coalesce, 3-dimensional tertiary structures materialize in presence of hydrophobic, ionic and disulfide bonds along with weak Van der Waals interactions (33). Non-covalent interactions (ionic, hydrogen and hydrophobic) could be disrupted in presence of a hydrophobic environment (SDS buffer, chaotropes like urea). Chaotropes alter solvent properties and exert profound effects on molecule interactions by unfolding hydrophobic protein cores, disrupt hydrogen bonds and hydrophobic interactions. This is made possible by these chemicals by decreasing ionic interactions, denaturing and altering the dielectric point [30]. However, the use of urea could lead to the formation of inclusion compounds, precipitation in presence of detergents or formation of ammonium cyanate with increased temperature or carboxylation due to change in pH [31].

Detergents are also widely used in cell lysis buffers where they play a pivotal in extraction, purification, and manipulation of membrane proteins; their amphiphilic nature allows them to interact with hydrophobic membrane proteins to keep them water-soluble outside of their native bilayer environment [32] [33]. There has not been a golden standard for using the kind of detergent suitable for a specific sample type. All these processes of lysis, dissolution and extraction can be achieved at the same step along with mechanical stimulus like vortexing or sonication. Lysis is normally followed by either sub-fractionation (ultra-centrifugation) of cellular particles followed by solubilization by detergents or removal of undissolved / unbroken cellular compartments and membrane structures by centrifugation (normally low speed, upto 10,000 g). Detergents are not compatible with downstream processing; therefore, it is of the highest priority to have them removed. Some detergents will interfere with enzymatic digestion, and most will interfere with reverse-phase separations and mass spectrometry, sometimes damaging instruments and irreversibly ruining columns [34]. Removal of unwanted cellular material, such as lipids and genomic DNA prevents signal suppression, chromatographic interference, and presents a much cleaner and clearer spectrum [35]. Removal could be achieved by precipitation, ultra-filtration or by washing. Precipitation is not the optimal technique to purify micro-samples from detergents or contaminants as they might lead to disastrous loss of sample material. The loss of sample could occur during precipitation or during dissolution of the pellet.
Care should be taken in case of microscale proteomic samples where there could be catastrophic sample losses while expunging the detergents. The same goes for selection of mechanical stimulus, where the most appropriate should be chosen intending minimum sample losses. Therefore, it is ideal to use methods with the least surface, like an immersion sonicator or resort to vigorous vortexing, thereby minimizing loss and maximizing protein concentration.

**Fractionation**

Nevertheless, sensitivity and resolution of today’s state of the art technologies have reached sky high to detect low abundant and variable species. However, reproducibility and variable dynamic ranges, of highly complex samples pose a humongous challenge. Highly abundant proteins cause ‘the suppression effect’ which could make it difficult, rather impossible for the detection of low represented molecules [36]. In order to break down the complexity and enable enrichment of low abundant species, fractionation is of prior importance. This is possible at different levels, either at the organelle level, protein level or even peptide level.

When tissues and cells are considered, organelle level separation could be achieved by cell fractionation techniques. Care should be taken to break up either inter-cellular structures in case of tissues or plasma membrane in cells. Vortexing, sonication, osmotic shock or by forcing cells through a needle could break up the cell membrane. Different components of the cell can be separated on basis of density by ultracentrifugation at different speeds. Sucrose gradients could be used to achieve a finer degree of separation. Paper III fractionated gill tissues into cytosolic, membrane and nuclei cell fractions before protein digestion.

The classical approach for protein separation is based on molecular weight where proteins are separated on a denaturing gel. Simultaneously, this could also aid in protein purification. One of the main impediments associated to this technique has been the incomplete recovery of peptides by in-gel digestion. This also possess insufficient resolving power to separate huge number of proteins present in biological fluids. To counter these flaws, affinity purification, immunoprecipitation, pull-downs and chromatography have been considered in our study.

**Filter assisted sample preparation (FASP)**

Care should be taken while manipulating microscale samples as mentioned above so as to minimize losses by the least handling possible. This makes filter aided sample preparation or FASP attractive. It is an efficient means for a wide range of proteins and shows great efficacy in washing away contaminants and detergents. FASP is
known for its flexibility as it could be used in conjunction with other preparations and methods. Paper II utilizes this method for its high throughput and reproducibility, while paper III uses it downstream after cell fractionation. In both the studies, proteins were effectively solubilized using SDS; afterwards exchanged with urea buffer (helps maintaining proteins denatured in solution, chaotropic agent as mentioned above), whose concentration is lowered for effective trypsinization. Since the proteins are being immobilized on a membrane, it also helps in concentrating proteins samples. Utilizing less amount of sample material helps us minimize costs too.

**Protein Digestion**

Trypsin is an endopeptidase, secreted in the small intestine that breaks down proteins or polypeptides into shorter chains called peptides. This has been the most popular choice, even though other enzymes and chemicals exist for digestion. This is because, it cleaves proteins into 700 – 1500 Da peptides which is ideal for MS analysis by hydrolyzing peptide bonds on the carboxyl side of arginine and lysine residues except when followed by a proline [37]. It is also considered to be comparatively stable, and potent under different conditions. The limitations associated with trypsin are that it works optimally under a pH range of 7.5 – 8.5 and does not survive strong detergents like SDS. In addition, trypsin leaves undigested sites especially at lysine residues. To counter this issue, a cocktail or sequential digestion with Lysine-C could be effective. Protein folding also prevents effective digestion for which proteins are reduced (reducing agents like dithiothreitol, DTT) and later alkylated by iodoacetamide (IAA). C18 ziptips could be utilized for concentrating and desalting samples prior to MS analysis [38].

**Chromatography**

Chromatography could be exploited at different levels, either at the protein level or the peptide level. In papers I to IV, liquid chromatography (LC) has been exploited to fractionate peptides. The LC was coupled on-line to an automated MS/MS setup, named MudPIT (multi-dimensional protein identification technology) [39]. Reversed phase (RP) HPLC of peptides are accomplished in the presence of an ion-pairing and organic modifier reagent [40]. RP-HPLC has been considered the apt tool because of [41]

a. Excellent resolution for closely related molecules.

b. Experimental ease as it does not need any clean-up and could be connected on-line to a mass spectrometer (MS).

c. Comparatively high recoveries and productivity.
Tandem MS was made possible after the introduction of ESI (electrospray ionization) which allows the direct introduction of peptides into the MS as they are eluted from the liquid chromatographic column. It has been associated with some drawbacks, especially when the sample preparation involves detergents. Peptides with less than 30-40 amino acid residues could be separated according to their hydrophobicity [42]. We have employed a RP-HPLC setup by the sequential elution from a preparative C18 (n-octadecyl) column, followed by an analytical C18 column.

2.4 Mass Spectrometry

Misnomer to what it is named as, a mass spectrometer does not measure the mass; rather the mass to charge ratio (m/z value) of ions or even some parameter from which m/z can be extrapolated. This is normally represented as mass spectrum (Daltons per unit charge) which is a plot with ion abundance versus m/z value of respective ions. From this information, molecular mass can be determined, revealing the molecular composition.

There are 3 principal components in a MS:

1. Ionization source: the analyte is transformed into gas phase ions in vacuum by the ionization source. This is done in two sequential steps, where it is first volatilized and then ionized. This approach limits the ionization process and the types of analytes that can be analyzed intact to relatively low-molecular-weight compounds that are thermally stable [43]. There are different methods for ionizing a neutral molecule like electron capture, electron ejection, protonation, deprotonation, cationization or by transferring a charged molecule from a condensed phase to the gas phase [44]. Peptides are ionized usually by protonation, by which a proton is added, resulting in a net positive charge. The different sources for ionizing non-volatile, thermally labile compounds are ESI (electrospray ionization) or MALDI (Matrix-assisted laser desorption/ionization).

ESI utilizes electricity for assisting the transfer of ions to gas phase from the solution where neutral compounds are protonated or cationised [45]. This is accomplished by (1) dispersal into droplets (2) evaporation of solvent (3) and ion ejection from the charged droplets. Ensuring a high sample flow rate by a nebulizing gas like nitrogen, the charged droplets pass down a pressure-potential gradient, leading to a decrease of droplet size and increase in surface charge density, ultimately reaching the critical point where surface ions on the droplets are ejected into the gas phase [45].

2. Mass analyzer: the precision of measurement in MS instruments is dictated by the ability of an analyzer to resolve two adjacent peaks called resolution.
Analyzers could be broadly divided into either beam or trapping ones. The former analyzer deals with passing the ions in a beam through the analyzing field into the detector; while the latter traps ions [43].

The ion trap mass analyzer possesses advantages of being compact and cheaper, where electrodes form a cavity making it possible to trap ions. The voltages applied to the electrodes could be controlled to trap and eject ions according to their m/z values [45]. While performing tandem MS using an ion trap, fractionation could be implemented by collision induced dissociation (CID). This is a relatively soft ionization technique where a kinetically excited precursor molecule collides with a buffer gas, transferring internal energy to impart breakage of bonds, leading to the formation of product ions [46]. Thereby, a fragmentation spectrum is generated for each product ion.

(3) Detector: the fragment ions are electronically analyzed in the detector. The Orbitrap is an axially-symmetrical mass analyzer consisting of a spindle-shape central electrode surrounded by a pair of bell-shaped outer electrodes. It traps ions electrostatically around a central spindle-shaped electrode and then analyzes their m/z values as they move across the spindle with different harmonic oscillation frequencies [47].

The Orbitrap Velos Pro mass spectrometer was composed of a Velos pro MS and an Orbitrap Analyzer. Ionization by CID and fragmentation is conducted in a linear ion trap. It has a curved linear trap for immediate storage of ions which injects similar m/z ions in short pulses into the Orbitrap mass analyzer.

