DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES
DER FAKULTÄT FÜR CHEMIE UND PHARMAZIE
DER LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

MS-based Quantitative Proteomics for Molecular Cancer Diagnostics

von

Sally Deeb

aus Dubai, Vereinigte Arabische Emirate

2014

Erklärung

Diese Dissertation wurde im Sinne von § 7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Matthias Mann betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 20.03.2015

.....

Dissertation eingereicht am 18.03.2014

1. Gutachter: Prof. Dr. Matthias Mann

2. Gutachter: Prof. Dr. Marc Schmidt-Supprian

Mündliche Prüfung am 16.04.2014

Summary

The last decade has witnessed considerable improvements in the molecular characterization of cancer. This is beginning to catalyze a shift from non-specific cytotoxic drugs towards targeted therapies. One of the fundamental prerequisites of such targeted therapeutic approaches is the ability to determine which patients are most likely to benefit from a particular agent. However, in many cancer types it is still a challenge to distinguish molecularly distinct entities. Diagnosis and classification of most cancers is still based on histological analysis of stained tissue sections or cells. The advent of modern technologies that allow global molecular profiling of individual tumors has increased the power to sub-classify cancers. The most celebrated examples are the sub-classification of breast cancers and diffuse large B-cell lymphomas (DLBCLs) using gene expression profiling.

Cancer is not a disease of single molecular aberrations, but involves dysregulation of multiple pathways governing hallmark processes such as cell death, proliferation, differentiation and migration. Mass spectrometry (MS)-based proteomics measures the functional players in a cell, the proteins. It provides a direct way to analyze signaling pathways and hallmark cancer processes. The focus of this thesis is to explore the possibility of using state-of-art proteomics to classify very closely related tumor subtypes. We selected a challenging system, the two histologically indistinguishable subtypes of diffuse large-B-cell lymphomas, the germinal center B-cell (GCB) and the activated B-cell (ABC) DLBCLs.

In the first project, the aim was to investigate whether the depth and quantitative accuracy attained with our MS-based proteomics platform were capable of distinguishing the two DLBCL subtypes. I compared the global protein expression profiles of five patient-derived ABC-DLBCL and GC-DLBCL cell lines each. For quantification, I employed the super-SILAC approach that was developed to enable accurate quantification of human tissue proteomes. The samples were analyzed using either a

fractionation approach (six fractions per proteome) and measured on a linear ion trap Orbitrap mass spectrometer or a single-run approach (no fractionation) with measurement on a quadrupole Orbitrap instrument. I achieved robust segregation of the two subtypes using both platforms. Drivers of the segregation included many proteins known to be differentially expressed between the subtypes and I identified differential NF-κB signaling which is one of the oncogenic hallmarks of ABC-DLBCL, indicating that the analysis captured the underlying biology.

In the second project I aimed at investigating the possibility of targeting the cell surface proteome to segregate the two B-cell lymphoma subtypes. I took advantage of the recently developed N-glyco-enrichment approach to target plasma membrane proteins. Using the quantitative super-SILAC approach, I was again able to segregate the cell line system, showing for the first time that it is possible to differentiate tumors based on profiles of their post-translational modifications (PTMs). Reassuringly, cell surface proteins that had been identified as markers of segregation in the first study were re-identified here. Remarkably, this analysis even pinpointed differential signaling pathways between the subtypes based on differences in membrane proteins.

The last project of the thesis evaluates our proteomics platform in the characterization of primary DLBCL patient material. Working with tissue samples is challenging because of their high complexity as well as the need to extract proteins from formalin fixed paraffin embedded (FFPE) material, in which most tumors are stored in bio banks. In this project, I reached a very good depth of more than 9,000 proteins from 20 FFPE DLBCL tissues, sufficient to segregate the subtypes and to highlight important biological differences.

In summary, classification of cancer patients into molecularly distinct subtypes has not been a straight-forward task. There is a definite need for robust and reliable tools. Using state-of-the-art technology, I have demonstrated successful proof-of-principle applications of MS-based proteomics for the classification of the difficult-to-segregate subtypes of DLBCL based on their protein and PTMs expression profiles.

Table of Contents

1. INTRODUCTION

1.1 Mass spectrometry (MS)-based clinical proteomics

Proteins are the functional units of a cell. All cellular states in the human body, whether normal or altered, are ultimately dependent on protein expression and regulation. Knowledge of protein expression and its dynamic alterations are therefore paramount to understanding the biology of health and disease. About twenty thousand human protein-coding genes have been cataloged [1]. However, splicing isoforms as well post-translational modifications expand the number of different protein products or "proteoforms" dramatically. "Proteomics" is the study of all proteins present in a cell, tissue or organism, as well as their changes under different conditions. Proteomics is thought to have great potential for clinical applications, provided the correct tools, both conceptual and technological, are available.

In-depth profiling of complex protein mixtures such as human tissues or biofluids requires high specificity, sensitivity and throughput. Several technologies have been developed for proteome analyses of clinical samples such as two-dimensional (2D) gel electrophoresis [2, 3], protein- and antibody-based microarrays [4-6], and liquid chromatography-mass spectrometry (LC-MS). Since LC-MS allows very accurate mass measurements of molecules in a sample as well as sensitive detection of variations in their composition and abundance, it has become the 'gold standard' for proteome measurements [7]. MS-based proteomics platforms offer highly sensitive analytical capabilities, relatively large dynamic range and reasonable throughput that make them in principle amenable to clinical applications. Global shotgun proteomics studies using LC-MS of human plasma/serum were published as early as 2002. [8]. Despite the difficulties imposed by the large complexity of biofluids, the less invasive nature of sample collection makes them an attractive option from a clinical perspective. The interest in identifying

disease related proteins is reflected in many publications that utilized LC-MS to study human biofluids (see for example refs. [9, 10]).

Clinical applications of MS-based approaches cover a broad range of biomedical and biological questions. The versatility of such applications is emerging in recent studies such as single cell analysis of blood cells using so called CytoTOF technology [11] and the identification of gram-negative bacilli in various clinical samples [12, 13]. Functional proteomics studies that aim at deciphering protein-protein interactions or deregulated signaling pathways involved in disease pathogenesis play an important role in understanding the underlying biology. Clearly, clinical proteomics has the potential to provide a functional understanding of diseases at the molecular level which is the key for advancing translational studies especially those related to personalized medicine. Furthermore, MS-based proteomics exclusively provides tools to investigate large-scale variations at the level of post-translational modifications (PTMs) whose role in crucial disease pathogenesis processes has become more and more evident [14]. Investigations at the level of PTMs hold great promise to provide new understandings of disease pathology. Phosphoproteomics, for instance, contributed to examining how the epidermal growth factor receptor (EGFR) and downstream signaling networks are involved in rapid cell proliferation and diffused invasion in glioblastoma [15]. In another example, MS-based quantitative phosphoproteomics was used to compare the activation of primary $CD4^+$ T cells of type 1 diabetes-prone and -resistant mice, thereby mapping signaling differences that may underlie the autoreactive phenotype of T cells against beta-cells [16]. However, biomarker development remains a primary aspiration for translational application of MS-based proteomics. Alterations in protein expression may be early indicators of disease or therapeutic targets for intervention and drug development [17]. Biomarker identification for early disease diagnosis, prognosis and targeted therapies is a broad field that may improve disease management at a personalized level. Despite this great potential, success remains modest due to a plethora of challenges, which will be further discussed in the following chapter.

1.2 Challenges in MS-based clinical proteomics

Despite being one of the most sensitive analytical methods, which can handle thousands of proteins simultaneously, MS-based proteomics has so far had a modest contribution in the clinical field especially in the area of protein biomarker discovery [18].

There are tremendous clinical benefits from the identification of biomarkers such as the early, non-invasive diagnosis of severe diseases like in the cases of C reactive protein [19] and troponin I [20] for myocardial infarction or prostate specific antigen for prostate cancer [21]. In a further step, biomarkers can classify patients with the same disease into molecular taxonomies that are clinically distinct. For instance, chronic myeloid leukemia patients with the BCR–ABL fusion gene respond to treatment with the tyrosine kinase inhibitor, Imatinib [22], whereas patients without that biomarker do not. Other applications of biomarkers include monitoring the activity of diseases as well as directing targeted therapies or assessing response to drugs. As the benefits and the importance of biomarkers are widely realized, there are large efforts both in academic and industrial settings for the identification of novel protein biomarkers. However, the outcomes of such efforts and investments have not been as successful as anticipated. In fact, few novel protein biomarkers are used in clinical practice and the average rate of introduction of new protein analytes approved by the US Food and Drug Administration has fallen to one per year since 1998 [18]. A variety reasons account for this slow rate, beginning with the long and difficult steps from candidate discovery to the establishment of a clinical assay.

1.2.1 The dynamic range challenge

The type of biological material used in biomarker studies represents the first challenge. As the single most informative tissue for assessing an individual's health, blood has been the focus of many biomarker discovery projects. Blood has been described as a circulating representation of all body tissues and of all processes whether physiological or pathological due to its direct or indirect interaction with the entire cell

complement of the body [23]. Additionally, blood is easily accessible. Despite all these favorable factors, blood is one of the most difficult biological samples to characterize due to its huge complexity as well as vast differences in protein concentrations. The dynamic range of protein concentrations in human plasma, for instance, spans over 10 orders of magnitude [24] (Figure 1). In addition, protein biomarker candidates may be present at the lower end of the plasma protein concentration range, exacerbating the problem. Such cases include ischemic heart disease and cancer where biomarkers are thought to originate from leakage or secretion from the diseased tissues and are greatly diluted in circulating blood [18]. The analytical challenges imposed by the complexity of blood or plasma counterbalance its benefits of being a comprehensive diagnostic material.

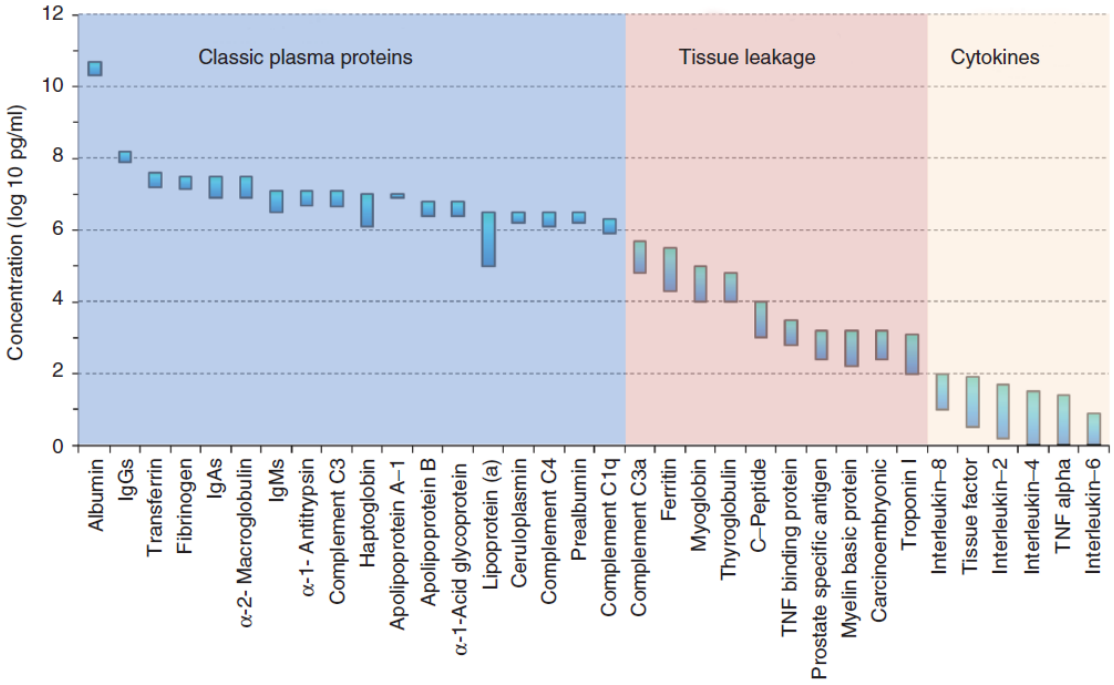


Figure 1. **Protein dynamic range in blood plasma**. The range of plasma protein abundances is illustrated using 34 proteins representing the most to least abundant. The dynamic range spans over 10 orders of magnitude. Adapted from [17].

For the aforementioned reasons, other options for biomarker discovery have gained attention. This included tissues as well as alternative biofluids such as cerebrospinal fluid [25], bronchoalveolar lavage fluid [26], saliva [27] and urine [28]. The proximity of some of these fluids to the source of disease may make them a more concentrated pool of potential biomarkers.

Techniques such as immunoaffinity capture have already been shown to be effective methods for the detection and quantification of selected protein biomarker candidates at picogram/milliliter levels in blood. Many disease biomarkers are thought to be present at this concentration level. However, while single or lowly multiplexed protein assays such as enzyme linked immune assays (ELISA) are well established in the clinic, the comprehensiveness of antibody or other protein profiling arrays makes them as yet unsuitable for real *de novo* discovery efforts. Unbiased screening requires increasing the 'content' for such arrays; a daunting task that needs intensive resources and efforts. The limitations of these affinity approaches leave MS as the principal technology for unbiased discovery of novel candidate biomarkers [18].

Several strategies compatible with MS have been developed to address the dynamic range challenge and to allow biomarker discovery from the plasma proteome [29]. For example, antibody-based depletion of abundant proteins somewhat reduces the dynamic range of plasma proteins ahead of MS analysis. However, there may be concomitant loss of low abundance proteins like cytokines [30]. Alternatively, extensive, multidimensional fractionation approaches may reduce sample complexity. Since a single blood sample expands to many more 'samples' each requiring hours of instrument analysis time this strategy severely limits sample throughput, which represents the second major challenge of clinical proteomics.

