Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Μηχανολόγων Μηχανικών
Τομέας Μηχανολογικών Κατασκευών & Αυτομάτου Ελέγχου

MECHANISM-BASED BIOMARKER DISCOVERY

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Asier Antoranz Martinez
ΔΙΠΛΩΜΑΤΟΥΧΟΥ ΜΗΧΑΝΟΛΟΓΟΥ ΜΗΧΑΝΙΚΟΥ ΕΜΠ

ΕΠΙΒΛΕΠΩΝ:
Λ.Γ. Αλεξόπουλος
Αν. Καθηγητής, ΕΜΠ

Αθήνα, Μάιος 2019

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Μηχανολόγων Μηχανικών
Τομέας Μηχανολογικών Κατασκευών & Αυτομάτου Ελέγχου

MECHANISM-BASED BIOMARKER DISCOVERY

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Asier Antoranz Martinez
ΔΙΠΛΩΜΑΤΟΥΧΟΥ ΜΗΧΑΝΟΛΟΓΟΥ ΜΗΧΑΝΙΚΟΥ ΕΜΠ

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ:
Λ ΑΛΕΞΟΠΟΥΛΟΣ, Αν. Καθ. ΕΜΠ (Επιβλέπων)
J. Saez-Rodriguez, Καθ. Heidelberg University
F. Planes, Καθ. Tecnun-Universidad de Navarra

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:
Λ ΑΛΕΞΟΠΟΥΛΟΣ, Αν. Καθ. ΕΜΠ (Επιβλέπων)
J. SAEZ-RODRIGUEZ, Καθ. Heidelberg University
F. PLANES, Καθ. Tecnun-Universidad de Navarra
Χ. ΠΡΟΒΑΤΙΔΗΣ, Καθ. ΕΜΠ
S. WAIKAR, MD, Brigham and Women’s Hospital
Α. ΒΛΑΧΟΥ, Ερευν. Γ. Bioacademy
Α. ΧΑΤΖΗΙΩΑΝΝΟΥ, Ερευν. Β. ΕΙΕ

Αθήνα, Μάιος 2019

Η έγκριση της διδακτορικής διατριβής από την Ανώτατη Σχολή Μηχανολόγων Μηχανικών του Ε.Μ. Πολυτεχνείου δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα (Ν. 5343/1932, Άρθρο 202)

Prologue

The research for this dissertation was carried out under the supervision of the Associate Professor of the Mechanical Engineering Faculty of NTUA, Leonidas G. Alexopoulos, and supported by the European Union Horizon 2020 grant Systematic Models for Biological Systems Engineering (SyMBioSys) MSCA-ITN-2015-ETN #675585.

During my PhD, I visited and collaborated with the following institutes/organizations: The Faculty of Medicine at the RWTH Aachen University in Aachen, Germany (prof. Julio Saez-Rodriguez, now Heidelberg University), the Department of Chemical Engineering at Imperial College London in London, United Kingdom (prof. Athanasios Mantalaris).

I would like to thank the following people/institutions/organizations for their help. Initially the international scientific community and the open-minded community for their self-serving contribution to tools, data and answers to technical questions. All members of the Biomechanics Laboratory of the Mechanical Engineering Faculty of the NTUA between 2015 and 2019 and especially, Dimitris Tzeranis, Dimitris Messinis, Giorgos Kanakaris, Jan Rozanc, Michael Neidlin, Danae Zareifi, and Christos for our cooperation and / or constructive discussions. All employees from ProtATonce LtD between 2015 and 2019 and specially Angeliki Mina, Nikos Tsolakos, Vaia Pliaka, Michalis Fragos and Sophia Stamatatou for their support and help during these years. All members from SyMBioSys, specially Joana Saldida, Romuald Gyorgy, Aurelien Dugourd, Michail Klontzas, Enio Gjerga, Jake Pitt, Nikolaos Tsiantis, Roman Rainer, Oulafemi Ademola Bolaji, Lucian Gomoescu, Sara Barbosa, Oshvaldo Kim, Jose Ferreira Rodrigues de Morais, and Maria Masid Barcon for making these years much more fun. All the staff members and specially Athanasios Mantalaris for his hospitality and Julio Saez-Rodriguez for his hospitality, leadership, help and guidance during my PhD. All the colaborators from the University of Athens, specially Ilias Doulamis and Aspasia Tzani for the delicious tsoureki which helped to make Athens even sweeter. Sushrut Waikar and Benajmin Humphreys for their help and support from the beginning of the project. The members of the three-member advisory committee Francis Planes for his guidance support and motivation all these years and the members of the seven-member committee of inquiry for their criticism. The members of the Secretariat of the School of Mechanical Engineering and especially Dimitra Poula for her support over the years.

Finally, I would like to thank the following: My supervisor, Leonidas Alexopoulos, for the guidance, motivation, and spiritual freedom he provided me. Teodore Sakellaropoulos, with whom I worked in the NTUA and in ProtATonce, for our discussions, dinners at Mexicanos and philosophical talks, which gave us all the ideas of this thesis. Francesca, who supported me physically and mentally ever since we met. My family, my mother and my sister who are always there whenever I need a smile, and my father who I wish could see everything has happened these last 10 years. All the people mentioned above, in different ways, offered far more than I could give them back, which is why I am grateful to them. The present work would not have been possible without them.

Asier Antoranz
Athens, April 19

Summary

Biomarkers are cornerstones of healthcare spanning a variety of applications from disease diagnosis to stratification and prediction of likely outcome. Despite significant efforts that have identified thousands of potential biomarkers, their translation into clinical practice remains poor, averaging 1.5 per year across all diseases. This inefficiency primarily results from the lack of connection of the candidate biomarkers with the underlying pathophysiological mechanisms that they monitor which results in poor reproducibility in their developmental pipeline. On top of these limitations, the current single-biomarker-to-single-disease approach does not capture the multifactorial nature of complex diseases like Chronic Kidney Disease (CKD). CKD is a major public health problem that affects approximately to 14% of the general population and requires asymptomatic, early-stage, and disease-specific, biomarkers to deliver more precise diagnostic and predictive information.

Here we propose, and experimentally validate, a biomarker discovery pipeline that aims to identify molecular signatures that not only allow to discern different CKD subtypes but also capture the underlying biology of the disease. To that end, first, we integrate protein-protein interaction networks with annotated gene-sets into a knowledge-base that captures plethora of information about biological entities and their interactions. This model is then fed with CKD transcriptional data to generate a disease specific model. Third, the model is analysed using different state-of-the-art methods (e.g.: network analysis, pathway analysis) which result in a molecular profile for each protein capturing different disease biology. Relevant features are then selected and optimized by training an elastic network model on plasma samples which is used to predict biomarker performance. Finally, the resulting individual candidates are integrated into a biomarker panel with increased performance and stability using linear discriminant analysis as machine learning integrative method. Results show that our holistic approach can find biomarkers that are associated with disease mechanisms while keeping competent predictive abilities.

Extended Summary

Modern high-throughput omic technologies like the widely used DNA microarrays in combination with the emergence of computational methods for multi-omics data analysis are allowing biomedical researchers to better understand complex diseases by unravelling their underlying mechanisms. This new field, Systems Biology, promotes the holistic description of the studied systems as opposed to the “traditional” approach that focuses on the study of its elements individually. DNA microarray studies, now collect the vast majority of protein encoding genes, thus allowing to capture most pathophysiological phenomena. Therefore, modern computational methods exploit the plethora of topological, functional, and contextual information about biological entities which are generally comprised into databases (protein-protein interaction networks, functionally correlated gene sets, etc.).

Complex and heterogeneous diseases like Chronic Kidney Disease (CKD) are to benefit most from these system-based approaches due to their high degree of polygenicity (groups of genes that can act together to produce an observable variation). CKD is a major public health problem that affects to approximately 14 percent of the general population according to the National Institute of Diabetes and Digestive and Kidney Diseases. Despite its prevalence, CKD is a syndrome that is heavily understudied, according to NIH data the investment in kidney research per affected patient is 10% of that of aids and 1% of that of cancer in the United States. CKD is an asymptomatic disease in its early stages and can go undetected until it is very advanced, when patients may experience weakness related to anaemia and proteinuria (presence of large quantities of protein in urine which often causes the urine to become foamy). All modern CKD categorizations (Risk Injury-Failure-Loss-End Stage Kidney Disease [RIFLE], Acute Kidney Injury Network [AKIN], or Kidney Disease, Improving Global Outcomes [KDIGO]) use the estimated Glomerular Filtration Rate (eGFR) to diagnose kidney injury. Despite a single measurement of eGFR and proteinuria are generally used to identify CKD, the abnormalities should persist for at least 3 months in order to complete its diagnosis. In adults, the most common ways to calculate eGFR are using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) and the Modification of Diet in Renal Disease (MDRD) equations, both of which use serum creatinine (sCr) as biomarker to calculate eGFR. However, this measurement lacks sensitivity (kidney damage is not the only factor that affects sCr levels) and fails on further characterizing CKD into subcategories (sCr its indifferent to the pathophysiology of the injury). On the other hand, second generation biomarkers like neutrophil gelatinase-associated lipocalin (NGAL), and kidney injury molecule-1 (KIM-1) have refined the diagnosis of CKD but fail to subfractionate CKD further.

Predictions based on single-molecules are usually ineffective for complex diseases and fail in their progress to clinical trials due to their lack of connection to the mechanism of the disease. On the other hand, most of commercially available products use panels of genes like the MammaPrint® test, a panel of 70 genes for breast-cancer prediction. In the case of CKD, although recent studies have shown that gene expression in the kidney and protein signatures in external biofluids (serum, urine, etc.) can distinguish different forms of CKD, implying a complex system responding to different factors; such a test including disease-specific biomarkers does not exists yet and may be required. Hence, mechanism-based biomarker discovery aims to identify molecular signatures that not only capture the underlying biology of the disease, but also allow to identify its diversity distinguishing the subtype of origin and permitting more effective therapies.

Methods

The present project deals with a proteomic biomarker discovery scheme that has as input gene expression data from patients. CKD is used as case study.

In gene expression, the intensity is defined as the number of RNA copies that is proportional to the concentration of the protein produced. For the purposes of this thesis, microarray DNA type measurements were used from public databases (NephroSeq, Gene Expresion Omnibus). These technologies have the ability to measure the expression of a predefined number of genes in many samples.

In the case of proteomics instead, for the present study, the measurement of concentrations of proteins was performed by the bead-based sandwich ELISA technique which allows measuring a predetermined number of proteins in many samples.

The pipeline presented in this thesis combines prior knowledge from publicly available biological models databases with state-of-the-art algorithms that aim to capture meaningful biology for the disease. The goal is to suggest a number of proteins as candidate biomarkers to be measured in blood samples that are able to diagnose a disease subtype, and use machine learning methods to integrate them into a panel of biomarkers with increased performance and stability. The cornerstone is to molecularly determine the underlying pathophysiological mechanism causing CKD and detect it before the damage is done. Briefly, our methodology includes 5 steps:

1) Integrate protein-protein interaction (PPI) networks with annotated gene-sets to construct the knowledge-base (KB) which comprises contextual information about the biological entities participating in our model. There exist several alternatives in both cases for the selection of the biological model, each of which being different from the previous one. A recursive motto in this thesis is that of the 'wisdom of the crowds', meaning that when several options are available without any apparent reason to prioritize any of them, their combination is probably closer to the truth. Consequently, for both biological models, we chose those that encapsulate these ideals in their construction by integrating several sources into a single database, that is, the Molecular Signatures database (MSigDB) for the gene sets, and OmniPath for the PPI.

2) Feed this KB with CKD data obtained from the Gene Expression Omnibus (GEO) to create disease-specific models. CKD data was obtained from different studies which needed to be merged before using them for any type of analysis. Gene expression microarray measurements can be affected by small differences in any number of non-biological variables, so different instruments, technicians, reagent lots, or even days in which the experiments were carried out can affect the data. Heterogeneous datasets make integration challenging, so different steps removing incomplete features, observations, and under-represented disease groups were applied. Batch effect adjustment was addressed using ComBat.

3) Apply different state-of-the-art omic-analysis methods (network analysis, pathway analysis, etc.) to describe a molecular profile for each biological entity included in the model. Our holistic approach uses three different types of analysis that study the system in three different biological levels: molecular-level, pathway-level, and network-level. For the molecular-level, the integrated gene expression dataset was analysed with Linear Models for Microarray Data (Limma), a gold-standard for the analysis of microarray data. For the pathway-level, the

method of choice was piano, a gene set analysis method that integrates 11 different algorithms in its pipeline. For the network-level, a custom method was applied which focuses in the surrounding area of each protein based on the PPI.

4) Select and optimize the features included in the above-described molecular profile by training an elastic network model on plasma samples. Apart from literature, we don't have information about which proteins (from the ones included in our model) are more likely to perform as biomarkers. We don't know either which of the obtained features in the analysis above are useful for their prediction. We tried to answer these questions by measuring a relatively small number of proteins and calculating their performance as biomarkers using also a custom equation. Our equation for biomarker performance integrates quantitative and qualitative metrics into a single value that is used to train a prediction model. Based on these predictions, a second round of proteins is measured for validation purposes. The motivation behind the use of blood samples is that, for monitoring human health, measuring protein biomarkers in blood is considered a very attractive solution because the pathology of almost every body tissue can affect the blood proteome on top of the simplicity of obtaining the sample.

5) Integrate the individual candidates into a biomarker panel with increased performance and stability. As described in the introduction of this section, biomarkers work best when integrated in panels. This is particularly true for polygenic diseases like cancer or CKD as in this study. Single proteins cannot capture the complexity and the different mechanisms that are affected by these pathologies. To that end, we exhaustively trained panels of size 1 to 7 trying all possible combinations of the measured proteins for each size. Then, the best performing panel was selected for each size to construct the Pareto front. Finally, the optimal panel was chosen based on the elbow criterion.

Results and discussion

Our results show that the pre-processing step, which involves, data integration and batch-effect removal is the most critical phase of the whole pipeline as most of the downstream analyses depend on the expression results. To a lesser extent, but also of great importance, the selection of biological models chosen to construct the KB largely affect the obtained results.

Our analysis also highlights one of the basic aspects of systems biology. Complex phenotypes are not explained by changes in individual molecules but rather small and coordinated changes in functionally correlated molecules originate those phenotypes. In fact, the results in the molecular-level do not correlate with biomarker performance in the way that the ones in the pathway- and network-level do. Moreover, the best performing proteins individually (TFF3 [BP = 96.08], and ICAM1 [BP = 90.24]) did not show a significant change in the transcriptomic level. In fact, from the introduced analyses, the network-level features are the ones correlating best with biomarker performance.

Regarding the performance of the proteins individually, our method was able to find strong biomarker candidates when gene-expression level statistics were strong (CKD vs healthy). On the other hand, the separation between disease subgroups at the gene level was not strong, and as such the biomarker diagnostic power was lower. This trend was observed in all the analysed levels (transcriptomics, pathway, network, proteomics). On this front, some proteins showed a moderate diagnostic ability to separate disease groups as is the case of Galectin 3 for Diabetic Nephritis (DN) (BP = 72.36) and MMP9 for Lupus Nephritis (LN) (BP = 73.66). In fact, these two proteins together with the previously mentioned ICAM are the ones that were combined for the conformation of the optimal panel.

Table of Contents

List of Figures

List of Tables

Chapter 1 Biomarkers

This chapter represents a summary of the biological and technical concepts surrounding biomarkers, including the different types and applications of biomarkers, their developmental pipeline, and how to evaluate their performance.

Introduction

Etymologically, the word biomarker derives from biological and marker, and even if this definition seems straightforward enough, there has been a lot of ambiguity around this term and different authors [1, 2] suggest that its meaning has been diluted by overuse. The most widely accepted definition of the term biomarker was proposed by the Biomarkers Definitions Working Group from the National Institutes of Health (NIH) in 2001 [3]. According to that group, a biomarker is defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention". Hence, in the broad sense, a biomarker may be a distinctive trait or feature with diagnostic, prognostic and/or predictive potential [1].

Currently, there is an overwhelming interest in biomarker research as the volume of literature dedicated to biomarker discovery, characterization and validation continues to increase every year [4]. From the first incorporation of the term biomarker in a journal article title in 1980 [5] [6] (from PubMed), the number of journal citations including the term biomarker has increased reaching more than 17,000 in the year 2000 and more than 53,000 in the year 2017 (Figure 1).

Types of biomarkers

Based on application

Biomarkers can perform diverse functions in medical research and are often classified according to their intended use. Biomarkers are most commonly applied to diagnosis, prognosis, and prediction [7].

Diagnostic biomarkers: Diagnosis is related with determining the current state of the patient and accurately identifying an existing, but unknown, disease state. Diagnostic biomarkers are those that discriminate subjects as having or not having a specific disease or condition, or more generally, distinguish between several possibilities (disease severity, staging) [7]. Therefore, diagnosis is a problem of classification which is liable to the same considerations attributed to classification problems. Examples of diagnostic biomarkers include elevated blood glucose concentration for the diagnosis of diabetes mellitus [8] or heart rate variability as a biomarker of fibromyalgia syndrome [9].

The differentiation between pathological and normal is not a trivial issue and involves the need to define "normal" to establish a reference interval to which patients' results can be compared [10]. The normal range, or reference interval, is the value expected for 95% of the population (National Cancer Institute; http://www.cancer.gov/dictionary/?CdrID=635450), although this cut-off value can be modified depending on each case. The establishment of a reference interval can require a large population so is usually not considered until the process of transition from biomarker discovery to validated assay [1]. Moreover, the concept of normal is confounded by the index of individuality and biological variation. The index of individuality is the intra-individual variation of a measurement while the biological variation represents the inter-individual disparity. Therefore, a diagnostic biomarker requires to characterize its "normal" levels within and between subjects and assess whether it varies by age, gender, and ethnicity.

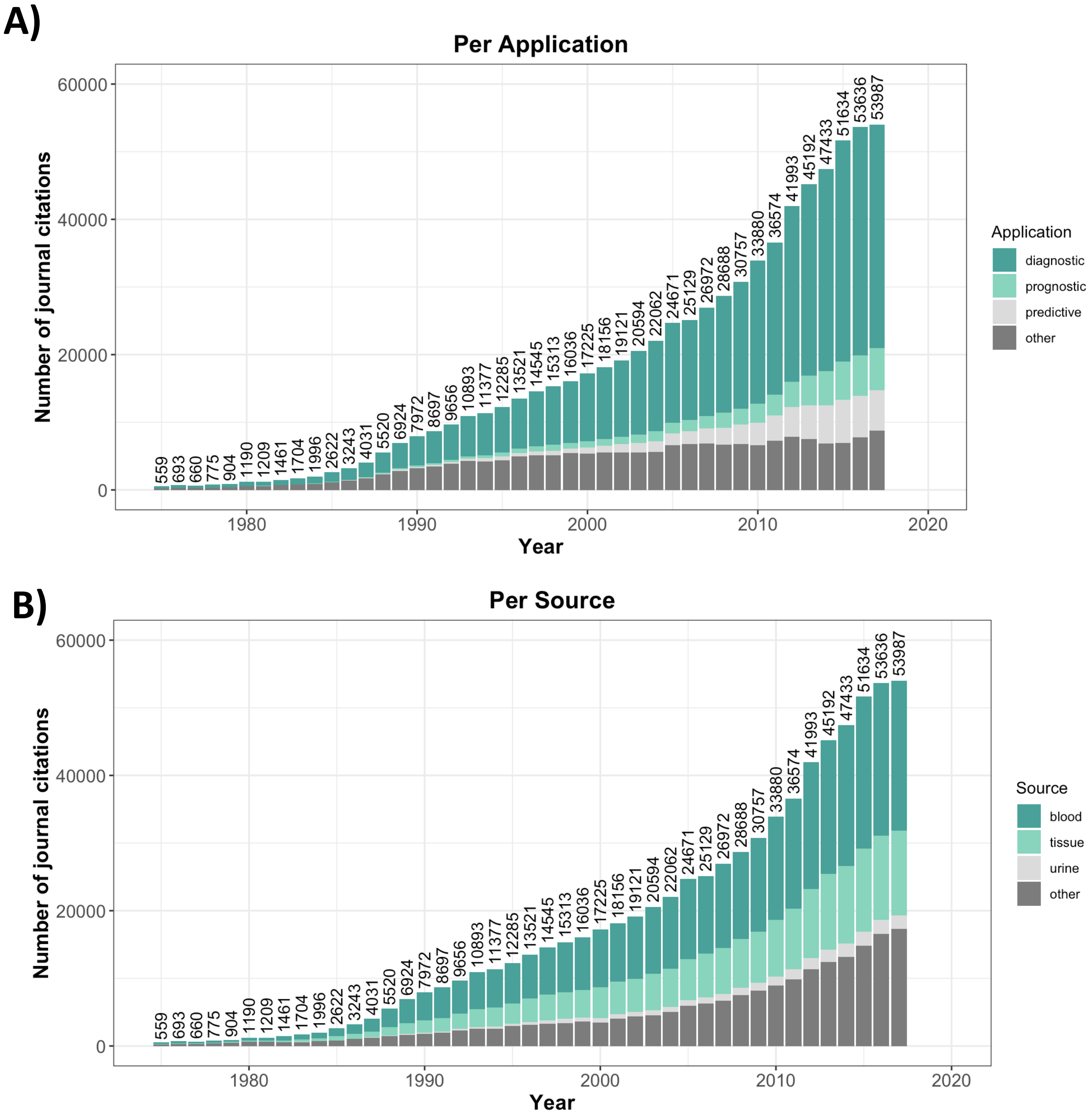


Figure 1 – Number of journal citations with the term biomarker*. Each bar corresponds to a citations per year; A) Classification per application (diagnostic, prognostic, predictive, other), B) Classification per biological source used (blood, tissue, urine, other).*

Prognostic biomarkers: Prognostic biomarkers are related to the prediction/risk of a future event [7]. These biomarkers add the element of time, and therefore, a stochastic element, one that is subject to chance. The outcome is not only unknown, but does not yet exist, distinguishing this task from diagnosis [11]. In diagnostic biomarkers, single measurements may be enough but in prognostics they are of little value. The problem is no longer only classification but also calibration and require a clinical assay to identify the likely course/outcome of a disease. For a prognostic biomarker to be beneficial it requires an available therapeutic option for influencing a negative outcome [1]. Examples of prognostic biomarkers include the Framingham risk score, which predicts the 10-year risk of cardiovascular disease [12] or the anatomic measurement of tumor shrinkage as prognostic biomarker for certain cancers [13].

Predictive biomarkers: Predictive biomarkers are related to identifying the group of patients that would be affected/resistant by/to a specific therapy. They refer to the use of a specific analyte or group of analytes to examine a large population for the purpose of identifying a specific pathophysiology. Therefore, prediction is a matter of screening. Predictive probability is largely dependent on prevalence, and somewhat less on sensitivity and specificity of the assay system [1]. Examples of predictive biomarkers include blood cholesterol concentrations for determination of the risk of heart disease [13] or PD-L1 expression as a predictive biomarker in cancer immunotherapy [14].

Although the same biomarkers may serve different purposes, most often different experimental designs, data types, and analytical methods are needed to assess biomarker performance in each case. For example, the performance of diagnostic markers can normally be established using data from a single time point, while prognostic marker assessment usually requires at least two time points [7].

Based on source

Another common classification of biomarkers is based on the source material in which they are measured. Biomarkers can be obtained from different sources like tissue samples or biological fluids such as blood, urine, or saliva.

Blood biomarkers: Blood is the most common source of biomarkers representing approximately 50% of the biomarker research efforts every year (figure, 1.B). There are many reasons to justify the use of blood as a source of potential biomarkers: blood is perhaps the most popular source for biological samples and is the preferred material for a final diagnostic test [1]. Moreover, it is reasonably easy to obtain samples; the samples are technically easy to process, and they are mostly considered homogeneous when compared to saliva or urine. Whole blood may serve directly as sample but more commonly the blood is fractionated into plasma or allowed to form serum. It must be emphasized that plasma and serum are not interchangeable terms and that there are gross analytical differences between these two biofluids [15]. Serum is the liquid fraction of whole blood that is collected after the blood is allowed to clot while plasma is produced when whole blood is collected in tubes that are treated with an anticoagulant. Serum is plasma without red/white cells nor clotting factors. An example of a blood biomarker is microRNA-210 in acute cerebral ischemia [16].

Urine biomarkers: Urine is a biological fluid that has recently drawn attention as a potential source of biomarkers. Urine biomarkers represent different advantages and disadvantages compared with those obtained from blood even if both fluids have a considerable history of use in clinical chemistry. The collection of urine is technically easier to perform as it does not require venepuncture; however, blood is likely a more reproducible source as issues such as flow rate do not need to be considered [1]. It is common to express the concentration of urine biomarkers relative to creatinine concentration [17] to correct for urine dilution. Creatinine, formed from creatine, is a by-product of musculoskeletal exercise which, in normal conditions, is secreted through urine and its levels are considered to be constant within each individual [18]. Examples of urine biomarkers include albumin as a biomarker for cardiovascular and renal disease [19], S100 as a biomarker for head injury in children [20], or VEGF as a biomarker for cancer [21].

Apart from the abovementioned biofluids, there are other (less investigated) biological fluids/excretory products of interest as biomarkers. Examples include salivary cortisol as a biomarker for stress [22], fecal waters as a biomarker for colorectal cancer [23], exhaled breath condensate as a biomarker for pulmonary function [24], and sweat as a biomarker for the diagnosis of cystic fibrosis [25].

Tissue biomarkers: Tissue biomarkers are not as attractive as biofluids due to a higher invasiveness of obtaining tissue samples and because a greater level of processing (tissue fixation, preparation of sections, staining, and imaging) is required. However, demonstration of a biomarker in a tissue section is an unequivocal evidence for the presence of said biomarker [1]. Advances in single cell technologies are of high interest in the area of tissue biomarkers as they can help to unravel the underlying mechanisms and better understanding of complex diseases. Examples of tissue biomarkers include HCCR oncoprotein as diagnostic biomarker for breast cancer [26] or CD63 as a prognostic biomarker in lung cancer [27].

Based on nature

Historically, biomarkers have often been either physical traits (e.g. skin pigmentation), or physiological metrics (e.g. heart rate). Nowadays, however, with the rapid progression of -omic technologies, the term has become a shorthand for molecular biomarkers [10]. Nevertheless, non-molecular biomarkers (specially imaging-based) are still being focused by the scientific community. Examples of these include blood pressure as a potential biomarker of the efficacy of angiogenesis inhibitor [28], positron emission tomography (pet) images as a potential biomarker for Alzheimer's disease [29], or magnetic resonance (MR) images as a potential biomarker for treatment response in oncology [30].

Molecular biomarkers: The rapid advancement of molecular biomarkers is partially due to the parallel development of -omic technologies that are allowing us to measure more, and more reliably, different molecules. Molecular biomarkers can be of various types including proteomic biomarkers [10], genomic biomarkers [31], metabolic biomarkers [32] and so on. Each type utilizes different technologies for its measurement and as a consequence a diversity of strategies have been implemented for their discovery. For a review about -omic technologies applied to kidney see [33].

Despite all the different types of biomarkers mentioned in the previous paragraphs, from this point onwards, we will use the term biomarker as a shorthand for diagnostic blood proteomic biomarker as these will be the focus of the present thesis. Genomic biomarkers are currently used for diagnosis of monogenic diseases like polycystic kidney disease, which are caused by a mutation in a single gene. However, complex diseases like chronic kidney disease or cancer are often polygenic (caused by small but coordinated changes in several different genes). In this way, proteomics has been the enabling technology absorbing most of the biomarker research efforts [1] as the protein domain is likely the most affected one in disease, response, and recovery. This makes proteomics the ideal territory for biomarker discovery as proteins provide a broad picture of patient phenotype [10]. Among all biofluids, human blood is the most attractive option as it contains more different proteins (estimated to contain at least 10,000 different proteins) [34]. Moreover, blood has been described as the most comprehensive human proteome, with the ability to represent physiological and pathological processes from all body tissues [34].

The biomarker development pipeline

Analogously to drug development, the biomarker pipeline is a multi-step process that typically involves numerous studies from preclinical, to confirmatory human trials, although the pharmaceutical industry has more than 100 times funding than the one of protein diagnostics [34]. A well-organized biomarker development project typically evolves from an unbiased experimental approach that emphasizes on comprehensive protein characterization via mass-spectrometry (MS) to a targeted approach that emphasizes on development of high-throughput antibody-based assays. The different steps in between differ in the experimental methods, as well as analytical tools used, and represent a tradeoff between the number of samples and analytes included [10] (Figure 2).

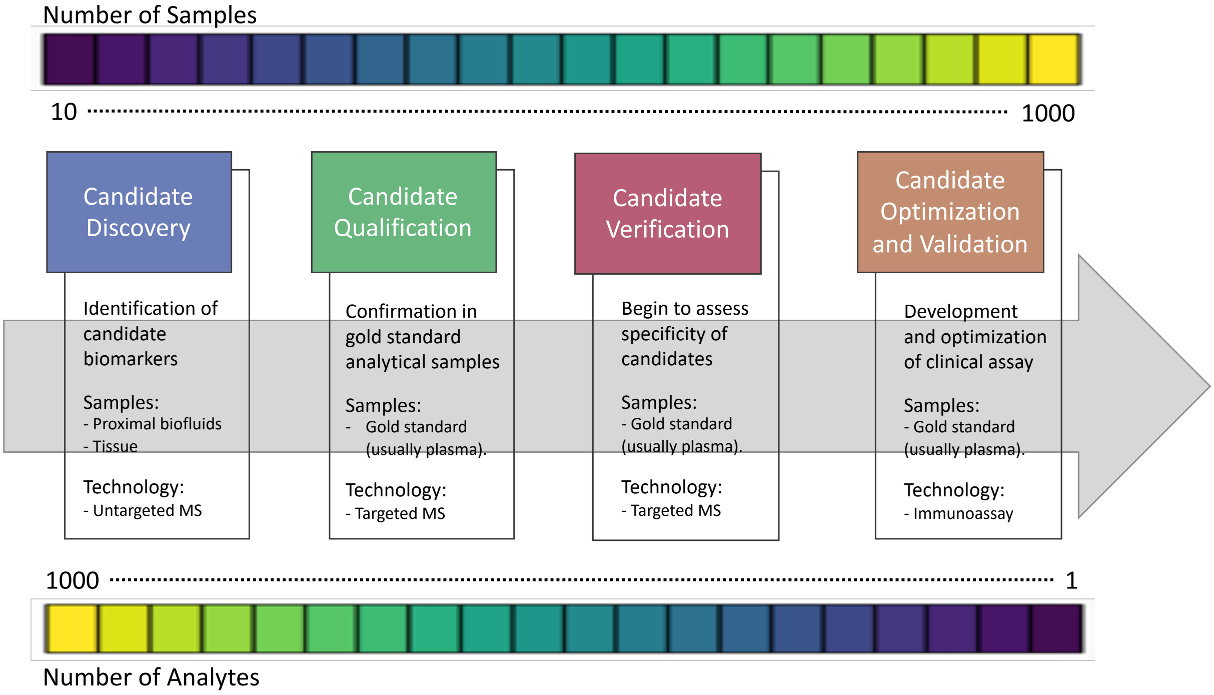

Number of Samples
10
1000
Candidate Discovery
Identification of candidate biomarkers
Samples:
- Proximal biofluids
- Tissue
Technology:
- Untargeted MS
Candidate Qualification
Confirmation in gold standard analytical samples
Samples:
- Gold standard (usually plasma).
Technology:
- Targeted MS
Candidate Verification
Begin to assess specificity of candidates
Samples:
- Gold standard (usually plasma).
Technology:
- Targeted MS
Candidate Optimization and Validation
Development and optimization of clinical assay
Samples:
- Gold standard (usually plasma).
Technology:
- Immunoassay
1000
1
Number of Analytes

Figure 2- The biomarker development pipeline. *Discovery involves the identification of hundreds of biomarker candidates differentially expressed via MS; Quantification emphasizes on the confirmation of differential expression in gold standard samples (usually plasma); Verification focuses on the assessment of specificity of the candidates; Validation concludes with the development of a clinical assay.*

Candidate discovery: Discovery is the unbiased and semiquantitative process by which the differential expression of proteins between specific conditions is first defined [10]. Discovery is not limited to specific species (e.g., animal models or cell lines) or biological materials (tissue, blood, etc.) and typically involves a simple two-group comparison (e.g., healthy vs disease) avoiding considerations about possible co-founding factors (gender, age, ethnicity, etc.). The output of the discovery phase is a list of proteins differentially expressed between the conditions of interest, which can be expanded with proteins drawn from the literature. Differential expression is typically assessed via semiquantitative untargeted MS. The discovery phase is known to have a high false positive ratio; i.e., proteins that during discovery showed differential expression but upon further testing, in fact, they do not. Because of their high false-discovery rate, these proteins are called 'candidate biomarkers' rather than biomarkers.

Candidate qualification: Qualification translates the findings from discovery by confirming differential expression of the candidate biomarkers in binary comparisons using alternative, targeted methods and human plasma samples [10].

Candidate verification: Verification extends the analysis to hundreds of human plasma samples, now including not only binary comparisons, but a broader range of cases and controls. This increase in patient number allows to start assessing the biological variation of the candidates in the tested population and the possible co-founding factors affecting candidate levels [10].

Biomarker validation: Validation tests many thousands of samples and aims to capture the full variation of the targeted population [10]. Because the final goal is the development of an in-vitro diagnostic (IVD) test, or in some cases, a point-of-care (POC) device, biomarker validation is closely related with the optimization and the validation of the test assay in which the biomarker is being measured (technical validation precedes biological validation). These tests are typically immunoassays (e.g., ELISA) that require full characterization of the immunoaffinity reagents to meet the rigorous standards required for clinical tests [10].

We can see that besides the experimental shift from untargeted to targeted approaches, the biomarker pipeline includes another important state change. Discovery and qualification focus on the biological association between marker and phenotype, while assay optimization and validation also put emphasis on the test being used for the measurements which can dramatically enhance signal-to-noise, streamline processing and enhance throughput [10]. Both phases are essential as one of the main problems with blood proteomic biomarkers is the lack of reproducibility. This lack of reproducibility is directly linked with the record of transition from discovered biomarker to clinical assay (suggested as 1 per year since 1998 [10]) which is, in part, due to the biomarker pipeline and the cost of obtaining data sufficient for approval, but more importantly to the lack of connection of the vast majority of biomarker candidates with the underlying biology of the system in question [1]. Too often the biomarker pipeline is oversimplified into a direct line relationship between hypothesis generation and testing [7]. However, it is important to determine whether the newly identified biomarkers are mere associations or real biomarkers of disease state representing pathophysiological processes (mechanistic biomarkers) [35]. The term biomarker then is not only a laboratory analyte, is more of a concept where the biomarker is closely related to the biology underlying its production [1]. The use of the term biomarker as a response to an environment stress or in response to a therapeutic intervention seems quite reasonable. However, the vast majority of statistical approaches fail con capturing the most important ingredient, causality [36]. Statistical correlations solely do not make biomarkers. I cannot emphasize this point strongly enough as an understanding that capturing the underlying biology of a disease with a biomarker will increase the chances of developing an assay, robust, reproducible and with potential to progress into the clinic.

Mechanism Based Biomarker Discovery

Traditionally, biomarker discovery was hypothesis driven and therefore related to extensive biological research to characterize pathological processes for identification of potential biomarker candidates. In contrast, the rapid development of -omics technologies over the last years initiated a paradigm shift into data-driven hypothesis-free multi-parametric profiling approaches for biomarker discovery [37]. However, as previously mentioned, this data-driven approach does not capture causal relationships between disease and marker showing little success as very few new biomarkers were introduced into clinical practice (averaging 1 per year) [38]. Therefore, a new paradigm shift is taking place, and modern biomarker discovery exploits existing knowledge about biological entities and their interactions as well as case-specific data to characterize pathological processes related to disease phenotype and identify surrogate endpoints that respond to alterations in those processes.

Complex and heterogeneous diseases like CKD are highly polygenic, rarely originated due to alterations in individual genes/proteins. In these cases, individual variants usually have a small effect size; thus, they make a limited contribution to the phenotype independently. In fact, the phenotype of complex diseases is more likely explained by complex physiological interactions, where multiple functionally connected variants act in concert causing a particular phenotype [39]. Therefore, this new paradigm shift encourages the analysis of complex phenotypes at various levels including simpler levels like alterations in single genes, and higher and more complex levels as in the case of gene sets, pathways, and biological networks.

In this context, pathways (represented as gene sets) embody one of the simplest and most commonly used units to model functionally correlated biological entities. In this case, similarly to aggregating genes from genetic variants, genes are grouped into sets. One of the main benefits of working with pathways instead of individual entities is that pathways represent biological processes enhancing the interpretability of the generated results. Therefore, pathway analysis has become the first choice for gaining insights into the underlying biology of differentially expressed genes and proteins [40]. We are no longer only interested in big changes in individual entities, but also in relatively small changes in groups of biologically/functionally correlated ones.

In pathway analysis, the unit of analysis is the gene set, which is any group of genes that shares a particular property, and the aim is to determine if the property has a role in the phenotype under study [41]. During the last years, a lot of different pathway analysis methods have been published each of then trying to tackle different limitations of the generic idea of pathway analysis including: biological crosstalk [42, 43], gene weights [44], pathway topology (for a review see: [45]), or different multilevel omics data [46], which has caused their aggrupation based on generation (Table 1) [40]. However, none of the available methods acts as gold standard as none of them has been proven to work better than the others. Therefore, a common practice is to use a number of different methods and find consensus among their results, a practice that has proven to outperform methods individually [47].

Table 1 - Generational classification of pathway analysis methods. *Pathway analysis methods can be classified in three generations, each following a different scheme.*

Generation	**Scheme**
1st generation, Over-Representation Analysis (ORA)	• Differential expression analysis • Threshold-based differentially expressed (DE) Genes • Number of DE genes relative to the number of genes in each pathway
2nd generation, Functional Class Scoring (FCS)	• Differential expression analysis • Gene-level statistics • Pathway-level statistics
3rd generation, Pathway Topology (PT) based	• Differential expression analysis • Pathway topology • Pathway-level statistics

The two-tier structure of gene set analysis (apart from differential expression analysis) allows to further classify all these methods according to their statistical properties, i.e., null hypothesis and the gene-level statistic (Table 2). Different results depend basically in these two properties with null hypothesis being more significant than test statistic [48] [49].

Table 2 Statistical classification of pathway analysis methods. *The two most significant properties affecting the obtained results are the null-hypothesis and the gene-level statistic.*

Based on null hypothesis		**Based on gene statistic**	
Type	**Description**	**Type**	**Description**
Self-contained analysis **θ_S > Ø?**	Hypothesis that only considers results within a gene set of interest.	**Fold-change (FC)**	Statistic describing how much a quantity changes from an initial to a final value.
Competitive analysis **θ_S > θ_0?**	Hypothesis that compares the results from genes within a gene set of interest with results from genes outside the gene set.	**T-value**	Statistic describing the size of the difference relative to the variation in the sample data.
		P-Value	Statistic describing the evidence against the null hypothesis.

Biological networks represented as protein-protein interaction (PPI) networks, on the other hand, embody a higher level of complexity where instead of aggregating groups of genes into sets one can analyze entire biological systems. Most of network analysis methods focus on identifying altered subnetworks (modules) for which lots of methods are available. These methods can be broadly classified in three groups. The first group uses a series of candidate subnetworks based on prior knowledge to then select the higher-scoring candidates based on their node values. Examples include MUFFIN [50], and NetSig [51] The second group of methods jointly analyzes network topology and node scores in order to identify subnetworks that may be affected by a particular phenotype. Examples include HotNet2 [52]. Finally, the third group of methods incorporate additional information about the interactions between the nodes, thus using more complex representations for the interactions between proteins. These methods need a more complex prior knowledge network. Examples include PARADIGM [53] and HIT'nDRIVE [54].