Data processing and analysis

A RP-HPLC online with a tandem MS rapidly generates fragment ion spectra or a CID spectrum (the choice of fragmentation was CID) as soon as the analytes elute form the C18 column based on their hydrophobicity. This kind of a data-dependent MS/MS analysis calls for detection of peptide ions of a specific m/z value in the MS scan called precursor ions. An MS/MS spectrum or the fragment ion spectrum is obtained when each precursor ions are isolated and fragmented by CID. A mass spectrum would be generated against the CID induced peptide fragmentation depending on parameters of CID along with other sequence specific characters. This results in a graphical display called mass spectrum with relative abundance or intensity of ions against the m/z values [45].

A protein database search engine like SEQUEST analyses in the spectrum domain where it generates virtual spectra from database sequences to find the best match against the observed spectrum [48]. SEQUEST first selects peptides from the database based on the precursor-ion mass. According to the enzyme used, the algorithm
creates a list of m/z values at which the fragment ions occur. The spectrum was searched for ions at these locations and is summed up, along with a bonus for successive fragment ions. These factors with the percentage of ions found versus the expected are summed up to produce a score [48]. It also generates values like XCorr (strong measure of accuracy), coverage and DelCn. High XCorr values indicates the comparison to be significantly greater than the background and therefore, the 1st hit would have the highest XCorr value. The coverage is a percentage of the ratio between number of amino acids found in the peptides and the total amino acids in the protein sequence. DelCn ≥ 0.1 is good but a crude calculation which give a relative idea about how different the first hit is from the second.
3. Proteomics in Nanomedicine and risk assessment

There has been a tremendous increase in the use of nanoparticles (NPs) during recent years, be it in nanomedicine, electronics, cosmetics, food, packing, construction etc. This large-scale increase in usage has also increased the exposure of humans and environment to NPs. Therefore, it is of utmost importance to understand the interactions of these NPs with biological systems. When a NP enters a biological media, there would be competition between different biomolecules to get adsorbed on its surface. Blood contains about 1000 different proteins with abundances ranging about 12 orders of magnitude [49] along with other biomolecules like lipids. According to Vroman’s effect, this competition results in the adsorption of the most abundant proteins, which are later replaced by proteins with the highest affinities [50]. This layer of proteins adsorbed on the surface of NPs is called the protein corona (PC). The PC is composed of different layers, the first being those that are irreversibly fixed called the hard corona, while the loosely fixed outside layers are called the soft corona [51]. Besides size and shape, the other defining element of NPs in a biological medium is the hard corona [49], which could significantly evolve when they pass through different compositions of proteins, according to the above mentioned Vroman effect making corona a non-static entity. Factors affecting the evolution of PC include the physiochemical properties (shape, size, surface charges, composition, surface functionalization, etc.), nature of biological composition, and duration of exposure [52]. Evidently, the PC transforms the size of NP along with the interfacial composition leading to revamping of bio-physiochemical properties.

Qualitative and quantitative evaluation of risks posed to the environment and humans by the presence of actual or potential pollutants is termed as health risk assessment [53]. A comprehensive risk assessment considers the exposure to the hazard, risk to benefit ratio, effects due to the exposure and to an extend explain uncertainties [54]. A novel sub-field in proteomics called toxicoproteomics focuses on these changes to predict potential effects, explain mechanisms of action decreasing uncertainties, identifying toxicity, its levels and effects, and even cross-species extrapolation [55]. There are various strategies employed in toxicoproteomics like protein expression profiling where molecular signatures in terms of relative protein levels are analyzed on a global scale. With the advent of state of the art techniques, minor perturbations in protein structures and protein-protein interactions could be tracked and attributed to the stress response and therefore to toxicity. Besides, proteomics enables it to be high throughput aiding analysis at different concentrations and time periods. Even though, this protein profile generated is the linchpin of risk assessment, it provides very less information related to the global effects on cell, tissue, organism or ecological level. However, bioinformatics aids proteomics in generating maps which could extrapolate the protein patterns to biological functions.
Since corona plays an important role in toxicity, it could also be an important factor in nanomedicine. Graphene oxide, when incubated with fetal bovine serum (FBS) was found to have lower cytotoxicity [56]. Meanwhile, silver nanoparticles in human blood turned out to more cytotoxic with the PC [57]. Apart from risk assessment, PC studies could hugely benefit the development of target specific NP drug carriers. Functionalization by ligands specific to the target are designed on the NP surface, which could be blocked by the proteins composed in the PC [58].

Cancer represents one of the best examples of a disease where the adequacy of delivery of chemotherapeutics with highly potent, yet toxic, mechanisms of action can mean the difference between efficacious responses and severe morbidity [59]. Traditional kinds of treatments include surgery, chemotherapy, radiation, or immunotherapy [60]. Other than that, a treatment that precisely target cancer cells are being developed called targeted therapy based on the differences between normal and cancer cells [61]. The main requirements for effective targeted drug delivery systems are: retain (pharmacokinetics and efficient drug loading), evade (biological barriers), target (bio-distribution), activity and release [62]. Targeted therapy focuses on specific kinases or receptors that are overexpressed in cancer cells or tissues. However, clinical results for many of those molecular target drugs have been disappointing [63-65]. Therefore, there is an increasing need to focus on PC, which might be the determining factor in resolving issues related to NP based drug delivery.
4. Proteomics in environmental assessment

Studies in environmental research have tried to link ecological function to protein diversity in complex ecological communities. Advanced technology in high throughput proteomics and bioinformatics, strengthened with appropriate statistics could help cataloging proteins and correlate with corresponding changes in the environment. The proteome is a dynamic interplay of proteins which respond to rapid environmental cues even without the need of transcriptional activation in many cases accomplished with post translational modifications (PTMs) and protein-protein interactions (PPIs) [66]. Thus, studying the proteome at a specific time would help us translate environmental conditions to subcellular biochemical responses. Despite the perks of proteomics as a tool for environmental assessment, there exist a number of obstacles ranging from sample collection to data analysis. The absence of fully sequenced databases for non-model organisms makes it challenging for identification, characterizing PPIs and changes in PTMs.

Blue Mussel (*Mytilus edulis*) is an ecologically important and commercially relevant filter feeding bivalve [67], whose gills are the major site of interaction with the outside environment, mediated by the creation of water currents in the pallial cavity [68]. Apart from gas exchange, food is also trapped which are sorted and transported to the mouth via the gill epithelia [69]. Mussels can filter up to 5 liters of seawater an hour exposing their gills to stressors and pollutants [67]. As a general biological defense of many marine and freshwater organisms against environmental toxicants, mussels are highly tolerant to xenobiotics, which tend to accumulate in large concentrations [70] [71]. In a study involving sediments, amphipods and algae, the relative uptake declined with increase of exposure (100 – 1000 µg/L), whereas the mussels exhibited the contrary with larger accumulation [60].

In addition, mussels are sessile and found attached to rocky substrates, allowing spatial localization of marine pollution and reflecting changes in the contamination of the environment from which they originate [67]. They are exposed to both dissolved and particulate forms of contaminants [72] [73]. These characteristics make them ideal candidates for bioindicators of marine pollution and could be very useful as a tool in ecotoxicological and proteomic studies. These organisms are used as sentinels in a number of biomonitoring projects including the mussel watch project which is the longest running continuous contaminant monitoring program in U.S. coastal and Great Lakes waters [74]. Many indicators based on general biomarkers (lysosomal stability, peroxisome proliferation etc); specific biomarkers (acetylcholinesterase activity, or metallothionein content) or genotoxicity biomarkers like micronuclei or
DNA damage have been exploited before in bivalves [71]. In mussels, cells exposed to pollutants responded by peroxisomal proliferation and this was developed into a proteomics-based method for marine pollution assessment [75]. Moving from a reductionist approach, high throughput methods with the introduction of 2-D gels and LC [76] have enabled to simplify the technical complexity and gain robustness for its applicability to marine pollution monitoring, especially tracking ecological disasters such as large marine oil spills [77]. However, the lack of sequenced genomes from sentinel organisms and environmental models have seriously delayed the development of the methodological opportunities and the acceptance of established ones by stakeholders and regulators or OECD guidelines. With this study, we aim to coalesce different omic approaches to effectively track biomonitoring in the molecular level.

![Diagram of pharmaceutical compounds](image-url)

**Figure 4**: The origin and fate of pharmaceutical compounds in the environment [78].

**Pharmaceuticals as pollutants**

Many pharmaceuticals have been identified in the environment and this has been a growing concern for both human health, and the ecosystem. These substances have been found in effluents from hospitals, sewage, sewage treatment plants, in surface water, ground water, and even in drinking water (Figure 4) [78]. Beta-adrenergic
blocking agents (β blockers) are highly used, prescription drugs used to treat cardiac illnesses or disorders such as hypertension and arrhythmia [79]. Like many other pharmaceutical compounds, these beta-blockers too are partially metabolized in the human body and are released into the environment as a mixture of unaltered molecules with its metabolites along with human waste [80]. Beta blockers are important pharmaceutical compounds for which little environmental adsorption data exists [81]. Like many pharmaceuticals, beta-blocker compounds have the potential to be highly persistent [82, 83] and toxic in the environment. They are normally found in very less concentrations in the range of ng/l to µg/l [84].

The ultimate fate of these compounds in the environment could depend on many factors, including their adsorption rates on sediments, soil or biological surfaces. This could affect their transport, degradation time and even reduce chances of exposure to these compounds. Hydrophobicity was found to be a good measure of adsorption and propanolol (PROP) was the one of the most hydrophobic beta blockers studied (compared with metoprolol and nadolol) [81]. PROP is a commonly used β-blocker that inhibits β1 and β2 adrenergic receptors in the heart, aimed at decreasing heartbeat and contractility. It is considered to be fairly persistent [85], bioaccumulative [86], water soluble and possess low volatility [87].