1.2.2 The throughput challenge

The extent of human and disease variability is one of the most pressing challenges in clinical studies. The biomarker discovery pipeline (see Box for a summary on the biomarker pipeline), for instance, is a multi-phase approach that in the first two phases - discovery and qualification - requires samples in the order of 10s. However, the later phases of verification and validation often require large scale projects that may necessitate patient samples in the order of 100s and even 1000s [18]. These large cohorts are necessary to deal with normal human genetic heterogeneity as well as disease heterogeneity. Statistical significance is achieved only with high throughput measurements, especially if effects are small. The need for such large cohorts also adds the complication of establishing standardized protocols for the collection and storage of the samples [17]. The quality of the samples collected has turned out to be a crucial factor in determining the outcome of biomarker studies and should be addressed carefully.

In the discovery phase, MS is currently the main qualifying technology for unbiased comprehensive screening of differential protein abundance between different states. What qualifies a protein to be a candidate biomarker is its consistent differential expression profile (abundance) between two states which in this case would be health and disease. At the other end of the pipeline, where large cohorts of patient samples need to be analyzed, targeted approaches that allow the monitoring of a sufficient number of candidates have been recommended [18]. The development of high-quality protein antibody assays is tedious, expensive and far from being straightforward. Such limitations would seem to argue for targeted MS-based platforms. However, a global MS-based approach capable of reaching the depth required to monitor interesting candidates would be even more ideal. This is because there is much effort involved in the optimization for each candidate in the targeted approach and the fact that by its nature it ignores most of the information in the proteome.

The Biomarker Pipeline

The biomarker identification pipeline is a multi-phased approach, summarized here based on the guidelines provided by Rifai *et al.* in 2006. It begins with a discovery phase that is usually an unbiased, semi-quantitative process by which the differentially expressed proteins are identified. In this phase, model systems such as cell lines, mouse models as well as a variety of human biological materials can be employed. The result is often a list of candidates with a high false discovery rate, especially when not employing state of the art technology. This is due to low sequencing frequency of low abundant proteins and the fact that the signal at the lower bounds of the dynamic range may randomly exceed detection limits, resulting in artificial differences.

The second phase (qualification) aims at checking if the differential expression of the candidate biomarkers is observable using other methods like targeted approaches. It ideally also confirms their differential expression in blood plasma in case the discovery phase was performed using other model systems. These first two phases (discovery and qualification) are focused on confirming the candidate biomarker's sensitivity, which is the likelihood that a disease sample will test positive.

The third phase which is the verification phase aims to capture all kinds of variations whether they are genetic, environmental, biological or stochastic. Large numbers of human plasma samples (100s) are analyzed. This phase is concerned with the specificity of the candidate, which is the likelihood that an unaffected sample will test negative. The few candidate biomarkers that pass the verification phase move to the validation phase for assay optimization. An immunoassay may have to be tested against thousands of samples that fully reflect the targeted population's variations. Finally, the immunoassay is refined to meet the standards of clinical tests.

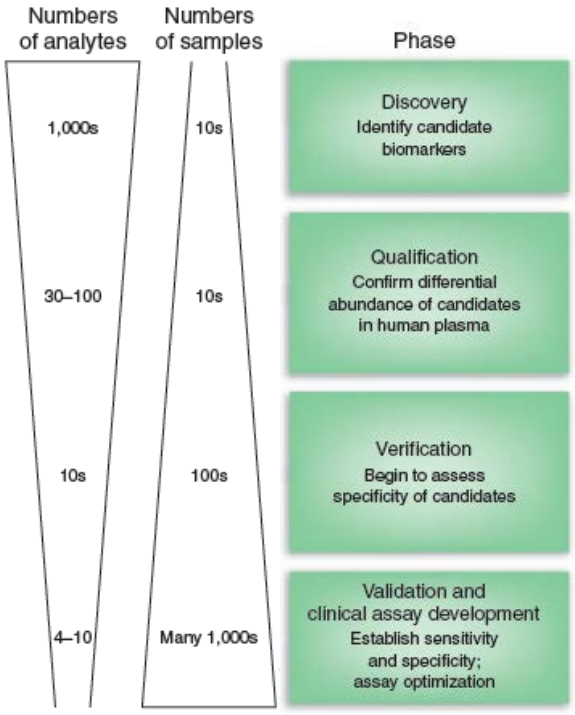


Figure 2. **Phases of the biomarker identification pipeline.** Biomarker identification requires at least four phases where the number of candidate analytes or proteins is reduced along the process. Adapted from [18].

As mentioned above, the outcome of discovering and validating biomarkers by MS-based methods has been disappointing. One contributing factor may have been premature application of the technology when it was not ready to answer challenging

clinical questions. The last decade has witnessed exciting developments in the MS field at all levels, starting from sample preparation to instrumentation to data acquisition and analysis. The field has also witnessed some paradigm shifts regarding the approaches used in clinical proteomics after researchers realized the types of challenges that need to be met. In the discovery phase of biomarkers, for instance, it has been suggested to substitute plasma samples with proximal fluids, which are biofluids in close proximity to the site of the disease. Potential biomarkers are likely to be enriched in proximal fluids before they are diluted in the blood. In ovarian cancer, for example, it has been shown that marker concentration is higher in ovarian cyst fluid and ascites fluid compared to plasma [31]. Other disease model systems such as cell lines or genetically homogenous animals may provide an attractive alternative that dampens the noise from genetic and environmental variation in discovery phase projects [18]. In addition, the concept of identifying the single 'magic' marker is not the ultimate goal anymore. There is growing consensus that panels of markers will be required for most applications and they appear to have several advantages [32, 33]. Other attractive approaches, which are also gaining momentum, take advantage of the large biochemical diversity of the proteome and aim for enriching specific classes of proteins of diagnostic and therapeutic value for the disease in question. This includes enriching phosphoproteins to study disease-specific deregulated signaling pathways, or glycosylated proteins to explore disease-specific changes in the cell surface proteome.

The largest developments in the proteomics field have been at the instrument level. Major performance enhancements in speed and accuracy of mass spectrometers now allow almost complete coverage of complex biological samples such as human cells [34]. Unprecedented depths of complex proteomes are even attained using single shot measurements where by definition no fractionation is required [35, 36]. Such leading-edge developments in the field of MS-based quantitative proteomics will be discussed systematically in the next chapter.

1.3 Developments in MS-based quantitative proteomics

A mass spectrometer can be used to analyze proteins either in their intact form (top-down proteomics) or as digested peptides (bottom-up proteomics). A top-down approach can in principle fully sequence the protein of interest, theoretically allowing the characterization of all its corresponding isoforms as well as co-occurring post-translational modifications (PTMs). However, this approach is not widespread due to several limitations. Intact proteins produce multiply charged ions which results in difficult to deconvolve, highly complex MS/MS spectra. In addition, the difficulty to handle intact proteins, especially insoluble ones, makes pre-MS separation techniques for reducing complexity challenging in top-down approaches. Peptides often have better ionization efficiencies, produce less complex MS spectra and easier to interpret fragmentation spectra. Therefore bottom-up approaches are much more popular in a wide range of applications encompassing both simple and complex samples. In case of complex mixtures, after enzymatic digestion, the sample is often first separated using chromatographic techniques such as reverse phase or ion exchange as well as other techniques such as isoelectric focusing. Identification of peptide sequences relies on the information in the fragmentation spectra that result from fragmenting the isolated peptide ions in the mass analyzer. The interpretation is usually not done manually but instead through an automatic search against a database containing theoretical fragmentation spectra. A step-wise representation of a bottom up workflow is represented in Figure 3. There are several advantages associated with a bottom-up approach like the possibility of coupling separation techniques ahead of the MS in an automated manner. In addition, a variety of quantification approaches and powerful software have been developed, making the approach applicable to a wide variety of biological questions. The main drawback of the bottom-up approach is the incomplete sequence coverage of the proteins identified. The fact that only a part of the protein is covered by the sequenced peptides makes it difficult to distinguish protein isoforms or PTMs in regions of the protein that are not covered.

One of the major developments that allowed MS to gain momentum in its application for biomolecules was the introduction of two soft ionization techniques: electrospray ionization (ESI) [37] and matrix assisted laser desorption ionization (MALDI) [38]. Previous methods were energetically damaging for peptides and other biomolecules which made it impossible to vaporize and ionize them in an intact form, in contrast to the gentle methods. John B Fenn and Koichi Tanaka were awarded a share of the noble prize in Chemistry in the year 2002 for their work on developing ionization methods for large biomolecules. Further developments in ESI were pursued by Matthias Wilm and Matthias Mann, who showed that the flow rate in ESI can be reduced to the nanoliter range without loss of signal (they reduced it from 2-10 ul/min to 20 nl/min) [39]. This resulted in improved measurement sensitivity (attomole range). Furthermore, since ESI ionizes peptides out of a solution, it can be coupled to liquid based chromatographic separation. For these reasons, ESI has become the method of choice for the analysis of complex protein mixtures. In contrast, MALDI generates mainly single charged ions which results in inefficient fragmentation and consequently insufficient peaks for peptide identification.

Bottom-up MS-based proteomics workflows in which peptides are initially separated in an online mode using high performance liquid chromatography (HPLC) and then electrosprayed directly into the MS have become the mainstream workflow for highly complex biological samples (Figure 3). A plethora of sample preparation techniques, types of instruments and data analysis software are available today. Each step in the workflow is crucial in determining the quality of the outcome, which for clinical applications needs to be of a very high standard.

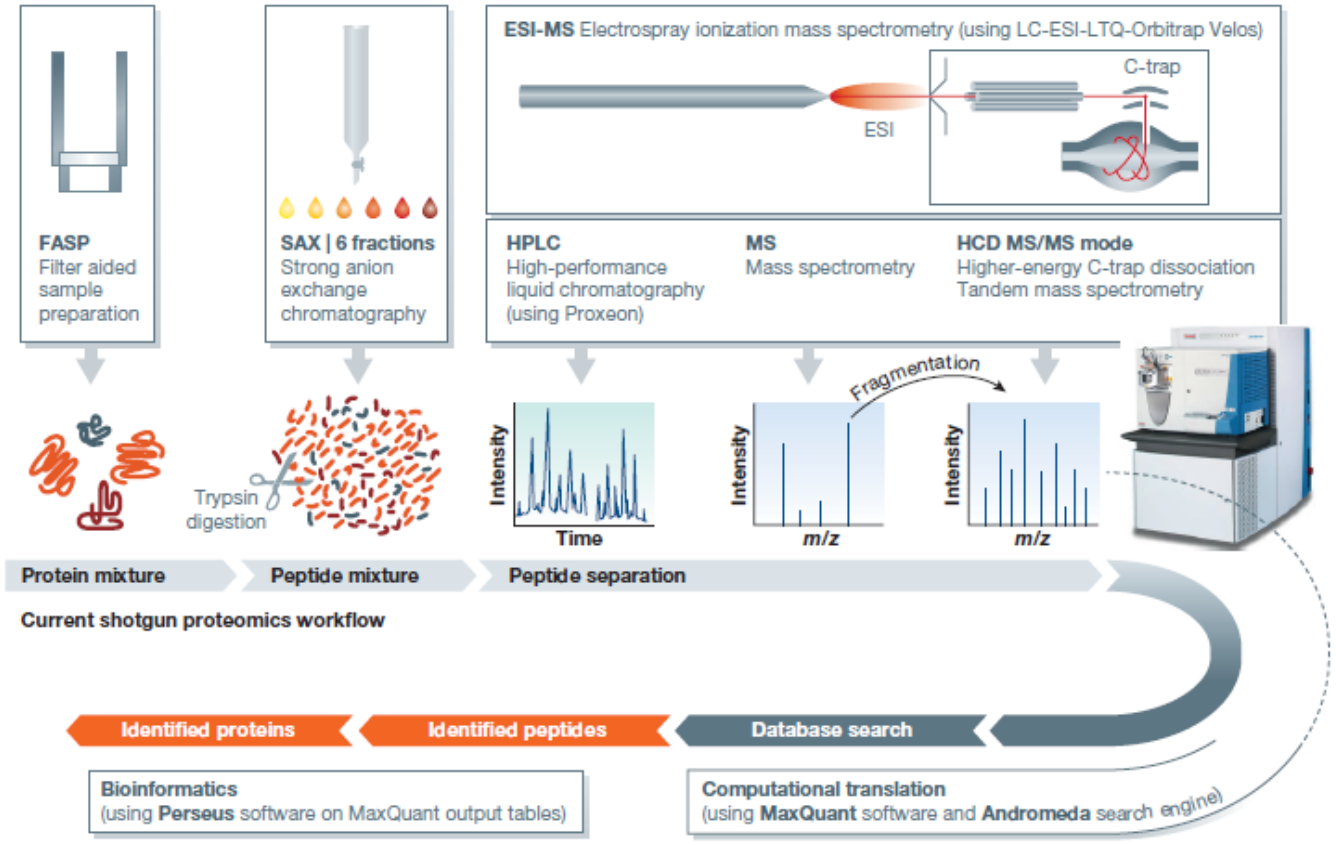


Figure 3. **Example of a bottom-up proteomics workflow**. Proteins extracted from biological samples such as tissues or cells are digested into peptides using sequence-specific enzymes such as trypsin. Fractionation steps may be applied at the protein or peptide level to enhance the coverage and dynamic range. Enrichment for specific post-translational modifications may also be performed using specialized approaches. The resulting peptides are separated by high-performance liquid chromatography (HPLC) and electrosprayed into the mass spectrometer. The peptides are measured in a data-dependent acquisition mode: after a full scan, a preselected number of the most intense ions are selected for fragmentation and corresponding MS/MS spectra are generated. In the computational section of the workflow, a database search is performed for peptide identification and protein inference. Regulated proteins are determined using different quantitative strategies followed by statistical analysis. From [40].