Biomarker performance evaluation

“Validation” is an extensively used term in biomarker development to refer to the validation of the biomarker as opposed to the validation of the assay. If we are to use KIM-1 as a biomarker for acute kidney injury, we have to make sure that we have a robust and reliable way to measure such marker. The validation of a biomarker is unrelated to the validation of the assay for the biomarker, which is associated with the accuracy, reproducibility, and robustness of the assay as well as its applicability (invasiveness, etc.) [1]. Thus, there is first the identification and validation of the biomarker, and then the validation of the assay of the biomarker in a clinical setting.

To recognize the role of statistics, we must first understand that just as there are multiple steps in biomarker development, there are diverse styles of data analysis. One of the most broadly accepted discriminations is the disparity between exploratory data analysis (EDA) and confirmatory data analysis (CDA) [55]. EDA is hypothesis-free, flexible, and gives a sense of “what seems to be going on”, while CDA is hypothesis-driven and should be reserved for instances in which the number of samples and the data quality is high [7]. Most often there is not a known set of optimal statistical methods or a single analytical pipeline that works best. On the contrary, it is a good practice to test our hypotheses using alternative mathematical models and different patient cohorts to validate our results as “confirmation comes from repetition”. If the information is contained in the data, most of the available analytical methods should be able to retrieve it. Until reaching validation when CDA becomes mandatory, all the previous steps of the biomarker development pipeline can be satisfied with EDA that can be further categorized as qualitative and quantitative evaluations. Quantitative evaluation includes differential expression analysis most commonly used in transcriptomics analyses. In the qualitative side, however, particularly in the discovery phase, ROC analysis remains the gold-standard.

Diagnostic predictability (statistical classification)

In binary cases when the population under study can take only two outcomes (e.g., healthy vs disease) it is natural to think of performance in terms of the proportion of correctly classified patients (accuracy). For tests that take continuous values (e.g., analyte concentration) it is first required to dichotomize the output values using a threshold. Results from such a test are typically collected in a confusion matrix (Table 3). Despite looking fairly simple, confusion matrices open a myriad of ways to analyze them deriving a large number of performance indicators each communicating different information. Some of the most common indicators extracted from these tables include sensitivity, the proportion of positive samples detected as positive; and specificity, the proportion of negative samples detected as negative [56]. Indicators from the confusion matrix can be classified in threshold-based and threshold-free. Table 4 includes a comprehensive list of the threshold-based performance indicators that can be obtained from a confusion matrix.

Table 3 - Confusion matrix for the binary case.

		REAL		
		Positive	**Negative**	**Total**
Predicted	**Positive**	TP	FP	PP = TP + FP
	Negative	FN	TN	PN = FN + TN
	Total	RP = TP + FN	RN = FP + TN	N = TP + FP + FN + TN

Table 4 - Threshold-based performance indicators.

Statistic	Formula
Prevalence	$P = {}^{RP}/_{N}$
Sensitivity, recall, probability of detection, true positive rate	$TPR = {}^{TP}/_{RP}$
Specificity, selectivity, true negative rate	$TNR = {}^{TN}/_{RN}$
Fall-out, probability of false alarm, false positive rate	$FPR = {}^{FP}/_{RN}$
Miss rate, false negative rate	$FNR = {}^{FN}/_{RP}$
Precision, positive predictive value	$PPV = {}^{TP}/_{PP}$
False discovery rate	$FDR = {}^{FP}/_{PP}$
False omission rate	$FOR = {}^{FN}/_{PN}$
Negative predictive value	$NPV = {}^{TN}/_{PN}$
Accuracy	$ACC = {}^{(TP + FN)}/_{N}$
Error rate	$ERR = {}^{(FP + FN)}/_{N}$
Matthew's correlation coefficient	$MCC = \dfrac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}$
Positive likelihood ratio	$LR(+) = {}^{sensitivity}/_{(1 - specificity)}$
Negative likelihood ratio	$LR(-) = {}^{(1 - specificity)}/_{sensitivity}$
Diagnostic odds ratio	$DOR = {}^{LR(+)}/_{LR(-)}$

The choice of an adequate threshold is a crucial part in the creation of a test and thus, the confusion matrix. Generally, researchers may choose a threshold that maximizes/minimizes a specific indicator of performance (e.g., maximum accuracy), although occasionally, researchers have particular reasons to favor a particular threshold. These reasons are normally of biological nature rather than statistical ones. However, the selection of a threshold is usually subject to the group of samples included in the study and can be problematic if this number is low (see the need of cross-validation). There are ways to go around this problem and assess the performance of a marker over a range of possible thresholds which is one way to motivate analysis of a receiver-operator characteristic (ROC) curve [57]. To generate a ROC curve, instead of creating a confusion matrix for every possible threshold, the indicators of interest (usually sensitivity and 1-specificity) are depicted for different thresholds and plotted in the XY axes generating the curve. This curve includes all relevant thresholds, without any specific threshold being made explicit on the plot. Different biomarkers can be shown in the same plot to compare their performance (Figure 3). One of the most attractive characteristics of ROC curves is that any monotonic transformation of the biomarker values (e.g., scaling, centering, normalization, log-transformation, standardization, etc.) will result in the same ROC curve [7]. This is because ROC curves are based on the order of the samples and not their actual differences in value (qualitative versus quantitative analysis). The most common threshold-free indicator of performance extracted from the ROC curve is the area under the curve (AUCROC, c-statistic, c-index) which combats the arbitrariness of comparing results at a single threshold. The AUCROC is essentially a weighted average of the sensitivity estimates over all the thresholds across the 1-specificity values and is one way to collapse information of the whole curve into a single performance indicator [7].

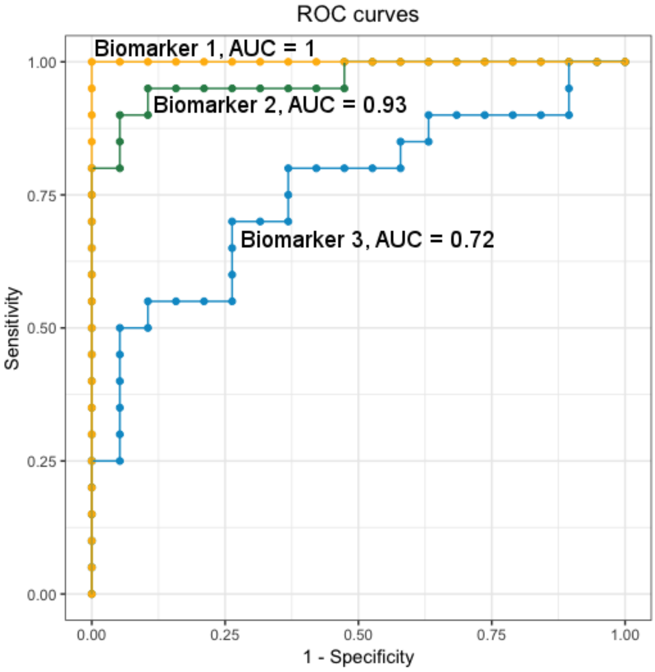


Figure 3 – ROC curve example. *The classificatory power of a biomarker can be expressed using threshold-based (sensitivity, specificity, etc.) or threshold-free (AUC) approaches. ROC curves can also be used to compare the performance of different biomarkers.*

Class-Imbalance

Class imbalance, a difference in the number of positive and negative instances where usually the negatives outnumber the positives can have a huge impact in the development and evaluation of predictive models and classifiers in many bioinformatic studies. The visual interpretation of a ROC curve in the context of a heavily imbalanced dataset can be misleading with respect to the performance of a biomarker. Precision-recall (PR) curves, on the other hand, are more robust in cases where the negatives outnumber the positives by evaluating precision (the fraction of true positives among positive predictions) versus sensitivity [58]. Other less popular options include concentrated ROC (CROC), and Cost Curves (CC). However, the precision-recall curve loses its robustness when the positives outnumber the negatives (in such cases the negative predictive value can be used). On the other hand, the Matthew's Correlation Coefficient (MCC) is an indicator of performance calculated from all four values of the confusion matrix that remains robust for any type of imbalanced dataset. While features like precision or negative predictive value are easier to interpret, they cannot be generalized for every imbalanced dataset. Moreover, the MCC shows improved reproducibility than traditional indicators of performance (e.g., accuracy) when facing the problem of choosing the optimal threshold using different datasets (Figure 4). We proved this by performing an in-silico experiment using two different populations (Positive, and Negative). Positive follows a normal distribution with average of 10 and standard deviation of 1 while Negative follows a normal distribution with average of 15 and standard deviation of 2 (Figure 4.A). Both populations were samples using three different experimental conditions (Figure 4.B): one where the number of positives outnumbers the number of negatives (P>N), one where the number of negatives outnumbers the number of positives (N>P), and one where both populations have the same number of samples (N=P). We calculated the optimal threshold maximizing ACC or MCC (Table 4) for all the experimental conditions and found that using MCC as statistic for the problem of optimal threshold returns more consistent results in all cases across 100 iterations (Figure 4.C).

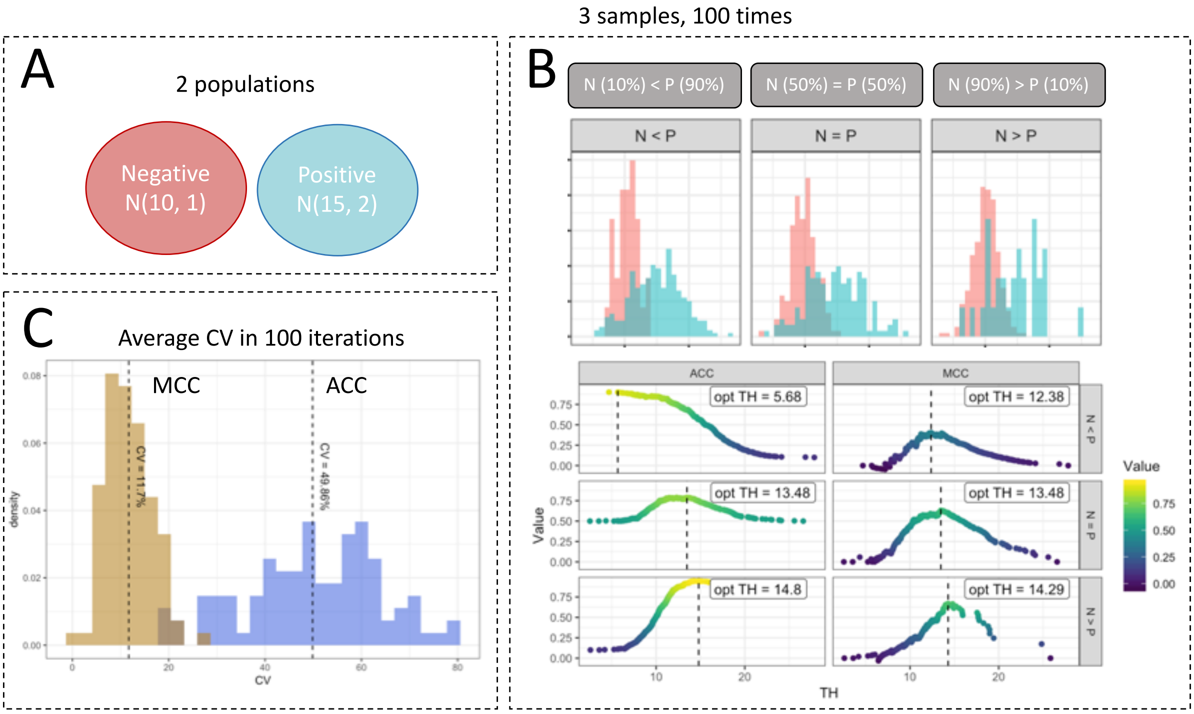


__Figure 4 - Optimal threshold selection.__ The optimal threshold usually is the one that optimizes sensitivity and specificity, but depending on the diagnostic test, different relative costs are given to misclassifying diseased and non-diseased individuals. Here we performed an in-silico experiment where we samples two different populations (A) using three different approaches: one where the number of positives outnumbers the number of negatives (P>N), one where the number of negatives outnumbers the number of positives (N>P) and one where the number of positives is the same as the number of negatives(P=N) (B). For each approach, we calculated the optimal threshold using the maximum of ACC or MCC and found that MCC returned more consistent results across 100 iterations (C).

Biomarker Panels

ROC curves are a great tool to assess qualitative performance of individual biomarkers or compare between different biomarkers. However, in biomarker development, though single markers may serve in selected cases, there is a growing consensus that multiple markers used either individually or as part of integrated panels will be required for most applications. The most relevant question, then, is not the performance of a single marker in isolation, but whether if another marker adds new information to the classification model [7]. The basic strategy to answer this question is to identify a set of metrics that quantify the discriminatory ability of the panels with and without the biomarker in question and evaluate if the gain in performance is significant. There are many ways to combine values from markers. To find an indicator that reliably represents the performance of the panel we need to construct a model. When a mechanistic model is not available, as is most often the case, a mathematical model can be used. [7]. Examples of mathematical models often used in classification problems include: support vector machines, logistic regression, random forest, etc. Remarkably, the more markers we include in the model, the mode complex the model becomes. Following the problem-solving principle known as the Occam's razor, usually we want "the simplest model that explains the data". A common strategy to find the optimal panel size is to plot the performance of the panel in the Y axis and the size in the X axis constructing what is known as the Pareto front. A valid rule of thumb for the optimal size is to choose the point in the curve with maximum convergence/concavity (elbow criterion). It is the responsibility of the statistician to choose the appropriate performance parameter(s) that best adapt to the specific problem/test.

The need of cross-validation

The evaluation of a biomarker using a mathematical model that has been optimized to give maximum performance for a given dataset is liable to be too optimistic compared to the performance it would show in a different study [7]. Optimizing a mathematical model for an entire dataset may cause not only to fit the information contained in the analyte (signal), but also the experimental error (noise). This is called bias due to overfitting. Re-sampling methods like bootstrap and cross-validation can be helpful diminishing this effect and bringing our model closer to the reality it aims to represent.

Analytical evaluation of clinical assays

Once the medical utility of an analyte as a biomarker has been verified, the analytical performance of the clinical assay has to be examined and optimized. A common practice for establishment of an analytical performance goal takes into consideration both the index of variation (within-subject variation, CV_i), and biological variation (between-subject variation, CV_g), and stipulates that imprecision should be smaller than $0.5 \cdot CV_i$ and bias smaller than $0.25 \cdot ({CV_i}^2 + {CV_g}^2)^{1/2}$ [10].

Indicators of precision: According to the International Standards Organization (ISO), precision is defined as the closeness of agreement among results of measurements performed under stipulated conditions [59]. Repeatability and reproducibility are indicators that express precision quantitatively. Repeatability refers to measurements that are performed under the same conditions, whereas reproducibility is used to describe measurements performed under different conditions. Even if these are known as indicators of precision, in reality they measure imprecision, which reflects random error and is expressed as a standard deviation (SD) or a coefficient of variation (CV%) and is evaluated using pools with different analyte concentrations (low, normal, high) [10]. In general, assays show smaller imprecision at higher concentrations of the analyte because of assay designs that generally result in direct proportionality between analyte concentration and generated signal. Therefore, the range of analyte concentrations tested should not only cover the span of the standard curve but more importantly the clinically relevant range.

Indicators of accuracy: Accuracy is defined as the closeness of agreement between the value of an analyte measurement in a sample and the true concentration of the analyte in that sample. Trueness and accuracy are similar, but not identical, as trueness is the closeness of agreement between the average value of different measurements of the same sample and the true concentration of the analyte in that sample [60]. Trueness reflects bias, a measure of systematic error, whereas accuracy reflects uncertainty, which comprises both random and systematic errors. Trueness can be evaluated by comparing the values obtained by the new test with those determined by a reference method, using unbiased regression analysis (Deming) and residual plots (Bland-Altman) or by measuring reference standards (samples of known concentration value) [10]. For novel analytes, reference standards do not usually exist, and alternative methods like recovery can be applied (Figure 5). In recovery experiments, different concentrations of the analyte of interest are added to a set of samples and the recovery is determined by comparing the measured concentration in the samples with and without added analyte. Ideally, a recovery of 100% should be seen which means that the difference in concentration between the samples with and without added analyte should be 100% of what was added. As the specificity of the antibodies used was established earlier by western blot analysis and immunostaining, no significant cross-reactivity with other proteins is expected. Even so, it can happen that the added recombinant/purified protein does not reassemble the naturally occurring one and that the recovery is far from 100%. Therefore, it is imperative to demonstrate that the newly developed assay is not affected by endogenous or exogenous interferences [61].

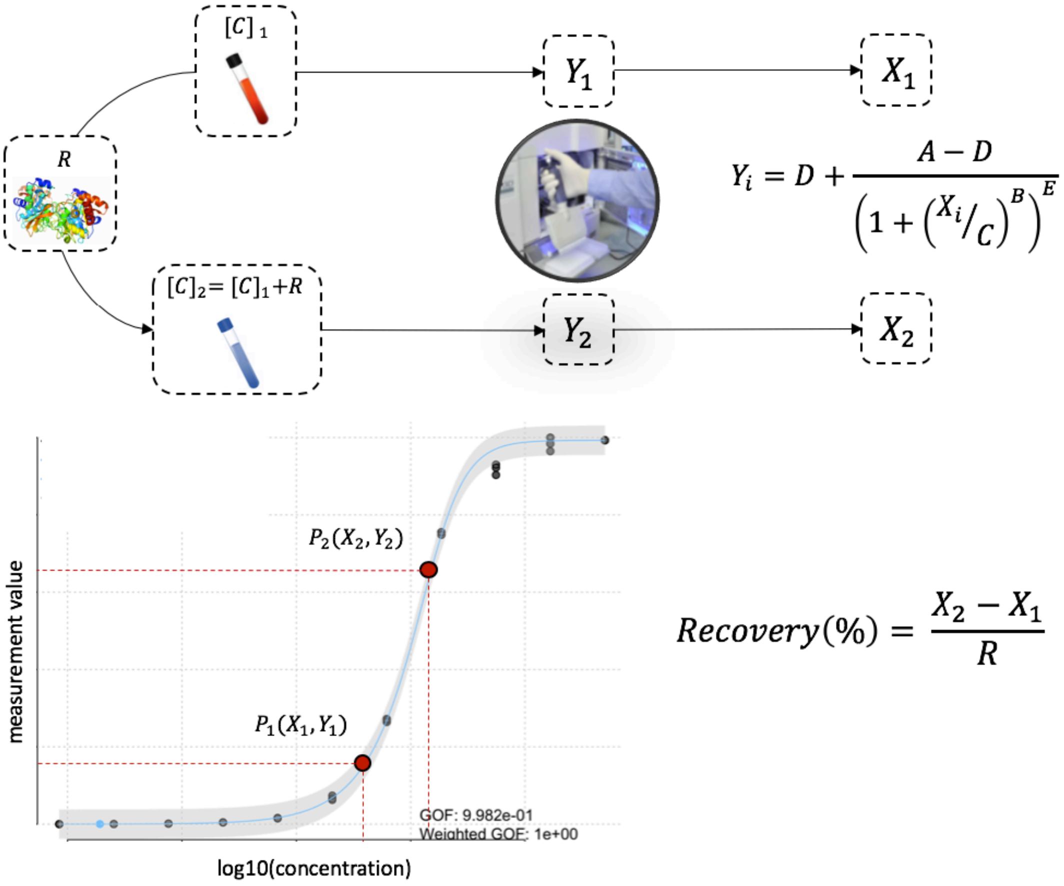


Figure 5 - Recovery experiment. *A set of samples of unknown concentrations 〖[C]〗_1 and 〖[C]〗_2 are measured with and without adding a known concentration(s) of a particular protein R. The measured values Y_1 and Y_2 are reversed to their corresponding concentrations X_1 and X_2 using the standard curve (usually modelled by a 5-parameter logistic regression).*

Recovery is then defined as the ratio between the difference in the recovered concentrations X_1 and X_2 and the originally added known concentration(s) R.

Indicators of analytical measurement range: The standard curve is the mathematical model that allows to estimate the concentration of analyte in a sample from its reported measurement value, but there is more: the standard curve also contains information about the measurement and linear ranges of a particular assay. The measurable range of an assay is determined by the upper and lower limits of detection. The lower limit of detection (LLOD) is defined as the lowest analyte concentration likely to be reliably distinguished from the measured values obtained from blank samples (plasma or serum, in this case, that does not contain the protein of interest) [62]. Linearity refers to the range of analyte concentration that exhibits a constant relationship between observed and expected measurements [63]. Linearity and limits of detection must be assessed to determine the range of measurand values over which measurements can be performed with acceptable precision and accuracy [64].

References

[1] Lundblad, R.L., 2016. Development and application of biomarkers. CRC Press.

[2] DeCaprio, A.P., Introduction to toxicological biomarkers, in Toxicological Biomarkers, A.P. DeCaprio (ed.), Taylor & Francis, New York, 2006.

[3] Biomarkers Definitions Working Group, Atkinson Jr, A.J., Colburn, W.A., DeGruttola, V.G., DeMets, D.L., Downing, G.J., Hoth, D.F., Oates, J.A., Peck, C.C., Schooley, R.T. and Spilker, B.A., 2001. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical pharmacology & therapeutics, 69(3), pp.89-95.

[4] Ptolemy, A.S. and Rifai, N., 2010. What is a biomarker? Research investments and lack of clinical integration necessitate a review of biomarker terminology and validation schema. Scandinavian Journal of Clinical and Laboratory Investigation, 70(sup242), pp.6-14.

[5] Paone, J.F., Phillip Waalkes, T., Robinson Baker, R. and Shaper, J.H., 1980. Serum UDP-galactosyl transferase as a potential biomarker for breast carcinoma. Journal of surgical oncology, 15(1), pp.59-66.

[6] Webb, K.S. and Lin, G.H., 1980. Urinary fibronectin: potential as a biomarker in prostatic cancer. Investigative urology, 17(5), pp.401-404.

[7] Vaidya, V.S. and Bonventre, J.V. eds., 2010. Biomarkers: In medicine, drug discovery, and environmental health. John Wiley & Sons.

[8] American Diabetes Association, 2010. Diagnosis and classification of diabetes mellitus. Diabetes care, 33(Supplement 1), pp.S62-S69.

[9] Staud, R., 2008. Heart rate variability as a biomarker of fibromyalgia syndrome. Future rheumatology, 3(5), p.475.

[10] Rifai, N., Gillette, M.A. and Carr, S.A., 2006. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. Nature biotechnology, 24(8), p.971.

[11] Decreased, G.F.R., 2013. Definition and classification of CKD. Kidney International, 3, pp.19-62.

[12] Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H. and Kannel, W.B., 1998. Prediction of coronary heart disease using risk factor categories. Circulation, 97(18), pp.1837-1847.

[13] Biomarkers Definitions Working Group, Atkinson Jr, A.J., Colburn, W.A., DeGruttola, V.G., DeMets, D.L., Downing, G.J., Hoth, D.F., Oates, J.A., Peck, C.C., Schooley, R.T. and Spilker, B.A., 2001. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical pharmacology & therapeutics, 69(3), pp.89-95.

[14] Patel, S.P. and Kurzrock, R., 2015. PD-L1 expression as a predictive biomarker in cancer immunotherapy. Molecular cancer therapeutics, 14(4), pp.847-856.

[15] Young, D.S. and Bermes, E.W.J., Specimen collection and process: Sources of biological variation, in Tietz Textbook of Clinical Chemistry, 3rd edn., C.A. Burris and E.R. Ashwood (eds.), W.B. Saunders, Philadelphia, PA, Chapter 2, pp. 42–72, 1999.

[16] Zeng, L., Liu, J., Wang, Y., Wang, L., Weng, S., Tang, Y., Zheng, C., Cheng, Q., Chen, S. and Yang, G.Y., 2011. MicroRNA-210 as a novel blood biomarker in acute cerebral ischemia. Front Biosci (Elite Ed), 3(3), pp.1265-1272.

[17] Price, C.P., Newall, R.G., and Boyd, J.C., Use of protein: Creatinine ratio measurements on random urine samples for prediction of significant proteinuria: A systematic review, Clin. Chem. 51, 1577–1586, 2005.

[18] Lamb, E., Newman, D.J., and Price, C.P., Kidney function tests, in Tietz Textbook of Clinical Chemistry and Molecular Diagnostics, 4th edn., C.A. Burtis, E.R. Ashwood, and D.E. Burns (eds.), Saunders, St. Louis, MO, Chapter 24, pp. 797–835, 2006.

[19] Gansevoort, R.T., Brinkman, J., Bakker, S.J. et.al., Evaluation of measures of urinary albumin excretion, Am. J. Epidemiol. 164, 725–727, 2006.

[20] Pickering, A., Carter, J., Hanning, I., and Townend, W., Emergency department measurement of urinary S100B in children following head injury: Can extracranial injury confound findings?, Emerg. Med. J. 25, 88–89, 2008.

[21] Hayward, R.M., Kirk, M.J., Sproull, M. et.al., Post-collection, pre-measurement variables affecting VEGF levels in urine biospecimens, J. Cell. Mol. Med. 12, 343–350, 2008.

[22] Hellhammer, D.H., Wüst, S., and Kudielka, B.M., Salivary cortisol as a biomarker in stress research, Psychoneuroendocrinology 34, 163–171, 2009.

[23] Pearson, J.R., Gill, C.I., and Rowland, I.R., Diet, fecal water, and colon cancer – Development of a biomarker, Nutr. Rev. 67, 509–526, 2009.

[24] Hunt, J., Exhaled breath condensate: An overview, Immunol. Allergy Clin. North Am. 27, 587–596, 2007.

[25] Webster, H.L., Laboratory diagnosis of cystic fibrosis, Crit. Rev. Clin. Lab. Sci. 18, 313–338, 1983.

[26] Jung, S.S., Park, H.S., Lee, I.J. et.al., The HCCR oncoprotein as a biomarker for human breast cancer, Clin. Cancer Res. 11, 7700–7708, 2005.

[27] Kwon, M.S., Shin, S.H., Yin, S.H. et.al., CD63 as a biomarker for predicting the clinical outcomes in adenocarcinoma of the lung, Lung Cancer 57, 46–53, 2007.

[28] Levy, B.I., 2009. Blood pressure as a potential biomarker of the efficacy angiogenesis inhibitor.

[29] Pupi, A., Mosconi, L., Nobili, F.M. and Sorbi, S., 2005. Toward the validation of functional neuroimaging as a potential biomarker for Alzheimer’s disease: implications for drug development. Molecular Imaging and Biology, 7(1), pp.59-68.

[30] Hamstra, D.A., Rehemtulla, A. and Ross, B.D., 2007. Diffusion magnetic resonance imaging: a biomarker for treatment response in oncology. Journal of clinical oncology, 25(26), pp.4104-4109.

[31] Gormally, E., Caboux, E., Vineis, P. and Hainaut, P., 2007. Circulating free DNA in plasma or serum as biomarker of carcinogenesis: practical aspects and biological significance. Mutation Research/Reviews in Mutation Research, 635(2-3), pp.105-117.

[32] James, S.J., Cutler, P., Melnyk, S., Jernigan, S., Janak, L., Gaylor, D.W. and Neubrander, J.A., 2004. Metabolic biomarkers of increased oxidative stress and impaired methylation capacity in children with autism. The American journal of clinical nutrition, 80(6), pp.1611-1617.

[33] Hanna, M.H., Dalla Gassa, A., Mayer, G., Zaza, G., Brophy, P.D., Gesualdo, L. and Pesce, F., 2017. The nephrologist of tomorrow: towards a kidney-omic future. Pediatric Nephrology, 32(3), pp.393-404.
[34] Anderson, N.L., 2010. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. Clinical chemistry, 56(2), pp.177-185.
[35] Fassett, R.G., Venuthurupalli, S.K., Gobe, G.C., Coombes, J.S., Cooper, M.A. and Hoy, W.E., 2011. Biomarkers in chronic kidney disease: a review. Kidney international, 80(8), pp.806-821.
[36] Pourret, O., Naïm, P. and Marcot, B. eds., 2008. Bayesian networks: a practical guide to applications (Vol. 73). John Wiley & Sons.
[37] Prikryl, P., Vojtova, L., Maixnerova, D., Vokurka, M., Neprasova, M., Zima, T. and Tesar, V., 2017. Proteomic approach for identification of IgA nephropathy-related biomarkers in urine. Physiological research, 66(4), pp.621-632.
[38] Anderson, N.L., 2010. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. Clinical chemistry, 56(2), pp.177-185.
[39] Keller, B.J., Martini, S., Sedor, J.R. and Kretzler, M., 2012. A systems view of genetics in chronic kidney disease. Kidney international, 81(1), pp.14-21.
[40] Khatri, P., Sirota, M. and Butte, A.J., 2012. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS computational biology, 8(2), p.e1002375.
[41] De Leeuw, C.A., Neale, B.M., Heskes, T. and Posthuma, D., 2016. The statistical properties of gene-set analysis. Nature Reviews Genetics, 17(6), p.353.
[42] Nam, D., 2010. De-correlating expression in gene-set analysis. Bioinformatics, 26(18), pp.i511-i516.
[43] Wu, D. and Smyth, G.K., 2012. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic acids research, 40(17), pp.e133-e133.
[44] Dong, X., Hao, Y., Wang, X. and Tian, W., 2016. LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. Scientific reports, 6, p.18871.
[45] Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichita, C. and Draghici, S., 2013. Methods and approaches in the topology-based analysis of biological pathways. Frontiers in physiology, 4, p.278.
[46] Sass, S., Buettner, F., Mueller, N.S. and Theis, F.J., 2013. A modular framework for gene set analysis integrating multilevel omics data. Nucleic acids research, 41(21), pp.9622-9633.
[47] Alhamdoosh, M., Ng, M., Wilson, N.J., Sheridan, J.M., Huynh, H., Wilson, M.J. and Ritchie, M.E., 2017. Combining multiple tools outperforms individual methods in gene set enrichment analyses. Bioinformatics, 33(3), pp.414-424.
[48] De Leeuw, C.A., Neale, B.M., Heskes, T. and Posthuma, D., 2016. The statistical properties of gene-set analysis. Nature Reviews Genetics, 17(6), p.353.
[49] Fridley, B.L., Jenkins, G.D. and Biernacka, J.M., 2010. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. PLoS One, 5(9), p.e12693.
[50] Cho, A., Shim, J.E., Kim, E., Supek, F., Lehner, B. and Lee, I., 2016. MUFFINN: cancer gene discovery via network analysis of somatic mutation data. Genome biology, 17(1), p.129.
[51] Horn, H., Lawrence, M.S., Chouinard, C.R., Shrestha, Y., Hu, J.X., Worstell, E., Shea, E., Ilic, N., Kim, E., Kamburov, A. and Kashani, A., 2018. NetSig: network-based discovery from cancer genomes. Nature methods, 15(1), p.61.
[52] Leiserson, M.D., Vandin, F., Wu, H.T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M. and Lawrence, M.S., 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature genetics, 47(2), p.106.

[53] Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D. and Stuart, J.M., 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics, 26(12), pp.i237-i245.

[54] Shrestha, R., Hodzic, E., Sauerwald, T., Dao, P., Wang, K., Yeung, J., Anderson, S., Vandin, F., Haffari, G., Collins, C.C. and Sahinalp, S.C., 2017. HIT'nDRIVE: patient-specific multidriver gene prioritization for precision oncology. Genome research, 27(9), pp.1573-1588.

[55] Rizvi, M.H., Rustagi, J.S. and Siegmund, D. eds., 2014. Recent advances in statistics: papers in honor of Herman Chernoff on his sixtieth birthday. Academic Press.

[56] Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L.M., Moher, D., Rennie, D., De Vet, H.C. and Lijmer, J.G., 2003. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Clinical chemistry, 49(1), pp.7-18.

[57] Zweig, M.H. and Campbell, G., 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clinical chemistry, 39(4), pp.561-577.

[58] Saito, T. and Rehmsmeier, M., 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, 10(3), p.e0118432.

[59] International Standards Organization. Statistics-vocabulary and Symbols – Part 1: Probability and General Statistical Terms.

[60] International Standards Organization. Accuracy (Trueness and Precision) of Measurement Methods and Results.

[61] Dimeski, G., 2008. Interference testing. The Clinical Biochemist Reviews, 29(Suppl 1), p.S43.

[62] Armbruster, D.A. and Pry, T., 2008. Limit of blank, limit of detection and limit of quantitation. The Clinical Biochemist Reviews, 29(Suppl 1), p.S49.

[63] Burtis, C.A., Ashwood, E.R. and Bruns, D.E., 2006. Textbook of clinical chemistry and molecular diagnostics. SA Bertis, ER Ashvud, D. Bruns.

[64] Jhang, J.S., Chang, C.C., Fink, D.J. and Kroll, M.H., 2004. Evaluation of linearity in the clinical laboratory. Archives of pathology & laboratory medicine, 128(1), pp.44-48.

Chapter 2 Chronic Kidney Disease

This chapter introduces Chronic Kidney Disease (CKD). It focuses on the development of renal damage, its diagnosis, and categorization.

Introduction

The kidneys are two bean-shaped organs that regulate the volume of extracellular fluid by removing excess of water and wastes from the blood, through the nephrons (the microscopic structural and functional units of the kidney), to form urine. These wastes come from the normal breakdown of active muscle and from food consumption. Additionally, kidneys help on the regulation of osmolarity, ion concentrations, and pH, and are responsible of the production of certain hormones (e.g., erythropoietin). Chronic kidney disease (CKD) is a syndrome that describes the gradual loss of kidney function and/or alterations on its structure. This loss is often recognized by reduced glomerular filtration rate (GFR) which is clinically identified by increased levels of serum creatinine (sCr). Creatinine is a waste product of musculoskeletal exercise that in healthy conditions is secreted through the urine but in presence of kidney damage builds-up increasing its serum concentration.

CKD is a major public health problem that affects to approximately 14% of the general population [1], although this percentage varies in different regions of the world and age groups, being significantly higher in people over 65 years of age. Despite its prevalence, CKD is a heavily understudied syndrome as NIH investments in kidney research represent less than 1% of Medicare costs for kidney diseases [2]. CKD as a cause of death is increasing and now represents 1.35% of the global burden of disability-adjusted life years lost, being ranked 17th in 2015 [3]. In high-income countries, CKD is most commonly associated with diabetes mellitus, hypertension and obesity; for example, 30-40% of those with diabetes also have CKD [4]. In low-income countries, however CKD is often accompanying inappropriate use of medications and infectious diseases [5].

Development of renal damage

The number of nephrons a person has is set at birth and averages 950,000 nephrons per kidney. After birth, no new nephrons can be generated, and any loss would require the remaining nephrons to work harder to compensate and maintain kidney function. CKD is attributed to the rupture of basement membranes including abnormal epithelial regeneration, which endorses nephron loss. Reductions in the total renal mass cause glomerular hypertrophy in the functional nephrons to accommodate renal demands [6]. However, this hypertrophy causes additional nephron losses triggering a vicious cycle that results in end stage renal disease (ESRD) if the process is not decelerated [5]. Moreover, the effects of shear stress on the remaining nephrons eventually result in the development of sclerotic lesions in the glomeruli and nephron atrophy. These lessons cause the filtration of molecules that, in healthy conditions, remain in the bloodstream. Therefore, advanced CKD not only causes insufficient filtration rate but also is associated with systemic complications such as electrolyte abnormalities, metabolic acidosis, anemia, mineral bone disorder, hypertension and hyperuricemia [5].

During renal damage development, it is unknown if all nephrons respond to an injurious stimulus or only a particular subgroup is affected. As the function of the different nephrons in the kidney can be considered independent, GFR can be expressed by the equation:

$$GFR_{total} = GFR_{single-nephron} \cdot N_{nephrons}$$

GFR_{total}: GFR of the whole renal mass (all nephrons).
$GFR_{single-nephron}$: the average GFR of individual nephrons.
$N_{nephrons}$: the number of operatory nephrons.

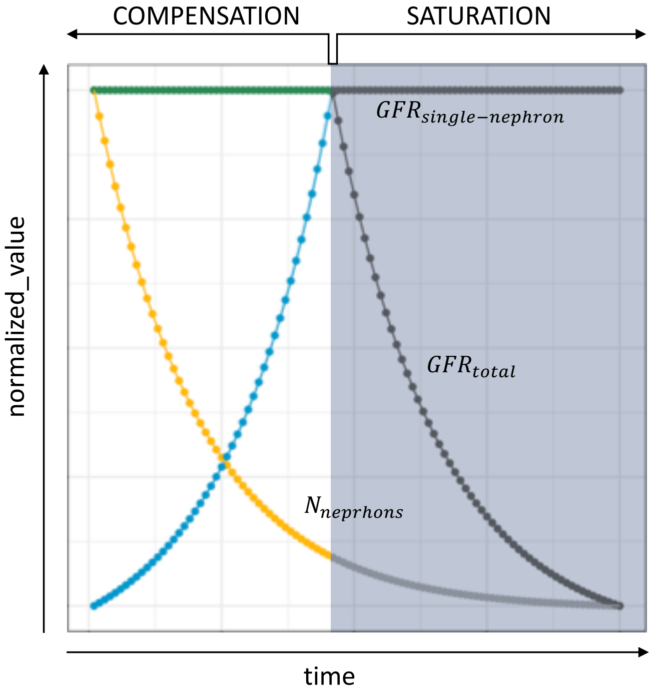


Figure 6 - Renal damage development model. *As the number of functional nephrons begins to reduce, the remaining working nephrons need to work harder to keep up with renal demand (compensation). However, the work a single nephron can do is not unlimited and only when functional nephrons have reached their limit, we can observe a drop in total GFR (saturation).*

This equation implies that when $N_{nephrons}$ decays, GFR_{total} will be constant as long as $GFR_{single-nephron}$ increases. However, $GFR_{single-nephron}$ cannot increase indefinitely. Therefore, only when $GFR_{single-nephron}$ has reached its maximum GFR_{total} will start decreasing. By contrast, a decline in GFR_{total} implies that $N_{nephrons}$ has significantly declined, possibly with the remaining functional nephrons operating at their maximum [5]. In fact, sCr does not increase until approximately 50% of the renal mass is lost, suggesting a compensatory but saturable system [7] (Figure 6). Also, it is unknown whether if all cells in the nephrons are affected by a given injury or if only specific cells are disturbed. Moreover, it is unknown if these responses are stimulus-specific or common to all forms of injury.

This mechanism can explain why CKD is an asymptomatic disease in its early stages and can go undetected until it is very advanced, when patients may experience weakness related to anaemia and proteinuria (presence of large quantities of protein in urine which often causes the urine to become foamy) [5].