Effects of PROP

PROP is hydrolyzed in aqueous medium to give 1,4-naphthoquinone and a remaining side chain [88]. These derivatives exhibited lower affinities than the parent compound, when they were examined for their affinities to β-adrenergic receptors in rat cerebral cortex membranes [89]. Survey on effluent water in the Baltic sea runoff area have detected PROP in concentrations ranging from 0.01 to 0.03 µg/L [85] while in other water bodies, it had been found as high as 0.29 g/L [90]. PROP, when exposed to concentrations from 12.5 µg/L to 125 µg/L to zebra fish embryos and sea urchin embryos from 1.25 to 12.5 µg/L induced significant mortality rates and all concentrations above 5 µg/L resulted in statistically significant abnormality rate for both groups [91]. As mentioned in paper III, mode of action of PROP may be similar to an extent with both humans and mussels because of evolutionary conservation of molecular targets [90].

A serendipitous discovery of PROP gave great promise for the treatment of infantile hemangiomas where it caused vasoconstriction; decreasing expression of the genes for vascular endothelial growth factor (VEGF) and basic fibroblast growth factor (bFGF), which contribute to angiogenesis; and triggering apoptosis of capillary...
endothelial cells [92]. However, the responses have been found to vary in perivascular cells and hemangioma derived stem cells [93].

PROP is extensively metabolized into propranolol glucuronide, naphthyloxylactic acid glucuronic acid, and sulfate conjugates of 4-hydroxy propranolol with most of them appearing in the urine. Clearance of propranolol is reduced with aging due to decline in oxidation capacity [28]. Blood levels and/or toxicity of propranolol may be increased by co-administration with many drugs and inhibitors. Studies have shown interactions with many cardiovascular drugs and calcium channel blockers [28]. PROP’s metabolism involves multiple pathways in the cytochrome P-450 system (CYP2D6, 1A2, 2C19), co-administration with drugs that are metabolized, or affect the activity (induction or inhibition) of one or more of these pathways leading to clinically relevant drug interactions [29].

In addition to pollutants, there are other abiotic factors like salinity that could affect growth, behavior, filtration activity, development and feeding of filter feeders like mussels [28]. The inner Baltic sea has been found to have belts of low salinity and a marked decline in the size of *M. edulis* was observed in those areas [29].

**Why Microcosm?**

Microcosms or miniaturized ecosystems are artificially assembled communities allowing investigation under controlled conditions [94]. Standardized, multi-trophic level microcosms can be a very useful tool in risk assessment [95]. Reproducibility, both within and between laboratories provides the extra edge compared to other methods and taxonomic uncertainties could also be avoided [95]. One of the biggest advantages of a microcosm is its relative degree of complexity, where it is complex enough to exhibit interactions between and within trophic levels, meanwhile keeping it simple in comparison to natural systems. Furthermore, being in the laboratory makes it permissible to introduce various factors that might not be present in the natural ecosystem especially ‘mixture effects’ with either pharmaceuticals or other factors. The risk of unpredicted, unregulated ‘mixture effects’ is low during their intended use but since the potential for their release into the environment is high, there is a risk of such chemicals being components of toxic mixtures [96] [97]. In the meantime, this could be detrimental in case of unpredictable factors, which could be countered only in the natural ecosystem. In addition, a proteomic risk assessment or a toxicoproteomic analysis using a microcosm helps multi component mixture analysis on the cellular and biological level.
5. Proteomics, B-cells and cancer

Acute lymphoblastic leukemia (ALL) in children is one of the most common malignancies which leads to abnormal lymphocytes or white blood cells (WBC) in the bone marrow. It has been found to occur most commonly, either under the age of 15 or over the age of 45 [98]. ALL progresses by replacing healthy lymphocytes with defective ones that couldn’t mature properly, later invading other parts of the body through the bloodstream.

Hematopoietic stem cells (HSCs) have the ability to differentiate into various specialized progenies [99]. They undergo autonomous self-renewal by dividing symmetrically, or differentiate into lineage-specific cells upon asymmetrical division in accordance with the environmental cues that they receive [100]. These cues that they receive might be coordinated by growth factors, neighboring cells or the local microenvironment called niche where stem cells resides [101]. The yolk sac is the first site of hematopoiesis in both murine and human hematopoietic ontogeny and from there, hematopoiesis progresses to the aorta-gonad-mesonephros region and the fetal liver, before shifting to the bone marrow, where it functions as the lifelong reservoir of HSCs [101]. HSCs give rise to lymphoid and myeloid progenitor cells, which produce all the mature blood and immune cells of the body [102]. They undergo restriction on stage and lineage as they are driven down through specific pathways leading to distinct cell types where differentiation take place irreversibly. Studies have shown that primitive hematopoietic progenitors are uniformly scattered throughout the bone marrow [103]. B-cells develop in the bone marrow from hematopoietic stem cells (HSCs) before they migrate to peripheral lymphoid organs through the blood. The mature B-cells following the activation by antigen in peripheral lymphoid organs develop into end-stage memory B-cells or plasma cells and returns back to colonize the bone marrow [104]. Approximately 0.01% of total nucleated cells in the murine bone marrow are HSCs, hence it becomes imperative in B-cell research to purify HSCs and the different cell stages of HSC differentiation; which could be made possible based on expression of cell surface markers by magnetic-activated cell separation or multi-parameter fluorescence-activated cell sorting [105].

B-cells

B-cells are highly specialized systemic cells which are an intrinsic part of the adaptive or acquired immune system found in vertebrates produced continuously throughout the life. These cells perform a number of critical immune functions apart from producing antibodies like generating immunological memory, antigen presentation and regulatory cytokine production [106]. B-cells differentiate into memory B-cells and plasma cells which are responsible for humoral immunity by producing antibodies. They have a complex life cycle, and it is important to consider their developmental pathways to understand the ramifications of different intervention
points [107]. B-cell development has been studied extensively in mice; although not so different in humans. Earlier studies have illustrated the similarities of humans and mice in B-cell development, except the dependence of pro- and pre-B-cells on Interleukin-7 (IL-7) for survival and differentiation [108].

B-cell development occurs within a complex bone-marrow microenvironment which is found in the medullary cavities of bones along with adherent cells called stromal cells, which create microenvironmental niches that maintain blood-cell viability and supply the requisite factors for their development [96]. Upregulation of the surface marker, CD19 marks the transition of the pre/pro B-cells to pre B-cells along with the rearrangement of immunoglobulin (Ig) heavy chain gene locus [109]. The IgM+ cells leave the bone marrow to lymphoid organs where they further develop into mature B-cells.

Transcription factors and regulatory surface receptors including IL-7, Fms-like tyrosine kinase 3 ligand (Flt3-L), PU.1, IKZF1 (IKAROS family zinc finger 1), TCF3, EBF1 (early B-cell factor 1), PAX5 (paired box gene 5) and IRF8 (interferon regulatory factor 8) play a vital role as signaling molecules for the B lineage commitment along with maintenance of cell fate [6]. Preceding the differentiation of CLP, low levels of transcription factor PU.1 [110] is involved in expression of IL-7 signaling components which are essential for EBF1-dependent lineage restriction [111, 112] (Figure 5). Studies have shown that IKZF1 upregulates growth factor independent protein-1 (Gfi1) which in-turn was responsible for the down modulation of PU.1 [113]. B-cell lymphopoiesis is marked by the initiation of transcription of genes associated with committed B-cells including the E2a and Ebf1 genes encoding two essential transcription factors, TF3 and EBF1 [114]. Both factors cooperate in the upregulation of genes like Pax5, Rag1, Rag2, IgH1 and Vpreb [115-117]. Other direct targets of EBF1 are Foxo1 and Pou2af1 [118]. The CLP stage still retain T-cell development potential where the levels of PAX5 is still low [119]. Later in pro B-cells, expression levels of PAX5 escalates and leads to repression of Notch1 expression crucial for maintained T-cell potential [120]. PAX5 acts downstream of EBF1 and TCF3, and controls pre-BCR formation by promoting V to DJ recombination and is essential for upregulation of CD19 and Ig gene expression [121, 122]. Overexpression of EBF1 promotes commitment to B-cell lineage even in PAX5 deficient pro B-cells [123]. A combined effort with the hierarchical expression and collaborative action of transcription factors along with epigenetic modulators leads to a committed lineage of B-cells.

B lineage commitment is riveted around transcription factor competition and cross-regulation [124]. Transcriptional regulation plays a critical role in B-cell differentiation with expression of distinct sets of genes at discrete stages resulting in the initiation of lineage differentiation [125]. It involves the onset of B-cell transcriptional program and the loss of other immune cell potentials for stem cells to acquire B-cell identity [126]. Key features of B lineage commitment are linked to the initiation of gene
rearrangements at the Ig heavy and light chain loci and the regulated expression of several transcription factors, like EBF1 and PAX5, which promote B-cell commitment by repressing lineage-inappropriate gene expression and reinforcing B-cell specific gene expression [127] [128] [129] [130]. Certain DNA-binding proteins, such as IKZF1, PU.1, E2A and TCF3, FOXO1 are required primarily for the formation of more primitive non B-cell committed lymphoid progenitors, while other factors such as EBF1 and PAX5 have more direct roles in specifying the B-cell-specific gene-expression program [127] [131] [132].

Figure 5: Main transcription factors involved in B-cell development [119]

Disruptions in this delicate balance of activity contributed by stage- and lineage specific transcription factors can lead to catastrophic repercussions like a developmental arrest as observed in B-lineage acute lymphoblastic leukemia (B-ALL) [133]. The majority of ALL cases involve transformation of B lineage cells, with only a small percentage of cases resulting from T-cell transformation [125]. Mutations in EBF1 and PAX5, have also been identified in human ALL [125] [134]. Small perturbations in the self-reinforcing network of these transcription factors critical for B-cell development, interacting with many factors, directly and indirectly resulting in ALL [125].