1.3.1 Sample preparation

There are two commonly used sample preparation approaches for the conversion of proteins extracted from biological material into peptides compatible with MS analysis. The first approach relies on solubilization of the proteins using denaturing detergents followed by their separation and subsequent digestion in a polyacrylamide gel ('in-gel' digestion) [41]. This approach is robust and results in high purity peptides. However, it is time-consuming and the recovery of the peptides from the gel can be poor. Today it is mainly used for the identification of gel-separated proteins visualized in single bands. The second approach dispenses with gel-separation, uses strong chaotropic reagents such as urea or thiourea to solubilize the proteins followed by 'in-solution' digestion. It is straight-forward and more amenable to automation but not all proteins may be solubilized and impurities present in solution may hinder the digestion. A more recently developed method by Wisniewski *et al.* [42] combines benefits of the two approaches. It takes advantage of the possibility to exchange buffers when using an ultrafiltration device, and was hence termed filter-aided sample preparation (FASP) [42]. Very strong detergents such as sodium dodecyl sulfate (SDS) are used, which allow total solubilization of cells and tissues. This is followed by exchanging the SDS-based buffer with an 8 M urea buffer after loading the sample onto the filter unit. The workflow is easy to handle and results in pure peptides in a short period of time. This method is particularly advantageous in cases where hydrophobic proteins are targeted such as in the case of N-glycosylated proteins, which are generally localized to the plasma membrane.

Although the ideal sample preparation technique would involve minimum sample handling, the large diversity and complexity of the proteome make it indispensable to apply fractionation techniques to reach the depth required in some cases. The separation techniques are said to be "online" if they are directly coupled to the mass spectrometer. They can be performed at the protein or peptide level. In complex proteomes with large dynamic range, extensive fractionation approaches are often applied to achieve global and

in-depth coverage of the sample. In other applications, a particular class of proteins may be of interest. Enrichment strategies based on the proteins biological properties can then be applied, allowing in-depth coverage of this specific class. However, due to the enormous technological advancements specifically in MS instrumentation, the depth reached with single shot approaches is already sufficient to answer many critical biological questions (see below).

1.3.1.1 Main proteomic separation and fractionation approaches

The most widely used separation technique in MS-based proteomics is reversed phase high performance liquid chromatography (RP-HPLC). The online coupling of RP-HPLC to MS, typically by an electrospray interface, is called LC-MS or LC-MS/MS. However, HPLC can also be used for fractionation in an offline mode, for example, when MALDI is applied for peptide ionization.

At the protein level, common separation strategies are based on protein size, such as size exclusion chromatography (SEC) and sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). In the latter, the charged, denatured proteins are separated in a polyacrylamide gel after applying an electric field. In the above mentioned 'in gel' digestion method the gel is cut into slices, the proteins are digested using an endoproteinase such as trypsin, followed by extraction of the peptides [43]. Adding a second dimension of separation based on the isoelectric point of the proteins is the principle of 2-dimensional (2-D) gel electrophoresis [3]. Although, it was originally thought to be a promising technique for proteomics, it has proven unsuitable to detect, identify, and quantify large numbers of proteins in a sample [44]. Furthermore it has several limitations including low resolution, low reproducibility, a narrow dynamic range, and bias against hydrophobic proteins.

At the peptide level, fractionation is commonly performed by isoelectric focusing, ion exchange chromatography or affinity chromatography. The principle of ion exchange chromatography is employed in so-called StageTips [45], which allow clean up and fast

and sensitive fractionation in a simple device. In this approach, peptides are separated on an anion (or cation) exchanger that is assembled following the StageTip principle by stacking six layers of anion (cation) exchange disks into a micropipette tip. The number of elution steps from the column determines the resolution of the separation. Fractionating FASP-eluted peptides with a six fractions elution protocol from a StageTip-based anion exchanger already allows in-depth characterization of a complex mammalian proteome [46, 47].

1.3.1.2 Enrichment approaches

Many biological questions do not require global in-depth analysis of complete proteomes. In some cases, only a certain class of proteins with certain biological properties is of interest. Employing these biological properties often allows the development of enrichment protocols that target this particular protein class and, at the same time, reduces the complexity of the sample. An example is the enrichment of glycoproteins to explore the cell surface or of phosphopeptides to study signaling pathways. Such strategies could be based on antibodies, ionic interactions or affinity ligands (Figure 4). In addition to post-translational modifications, enrichment techniques can cover a range of biological properties such as cellular localization, specific protein-protein interactions, DNA/RNA–protein interactions.

It has become increasingly clear that cells extensively use PTMs as molecular switches for signal propagation, regulating diverse cellular aspects [48]. The global analysis of PTM sites is an exclusive domain of MS-based proteomics [14] and this requires the development of robust enrichment strategies for the PTM of interest. MS-based workflows have mainly addressed the PTMs with the greatest biological interest such as phosphorylation and glycosylation. For phosphopeptides, metal affinity chromatography using titanium dioxide [49] and/or anti-phosphotyrosine antibodies [50] are frequently employed. A mixture of lectins is commonly used for the enrichment of glycopeptides [51] . The specificity and enrichment obtained with these approaches are generally quite high [14]. This has allowed large scale studies on the role of PTMs in

important processes such as EGF signaling [52, 53] and the cell cycle [54]. The large interest in the biology of the phosphoproteome and the glycoproteome, as well as several other PTMs, has driven the development of their corresponding enrichment strategies. Over 300 different modifications have been reported to occur physiologically [55]. However, for most of these, enrichment strategies do not exist at all. The more powerful the enrichment strategies for such PTMs, the more we are able to understand the scope of their contribution to various biological processes. The ability to explore the ubiquitinome is a recent example: here the introduction of diglycine-specific antibodies now allows capture of the remnant modification following tryptic digestion of ubiquitinated proteins, making it possible to specifically enrich and analyze tens of thousands of ubiquitination sites [56, 57].

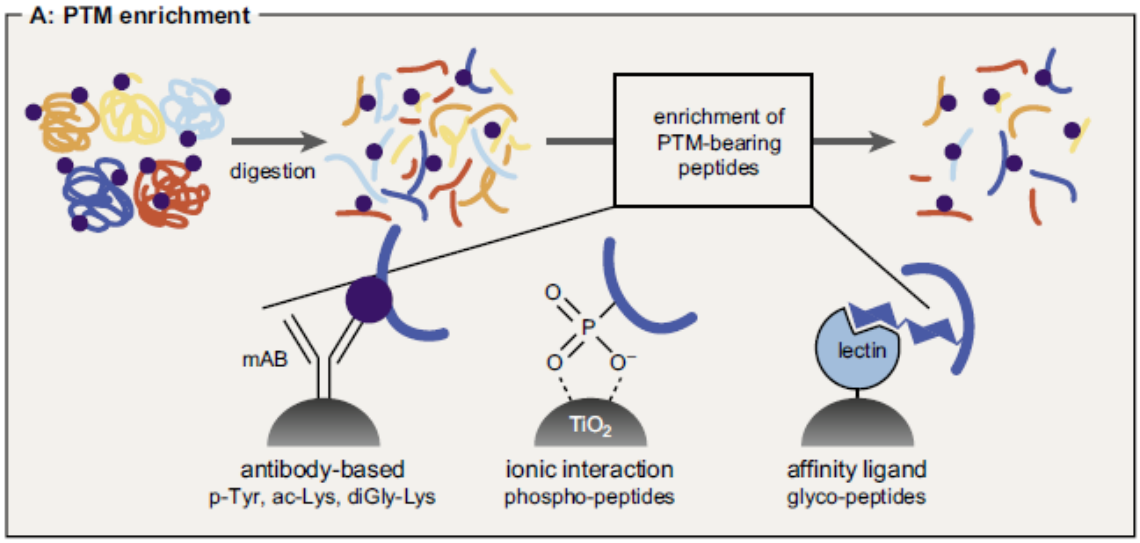


Figure 4. **Enrichment of PTMs**. PTM-bearing peptides or proteins are enriched using different strategies (antibody-based, ionic, and affinity-based interactions). Adapted from [48].

Compared to the proteome, the investigation of PTMs is more difficult both conceptually and technologically. When analyzing the whole proteome, the characterization of each protein is almost always based on several peptides. In contrast, when PTMs are analyzed, each peptide with the modification of interest stands on its own. Modified peptides are generally of low abundance and combined with their often

difficult to interpret fragmentation spectra, this makes identification more challenging. Localization of the modification with single amino acid resolution is another crucial piece of information to be extracted from the fragmentation spectra. Furthermore, modified peptides are normally present in lower amounts compared to their unmodified counterparts. Indeed, it is very beneficial to determine the sites' stoichiometry or occupancy to shed light on the possible functionality of these modifications in cells in specific conditions. The determination of phosphorylation stoichiometry on a large scale has recently been reported [54, 58].

All these difficulties, in addition to the extra sample handling steps involved, make PTM analysis a challenging task, especially when working with clinical samples. In the long-run it is conceivable that highly improved MS performance could make enrichment steps superfluous [14]. In any case, similar to the proteome, further simplification of the workflow will result in a more robust, applicable technology with great potential for an impact in the clinic.

1.3.1.3 Single-shot approaches

Fractionation and enrichment are particularly necessary and informative in cases such as organellar proteomics or low abundant PTMs. In the context of whole-proteome measurements they are typically applied with the aim of solving the dynamic range problem. With modern spectrometers, such approaches have diminishing returns. Even upon extensive fractionation, the relative abundance of proteins is altered by only a factor of 10 to 100, because proteins fractionation is rarely perfect. Hence, when using modern mass spectrometers, the de-enriched proteins are still easily detected. The increased number of sub-samples to be analyzed increases the total measuring time and limits the overall sensitivity [36]. Instead, other aspects of the workflow have proven to be more amenable to improvements. Specifically, these are the online chromatography setup preceding peptide analysis in the mass spectrometer and the mass spectrometers themselves. The chromatography setup is being pushed to its limits by employing high-pressure HPLC pumps, very small bead particles as column material, and relatively long

columns and gradients [59, 60]. These drastic measures have yielded high peptide-separation capacity and greatly increased the number of eluting peptides that the mass spectrometer can isolate and fragment. Remarkably, the combined power of such LC systems along with the most modern spectrometers (discussed in the following section) have now made it possible to reach an almost complete coverage of the yeast proteome with no upfront protein or peptide separation ("single-shot analysis") [35].

Two recent studies have attempted to characterize the human cell line proteome comprehensively, using extensive fractionation strategies [34, 61]. Both studies showed that at least 10,000 different protein coding loci are expressed in a typical human cancer cell line. Judged against the coverage of macromolecular complexes and pathways achieved, these numbers approach complete coverage, which may be attained at about 12,000 proteins [34]. In a more recent study to assess the single-shot approach, 11 mammalian cell lines were measured in single 4 h gradients. This resulted in the identification of around 8,000 proteins in each one of them, with a dynamic range exceeding 6 orders of magnitude [36]. Different protein abundance ranges showed enrichment for different biological functions. Clearly the single shot approach is capable of capturing a very large percentage of the mammalian proteome.

In a clinical context, the two main biological sample types that are of general interest are plasma and formalin-fixed paraffin-embedded tissues (FFPE). So far, there are few in-depth studies of human tissues. In a recent colon cancer study, around 7,500 proteins were identified across patient matched normal mucosa, primary carcinoma, and nodal metastases [62]. FFPE is the main form in which patient samples are stored in tissue banks. We have recently shown that proteins can readily be extracted from FEPE for both global proteomic and PTM studies [63]. This is crucial for clinical proteomics workflows when FFPE tissues are the material of interest and represents a clear advantage over other technologies that require RNA extraction from these samples, which is difficult and which may lead to low quality results.

Tissues have complex compositions, and answering clinical questions requires the analyses of many patient samples. Targeted approaches, in which the mass spectrometer is 'fed' a list of pre-defined peptides and its corresponding fragments, have so far been favored for post-discovery applications in biomarker discovery pipelines [18]. However, the approach is not suitable for discovery phase projects and requires reaching acceptable levels of specificity through extensive method optimization and process control. The single-shot approach combines advantages of targeted and shotgun approaches. Its sample and measuring time requirements are low but it still retrains the advantage of being an unbiased method. This may make it amenable to both discovery and validation steps of the biomarker pipeline.

1.3.2 Mass spectrometry instrumentation

There have been major developments in the hardware of mass spectrometers over the last decade and these have produced much faster, more accurate and more sensitive machines. The time to execute a basic measurement cycle, which typically consists of one survey mass spectrum followed by fragmentation spectra of the 10 most intense eluting peptides (known as "top 10" method), is currently around one second [64, 65]. The increased sensitivity of today's instruments is evident in their higher dynamic range, the ability to detect low abundance species in the presence of highly abundant ones. They have a high resolving power, allowing co-eluting peptides of similar mass to be distinguished and are therefore a prerequisite of their accurate quantification. Furthermore, high resolution makes it possible to achieve mass accuracies in the ppm and even sub-ppm range [66]. This has important benefits on the certainty of peptide identifications.

The three basic elements of a mass-spectrometer are the (1) source of ions, (2) mass analyzer, and (3) detector. The ion source ionizes the analytes and transfers them into the gas phase. This allows the generated ions to fly in the mass spectrometer, guided by a series of electric potential differences and radio-frequency (RF) fields until they reach the mass analyzer, where they are separated and analyzed. Different mass analyzers apply

different principles of separation and analysis but the basic concept of all of them is that the motion of an ion is dependent on its mass-to-charge ratio (m/z) under the effect of magnetic and/or electric fields. A detector then measures the signal and amplifies it. The acquired data is processed and eventually represented as a spectrum where the x-axis is the m/z of the ions and the y-axis corresponds to their relative intensity. In a simple view, the x-axis characterizes the peptide and the y-axis provides its abundance under the conditions being studied.