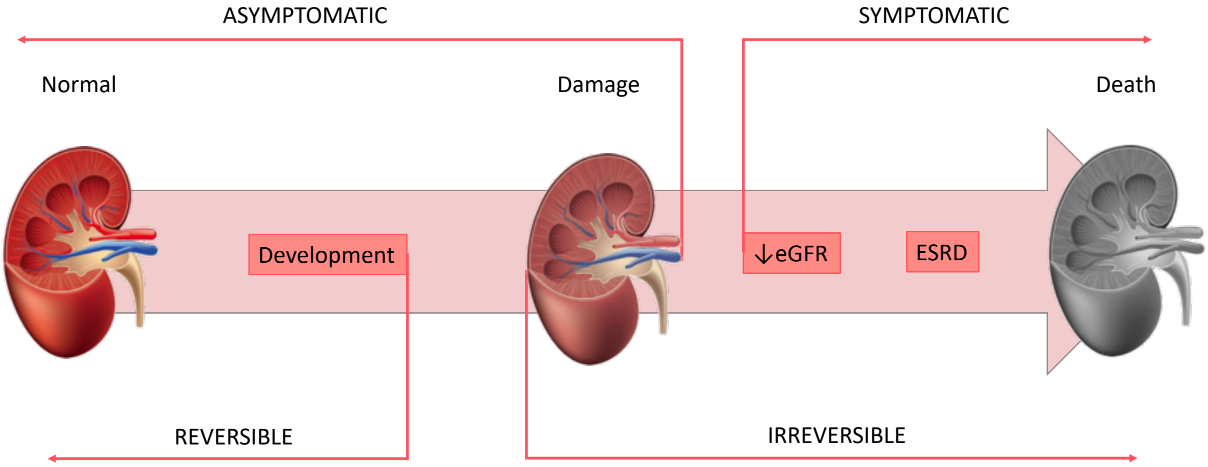


Figure 7 - CKD development model. *During the first stages, the disease remains asymptomatic given the compensatory nature of the kidney. However, when this system saturates, the disease becomes asymptomatic represented by a drop in eGFR and a rise of sCr.*

CKD diagnosis

Classic nephrology taught that kidney failure should be described based on the etiology and the anatomy of the injury. It was believed that identifying the site of injury and the mechanism of dysfunction should inform the correct treatment [7]. Nowadays, however, modern nephrology uses a single analyte, sCr, regardless of the type of injury or the related mechanisms. It is believed that an increase in this analyte, reflects some degree of injury in the kidney tubule. In fact, published guidelines for kidney disease: Risk-Injury-Failure-Loss-End Stage Kidney Disease (RIFLE) [8], Acute Kidney Injury Network (AKIN) [9], or Kidney Disease, Improving Global Outcomes (KDIGO) [10] mention that even minimal increases in sCr represent the initial phase of a wide spectrum of diseases related to kidney damage [7]. These diseases are comprised in the diagnostic term acute kidney injury (AKI), which is defined by increasing levels of sCr, regardless of the pathophysiology of origin or the anatomical site of injury. Most of the times, AKI episodes are self-limited, and kidney function returns to normal. However, AKI episodes, often show, more chronic renal damage than would be expected and subnormal renal function occurs late in the follow-up period. In fact, when AKI symptoms persist for more than 3 months, the patient is diagnosed with CKD which has made AKI to gain increasing recognition as a prelude to the development of CKD [7]. Moreover, this relationship is bidirectional, with CKD patients at substantially greater risk of suffering an episode of AKI [11]. Therefore, AKI and CKD are often understood as an integrated clinical syndrome [12], and nephrologists argue that the distinction between AKI and CKD may be artificial. The definition of kidney malfunction by the AKI paradigm has permitted the analysis, integration and comparison of population-based data sets which has grown awareness of the serious public health problem that AKI represents: the incidence continues to increase among hospitalized adults [13, 14], and the diagnosis is associated with significant morbidity and mortality [15] as there is evidence of AKI being responsible for approximately 2 million deaths annually worldwide [16].

Undoubtedly, the introduction of sCr in clinical practice has significantly improved the state of the art in CKD diagnosis. However, sCr does not represent a perfect reference standard as it may rise for reasons unrelated to kidney damage (e.g., fully reversible volume depletion [7]) implicating low specificity and may also fail to rise when kidney damage has been done involving low sensitivity. Moreover, as previously mentioned, sCr concentration does not increase until approximately 50% of the renal mass is lost, meaning that when the damage is not severe enough, sCr may not be changed, implicating late diagnosis. Finally, diagnosis by sCr is used to identify a generic form of CKD, despite CKD being a complex and heterogenous disease, denoting a biomarker that cannot fractionate the disease in subcategories. All these drawbacks indicate that renal biopsy still plays the crucial role in the definitive diagnosis of CKD [17, 18]. However, due to its invasiveness, it is not exempt of complications and a lot of research efforts are being focused on the identification of molecular biomarkers for diagnosis of CKD using less-invasive and easy-to-collect samples, such as urine and blood [18]. Recent data have shown that gene expression in tissue and protein signatures in urine can distinguish different kidney conditions, implying a complex system responding to different environmental stimuli [19]. In fact, some proteins are progressing through the biomarker development pipeline to the point of starting to be considered “second generation” biomarkers for CKD. For example, neutrophil gelatinase-associated lipocalin (NGAL), and kidney injury molecule-1 (KIM-1) have shown improved specificity when compared with sCr but also fail on discretizing further the disease [19, 20]. These efforts may guide us towards a molecular-based classification of renal diseases, where proteomics play a central role identifying the early pathological modifications and the molecular changes related to disease progression. However, new insights into the different mechanisms affecting kidney biology may be needed to redefine CKD in molecular terms.

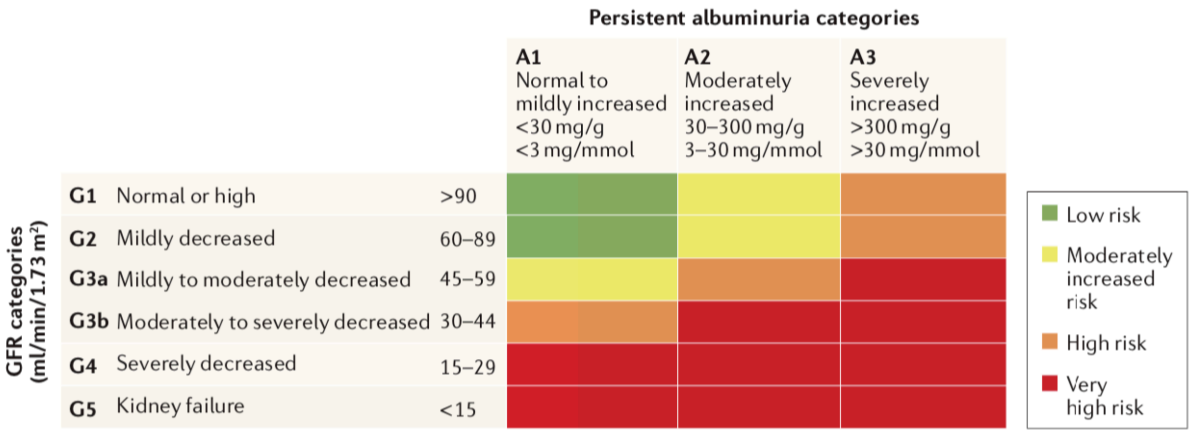
CKD categorization

It has been stated that in CKD, different mechanisms may affect different parts of cell physiology, but ultimately, all the variants, regardless of the initiating disease, converge in nephron loss which involves nonspecific wound-healing responses that include some degree of interstitial fibrosis. This common mechanism has allowed the stratification of the disease based on the degree of renal damage (severity). However, the clinical presentation of CKD also takes into account the pathogenesis of origin (primary disease) making it a bi-directional classification.

Degree of renal damage

All modern categorizations of CKD use a classification of severity based on the GFR, either estimated (eGFR) or measured (mGFR). Even if these two parameters can show differences, mGFR is calculated by urine clearance which is cumbersome and expensive making eGFR a more attractive alternative [21]. In adults, the most common ways to calculate eGFR are the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) [22] and the Modification of Diet in Renal Disease (MDRD) equations [23], both of which use sCr under steady-state conditions. Regardless of the equation used for the eGFR, CKD can be generally categorized in 5 stages with increased severity (Figure 3). Early stages of CKD (G1-G2) are usually asymptomatic, but from CKD G3 on, patients may experience symptoms like weakness related to anemia and polyuria (excessive urination) [5]. When a patient progresses to G5 and ESRD, kidneys cannot satisfy renal demands and haemodialysis or kidney transplantation are required. Patients receiving a kidney transplant have better prospects as the 5-year survival is around 86-93%, comparing to those receiving dialysis whose life expectancy is one third of that of the age and sex-matched general population [5].

KDIGO [10], in 2012 further detailed the classification of CKD based on severity, incorporating the level of albuminuria (which can be approximated by proteinuria as serum albumin accounts for 55% of the total protein in blood plasma [24]) expanding the G1-G5 staging into a 2D matrix. This matrix also incorporates prognostic information about the risk of CKD progression and related complications [5]. The use of these two parameters for CKD is complementary as eGFR is a well-stablished indicator of renal excretory function (tubular damage) while albuminuria is a marker of renal barrier dysfunction (glomerular damage) [5]. Albuminuria (given as a ratio to creatinine, in mg/g), is most commonly determined using the urinary albumin concentration test [5] and KDIGO classifies the continuous measurement into 3 categories with increasing severity (Figure 8).



			Persistent albuminuria categories		
			A1 Normal to mildly increased <30 mg/g <3 mg/mmol	A2 Moderately increased 30–300 mg/g 3–30 mg/mmol	A3 Severely increased >300 mg/g >30 mg/mmol
GFR categories (ml/min/1.73 m^2)	G1 Normal or high	>90	🟩	🟨	🟧
	G2 Mildly decreased	60–89	🟩	🟨	🟧
	G3a Mildly to moderately decreased	45–59	🟨	🟧	🟥
	G3b Moderately to severely decreased	30–44	🟧	🟥	🟥
	G4 Severely decreased	15–29	🟥	🟥	🟥
	G5 Kidney failure	<15	🟥	🟥	🟥

🟩 Low risk
🟨 Moderately increased risk
🟧 High risk
🟥 Very high risk

__Figure 8 - Kidney Disease classification based on the KDIGO paradigm (from KDIGO [Kidney Disease: Improving Global Outcomes (KDIGO) Acute Kidney Injury Work Group].__ The 2D matrix incorporates the level of albuminuria and the glomerular filtration rate (eGFR) to describe the risk of patients with chronic kidney disease progressing to end-stage renal disease and other adverse effects.

The two metrics used in KDIGO matrix (GFR and albuminuria) are used to classify a 'generic' form of CKD, but adding the underlying cause is highly desirable and is critical for optimal management of the disease [5].

Pathogenesis of CKD

Glomerulonephritis (GN) is one of the most common causes leading to CKD [18]. There are two main types of glomerulonephritis: primary and secondary. Primary glomerulonephritis affects the kidneys directly, while in secondary glomerulonephritis the kidneys are damaged as a result of another illness [25]. Of the patients who come to ESRD, 15-20% are due to primary glomerulonephritis, 55-60% to secondary glomerulonephritis (30% diabetic, 15-20% hypertensive/vascular, 10% immune), 10% to autosomal dominant polycystic kidney disease (ADPKD), 10% to interstitial nephropathy, and the remaining 5-10% remain unclear [26]. Although they share some common mechanisms, there are many primary diseases that lead to CKD, each with its own specific pathophysiology. Therefore, here we will focus on those used in our experiments, i.e., immunoglobulin A nephropathy (IgAN), diabetic nephropathy (DN), and lupus nephritis (LN).

Primary GN

Immunoglobulin A nephropathy: IgA nephropathy (IgAN) involves glomerular damage and is the most common GN worldwide [27]. Its pathogenesis involves an abnormal immunitary response to antigen encountered in the upper respiratory/gastrointestinal tract. The exposition of antigens at mucosal sites stimulates the production of IgA which entails the generation of anti IgA antibodies, immunocomplex formation, and renal deposition [26]. Recent studies [28], have shown that the galactose-lacking pIgA1 serum levels are often increased in IgAN patients. However, the low sensitivity/specificity of this marker doesn't allow the substitution of the current diagnostic practice (demonstration of mesangial IgA-dominant staining by immunofluorescence or immunohistochemistry), which requires renal biopsy [29]. Nevertheless, researchers are employing many non-invasive proteomic techniques for detecting alternative diagnostic biomarkers both in urine and serum samples [17, 30].

Secondary GN

Diabetic nephropathy: DN is the most common secondary glomerulonephritis accounting for 30% of the patients reaching ESRD in CKD, representing the first cause of chronic kidney failure. In DN all the kidney compartments are affected including glomeruli, tubuli, interstitium and vasculature. Its pathogenesis involves vascular damage generated by protein glycosylation, sorbitol production, increased glomerular filtration, and finally glomerular hypertension and hypertrophy [26]. In the tubuli, the pathology includes interstitial fibrosis, inflammatory infiltration and atrophy. DN is clinically presented in stages based on the urinary albumin excretion (UAE): microalbuminuria (20µg/min $\leq$ UAE $\leq$ 199µg/min) and macroalbuminuria (UAE > 200µg/min) [31]. Although diabetic nephropathy has been classically defined by the presence of proteinuria in diabetic patients, proteinuria explicitly is not specific of DN making the identification of novel molecular biomarkers for DN an attractive option.

Lupus nephritis: LN is a secondary glomerulonephritis affecting both the glomeruli and the tubuli. LN represents a frequent and severe manifestation of systemic lupus erythematosus (SLE, a multisystem connective tissue disease), appearing in about 75% of patients with this disease [32, 33]. The pathogenesis involves immunocomplex deposition from autoantibodies (anti-phospholipids, anti-DNA, anti-SM, anti-nucleus) and its pathology differentiates 6 classes based on histology [26]. This level of complexity represents a clinical challenge in choosing the best therapeutic strategy [34], which can be only achieved by renal biopsy [33, 35, 36]. To avoid the invasiveness of renal biopsy, many efforts are being made to identify reliable biomarkers in non-invasive samples like urine/serum that allow monitoring of the disease thus facilitating an early diagnosis and intervention that can improve favorable outcomes [37, 38].

Biomarkers in CKD

We have seen how when nephrons shut down, the active ones work harder to compensate which carries out further complications. Indicators in clinical practice like GFR, proteinuria, or albumin-creatinine ratio (ACR) are useful for later stages of the disease but fail on its early detection given the compensatory abilities of the kidney [39]. The lack of reliable molecular biomarkers makes renal biopsy to remain the gold standard to identify the primary disease leading to CKD. However, the unprecedented availability of molecular data is attracting the focus of several research efforts to identify novel biomarkers. It is important to determine whether the newly identified biomarkers are purely associations or real biomarkers of underlying pathophysiological processes (mechanistic biomarkers) [19]. These biomarkers can be classified according to the main mechanism they reflect including: GFR, proteinuria, glomerular damage, tubular damage, among others. Here we will review the main biomarkers for each mechanism (for an extensive review, read [19]).

Biomarkers for kidney function (eGFR)

Creatinine (Cr): creatinine is a by-product of musculoskeletal exercise that in healthy conditions is secreted through the kidneys to the urine. Conversely, in CKD, this filtration is affected which causes creatinine to build up in blood increasing its serum levels. Having a blood test for creatinine is the first step in checking kidney function. However, serum creatinine levels are also affected by factors independent of kidney disease like muscle mass and diet.

Cystatin C: Cystatin C is a small protein that is produced by the cells in the body. In healthy conditions, Cystatin C is filtered and metabolized after tubular absorption, but in CKD is accumulated in blood. Cystatin C is alternatively used to calculate eGFR in cases of overweighed patients or with high muscular mass as cystatin C is not influenced by these factors but is affected by inflammation and diabetes among others [5].

Biomarkers for CKD progression (proteinuria)

Albumin (Alb): albumin is a protein produced by the liver which our body uses mainly for tissue regeneration. In healthy conditions, albumin stays in the bloodstream. However, in CKD, albumin is secreted to urine. Indeed, increased urinary albumin excretion is an established marker of CKD progression [40] reflecting both glomerular and tubular damage.

Biomarkers for glomerular damage

Podocin: Podocin is a protein component of the podocytes, cells in the Bowman's capsule in the kidneys that wrap around the capillaries of the glomerulus [41]. It functions as the filtration barrier of the kidney glomerulus. Mutations in the podocin gene NPHS2 can cause glomerulonephritis [42]. Urinary podocin is elevated in diabetic nephropathy and active lupus nephritis [19].

Biomarkers for tubular damage

Kidney Injury Molecule 1 (KIM-1): KIM-1 is a transmembrane tubular protein with unknown function, undetectable in healthy kidneys, but highly expressed in experimental and clinical kidney injury [43]. KIM-1 expression is significantly increased in tissue and urine in AKI and kidney disease. The correlation between tissue and urine KIM-1 is unknown. Studies suggest that KIM-1 may be an indicator of AKI to CKD transition [44].

Neutrophil gelatinase-associated lipocalin (NGAL): NGAL is a protein highly expressed in the tubular epithelium of the distal nephron which is secreted from tubular epithelial cells after tubular injury [19]. It has two molecular forms in the urine: the dimeric form that comes from neutrophils, and the monomeric form which is generated by the kidney tubular epithelial cells [45]. Differentiating these two forms may enhance the performance of NGAL as a biomarker even further.

As seen, eGFR (based on sCr), especially when combined with proteinuria, is the most widely spread biomarker for CKD and one of the few accepted for clinical practice, but these have limitations [19]. Second generation biomarkers like podocin, NGAL, and KIM-1 have refined the diagnosis of CKD by capturing different pathologic mechanisms but fail to sub fractionate CKD further [7]. Although elements of CKD are common to all primary diseases, particularly in the late stages, there are specific pathophysiological mechanisms unique to each underlying renal etiology [19]. For this reason, early, more sensitive, biomarkers are required. However, it is unlikely that a single biomarker will be able to capture all the complexity of this heterogeneous disease. Hence, a panel measuring multiple biomarkers including disease-specific biomarkers may be required for the molecular characterization of CKD [19].

Conclusions

Undoubtedly, the standardization of CKD and the staging system based on measurements of eGFR and proteinuria that can be reliably applied to a large population at a reasonable cost, represents a major success of modern nephrology and an attractive focus for research efforts [46]. However, this classification system represents a reductionist approach that does not capture the heterogeneity and complexity of CKD. Although recent efforts have discovered a new generation of biomarkers that have laid the groundwork for capturing the different mechanisms in CKD, the heterogeneity of these studies and the large number of biomarkers identified make it difficult to translate these findings into clinical practice [7]. Moreover, predictions based on single-genes are usually ineffective for complex diseases and fail in their progress to clinical trials due to their lack of connection to the mechanism of the disease. On the other hand, most of commercially available products use panels of genes like the MammaPrint® test [47], a panel of 70 genes for breast-cancer prediction. In the case of CKD, one of the first applications of integrative panels on the urine of CKD patients allowed the identification of a 273-peptide classifier capable of discriminating between CKD patients and healthy controls with a sensitivity of 98.7% and a specificity of 100% [48]. Although its satisfactory performance as a diagnostic tool, the large number of analytes included makes the test expensive and its capability to discriminate among different primary diseases has not been tested. Therefore, a test including disease-specific biomarkers does not yet exist and may be required [19].

References

[1] https://www.niddk.nih.gov/health-information/health-statistics/kidney-disease
[2] https://report.nih.gov/categorical_spending.aspx] [https://www.usrds.org/adr.aspx
[3] Jager, K.J. and Fraser, S.D., 2017. The ascending rank of chronic kidney disease in the global burden of disease study. Nephrology Dialysis Transplantation, 32(suppl_2), pp.ii121-ii128.
[4] Maric-Bilkan, C., 2013. Obesity and diabetic kidney disease. Medical Clinics, 97(1), pp.59-74.
[5] Romagnani, P., Remuzzi, G., Glassock, R., Levin, A., Jager, K.J., Tonelli, M., Massy, Z., Wanner, C. and Anders, H.J., 2017. Chronic kidney disease. Nature Reviews Disease Primers, 3, p.17088.

[6] Chawla, L.S. and Kimmel, P.L., 2012. Acute kidney injury and chronic kidney disease: an integrated clinical syndrome. Kidney international, 82(5), pp.516-524.

[7] Kiryluk, K., Bomback, A.S., Cheng, Y.L., Xu, K., Camara, P.G., Rabadan, R., Sims, P.A. and Barasch, J., 2018, January. Precision medicine for acute kidney injury (AKI): redefining AKI by agnostic kidney tissue interrogation and genetics. In Seminars in nephrology (Vol. 38, No. 1, pp. 40-51). WB Saunders.

[8] Bellomo, R., Ronco, C., Kellum, J.A., Mehta, R.L. and Palevsky, P., 2004. Acute renal failure–definition, outcome measures, animal models, fluid therapy and information technology needs: the Second International Consensus Conference of the Acute Dialysis Quality Initiative (ADQI) Group. Critical care, 8(4), p.R204.

[9] Mehta, R.L., Kellum, J.A., Shah, S.V., Molitoris, B.A., Ronco, C., Warnock, D.G. and Levin, A., 2007. Acute Kidney Injury Network: report of an initiative to improve outcomes in acute kidney injury. Critical care, 11(2), p.R31.

[10] Kellum, J.A., Lameire, N., Aspelin, P., Barsoum, R.S., Burdmann, E.A., Goldstein, S.L., Herzog, C.A., Joannidis, M., Kribben, A., Levey, A.S. and MacLeod, A.M., 2012. Kidney disease: improving global outcomes (KDIGO) acute kidney injury work group. KDIGO clinical practice guideline for acute kidney injury. Kidney international supplements, 2(1), pp.1-138.

[11] Garg, A.X. and Parikh, C.R., 2009. Yin and Yang: acute kidney injury and chronic kidney disease.

[12] Chawla, L.S. and Kimmel, P.L., 2012. Acute kidney injury and chronic kidney disease: an integrated clinical syndrome. Kidney international, 82(5), pp.516-524.

[13] Ishani, A., Xue, J.L., Himmelfarb, J., Eggers, P.W., Kimmel, P.L., Molitoris, B.A. and Collins, A.J., 2009. Acute kidney injury increases risk of ESRD among elderly. Journal of the American Society of Nephrology, 20(1), pp.223-228.

[14] Anderson, S., Eldadah, B., Halter, J.B., Hazzard, W.R., Himmelfarb, J., Horne, F.M., Kimmel, P.L., Molitoris, B.A., Murthy, M., O'Hare, A.M. and Schmader, K.E., 2011. Acute kidney injury in older adults. Journal of the American Society of Nephrology, 22(1), pp.28-38.

[15] Bucaloiu, I.D., Kirchner, H.L., Norfolk, E.R., Hartle II, J.E. and Perkins, R.M., 2012. Increased risk of death and de novo chronic kidney disease following reversible acute kidney injury. Kidney international, 81(5), pp.477-485.

[16] Murugan, R. and Kellum, J.A., 2011. Acute kidney injury: what's the prognosis?. Nature Reviews Nephrology, 7(4), p.209.

[17] Prikryl, P., Vojtova, L., Maixnerova, D., Vokurka, M., Neprasova, M., Zima, T. and Tesar, V., 2017. Proteomic approach for identification of IgA nephropathy-related biomarkers in urine. Physiological research, 66(4), pp.621-632.

[18] L'Imperio, V., Smith, A., Chinello, C., Pagni, F. and Magni, F., 2016. Proteomics and glomerulonephritis: A complementary approach in renal pathology for the identification of chronic kidney disease related markers. PROTEOMICS–Clinical Applications, 10(4), pp.371-383.

[19] Fassett, R.G., Venuthurupalli, S.K., Gobe, G.C., Coombes, J.S., Cooper, M.A. and Hoy, W.E., 2011. Biomarkers in chronic kidney disease: a review. Kidney international, 80(8), pp.806-821.

[20] Xu, K., Rosenstiel, P., Paragas, N., Hinze, C., Gao, X., Shen, T.H., Werth, M., Forster, C., Deng, R., Bruck, E. and Boles, R.W., 2017. Unique transcriptional programs identify subtypes of AKI. Journal of the American Society of Nephrology, 28(6), pp.1729-1740.

[21] Stevens, L.A., Coresh, J., Greene, T. and Levey, A.S., 2006. Assessing kidney function—measured and estimated glomerular filtration rate. New England Journal of Medicine, 354(23), pp.2473-2483.

[22] Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF, 3rd, Feldman HI, et al. A new equation to estimate glomerular filtration rate. Ann Intern Med. 2009;150(9):604-12.

[23] Levey AS, Coresh J, Greene T, Stevens LA, Zhang YL, Hendriksen S, Kusek JW, Van Lente F; Chronic Kidney Disease Epidemiology Collaboration. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. Ann Intern Med. 2006 Aug 15;145(4):247-54.

[24] https://www.britannica.com/science/serum-albumin

[25] https://medbroadcast.com/condition/getcondition/glomerulonephritis

[26] Lager, D.J. and Abrahams, N., 2012. Practical Renal Pathology, A Diagnostic Approach E-Book: A Volume in the Pattern Recognition Series. Elsevier Health Sciences.

[27] D'amico, G., 1987. The commonest glomerulonephritis in the world: IgA nephropathy. QJM: An International Journal of Medicine, 64(3), pp.709-727.

[28] Moldoveanu, Z., Wyatt, R.J., Lee, J.Y., Tomana, M., Julian, B.A., Mestecky, J., Huang, W.Q., Anreddy, S.R., Hall, S., Hastings, M.C. and Lau, K.K., 2007. Patients with IgA nephropathy have increased serum galactose-deficient IgA1 levels. Kidney international, 71(11), pp.1148-1154.

[29] Julian, B.A., Wittke, S., Haubitz, M., Zürbig, P., Schiffer, E., McGuire, B.M., Wyatt, R.J. and Novak, J., 2007. Urinary biomarkers of IgA nephropathy and other IgA-associated renal diseases. World journal of urology, 25(5), pp.467-476.

[30] Kaneshiro, N., Xiang, Y., Nagai, K., Kurokawa, M.S., Okamoto, K., Arito, M., Masuko, K., Yudoh, K., Yasuda, T., Suematsu, N. and Kimura, K., 2009. Comprehensive analysis of short peptides in sera from patients with IgA nephropathy. Rapid Communications in Mass Spectrometry: An International Journal Devoted to the Rapid Dissemination of Up-to-the-Minute Research in Mass Spectrometry, 23(23), pp.3720-3728.

[31] Gross, J.L., De Azevedo, M.J., Silveiro, S.P., Canani, L.H., Caramori, M.L. and Zelmanovitz, T., 2005. Diabetic nephropathy: diagnosis, prevention, and treatment. Diabetes care, 28(1), pp.164-176.

[32] Cervera, R., 2006. Systemic lupus erythematosus in Europe at the change of the millennium: lessons from the "Euro-Lupus Project". Autoimmunity reviews, 5(3), pp.180-186.

[33] Pagni, F., Galimberti, S., Goffredo, P., Basciu, M., Malachina, S., Pilla, D., Galbiati, E. and Ferrario, F., 2013. The value of repeat biopsy in the management of lupus nephritis: an international multicentre study in a large cohort of patients. Nephrology Dialysis Transplantation, 28(12), pp.3014-3023.

[34] Gurevitz, S., Snyder, J., Wessel, E., Frey, J. and Williamson, B., 2013. Systemic lupus erythematosus: a review of the disease and treatment options. The Consultant Pharmacist®, 28(2), pp.110-121.

[35] Moroni, G., Quaglini, S., Radice, A., Trezzi, B., Raffiotta, F., Messa, P. and Sinico, R.A., 2015. The value of a panel of autoantibodies for predicting the activity of lupus nephritis at time of renal biopsy. Journal of immunology research, 2015.

[36] Weening, J.J., D'Agati, V.D., Schwartz, M.M., Seshan, S.V., Alpers, C.E., Appel, G.B., Balow, J.E., Bruijn, J.A., Cook, T., Ferrario, F. and Fogo, A.B., 2004. The classification of glomerulonephritis in systemic lupus erythematosus revisited. Journal of the American Society of Nephrology, 15(2), pp.241-250.

[37] Liu, C.C., Manzi, S. and Ahearn, J.M., 2005. Biomarkers for systemic lupus erythematosus: a review and perspective. Current opinion in rheumatology, 17(5), pp.543-549.

[38] Li, Y., Fang, X. and Li, Q.Z., 2013. Biomarker profiling for lupus nephritis. Genomics, proteomics & bioinformatics, 11(3), pp.158-165.

[39] Keller, B.J., Martini, S., Sedor, J.R. and Kretzler, M., 2012. A systems view of genetics in chronic kidney disease. Kidney international, 81(1), pp.14-21.

[40] Taal, M.W. and Brenner, B.M., 2008. Renal risk scores: progress and prospects. Kidney international, 73(11), pp.1216-1219.

[41] Ingelfinger, F., 1999. Dorland's Medical Dictionary.

[42] Podocin inactivation in mature kidneys causes focal segmental glomerulosclerosis and nephrotic syndrome.

[43] van Timmeren, M.M., van den Heuvel, M.C., Bailly, V., Bakker, S.J., van Goor, H. and Stegeman, C.A., 2007. Tubular kidney injury molecule-1 (KIM-1) in human renal disease. The Journal of pathology, 212(2), pp.209-217.

[44] Ko, G.J., Grigoryev, D.N., Linfert, D., Jang, H.R., Watkins, T., Cheadle, C., Racusen, L. and Rabb, H., 2010. Transcriptional analysis of kidneys during repair from AKI reveals possible roles for NGAL and KIM-1 as biomarkers of AKI-to-CKD transition. American Journal of Physiology-Renal Physiology, 298(6), pp.F1472-F1483.

[45] Cai, L., Rubin, J., Han, W., Venge, P. and Xu, S., 2010. The origin of multiple molecular forms in urine of HNL/NGAL. Clinical Journal of the American Society of Nephrology, 5(12), pp.2229-2235.

[46] Eckardt, K.U., Coresh, J., Devuyst, O., Johnson, R.J., Köttgen, A., Levey, A.S. and Levin, A., 2013. Evolving importance of kidney disease: from subspecialty to global health burden. The Lancet, 382(9887), pp.158-169.

[47] Van't Veer, L.J., Dai, H., Van De Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., Van Der Kooy, K., Marton, M.J., Witteveen, A.T. and Schreiber, G.J., 2002. Gene expression profiling predicts clinical outcome of breast cancer. nature, 415(6871), p.530.

[48] Good, D.M., Zürbig, P., Argiles, A., Bauer, H.W., Behrens, G., Coon, J.J., Dakna, M., Decramer, S., Delles, C., Dominiczak, A.F. and Ehrich, J.H., 2010. Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease. Molecular & cellular proteomics, 9(11), pp.2424-2437.

Chapter 3 Methods

This chapter introduces the methodology followed in the study. This PhD thesis exhibits a proteomic biomarker discovery scheme that requires as input transcriptomic data from patients. CKD is used as case study. The pipeline combines biological models extracted from publicly available databases with state-of-the-art algorithms that aim to capture meaningful biology for the disease. The goal is to suggest a number of proteins as candidate biomarkers to be measured in blood samples that are able to asymptomatically diagnose a myriad of subcategories, and use machine learning methods to integrate them into a panel of increased performance and stability. The cornerstone is to molecularly determine the underlying pathophysiological mechanism causing CKD and detect it before the damage is done. Briefly, our methodology includes 5 steps (Figure 9):

1) Integrate protein-protein interaction networks with annotated gene-sets to construct our knowledge-base (KB) which comprises contextual information about the biological entities participating in our model.
2) Feed this KB with CKD data obtained from the Gene Expression Omnibus (GEO) to create disease-specific models.
3) Apply different state-of-the-art omic-analysis methods (network analysis, pathway analysis, etc.) to describe a molecular profile for each biological entity included in the model.
4) Select and optimize the features included in the above-described molecular profile by applying an iterative active-learning approach in order to identify the best quality candidate biomarkers. The evaluation of performance as biomarkers is assessed by using plasma samples.
5) Integrate the individual candidates into a biomarker panel with increased performance and stability.

The chapter described below does not follow the order presented in the scheme above. Instead, it follows the methodology in the same order it evolved: from the acquisition of the gene expression data and their analysis, to the construction of the molecular profile, training of the model and posterior validation using proteomics data.

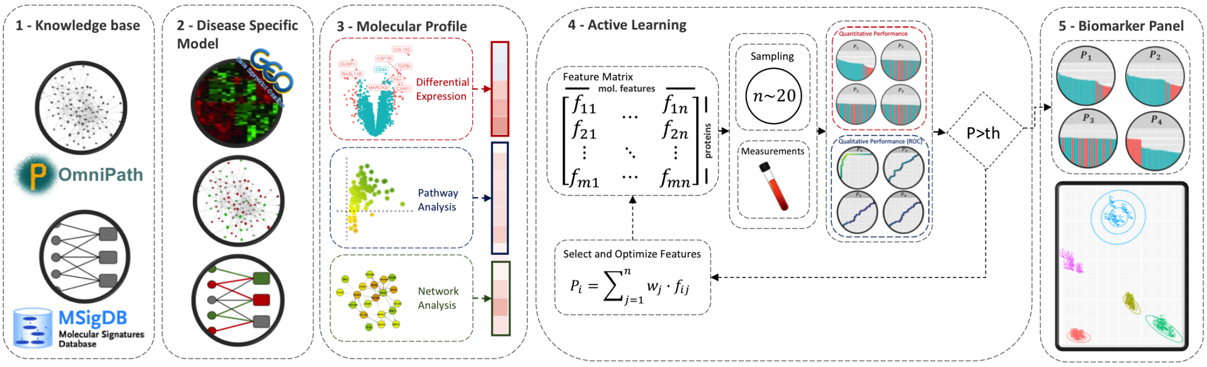


***Figure 9 - Schematic representation of the methodology.** Briefly our methodology follows 5 steps: 1) Construction of the knowledge base, 2) Generation of the disease specific model, 3) Construction of the molecular profile, 4) Construction of the predictive model, 5) Generation of the biomarker panel.*

Data Acquisition and Pre-Processing

The acquisition and pre-processing steps involve everything from downloading the gene expression data from public databases, through the pre-process it to make it comparable, to its integration it in a single dataset.

Data Collection

Gene expression data was obtained from studies reported in Nephroseq V3 (https://nephroseq.org/resource/main.html), a rapidly growing database that collects gene expression studies in renal disease. From the studies included in the database, those meeting the following criteria were initially chosen:

- Expression is measured in human samples.
- Expression is measured in glomerular or tubular tissues.
- Expression is measured using Affymetrix technology.
- Samples do not belong to transplants.
- Dataset includes at least two groups.

From the selected studies, raw expression data was downloaded from the Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo/). For each GEO series (GSE), the raw data (CEL files) were downloaded and the expression values normalized using the Robust Multi-array Average (RMA) algorithm. Genetic variants were aggregated into genes as genes are usually represented by the most abundant trait to characterize the whole gene (HUGO Gene Nomenclature Committee, HGNC format). DNA-methylation probes mapping to the same gene were median centered to obtain a single value per gene. Patient related data was obtained from Nephroseq.

Data Filtering

Data was filtered in different levels following the subsequent criteria:

- **Expression data:** Only genes appearing in all samples were kept for further analysis.
- **Patient data:** Incomplete patient data was removed by removing those features not included in at least 75% of the samples.

- **Samples:** Under-represented sample groups (samples from the same experiment, platform, tissue, and disease group) were removed by filtering those groups with less than 5 samples in each tissue, thus reducing the sparsity of the dataset.

Data Integration

Expression data from individual experiments was evaluated for batch effects using Principal Component Analysis (PCA). Evaluation of batch existence was assessed by visual inspection. Intra-experiment expression data was adjusted for batch effect using the Combining Batches of Gene Expression Microarray Data (ComBat) algorithm [1], a method that uses an empirical Bayes framework to adjust for batch effects. Combat was the method of choice as it has shown increased performance when compared with five other well-known batch effect adjustment methods in independent studies [2]. The adjustment is calculated without setting any covariates because adjusting for batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses [3].

Secondly, expression data from different GSEs was integrated for each tissue. A bipartite graph was created joining GSE and disease groups. Disconnected GSEs or disease groups were not included in the integration. Evaluation of batch presence and correction was again evaluated using PCA and ComBat (without covariates).

From the integrated dataset, only those patients belonging to the diabetes nephritis (DN), healthy living donor (HLD), immunoglobulin A nephropathy (IgAN), and lupus nephritis (LN) groups were filtered as these are the groups that were also experimentally validated.

Generation of Molecular Profile

The generation of the molecular profile involves all the different analyses used to obtain biological features from the expression data using state of the art analytical methods. This includes gene-level statistics (differential expression analysis), pathway-level statistics (pathway analysis), and network-level statistics (network analysis).

Pathway Database (MSigDB)

The Molecular Signatures Database (MSigDB), a collection of human signaling pathways, was the database of choice for the pathway definitions. We selected the curated gene sets (C2), canonical pathways (CP) dataset (version 6.2 updated July 2018), as it contains gene sets curated from various online pathway databases such as Biocarta, Reactome, KEGG, and PID. Gene sets were further classified based on their size into two groups using Gaussian mixture models. Large gene sets representing general biology were removed from further analysis.

Protein-Protein-Interaction Database (OmniPath)

Protein-protein-interactions (PPIs) were obtained from OmniPath [4], a database that collects protein-protein-interactions from different public databases (String, HPRD, etc.). Inconsistencies between the different databases were excluded by including only those interactions that were reported in at least 3 external sources.

Gene-Level Statistics (GLS)

Differential expression analysis was performed using Limma [5]. Limma first fits a linear model for each gene and then computes the estimated coefficients and standard errors for a given set of contrasts to finally calculate differential expression using empirical Bayes statistics. Contrasts act as a method for defining comparisons between groups (Case and Control). In this study, given that the number of groups is larger than two, we build a contrast for each disease group using as a control a pool of all the remaining disease groups (DN vs HLD+IgAN+LN, for example).

From Limma analysis, the logarithm in base 2 of the fold-change (logFC) and the minus logarithm in base 10 of the p-value (logP) were included in the molecular profile.

Pathway-Level Statistics (PLS)

The second set of molecular features are the pathway-level statistics. These features come from first running a gene-set enrichment (pathway analysis) method, and second integrating the pathway enrichment scores back to the genes. In this section we will be the terms gene-set analysis and pathway analysis indistinctively. However, in the strict sense, this only holds true when pathways are represented as gene sets.

Gene-Set Analysis: Gene-set enrichment analysis was performed using piano [6], a pathway analysis method that allows to use any number of 11 different gene set analysis methods and offers different alternatives to integrate the results. Each method requires a user-defined gene-level statistic (GLS) as input. For the present study, the following 10 methods were chosen using the GLSs in the parenthesis:

- Fisher's combined probability test (p-value)
- Stouffer's method (p-value)
- Reporter features (p-value)
- Parametric Analysis of Gene-set Enrichment (PAGE) (t-value)
- Tail Strength (p-value)
- Gene Set Enrichment Analysis (GSEA) (t-value)
- Mean statistic (fold-change)
- Median statistic (fold-change)
- Sum statistic (fold-change)
- MaxMean statistic (t-value)

For each method, each pathway is represented by a set of 5 p-values including different directionality classes:

- **Distinct-up:** Up-regulation (up and down are cancelled out).
- **Mixed-up:** Up-regulation (up and down are not cancelled out, a pathway can be mixed-up and mixed-down regulated at the same time).
- **Non-Directional:** Differentially regulated (regardless the sign).
- **Mixed-down:** Down-regulation (up and down are not cancelled out).
- **Distinct-down:** Down-regulation (up and down are cancelled out).

Consensus scoring of the gene sets is achieved as described in the original paper [6], which consists of aggregating the ranks of the gene sets based on p-values rather than aggregating the p-values directly. The median of the ranks of a given gene set is used as the consensus score.

Map pathways to genes: For each gene, the collection of pathways in which it participates was obtained by recycling the same biological model used in gene-set analysis (MSigDB). These p-values were combined using Hartung's method [7] which incorporates the correlation among p-values via introducing the correlation matrix in the Z-transform test.

From this analysis, the minus logarithm in base 10 of each of the 5 p-values as well the logarithm in base 2 of the number of pathways in which each gene participates were included in the molecular profile.

Network-Level Statistics (NLS)

The third set of molecular features are the network-level statistics. These features come from applying a custom network analysis method that integrates differential expression and topological information.

In this study, we used a custom guilt-by-proximity approach in which the network-score for each protein is calculated based on the GLSs of the neighboring nodes as described in Eq. 1:

$$NLS_i = \frac{1}{log_2(C_i + 1)} \left(\sum_{j=1}^{DM} \left(\frac{1}{D_j} \cdot \sum_{\delta=1}^{D_j} \left(\frac{GLS_\delta}{j^2} \right) \right) \right)$$

where C_i is the centrality of the node i (betweenness), DM is the maximum distance from any node δ connected by a distance j to the source i, D_j is the number of nodes in distance j from the source i, and GLS_δ is the GLS value of the node δ in distance j from the source i. The distance j between nodes i and δ is calculated using Dijkstra algorithm for the shortest path problem. When applying the formula using logFC as GLS, we take the absolute value of logFC so the positive and negative values do not cancel out when applying their summation.

From network analysis, the network scores using both GLSs as well as the logarithm in base 2 of the degree of each entity in the network were included in the molecular profile.

Map to Protein Identifiers

HGNC symbols were mapped using the UniProt database (https://www.uniprot.org/) (taxonomy 9606, Homo Sapiens) to protein identifiers (UNIPROTKB). We filtered-in those proteins that were reviewed (existence experimentally probed, not only predicted). Different HGNC symbols mapping to the same UNIPROTKB symbol were mean centered while HGNC symbols mapping to several UNIPROTKB symbols were duplicated.