Many studies have focused on specific genes encoding transcription factors that are critical for initiating and managing B lineage specificity like TCF3, EBF1, PAX5 and B-cell specific activator protein [135] [127]. EBF1 have been shown to support B-cell differentiation by coordinating the expression of downstream transcription and signaling factors with the activation of the immunoglobulin gene recombination machinery [132]. Furthermore, EBF1 function is related to the regulation of genes involved in DNA repair that contributes to the malignant transformation in a dose-dependent manner [136]. EBF1 has been found to be crucial for B-cell lineage specification, as in its absence lymphoid progenitor cells fail to initiate transcription of
B-lineage genes in collaboration with a hierarchy of partner proteins like TCF3, RUNX1 and PAX5 [130] [118].

The expression of PAX5 is induced by TCF3 and EBF1, and is involved in recruiting chromatin-remodeling, histone-modifying and transcription-factor complexes to its target genes to activate the transcription of B-cell-specific genes, and to silence lineage-inappropriate genes [128, 137]. A B-lineage specific transcriptional program is initiated even in the absence of PAX5, but those cells were not stably committed and stimuli like cytokines or Notch signaling could drive them to alternate fates [138, 139]. PAX5 render a dual role by repressing B lineage ‘inappropriate’ genes and simultaneously activating B lineage–specific genes [140].

Studies have proven combined heterozygous deletions of EBF1 and PAX5 being preconditioned for the development of lymphoid leukemia with phenotypic features resembling B-ALL in humans [136]. They act in concert with other factors to establish the B-cell specific transcription profile; concurrently repress alternative cell fates and lock cells into the B-lineage fate [132]. These factors were expressed and operated in a sequential manner during the differentiation process [138]. Ungerbäck et. al. demonstrated that PAX5 and EBF1 collaborate to reduce the impact of active Notch signaling, thereby maintaining B-lineage identity in early progenitor cells [128].
The proteins present in the environment dictate the PC, adsorbed on the surface of NPs. Many studies have shown that PC induces sequential immune responses, leading to the uptake of NPs by phagocytes of the reticuloendothelial system (RES) and rapid clearance from the circulation [141, 142]. Functionalization could change surface properties like hydrophobicity and charge, but does not completely eliminate the formation of protein corona on NPs [143]. It was believed that polyethylene glycol (PEG) could prevent binding of plasma proteins increasing their circulation half-life and ultimately leading to greater accumulation in tumors via the EPR (enhanced permeability and retention) effect. Paradoxically, long-circulating PEGylated liposomes had an increased clearance from circulation mediated by anti-PEG IgM [144]. Recently, it was revealed that the stealth effect of this PEG functionalization could be explained by the presence of some specific proteins in the PC rather than the absence of all [145]. More than any physical or chemical property, it was the surface proteins that determined how the NP behaved in a biological protein rich medium.

Current testing methodologies are focused on NP synthesis and characterization involving *in-vitro* and *in-vivo* studies. These testing strategies should be able to predict how the NP would behave in a biological medium. This prediction should be able to reduce steps and provide an alternate route to understand whether the NP in focus, upon administration *in-vivo*, could reach the action site and perform the functions its designed for. Concurrently, it should also help us understand the destiny and how it might affect cellular networks and molecular pathways. This understanding might help to prove how toxic and stable these particles could be and where they would ultimately end up in the body.

Our aim was to unravel the ‘first face’ of NPs presented to cells by analyzing the qualitative composition and quantitative distribution of the most exposed or the outermost proteins in the PC. We intend to achieve this by differentially tagging surface proteins using biotin. This would represent the ‘true identity’ of the PC rather than considering the whole corona. This surface of the hard corona is responsible for the interactions with the dynamic soft corona which has much less of an identity and might not play a pivotal role in the behavior of the NP in a biological system.
To investigate our hypothesis, we characterized 10 nm TiO$_2$ NPs and 30 nm TiO$_2$ NPs. The PC was isolated by subjecting the whole TiO$_2$ NP-PC complex to FASP. This acted as controls aiding the discrimination of non-surface proteins compromising the PC. To differentiate between the outermost exposed proteins or epitopes on the surface of the TiO$_2$ NP-PC complex, a tagging was utilized by labelling with EZ-link TM sulfo-NHS-SS biotin. Sulfo-NHS-SS biotin is a thiol-cleavable amine-reactive biotinylation reagent that contains an extended spacer arm to reduce steric hindrances associated with avidin binding [146, 147]. Being water soluble, makes it easy to work without the need of DMSO or other organic solvents. The cleavable spacer arm helps pull down using streptavidin affinity columns. Once labelling has been achieved, the NPs underwent on-bead digestion to achieve a whole PC digest using trypsin protease enzyme. This mixture of biotinylated and non-biotinylated peptides from the PC were incubated in NeutrAvidin agarose resin for the purification facilitating specific pull down of labelled peptides. Both unlabeled controls and labeled samples were analyzed using a reverse-phase nano LC coupled to a hybrid ion trap mass spectrometer. The peptides were fractionated initially using an nLC Easy II LC system by C18 reverse-phase chromatography to achieve maximum separation before it is electrosprayed online into the LTQ orbitrap Velos Pro.

For analysis of MS RAW files, they were searched against a bovine uniport database with 23848 entries along with contaminants. Proteome discoverer (Thermo Scientific, version 1.4) pipeline was utilized with the database search made possible with SEQUEST search engine. MSF files were imported to Scaffold software (4.3.4, Proteome Software Inc.) where statistical tests, validation and quantitation were achieved. A second database search was also performed using X! Tandem for validation.

The greater hydrodynamic size of TiO$_2$ NPs in water and media indicated aggregation which was rigid enough not to be broken down by sonication. This was observed in other studies also where the agglomerate size varied from 770 nm to 1052 nm depending on the type of medium which also led to a dramatic decrease of zeta potentials to about -10 mV [148-150]. 10 nm NPs were found to have a higher hydrodynamic size compared to 30 nm which could be reasoned because of the high ionic strength compressing the electrical double layer and the magnitude of the repulsive barrier decreases whereby resulting in the domination of universally attractive van der Waals forces [148].

This method helped us to identify and quantify the entire NP-PC and the subset of biotin labelled surface proteins. This also enabled in reiterating the existence of a multilayered PC, one part being buried under a layer of surface proteins accessible for
Biotin labeling. This distinction helped reduce the number of proteins considerably compared to the entire PC. Different sizes of TiO$_2$ NPs (10 and 30 nm) exhibited different distribution of proteins both in the whole PC and the biotinylated surface proteins. In 30 nm NPs, 80% of surface PC proteins comprised half of the proteins constituting the whole PC. Meanwhile, in 10 nm NPs, 80% of surface PC proteins were made up of the 6 most abundant proteins comprising the entire PC. The testing of both sized NPs indicated a different pattern in ranking of the most abundant proteins. This reiterates the fact that whole PC and surface PC may differ with material and sizes. We observed some of the most abundant proteins missing from the surface which verified the lack of homogeneous distribution of proteins along the length of various layers of the PC. The thickness of the protein corona depends on many factors including particle size, protein concentration, and surface properties of the NP \cite{151}. For example; in the plasma, proteins maintain a hydrodynamic diameter of about 3-15 nm making the corona too thick to be composed of a single layer \cite{151}. It consists of a layer of primary binders that bind to the NP directly and the secondary binders attach to the primary binders by protein-protein interactions, as per a model by Simberg et al. \cite{152}.

Results in paper I (figure 2) indicated that low abundant proteins comprised of 20% of the total surface PC and this was also particle and size dependent as the whole PC. Meanwhile, figure 2D in paper I indicates that the proteins contained on the surface might not be based on protein properties like molecular mass and calculated isoelectric points.

This study does help us understand the distribution pattern of a NP-PC and if this could be used as a potential benchmark for the prediction of NP targeting and uptake. Our aim was to obtain qualitative and quantitative differences between the entire PC and the subset of proteins or moiety of proteins on the surface of the PC. This would prove the fact that considering the whole PC might not be the best strategy forward as it was just the subset of PC surface proteins that ‘talk’ to the cell or other proteins which is selectively labelled using biotin in our study. Consequently, all biotinylated peptides obtained might be either surface proteins or moieties of a surface protein. Concurrently, this information may generate credibility, reliability and simplify NP profiling based on PC. Understanding of this nano-niche could aid evaluation of different novel formulations or surface functionalizations involving nanomaterials or NPs with the intent of specific targeting in drug delivery or nanomedicine.
Different sizes of TiO$_2$ NPs have been in use for both medical and everyday applications. In recent years, focus of research have progressed from material composition of NPs to surface functionalization. The potential impact of PC on the cells and other biomolecules have least been studied. Although multiple delivery routes are possible, the majority of the clinically-relevant nanocarriers, such as anti-cancer and anti-inflammatory nano-drugs, are expected to require intravascular administration [153]. Thus, a rational design of drug carriers is necessary to limit their cyto- and genotoxicity, prevent their clearance by reticuloendothelial system, increase the circulation half-life of their cargo, and to deliver a sufficient payload of therapeutics to the diseased regions [153]. These NPs possess a PC which would be the primary line of interaction between the NP and the cell.

As explained earlier, intravascular route is one of the most preferred routes for nanomedicine. So, the interaction of these NPs with the endothelium is inevitable. But in cases of drug delivering long-circulating reactors or slow releasing systems, it is required that the drug carrier does not interact with the endothelium [154]. Drug carriers have been designed for this specific purpose but their application is impeded by the complex PC adsorbed on the surface. This complicates any prediction of cell interactions, bio-distribution, toxicity and in particular, the unspecific uptake of nanocarriers have been a major obstacle in clinical studies [155]. The PC confers these particles with a biological identity and this is dependent on many factors including the environment which could be detrimental, especially when they may get involved in various cellular processes. Moreover, these underlying toxic effects of TiO$_2$ NPs are very poorly studied especially in the transcript and protein level.