1.3.2.1 Mass analyzers

The mass analyzer is the core of a mass spectrometer. Its basic characteristics are mass range, mass accuracy, resolution and sensitivity. Further important aspects are its dynamic range, analysis speed, ion transmission and fragmentation capabilities. There is a variety of mass analyzers that apply different principles and that have different characteristics. Modern instrumentation combines the benefits of at least two mass analyzers in one platform - so-called hybrid instruments (here defined as the combination of two mass analyzers that could in principle be used independently). The most generally used analyzers are the linear ion trap, Orbitrap analyzer and TOF mass analyzers. Two common hybrid platforms are the quadrupole time-of-flight (TOF) and the linear ion trap or quadrupole – Orbitrap (LTQ-Orbitrap or Q Exactive) instruments. The latter were used exclusively in this thesis and, hence, will be discussed below.

Linear ion trap – Previously, three dimensional ion traps were widespread in proteomics but today, so called linear or two-dimensional (2-D) quadrupole ion traps have now taken over, because of their much higher ion capacity. It is a highly versatile mass analyzer capable of isolating, storing and fragmenting ions. In common with most other ion traps it measures the mass-to-charge ratio based on the stability of ion trajectory in oscillating electric fields. It consists of four hyperbolic rods. Each of them is split into three axial sections, allowing spatially separated manipulation of ion populations. The rod sections are electrically isolated with a discrete DC level generating an axial trapping field. Opposing rods are paired and an anti-phasic radio frequency (RF) voltage is applied

to the rod pairs generating a radial trapping field. The combination of axial and radial trapping results in the ions arranging themselves in a linear string. By applying dynamic field to the trapped ion population particular m/z values can be isolated or activated for fragmentation. A supplemental AC voltage is applied across one of the rod pairs (marked by X in Figure 5) for ejection of ions [67]. These X rods have a slit for the ejection of the ions and their subsequent detection by electron multipliers (Figure 5).

Applying these principles, in a full scan, all of the ions are collected and then ejected. In an MS/MS scan, isolation of a single ion is performed by resonance ion ejection, i.e. all ions with mass-to-charge ratios higher and lower than the ion of interest are ejected from the trap. To reduce the motion and dispersion of ions, helium is present at a low pressure in the trap as a dampening gas. When the axial AC voltage is applied, ion trajectories are amplified, leading to more frequent collisions with the helium molecules and to eventual fragmentation, a process called collision induced dissociation (CID). Linear ion traps are characterized by their high sensitivity and high sequencing speed. Compared to high resolution mass analyzers, however, their mass accuracy and mass resolution are relatively low. In hybrid platforms, linear ion traps are often used for fragmentation because of their fast cycle times and the smaller number of ions required for MS/MS in that device.

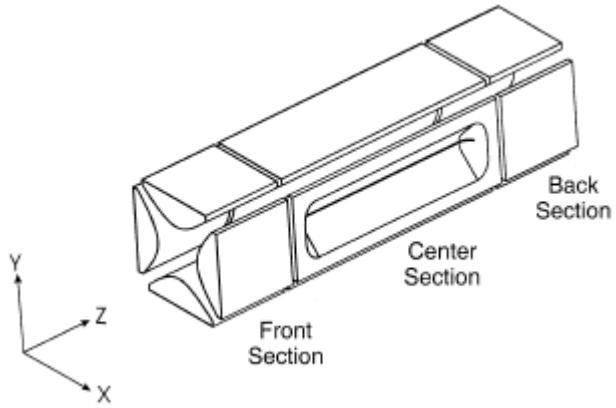


Figure 5. **Basic design of a linear ion trap.** A linear ion trap consists of four hyperbolic rods each divided into three axial sections. Adapted from [67].

Orbitrap - The Orbitrap is the most recently developed mass analyzer, and remarkably it uses an entirely new principle. It was presented in the year 2000 by Alexander Makarov [68] and has quickly become an indispensable tool in the proteomics field. Unlike the linear ion trap, which uses oscillating electric fields, it employs electrostatic fields to trap and measure ions. The Orbitrap cell consists of a central spindle-like electrode surrounded by an outer barrel-like electrode split into two halves isolated by an insulating ceramic ring (Figure 6). The electrostatic field that ions experience inside the Orbitrap forces them to move in stable trajectories, combining orbital movement around the central electrode with axial oscillations along the z-axis resulting in an intricate spiral.

One of the major challenges in the development of Orbitrap instruments was the process of ion injection into the cell. A linear ion trap was first employed but this was later replaced by a more efficient curved RF-only quadrupole which is in the shape of the letter "C" (*C-trap*). The ions are injected from the C-trap as a focused package by high voltages and are then accelerated to high kinetic energies before entering the Orbitrap cell [69]. The ions are injected off-center to the outer electrode through a small aperture. The oscillation in radial direction is the result of the electrostatic attraction towards the central electrode which is counter-balanced by the centrifugal force that arises from the initial tangential velocity of the ions. The axial oscillation is controlled by the unique shape of the trap in the z-direction. Applying an electric field to each side generates an axial electrostatic field with purely harmonic potential. This axial field is zero at the equator plate and increases with the distance from the center. Importantly, the frequency of the harmonic axial oscillations is completely independent of the initial energy and spatial spread of ions. It is only dependent on the m/z ratio and thus, it is used to derive them. The axial frequencies are measured by the acquisition of the image current transients detected on the split outer electrodes. The generated time-domain signal (transients) can be converted into mass-to-charge spectrum using a fast Fourier transform algorithm [70].

The Orbitrap analyzer is characterized by its high resolution (up to at least 150,000), high mass accuracy (routinely better than 2 to 5 ppm) and a large dynamic range (greater than 10^3) [70, 71]. The mass accuracy can even be improved to the sub-ppm range by using ambient air ions for real-time recalibration [72] or software-based recalibration [73]. The benefits of high mass accuracy are particularly prominent for the accurate quantification of low abundant proteins in complex biological samples such as tissues or plasma. Because the Orbitrap is not capable of ion fragmentation, it is coupled to ion selection and fragmentation devices. These platforms are having a tremendous impact in proteomics research.

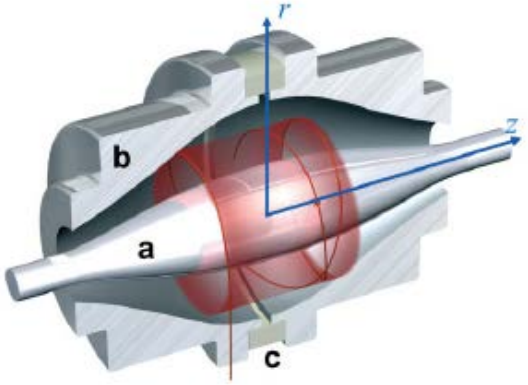


Figure 6. **Design of the Orbitrap.** A cross-sectional scheme of an Orbitrap cell shows blue arrows indicating radial (r) and axial (z) directions. The red arrow indicates ion movement. The cell consists of a spindle-shaped central electrode (a) surrounded by two halves of the outer electrode (b) which are electrically isolated by a ceramic ring (c). Adapted from [74].

1.3.2.2 The Orbitrap family of mass spectrometers

Currently there are five different instruments types equipped with the Orbitrap cell. Three of them are hybrid configurations where high resolution Orbitraps analyzers are combined with low resolution linear traps. These are the LTQ Orbitrap, Orbitrap Velos and Orbitrap Elite. The remaining two instruments are the Exactive and the Q Exactive which are benchtop instruments where the Orbitrap cell is the sole mass analyzer.

The LTQ Orbitrap was the first instrument to incorporate the Orbitrap mass analyzer [69]. It is a hybrid instrument where the Orbitrap records survey or full scans across a broad mass range of precursor ions, and a linear ion trap rapidly acquires fragment or MS/MS spectra (Figure 7A). In a recording cycle, the ions are first guided through the ion optics and the linear ion trap to the C-trap where they are accumulated. A compacted package of ions is then transferred from the C-trap to the Orbitrap for the recording of the high resolution survey scan. In a data dependent acquisition mode, the N (usually five or ten) most abundant ions from the full scan are chosen for fragmentation from the survey scan (TopN methods). Selected peptide precursor ions are sequentially isolated and subjected to fragmentation by CID. Fragments can be measured in the linear trap, which occurs in parallel to the acquisition of the high-resolution full scan of precursor masses in the Orbitrap. Based on the difference in resolution between the full and fragmentation scans, this mode of operation is referred to as the high-low strategy. Alternatively, the fragments can be passed on to the Orbitrap analyzer. Since this results in high resolution for both precursors and fragments it is called high-high strategy. An upgrade of this instrument (Orbitrap XL) contains a dedicated collision cell at the far side of the C-trap. In this cell, ions are dissociated with higher energies compared to the ion trap and this mode is called higher energy C-trap fragmentation (HCD) [75].

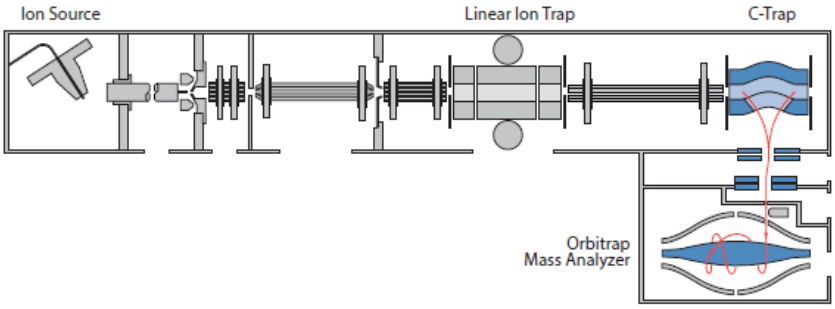
The second instrument was the Orbitrap Velos [76], which has the same basic design as the LTQ Orbitrap. However, major modifications were introduced at the front end (Figure 7B). An S-lens now replaces the tube lens allowing up to ten-fold better transmission of ions into the instrument, thus enhancing sensitivity. There is a dual linear ion trap instead of a single one. They are placed consecutively where the first one operates at a higher pressure. This allows very efficient trapping, isolation and fragmentation of ions in the first trap. The second linear trap is operated at lower pressure allowing faster acquisition of mass spectra.

Although the Orbitrap XL was already capable of performing HCD fragmentation, it was not until the introduction of the Orbitrap Velos that the high-high strategy actually

became feasible. In the Orbitrap XL, ion transfer to the HCD cell was inefficient and a large number of ions had to be accumulated. The drastically improved sensitivity and speed of the Orbitrap Velos have made HCD fragmentation and the high-high strategy feasible for standard proteomics workflows. The high resolution, high accuracy measurements of both full scan and fragmentation scans in the Orbitrap has great benefits in increased confidence in the matching of spectra, and thus peptide identifications.

The third version of this hybrid instruments family is the Orbitrap Elite [77] whose major improvement has been in the Orbitrap analyzer itself. The inner diameter of the outer electrode was reduced from 30 to 20 mm. This compact, high field Orbitrap analyzer increases resolving power twofold. An enhanced Fourier transform algorithm further doubles the resolving power. Besides its obvious benefits to the bottom up approach this high resolution is important for resolving the different charge states of intact proteins in top-down approaches.

A) LTQ Orbitrap



B) Orbitrap Velos

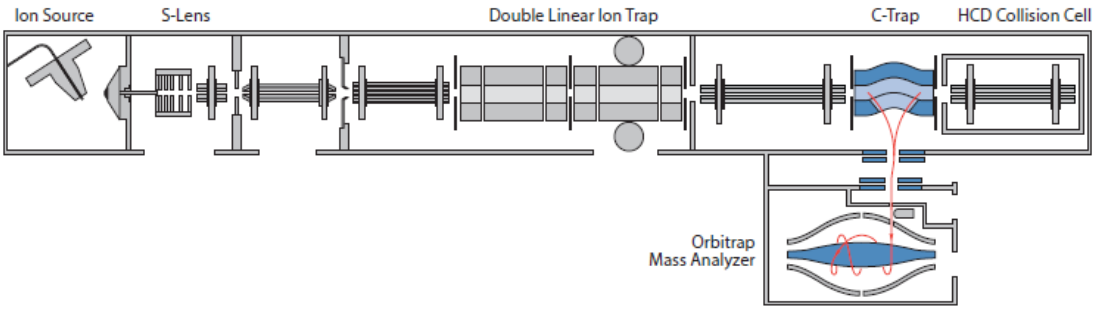


Figure 7. **Hybrid instruments in the Orbitrap family.** A) The LTQ Orbitrap was the first hybrid instrument containing the Orbitrap mass analyzer. Adapted from [72]. B) The Orbitrap Velos was the second hybrid instrument with major improvements in the front end and linear ion trap. Adapted from [76].

The latest members of the Orbitrap family are the Exactive and the Q Exactive. The Exactive consists solely of an Orbitrap analyzer and therefore can only perform MS scans or non-mass selective fragmentation of the entire mass range (all ion fragmentation) [78]. It was developed for small molecule applications [79]. The Q Exactive adds an additional quadrupole to this basic design [65] (Figure 8). This enables ion selection and isolation to perform data dependent acquisition. The Orbitrap is the only mass analyzer and therefore both the full and MS/MS scans are measured with high resolution (similar to HCD experiments on an Orbitrap Velos). The shorter ion path, the lack of a linear ion trap, parallel fragmentation and read out of the fragments and further developments in software and electronics greatly improved the sensitivity and speed of the Q Exactive. Due its simple design the Q Exactive is a benchtop instrument; this is an important step towards making mass spectrometry a more readily available technology to answer clinical and biological questions.