Feature Integration

The molecular profile of each protein was integrated into a feature matrix. Only the proteins containing data for all the features (complete observations) were kept for further analysis.

Feature Selection

Non-informative or highly correlating features were removed using three approaches:

- **Estimation of predictor variance:** we evaluated the coefficient of variation (CV) of each predictor individually and selected those features with at least 40% of CV.
- **Estimation of predictor correlation:** we evaluated the correlation coefficient of each pair of features and selected those with a Pearson correlation coefficient smaller than 0.85.
- **Estimation of linear dependencies:** we used the QR decomposition of the feature matrix to find sets of linear combinations between the features (if they exist).

Secretion Prediction

The feature matrix was expanded using information about the subcellular location of each protein. Known subcellular locations were obtained from the UniProt database while the unknown ones were predicted using a prediction model developed together with C. Fotis (unpublished) (Figure 10) and similar to [8]. Briefly, the method tries to predict if a protein is going to be secreted from characteristic sequences that can be identified in the amino acid chain of a protein following the subsequent steps:

- For training, evaluation and testing of the model the manually curated dataset of human proteins was downloaded from UniProt.
- Subcellular locations for known proteins were classified in three different categories: intracellular, membranous, extracellular.
- Protein features were created as to capture the information needed for cellular secretion (Figure 10.A). For a protein to be secreted through a traditional secretory pathway [9, 10] it must possess a characteristic signal peptide and must lack a transmembrane domain.
- Five different methods were used for secretion prediction predicting different were used to predict the existence of signaling peptides given its sequence of amino acids for each protein: SignalP [11], TargetP [12], TMHMM [13], Phobius [14] and WolfPSORT [15]. Briefly, SignalP is 3 layer feed forward neural network that outputs the probability of the protein having a Signal peptide in its amino acid sequence; TargetP uses a similar approach to predict subcellular location; TMHMM, is a hidden Markov Model that outputs the probability of the protein having a membrane domain inside its amino acid sequence; Phobius, is a hidden Markov Model that predicts the presence of both signal peptides and membrane domains inside a protein's amino acid sequence; and WolfPSORT is a KNN model that uses the presence of functional domains inside a protein's amino acid sequence to predict its subcellular location.
- The performance of each individual method was individually validated using the Uniprot dataset.
- The output of each method was integrated into a feature vector describing the presence of signal peptides for each protein. This feature vector then served as input for training and testing a random forest binary classifier that predicts protein secretion. The classifier was trained and optimized using a 10-fold cross validation scheme.

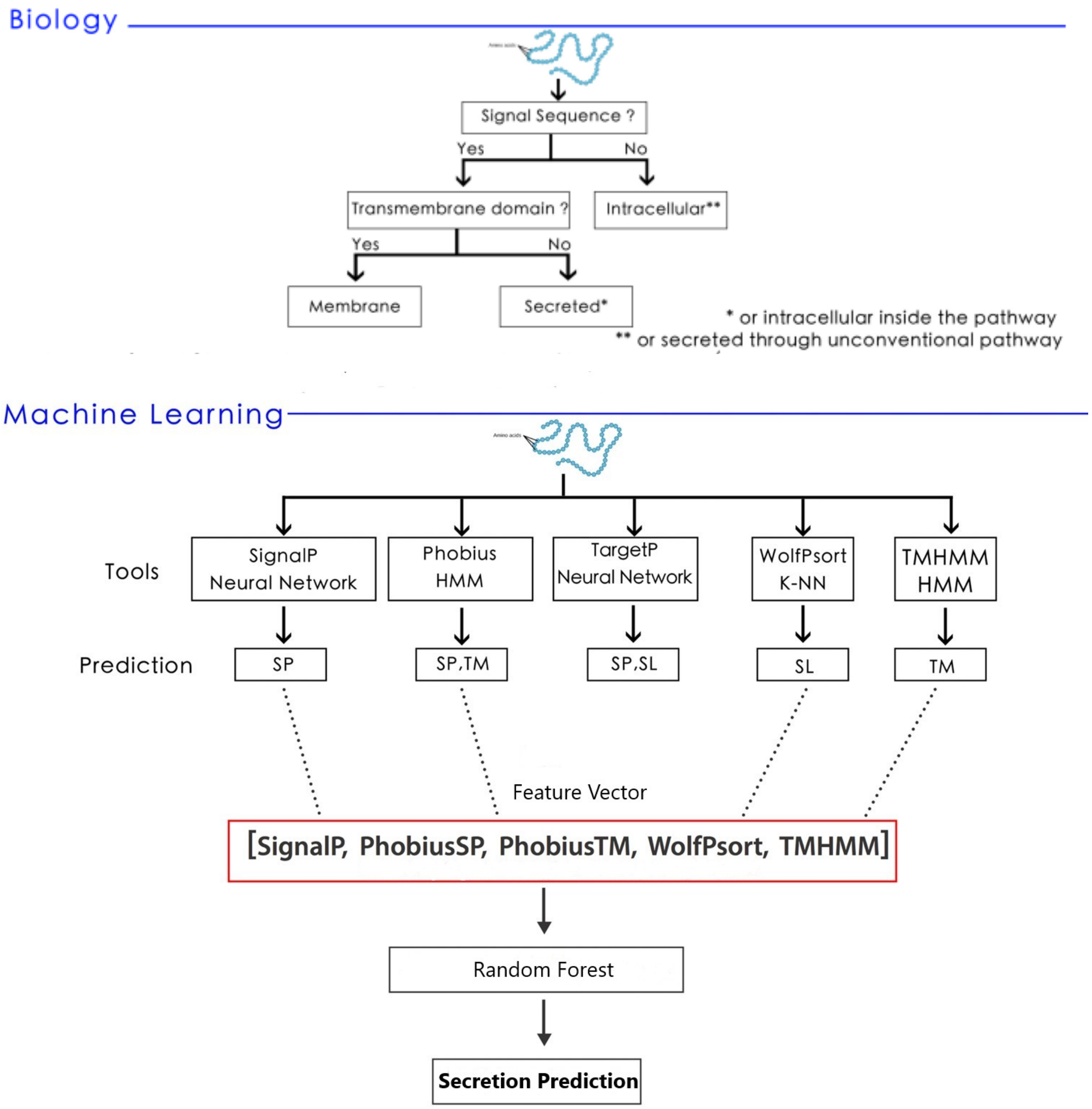


Figure 10 - Secretion prediction. *A) Biology. A protein is secreted through a conventional pathway if it possesses a signal sequence and lacks a transmembrane domain. B) Machine learning. This biology was captured using a combination of 4 methods to predict whether if a protein is secreted or not.*

Experimental Design

This section of the methods involves everything related to the proteomic experiments performed, from acquisition to measurement. It also describes the measured analytes and some quality control checkpoints introduced.

Sample acquisition

79 human plasma samples (20 DN, 19 HLD, 19 IgAN, and 21 LN) were obtained from patients enrolled from Brigham and Women's Hospital (BWH, Boston, US), Beth Israel Deaconess HealthCare (BIDMC, Boston, US), and Massachusetts General Hospital (MGH, Boston, US) undergoing native kidney biopsy and venipuncture. Diagnosis was performed based on histopathology (2 expert pathologists blinded to diagnosis).

Sample preparation

Plasma samples were heat-inactivated to destroy complement prior to performing dual-antibody Luminex assays. Samples were stored at -80°C. They were thawed at room temperature and gently mixed after thawing. Then, they were placed in a 56°C water bath for 30 min and gently mixed after the incubation. The inactivated plasma samples were diluted 1 to 10 with Low cross normal buffer (CANDOR) before the measurements.

xMAP technology – Proteomics

Proteomics measurements were performed in triplicates using the Luminex xMAP technology. Briefly, Luminex xMAP is an instrument for multiplex detection of proteins in a single biological sample using bead-based sandwich immunoassays (ELISA-like). It is based on polystyrene or paramagnetic microspheres (beads) that are internally dyed with red and infrared fluorophores of differing intensities. Each dyed bead is given a unique number (bead region) allowing the differentiation of one bead from another. Multiple analyte-specific beads can then be combined in a single well of a 96-well microplate-format assay to detect and quantify multiple targets simultaneously, using the Luminex instrument for analysis. The system can simultaneously detect many targets in a single sample depending on the system design. Using a dual laser system, the signature of each bead is identified, and the presence and intensity of reporter associated with the bead is detected. The intensities are usually reported by the median of all the beads with the same ID. This gives information about both the identity and concentration of targets in the sample. Proteins including antibodies, ligands, and nucleic acids specific to the targets can be coupled to the beads.

As part of this PhD thesis, 24 (14 training, 10 validation) custom dual-antibody Luminex assays were developed using ProtATonce's (Athens, Greece) multiplex assay service (Table 5). 2 to 4 antibodies were selected and cleaned up from amine containing buffers and carrier proteins that interfere with the coupling procedure. All antibodies were tested pair-wise as capture and as detection antibody. Capture antibodies were coupled to the beads whereas detection antibodies were biotinylated (if not already biotinylated). Quality control confirmed biotinylation and coupling efficiency. In each assay, the optimal capture/detection antibody pair was selected based on signal-to-noise ratio measurement.

All the different capture antibodies were coupled to Luminex magnetic beads, different bead regions, and were detected using analyte-specific biotinylated antibodies that bind to the appropriate epitope of the immobilized analyte. The 24 conjugated capture antibodies and the 24 biotinylated detection antibodies were multiplexed to create the bead mix and the detection mix, respectively. 50μl of the coupled beads (bead mix) were incubated with the samples on a flat bottom 96-well plate on a shaker at 900 rpm for 90 minutes at room temperature. Then, detection mix was added, and the samples incubated on a shaker at 900 rpm for 60 minutes at room temperature. The final step was the addition of freshly prepared SAPE solution (Streptavidin, R-Phycoerythrin conjugate, Cat Nr: S866, Invitrogen) for the detection of the signal. 15 minutes after the incubation with SAPE, samples were measured with the Luminex FlexMAP 3D instrument.

***Table 5 - Developed assays**. 24 custom dual-antibody Luminex assays were developed using ProtATonce's multiplex assay service.*

Protein Name	**UNIPROTID**	**Catalog Number**	**Provider**	**Round**
ANXA2	P07355	DYC3928-2	R&D Systems	Training
CCL2	P13500	900-K31	Peprotech	Training
CCL5	P13501	900-K33	Peprotech	Validation
CFH	P08603	DY4779	R&D Systems s	Training
COL3A1	P02461	DY6220-05	R&D Systems	Training
CNTF	P26441	900-K158	Peprotech	Training
CXL11	O14625	900-K151	Peprotech	Validation
DEFB1	P60022	900-K202	Peprotech	Training
GAL3	P17931	DY1154	R&D Systems	Validation
GROA	P09341	900-K38	Peprotech	Training
hsTropT	P45379	MAB1874	R&D Systems	Training
ICAM1	P05362	900-K464	Peprotech	Training
IL1A	P01583	900-K11	Peprotech	Validation
IL1B	P01584	DY201	R&D Systems	Validation
IL20	Q9NYY1	900-K224	Peprotech	Training
IL6	P05231	900-K16/ DY206-05_Det	Peprotech/R&D Systems	Validation
MMP1	P03956	DY901B	R&D Systems	Validation
MMP9	P14780	DY911	R&D Systems	Validation
NPHS1	O60500	DY4269-05	R&D Systems	Validation
NRG1	Q02297	900-K316	Peprotech	Training
RETN	Q9HD89	900-K235	Peprotech	Training
TFF3	Q07654	RD191160200R	Biovendor	Training
TNF10	P50591	900-K141	Peprotech	Validation
TNFA	P01375	900-K25	Peprotech	Training

Proteomics – Quality control

Proteomic measurements were accepted based on the following criteria:

- **Bead Counts QC:** A measurement is accepted if the corresponding bead count is above 30 [16].
- **Replicate QC:** A measurement is accepted if is less than 3 median absolute deviations from the median MFI of the replicates.
- **Analyte distribution QC:** A measurement is accepted if is less than 3 standard deviations from the mean MFI of the analyte.

All the valid measurements were median centered. If all the triplicates for a particular sample were invalid, the value was substituted by NA. Patients with more than 1 NA were considered outliers and removed from the analysis. For patients with a single NA, the value was substituted by the median for that analyte.

Background Evaluation: Valid measurements were compared versus blank/background measurements, (measurements performed with antibody but without patient sample). We tested that patient measurements for each protein were significantly higher than those of the background using one-tiled t-tests. Obtained p-values were adjusted for multiple comparisons using false-discovery-rate (FDR) method. Those proteins with an adjusted p-value over 0.01 were removed from further analysis.

Downstream analysis

Exploratory analysis

Data was visually explored by applying three unsupervised machine learning clustering methods: PCA, tSNE, and Hierarchical Clustering (HC).

Definition of biomarker performance

Our definition of biomarker performance (BP) integrates two types of metrics: quantitative (differential expression analysis, -log10(p-value), log2(FC)) and qualitative (ROC analysis, AUC).

$$BP = 75 \cdot AUC + 15 \cdot \widehat{logFC} + 10 \cdot \widehat{logP}$$

Where BP represents biomarker performance, AUC represents the area under the ROC curve, and logFC and logP represent the log2 of the fold-change and -log10 of the p-value from differential expression analysis after some transformations (quantitative):

$$\widehat{logFC} = min(|log_2(FC)|, 2)$$

$$\widehat{logP} = \min(-log_{10}(p), 2)$$

The justification behind these transformations is that while AUC is a parameter that naturally ranges between 0 and 1 (with 0.5 representing the random case), the log2(FC) ranges between -inf to +inf and the -log10(p-value) ranges between 0 and +inf. Therefore, to keep our BP score between 0 and 100, we trimmed the log2(FC) between -2 and 2 considering that a 2-fold increase/decrease in the expression was enough and took its module because we are interested in absolute differences regardless of their direction. For the p-value we put the threshold in 0.01 for the same reason.

Model training

Features from the feature matrix were z-normalized and trimmed in the [-5, +5] range to avoid strong outlier influence. Biomarker performance (BP) was predicted using elastic network models for the glomerular and the tubular tissues as implemented in the glmnet R package [17].

$$min_{\beta_0,\beta}\left(\frac{1}{N}\sum_{i=1}^{N}\omega_i l(y_i,\beta_0+\beta^T x_i)+\lambda\left[(1-\alpha)\frac{||\beta||_2^2}{2}+\alpha||\beta||_1\right]\right)$$

Here, l(y,x) is the negative log-likelihood contribution for the observation I while β represents the weights given to the features x. The elastic-net penalty is controlled by α which connects the extreme cases of Lasso (α = 1), and Ridge (α = 0). The parameter λ controls the strength of the penalty in the optimization function. In this study, λ was selected so that gives the minimum mean cross-validated error (λ.min).

Elastic networks represent a regularized regression method that combine the L1 (Lasso, size) and L2 (Ridge, performance) penalties of the Lasso and Ridge regressions [18]. Model training was performed under a 10-times nested 10-fold cross-validation scheme which is less likely of overfitting the data than non-nested cross-validation [19]. Stratification was performed so that each fold kept the same proportion of samples for each subtype.

Weights for each fold were mean-centered and used to predict biomarker performance in the entire dataset. The final prediction was obtained averaging the prediction for the glomerular and the tubular models.

Model validation

10 proteins were selected for validation based on several criteria: predicted biomarker performance, literature, and antibody availability to report the protein of interest. Samples were pre-processed, measured and analyzed as in the training set. Model validation was then performed comparing the obtained residuals (differences between the predicted and measured biomarker performance) of each protein in both the training and validation sets.

Integration into biomarker panel

Combinations of candidate biomarkers into panels with increased performance were then evaluated. Linear discriminant analysis (LDA) was the method of choice for candidate integration. Briefly, panels of size 2-7 were exhaustively trained using a 5-fold cross validation scheme. Then, the Pareto front was constructed by choosing the best performing panel for each size. Here, best was defined using a composite score reflecting both quantitative and qualitative indicators. For the quantitative indicator, the posterior residual probabilities of the samples (probability of misclassification) was chosen, while panel accuracy (percentage of correctly classified samples) was chosen as the qualitative indicator. Finally, both indicators were ranked, and the rankings averaged to select the best panel for each size.

The optimal panel size was selected based on the elbow criterion on the previously generated Pareto front. We again used quantitative and qualitative indicators of panel performance. For the qualitative indicator the panel percentage of correctly classified samples was reused. For the quantitative indicator: first, samples were down-scaled to the two first components returned by LDA (LD1 and LD2); second, each subgroup was modelled as a two-dimensional normal distribution; and third, the Bhattacharyya distance [20], was calculated between each pair of distributions.

$$D_B(i,j) = \frac{1}{8}\left(\boldsymbol{\mu_i} - \boldsymbol{\mu_j}\right)^T \boldsymbol{\Sigma^{-1}}\left(\boldsymbol{\mu_i} - \boldsymbol{\mu_j}\right)\frac{1}{2}\left(\frac{det\ \boldsymbol{\Sigma}}{\sqrt[2]{det\ \boldsymbol{\Sigma_i} \cdot det\ \boldsymbol{\Sigma_j}}}\right)$$

Where $\mathrm{D_B}(\mathrm{i}, \mathrm{j})$ is the Bhattacharyya distance between distributions *i* and *j*, $\boldsymbol{\mu}_\mathbf{i}$ is the mean of the distribution *i*, $\boldsymbol{\Sigma}_\mathbf{i}$ is the covariance of the distribution *i*, and $\boldsymbol{\Sigma} = \frac{\boldsymbol{\Sigma_1}+\boldsymbol{\Sigma_2}}{2}$.

The assays measuring the proteins included in the optimal panel were further characterized by constructing the calibration curves that allow to map the MFI values to molarity (ng/ml) (Appendix 1).

Correlation with eGFR

Proteomic measurements were also analyzed for correlation with the estimated glomerular filtration rate (eGFR) by fitting linear models between the mean fluorescence intensity values (MFIs) and the eGFR.

Implementation as shiny app

The pipeline described here was implemented as an interactive online application in Shiny (from RStudio): https://asierantoranz91.shinyapps.io/MBBD. The app implements a reduced version to speed up the calculation of the results. Specifically, the calculation of the pathway-level-statistics has been implemented using 2 of the 10 applied methods (Fisher's and Stouffer's) and can only handle 2-group comparisons (case vs control).

References

[1] Johnson, W.E., Li, C. and Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 8(1), pp.118-127.

[2] Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L. and Liu, C., 2011. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PloS one, 6(2), p.e17238.

[3] Nygaard, V., Rødland, E.A. and Hovig, E., 2016. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics, 17(1), pp.29-39.

[4] Türei, D., Korcsmáros, T. and Saez-Rodriguez, J., 2016. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nature methods, 13(12), p.966.

[5] Smyth, G.K., 2005. Limma: linear models for microarray data. In Bioinformatics and computational biology solutions using R and Bioconductor (pp. 397-420). Springer, New York, NY.

[6] Väremo, L., Nielsen, J. and Nookaew, I., 2013. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic acids research, 41(8), pp.4378-4391.

[7] Hartung, J., 1999. A note on combining dependent tests of significance. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 41(7), pp.849-855.

[8] Min, X.J., 2010. Evaluation of computational methods for secreted protein prediction in different eukaryotes. JPB, 3(5).

[9] Nickel, W., 2010. Pathways of unconventional protein secretion. Current opinion in biotechnology, 21(5), pp.621-626.

[10] Rabouille, C., 2017. Pathways of unconventional protein secretion. Trends in cell biology, 27(3), pp.230-240.

[11] Petersen, T.N., Brunak, S., Von Heijne, G. and Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature methods, 8(10), p.785.

[12] Emanuelsson, O., Brunak, S., Von Heijne, G. and Nielsen, H., 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nature protocols, 2(4), p.953.

[13] Chen, Y., Yu, P., Luo, J. and Jiang, Y., 2003. Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. Mammalian Genome, 14(12), pp.859-865.

[14] Käll, L., Krogh, A. and Sonnhammer, E.L., 2007. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. Nucleic acids research, 35(suppl_2), pp.W429-W432.

[15] Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J. and Nakai, K., 2007. WoLF PSORT: protein localization predictor. Nucleic acids research, 35(suppl_2), pp.W585-W587.

[16] http://www.bio-rad.com/webroot/web/pdf/lsr/literature/10032257.pdf

[17] Hastie, T. and Qian, J., 2014. Glmnet vignette. Retrieve from http://www. web. stanford. edu/~ hastie/Papers/Glmnet_Vignette. pdf. Accessed September, 20, p.2016.

[18] Zou, H. and Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), pp.301-320.

[19] Cawley, G.C. and Talbot, N.L., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11(Jul), pp.2079-2107.

[20] Bhattacharyya, A., 1943. On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math. Soc., 35, pp.99-109.

Chapter 4 Results

This chapter introduces the results obtained in the present study. It includes results related to data acquisition and pre-processing, to the obtained molecular profile, experimental results (training), model construction, and experimental results (validation).

Data Acquisition and Pre-Processing

Data Collection

Nephroseq included 32 studies at the time of collection (November 2017), 14 of which met the criteria described in the methods (Table 6).

Table 6 - Studies selected from Nephroseq. *The reported columns are study name, organism, tissue, platform, disease group, and GEO series. In the "Groups" column AH stands for Hypertensive Nephropathy, CKD for Chronic Kidney Disease (generic), DN for Diabetic Nephropathy, FSGS for Focal-Segmental Glomerulosclerosis, HLD for Healthy Living Donor, IgAN for Immunoglobulin A Nephropathy, IgM for Immunoglobulin M nephropathy, LN for Lupus Nephritis, MCD for Minimal Change Disease, MGN for Membranous Glomerulonephritis, MPGN for Membranous Proliferative Glomerulonephritis, NPHSCL for Nephrosclerosis, TBMD for Thin Basement Membrane Disease, TN for Tumor Nephrectomy, VCL for Vasculitis.*

Study	Organism	Tissue	Platform	Groups	GEO Series
Berthier Lupus Glom	Homo Sapiens	Glomeruli	GPL96	HLD, LN	GSE32591
Berthier Lupus TubInt	Homo Sapiens	Tubulointerstitium	GPL96	HLD, LN	GSE32591
Hodgin FSGS Glom	Homo Sapiens	Glomeruli	-	FSGS, HLD	-
Ju CKD Glom	Homo Sapiens	Glomeruli	GPL96, GPL570	DN, FSGS, MCD, TN, VCL	GSE47185
Ju CKD TubInt	Homo Sapiens	Tubulointerstitium	GPL96, GPL570	DN, FSGS, MCD, MGN, TBMD, TN	GSE47185
Ju CKD TubInt 2	Homo Sapiens	Tubulointerstitium	GPL570	AH, CKD, DN, FSGS, IgAN, IgM, LN, MCD, MGN, MPGN	GSE69438
Neusser Hypertension Glom	Homo Sapiens	Glomeruli	GPL96	NPHSCL, TN	GSE20602
Reich IgAN Glom	Homo Sapiens	Glomeruli	GPL96	HLD, IgAN	GSE35488
Reich IgAN TubInt	Homo Sapiens	Tubulointerstitium	GPL96	HLD, IgAN	GSE35488
Sampson Nephrotic Kidney Glom	Homo Sapiens	Glomeruli	GPL17692	FSGS, MCD, MN, Other	GSE68127
Sampson Nephrotic Kidney TubInt	Homo Sapiens	Tubulointerstitium	GPL17692	FSGS, MCD, MN, Other	GSE68127
Schmid Diabetes TubInt	Homo Sapiens	Tubulointerstitium	GPL96, GPL570	AH, HLD, IgAN, LN	GSE32591, GSE37455, GSE35488

Woroniecka Diabetes Glom	Homo Sapiens	Glomeruli	GPL571	DN, HLD	GSE30122
Woroniecka Diabetes TubInt	Homo Sapiens	Tubulointerstitium	GPL571	DN, HLD	GSE30122

From these studies, all but the Hodgin FSGS Glom have matching publicly available expression data in the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/). For both Sampson Neprhotic Kidney studies, the annotation is incomplete. Therefore, these 3 studies were excluded from further analysis reducing the number of studies to 11. Furthermore, the Ju CKD studies re-use the expression data of the Reich IgAN, Berthier Lupus, and Schmid Diabetes studies. Consequently, the raw expression data of the 7 independent datasets was downloaded directly from GEO:

- **GSE20602:** Human Nephrosclerosis Triggers a Hypoxia-Related Glomerulopathy [1].
- **GSE30122:** Transcriptome Analysis of Human Diabetic Kidney Disease. [2, 3].
- **GSE32591:** Expression data from human with lupus nephritis (LN) [4].
- **GSE35488:** Expression data from human with IgA nephropathy (IgAN) [5].
- **GSE37463 (superseries of GSE37455):** Cross-species transcriptional network analysis defines shared inflammatory responses in murine and human lupus nephritis [4].
- **GSE47185:** In silico nano-dissection: defining cell type specificity at transcriptional level in human disease [6].
- **GSE69438:** Tissue Transcriptome Driven Identification of Epidermal Growth Factor as a Chronic Kidney Disease Biomarker Gene expression analysis of peripheral blood cells in patients with chronic kidney disease [7].

To these datasets, another GEO series (GSE) dataset was added with matched patient data and satisfying our inclusion criteria stated above but not present in Nephroseq V3 at the time of collection:

- **GSE93798:** Transcriptomic and proteomic profiling reveal insights of mesangial cell function in patients with IgA Nephropathy [8].

The global dataset integrating all samples contained 552 patients and 21656 different HGNC symbols across all samples.

Data Filtering

Expression data: From the 21656 different HGNC symbols mapped from the probes 12549 were common to all samples.

Patient data: From the 33 different features obtained, 8 were present in at least 75% of the samples including:

- **GSE series:** GEO ID of the dataset (100% of the samples).
- **GPL platform:** GEO ID of the platform used to measure the sample (100% of the samples).
- **Accession number:** GEO ID of the sample (100% of the samples).
- **Age:** Age of the patient at the time of biopsy (91.3% of the samples).

- **GFRMDRD:** estimated glomerular filtration rate (eGFR) using the 4-variable MDRD equation at the time of biopsy (87.32% of the samples).
- **Disease group:** Diagnosed condition (100% of the samples).
- **Sex:** Male (0) or female (1) (92.39% of the samples).
- **Tissue:** Tissue from where expression data was obtained (Glomeruli or Tubulointerstitium) (100% of the samples).

Samples: Sample distribution per experiment, platform, disease, and tissue is shown in Figure 11.A. All the under-represented groups (less than 5 samples in each tissue) were removed from further analysis (Figure 11.B). This resulted in the removal of 44 patients from the global dataset to a total of 508.

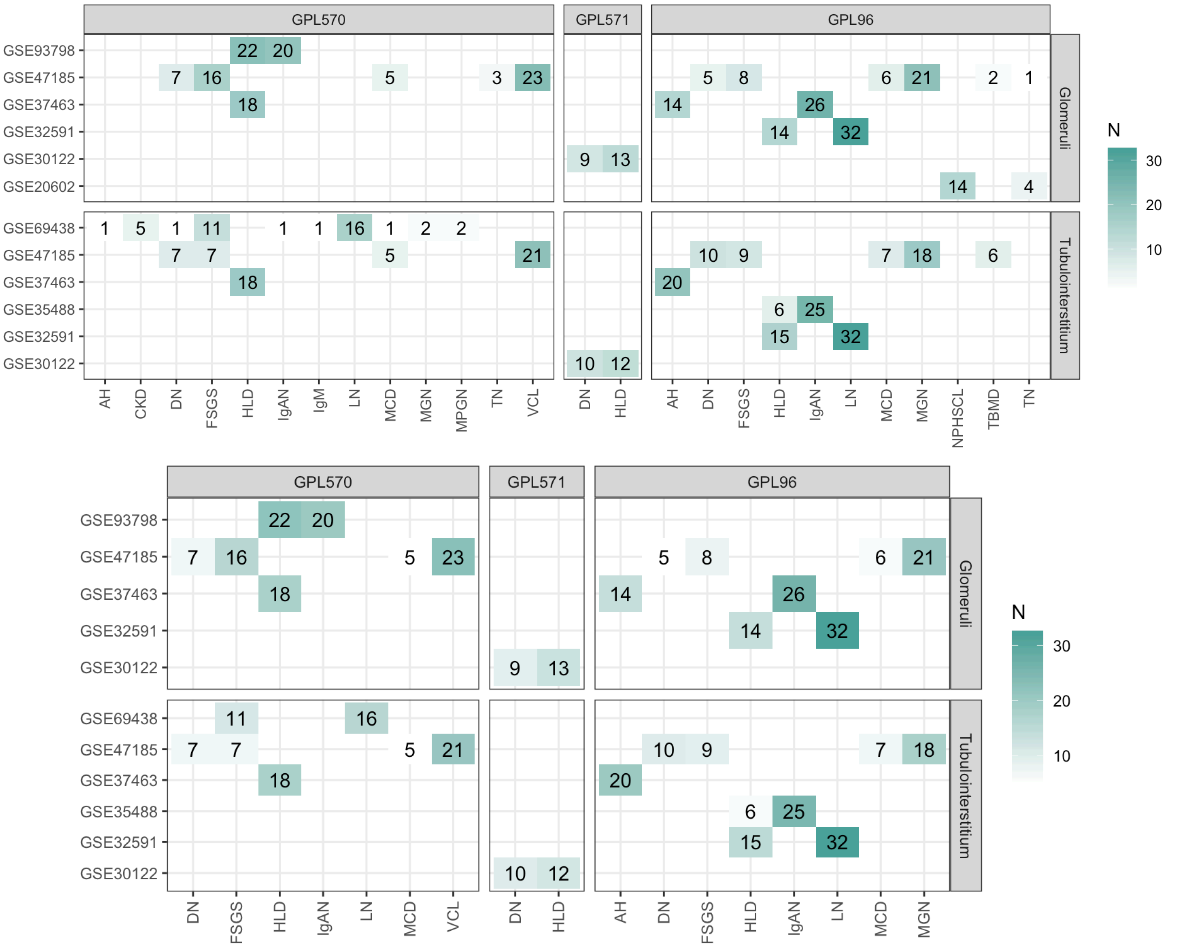


*Figure 11 - **Sample distribution per experiment (y-axis), platform (x-facet), disease (x-axis), and tissue (y-facet).** A) All samples, B) After removing under-represented groups. The number inside each tile represents the number of patients for that particular quadruplet. AH stands for Hypertensive Nephropathy, DN for Diabetic Nephropathy, FSGS for Focal-Segmental Glomerulosclerosis, HLD for Healthy Living Donor, IgAN for Immunoglobulin A Nephropathy, LN for Lupus Nephritis, MCD for Minimal Change Disease, MGN for Membranous Glomerulonephritis, and VCL for Vasculitis.*

Data Integration

Intra-Experiment: Figure 12 shows the projection of each dataset in the first two components after PCA before (A) and after (B) ComBat. The variance explained for each subplot is included in Table 7. Figure 12.A shows that GSE37463 and GSE47185 have internal batch effects due to platform while GSE93798 has an internal batch effect due to the day in which the measurements were carried out.

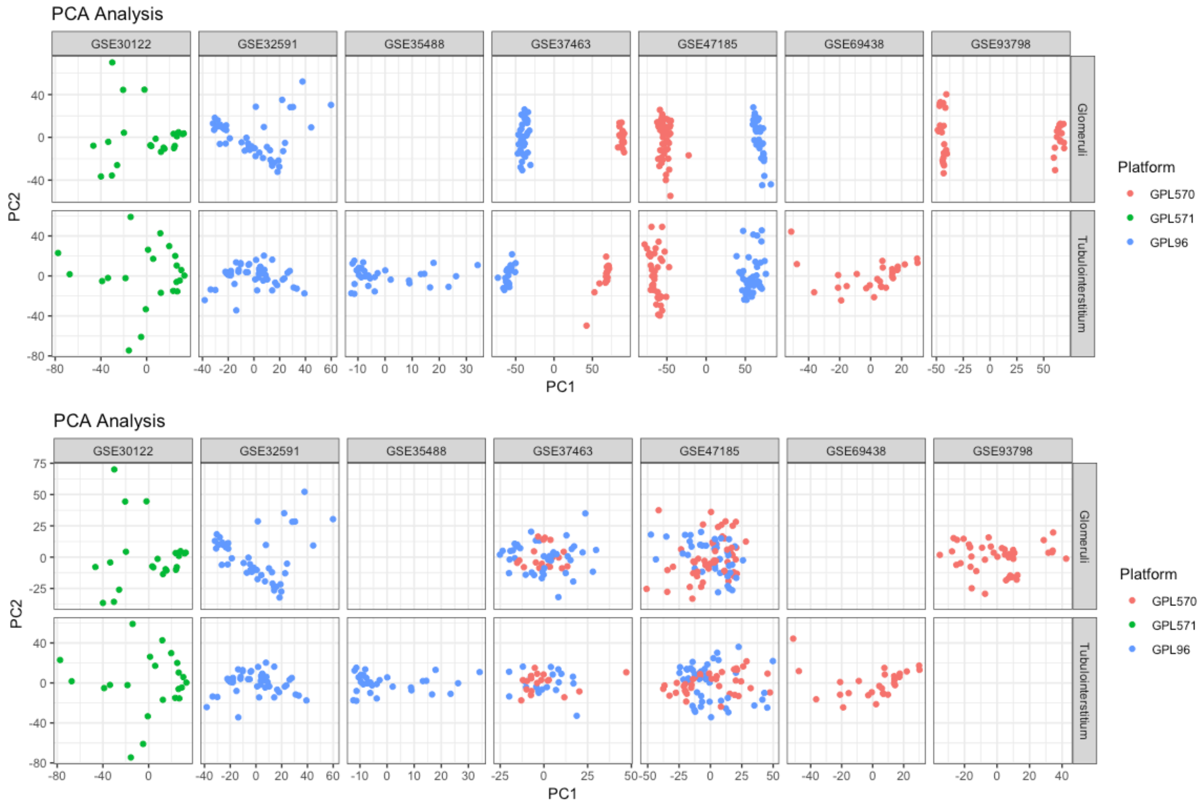


Figure 12 - Batch effect correction (intra-experiment). *Each dataset is represented by a projection of its samples in the first two components after Principal Component Analysis (PCA). A) Before applying ComBat, B) After applying ComBat. Evaluation of batch existence was assessed by visual inspection. Three experiments were identified containing batches: GSE37463 (Platform), GSE47185 (Platform), GSE93798 (Day).*

Table 7 - PCA variance. *Variance explained for PC1 and PC2 in all the subplots from Figure 12.*

Experiment	Tissue	Before		After	
		PC1 (%)	PC2 (%)	PC1 (%)	PC2 (%)
GSE37463	Tubulointerstitium	82,24	2,97	19,07	12,77
GSE37463	Glomeruli	77,14	3,85	16,88	12,76
GSE30122	Tubulointerstitium	27,20	25,40	27,20	25,40
GSE30122	Glomeruli	23,93	21,93	23,93	21,93
GSE47185	Tubulointerstitium	70,85	8,20	27,68	14,97
GSE47185	Glomeruli	63,62	5,56	15,34	12,05
GSE69438	Tubulointerstitium	27,05	12,31	27,05	12,31
GSE32591	Tubulointerstitium	28,85	10,68	28,85	10,68
GSE32591	Glomeruli	26,05	17,68	26,05	17,68
GSE35488	Tubulointerstitium	24,98	12,34	24,98	12,34
GSE93798	Glomeruli	61,06	7,33	21,17	7,55

Inter-Experiment: Figure 13shows the resulting bipartite graph of GSEs and disease groups. All the pairs are connected to the main graph, thus no GSEs need to be removed for this reason. The PCA projections for the integrated dataset are shown in Figure 14. From the integrated and corrected dataset, 317 samples (Figure 15) belonging to the diabetes nephritis (DN), healthy living donor (HLD), immunoglobulin A nephropathy (IgAN), and lupus nephritis (LN) groups were included (191 samples were filtered out). Expression and patient-related data are included in supplementary files 1 and 2.