Although a simple monolayer, a healthy endothelium is very much optimally placed and can respond to many factors that help regulate the vascular tone, cell adhesion, thromboresistance, smooth muscle cell proliferation, and vessel wall inflammation [156]. The function as a semipermeable barrier is one factor of utmost importance to these cells for the regulation of tissue-fluid homeostasis, transport of nutrients and migration of blood across the barrier [157] and this is maintained by transcellular and paracellular transport mechanisms across the cell layer. This critical function is made possible by the orchestrated activity of many protein complexes at the inter-
endothelial junctions like adherens junctions (AJs), tight junctions (TJs) and gap junctions (GJs) [158].

TJs act as a fence which prevents the mixing of membrane proteins between the apical and basolateral membranes; and a gate which controls the paracellular passage of ions and solutes in-between cells [159]. They contain two types of transmembrane proteins, occludins and claudins, which confer these functions, and other associated cytoplasmic proteins that may link tight junctions to the actin-cytoskeleton and the adherens junction [159]. Meanwhile, AJs play important roles in cell–cell adhesion, regulation of the actin cytoskeleton, intracellular signaling and transcriptional regulation by interactions between the cadherin superfamily (VE-cadherin) and the catenin family members (p120-catenin, β-catenin, α-catenin, α-actinin) (figure 6) [159] [160]. Another participant, ARP2/3 complex and the nucleating-promoting factors play important roles in the maintenance of endothelial barrier function and junction remodeling of established endothelial cell junctions (figure 6) [161].

Figure 6: Illustration of AJ complex: VE-cadherin acts as the major structural component facilitating the proper assembly of AJs and adhesion of adjacent cells in a Ca\textsuperscript{2+} dependent manner. The cytoplasmic tails of VE-cadherin are composed of two domains, the juxta-membrane domain (JMD) and the COOH domain. p120 catenin binds to JMD while β-catenin binds to CTD domain to which α-catenin binds, which links the Cadherin-catenin complex to actin cytoskeleton. AJs anchor to the actin cytoskeleton via α-catenin, α-actinin along with other proteins.

Here, we tried to understand how TiO\textsubscript{2} NPs affected human umbilical endothelial cells (HUVEC) using a combination of two ‘omic’ approaches, transcriptomics and mass spectrometry based proteomic approaches. Both approaches were used on the same sample to identify the effects of exposure to sub lethal concentrations of two sizes of TiO\textsubscript{2} NPs, 30 nm and 1-3 nm or Ultra small NPs (USNPs) had on the transcript level and
on the protein level. This approach helped us to investigate the correlation between NP size and its effects on the dynamics of endothelial proteome and transcriptome.

Human dermal microvascular endothelial cell line (HMEC-1) was cultured in MCD media 131 along with supplements. The NPs were prepared in complete cell culture media, sonicated and vortexed to avoid NP aggregation. Sub lethal concentration levels were ensured with no cytotoxicity at 0.2 and 10 mg/l and exposed for 1 hr and 24 hrs. The control samples were untreated cells cultivated under identical conditions without exposure to NPs.

RNA was isolated, whose quantity and quality were estimated by measuring the absorbance at 280 and 260 nm, and the intactness by using the Agilent 2100 BioAnalyzer. RNA samples were submitted to Applied Biosystems SOLiD high-throughput sequencing at the Science for Life Laboratories (Scilifelab). 28S rRNA/18S rRNA ratio ≥2 was ensured for RNA integrity. Transcript analysis were performed and fold changes between conditions were estimated using transcript relative abundance log2 ratios. Genes were considered differently expressed if there was (i) a more than 2-fold change in TRA log2 ratio, (ii) if the P-value was 0.01 or less and (iii) if at least 20 reads mapped the gene.

For proteomics, the cells were frozen right after the exposure. These cells were then lysed using a buffer containing a cocktail of detergents along with protease and phosphatase inhibitors. The samples were then incubated to facilitate complete lysis, prior to be processed using FASP as described earlier. The peptides obtained from FASP were desalted using zip tips and later reconstituted in 0.1 % formic acid. The concentration was measured in the Nanodrop before being analyzed on the reverse phase nano-LC (liquid chromatography) coupled to a hybrid LTQ Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific Inc.). Quantitation was based on top 3 precursor intensity values which were normalized and Mann-Whitney test (P value < 0.05) was used for analyzing the statistical significance in Scaffold (version Scaffold_4.3.4, Proteome Software Inc., Portland, OR). R programming was employed for all statistical analysis, heat maps and volcano plots for both transcriptomics and proteomics. Pathway analysis was carried out using GO annotation, REACTOME, Panther and KEGG in REVIGO.

The transcriptomics analysis demonstrated that the exposure to 30 nm TiO₂ NPs induced a higher number of differentially expressed transcripts both for 1 h (25 genes) and 24 h (135 genes), compared with the exposure to TiO₂ USNPs (15 genes for 1 h and 50 genes for 24 h). There were 10 common transcripts for 30 nm NPs between short and long exposures while nothing common for USNPs. The 24 h exposure led to
27 common transcripts between 30 nm and USNPs. Stress responses was evident in the short time while signaling and protein folding seemed affected by prolonged 24 hr exposure.

Protein and peptide identifications were filtered with a high confidence interval and a minimum of 2 peptides were considered as reliable identifications (0.1 FDR). Top 3 peptide precursor intensities were quantified as they offered accuracy while providing label-free quantification and helped effective comparison across samples. A total of 1107 and 1345 proteins were identified upon exposure to USNPs and 30 nm TiO\textsubscript{2}-NPs respectively. Under the exposure to USNPs, 255 proteins were differentially expressed (23% of the total number of identified proteins). Meanwhile, 50 were found to be differentially expressed in case of 30 nm exposure (3.71 % of total proteins). Our earlier studies had confirmed that 30 nm NPs recruited high abundant serum proteins such as hemoglobins, tetranection, alpha-2- HD-glycoprotein, and low abundant ones such as trombospondin, inter-alpha-trypsin inhibitor in their PC \cite{162}.

This study revealed the down-regulation of \(\alpha\)-catenin (CTNA1), Ras GTPase-activating-like protein (IQGAP1), \(\alpha\)-actinin 1 (ACTN1), \(\alpha\)-actinin 4 (ACTN4), protocadherin 12 (PCDH12), vimentin (VIME), and ARP2/3 along with other cytoskeletal proteins that play a part in endothelial cell-cell adhesion at the AJs (figure 6). Proteins closely related to these complexes like integrin, Rho GTPases, SRC 8 Cortactin (CTTN), and CDC42 have also been found to be downregulated (figure 6). This indicates that TiO\textsubscript{2} USNPs might affect ARP2/3-induced actin assembly as SRC promotes actin polymerization by simultaneously binding to ARP2/3 complex and F-actin, which has the dual effect of activating ARP2/3 complex for actin nucleation and stabilizing filament branches created by ARP2/3 complex. The growing actin filaments requires capping by F-actin capping protein (CAPZ) and lesser turnover of those proteins, as observed in our study might affect recruitment of new actin for assembly. Its capping activity is enhanced in the presence of tropomyosin \cite{163}, which suggests that the two proteins function as a complex to stabilize a filament which would be disrupted with the downregulation of these two proteins. Other key players in cellular stability, affected by the NP exposure are Vimentin (VIME), IQGAP1, both \(\alpha\) and \(\beta\)-tubulin and tubulin-folding cofactor (TBCB). The under expression of ezrin, one of the proteins from ERM (Ezrin, Radixin, Moesin) family that are linkers between membrane anchorage proteins such as VE-catherin and cytoskeletal proteins indicates that the exposure to NPs result in the destabilization of cell adhesion and cell polarity.

The size of TiO\textsubscript{2} NPs is factor that determines the cellular response. This data reveals that larger 30 nm TiO\textsubscript{2} NPs, at the same concentration cause a differential cellular response on cytoskeletal proteins like \(\beta\)-tubulins and thrombospodin-1(TSP1) which
are strongly up-regulated while proteins like calmodulin (CALM) are down-regulated. This might affect the stability of AJs as it the complex, E-cadherin: β-catenin: α-catenin which anchors adjacent cells with actin cytoskeleton.

The omics analysis of endothelial cells exposed to cell culture media containing TiO₂ NPs have provided understanding of the effects of TiO₂ NPs on the permeability at different levels; paracellular traffic, transcellular traffic; and the crosstalk between both transport pathways at the cellular cytoskeletal level.
**Paper III**

**SHOTGUN PROTEOMICS TO UNRAVEL MARINE MUSSEL (*Mytilus edulis*) RESPONSE TO LONG-TERM EXPOSURE TO LOW SALINITY AND PROPRANOLOL IN A BALTIC SEA MICROCOSM**

Intensive research has been going on for many years on pharmaceuticals but very less have been invested in understanding its effects from accumulation, both individually and in combination with other chemicals and conditions. Our study aimed to evaluate individual and ‘mixture effects’ (pharmaceutical + salinity) of PROP in the gills of the blue mussel at the molecular level using shotgun proteomics. The microcosm experiment [164] with the model communities was exposed to 0 (control) and 1000 µg/L PROP (P1000) along under normal salinity (NS: 6.5 psu) and low salinity (LS: 5.5 psu).

After a 4-week exposure, gills from the Blue mussels were homogenized and fractionated into nuclei, membrane and cytoplasmic fractions (figure 7). Proteins in each of these fractions were processed using filter aided sample preparation (FASP), where these crude fractions were digested into tryptic peptides immobilized on a membrane. FASP was superior in comparison to traditional methods in various stages of sample preparation as explained below:

1. It allowed detergent based lysis in high concentrations, which enabled solubilization of protein complexes.
2. Effective clean-up of detergents, salts and unwanted small molecules downstream.
3. The ease and swift preparation time along with its least handling needs made FASP as one of the best-suited methods for mass spectrometry based analysis.