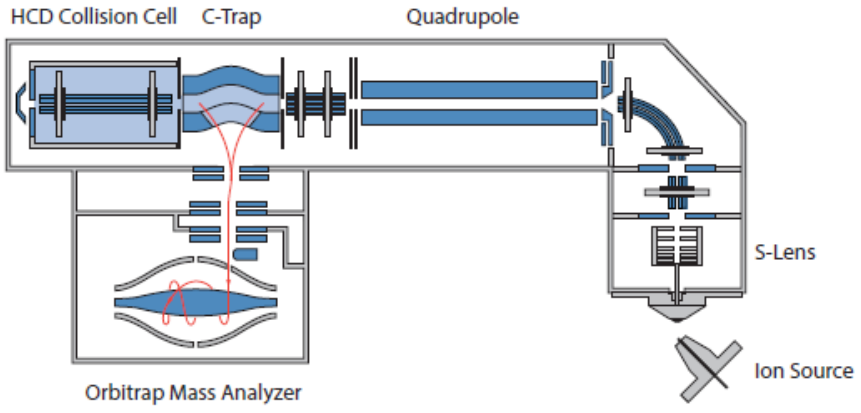


Figure 8. **The Q Exactive mass spectrometer.** The Q Exactive is a benchtop instrument that contains the Orbitrap cell as the sole mass analyzer. From [65].

1.3.3 Quantification strategies

It has now become very clear that for answering almost any biological and clinical question, it is insufficient to determine the mere presence or absence of a protein. Instead, quantitative information about protein changes is necessary to study the effect of perturbations in a biological system. Because of different peptide properties, such as their ionizability, peptide concentrations cannot be inferred directly from the signal intensities of peptide ions as they are recorded by a mass spectrometer. This has led to the development of a plethora of strategies to complement MS giving rise to quantitative proteomics.

Quantitative measurements can be either absolute or relative. Absolute quantification is the measurement of the absolute amount of a protein in a specific sample and is usually given as the concentration of the protein or its copy number per cell. Relative quantification is the measurement of the relative change in protein amount between states and is generally given as a fold change. Whether absolute or relative, quantification strategies can be divided into two major categories. The first one introduces stable isotopes to generate a mass difference between two samples (label-based quantification). The second strategy, label-free quantification, dispenses with this step and is therefore more straight-forward (Figure 9).

Label-based quantification – The basic principle underlying stable isotope labeling is that, except for their mass, the physical and chemical properties of a labeled peptide remain the same as for the natural peptide. Therefore, they have the same behavior during chromatographic separation and during MS analysis. The differences in their masses are easily resolved in modern mass spectrometer. Consequently, relative or absolute quantification of a peptide or protein can be performed by comparing intensities between natural and labeled peptide forms in the same sample. Quantification can be performed at the MS and the MS/MS level. Commonly used heavy isotopes are those of carbon (^{13}C), nitrogen (^{15}N) and oxygen (^{18}O). These labels are almost always introduced chemically or metabolically.

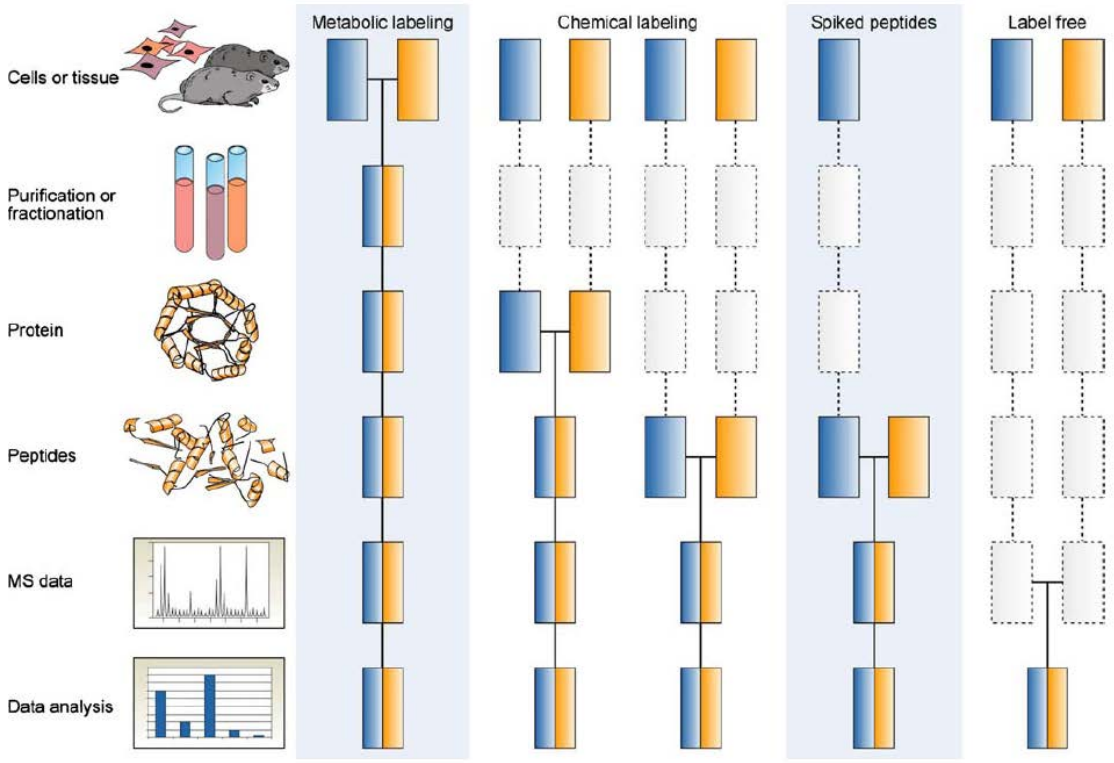


Figure 9. **Quantitative proteomics strategies.** In labeling strategies, peptides derived from different biological samples are labeled with isotopes. In metabolic labeling, samples can be pooled at a very early stage during sample preparation, unlike in chemical labeling where pooling is most commonly performed after proteolytic digestion. For label-free approaches, quantification is performed computationally between different experiments. From [80].

Chemical labeling strategies target the reactive side chains or protein/peptide termini. They can be applied at the protein or peptide level, but because of its overwhelming preponderance, only the peptide level will be discussed further (Figure 9). One of the earliest labeling strategies is called ICAT (isotope-coded affinity tags) [81]. The ICAT reagent consists of a thiol-specific reactive group, a linker which contains the isotope label and a biotin affinity tag for enrichment. Therefore, it reacts with cysteine-containing peptides that can be then enriched using the biotin tag for MS analysis. Not all proteins have cysteines and most peptides cannot be quantified, therefore ICAT is not in general use in proteomics. In dimethyl labeling [82], all primary amines are converted to dimethylamines. This reaction adds two methyl groups to all lysine side chains and all

free N-termini, thus in principle labeling all peptides. For quantification at the MS/MS level, the most commonly used reagents are TMT (tandem mass tag) [83] and iTRAQ (isobaric tags for relative and absolute quantification) [84]. Both reagents use isobaric tags consisting of a reporter group, a mass balance group, and a peptide-reactive group (NHS ester). The overall nominal cumulative mass of reporter and balance components are kept constant using differential isotopic enrichment, hence the name isobaric. The labeled peptides generate a single ion cluster at the MS level. The different reporter ions are released upon fragmentation and are used for relative quantification. TMT and iTRAQ can easily be multiplexed and they can be applied to any sample type, making them popular choices in quantitative proteomics. One of the limitations of these methods is that mixing of samples occurs late in the workflow, which can introduce variability and systematic errors.

Metabolic labeling approaches introduce the isotope labels at the cellular or organismal level (Figure 9). One of the most popular metabolic labeling methods is stable isotope labeling with amino acids in cell culture (SILAC) [85]. In this approach, cells are grown in isotopically labeled amino acids (typically arginine and lysine). After at least five doublings, the heavy amino acids are incorporated into almost the entire cellular proteome. Choosing arginine and/or lysine ensures that the peptides cleaved by trypsin or LysC incorporate at least one labeled amino acid. A classical SILAC experiment would usually involve the comparison of two cellular states one of which is labeled with the 'light' or wild type amino acid and one with the 'heavy' form. Comparing the intensities of the isotope clusters of light and heavy peptides yields the relative amount of the peptide and thereby proteins between the two states in question. It is also possible to label and analyze three samples together (designated light, medium and heavy SILAC labeling). Expanding the experiment to four or five labels is prevented by the repertoire of available labeled amino acids. Deuterated (^{2}H) peptides, for instance, can show retention time shifts that hinder accurate quantification. Metabolic labeling, in general, and SILAC, in particular, is considered to have the highest quantification precision and accuracy.

Samples can be pooled at the earliest step of the proteomic workflow, minimizing the systematic error that can arise from separate sample handling.

Applications of the SILAC approach have lately expanded into many biological questions beyond simple expression differences between cultured cells. It can now be used to study protein turnover and translation rates (pulsed SILAC) [86, 87], protein-protein interactions [88] as well as time-resolved analysis of signaling pathways [52]. Human biological samples cannot be SILAC-labeled, however, the recently developed super-SILAC approach sidesteps this limitation by using a mixture of SILAC labeled cell lines that is used as an internal standard for quantifying proteins in tissues [89]. The basic idea is that the mixture is to represent the complexity of the tissue proteome as closely as possible. Indeed, a mix of labeled cell lines achieved a higher quantification accuracy compared to using one cell line when studying breast cancer tissues. A super-SILAC mix of labeled breast cancer cell lines was generated by rationally selecting four breast cancer cell lines from different stages of the tumor and adding to this a mammary epithelial cell type. Equal amounts of the lysates from the five labeled cell lines were combined. This super-SILAC mix was then spiked in at a 1:1 weight/weight ratio into every sample to be analyzed. In principle, this approach allows the comparison of any number of breast cancer tissues. The first step is to quantify the endogenous tissue proteins against the spiked-in standard. Since the same quantity of standard is spiked into all samples, this allows the comparison of all samples against each other. Therefore, super-SILAC mixtures can be considered reference standards that when spiked in fixed ratios allow the comparison of any number of tissue samples (Figure 10).

Label-based approaches can also be used for absolute quantification by spiking a known quantity of labeled standard into the sample. This can be done at the peptide level as in the case of AQUA (for absolute quantification) where labeled synthetic peptides are used [90]. Alternatively, labeled protein fragments or full length proteins can be spiked in before the digestion step [91-93], to control for the variability introduced during sample preparation, such as missed cleavages and protein adsorption. These approaches are by

their nature two-step processes, which require accurate quantification of the standard first.

Label-free quantification – As the name indicates, label-free quantification includes all approaches used to quantify proteins which do not involve introducing stable isotopes. Compared to other techniques, label-free quantification is considered less accurate, especially when the overall experimental process is taken into account [80]. There is higher variability during each of the steps of sample preparation and measurements since they are performed separately for the samples. However, since the approach is inexpensive, technically straightforward, and allows studying any number of samples, it is inherently attractive. A very simple form of relative label-free quantification is spectral counting, in which the number of spectra identifying each protein serves as a proxy for its abundance. Proteins with few sequenced peptides cannot be quantified accurately, making this approach unreliable for low abundance proteins. Intensity-based label-free algorithms such as the ones in the MaxQuant platform [66], combined with high resolution data, can provide much more accurate results especially if they include advanced normalization steps which correct for experimental variability (Cox et al., in revision).

Acknowledging the importance of developing such reliable data analysis tools, intensive normalization steps are now implemented, which allow the comparison of large numbers of samples such as those required for clinical studies. An even more advanced algorithm, which combines label-based and label-free calculations (hybrid algorithm), has also been implemented. In this strategy, quantification is done by SILAC pairs or by label free methods as appropriate. In this way, the accuracy of SILAC is retained while it is still possible to quantify proteins that are present in the sample but absent from the super-SILAC mix.

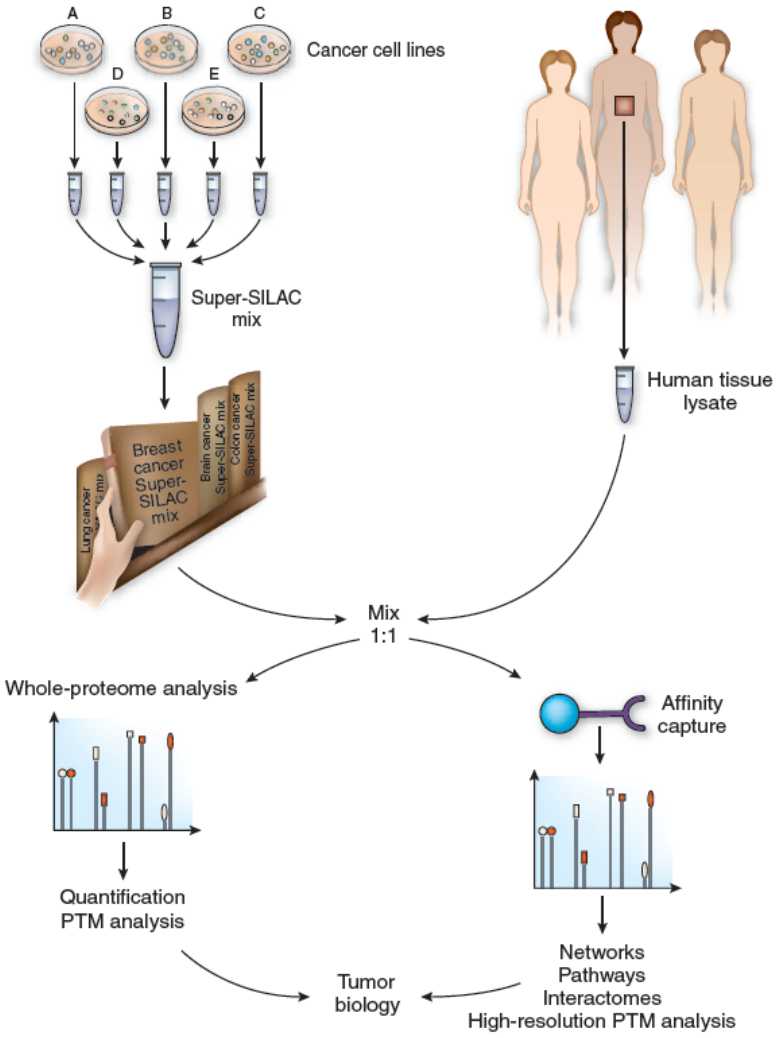


Figure 10. **Super-SILAC workflow to study cancer tissues.** The super-SILAC approach allows the comparison of any number of healthy and patient tissues. In the example depicted in the workflow, five breast cancer cell lines, labeled with heavy arginine and lysine, were used to generate a breast cancer super-SILAC mix. The super-SILAC mix was spiked in at a 1:1 ratio with the lysate of a human tissue. A standard mix can be generated for any type of cancer. From [94].