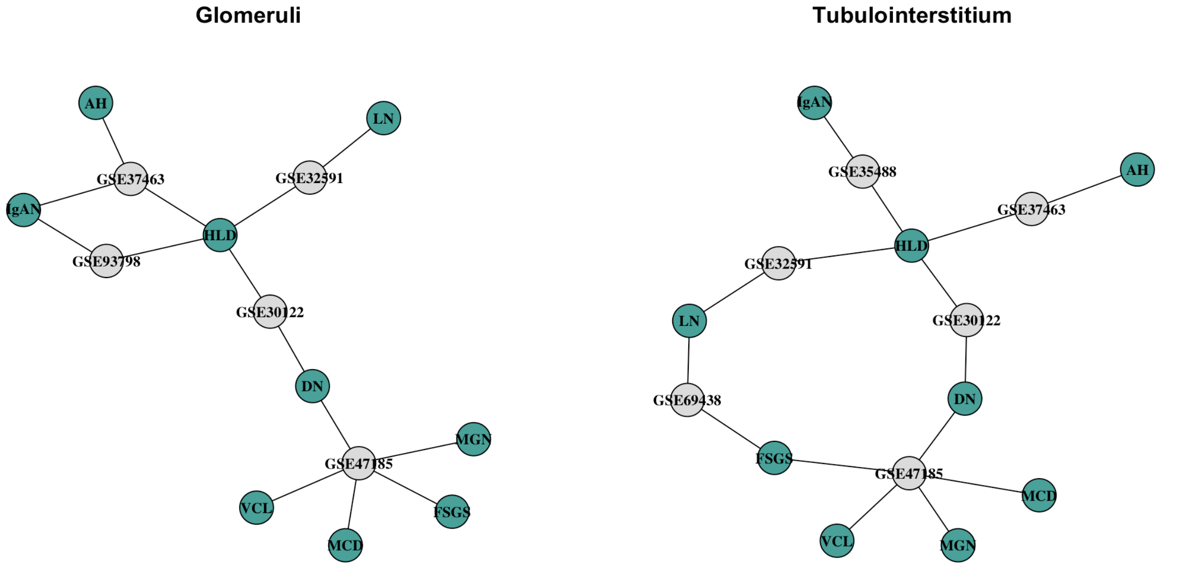


*Figure 13 – **Bipartite graph representing GSE-Disease pairs for each tissue (left: Glomeruli, right: Tubulointerstitium).** All pairs are connected to the main graph meaning that all the GSEs contain at least one disease group in common with another GSE.*

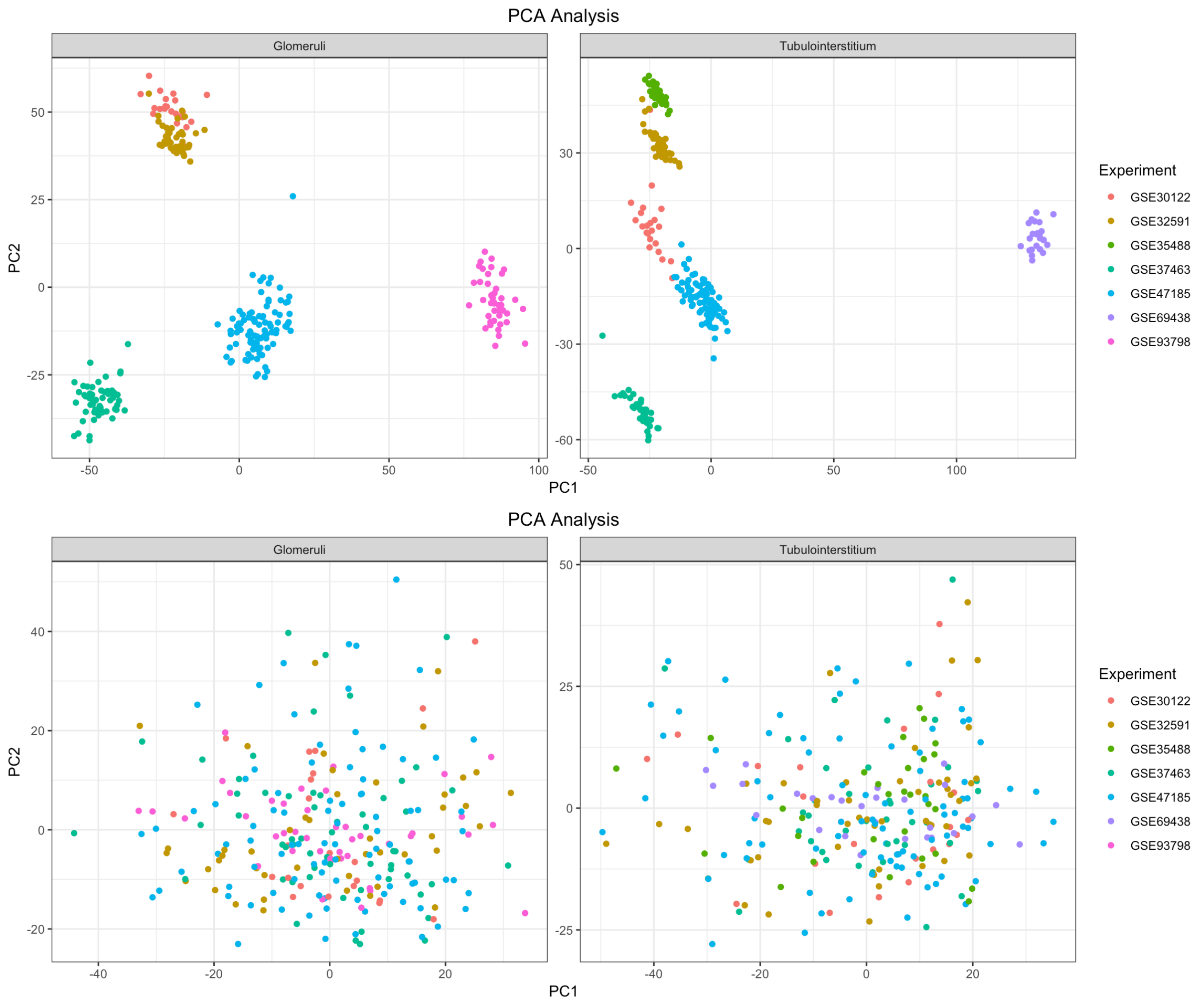


***Figure 14 - Batch effect correction (inter-experiment)**. All datasets are integrated and represented by a projection of their samples in the first two components after Principal Component Analysis (PCA). A) Before applying ComBat, B) After applying ComBat. Evaluation of batch existence was assessed by visual inspection. GSE series was identified as the only remaining artifact introducing batch effect.*

***Table 8 – PCA variance.** Variance explained for PC1 and PC2 in all subplots of Figure 14.*

Tissue	Before		After	
	PC1 (%)	PC2 (%)	PC1 (%)	PC2 (%)
Tubulointerstitium	41,29	18,69	21,32	11,84
Glomeruli	36,37	17,14	12,07	10,58

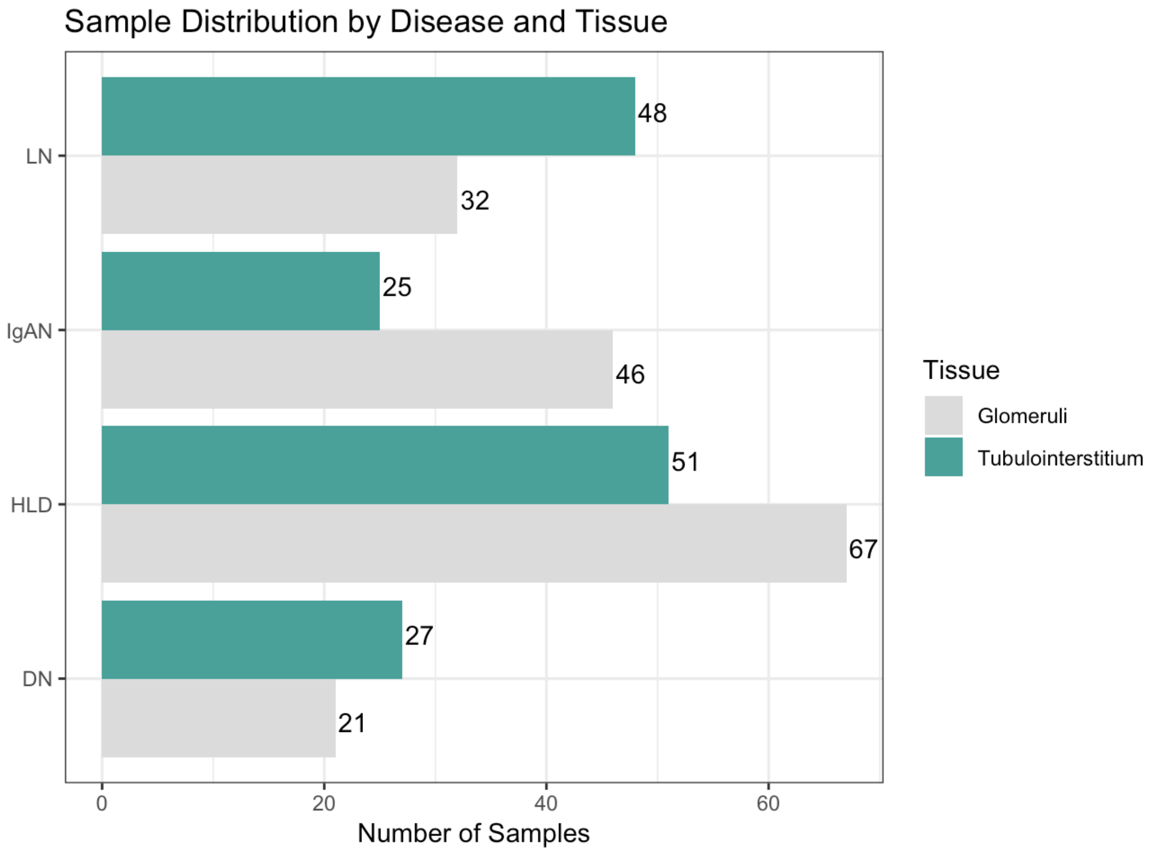


Figure 15 - Sample distribution by disease group and tissue. *DN stands for diabetic nephritis, HLD for healthy living donor, IgAN for immunoglobulin A nephropathy, and LN for lupus nephritis.*

Generation of Molecular Profile

Pathway Database (MSigDB)

The MSigDB C2 CP version 6.2 dataset originally included 1329 gene sets mapping to 8904 HGNC IDs connected by 67622 interactions (50.88 genes per pathway, 7.59 pathways per gene). Classification by gene-set size yielded a threshold of 75 genes per set, thus, only those sets with less than 75 were included in the final model. This resulted into a final model of 6613 genes mapped to 1124 terms by 32853 connections (29.22 genes per pathway, 4.96 pathways per gene).

Protein-Protein-Interaction Database (OmniPath)

The original OmniPath database included 7990 nodes and 49070 edges (6.14 edges per node). After removing inconsistencies, the reduced network contained 5927 nodes connected by 31042 edges (5.24 edges per node).

Gene-Level Statistics (GLS)

Differential expression analysis results are typically represented as volcano plots (Figure 16) where the X axes represents the log2 of the fold change between the case and the control and the Y axes represents the -log10 of the p-value of the Limma model (essentially ANOVA). The complete list of GLSs is included in supplementary file 3.

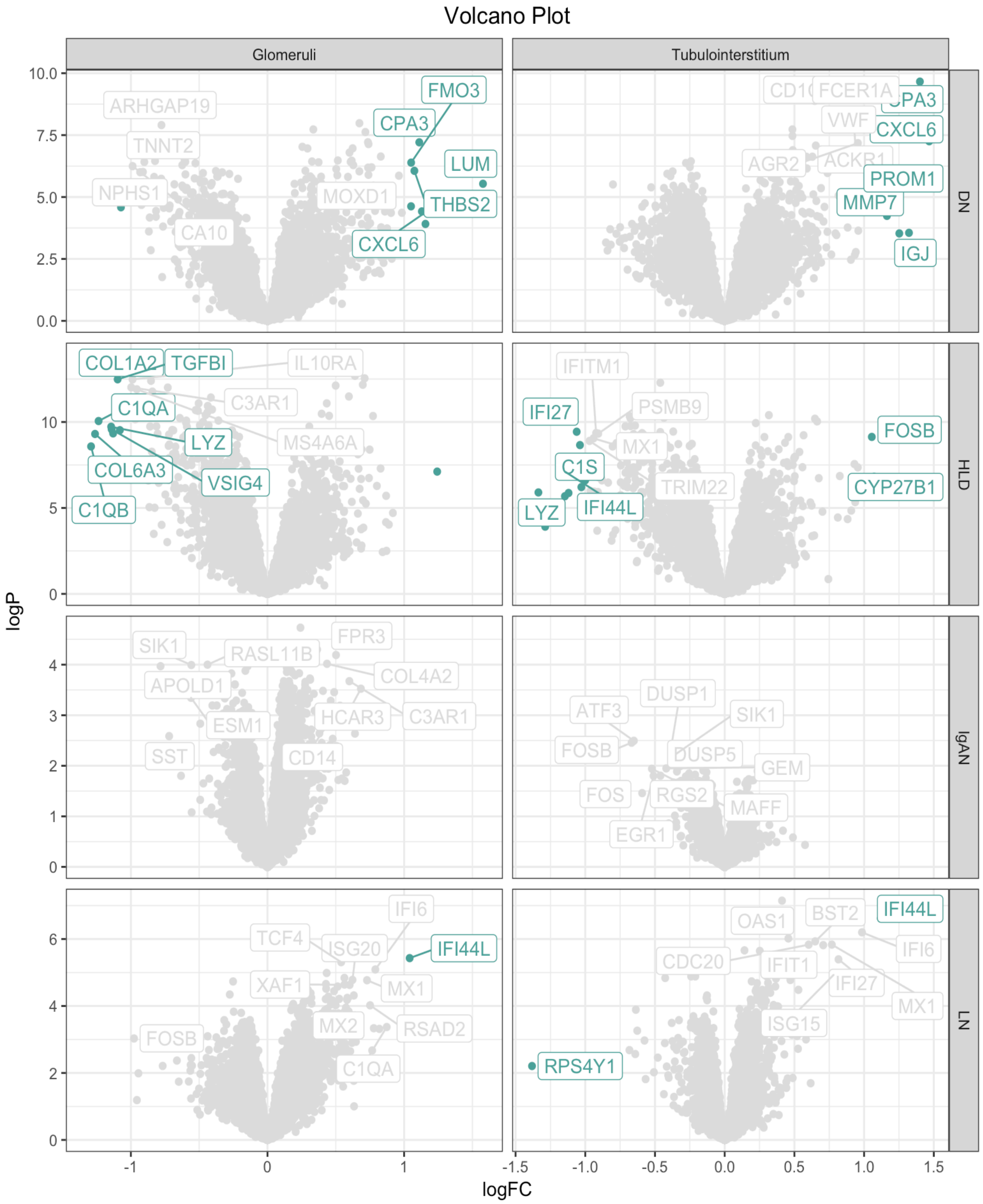


__Figure 16 - Differential expression analysis results.__ Each subplot represents a volcano plot where the x axes represents the log2 of the fold change (logFC) and the y axes the -log10 of the p-value (logP) calculated as described in the methods. A gene is considered to be differentially expressed if the absolute logFC is larger than 1 and the p-value is smaller than 0.05 (logP > 1.3). Differentially expressed genes in each case are represented in green.

Pathway-Level Statistics (PLS)

Gene-Set Analysis: Most affected pathways for each tissue and contrast are summarized in Figure 17 (for the complete list, see supplementary file 4).

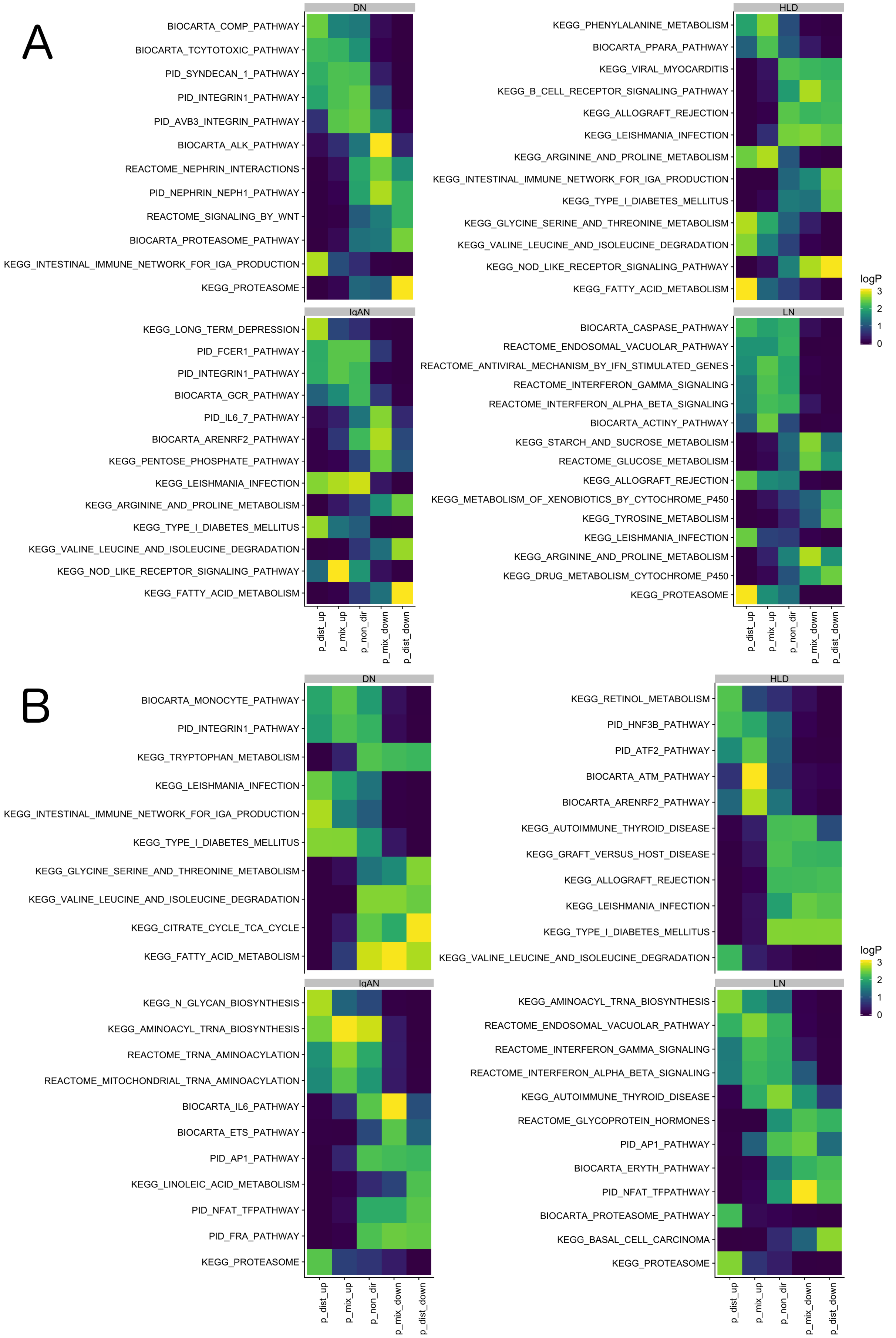


***Figure 17 - Most affected pathways**. A) Glomeruli, B) Tubulointerstitium. For each disease group, the three most affected pathways are shown for each p-value type.*

Map pathways to genes: Most affected genes in the pathway level are summarized in Figure 18.A (for the complete list, see supplementary file 5). Figure 18.B shows an example for the pathways where HLA-C participates using tubulointerstitial and DN contrast data. Figure 18.C shows a volcano plot for one of its pathways (KEGG_TYPE_I_DIABETES_MELLITUS) for the same data.

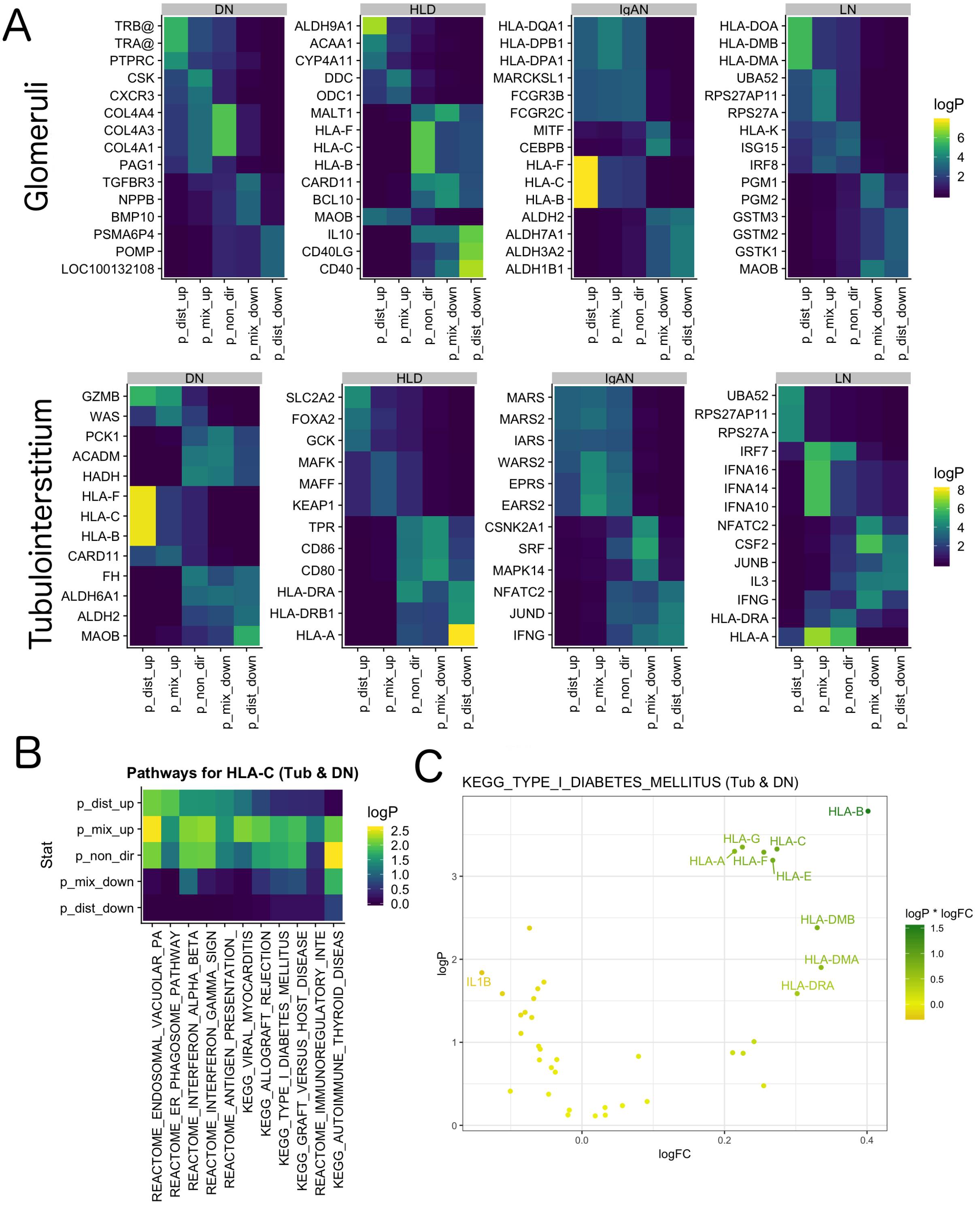


Figure 18 - Pathway-Level Statistic (PLS) results. A) Most affected genes. For each tissue and disease group, the three most affected genes are shown for each p-value type. B) Heatmap representing the set of 5 p-values for all the pathways where HLA-C participates using tubulointerstitial data for DN contrast. C) Volcano plot for KEGG_TYPE_I_DIABETES_MELLITUS pathway (one of the pathways where HLA-C participates) using tubulointerstitial data for DN contrast.

Network-Level Statistics (NLS)

The most affected genes in the network level as described in the methods are summarized in Figure 19 (for the complete list, see supplementary file 6). Figure 20 shows the Local environment (nodes in distance 1) of the 5 most affected genes (AIM2, C1QC, NLRC4, and NLRP2) for glomerular tissue and HLD contrast.

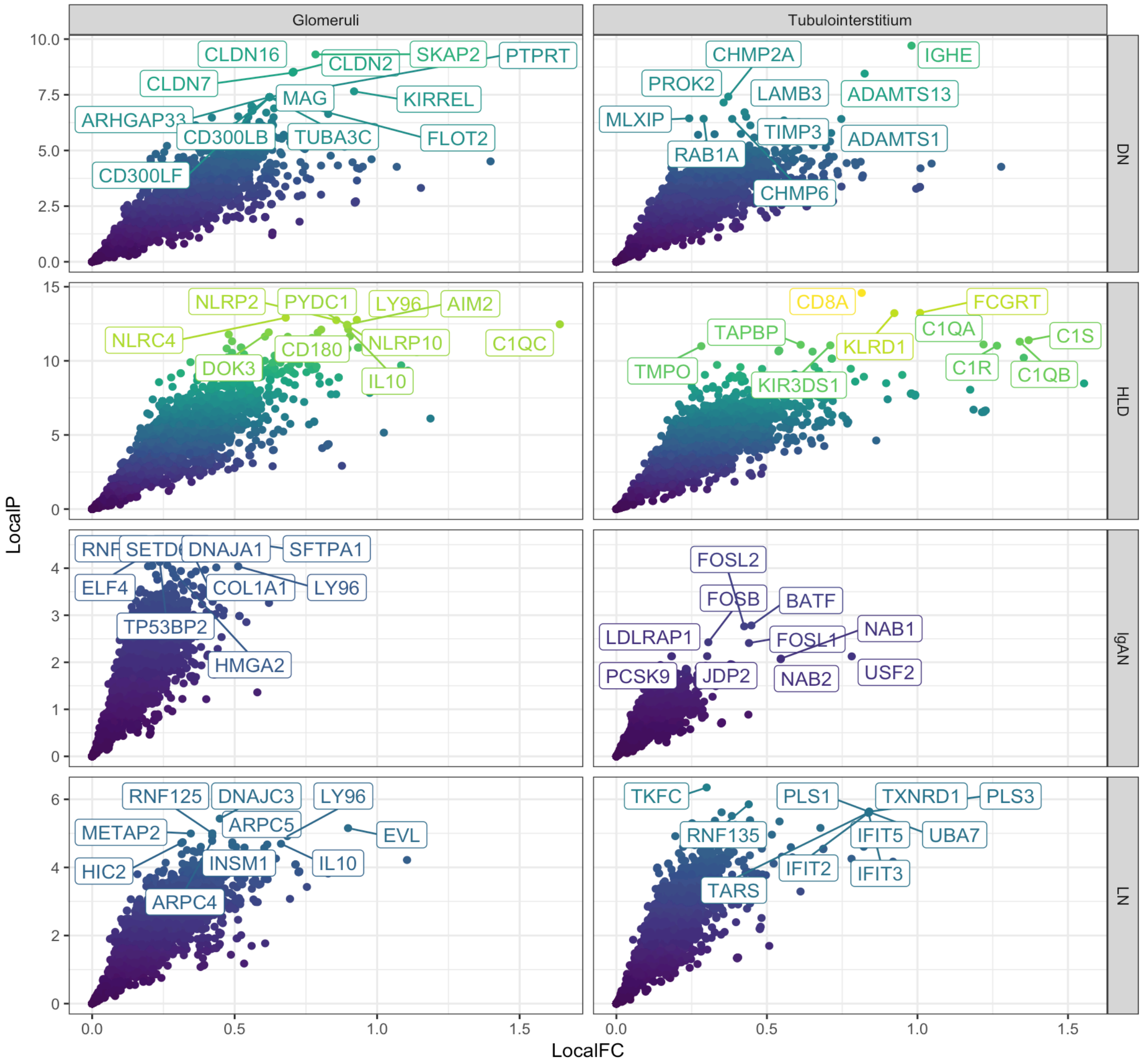


__Figure 19 - Network-Level Statistic (NLS) results.__ A) Most affected genes. Each tissue and disease group is represented by a volcano plot of the obtained scores. Each subplot contains labels for the 10 genes further from the origin (0,0 point) in each case.

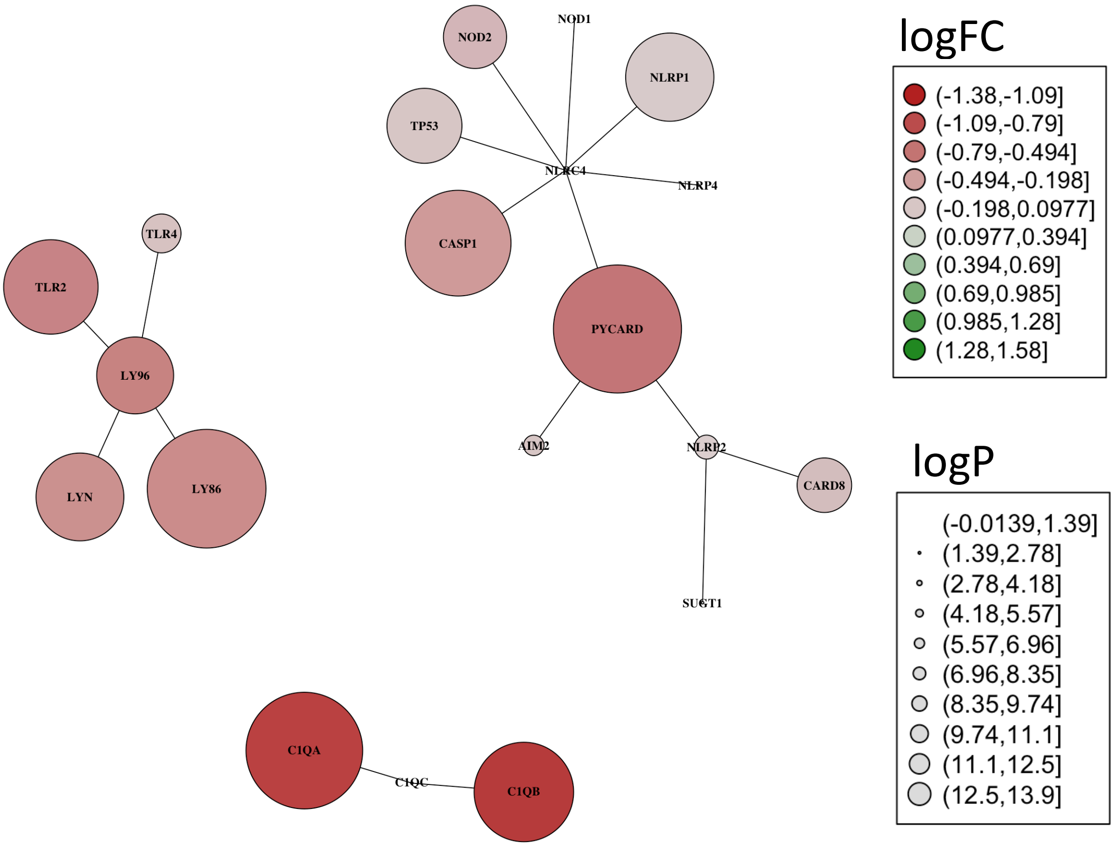


Figure 20 - Glomeruli HLD local environment. *The 5 most affected genes from NLS are projected together with their adjacent nodes based on the PPI network. These nodes cluster in 3 different affected areas. Nodes are colored by the logFC from differential expression analysis. Node size correlates with the logP from differential expression analysis.*

Map to Protein Identifiers

The 14350 unique HGNC symbols included in the union of all three models (GLS, PLS, NLS) were mapped to 13743 unique reviewed Uniprot IDs (UNIPROTKBs).

Feature Integration

Figure 21 shows the Venn diagram of the different UNIPROTKBs included in each model (GLS: 12279, PLS: 6551, and NLS: 6119). The 3587 proteins included in all three models (triple intersection) were used for further analysis.

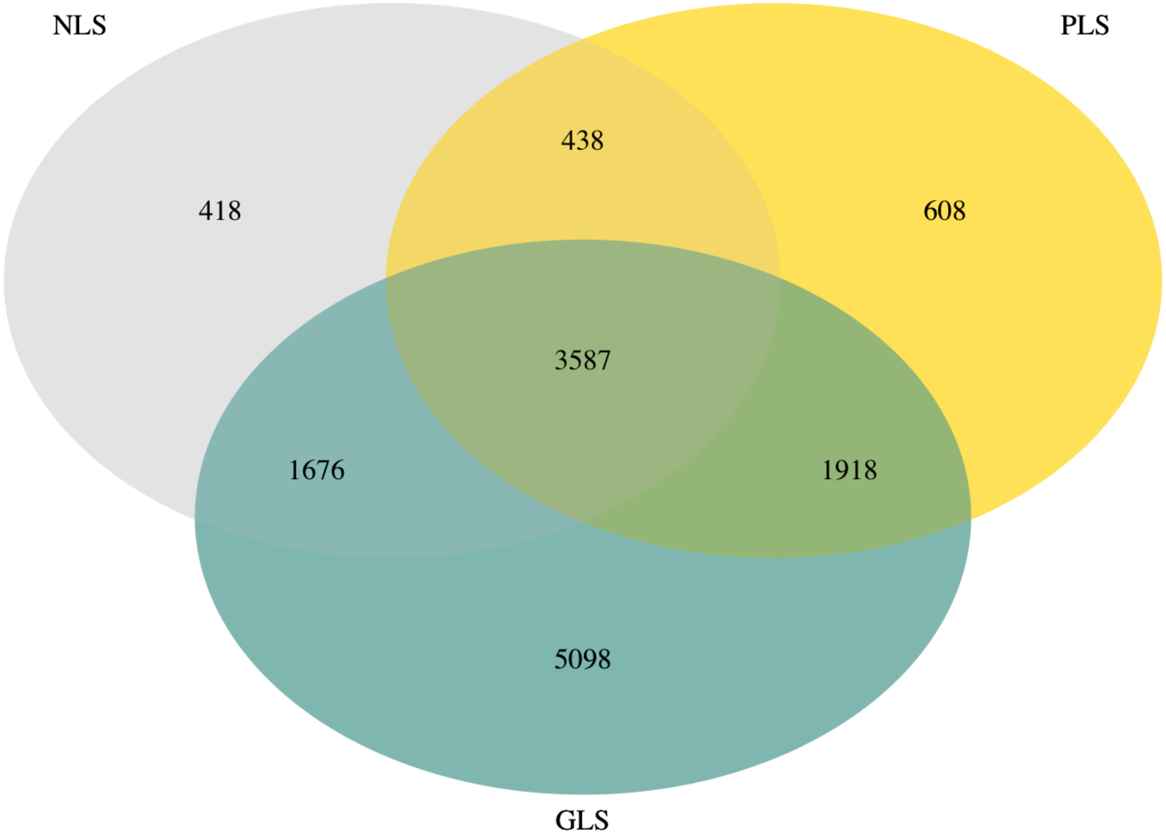


Figure 21 - Venn diagram. The number of different proteins included in each model is represented as a finite set. Only those proteins included in the triple intersection were kept for further analysis.

Feature Selection

Table 9 shows the coefficient of variation (CV) for each feature, tissue, and disease group. None of the coefficients is below the 40% threshold of exclusion, thus all of them were kept in the matrix.

***Table 9 - Feature selection.** Estimation of predictor variance. For each tissue and disease group the coefficient of variation (CV) of each feature is calculated across all proteins.*

Tissue	*Feature*	*DN*	*HLD*	*IgAN*	*LN*
Glomeruli	gls_logFC	8580,76	775,57	598,47	1163,25
Glomeruli	gls_logP	101,67	119,44	95,64	86,14
Glomeruli	nls_localFC	48,67	55,28	46,53	46,32
Glomeruli	nls_localP	47,45	52,76	45,00	41,02
Glomeruli	nls_logC	51,97	51,97	51,97	51,97
Glomeruli	pls_logN	81,00	81,00	81,00	81,00
Glomeruli	pls_p_dist_down	89,14	111,92	107,96	106,48
Glomeruli	pls_p_dist_up	111,10	104,33	172,49	95,16
Glomeruli	pls_p_mix_down	88,13	102,40	97,51	95,31
Glomeruli	pls_p_mix_up	98,04	96,31	99,66	95,87
Glomeruli	pls_p_non_dir	91,58	140,76	90,95	89,06
Tubulointerstitium	gls_logFC	972,47	708,70	1278,43	926,56
Tubulointerstitium	gls_logP	96,54	108,80	92,60	88,81
Tubulointerstitium	nls_localFC	50,74	53,01	47,51	47,89

Tubulointerstitium	nls_localP	44,23	45,16	42,98	43,23
Tubulointerstitium	nls_logC	51,97	51,97	51,97	51,97
Tubulointerstitium	pls_logN	81,00	81,00	81,00	81,00
Tubulointerstitium	pls_p_dist_down	106,87	132,41	103,49	109,27
Tubulointerstitium	pls_p_dist_up	169,58	110,99	93,62	101,14
Tubulointerstitium	pls_p_mix_down	109,20	100,72	100,22	105,45
Tubulointerstitium	pls_p_mix_up	102,47	94,63	87,64	136,58
Tubulointerstitium	pls_p_non_dir	95,43	99,53	84,38	118,19

Figure 22 shows the correlation matrix for the features across all tissues and disease groups. The largest correlation coefficient was between the network-level-statistics localFC and localP with R = 0.81, below the 0.85 threshold declared in the methods.

QR decomposition revealed that none of the features can be explained as a linear combination of the others.

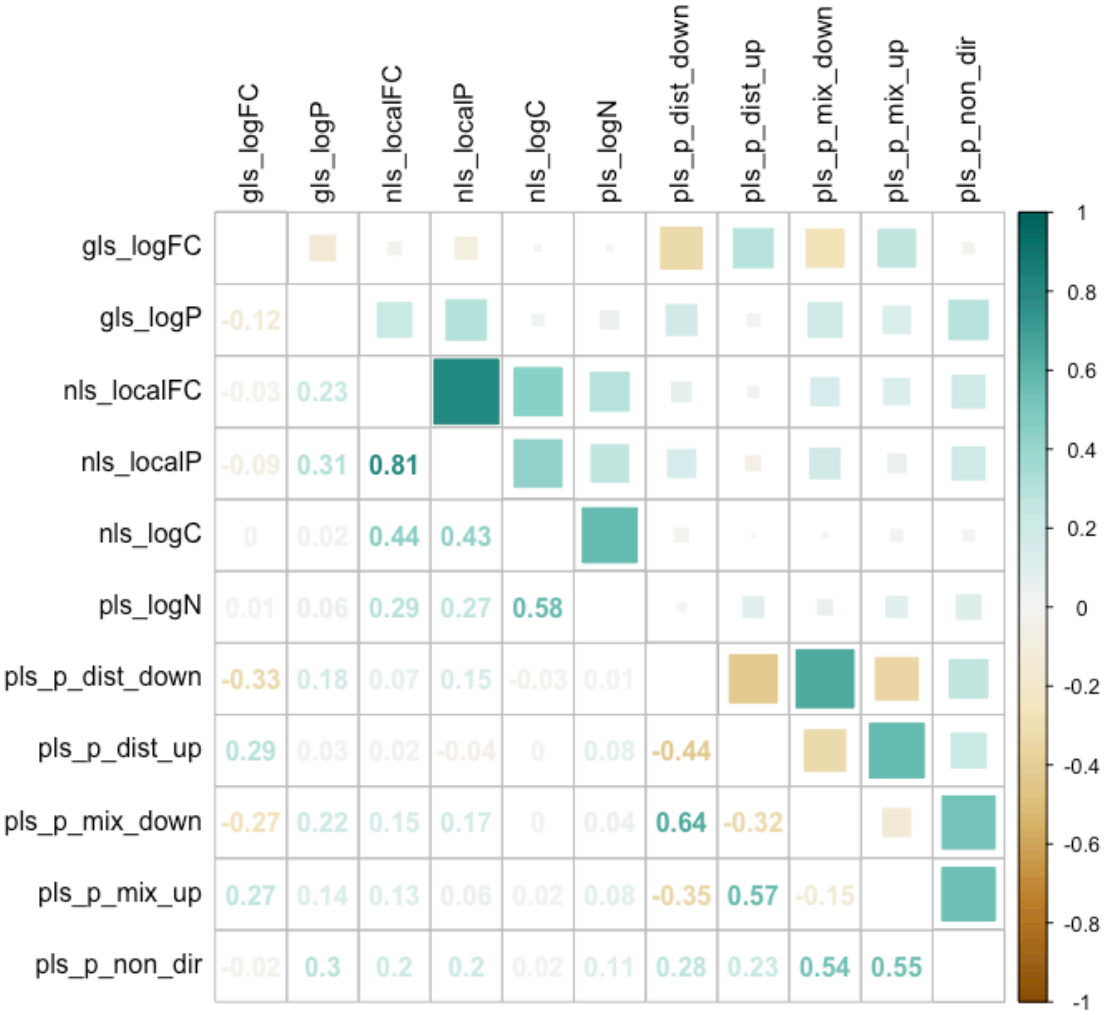


Figure 22 - Correlation matrix. *Pearson correlation coefficient was calculated between each pair of features across all features and disease groups.*

Secretion Prediction

The downloaded dataset contained rich information for 20183 proteins including, when available, the subcellular location and presence of signal sequences. For training, 3950 proteins were removed due to unknown subcellular location. In addition, 3959 proteins that had supporting evidence of being situationally intracellular or extracellular were removed leaving 12274 proteins for training and testing.

Following cross validation and parameter tuning the final model exhibited a Matthews correlation coefficient of 0.849 outperforming by 0.09 points other majority vote-based methods. Regarding protein secretion alone it also outperformed by 0.15 points other machine learning methods which predict the subcellular location of a protein given its amino acid sequence.

Experimental results

Sample acquisition

Baseline patient characteristics are specified in Table 10 (for patient specific information, see supplementary file 7).

Table 10 - Sample acquisition. *80 samples belonging to 3 disease groups and control were included in the study.*

Group	**DN**	**HLD**	**IgAN**	**LN**
Number of patients (N)	20	19	19	21
Age (years)	51.2±12.3	50.0±19.2	47.9±12.4	46.4±17.3
Female (%)	50%	45%	52.6%	66.7%
eGFR (ml/min/1.73m^3)	46.5±26.6	NA	56.3±33.2	58.7±34.6

Proteomic results – training set

The raw Mean Fluorescent Intensities (MFIs) representing the proteomic measurements for the training set are contained in supplementary file 8. 425 replicate measurements were removed based on QC (14.93% of the total measurements), 5 (0.18%) due to bead counts, 299 (10.50%) due to replicate QC and 121 (4.25%) due to analyte distribution QC. Figure 23 shows the number of remaining replicates for each sample/analyte pair before median centering.

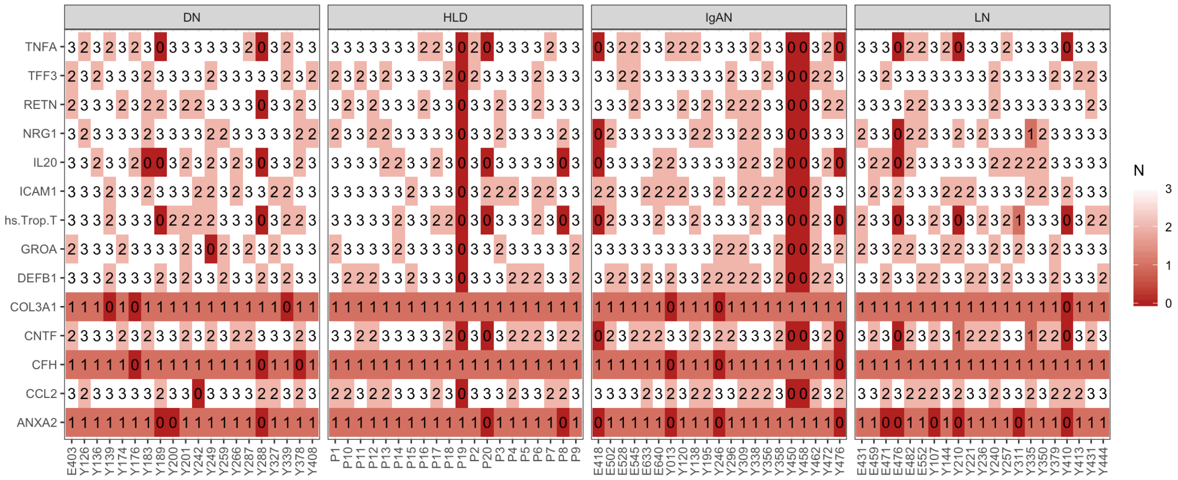


Figure 23 - Number of replicates per sample and analyte pair – training set. *425 replicates were removed due to quality control checks.*

10 patients (E418, E476, P19, P20, Y189, Y288, Y410, Y450, Y458, and Y476) were removed due to loss of information (more than 1 NA). Table 11 shows the resulting p-values from background evaluation after running one-tailed t-test and FDR for multiple comparisons. CNTF, ANXA2, and hs.Trop.T did not show significant differences from background measurements and were removed from further analysis.