The peptides were fractionated using reverse-phase Easy nLC 1000 on a two-column setup with a trap and analytical column. A step gradient with a total elution time of 70 minutes nano-sprayed the peptides into the Orbitrap Velos Pro where 20 most abundant ions were selected and fragmented by CID. Each of the biological replicates were analyzed in triplicates. All raw files were processed using SEQUEST with an FDR probability of 95 % and 99 % for peptides and proteins respectively. The processed files for nuclei, cytosolic and membrane fractions were merged in Proteome discoverer before being exported to Scaffold software. Proteins were quantified using spectral abundance to view the effects of PROP and salinity.
Figure 7: Schematic illustration of sample preparation. Tissues were processed and fractionated into 3 fractions enriched in nuclei (fraction N), membranes (fraction M), and cytoplasmic contents (fraction C).

A total of 2071 proteins were identified with 0.01% decoy FDR. 587 proteins were found consistent with all biological and technical replicates. This unsequenced model organism convinced us to employ a transcriptome database of *M. galloprovincialis* where a tryptic peptide database was generated from the transcriptome sequences with which identifications were based upon, along with Uniprot protein database. *M. galloprovincialis* and *M. edulis* have a very similar genome, except for some very specific markers like first intron of the *mac-1* actin gene [165] [166]. Analysis of this kind drastically increased sensitivity of peptide identification, reduced ambiguity and enabled the detection of known and novel peptide variants [167]. Principal component analysis and hierarchical clustering analysis revealed discrimination of all 4 groups and that PROP had more adverse effects under low salinity conditions. Multivariate and univariate analysis revealed 44 and 32 proteins respectively, to exhibit significant differences among the different groups. Annotation of protein sequences had to be retrieved to perform GO analysis. The predicted protein sequences from the transcriptome was blasted against sequences in NCBI by which homologous proteins could be retrieved along with their GO terms. GO analysis was performed on 74 proteins, which were found to be significantly different with exposure and conditions applied. The difference in salinity created osmolarity differences, which in turn affected nucleic acid binding, with regards to molecular function; while metabolism was most affected under biological processes.
Figure 3 and 4 in paper III depicts the differential expression analysis. Low salinity stress affected microtubules in cilia, metabolism, muscle contraction, intra cellular transport, signaling pathways, defense mechanisms, transcription and various cellular processes. Meanwhile PROP disturbed the expression of cytoskeletal proteins, translation, various metabolic processes, and cellular processes like oxidative phosphorylation. The combined effect of the stressors modulated the expression of structural proteins, especially proteins related to cilia functions and down regulated membrane protein transporters. Both stressors were found to create ‘mixture effects’, catapulting contributions from PROP. This might be because of impairment of the defense mechanism brought upon by the down regulation of ABC transporter proteins responsible for elimination of toxic substances.

When analyzing an environmentally relevant compound in the laboratory, there is a need for high concentrations for observing relevant effects. That was the reason why we had to include concentrations of both extremes. Moreover, pharmaceuticals have a high tendency to interfere with environmental matrices, especially humic acid and other nitrogen species affecting degradation and photoproduct formation [80]. Additionally, the role of mussels as bioindicators of marine pollution could be more interesting in conjunction with proteomics as this could provide cellular, genetic, biochemical and molecular level cues for the evidence of exposure to one or more pollutants. In 2004, environmental risk was added in the risk-benefit analysis for veterinary medicinal products [168]. Hence, veterinary medicines were not authorized in the EU in case of any environmental risk meanwhile human medicinal products could still be launched with an environmental risk [169]. According to WHO, increasing concentrations of pharmaceuticals in water bodies are to be expected in the coming years because of increasing population [169]. Therefore, there is pivotal need for these substances to identified and prioritized using a combination of exposure and adverse effect data. In addition, abiotic factors like climate changes leading to salinity could act cohesively leading to unintentional, unexpected and disastrous effects. Top-down toxicity studies, where a ‘mixture effect’ of stressors could be considered for their composite effect.

Toxic responses involve effects at all, including DNA, RNA, proteins, metabolites and even interactions between them. Therefore, it is pivotal to understand the response at all -omic levels and integrate the data in order to unravel key mechanisms of toxicity. As far as a simplified risk assessment is concerned, proteomics serves to be one of the best tools for undermining the role of functional entities for all molecular processes.
Paper IV

IDENTIFICATION OF INTERACTOMES FOR EBF1 AND PAX5 UNRAVEL MOLECULAR MECHANISMS UNDERLYING THE ESTABLISHMENT OF EPIDERMAL LANDSCAPES IN B-CELL DEVELOPMENT

Biological systems possess subcellular architectures that are dynamic and responsive to varying needs and conditions leading to specific spatial networks of macromolecules capable of signaling, creating phase changes that define functional regions in cells and generates cellular responses to environmental changes [1]. Cellular proteins exist and operate in an extremely crowded environment and this has assisted in the formation of weak but functionally important protein-protein interactions [1]. These spatially restricted multiprotein complexes have made proteomic characterization challenging using traditional methods like yeast two hybrid or immunoprecipitation [2]. During signaling, some proteins would be involved in direct interaction and some indirectly with its immediate partners in the cluster. To investigate multiprotein signaling processes, it is pivotal to decode the proximal interaction partners (PXIs), along with its immediate or direct interactor.

Interpretation of how transcription factors restrict cell fate of B-cells and drive lineage differentiation is important to understand in the development of hematopoietic cancer, and elucidating this information may provide important access points for therapeutic intervention. Earlier studies have shown EBF1 and PAX5 to be directly involved in the coordinated activation and repression of target genes involved in B-cell differentiation leading to the establishment of a lineage specific pathway warranting the suppression of alternate cell fates simultaneously. The understanding of these cellular mechanisms at the molecular level requires unraveling of the interactomes of EBF1 and PAX5, especially while dealing with such complex stage and lineage dependent regulatory network. In this study, we employed our multifaceted tool, proteomics to identify proximal proteins of EBF1 and PAX5. This was made possible by the ‘proximity dependent biotin identification’ or BioID.

Protein-protein interactions (PPIs) and BioID

Proteins rarely act alone. They team up with other proteins or biomolecules, transforming into complex molecular machines. They possess intricate physicochemical dynamic connections to undertake biological functions at both cellular and systems levels [170]. A critical step towards unraveling the complex molecular relationships in living systems is the mapping of protein-to-protein physical
“interactions” and this led to the creation of another ‘sub-omic’ field in proteomics called the ‘interactome’ [171]. PPIs can be classified as permanent or transient according to how strong and reversible they were [172]. Transient PPIs could be further classified as weak or strong according to their binding affinity. The weak transient complexes show a dynamic mixture of different oligomeric states in vivo, characterized by a dissociation constant (K_D) in the micromolar range and lifetimes of seconds meanwhile the strong transient complexes change their quaternary state only when triggered and are stabilized by binding of an effector molecule. Those interactions may last longer and have a lower K_D in the nanomolar range [173].

BioID helps us track all protein associations in vivo (figure 7). This is accomplished by the fusion of a promiscuous biotin ligase to a protein of interest expressed in living cells, thereby labeling proximate proteins during a defined labeling period [174]. It was critical to understand the behavior of the desired proteins, EBF1 and PAX5 in the specific state and lineage. This fusion of BirA* to the proteins of interest could open doors to reveal a wealth of data, by characterizing protein complexes including transiently associated protein clusters. This technique exploits the ability of biotin protein ligase enzyme that is capable of covalently labeling a protein target.

Figure 7: Workflow for BioID [175] copyright nature reviews 2013
Bir A is a biotin ligase enzyme from *Escherichia coli* that regulates the biotinylation of a subunit of acetyl-CoA carboxylase and acts as a transcriptional repressor for the biotin biosynthetic operon [176]. The BirA acceptor-peptide system takes advantage of the extreme specificity of BirA in biotinylating its substrate peptide [177]. Normally wild-type BirA only transfers biotin to a substrate bearing a specific recognition sequence. A mutant engineered form of BirA* (R118G) with refined substrate specificity and much reduced affinity for biotinoyl 5'-AMP was utilized [178]. BirA* is promiscuous in that it activates biotin for transfer in the absence of a substrate peptide, and the activated biotinoyl-5'-AMP is free to diffuse away from the enzyme and covalently modify primary amines of nearby proteins and since biotinoyl-5'-AMP is highly reactive and short-lived, the zone of modification in BioID is thought to extend only about 10 nm in radius around the BirA*-tagged protein thereby creating biotinylated proximity partner proteins [179]. Once biotinylation of proximity partners were achieved, cells were lysed under denaturing conditions. These biotinylated proteins are selectively pulled down on agarose beads coated with streptavidin. The beads are washed repeatedly to minimize background and reduce non-specific binding. Subsequently, the proteins are digested on beads using trypsin and identified using LC-MS/MS.

The precise control from gene expression to functional protein was made possible by the fine segregation of genetic material with the transcriptional machinery in the nucleus and translational machinery, along with the metabolic systems in the cytoplasm by the nuclear membrane. A selective bidirectional transport facilitates gene expression, signal transduction and cell cycle progression. This physical separation also necessitates the existence of a molecular machinery that specifically recognizes a cargo in one compartment, translocate it through the nuclear pore, and finally releases it in the other compartment [180]. This nuclear transport systems was first proposed following the characterization of a nuclear targeting signal in the simian virus 40 (SV40) large T antigen more than 20 years ago [180]. The proteins destined for nuclear pore translocation contains specific targeting sequences called nuclear localization signals (NLSs) [180]. Thus, we used BirA* harboring NLS from SV-40 on its N-terminus as controls.