Label-free approaches that can estimate absolute amounts without labeled standards have also been developed. Intensity-based absolute quantification (iBAQ) [95], for instance, uses the summed intensities of the peptides identifying a protein as a proxy for its abundance and the number of theoretical peptides as a protein-specific correction factor. A spiked-in non-labeled standard of accurately quantified proteins from a different

organism, introduced before sample preparation, allows a linear extrapolation of the absolute amounts of all identified proteins.

In short, the technology at all levels, and particularly at the bioinformatics level, has taken huge steps towards overcoming the depth and the high-throughput challenges. Thus, the technology has recently become more amenable to answer critical clinical questions such as classifying cancer patient into different disease entities.

1.4. Molecular cancer diagnostics

In addition to dramatic technological advances in the molecular characterization of cancers, it is very important to be able to reliably identify those patients who are most likely to benefit from a particular agent. This has led to the rise of companion diagnostics as an important component of targeted therapeutics. It has long been recognized by oncologists that every human cancer is comprised of biological subsets that differ in clinical behavior. Each patient is diverse in clinical presentation, prognosis and response to treatment. They also differ in their risk of recurrence, second malignancy and long-term complications of treatment [96]. Recent technological advances allow large-scale analyses of individual cancers at the level of the genome, transcriptome and proteome [97]. This allows scientists and clinicians to better understand the biological heterogeneity of human cancer at the molecular level. The potential benefits of improved molecular characterization of individual cancers are enormous [97]. In fact, it is the molecular understanding of cancer causation and progression that led to the discovery and development of molecularly targeted drugs that enabled a personalized approach to cancer treatment [98]. The hope is that these developments will allow a shift from non-specific cytotoxic drugs, which damage both tumor and normal cells, to more precise drugs. Clinical success in targeting molecularly defined subsets of several tumor types is already evident. Early experiences include the HER2 antibody trastuzumab in breast cancer [99], the BCR-ABL inhibitor imatinib in chronic myeloid leukemia [22], and the EGFR kinase inhibitors gefitinib and erlotinib in non-small-cell lung cancer (NSCLC) [100-

102]. Remarkably, the successful development of all of these targeted drugs was highly dependent on patient selection using predictive biomarkers [103]. The ability to select patients benefits clinical trials in the evaluation process of the drugs as well as in clinical practice. In fact, the lack of an efficient strategy to evaluate targeted agents in patients is a key reason why there is only a modest number of similarly successful targeted therapies [97]. In short, the ability to categorize patients into subtypes with associated treatments is the most important step toward the goal of personalized medicine. The challenge lies in discovering cancer biomarkers or other stratification strategies for patient subtyping.

1.4.1 Types of cancer biomarkers

There are three types of biomarkers that are important for the rational development of anticancer drugs and the clinical management of patients, predictive, prognostic and pharmacodynamic biomarkers [97]. There has been a large focus on predictive markers, used to assess response; that is the probability that a patient will benefit from a particular treatment. As a prominent example, breast cancer patients with amplification in ERBB2 (also known as HER2 or NEU) benefit from treatment with trastuzumab (Herceptin) [99] and patients over-expressing the oestrogen receptor benefit from treatment with tamoxifen instead [104, 105]. Predictive chromosomal translocations as occurring in some leukemia patients can also be biomarkers. Patients with PML–RARA translocation respond to all-*trans* retinoic acid [106], whereas those with the Philadelphia chromosome (BCR–ABL fusion gene) respond to imatinib mesylate (Gleevec or Glivec) [22]. Often mutations in specific genetic regions have predictive power and genotype-based analysis is required. Examples include mutations in the kinase domain of the epidermal growth-factor receptor (EGFR) that predict the sensitivity of lung tumors to erlotinib or gefitinib [100-102]. Conversely, distinct mutations in KRAS predict that patients with lung cancer will fail to respond to these inhibitors [107].

The second type of biomarkers that can influence treatment choice are prognostic biomarkers. These allow the prediction of the natural course of individual cancers distinguishing ‘good outcome’ tumors from ‘poor outcome’ tumors. Such distinctions can

guide how aggressive the treatment should be [97]. The most prominent example is also in breast cancer but in this case it involves gene-expression signatures [108]. These signatures estimate the probability of the original breast cancer recurring after it has been resected [97]. They are marketed for clinical use as Oncotype DX (Genomic Health) [109] and Mamma Print (Agendia) [110]. This signature is used to decide which patients should receive systemic therapy after surgery with the aim to eliminate any remaining tumor cells and reduce the risk of relapse and which patients can be spared such invasive treatment.

The third type of biomarkers are the pharmacodynamic biomarkers. These measure the treatment effect of a drug on the tumor or the host. This can help to assess the efficacy of the drug and to guide dose selection in the early stages of its clinical development.

1.4.2 Cancer molecular profiling and biomarker discovery technologies

The topic of cancer biomarkers is a broad one with a large associated literature. From the examples discussed above it is clear that the term "biomarker" encompasses a wide range of molecular forms. It can be a gene amplification, a translocation, a mutation, a protein overexpression or even a gene expression signature. This diversity is the result of the different and rapidly evolving technologies employed for the molecular characterization of tumors at all levels. These include technologies that measure changes in content or sequence of DNA (genome), expression of messenger RNA (transcriptome), and production of proteins (proteome).

Technologies for analyzing the cancer genome - The human genome project, which was completed using first generation "Sanger" sequencing, generated demand for cheaper and faster sequencing methods. So called 'next generation sequencing' platforms are capable of performing massively parallel sequencing, i.e. the simultaneous sequencing of millions of DNA fragments [111]. In addition, for many types of alteration in the DNA, specialized detection methods have been developed, including those to study single

nucleotide variants, small insertions/deletions, chromosomal rearrangements, gene fusions, alternatively spliced transcripts, chromosomal copy number alterations, and detection of foreign DNA (such as from viruses) [112].

Since most biomarkers previously discovered in cancer were either single genetic mutations that drove the cancer ('driver mutation'), gene amplifications or translocations, efficient and affordable strategies were first developed for their detection. Strategies like whole-exome sequencing, which target protein-coding genes, or targeted sequencing, which focuses on hotspots for disease causing mutations, have been broadly applied. Currently, the costs of whole genome sequencing are plummeting [112], making the technology much more accessible. This has dramatically increased the amount of research involving large-scale sequencing.

Technologies for analyzing the cancer transcriptome - Until recently, the most commonly used technology to analyze gene expression profiles were microarrays. The basic principle of microarrays is the hybridization of cellular RNA that has been converted to cDNA to fixed probes followed by its detection using fluorescence. However, the advent of next generation sequencing has provided a strong alternative to microarrays. In a single analysis, next generation sequencing of RNA (RNA-seq) can provide information on the entire transcriptome of a sample with much higher depth and quantitative accuracy than microarrays.

Technologies for analyzing the cancer proteome - As discussed in Chapter 3, MS-based quantitative proteomics has evolved to become the method of choice for global in-depth profiling of proteomes of any biological system. Previously used methods such as 2D gels had proven to have low accuracy and depth and protein arrays are far from providing comprehensive coverage. That said, early application of MS-based approaches to characterize tumors used immature technology, generally resulting in poor outcomes. Quantitative approaches have only recently reached the stage where almost complete and accurate measurements of complex proteomes are possible. Such in-depth accurate

measurements are a requirement for the global molecular profiling of individual tumors at the proteome level.

To date, most biomarkers with clinical applications are single mutations, translocations, or genetic amplifications [113]. However, it has now become evident to cancer biologists that these single biomarkers are likely to be insufficient to select the optimal targeted therapies. In cancer, signaling pathways important for tumor growth and survival are deregulated by multiple cellular changes rather than a single modification. Alternative compensatory mechanisms that continue to promote cell proliferation and survival become activated by various targeted regimens, for instance in leukemia patients who develop resistance to Imatinib [114]. Therefore, biomarkers that better reflect the complex molecular aberrations of a given single tumor are needed. The recent development of technologies that allow measuring global molecular profiles of tumors allows researchers to screen the whole genome, proteome, transcriptome, and metabolome for new biomarkers. Global molecular screening can result in genetic, proteomic or metabolic profiles, or 'signatures' that may better classify the tumor and that can guide development of rationally designed combination therapies. Since somatic mutation profiles being are challenging to work with, most prior attempts to stratify tumors with molecular profiles have used mRNA expression data.

1.4.3 A success story of gene expression profiling: subtyping of diffuse large B-cell lymphomas based on cell-of-origin

Multiple stages of normal B-cell development can give rise to malignant lymphomas, which exploit the regulatory biologic features of normal B-cells for their own advantage. When a mature naïve B-cell is stimulated with a T-cell dependent antigen, a germinal center reaction is initiated. A germinal center B-cell is at a quasi-stable stage characterized by two forms of genetic modifications that alter the B-cell receptor (BCR). These are 'class-switch recombinations', which changes the immunoglobulin heavy-chain class from IgM to IgG, IgA or IgE, and 'somatic hypermutations' which expand the

repertoire of BCRs through the induction of immunoglobulin-variable-region mutations. Both types of modifications involve DNA editing processes which require activation-induced cytidine deaminase (AID) [115]. Within the dark zone of the germinal center, the B cells are rapidly dividing with a non-cleaved nucleus (centroblasts). These cells expand at a remarkable rate. Periodically, they travel to a sub-compartment of the germinal center (the light zone) where they develop into centrocytes (non-dividing B-cells with a cleaved nucleus). The light zone is rich in follicular dendritic cells and follicular helper T cells. As a result of stimulation by an antigen on dendritic cells and CD40 ligand on T cells, centrocytes may be rescued from cell death. They can revert to centroblasts and re-proliferate, or they can differentiate into memory B-cells or plasma cells. A germinal center reaction where rapidly dividing cells are undergoing genetic rearrangements is essential for a normal and fast immune response. However, the germinal center reaction can also give rise to many types of lymphomas. The genetic modifications essential for a normal immune response are also a source of DNA damage, which can become pathologic in lymphomas (Figure 11) [116]. In fact, studies have shown that many types of lymphomas such as diffuse large B-cell lymphoma (DLBCL), follicular lymphoma and Burkitt's lymphoma are derived from germinal center B-cells [117]. The malignant cells of these lymphomas carry the differentiation program of the normal B-cells they are derived from [118-120], but they also acquire oncogenic abnormalities that subvert the normal program. It is these oncogenic molecular differences that are refining the subgrouping of these lymphomas, which prior to the advent of molecular profiling technologies was based mainly on morphological and histological characteristics of the tumors. One of the most prominent examples is the ability to molecularly classify three histologically indistinguishable subtypes of DLBCLs using microarrays [121]. This has great promise for targeted therapeutic regimens for each of the subtypes.

Diffuse large-B cell lymphoma (DLBCL) comprises the largest percentage (30-35%) of B-cell non-Hodgkin lymphomas. It is biologically aggressive with a wide range of clinical presentations. The current treatment regimens can cure more than half of the cases

[122]. However, one-third of the patients have refractory disease or relapse after treatment and eventually succumb to the disease [123]. The salvage treatment for these patients is autologous stem cell transplantation (ASCT), which, at present, has poor success rates. There is an urgent need for more targeted approaches that take into account the underlying molecular heterogeneity responsible for the diverse clinical responses [121].