Table 11 - Background Evaluation – training set. *Sample MFIs were compared to blank/background measurements using one-tailed t-tests. P-values were adjusted for multiple comparisons using false-discovery-rate (FDR) method.*

Analyte	**p.value**	**p.adj**
CNTF	8,89E-01	8,89E-01
ANXA2	1,40E-01	1,51E-01
hs.Trop.T	4,51E-02	5,26E-02
IL20	1,85E-04	2,36E-04
NRG1	4,20E-05	5,88E-05
TNFA	2,07E-07	3,21E-07
TFF3	4,60E-17	8,05E-17
RETN	2,21E-17	4,41E-17
GROA	6,90E-18	1,61E-17
CFH	5,08E-18	1,42E-17
CCL2	1,14E-26	4,01E-26
COL3A1	2,33E-30	1,09E-29
ICAM1	9,28E-48	6,50E-47
DEFB1	1,58E-49	2,22E-48

Exploratory analysis: Results from PCA, tSNE, and HC are summarized in Figure 24-C. Figure 24.D shows violin plots for all the analytes colored by disease group.

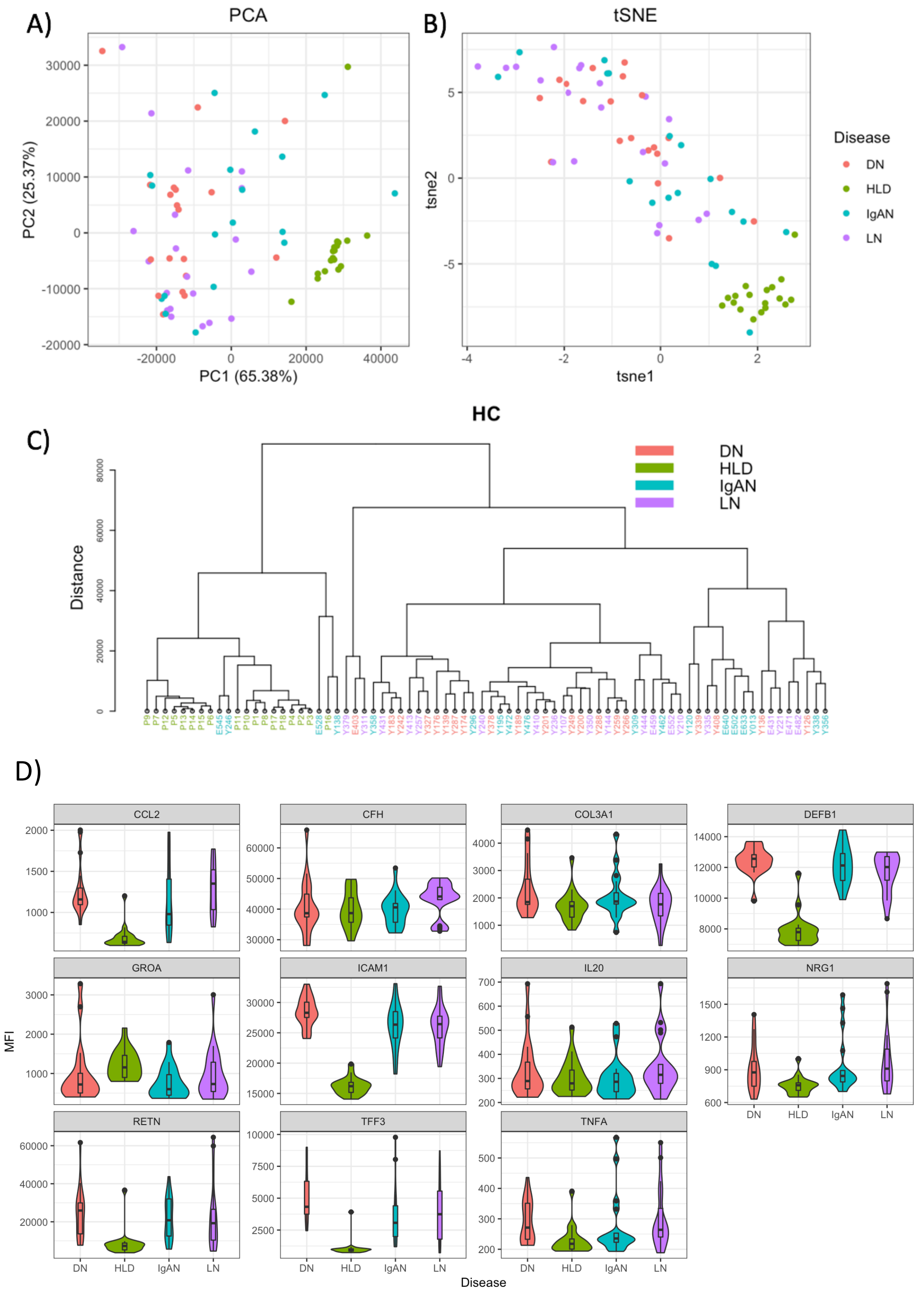


Figure 24 - Exploratory Analysis – training set. A) Principal Component Analysis (PCA), B) t-stochastic neighbor embedding (t-SNE), C) Hierarchical Clustering (HC), D) violin plots.

Candidate Biomarker Performance: Figure 25 summarizes results from both quantitative and qualitative analyses (for a complete list, see supplementary file 9).

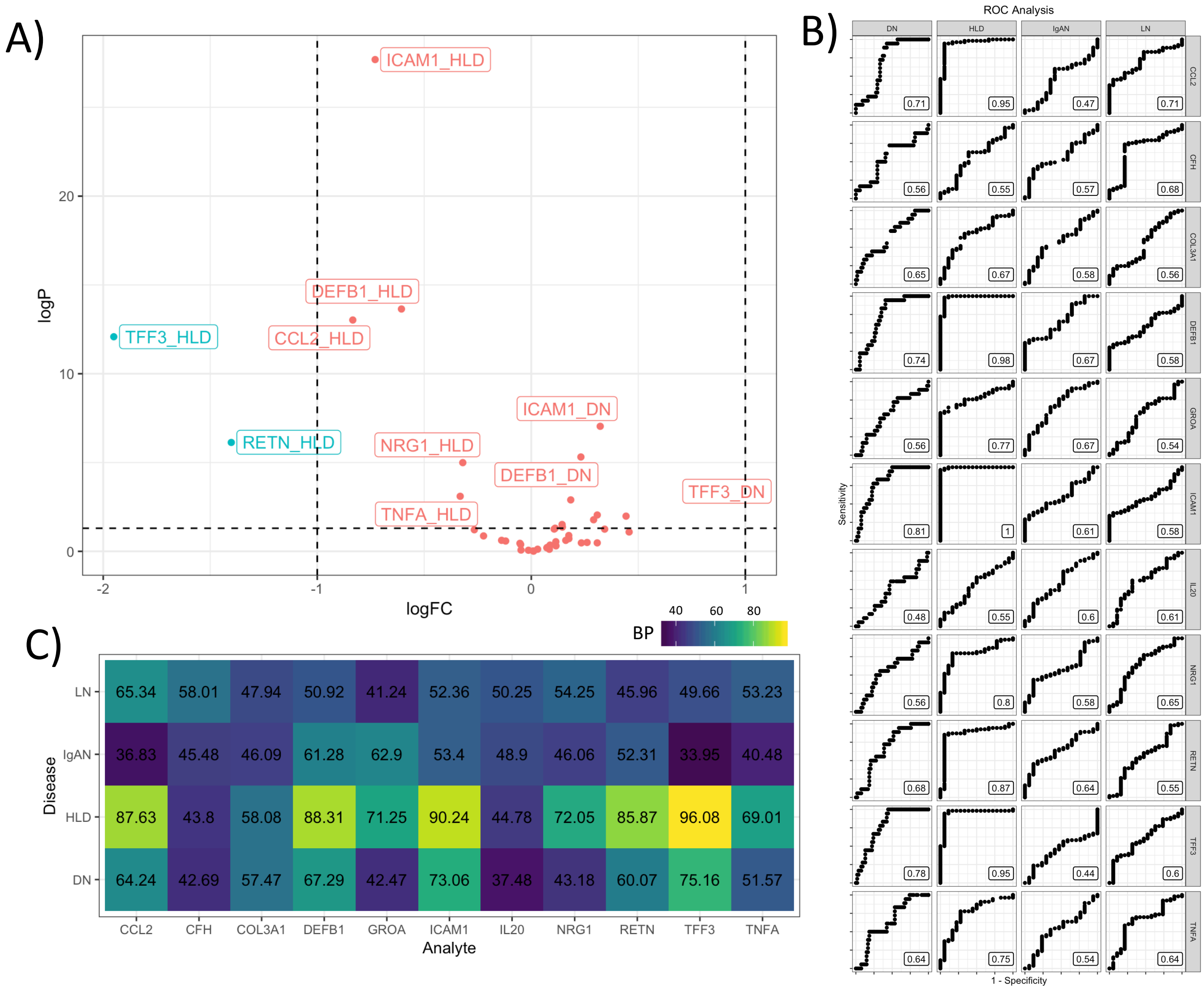


Figure 25 – Candidate biomarker performance – training set. *A) Quantitative performance. MFIs were compared in a contrast-based manner similarly to differential expression analysis of genes. B) Qualitative results. Area under the ROC curve (AUC) was used as performance metric. C) Biomarker performance (BP). BP was calculated integrating quantitative and qualitative metrics as described in the methods.*

Model training

Correlation between the molecular features defined in the molecular profile and biomarker performance are summarized in Figure 26. Biomarker performance was predicted using elastic networks. Figure 27 shows the obtained weights from the elastic network (lambda min) in the cross-validated model. These weights were used to predict the biomarker performance of all the proteins in our model (feature matrix). Glomerular and Tubular correlations showed a Pearson correlation coefficient of 0.62 (RMSE = 5.6, p-value < 2e-16) (Figure 28). The final predicted biomarker performance was obtained by averaging the predictions from the glomerular (Glom) and tubular (Tub) models. These predictions were used to compare the obtained biomarker performance in the measured proteins with the predicted one (Figure 29). The obtained root-mean-square error (RMSE) was equal to 12.85 while the mean absolute error (MAE) was equal to 10.08.

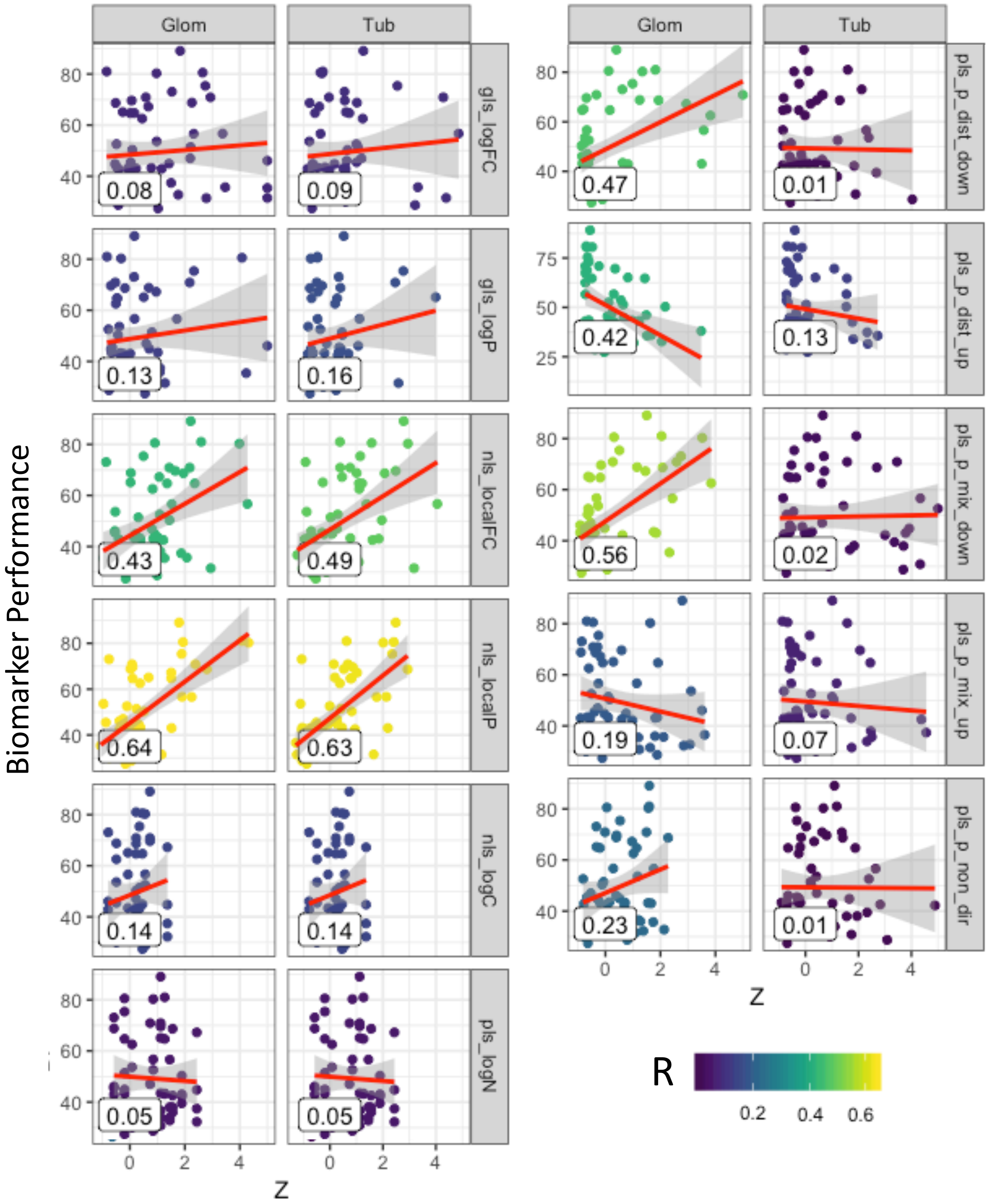


__Figure 26 - Biomarker performance (BP) correlation with molecular features.__ The label inside and the color for each subplot represents the Pearson correlation coefficient between the feature and biomarker performance.

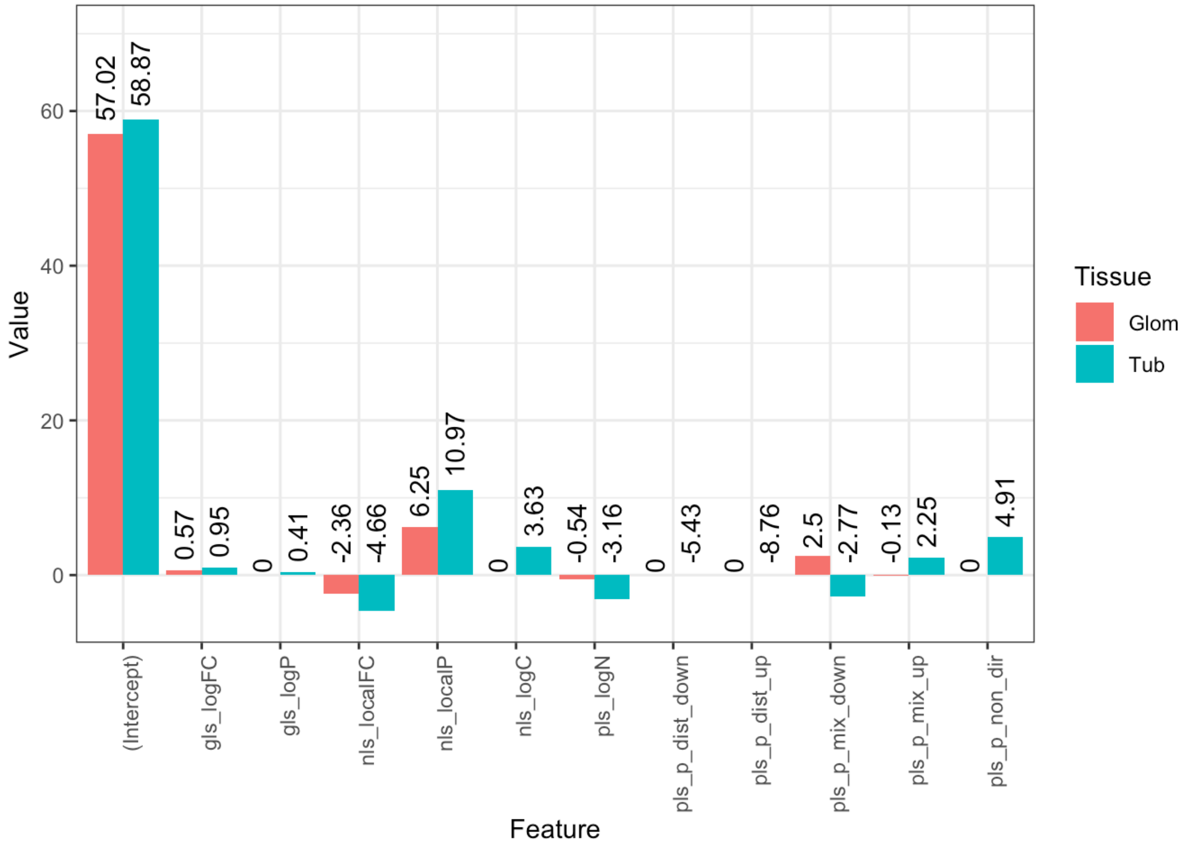


Figure 27 - Elastic network, cross validated weights. *Each weight is represented by the mean and the standard deviation across all folds.*

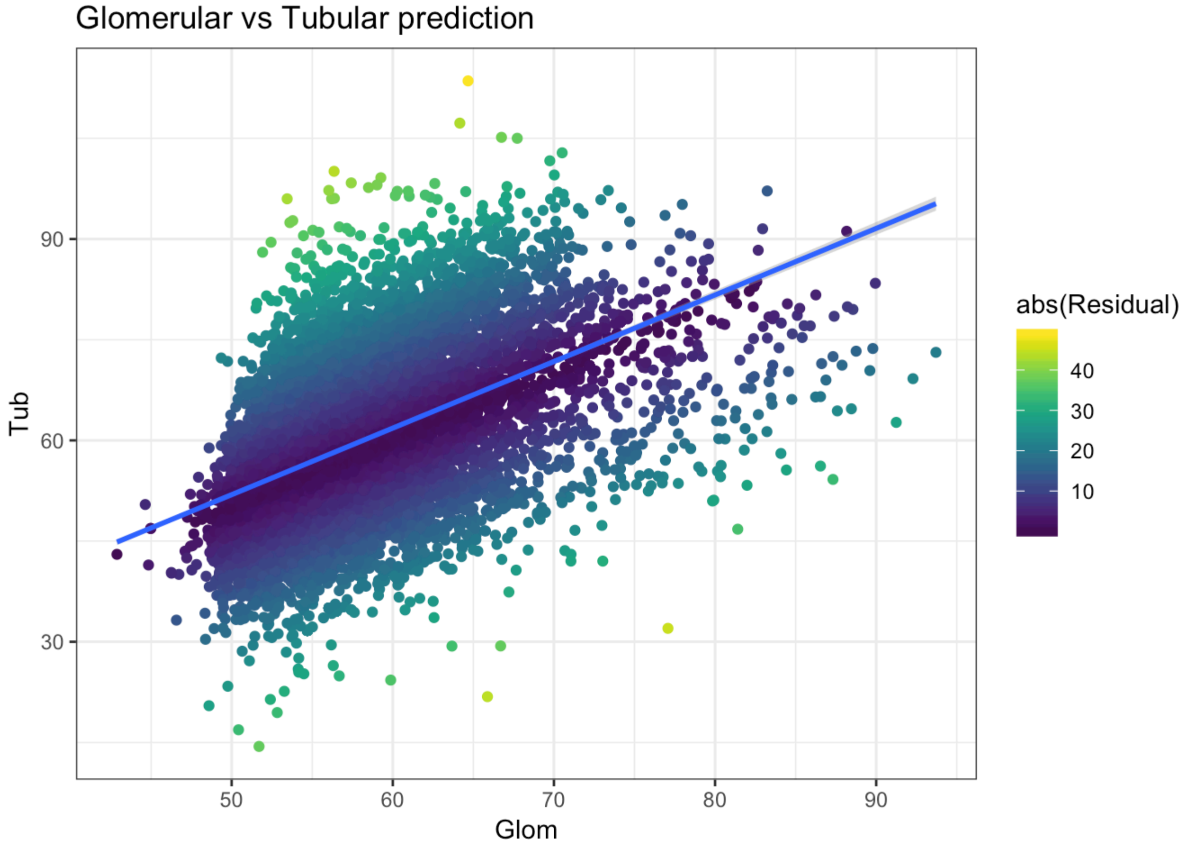


Figure 28 - Glomerular versus tubular prediction. *X-axes represents the predicted biomarker performance in glomerular tissue (Y-axes, tubulointerstitial). The blue line represents the linear model obtained from both predictions. Correlation analysis showed a Pearson correlation coefficient of 0.62 (RSE = 5.6, p-value < 2e-16).*

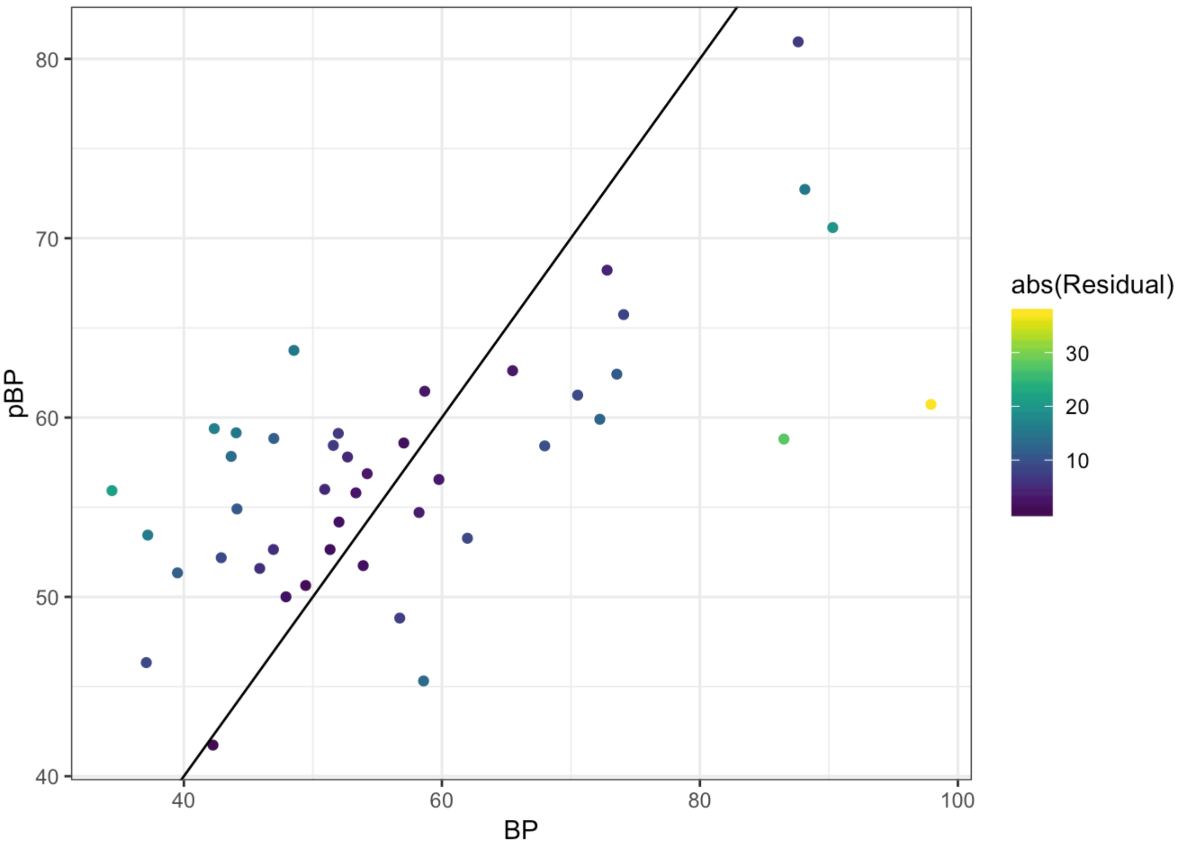


Figure 29 - Training set performance. Biomarker performance from the proteomic measurements (BP) was compared against the predicted biomarker performance (pBP).

Proteomic results – validation set

The raw Mean Fluorescent Intensities (MFIs) representing the proteomic measurements for the validation set are contained in supplementary file 10. 367 sample measurements were removed based on QC (17.95% of the total measurements), 1 (0.05%) due to bead counts, 227 (10.99%) due to replicate QC, and 140 (6.81%) due to analyte distribution QC. Figure 30 shows the number of remaining replicates for each sample/analyte pair before median centering.

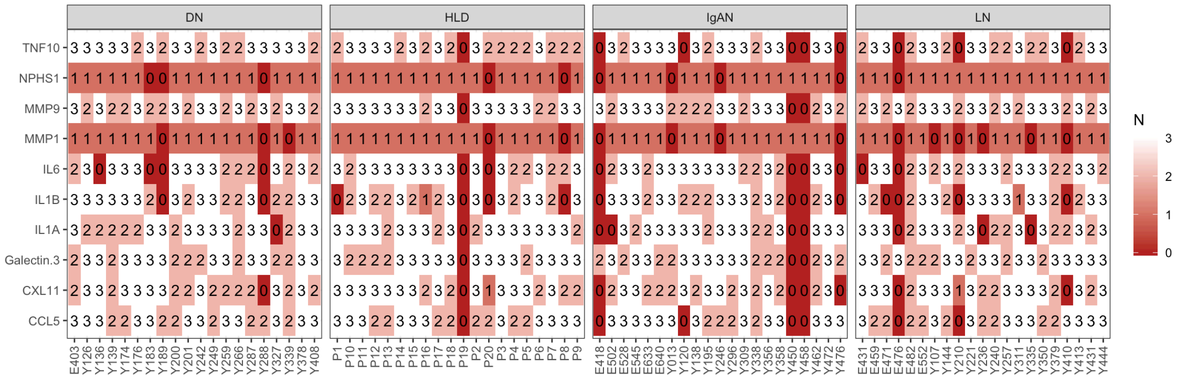


Figure 30 - Number of replicates per sample and analyte pair - validation set. 367 replicates were removed due to quality control checks.

11 patients (E418, E476, P19, P20, Y189, Y210, Y288, Y410, Y450, Y458, and Y476) were removed due to lack of information (more than 1 NA). From the 11 patients, all but Y210 were

removed in the training set for the same reason. Table 12 shows the resulting p-values from background evaluation after running one-tailed t-tests and FDR for multiple comparisons. IL1B, IL1A, and MMP1 did not show significant differences when compared to background measurements thus were removed from further analysis.

Table 12 - Background Evaluation – validation set. *Sample MFIs were compared to blank/background measurements using one-tailed t-tests. P-values were adjusted for multiple comparisons using false-discovery-rate (FDR) method.*

Analyte	**p.value**	**p.adj**
IL1B	1,00E+00	1,00E+00
IL1A	9,77E-01	1,00E+00
MMP1	9,16E-01	1,00E+00
IL6	6,89E-04	9,84E-04
CXL11	5,44E-08	9,07E-08
CCL5	5,99E-10	1,20E-11
NPHS1	9,47E-12	2,37E-11
TNF10	6,38E-12	2,13E-11
MMP9	6,46E-19	3,23E-18
Galectin.3	4,77E-31	4,77E-30

Exploratory analysis: Results from PCA, tSNE, and HC are summarized in Figure 31.A-C. Figure 31.D shows violin plots for all the analytes colored by disease group.

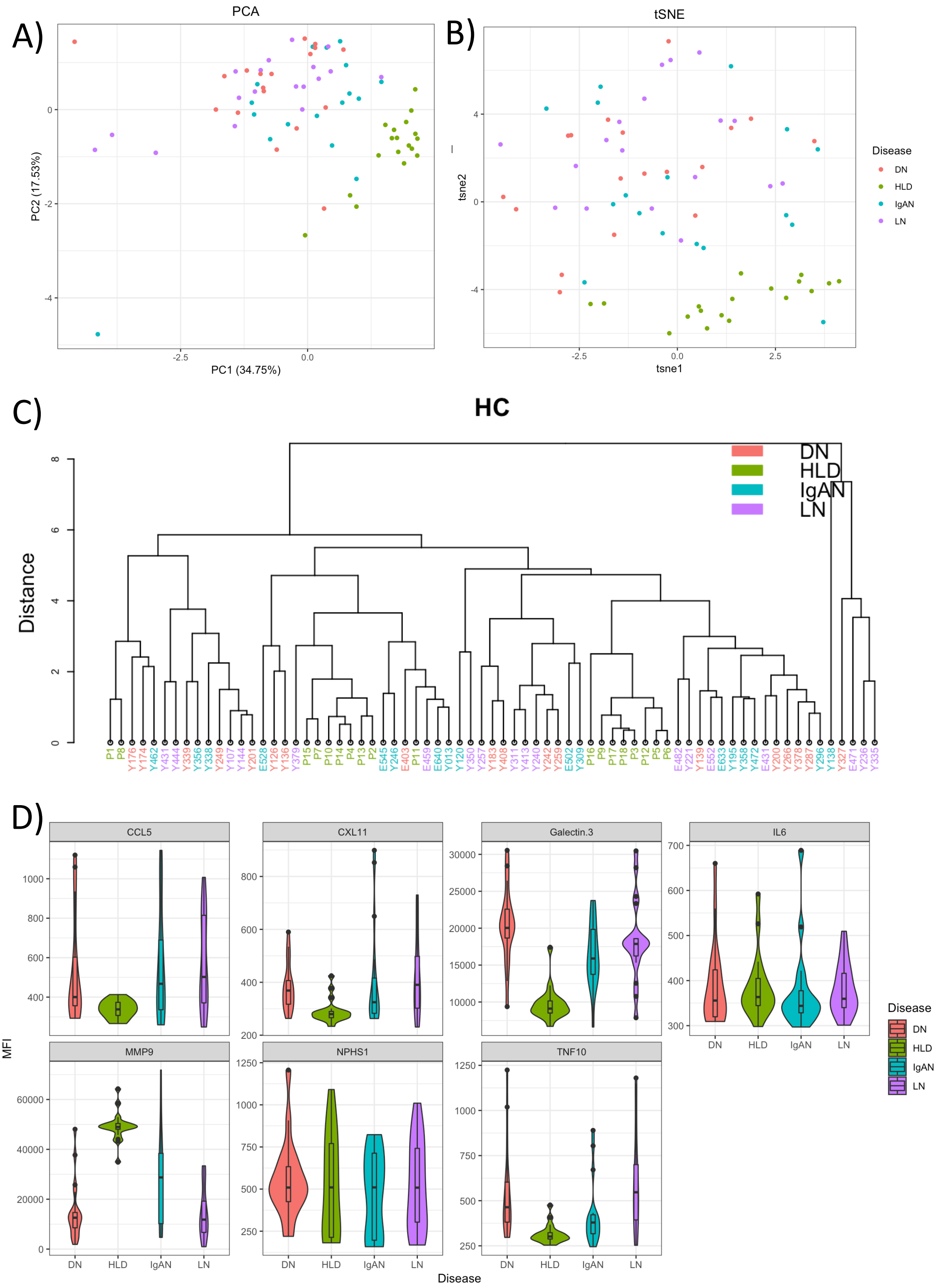


__Figure 31 - Exploratory Analysis – validation set.__ A) Principal Component Analysis (PCA), B) t-stochastic neighbor embedding (t-SNE), C) Hierarchical Clustering (HC), D) violin plots.

Candidate Biomarker Performance: Figure 32 summarizes results from both quantitative and qualitative analyses (for a complete list, see supplementary file 11).

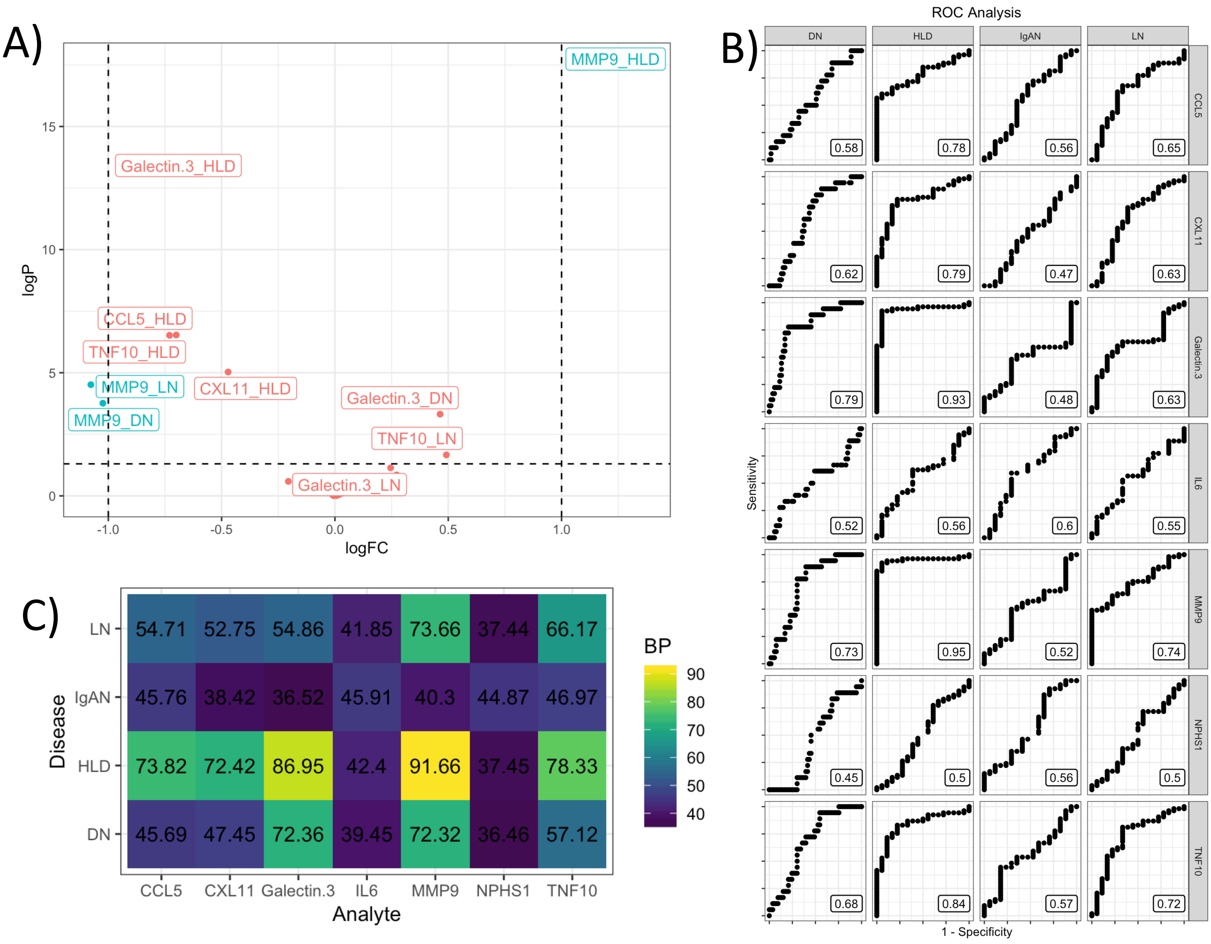


*Figure 32 – **Candidate biomarker performance – validation set.** A) Quantitative performance. MFIs were compared in a contrast-based manner similarly to differential expression analysis of genes. B) Qualitative results. Area under the ROC curve (AUC) was used as performance metric. C) Biomarker performance (BP). BP was calculated integrating quantitative and qualitative metrics as described in the methods.*

Model testing

For the model testing, we compared the measured biomarker performance of the proteins that were part of the validation set with their predicted counterparts (Figure 33). These candidates returned a root-mean-square error (RMSE) of 15.87 and a mean absolute error of 13.68. Figure 34 shows the comparative in absolute error distribution between training and test sets. We compared these distributions using a two-tailed t-test which yielded a p-value of 0.31, i.e., the residuals from the validation set were not significantly different to those of the training set.

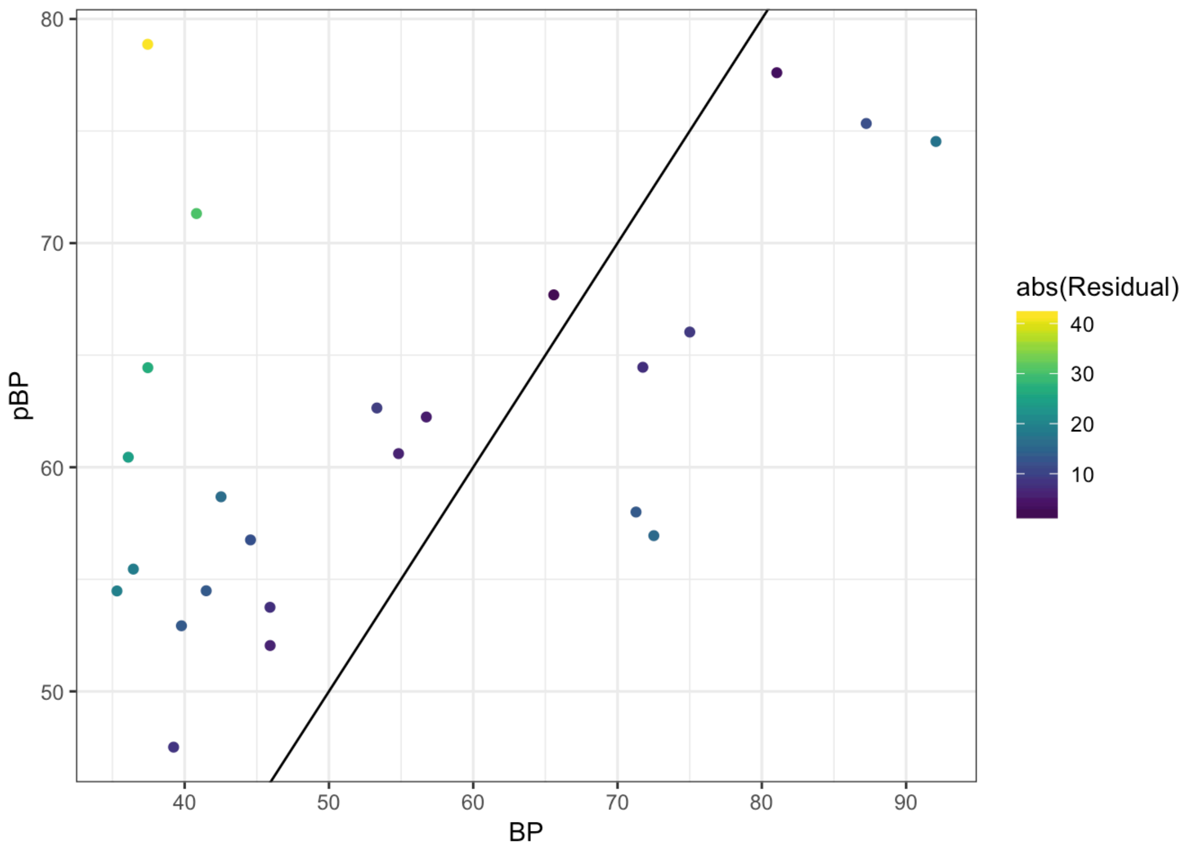


*Figure 33 - **Validation set performance.** Biomarker performance from the proteomic measurements (BP) was compared against the predicted biomarker performance (pBP).*

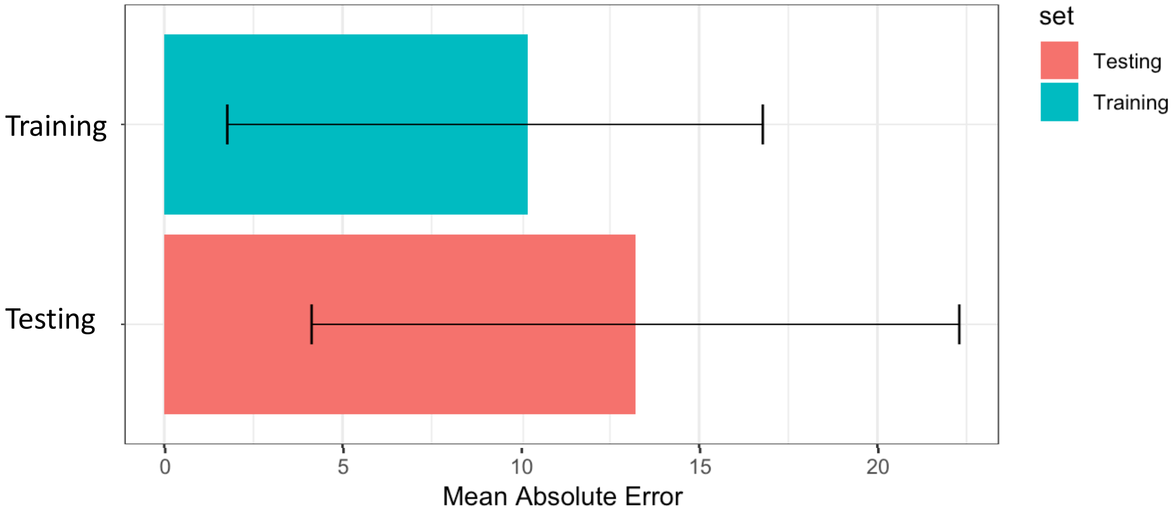


***Figure 34 - Mean absolute error comparative.** Training = 10.08 ± 6.60; test = 13.68.7 ± 9.2.*

Downstream analysis

Integration into biomarker panel

Combinations of candidate biomarkers into panels with increased performance were then evaluated. Figure 35 shows the correlation matrix between the measured analytes in both the training and validation sets. Figure 36 shows the resulting Pareto front from LDA using a 5-fold cross-validation scheme. The elbow criteria revealed that the optimal panel size was 3 (ACC = 0.69, BD = 1.07). The panel of size 3 was formed by the proteins: Galectin3, ICAM1, and

MMP9. Figure 37 shows the rotation vectors for LD1 and LD2, the confusion matrix, and the projected values of the samples in LD1 and LD2.