Fetal liver pro-B-cells from Pax5-/ and Ebf1-/ fetal livers were cultured on OP9 stroma cells in Opti-MEM cell culture media along with additives which included cytokines like IL-7. Murine Abelson virus-transformed cell line 230-238 pro B-cells, and IL-3 dependent pre-B-cell line BaF3 were cultured in supplemented RPMI1640. EBF1D430 (EBF1 lacking the carboxy-terminal transactivation domain) or PAX5 (full length cDNA, Genscript) was cloned into a retroviral MigR1 vector expressing a green
florescent protein (GFP) reporter (an efficient means to deliver expression constructs to mammalian cells) and carrying BirA* resulted in the fusion of BirA* to the C-terminal of bait proteins. Transfection was accomplished and infection was performed into 230-238 cells. Stable transfection was confirmed by FACS for stable GFP+ cells. The cells were grown to 70% confluency before biotin was spiked in the media. After 24 hours, cells were lysed in highly reducing RIPA buffer and sonicated to ensure dissolution of immiscible proteins. The biotinyalted proteins were pulled down using streptavidin-sepharose beads and washed thoroughly. On-bead digestion was achieved using trypsin to obtain peptides. The peptides were analyzed in an Orbitrap Velos Pro mass spectrometer. Reverse phase liquid chromatography was performed with a polar mobile phase and a non-polar stationary phase to achieve prior separation before being ionized by electrospray ionization. After fragmentation by collision induced dissociation (CID), M/Z values are calculated and subsequently analyzed using MS/MS.

Data Analysis

Data analysis was achieved using Trans-proteomics pipeline (TPP). Unprocessed ‘.raw’ files were converted into open format, ‘.mzXML’ files using proteowizard [181] and database searches were performed by Comet search engine. The open source search engine, Comet started off as an academic version of SEQUEST. TPP aimed to come up with an open source tool suite for a complete start-finish shotgun proteomic dataset processing. It was a suite of tools for conversion of formats, sequence database searching, spectral library searching, statistical testing, validation and protein abundance calculation.

TPP workflow:

1. Conversion of vendor specific to vendor neutral format: the output file format of an MS/MS experiment would be vendor specific and in our case it was Thermo ‘.raw’ format. Since the input file format of a TPP workflow was mzXML, an msconvert command line tool had to be installed outside TPP. This architecture independent file format helps standardized representation of MS data from various vendors and instruments thereby, removing the burden of having to support multiple native formats from the developers of analytical applications [182].

2. Protein database search: assign amino acid sequences to the fragmentation spectra from the mzXML file. This was made possible by Comet [183]. They were searched against a mouse database with reverse decoy sequences and contaminants (85828 sequence entries). Monoisotopic masses were considered for the calculation of masses of precursor and fragment ions with
a 10ppm precursor tolerance and 0.6 Da fragment tolerance. The FDR considered was 0.01 with a minimum peptide length of 6.

3. Validation of searches: the validation of the searches was made possible by Peptideprophet which provided statistical validation by calculating a score. This helped in identifying the quality of the identification [184]. iProphet further provides statistical refinement for the Peptideprophet results. The use of iProphet in the TPP increased the number of correctly identified peptides at a constant FDR as compared with both PeptideProphet and Percolator [185]. This resulted in a pepXML file that reports the probability for every PSM as well as the results of the modeling, including a representation of the receiver operating characteristic (ROC) curves [186].

4. Protein quantitation: the INTERACT program in TPP allowed the calculation of relative abundance of the precursor ions using XPRESS software.

5. Protein assignment: was accomplished by ProteinProphet by creating the simplest list of proteins that can explain all the peptide observations leading to the generation of protXML files [186].

All identifications were confirmed with identical parameters in Proteome discoverer (Thermo Scientific, version 1.3) coupled with SEQUEST algorithm (Thermo Fisher Scientific, San Jose, CA, USA; version 1.4.0.288) and X! Tandem (CYCLONE (2010.12.01.1).

Further analysis was carried out with the analyst module of ProHits. The simplified version of the software suite ("ProHits Lite"), consisted of the analyst module and admin office [187]. The analyst module helped to organize data by classifying them under bait and prey or controls and experiments, once the search results were imported into the Analyst module. Using Prohits, non-specific identifications and contaminants were filtered as background. Significance Analysis of INTeractome (SAINT) software was embedded in Prohits, assigning confidence scores to protein-protein interaction data. SAINT utilizes spectral counts for each prey-bait pair, which are modeled with a mixture distribution of two components representing true and false interactions [188]. It builds a large number of mixture distributions by pooling data (separate mixture distributions for individual prey-bait pairs), but all models are interconnected through the shared abundance parameters [188]. The outputs from SAINT were exported to Cytoscape (Version 3.5.1) for visualization.

BioID revealed 239 PAX5 and 104 EBF1 interaction partners with a cut off of 0.85 (≤1% FDR) SAINT score with an overlap of 40 shared between the two (table S1, paper IV) as in paper IV, figure 1A. 60% of proximal proteins were accounted for in curated databases according to STRING (version 10.5) (figure S1). Only 4 EBF1 and 6 PAX5 interactions were identified as experimentally verified in Uniprot database (figure S1).
Meta-analysis of microarray data from the expression of PXIs using Gene expression commons revealed about 80% of genes encoding the PXIs were broadly expressed in the hematopoietic system (figure S2).

GO molecular function analysis revealed both interaction partners involved in DNA binding and transcriptional cofactor activity (figure 1B, 1C). Figure 1C illuminate the fact that EBF1 and PAX5 interactors act both as transcriptional co-activators and co-repressors. Functional analysis identified protein members of complexes related to chromatin binding and remodeling (figure S2-S4, 2A). Figure 2E analyzed the levels of histone modifications at shared or unique EBF1 and PAX5 sites, revealing that H3K27 acetylation was significantly higher. The components of chromatin remodeling complexes like BAF, INO80, CHD and ISWI were identified (figure 2F). The BAF complex components were shared between PXIs of PAX5 and EBF1, verified by ChIP-seq analysis which suggested a collaborative mechanism that might exist to generate histone modifications.

EBF1 and PAX5 biotin tag identified FLI1 and IKZF1 in both pull downs (table S1). PAX5 tagged samples co-precipitated EBF1, RUNX/CBFB, TCF7, CTCF and OCT4/POU2F1 (table S1, figure 3D) and these were enriched in PAX5 ChIP analysis too (figure 3C). This regulatory network was verified using Q-PCR for their ability to interact with promoters of EBF1 and PAX5. Proximity ligation assay (PLA) was employed to confirm the interaction in vivo.

Considering the labeling radius of about 10 nm, BioID analyzed distinct sets of candidate proteins that might be collaborating with either EBF1 or PAX5 separately or alone. Pro-B-cells exhibited enrichment in binding for FLI, IKZF1 and RUNX1 during ChIP-seq analysis. It has been demonstrated earlier that EBF1 with PAX5 along with other E26 transformation-specific (ETS) proteins like FLI1 and RUNX/CBFB play a role in the activation of the target genes [189] [190] [191].


60. Oncology, A.S.o.C., Understanding Targeted Therapy. 2015.


98. America, C.t.c.o.


135. Hagman, J. and K. Lukin, *Transcription factors drive B cell development*. (0952-7915 (Print)).


CONCLUSIONS AND FUTURE PERSPECTIVES
Proteomics have been used both for hypothesis-driven and exploratory analysis. The key outcome of our experiments was evidences that were measurable, incorporated with learning new cues for future research. By incorporating proteomics as a multi-faceted tool, we intended to improve and facilitate present norms and practices in the field of molecular biology, biomedical research and environmental testing. In addition, tandem MS generated from MudPIT tackled complexity at immense levels.

In all the papers, a gel free approach has been undertaken. Each of the study was trimmed according to the needs of the question under focus, keeping in mind the compatibility with downstream processing. Efficient lysis and solubilization required the use of detergents or and chaotropic agents. Because of the lack of a gel, where proteins could be clamped upon, removal of contaminants and detergents were not possible. One of the critical steps in our study related to proteomics has been sample preparation, where ideal techniques and methods had to be used in conjunction, aiding to crack open the doors of opportunities.

Nanomedicine offers a strategic leverage compared to traditional drugs and therapeutic biomolecules, by being able to cross biological barriers like plasma membrane or the blood-brain barrier. Meanwhile, it was never the naked NPs, but the PC that gifts NP's its biological significance dictating cellular uptake, bioavailability, and even toxicity. Proteomic analysis of the PC might serve as an analytical tool for all kinds of biomaterials that are intended to be used in nanomedicine. Our analysis indicates the PC to be multilayered and the surface of the PC were composed of high abundant and low abundant protein classes; the latter being particle and size dependent. This reiterates the importance of proteomic characterization and the concept of proteomic fingerprints for NPs used in nanomedicine. To decipher long term effects of NPs on cells, a shotgun approach on exposed endothelial cells revealed USNPs to have adverse effects compared to the bigger NPs leading to altered structural integrity, vehicular transport across the membrane and signaling pathways. These findings reveal that NPs interact and actively participate in the cell’s molecular processes. It would be of interest to investigate if these effects were maintained by the cells and in the calculation of a risk : benefit criteria. There is a need to integrate material sciences, molecular biology, surface chemistry, pharmacodynamics, and even mathematical modeling to join hands during the design of NPs in nanomedicine. Nanomedicine open doors for opportunities if long term, proteomically tested, biocompatible NPs that could be circulated in blood for a longer period of time, gathering low abundant biomarkers could be made possible. These particles could be later harvested and screened for these biomarkers. Similarly, drug carrying NPs could
be designed to circulate in blood, releasing the drug over time. All this could be made possible by understanding the surface corona and how they evolve over time and environment.