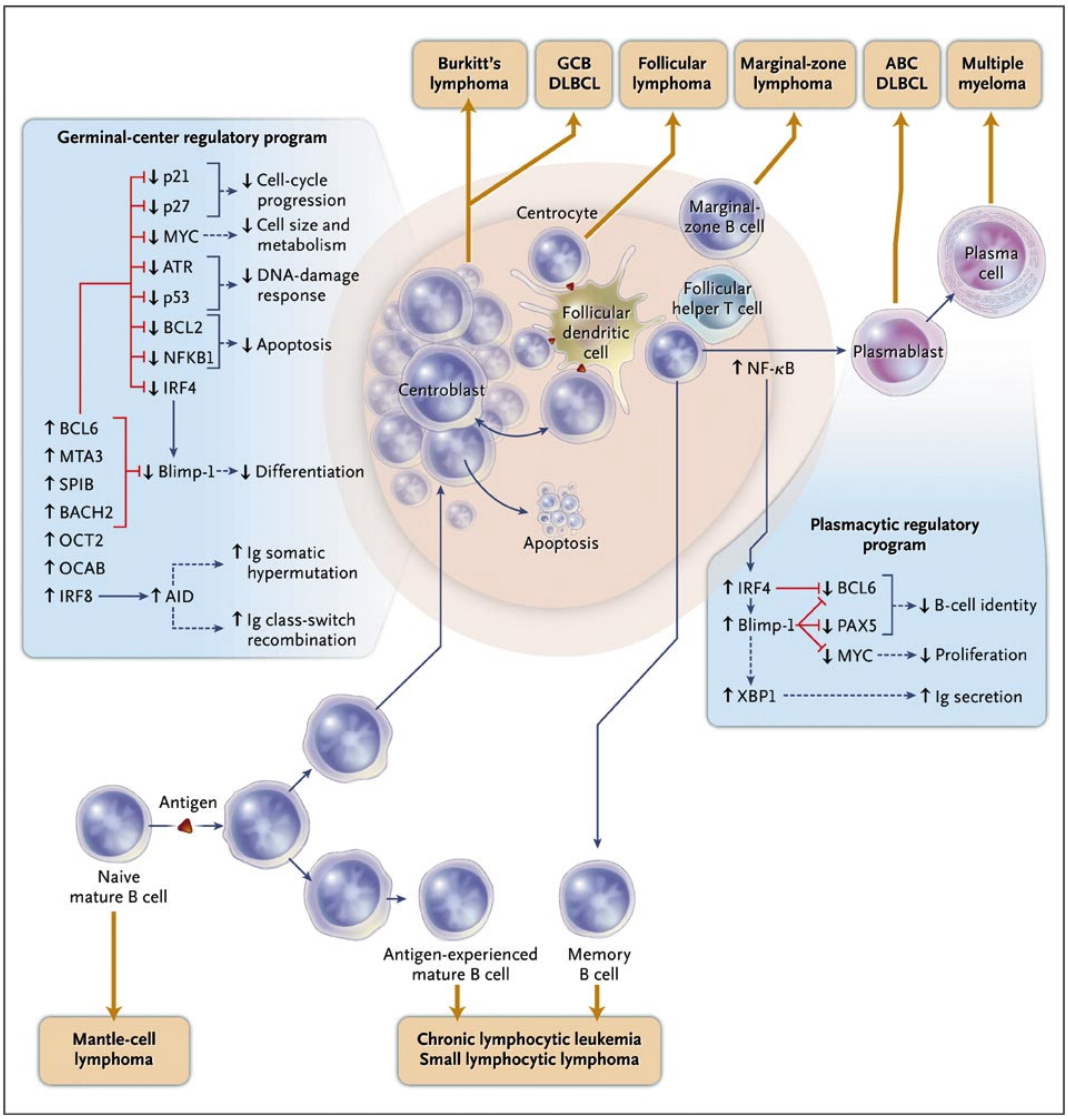


Figure 11. **B-cell differentiation and the putative origins of various non-Hodgkin's lymphomas**. Malignant lymphomas can arise at multiple stages of B-cell development. Based on gene expression studies, GC DLBCL is potentially derived from a germinal center B-cell and overexpresses genes characteristic of the germinal center

reaction. ABC DLBCL is likely derived from a post-germinal B-cell and is characterized by overexpression of genes that regulate the plasmacytic differentiation program. From [116].

Despite differences in clinical courses of patients, DLBCL was for a long time considered to be a single disease. Attempts to subgroup DLBCL based on morphology had largely failed as the precision of morphological diagnosis, even when supplemented with immunochemistry for a few markers, was insufficient to robustly define diagnostic subgroups [118]. The underlying molecular heterogeneity within morphologically indistinguishable tumors of DLBCL was first revealed using microarray-based gene expression profiling (GEP). This technology allowed the identification of subtypes that originated from B-lymphocytes at different developmental stages [118]. These are the germinal center B-cell (GCB), activated B-cell (ABC) and primary mediastinal B-cell lymphoma (PMBL) (Figure 12). Each subtype possesses a set of overexpressed genes that corresponds to a B-cell differentiation stage from which the tumor has potentially arisen. The GCB subtype, for instance, expresses hundreds of genes characteristic of germinal center B-cells. In addition, the malignant GCB cells continue to undergo somatic hypermutation. The ABC-DLBCL has the characteristics of a post-germinal B-cell and overexpresses genes that regulate plasmacytic differentiation. The third subtype, PMBL, is likely derived from a post-thymic B-cell and has a gene expression signature with more similarities to Hodgkin's lymphomas than with other DLBCL subtypes. Importantly, the cell-of-origin classification correlated with subgroups that have different prognosis. The majority of PMBL patients can be cured with an effective chemotherapeutic regimen (DA-EPOCH-R) [124]. When treated with R-CHOP, the most commonly used regimen for newly diagnosed DLBCL, the ABC and GCB subtypes show a 3-year overall survival rate of approximately 45% and 80%, respectively. Differences in prognosis as well as in their corresponding molecular profiles support the notion that ABC and GCB are distinct neoplasms that should be treated differently. Several studies focused on elucidating the different oncogenic mechanisms (genetic aberrations and oncogenic pathways) each

subtype is dependent on. This shows that such studies can play a major role in guiding targeted therapeutic strategies.

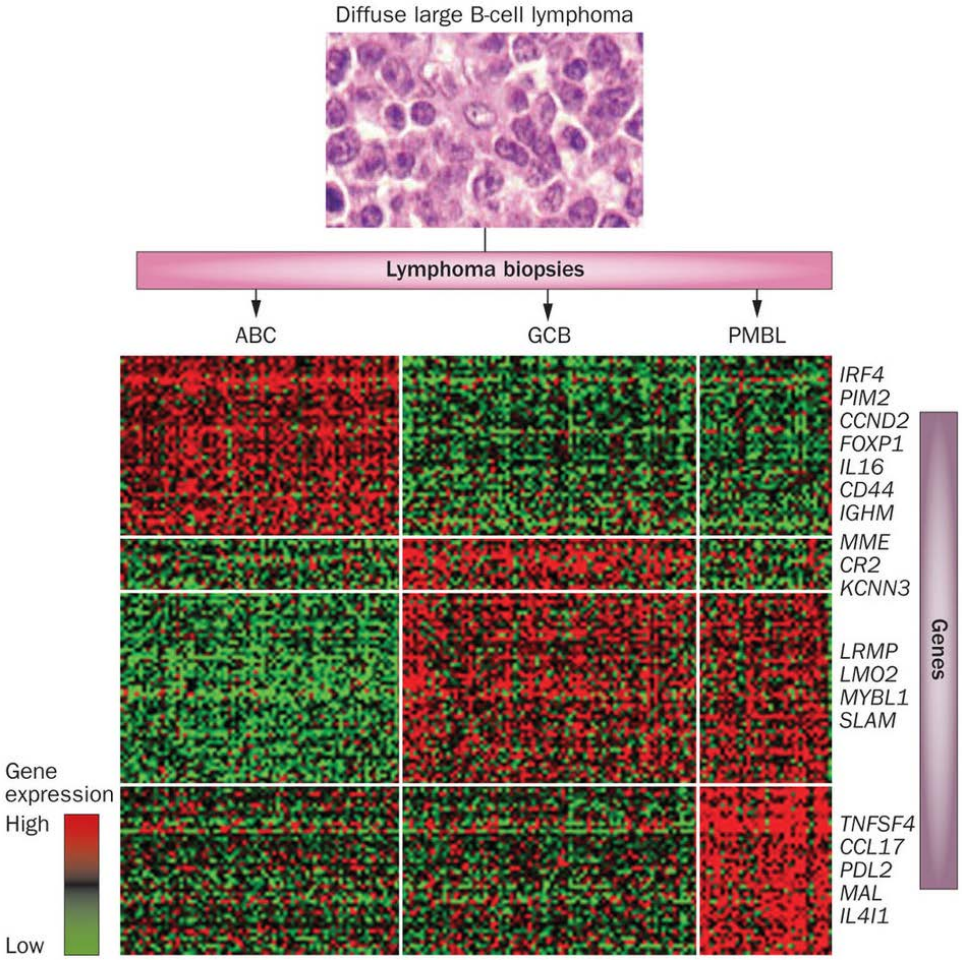


Figure 12. **Three distinct cell-of-origin subtypes of DLBCL discovered by gene expression profiling**. The three subtypes of DLBCLs overexpress genes that are characteristic of the B-cell differentiation stage from which they are derived. From [121].

Oncogenic mechanisms in lymphoma subtypes - In addition to a gene expression signature characteristic of germinal-center B cells, recent genomic studies have revealed some genetic lesions specific to GCB DLBCL. Examples include the t(14;18) translocation which is found in 34% of GCB DLBCL cases [125] and the deletion of the tumor

suppressor PTEN which is found in 11% of GCB DLBCL, but is never observed in ABC DLBCL or PMBL [126, 127]. In addition, a somatic point mutation in the EZH2 gene, which encodes histone-lysine N-methyl transferase, has been identified in the GCB subtype. This mutation was present in 18 out of the 83 GCB samples (22%) but in none of the 42 ABC samples [128]. The ABC subtype possesses a gene expression signature characteristic of normal B-cells that have been activated by BCR cross-linking. One of the pathogenic hallmarks of the ABC DLBCL subtype that was revealed using gene expression studies is the constitutive activation of the NF-κB pathway [129]. The NF-κB family of transcription factors control various cellular processes involved in tumor development such as cytokine secretion, cellular proliferation, angiogenesis, invasion and metastasis [130]. NF-κB signaling results in the upregulation of the transcription factor interferon regulatory factor 4 (IRF4), which drives plasmacytic differentiation. IRF4 expression is directly regulated by NF-κB transcription factors, which can be induced by both BCR and Toll-like receptors (TLRs) signaling pathways [131]. A unique characteristic of B-cells is the expression of both types of receptors, which provides them with the ability to integrate responses to a variety of stimuli [132]. Several activating mutations influencing the NF-κB pathway have been identified in subsets of patients with ABC DLBCL. For instance, mutation in caspase recruitment domain-containing protein 11 (CARD11) were found in 10% of ABC DLBCL cases [133]. B-cell lymphoma/leukemia 10 (BCL-10) and mucosa-associated lymphoid tissue lymphoma translocation protein (MALT1) and CARD11 form a signaling complex, which is required for the BCR-dependent activation of NF-κB signaling upon antigen stimulation [134]. In contrast, patients with wild type CARD11 activate NF-κB signaling through 'chronic active' BCR signaling [135]. In some cases, chronically active BCR signaling is associated with mutations in the B-cell co-receptor CD79B (21% of cases of ABC DLBCL) [135]. Mutations involving the TLR signaling pathway have also been identified. A specific point mutation in MYD88, an adaptor protein of TLR signaling, has been observed in 30% of cases of DLBCL [136]. Other mutations such as the biallelic deletion of TNFAIP3 (A20), which is a negative regulator of the NF-κB pathway, occurs in 30% of the ABC cases and can co-exist

with mutation in both MYD88 and CD79B. This suggests that A20 inactivation can play a role in enhancing both BCR and TLR signaling [137].

Genomic analysis of primary DLBCL tumors has revealed the tremendous molecular complexity of the disease. There are clear indications that devising targeted therapies requires more than just the identification of mutations that drive tumor development. Cooperating mutations as well as cross-talk in signaling pathways that confers drug resistance are all important factors to consider in devising targeted therapeutics. In some cases of ABC DLBCL, for instance, there were no genetic alterations associated with chronic BCR signaling, which is an important driver of lymphomagenesis in this subtype. A global understanding of aberrant signaling pathways in such cases may provide valuable insights.

Classification improvements - The cell-of-origin (COO) classification in DLBCL mentioned above did not fully reflect the differences in overall survival after chemotherapy among patients [118] . In an attempt to refine the classification, a follow up study was conducted by the same group where a molecular predictor of risk was constructed. It used genes with expression patterns that correlated with survival and identified four gene-expression signatures reflecting different biological attributes of the tumor. Genes mainly from the four signatures (germinal-center B-cell, proliferation, major-histocompatibility-complex class II (MHCII) and lymph node signatures) were included in the predictor. Representing the four signatures, the predictor was then minimized to seventeen genes and it was subsequently shown to have greater prognostic power than COO subgrouping of DLBCLs [138]. Following an independent change of the treatment regimens of DLBCL involving the addition of rituximab to combination chemotherapy, a further stratification study was conducted. This study showed that a multivariate model constructed from three gene-expression signatures (germinal-center B-cell, stromal-1, and stromal-2) predicted survival both in patients who received CHOP, and patients who received R-CHOP. Stromal-1 signature reflected extracellular matrix

deposition, and stromal-2 signature, which had an unfavorable prognosis, reflected tumor blood vessel density [139].

Since RNA extraction from FFPE samples is still difficult [140, 141], many immunohistochemistry-based algorithms have been developed in parallel to GEP studies. The basic principle of these algorithms is to simulate the results of GEP by classifying the patients into ABC-DLBCL and GCB-DLBCL subgroups [142]. They included 3 to 6 antibodies already available for immunohistochemistry and would provide a cost-effective replacement to GEP if proven successful [142]. The Hans algorithm, which was the first algorithm to be developed, uses antibodies against CD10, BCL6, and MUM1 and had 83% concordance with GEP [143]. Several algorithms aimed to improve the accuracy of the Hans algorithm [142]. However, a recent study that compared all IHC-based algorithms, concluded that GEP and not IHC-based algorithms accurately predict prognosis in DLBCL patients treated with immunochemotherapy [144]. The most recent prognostic model based on immunohistochemistry attempts to simulate the COO classification as well as the stromal signatures. It also includes the assessment of microvascular density as well as the International Prognostic Index (IPI) to stratify patients [141]. However, it has not yet to be proven in clinical settings.