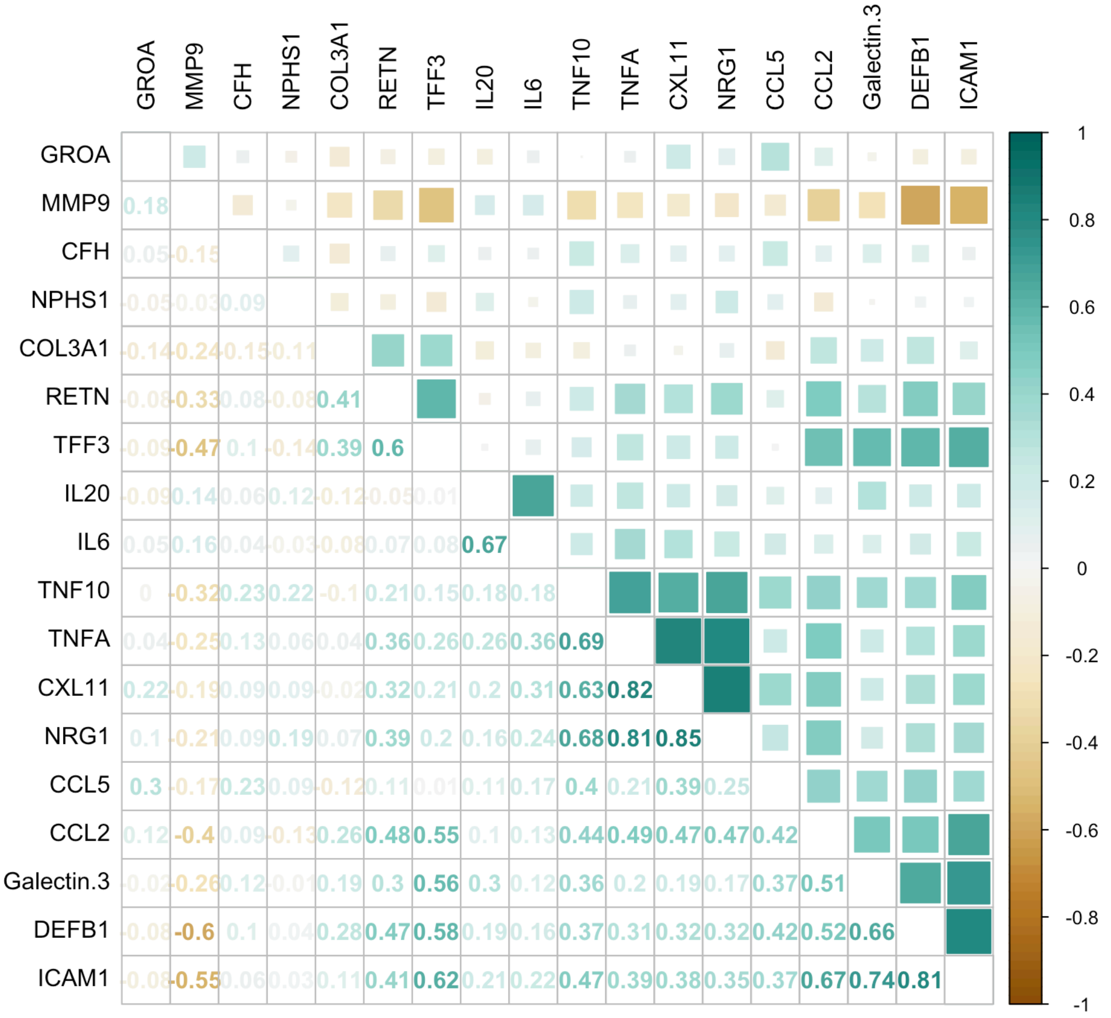


Figure 35 - Correlation matrix - analytes. *All measured proteins were pair-wise evaluated for correlation using Pearson's correlation coefficient.*

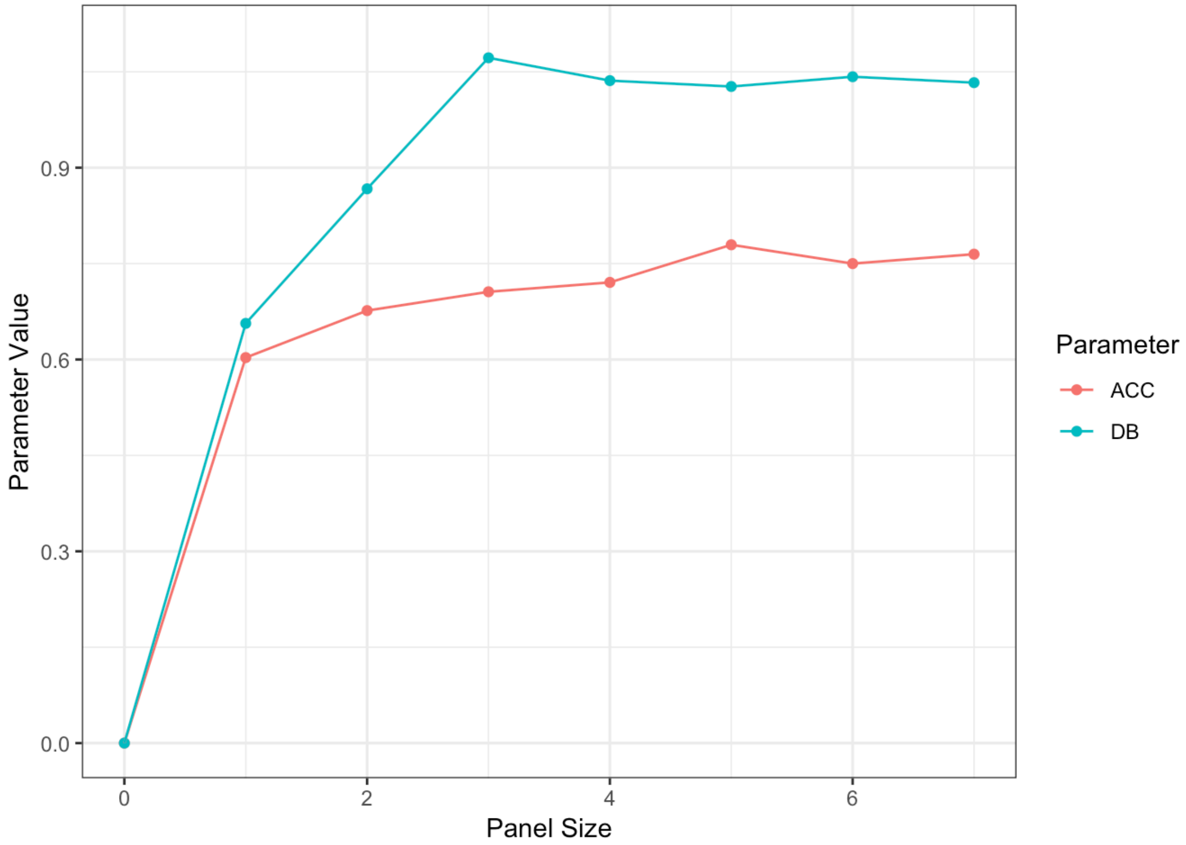
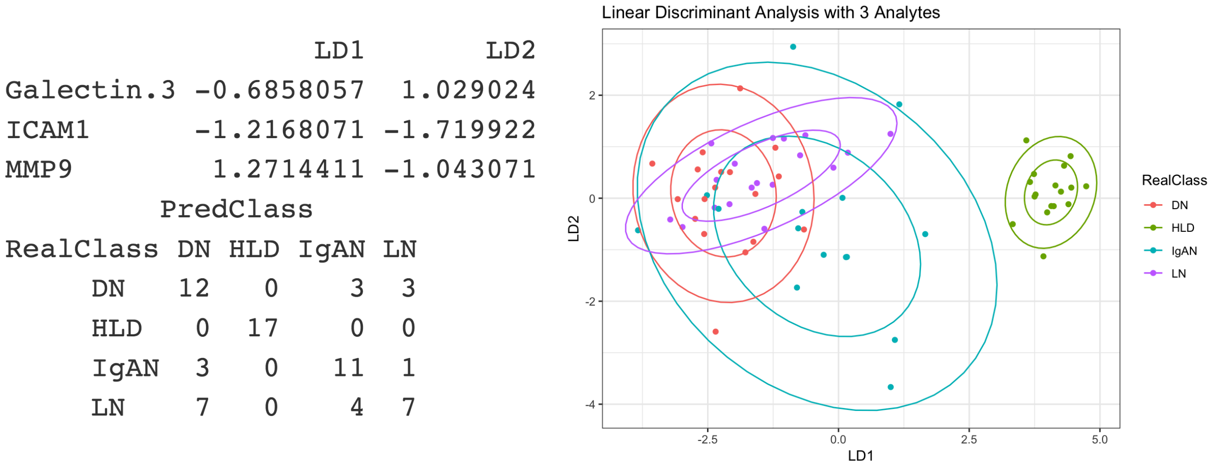


Figure 36 - Pareto front. *The Pareto front contains the optimal solution for the panel of each size (x-axes). The elbow criterion was applied for both quantitative (Bhattacharyya Distance, BD) and qualitative (Accuracy, ACC) parameters.*



```
                    LD1        LD2
Galectin.3 -0.6858057   1.029024
ICAM1      -1.2168071 -1.719922
MMP9        1.2714411 -1.043071
         PredClass
RealClass DN HLD IgAN LN
     DN    12   0    3  3
     HLD    0  17    0  0
     IgAN   3   0   11  1
     LN     7   0    4  7
```

Figure 37 – Optimal Panel. *Left) rotation vectors for LD1 and LD2 (top) and confusion matrix (bottom), Right) projected values of the samples in LD1 and LD2.*

Correlation with eGFR

Proteomic measurements were also analyzed for correlation with the estimated glomerular filtration rate (eGFR). From the 18 valid proteins, 4 showed significant correlation with eGFR: TFF3 (p-value = 2.41e-12), RETN (p-value = 1.68e-03), COL3A1 (p-value = 4.08e-03), and CCL5 (p-value = 4.84 e-02) (Figure 38) when fitting linear models between the mean fluorescence intensity values (MFIs) and the eGFR. After correcting for multiple comparisons only TFF3 (p-value = 5.79e-11), RETN (p-value = 2.02e-02), and COL3A1 (p-value = 3.26e-02) remained significant.

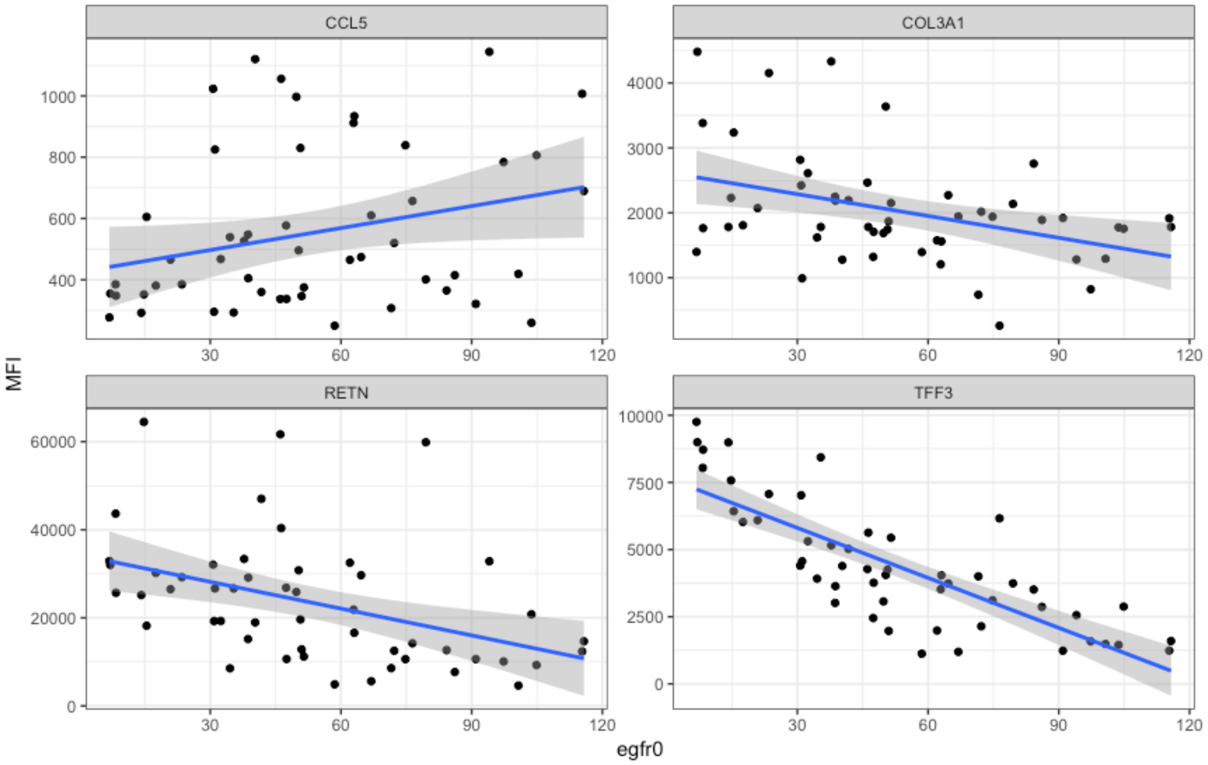


__Figure 38 - Correlation with eGFR.__ MFIs are shown as a function of eGFR. CCL5, COL3A1, RETN, and TFF3 showed significant correlation.

References

[1] Neusser, M.A., Lindenmeyer, M.T., Moll, A.G., Segerer, S., Edenhofer, I., Sen, K., Stiehl, D.P., Kretzler, M., Gröne, H.J., Schlöndorff, D. and Cohen, C.D., 2010. Human nephrosclerosis triggers a hypoxia-related glomerulopathy. The American journal of pathology, 176(2), pp.594-607.

[2] Woroniecka, K.I., Park, A.S.D., Mohtat, D., Thomas, D.B., Pullman, J.M. and Susztak, K., 2011. Transcriptome analysis of human diabetic kidney disease. Diabetes, 60(9), pp.2354-2369.

[3] Na, J., Sweetwyne, M.T., Park, A.S.D., Susztak, K. and Cagan, R.L., 2015. Diet-induced podocyte dysfunction in Drosophila and mammals. Cell reports, 12(4), pp.636-647.

[4] Berthier, C.C., Bethunaickan, R., Gonzalez-Rivera, T., Nair, V., Ramanujam, M., Zhang, W., Bottinger, E.P., Segerer, S., Lindenmeyer, M., Cohen, C.D. and Davidson, A., 2012. Cross-species transcriptional network analysis defines shared inflammatory responses in murine and human lupus nephritis. The Journal of Immunology, 189(2), pp.988-1001.

[5] Reich, H.N., Tritchler, D., Cattran, D.C., Herzenberg, A.M., Eichinger, F., Boucherot, A., Henger, A., Berthier, C.C., Nair, V., Cohen, C.D. and Scholey, J.W., 2010. A molecular signature of proteinuria in glomerulonephritis. PloS one, 5(10), p.e13451.

[6] Ju, W., Greene, C.S., Eichinger, F., Nair, V., Hodgin, J.B., Bitzer, M., Lee, Y.S., Zhu, Q., Kehata, M., Li, M. and Jiang, S., 2013. Defining cell-type specificity at the transcriptional level in human disease. Genome research, 23(11), pp.1862-1873.

[7] Ju, W., Nair, V., Smith, S., Zhu, L., Shedden, K., Song, P.X., Mariani, L.H., Eichinger, F.H., Berthier, C.C., Randolph, A. and Lai, J.Y.C., 2015. Tissue transcriptome-driven identification of epidermal growth factor as a chronic kidney disease biomarker. Science translational medicine, 7(316), pp.316ra193-316ra193.

[8] Liu, P., Lassén, E., Nair, V., Berthier, C.C., Suguro, M., Sihlbom, C., Kretzler, M., Betsholtz, C., Haraldsson, B., Ju, W. and Ebefors, K., 2017. Transcriptomic and proteomic profiling provides insight into mesangial cell function in IgA nephropathy. Journal of the American Society of Nephrology, 28(10), pp.2961-2972.

Chapter 5 Conclusions

The present project exhibits a proteomic biomarker discovery scheme that feeds on transcriptomic data using CKD as case study. The pipeline combines biological models extracted from publicly available databases with state-of-the-art algorithms that aim to capture meaningful biology for the disease. The goal is to suggest a number of proteins as candidate biomarkers to be measured in blood samples that are able to asymptomatically diagnose a myriad of subcategories and use machine learning methods to integrate them into a panel of increased performance and stability. The cornerstone is to molecularly determine the underlying pathophysiological mechanism causing CKD and detect it before the damage is done.

Our methodology follows several steps for each of which there exist several alternatives that potentially affect the final results. Therefore, the first part of the discussion consists of the justification behind the selected methodologies, while the second part focuses on the discussion of the results.

Methodological considerations

Data Acquisition and Pre-Processing

The pre-processing step involves everything from data collection to combination, which includes normalization, filtering, and integration. This step is critical as it affects the quality of all the downstream results regardless of the analytical methods chosen. It ensures that the data (coming from different sources) is comparable and can be used as a whole. If the input data is not good, the obtained results won't be good either regardless of the analytical tools and algorithms used, a concept described in computer science as garbage in, garbage out (GIGO) (Figure 39).

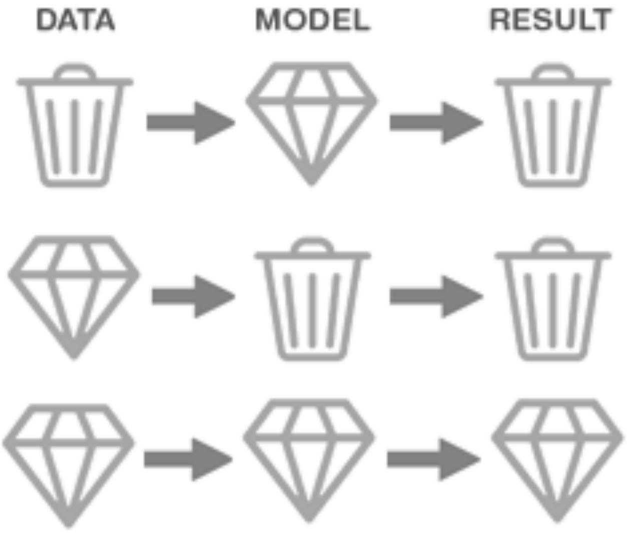


Figure 39 - Garbage in, garbage out (GIGO). Quality results are only obtained when both, the input data and applied analytical tools are of quality.

In this study, we are using public gene expression microarray data which measures the expression of thousands of genes in a single assay, using multiple probes to assay each transcript, making them a valuable data source for biomarker discovery and identification as they capture most of the pathophysiological phenomena, thus providing valuable information on disease subcategories, disease prognosis, and treatment outcome [1]. Microarray technology has become the gold standard for identifying genes or pathways whose expression changes between specific phenotypes.

The expression data was obtained from studies were reported in Nephroseq and obtained from the Gene Expression Omnibus (GEO) database. GEO supports community-derived reporting standards that facilitate integration of different datasets [2]. However, gene expression microarray measurements can be affected by small differences in any number of non-biological variables, so different instruments, technicians, reagent lots, or even days in which the experiments were carried out can affect the data.

When facing integration of different datasets, there are two general approaches: The first one involves meta-analysis or combination of the results from multiple studies; while the second one encompasses integration or combination of data from different studies. In this study, we opted for the latter given that the data was obtained from standardized databases which eases integration and because the heterogeneity of the samples to be combined (different studies included different disease groups) (Figure 11 - Results). If each experiment was analyzed individually, given the way contrasts have been built (case vs pool of controls), the control groups would be very different one from another, thus introducing bias due to experiment.

Data filtering was another critical stage to ensure the quality of the input dataset. Heterogeneous datasets make integration challenging, thus the different steps removing incomplete features, observations, and under-represented disease groups increase comparability of the datasets and facilitate their integration. The integrated dataset is subject to batch effect which was addressed using ComBat. In the same way that there are many non-biological variables that affect the data, there are many methods to alleviate batch effects ComBat was the method of choice as it has shown increased performance when compared with five other well-known batch effect adjustment methods in independent studies [3]. The adjustment is calculated without setting any covariates because adjusting for batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses [4].

Generation of the molecular profile

The study presented in this thesis involves an integrative holistic analysis in different biological scales: genes, pathways, and network (Figure 40). The analysis of the data in each scale returns a series of statistics (features) that are captured in a molecular profile representing a protein present in the Uniprot database. While the analysis in the gene-level is more straightforward with Limma being the gold-standard approach [5], analysis in the pathway- and network-levels is more complex, require of biological models for their implementation, and the selection of the best method is an open question.

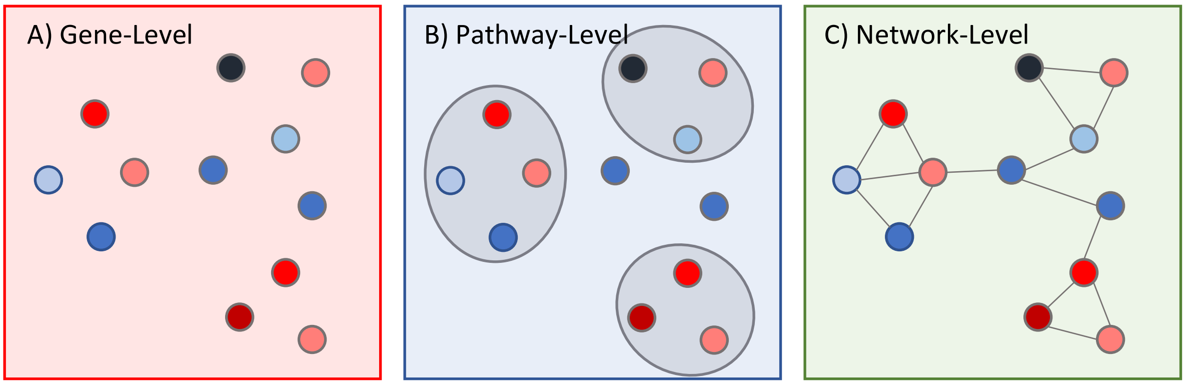


Figure 40 - Illustration of analysis in different biological scales. *A) Gene-Level, B) Pathway-Level, C) Network-Level.*

Perhaps it should be mentioned, that gene-level analysis revealed a relatively small number of differentially expressed genes with classical p-value and logFC thresholds (25 across all 8 subsets) (Figure 16 - Results). However, this is an expected result given that the methodology followed up to that point was very conservative (integration of datasets without disease co-variates). Moreover, the methodology presented here does not make threshold-based decisions and dichotomize the results between significant and non-significant but rather take the continuous values of the statistics for downstream analyses, approach that has been recently encouraged by the scientific community [6].

Pathway-level analysis: The number of degrees of freedom in this analysis is 3: the pathway database used, the method used for mapping gene-level statistics to pathway-level statistics, and the method used to combine p-values. While the number of pathway-database options is large, including broadly known the Kyoto Encyclopedia of Genes and Genomes (KEGG, https://www.genome.jp/kegg/) or Reactome (https://reactome.org/), the number of pathway-analysis methods is perhaps larger (only piano already includes 11). Kharti et al [7] reviewed the different approaches used in pathway analysis and classified them in three generations: Over-representation analysis (ORA), functional-class scoring (FCS), or pathway-topology-based (PT-based). The first two generations use pathways defined as gene-sets, while the third generation uses network-based representations of pathways. Intuitively, the methods and approaches used for each definition of pathway vary accordingly (for a review see: [8]). The use of different pathway-databases or analytical methods heavily affects the obtained results and, naturally, any downstream analyses performed. Given that there is not a clear understand of any individual pathway analysis method working better than the others, a common approach is to run a number of methods and then find consensus among the obtained results, which actually, outperforms methods individually [9].

In fact, both the database (MSigDB) and method of choice in this study (piano [10]), directly incorporate this logic as MSigDB contains gene sets curated by domain experts from various online databases such as Biocarta, Reacome, KEGG, and PID. Piano, instead, allows to use any number of 11 different gene set analysis methods and offers different alternatives to integrate the results. In this study, from the 11 available methods, the Wilcoxon rank-sum test was not included in the pipeline due to significantly longer computational times when compared with the rest of the methods. Using the classification system described in [7], these methods belong to the second-generation pathway analysis methods, functional class scoring (FCS). FCS methods rely on the assumption that although large changes in individual genes can have significant effects on the phenotype, weaker yet coordinated changes in sets of functionally correlated genes can also have significant effects. Second-generation methods act in two steps (two-tier structure): first, they calculate a gene-set statistic from the individual entities belonging to the set; and second, they assess whether if the calculated statistic is significant. FCS methods represent an improvement over the first-generation ones (Over-representation analysis, ORA) in the sense that they are not subject to user-defined thresholds to assess significance, but do not explode the pathway topology information in the way that third-generation methods do [7]. However, this information is included in the analysis in the network-level.

Using the status of a pathway as a biomarker may be impractical as the status of a pathway is generally obtained from a competitive analysis in which the score for a particular pathway is compared against an empirical background distribution built with the scores for the rest of the pathways. Therefore, we had to integrate enrichment statistics in the pathway-level back to the molecular entities. In the same way that a pathway is composed by several proteins (in the database encoded by their genes), a protein can appear in several pathways. Therefore, for each protein (encoded by its gene), the enrichment scores from the different pathways in which it participates were aggregated. We avoid the use of traditional score aggregation operators (mean, median, etc.) directly on the p-values as it can lead to loss of information [9]. We also avoid traditional p-value combination methods (Fisher, Stouffer, Tippet, sumZ, etc.) [9] because these methods may produce inaccurate results as they assume independency of the p-values to be integrated which is not the case here (different pathways can have proteins in common making the p-values correlated). Therefore, applying such methods would lead to inaccurate bimodal p-value distributions, and by definition, the p-values of null hypotheses should be uniformly distributed between 0 and 1 [11]. It has been demonstrated, that when correlation information is available, a weighting-capable method for integration of dependent p-values is performing best [11]. In this study, we used Hartung's method [12] which incorporates the correlation among p-values via introducing the correlation matrix in the Z-transform test. The resulting p-value distributions for the different scores was uniform in all the cases (Figure 41.A), unlike any other approach tested which resulted in bimodal p-value distributions (Figure 41.B, Tippet's method). Anticonservative distributions could me 'corrected' using multiple-hypothesis test correction (Bonferroni, FDR, etc.), but bimodal distributions usually mean that the wrong test has been selected [13, 14].

Network-level analysis: Integrative analysis of gene expression data in the context of biological networks such as protein-protein interaction (PPI) networks is increasingly becoming a major technique in systems biology, particularly in polygenic diseases where the phenotype cannot be explained by modifications in individual entities. However, most of published methods [15-18], focus on finding disease modules rather than assigning a score to each protein based on local information. Therefore, in this study, we propose a guilt-by-proximity approach in which a network-score is calculated for each protein which aggregates the gene-level scores for surrounding proteins based on the PPI network.

Our scoring function penalizes by the degree of the node in the graph because nodes of large degree represent hubs in the network which are more likely affected by any pathological/pharmacological condition, thus reducing the specificity of our biomarker candidates. Promising candidates are not hubs but are located in close proximity. For example, ICAM1, one of the best reported candidates to discriminate healthy samples from a generic definition of CKD (pool of DN, IgAN, and LN), is adjacent to EGFR, a hub in the PPI network (Figure 42).

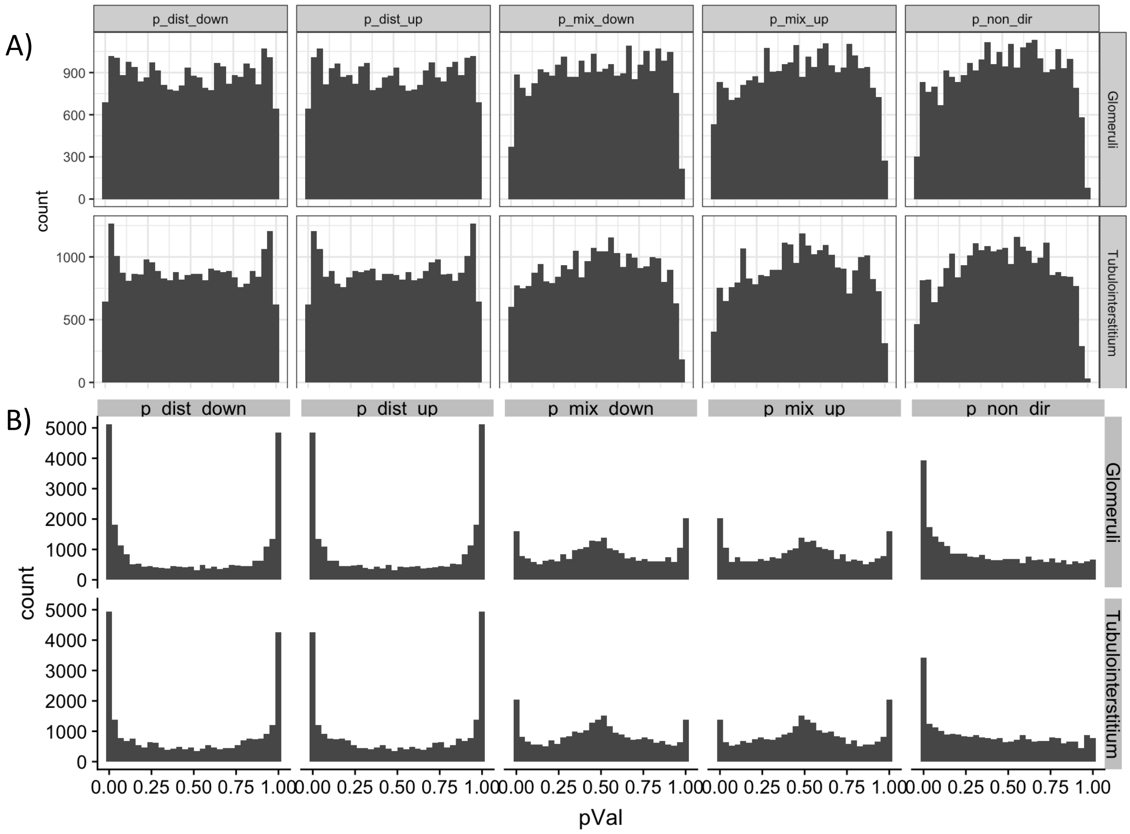


__Figure 41 - p-value distributions.__ P-value distributions after aggregation from different gene sets. A) Hartung's method, B) Tippet's method.

Our network-level analysis is also subject to the biological model used. The logic used here is similar to the one used in the pathway-level as Omnipath [19] is a database that collects protein-protein-interactions from different public databases (String, HPRD, etc.). In this case instead, in order to construct a high reliability prior knowledge network for downstream analysis, similarly to [20] inconsistencies between the different databases were excluded from the final network by including only those interactions that were reported in at least 3 external sources.

Map to protein identifiers: All the different analyses used HGNC nomenclature which is representative of genes. However, the goal is to obtain proteomic biomarkers. We have performed all the analyses in HGNC because not all the genes included in our models are protein encoding genes, but their expression may be affected by other pseudo-genes that do not encode for proteins. Therefore, after the construction of the molecular profile, the HGNC symbols were mapped to Uniprot IDs.

Feature selection/integration: Generally, there exist three strategies to deal with missing data: deleting the observations, deleting the variables, and imputing the missing values. The selection of the correct strategy depends on the structure of the data (number of observations, number of variables, amount of missing data):

- **Deleting observation:** If the number of observations is large in comparison to the number of variables and the different categories are sufficiently represented in the data, one may consider deleting some observations while preserving the power of the dataset.

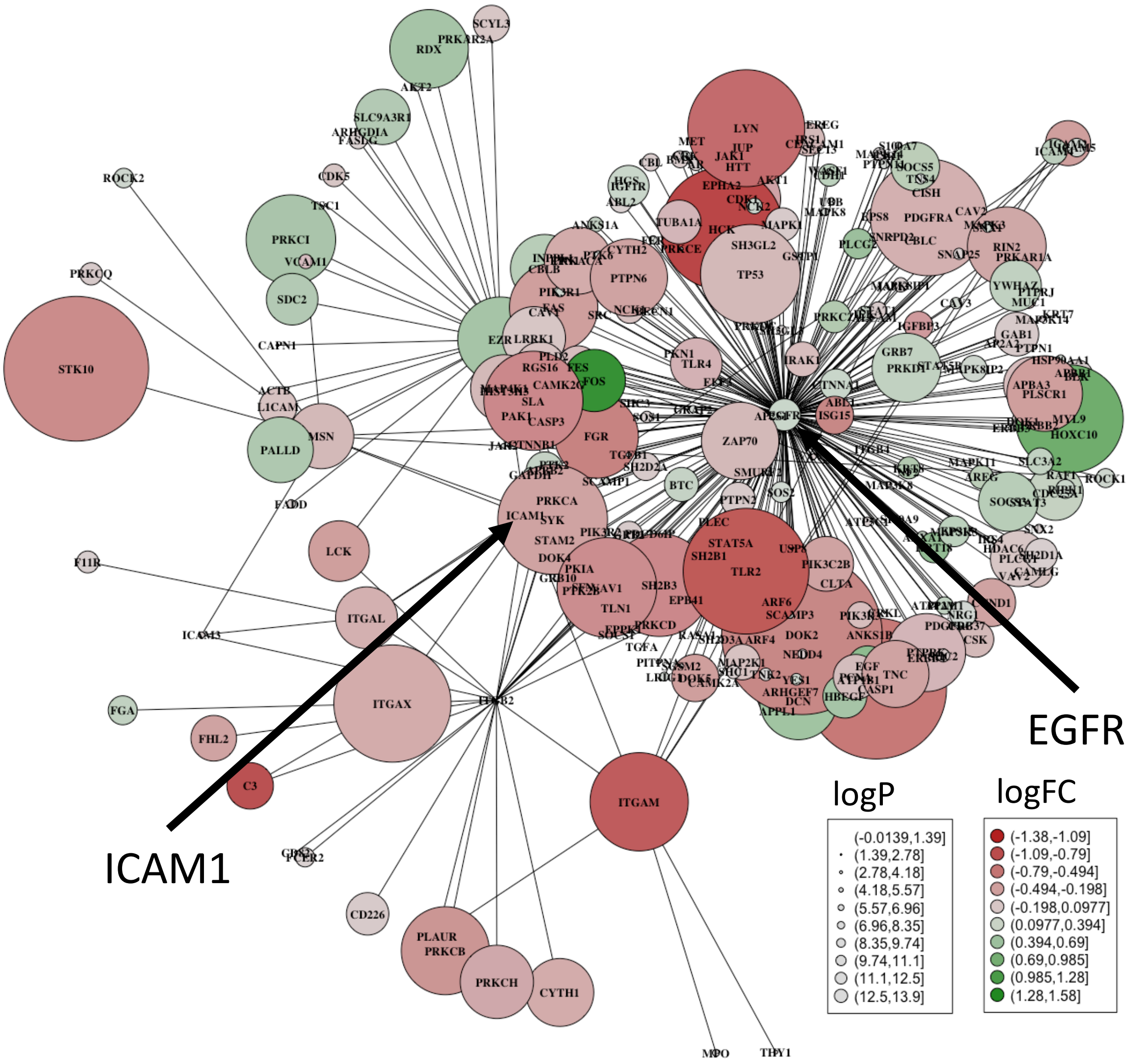


Figure 42 – ICAM1/EGFR network for glomerular tissue and HLD contrast. ICAM1 is an adjacent node of EGFR, a central hub of the protein-protein-interaction network. Both nodes are highlighted with black arrows.

- **Deleting features:** If most of missing values come from few variables and by removing these variables many observations can be saved, one would consider deleting variables instead.
- **Imputing values:** If the amount of non-missing data is enough to associate the different variables and get an idea of how they correlate, one may consider predicting the missing values using analytical approaches. Imputation with mean/median/mode is a crude way of treating missing values and may destroy the dataset. In complex cases, is usually best to use more advanced methods to predict missing values such as KNN [21], or MICE [22].

In this study, given that the number of observations (13743) is one thousand times larger than the number of features (11), and that the percentage of missing data represents 45.61% of the total data, we delete all the observations with missing data and proceeded only with the proteins containing data for all the features (complete observations). The complete observations were subject to different quality control checkpoints for feature selection because introducing non-informative or highly correlating features/predictors in a mathematical model, a common problem in biomedical datasets, can reduce the predictive performance or increase feature instability [23].

Secretion prediction: The subcellular location (extracellular, membranous, intracellular) of each protein was included as an indicative feature for potential users. We hypothesized that if a protein was extracellular it could be detected in blood samples. However, we did not filter to extracellular locations based on this feature because these predictions are valid only for the normal state. On the other hand, proteins that in normal circumstances are intracellular, may be extracellular in pathological conditions.

Experimental design

The molecular profiles were constructed using data obtained directly from two different tissues. However, samples were measured in blood samples. The motivation behind the use of blood samples is that, for monitoring human health, measuring protein biomarkers in blood is considered a very attractive solution because the pathology of almost every body tissue can affect the blood proteome [24] on top of the simplicity of obtaining the sample. A biomarker is more likely to progress to clinical settings if its measurement is minimally invasive.

The number of checkpoints for quality control of Luminex xMAP measurements may appear excessive with a lot of measurements lost. However, the number of patients lost was 3 for the training set and 12 for the validation set. Moreover, when constructing prediction models, particularly when the number of data points is enough, it is preferable to lose some good measurements rather than to introduce bad ones (better to lose signal than to introduce noise).