As discussed above, proteins are the closest to the phenotype and effects of stressors like NPs or pharmaceuticals would be best described at the protein level. Long term exposure to PROP in the context of low salinity was evaluated proteo-genomically on the gills of *M. edulis*. PROP affected transcription, protein synthesis, membrane trafficking, signaling, and cytoskeletal proteins. The effects of PROP were heightened under low salinity, disrupting cilia functions and membrane transport. This led to stressed mussels, indicating a relevant method of tracking possible effects on the ecosystem. This proteomic assessment could be extrapolated to understand various combinatorial effects of pollutants and other abiotic factors.

Apart from being an analytical tool in nanomedicine and environmental assessment, proteomics was exploited in revealing interaction partners for regulatory proteins like EBF1 and PAX5. BioID helped biotinylate proximal proteins around a radius of about 10 nm. A. BirA*-based biotinylation of proteins was achieved in living cells, making interaction studies more relevant. Thereby, loss of subcellular localization could be avoided. This technique was characterized by biotin-streptavidin interaction, one of nature’s strongest bonds, allowed extensive purification of target proteins. Additionally, this assisted the use of highly denaturing conditions during lysis and purification. Labelling of both stable and transient interactions was crucial in the study of stage and lineage specific factors. It would be worthwhile to delve into identifying the direct and indirect interactions to reveal deeper insights into the complex regulatory network.

Thus, we demonstrated the feasibility of employing proteomic techniques to unravel the outer layer of the protein corona; to understand effects of TiO$_2$ on endothelial cells; effects of a stressors on an organism; or even to decipher molecular mechanisms underlying B-cell development. In this age of systems biology, an integrated approach of proteomics with other omics would do wonders in the understanding of complex networks. We have tried to integrate transcriptomics along with other methods to confirm our findings. Besides transcriptomics, integrating epigenetic gene regulations could broaden our understanding of gene regulation and molecular mechanisms. The acceptance that all cells and systems are heterogenous and have led research to undertake homogenous or single cell analysis. This could unveil undiscovered subpopulations, mysterious expression profiles and help unearth new regulatory and signaling pathways. Meanwhile, analysis has taken another turn holistically towards systems biology with mathematical and computational models.
At night, we sleep to charge ourselves, at breakfast we fill our batteries, and exercise to keep our motors running. Our lungs chug in air, hearts work like clockwork, and the gears keep turning. Is the human body a complex machine? Or is a highly complex system playing by the rules of chaos?

The butterfly effect refers to the concept where small, inconsiderate causes could have humongous effects as you might have heard, ‘a bird in a pasture flaps its wings and a storm brews in the Atlantic’! Chaos theory explains the huge sensitivity of these systems to internal and external conditions. The human body is one such system, whose intricacies are a matter of wonder. There are tiny conditions that are magnified which contribute to the holistic life of an individual. According to Crystal Ives, one of the characteristics of a chaotic system is bifurcation where the system can gyrate from order to chaos and back again. Another characteristic is order without periodicity like a pair of butterfly wings. Even though, the system is governed by many such rules, they undergo changes, delays and feedback.

Many such chaotic systems make up the human body like the lungs or the beating of the heart or even signaling pathways. Though the heart beats in a clock-like rhythm, it never beats the same twice. Kim and Stringer called aging as the loss of deterministic chaos or complexity. Or ‘dynamic disorders’ like leukemia, seizures and heart attacks. There could be a time when we could extrapolate the various events to bring a measure to chaos. Thereby, gain more understanding into the human body and the environment. A time would come when we would be able to predict seizures, heart attacks and neuronal disorders. As said before, patterns on top of patterns.
Biological research has been undergoing a revolution especially in molecular biology where research started compartmentalizing with a series of -omes like the genome, transcriptome, proteome etc. They have grown and established as separate fields of research and have started growing into each other too. We have focused on proteomics which stands for the large-scale analysis of the protein compliment. It made more relevant to study proteins as it is them that determines the phenotype of the cell, rather than genes. In our study, depending on the question in focus, either specific proteins were isolated or in some cases all proteins were considered. They were first chopped into smaller units and separated to make it simpler for the later analysis.

New nanoparticles must be tested for optimal targeting in drug delivery or imaging in nanoparticles based nanomedicine. One of biggest challenges is to successfully target a specific tissue or cell. TiO$_2$ nanoparticles would be bestowed as a biological entity as soon as it enters a biological media due to the adsorption of proteins on its surface called protein corona. The cue lies in studying the surface of the interface between the nanoparticle and the biological fluid rich in proteins, rather than the whole interface itself and this helps us predict interaction and behavior of the nanoparticle in focus. A key route of entry for either intentional intervention in biomedicine or unintentional in case of pollutants, is the vascular system lined with endothelial cells. The proteins adsorbed on the surface of TiO$_2$ nanoparticles bestows them as a biological entity and could influence the biological fate of these particles. This protein corona could get involved in cellular and metabolic activities of the cell and may lead to disastrous effects. Analyzing both the corona and the biological environment could also help us design NPs, possessing both favorable targeting potential and non-toxic. Proteomics provides a fast and simple characterization by which a nanoparticle’s potential could be evaluated. A global proteomic analysis of all proteins in these cells enabled us to understand the effects of TiO$_2$ nanoparticles on endothelial cells by modulating their permeability, and their attachment to other cells affecting their integrity. We could also find that different sizes of the same NP had varied cellular responses.

Pharmaceuticals like propranolol act as pollutants when they end up in water bodies leading to direct and indirect effects especially in combination with climatic factors like low salinity. Marine mussels are ideal organisms for biomonitoring because of their capacity to accumulate and concentrate pollutants. Proteomics on the mussel gills revealed effects on protein production, structural framework and other cellular activities. Low salinity conditions arose as the key factor in modulating severity of
propranolol effects. These conditions combined, led to dreadful effects especially on key cellular functions.

Acute Leukemia has been associated with a pronounced disturbance in differentiation of B-cells, a type of blood cell playing a major role in immunity. Because of severe side effects associated with the treatment, it is of vital importance to unravel how transcription factors govern the fate of B-cells and drive lineage differentiation. Proteomics was also used to understand protein-protein interactions in the development of B-cells, proving yet once again as an ambidextrous tool. The proteins of interest were fused with an enzyme capable of tagging proximate and interacting proteins in living cells. Once the proteins were tagged, they were isolated using affinity capture and identified by MS.

The use of proteomics as an underlying tool in molecular research has helped us stay the closest to the phenotype, though it is dynamic, and complex, varying according to cells, stages and environmental conditions.
ACKNOWLEDGEMENTS
First and foremost, I would like to express my special appreciation to my supervisor, **Susana** for giving me this opportunity and to make this thesis possible. This would not have been possible without your positive outlook and encouragement in all ups and downs. You have been a great support and with unlimited patience and heaps of tolerance.

My sincere gratitude to my co-supervisors, **Micke**, your undying knowledge, enthusiasm and passion for science has really inspired me. You always have the door open for me with the right advice and guidance. **Mats**, who have been very supportive and considerate enough to spare time for me.

I gratefully acknowledge **Linköping university, the funding sources and all committee members**.

My ex-coworkers, **Sara, Christine, Paula, Alex, Binu** and **Viviana**. Thank you all so much for all collaborations, inputs and talks, both in and out of the lab. **Binu, Sara and Christine** had been a source of great advice and true friends as well. My time in lab wouldn’t have been as much fun without you guys.

I would like to also thank my collaborators in Linköping, **Kazuki, Rajesh, Mahadesh, Josephine**, and all other members of **Sigvardsson lab** for all the excellent discussions and inputs.

My collaborators from Toronto, **Brian Raught, Tharan Srikumar, Etienne Coyaud** – Thank you for the valuable suggestions and guidance during the collaboration. It has been an amazing learning experience with you all.

It goes without saying that I survived my PhD because of my friends. **Sandeep, Shibi, Binu, Neena, Varun, Deepak, Jithu, Shani, Jenna, Sumit, Mayur, Dawn, Midhu, Feroz, Caisy, Kamal, Janaki, Dinesh, Jaibu, Manil and Chamilly** – I cherish our times spent together and would like to thank you all for always being there for me. My sincere thanks to **Sandeep and Sumit** for sparing time for me, the advices and feedbacks. Hugely indebted to **Sandeep** for his company and being the ultimate Maggyver.

I thank the administration for helping and supporting me. **Annelie Lindström**, you had been a great help always. I also thank **Anna Weinhofer** and **Hakan Wiktander** for being our problem solvers.

A sincere thanks to my previous work colleagues **Björn, Rickard, Masha, Hanna, Lan and Patrik**. It was with you guys that I started this journey and I couldn’t have wished for a better start. Thank you **Björn** for being so extremely nice and helpful. **Masha** has always been the expert when it comes to proteins and MS. My sincere gratitude for all the suggestions and inputs.
My sincere and heartfelt gratitude to my mentors Alex, Cornelia and Jordi for sharing their wisdom, knowledge and extensive experience.

I would also like to thank Stefan Klintström, Charlotte Immerstrand, Anette Andersson and other forum members for all the fun times together in conferences, study visits and monthly meetings.

I also want to thank my floor mates Teja, Mike, Olof, Antonio, Catherine, Jordi and all others in floor 13 for their lovely company and lively conversations. Teja, many thanks for all the help with R software and the great company.

Mummy, Papa and Monu – thanks for your love and continuous prayers.

I am indebted to my family for all their unconditional love and support. Daddy, Mamma, it is your efforts, prayers and well wishes that has made me stand where I am today. My sister, Sonia, who has taught me success is no accident, rather it is more of hard work, perseverance, sacrifice and blessings of the Almighty.

Last but never the least, Meenu for her continued and unfailing love, support, patience and understanding. I greatly value your contribution and deeply appreciate your belief in me.

With gratitude and prayers
Papers

The articles associated with this thesis have been removed for copyright reasons. For more details about these see:

http://urn.kb.se/resolve?urn:nbn:se:liu:diva-139867