Our definition of biomarker performance is novel and integrates routinary indicators of performance used in biomarker discovery like ROC curves (AUC) with quantitative indicators of performance (log-fold-change, p-value). This integration tries to balance up the relatively low number of samples tested with modifications that rank-up robust biomarker candidates.

We also trained the obtained molecular profiles from each tissue individually and calculated a final score by aggregating the predictions obtained from each model to harmonize with our “wisdom of the crowds” analytical approach where combining multiple results enhance the performance. One of the greatest results obtained is that the predicted biomarker performances for both tissues were very correlated across all proteins (p-value < 2e-16) which gives us confidence about the robustness of our metric.

Biomarker discovery method: performance

The goal of this thesis was the characterization of kidney diseases at the molecular level. This characterization would assist the discovery of new biomarkers directly associated with the mechanism of the disease. The proposed pipeline performs an agnostic interrogation that aims to define different phenotypes in molecular terms. To that end, it needs the integration of knowledge in different levels (transcriptomics, pathways, network) to better understand the molecular relationships underlying the pathophysiology in a phenotype-specific manner to develop more reliable sets of biomarkers for diagnosis and personalize treatment of individual patients.

This type of systems-based approach is perhaps most useful in complex and heterogeneous diseases like Chronic Kidney Disease (CKD) or cancer given that they are better explained by alterations in molecular mechanisms rather than effects in individual genes [25]. In the case of CKD, current biomarkers like serum creatinine (sCr), NGAL, or KIM-1 either lack the sensitivity and specificity required, or cannot distinguish between different subtypes, or report alterations when the damage is done. The ideal set of biomarkers would be composed by a small number of analytes that are able to diagnose a myriad of subcategories before the damage is done (asymptomatic). Moreover, this set of analytes should be connected to the mechanism of the disease to minimize their attrition rate in clinical practice.

Therefore, recent efforts to predict outcome in cancer exploit the wealth of information about gene regulations, pathways, protein-protein-interactions, metabolic reactions and other types of relationships between biomolecules available in public databases [25-29]. In the context of phenotype prediction, these methods aggregate groups of biologically related features into scores that are often used in supervised machine learning methods in the expectation that they will provide increased performance and stability. However, the claim that scores obtained from aggregations represent an improvement over single genes has been recently challenged as more and more studies come up demonstrating that aggregation scores do not outperform single genes in prediction performance or stability [30, 31]. Moreover, if we check commercialized products that are being used as biomarkers, they use single-gene panels rather than MGs. Examples of this are Mammaprint® [32], a panel of 70 genes for breast-cancer prediction, Decipher® [33], a panel of 22 RNA markers to predict risk of metastases in prostate cancer, PAM50 [34] a set of 50 genes stratifying breast cancer patients into five 'intrinsic' molecular subtypes, and OncoTypeDX® [35], a set of panels for tumor profiling in breast, prostate, and colon cancer.

To obtain a set of predictive features in CKD, we used predefined pathways and protein-protein-interactions expertly curated and stored in public databases (MSigDB, OmniPath). However, most of these databases are not disease-specific so only partially represent the disease in question. Moreover, despite significant improvement over the years, a priori defined databases are still biased towards well-understood biological processes and few pathways are compiled and annotated for rare or specific sub-types of diseases [7]. We found a promising kidney-specific protein-protein-interaction network in TissueNet V2 [36] which has the benefit of being kidney specific but the limitation of not being curated by domain experts. Moreover, TissueNet would represent the only kidney-specific database included in our model thus losing the robustness and reliability acquired by integration of several curated databases. Therefore, TissueNet could be potentially used to expand and modify OmniPath, but not a substitute.

A successful model construction requires molecular patterns that consistently change across disease patients. Our model training revealed that features coming from the network-level have the biggest role in biomarker performance predictions as the network-level-statistic localP has the biggest weight in the predicted models for both tissues (Figure 27). Other authors [37] have already reported the relevance of network analysis as a potential tool for biomarker discovery which is encouraging the scientific community to move in this direction as several studies have been recently published using biological networks as means for biomarker discovery [38, 39].

Going back to the example shown in Figure 42, it appears that promising candidate biomarkers are located next to hubs in the network. The intercellular adhesion molecule 1 (ICAM1), the protein from the training set with the highest network-level score for the localP feature, is at the same time one of the best performing biomarker candidates to classify the HLD group from the rest of groups (generic CKD candidate). ICAM-1 is an endothelial- and leukocyte-associated transmembrane protein known for its importance in stabilizing cell-cell interactions and facilitating leukocyte endothelial transmigration. ICAM-1 ligation produces proinflammatory effects such as inflammatory leukocyte recruitment by signaling through cascades involving a number of kinases [40]. ICAM-1 is involved in FGF23 signaling, a mechanism that impairs neutrophil recruitment and host defence during CKD [41]. Based on our network model, ICAM-1 is located adjacent to epidermal growth factor (EGFR) which has also been associated to CKD [42]. In the gene level, none of the evaluated contrasts showed ICAM-1 as significantly differentially expressed. In the pathway level, ICAM1 participates in several significantly affected pathways like the BIOCARTA_MONOCYTE_PATHWAY gene set or the KEGG_VIRAL_MYOCARDITIS gene set. The KEGG_VIRAL_MYOCARDITIS gene set is particularly interesting because highlights the systematic nature of CKD and there exist several studies associating CKD with cardiovascular diseases [43].

The other two biomarker candidates appearing in our panel, the matrix metallopeptidase 9 (MMP9) and Galectin-3 were suggested by our biomarker discovery model (validation set) and have also been recorded in association with CKD.

The matrix metallopeptidase 9 (MMP9), is an enzyme involved in the degradation of the extracellular matrix, thus is also involved in several important functions within neutrophil actions. Peripheral blood levels of MMP9 have been found to be altered in patients with chronic kidney disease on conservative treatment and on hemodialysis [44]. It was also reported as to be related to renal function in patients with CKD [45, 46]. Similarly to ICAM1, MMP9 is not significantly affected in the transcriptomic level. However, it participates in several affected pathways including AP-1 pathway and FRA pathway, both of which have been associated with renal fibrosis [47-48].

Galectin-3 is a member of the beta-galactoside-binding protein family and plays an important role in cell-cell adhesion, cell-matrix interactions, macrophage activation, angiogenesis, metastasis, apoptosis. Galectin-3 is an established biomarker for cardiac fibrosis [49] and has also been linked to the development of CKD [50]. In our study, it was not significantly affected in any of the contrasts in the gene or pathway level. In the network-level instead, Galectin 3 is located close to ABCA1, an important hub in our network model. ABCA1 is a transporter that removes cholesterol from macrophages [51]. In fact, cholesterol accumulation is one of the main risk factors for cardiovascular disease (CVD) which again highlights the relationship between CVD and CKD.

As individual predictors, Galectin-3 is the best performing biomarker specifically for DN (BP = 72.36, AUC = 0.79), and MMP9 is the best performing biomarker for LN (BP = 73.66, AUC = 0.74). If these results are in fact not as strong as TFF3 (BP = 96.08, AUC = 0.95) or ICAM (BP = 90.24, AUC = 1) for the HLD group, they show a moderate correlation that is of biological interest. In fact, the obtained optimal panel misclassifies the most between the LN and DN groups (10 out of 21 misclassified samples). However, there is a biological reason that could explain this behaviour as both DN and LN are secondary glomerulonephritis while IgAN is a primary glomerulonephritis (see Chapter 2). If we would merge DN and LN groups together and use our panel to distinguish primary glomerulonephritis, secondary glomerulonephritis and control we would obtain a classification accuracy of 83.82% which starts to be of clinical relevance (Figure 43).

```
                  LD1        LD2
Galectin.3  0.6538969  1.149873
ICAM1       1.1720349 -1.720271
MMP9       -1.2684012 -1.011506


              PredClass
RealClass      HLD PrimaryGN SecondaryGN
  HLD           17         0           0
  PrimaryGN      0        11           4
  SecondaryGN    0         7          29
```

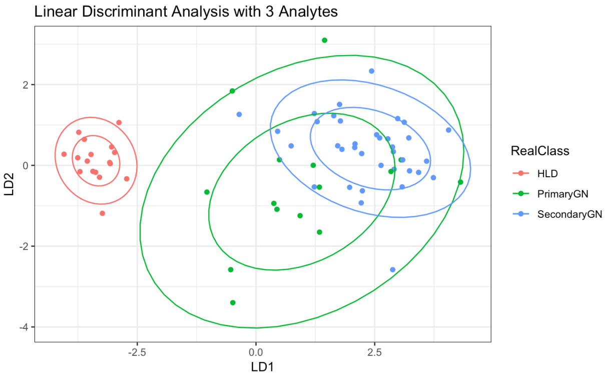


Figure 43 – Optimal Panel – glomerulonephritis groups. Left) rotation vectors for LD1 and LD2 (top) and confusion matrix (bottom), Right) projected values of the samples in LD1 and LD2.

From the proteins included in the validation set, Nephrin (NPHS1) is a protein associated with CKD that didn't show any significance in proteomic measurements. Nephrin is a protein necessary for the proper functioning of the renal filtration barrier which consists of fenestrated endothelial cells, the glomerular basement membrane, and the podocytes of epithelial cells. One possible reason for this poor performance is that, based on the Tissue Protein Atlas (https://www.proteinatlas.org/), Nephrin is a membranous protein almost exclusively present in kidney tissue (particularly in the nephrons). This fact, together with the low expression found across all samples makes us consider Nephrin as a promising marker for CKD but not in blood samples. In fact, NPHS1 has been found mutated in congenital nephrotic syndrome [55], a primary cause for CKD in children.

Finally, we also found 3 proteins (after multiple testing correction) that are correlated with eGFR: TFF3, RETN, and COL3A1. COL3A1 showed some predictive ability in DN (AUC = 0.65). COL3A1 is the gene that encodes for collagenase 3 alpha-1 chain. It has been recently shown that it is related with development of tubulointerstitial injury in DN [52] in the transcriptomic level. On the other hand, trefoil factor 3 (TFF3) was the protein that showed the most significant correlation with eGFR (p-value = 2.41e-12). Moreover, it was the protein that obtained the best biomarker performance score (individually) for HLD. It showed an AUC = 0.95, a logFC = -1.95 and a p-value = 3.09e-13. This protein has been extensively reported in literature as related not only to CKD [58] but more particularly to eGFR [59].

Recapitulating, we have introduced a computational pipeline that can find promising biomarker candidates in serum samples using as input gene expression data. Our pipeline uses mostly, but not only, already published algorithms and methods, but their integration and application is novel. The method identifies strong protein biomarkers when there is strong contrast at the gene expression data. The potential use of some of these proteins as biomarkers in CKD has been validated with limited number of blood samples. We also have shown (using our biological models and available literature) that these biomarker candidates are connected to mechanisms related to CKD. Finally, we have implemented a reduced version of the algorithm in shiny and deployed it online. Although in this study we have focused on CKD, the pipeline and the online app are non-specific and could potentially be used for any set of phenotypes with transcriptomic data.

REFERENCES

[1] Trevino, V., Falciani, F. and Barrera-Saldaña, H.A., 2007. DNA microarrays: a powerful genomic tool for biomedical and clinical research. Molecular Medicine, 13(9), p.527.

[2] Clough, E. and Barrett, T., 2016. The gene expression omnibus database. In Statistical Genomics (pp. 93-110). Humana Press, New York, NY.

[3] Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L. and Liu, C., 2011. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PloS one, 6(2), p.e17238.

[4] Nygaard, V., Rødland, E.A. and Hovig, E., 2016. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics, 17(1), pp.29-39.

[5] Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43(7), pp.e47-e47.

[6] Amrhein, V., Greenland, S. and McShane, B., 2019. Scientists rise up against statistical significance.

[7] Khatri, P., Sirota, M. and Butte, A.J., 2012. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS computational biology, 8(2), p.e1002375.

[8] Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichita, C. and Draghici, S., 2013. Methods and approaches in the topology-based analysis of biological pathways. Frontiers in physiology, 4, p.278.

[9] Alhamdoosh, M., Ng, M., Wilson, N.J., Sheridan, J.M., Huynh, H., Wilson, M.J. and Ritchie, M.E., 2017. Combining multiple tools outperforms individual methods in gene set enrichment analyses. Bioinformatics, 33(3), pp.414-424.

[10] Väremo, L., Nielsen, J. and Nookaew, I., 2013. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic acids research, 41(8), pp.4378-4391.

[11] Alves, G. and Yu, Y.K., 2014. Accuracy evaluation of the unified P-value from combining correlated P-values. PloS one, 9(3), p.e91225.

[12] Hartung, J., 1999. A note on combining dependent tests of significance. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 41(7), pp.849-855.

[13] Halsey, L.G., Curran-Everett, D., Vowler, S.L. and Drummond, G.B., 2015. The fickle P value generates irreproducible results. Nature methods, 12(3), p.179.

[14] Murdoch, D.J., Tsai, Y.L. and Adcock, J., 2008. P-values are random variables. The American Statistician, 62(3), pp.242-245.

[15] Beisser, D., Klau, G.W., Dandekar, T., Müller, T. and Dittrich, M.T., 2010. BioNet: an R-Package for the functional analysis of biological networks. Bioinformatics, 26(8), pp.1129-1130.

[16] Alvarez, M.J., Shen, Y., Giorgi, F.M., Lachmann, A., Ding, B.B., Ye, B.H. and Califano, A., 2016. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nature genetics, 48(8), p.838.

[17] Leiserson, M.D., Vandin, F., Wu, H.T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M. and Lawrence, M.S., 2015. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature genetics, 47(2), p.106.

[18] Ghiassian, S.D., Menche, J. and Barabási, A.L., 2015. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS computational biology, 11(4), p.e1004120.

[19] Türei, D., Korcsmáros, T. and Saez-Rodriguez, J., 2016. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nature methods, 13(12), p.966.

[20] Kirouac, D.C., Saez-Rodriguez, J., Swantek, J., Burke, J.M., Lauffenburger, D.A. and Sorger, P.K., 2012. Creating and analyzing pathway and protein interaction compendia for modelling signal transduction networks. BMC systems biology, 6(1), p.29.

[21] García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R. and Verleysen, M., 2009. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing, 72(7-9), pp.1483-1493.

[22] Royston, P., 2004. Multiple imputation of missing values. The Stata Journal, 4(3), pp.227-241.

[23] Toloşi, L. and Lengauer, T., 2011. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics, 27(14), pp.1986-1994.

[24] Anderson, N.L. and Anderson, N.G. (2002) The human plasma proteome history, character, and diagnostic prospects. Mol. Cell. Proteomics 1, 845–867.

[25] Alcaraz, N., List, M., Batra, R., Vandin, F., Ditzel, H.J. and Baumbach, J., 2017. De novo pathway-based biomarker identification. Nucleic acids research, 45(16), pp.e151-e151.

[26] Kim, S., Kon, M. and DeLisi, C., 2012. Pathway-based classification of cancer subtypes. Biology direct, 7(1), p.21.

[27] Lee, E., Chuang, H.Y., Kim, J.W., Ideker, T. and Lee, D., 2008. Inferring pathway activity toward precise disease classification. PLoS computational biology, 4(11), p.e1000217.

[28] Nevins, J.R., 2011. Pathway-based classification of lung cancer: a strategy to guide therapeutic selection. Proceedings of the American Thoracic Society, 8(2), pp.180-182.

[29] Hou, D. and Koyutürk, M., 2014. Comprehensive evaluation of composite gene features in cancer outcome prediction. Cancer informatics, 13, pp.CIN-S14028.

[30] Staiger, C., Cadot, S., Kooter, R., Dittrich, M., Müller, T., Klau, G.W. and Wessels, L.F., 2012. A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. PloS one, 7(4), p.e34796.

[31] Allahyar, A. and De Ridder, J., 2015. FERAL: network-based classifier with application to breast cancer outcome prediction. Bioinformatics, 31(12), pp.i311-i319.

[32] Van't Veer, L.J., Dai, H., Van De Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., Van Der Kooy, K., Marton, M.J., Witteveen, A.T. and Schreiber, G.J., 2002. Gene expression profiling predicts clinical outcome of breast cancer. nature, 415(6871), p.530.

[33] Karnes, R.J., Bergstralh, E.J., Davicioni, E., Ghadessi, M., Buerki, C., Mitra, A.P., Crisan, A., Erho, N., Vergara, I.A., Lam, L.L. and Carlson, R., 2013. Validation of a genomic classifier that predicts metastasis following radical prostatectomy in an at risk patient population. The Journal of urology, 190(6), pp.2047-2053.

[34] Parker, J.S., Mullins, M., Cheang, M.C., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z. and Quackenbush, J.F., 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology, 27(8), p.1160.

[35] Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., Walker, M.G., Watson, D., Park, T. and Hiller, W., 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. New England Journal of Medicine, 351(27), pp.2817-2826.

[36] Basha, O., Barshir, R., Sharon, M., Lerman, E., Kirson, B.F., Hekselman, I. and Yeger-Lotem, E., 2016. The TissueNet v. 2 database: A quantitative view of protein-protein interactions across human tissues. Nucleic acids research, 45(D1), pp.D427-D431.

[37] Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C. and Raphael, B.J., 2015. Pathway and network analysis of cancer genomes. Nature methods, 12(7), p.615.

[38] Zhou, M., Diao, Z., Yue, X., Chen, Y., Zhao, H., Cheng, L. and Sun, J., 2016. Construction and analysis of dysregulated lncRNA-associated ceRNA network identified novel lncRNA biomarkers for early diagnosis of human pancreatic cancer. Oncotarget, 7(35), p.56383.

[39] Seiwert, T.Y., Zuo, Z., Keck, M.K., Khattri, A., Pedamallu, C.S., Stricker, T., Brown, C., Pugh, T.J., Stojanov, P., Cho, J. and Lawrence, M.S., 2015. Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. Clinical cancer research, 21(3), pp.632-641.

[40] Luo, B.H., Carman, C.V. and Springer, T.A., 2007. Structural basis of integrin regulation and signaling. Annu. Rev. Immunol., 25, pp.619-647.

[41] Rossaint, J., Oehmichen, J., Van Aken, H., Reuter, S., Pavenstädt, H.J., Meersch, M., Unruh, M. and Zarbock, A., 2016. FGF23 signaling impairs neutrophil recruitment and host defense during CKD. The Journal of clinical investigation, 126(3), pp.962-974.

[42] Ju, W., Nair, V., Smith, S., Zhu, L., Shedden, K., Song, P.X., Mariani, L.H., Eichinger, F.H., Berthier, C.C., Randolph, A. and Lai, J.Y.C., 2015. Tissue transcriptome-driven identification of epidermal growth factor as a chronic kidney disease biomarker. Science translational medicine, 7(316), pp.316ra193-316ra193.

[43] Zoccali, C., Vanholder, R., Massy, Z.A., Ortiz, A., Sarafidis, P., Dekker, F.W., Fliser, D., Fouque, D., Heine, G.H., Jager, K.J. and Kanbay, M., 2017. The systemic nature of CKD. Nature Reviews Nephrology, 13(6), p.344.

[44] Pawlak, K., Mysliwiec, M. and Pawlak, D., 2011. Peripheral blood level alterations of MMP-2 and MMP-9 in patients with chronic kidney disease on conservative treatment and on hemodialysis. Clinical biochemistry, 44(10-11), pp.838-843.

[45] Chang, H.R., Yang, S.F., Li, M.L., Lin, C.C., Hsieh, Y.S. and Lian, J.D., 2006. Relationships between circulating matrix metalloproteinase-2 and-9 and renal function in patients with chronic kidney disease. Clinica Chimica Acta, 366(1-2), pp.243-248.

[46] Cheng, Z., Limbu, M., Wang, Z., Liu, J., Liu, L., Zhang, X., Chen, P. and Liu, B., 2017. MMP-2 and 9 in chronic kidney disease. International journal of molecular sciences, 18(4), p.776.

[47] Khan, Z. and Pandey, M., 2014. Role of kidney biomarkers of chronic kidney disease: An update. Saudi journal of biological sciences, 21(4), pp.294-299.

[48] Liu, Y., 2006. Renal fibrosis: new insights into the pathogenesis and therapeutics. Kidney international, 69(2), pp.213-217.

[49] Ho, J.E., Liu, C., Lyass, A., Courchesne, P., Pencina, M.J., Vasan, R.S., Larson, M.G. and Levy, D., 2012. Galectin-3, a marker of cardiac fibrosis, predicts incident heart failure in the community. Journal of the American College of Cardiology, 60(14), pp.1249-1256.

[50] O'Seaghdha, C.M., Hwang, S.J., Ho, J.E., Vasan, R.S., Levy, D. and Fox, C.S., 2013. Elevated galectin-3 precedes the development of CKD. Journal of the American Society of Nephrology, 24(9), pp.1470-1477.

[51] Reiss, A.B., Voloshyna, I., De Leon, J., Miyawaki, N. and Mattana, J., 2015. Cholesterol metabolism in CKD. American Journal of Kidney Diseases, 66(6), pp.1071-1082.

[52] Vivante, A. and Hildebrandt, F., 2016. Exploring the genetic basis of early-onset chronic kidney disease. Nature Reviews Nephrology, 12(3), p.133.

[53] Zeng, M., Liu, J., Yang, W., Zhang, S., Liu, F., Dong, Z., Peng, Y., Sun, L. and Xiao, L., 2019. Multiple-microarray analysis for identification of hub genes involved in tubulointerstial injury in diabetic nephropathy. Journal of cellular physiology.

[54] Du, T.Y., Luo, H.M., Qin, H.C., Wang, F., Wang, Q., Xiang, Y. and Zhang, Y., 2013. Circulating serum trefoil factor 3 (TFF3) is dramatically increased in chronic kidney disease. PloS one, 8(11), p.e80271.

[55] Fassett, R.G., Venuthurupalli, S.K., Gobe, G.C., Coombes, J.S., Cooper, M.A. and Hoy, W.E., 2011. Biomarkers in chronic kidney disease: a review. Kidney international, 80(8), pp.806-821.

Appendix 1 – Calibration Curves

The assays measuring the proteins included in the optimal panel (Galectin3, ICAM1, and MMP9) were further characterized by constructing the calibration curves that allow to map the MFI values to molarity (ng/ml). This requires the production of a calibration curve using known concentrations of recombinant protein for each assay. All recombinant proteins were added in a mix, standard mix, at initial target concentration 50 ng/ml. Then, 3-fold serial dilutions were performed with Low cross normal buffer (CANDOR), same buffer used for the sample dilution, for 15 wells including negative control (0 ng/ml). All measurements were performed in triplicates (9 replicates for samples of 0 concentration, blanks). Based on the known concentration of each recombinant protein in the standard mix and the Median Florescent Intensity units (MFI) obtained, the calibration curve of the assay was created; thus, the concentration of each target protein in the samples was calculated using the respective calibration curve. The construction of the curves, followed the subsequent steps:

Bead counts (QC): Measurements with less than 30 bead counts were removed from further analysis [1]. In these assays, none of the measurements had less than 30 counts (min = 100).

Outlier detection: Replicate measurements were evaluated for outliers. For each triplicate, a data point was considered an outlier if its absolute deviation from the median of the replicates was larger than 3 median absolute deviations (MADs):

$$Data\ Point \rightarrow \begin{cases} Outlier, & if \quad \left|x_{i,j} - \widehat{x_i}\right| > 3MAD_i \\ NotOutlier, & if \quad \left|x_{i,j} - \widehat{x_i}\right| \leq 3MAD_i \end{cases}$$

Where $\widehat{x_i}$ is the median of the MFIs of the triplicates for protein *i*. 17 replicates (11.11% of the total) were identified as outliers for this reason and removed from further analysis.

Limit of Blank (LoB): The LoB is defined as the highest value (MFI) to be found when replicates of a sample containing no analyte are tested. If we assume a Gaussian distribution of the raw analytical signals from blank samples, and set a 95% confident interval (CI), then the LoB is calculated as [2]:

$$LoB_i = \mu_{b,i} + 1.645\sigma_{b,i}$$

Where $\mu_{b,i}$ and $\sigma_{b,i}$ are the average and standard deviation of the MFIs for the blanks (0 concentration) for protein *i*. Table 13 shows the resulting LoBs for the three assays.

Table 13 - Limit of Blank (LoB). *For each assay, the average for replicates of concentration 0 (µb), their standard deviation (σb) and the LoB are represented.*

Assay	μ_b (MFI)	σ_b (MFI)	LoB (MFI)
Galectin.3	2037.89	263.82	2562.53
ICAM1	767.56	56.07	879.71
MMP9	102.78	17.97	138.72

Limit of Detection (LoD): The LoD is defined as the lowest analyte concentration likely to be reliably distinguished from the LoB and at which detection is feasible. If we make the same assumptions made for the calculation of the LoB (Gaussian, and CI of 95%), then [3]:

$$LoD_i = LoB_i + c_{\beta,i} \cdot \sigma_{lc,i} = LoB_i + \frac{1.645}{1 - \frac{1}{4f_i}} \cdot \sigma_{lc,i}$$

Where σ_{lc} is the standard deviation for the relevant low concentration for the protein *i*, i.e., a concentration in the range from LoB to approximately 4LoB, and f is the number of degrees of freedom of the estimated standard deviation σ_{lc} (number of data points minus number of concentrations). For each protein, the low concentration range is defined as those concentrations whose $\mu_i + 2\sigma_i < 4LoB$, thus:

$$\sigma_{lc.i} = \sqrt[2]{\frac{\sum_{j=1}^{m_i}(n_{j,i} \cdot Var_{j,i})}{N_{lc,i}}}$$

Where $n_{j,i}$ is the number of data points for concentration *j* and protein *i*, $Var_{j,i}$ is the variance for concentration *j* and protein *i*, $N_{lc,i}$ is the number of data points included in the low concertation range for protein *i*, and m_i is the number of concentrations measured for the low concentration range for protein *i*. Table 14 shows the resulting LoDs for the three assays.

Table 14 - Limit of Detection (LoD). *For each assay, the number of data-points in the low concentration range Nlc, their standard deviation (σlc), the coefficient (Cb) and the LoD are represented.*

Assay	N_{lc}	σ_{lc} (MFI)	Cb	LoD (MFI)
Galectin.3	14	146.64	1.68	2811.88
ICAM1	27	104.65	1.66	1053.61
MMP9	9	6.68	1.7	150.12

Select calibration range: Fitting of calibration curves is best when none of the different areas is over/under-represented [4]. Thus, for each protein few concentrations in each area of the calibration range were selected (lower asymptote, linear range, upper asymptote). For the upper asymptote, we selected the concentration with the maximum median MFI as the upper limit. For the lower asymptote, we selected the biggest concentration with median MFI smaller than the LoD. The concentration ranges used for curve fitting for each analyte are represented in Table 15. Figure 44 shows the semi-log curves correlating concentration and MFI for each assay.

Table 15 - Calibration range. *For each assay, the lowest and highest concentrations used to fit the curve are represented.*

Assay	lowConcentration (ng/ml)	highConcentration (ng/ml)
Galectin.3	0.0228	50
ICAM1	0.0038	62.5
MMP9	0.0038	250

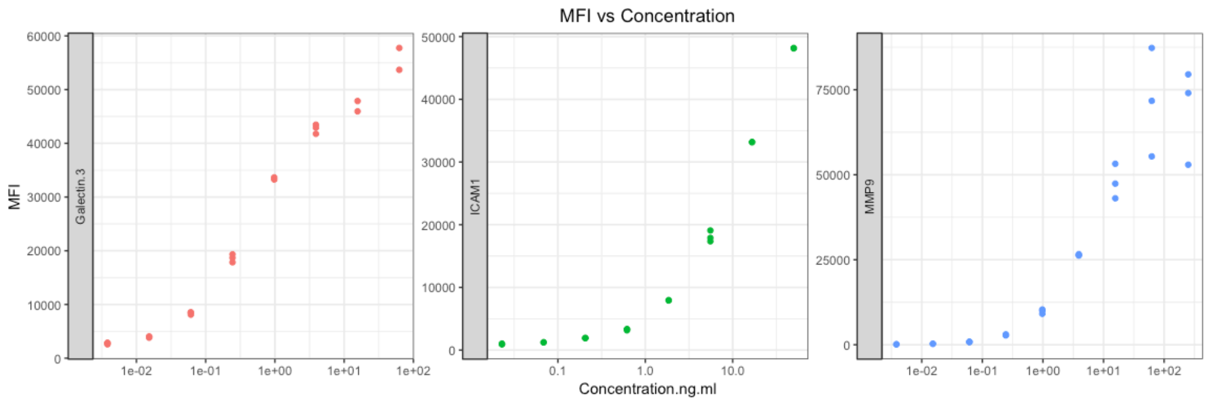


Figure 44 – Semi-log concentration ranges. *For each assay, the MFI is shown as a function of the log10 of the concentration*

Replicate precision: Concentrations whose replicates with a coefficient of variation of more than 25% were excluded from further analysis. In this case, none of the replicates showed a coefficient of variation higher than 25% (max = 22.32%) so none of the replicates was eliminated for this reason.

Calibration of the curves: Standard curves were calculated using 5 Parameter Logistic (5PL) curves as mathematical models:

$$F(x) = D + \frac{A - D}{\left(1 + \left(\frac{x}{C}\right)^{B}\right)^{E}}$$

Where, x represents a given concentration of recombinant protein, F(x) the reported MFI, and A-E the parameters of the curve:

- A: Minimum asymptote, response value at 0.
- B: Hill's slope, steepness of the curve
- C: Inflection point, the point on the curve where the curvature changes direction or sign. C is the concentration of analyte where F(x) = (D-A)/2.
- D: Maximum asymptote, response value at infinite.
- E: Asymmetry factor. When E = 1 the curve is symmetrical around the inflection point resulting in a four-parameter logistic regression (4PL).

In the present study, we used the R package nplr [5] which requires normalization of the MFIs in the [0-1] range:

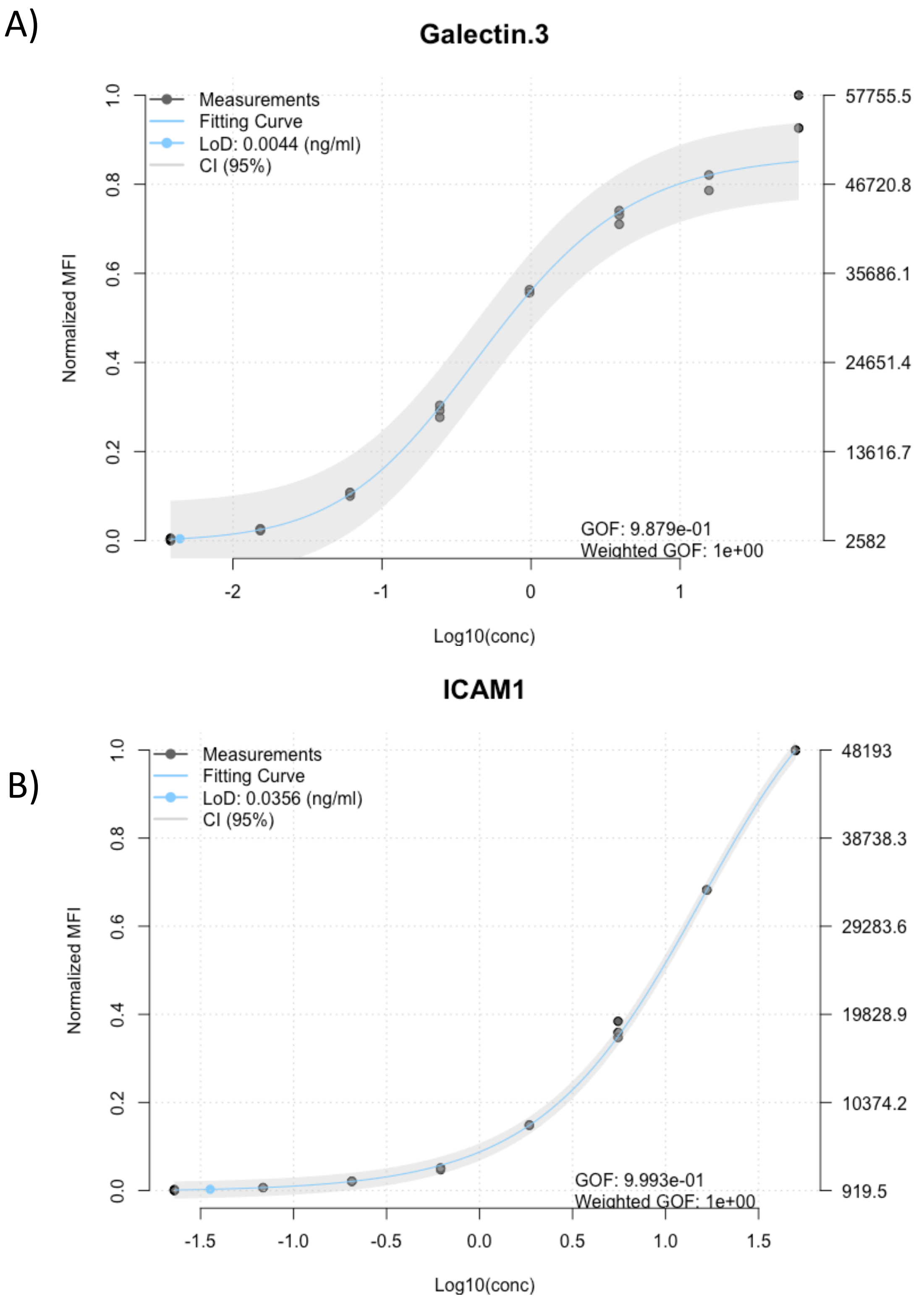
$$F(x)^{*}_{i,j} = \frac{F(x)_{i,j} - \min(F(x)_i)}{\max(F(x)_i) - \min(F(x)_i)}$$

Where $y_{i,j}$ is the MFI for a given protein *i* at a concentration *j*, and min/max y_i are the minimum/maximum values for each protein across all concentrations. The 5PL curve estimates the MFI for a given concentration. To solve the inverse problem (concentration for a given MFI) we need to invert F(x):

$$F'(y) = C\sqrt[B]{\sqrt[E]{\frac{A-D}{y-D}-1}}$$

The calibration curves obtained are shown in

Figure 45. Table 16 shows the coefficient values for each curve.



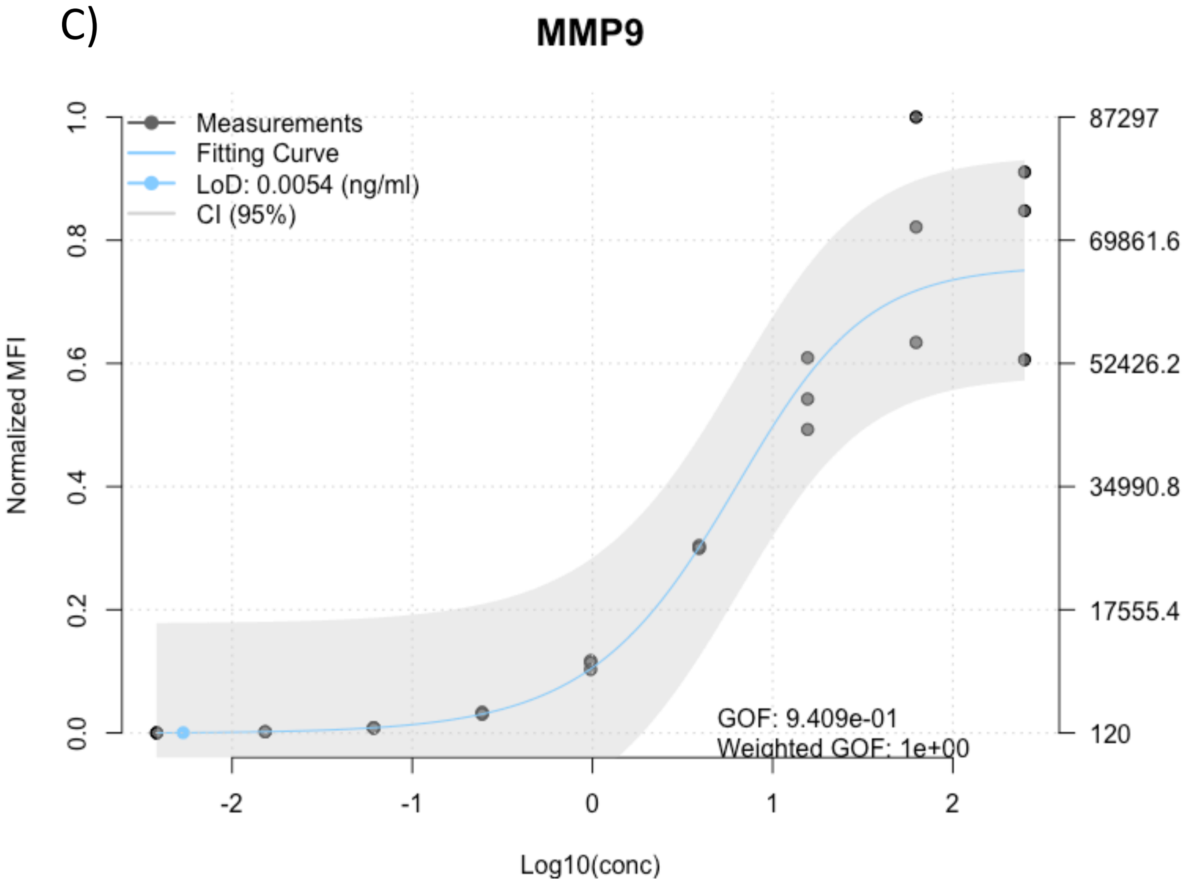


__Figure 45 - Calibration curves.__ A) Galectin.3, B) ICAM1, C) MMP9. The y-axes (right) shows the MFI and (left) the normalized MFI. The x-axes shows the concentration (ng/ml) in log10-scale. GOF represents the goodness of fit.

__Table 16 - Calibration coefficients.__ A: Minimum asymptote, B: Hill's slope, C: Inflection point, D: Maximum asymptote, E: Asymmetry factor.

Assay	**A**	**B**	**C**	**D**	**E**
Galectin.3	-0.0030	0.7958	-0.5777	0.8673	1.4547
ICAM1	-0.0021	1.1527	1.2951	1.2503	0.7598
MMP9	-0.0006	1.2635	0.9277	0.7584	0.7099

Quality control: the functional sensitivity of the assays (A) is calculated by measuring the relative error between the obtained MFI for a given concentration and the predicted MFI for the same concentration using the 5PL:

$$A = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \left|\frac{MFI_i - pMFI_i}{pMFI_i}\right|\right)$$

Where MFI$_i$ is the measured MFI for the concentration *i*, pMFI$_i$ is the predicted MFI for the concentration *i* based on the 5PL, and N is the number of concentrations measured. The functional sensitivity of the assays (A) together with LoB and LoD values in molarity (ng/ml) are represented in Table 17.

Table 17 - Funcional sensitivity of the assays. LoB, LoD, and functional sensitivity are represented for each assay.

Assay	LoB (ng/ml)	LoD (ng/ml)	Functional sensitivity (A)
Galectin.3	0.0019	0.0044	0.9664
ICAM1	0.0075	0.0356	0.9704
MMP9	0.0046	0.0054	0.9216

Assay validation including Limit of detection (LOD) and reproducibility were performed based on the European medicines Agency EMEA/CHMP/EWP/192217/2009 guideline on bioanalytical method validation.

REFERENCES

[1] http://www.bio-rad.com/webroot/web/pdf/lsr/literature/10032257.pdf
[2] Armbruster, D.A. and Pry, T., 2008. Limit of blank, limit of detection and limit of quantitation. The clinical biochemist reviews, 29(Suppl 1), p.S49.
[3] Shrivastava, A. and Gupta, V.B., 2011. Methods for the determination of limit of detection and limit of quantitation of the analytical methods. Chronicles of young scientists, 2(1), p.21.
[4] Fong, Y., Sebestyen, K., Yu, X., Gilbert, P. and Self, S., 2013. nCal: an R package for non-linear calibration. Bioinformatics, 29(20), pp.2653-2654.
[5] Commo, F. and Bot, B.M., 2016. R package nplr n-parameter logistic regressions.