Therapeutic opportunities – The above mentioned addition of rituximab (R), a chimeric monoclonal antibody targeting CD20, to combination chemotherapy with cyclophosphamide, doxorubicin, vincristine, and prednisone (CHOP) is the most recent breakthrough in the treatment of DLBCL patients. It has shown benefit for all DLBCL subtypes and is the most commonly used regimen in newly diagnosed cases [121]. Nevertheless, the tremendous improvements in understanding the molecular heterogeneity of DLBCL have opened the door for subtype-targeted therapeutic approaches. Several clinical trials are underway based on characteristic mutations and oncogenic pathways that are specifically activated in each subtype. As ABC DLBCL is the subtype with worst prognosis, improving therapeutic strategies by targeting its aberrant signaling pathways should be a high priority.

Targeted therapy in GCB DLBCL – The increased histone methyltransferase activity in DLBCL cell lines that harbor EZH2 mutations seem to be necessary for their proliferation. An oral inhibitor (E7438) which targets EZH2 is in early clinical testing for relapsed DLBCL of all subtypes [121]. In addition, targeting the PI3K/Akt/mTOR pathway is also under clinical investigation using several inhibitors [127]. A prominent target for GCB DLBCL is BCL-2. A second generation inhibitor of BCL-2 is in a phase II study including patients with relapsed B-cell lymphomas. This inhibitor lacks significant binding to Bcl-Xl, avoiding its co-inhibition that causes thrombocytopenia [121].

Targeted therapy in ABC DLBCL - Being the characteristic pathological hallmark of the ABC subtype, several clinical studies focus on the NF-κB pathway. Targeting this pathway via proteasome inhibition has shown selective efficacy in ABC DLBCL patients. This landmark study in patients with refractory or relapsed DLBCL revealed that bortezomib (a proteasome inhibitor) sensitized patients with the ABC subtype and not the GCB subtype to chemotherapy [145]. More selective and better tolerated proteasome inhibitors are under clinical investigations [121]. Lenalidomide is a new and promising drug that selectively kills ABC DLBCL cells through targeting IRF4 [146]. In a study of patients with relapsed or refractory DLBCL, lenalidomide demonstrated differential efficacy between non-GCB and GCB DLBCL, with the latter showing better outcome [147]. In addition, several drugs that target the NF-κB pathway at different stages are under investigation. Some target upstream regulators such as BTK, PKC-β and MALT1 [121].

1.4.4 MS-based proteomics: a promising tool for molecular cancer diagnostics

The successful use of genomic aberrations (BCR-ABL, ERBB2 and EGFR) as biomarkers influencing treatment decisions is a promising example of the translation of information from the cancer genome to clinical practice. However, many of these successes predate the current genome-wide, high-throughput technologies, and some of them instead resulted from decades of work on the molecular mechanisms involved. With the advent of more affordable and efficient next generation sequencing platforms, the

numbers of sequenced human tumors have exploded. The enormous amount of information generated has uncovered a hitherto unimagined level of complexity of the cancer genome. An emerging challenge is to interpret the information and to determine which molecular abnormalities contribute to cancer, and which are simply noise [148].

As mentioned above, there are many indications that biomarkers based on a single mutation, translocation, or genetic amplification are insufficient to design targeted therapeutic strategies. Signaling pathways that the tumor exploits for its own growth and survival are under the control of multiple cellular changes. A global molecular understanding of the tumor subtypes can help design combinations of novel agents to target the oncogenic drivers of each subset of disease. Attempts to stratify patients based on the entire mutation profile have been challenging. Somatic mutation profiles are inherently sparse [149]. In addition, there is remarkable heterogeneity in mutation frequency and spectrum within cancer types [150]. As a result most attempts to stratify tumors with molecular profiles have used mRNA expression data. Gene expression studies discovered informative subtypes in diseases such as glioblastoma, breast cancer and DLBCLs. Conversely, defining clinically relevant subtypes of colorectal adenocarcinoma using expression profiles was not equally successful and the subtypes derived did not correlate with patient survival and response to chemotherapy or any clinical phenotype [151]. Issues such as RNA sample quality and lack of reproducibility between biological replicates may have contributed to these results [152].

Integrating somatic tumor genomes with gene networks a recently developed network-based stratification (NBS) method classified cancers into informative subtypes by clustering together patients with mutations in similar network regions[149]. Such approaches are built on the insight that even if two tumors do not share common mutations, the affected networks may still be common. In fact, it has become more widely accepted that cancer is not a disease of individual mutations, nor of genes, but of gene combinations working in molecular networks [153, 154]. These networks correspond to cancer hallmarks such cell proliferation and apoptosis.

A more direct way to observe impacted biological networks and systems would be to explore the end product of the gene expression cascade, the proteins. Additionally, proteomics can in principle provide a detailed picture of the active and fully modified protein forms. There is mounting evidence that many quantitative measures of biological regulation cannot be predicted from transcript levels alone. For instance, the major factor controlling abundance of proteins appears to be the level of translation [95]. Proteomics measures regulation directly at the expression level of all proteins. It further allows characterization of such regulations at the level of protein-protein interactions and PTMs. These dimensions can in principle be explored either at the whole cell level or in individual subcellular compartments.

The potential of clinical proteomics is therefore self-evident. The general aim of thesis is to develop contemporary MS-based proteomics into a platform for reaching one of the fundamental goals of personalized medicine, tumor stratification.

2. RESULTS

2.1 Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles

2.1.1 Project aim and summary

Diffuse large B-cell lymphoma (DLBCL) is an aggressive malignancy of mature B lymphocytes. It is the most common subtype of non-Hodgkin's lymphoma and it is clinically heterogeneous. Attempts to sub-classify DLBCL on the basis of morphology, even when supplemented with immunohistochemistry for a few markers, have largely failed. Two molecularly distinct subtypes of DLBCL have been identified by gene expression profiling using a specialized DNA microarray (Lymphochip). Each of the subtypes expressed a set of genes characteristic of a particular B-cell development stage. Germinal center B-cell DLBCL (GCB-DLBCL) is a subtype that expresses genes characteristic of germinal center B cells, while a second subtype, activated B-cell DLBCL (ABC-DLBCL), expresses genes characteristic of *in vitro* activated peripheral blood B cells. Importantly, the molecular subtypes identified patient groups that differed in overall survival after chemotherapy.

Mass spectrometry-based proteomics offers unbiased methods to molecularly characterize tumors at the protein level, which is one step closer to the disease phenotype. Using and improving on recent developments in our MS-based proteomics platforms, we wanted to employ state-of-the-art technology to address tumor classification a challenging question of clinical relevance. We chose to work with DLBCL as a model system and to employ the super-SILAC approach for quantification. One of the important steps in this project was the empirical design of a general lymphoma super-SILAC mix that included cell lines with the most diverse protein expression profiles. The mix provided a spike-in standard that allowed the comparison of any number of lymphoma cell lines or tissues. We compared an extensive fractionation approach (where each sample was fractionated into

six sub-samples) with our recently developed single shot approach where no fractionation is performed. Interestingly, the in-depth profiling as well as the single shot measurements of the samples allowed clear segregation into their corresponding subtypes. Reassuringly, the study confirmed some of the known segregators as well as identified some novel ones. We discovered a signature of 55 proteins that was capable of strongly segregating the subtypes. It even highlighted the relative up-regulation of NF-κB target genes in the ABC-DLBCL subtype whose constitutive NF-κB signaling is an oncogenic hallmark.

2.1.2 Contribution

This project was initiated by Matthias Mann, Marc Schmidt-Supprian, who also provided supervison, and myself. I performed and optimized all sample preparation techniques and MS analysis methods and acquired data and analyzed it. I designed all figures and tables for the publication. I wrote the manuscript with the help of Matthias Mann and Marc Schmidt-Supprian.

2.1.3 Publication

This project was published in 2012 as a research article in Molecular and Cellular Proteomics:

Super-SILAC Allows Classification of Diffuse Large B-cell Lymphoma Subtypes by Their Protein Expression Profiles

Sally J. Deeb, Rochelle C. J. D'Souza, Juergen Cox, Marc Schmidt-Supprian, and Matthias Mann

Mol Cell Proteomics. 2012 May; 11(5)

Super-SILAC Allows Classification of Diffuse Large B-cell Lymphoma Subtypes by Their Protein Expression Profiles*[S]

Sally J. Deeb‡, Rochelle C. J. D'Souza‡, Jürgen Cox‡, Marc Schmidt-Supprian§¶, and Matthias Mann‡‖

Correct classification of cancer patients into subtypes is a prerequisite for acute diagnosis and effective treatment. Currently this classification relies mainly on histological assessment, but gene expression analysis by microarrays has shown great promise. Here we show that high accuracy, quantitative proteomics can robustly segregate cancer subtypes directly at the level of expressed proteins. We investigated two histologically indistinguishable subtypes of diffuse large B-cell lymphoma (DLBCL): activated B-cell-like (ABC) and germinal-center B-cell-like (GCB) subtypes, by first developing a general lymphoma stable isotope labeling with amino acids in cell culture (SILAC) mix from heavy stable isotope-labeled cell lines. This super-SILAC mix was combined with cell lysates from five ABC-DLBCL and five GCB-DLBCL cell lines. Shotgun proteomic analysis on a linear ion trap Orbitrap mass spectrometer with high mass accuracy at the MS and MS/MS levels yielded a proteome of more than 7,500 identified proteins. High accuracy of quantification allowed robust separation of subtypes by principal component analysis. The main contributors to the classification included proteins known to be differentially expressed between the subtypes such as the transcription factors IRF4 and SPI1/PU.1, cell surface markers CD44 and CD27, as well as novel candidates. We extracted a signature of 55 proteins that segregated subtypes and contained proteins connected to functional differences between the ABC and GCB-DLBCL subtypes, including many NF-κB-regulated genes. Shortening the analysis time to single-shot analysis combined with use of the new linear quadrupole Orbitrap analyzer (Q Exactive) also clearly differentiated between the subtypes. These results show that high resolution shotgun proteomics combined with super-SILAC-based quantification is a promising new technology for tumor characterization and classification. ***Molecular & Cellular Proteomics 11: 10.1074/mcp.M111.015362, 77–89, 2012.***

From the ‡Department of Proteomics and Signal Transduction, Max Planck Institute of Biochemistry, D-82152 Martinsried, Germany and the §Department of Molecular Immunology and Signal Transduction, Max Planck Institute of Biochemistry, D-82152 Martinsried, Germany

⌘ *Author's Choice*—Final version full access.

Received October 24, 2011, and in revised form, February 22, 2012

Published, MCP Papers in Press, March 21, 2012, DOI 10.1074/mcp.M111.015362

Clinical heterogeneity in terms of patient survival rates and response to therapy is a major challenge in cancer treatment. This difficulty partly stems from grouping together molecularly distinct tumor entities as one clinical type and treating them in the same manner. Transcript-based profiling technology enables the segregation of subtypes based on their gene expression signatures (1, 2). However, it is often difficult to interpret such signatures with respect to the biology of the disease (1). In addition, gene expression signatures do not provide information if or to what extent the detected transcript is translated into proteins, and it ignores the effects of post-translational modifications. An in-depth, high accuracy quantitative proteomics approach capable of revealing common and distinct functional features between tumor entities may provide valuable insights into cancer subtypes of potential clinical relevance.

MS-based proteomics has recently evolved into an important tool in mining deregulated signaling pathways in cancer because of its ability to move one step closer toward the cancer phenotype and because of substantial progress in technology and methodology (3, 4). These advances in MS now allow the identification of thousands of proteins in a single experiment as a result of enhanced sensitivity, accuracy, and speed of analysis (5–7). In addition, a variety of quantitative proteomic approaches can monitor expression changes of thousands of proteins and post-translational modifications in a reproducible manner (8, 9). Stable isotope labeling with amino acids in cell culture (SILAC)[1] is a particularly accurate method of quantitative proteomics (10, 11), but until recently it was limited to cell lines or animals that could be metabolically labeled with heavy amino acids. This limitation of SILAC in studying patient tumor samples has been overcome through the use of a mix of multiple SILAC-labeled cell lines as an internal standard, a technique called super-SILAC

[1] The abbreviations used are: SILAC, stable isotope labeling with amino acids in cell culture; ABC-DLBCL, activated B-cell-like diffuse large B-cell lymphoma; BCR, B-cell receptor; DLBCL, diffuse large B-cell lymphoma; GCB-DLBCL, germinal-center B-cell–like DLBCL; PCA, principal component analysis; FASP, filter-aided sample preparation; SAX, strong anion exchange; STAT, signal transducers and activators of transcription; KEGG, Kyoto Encyclopedia of Genes and Genomes.

(12). This mix achieved superior quantification accuracy compared with a single SILAC-labeled cell line (13). In particular, a narrow ratio distribution was obtained with 90% of proteins contained within an easily quantifiable 4-fold range between the tumor and the SILAC mix. We reasoned that this ability to quantify several thousand proteins with high accuracy might enable confident proteomic classification of tumors in different subtypes.

The subclassification of diffuse large B-cell lymphoma (DLBCL), the most common lymphoma in adults, by gene expression profiling was a major breakthrough because it resulted in the identification of two histologically indistinguishable subtypes that differ in their outcomes after multiagent chemotherapy (14). The germinal-center B-cell-like (GCB) subgroup has a gene expression signature characteristic of normal germinal center B-cells, whereas the activated B-cell-like (ABC) subgroup, being the one with worse prognosis, has a gene expression signature characteristic of B-cells activated through their B-cell receptor. One of the key pathways that are differentially activated between DLBCL subgroups is signaling to NF-κB family transcription factors, which are constitutively active in the ABC subgroup (15). In B-cells, NF-κB controls the expression of genes necessary for both proliferation and survival in response to stimulation, including antigen recognition by the B-cell receptor (BCR). The IRF4 transcription factor, an NF-κB target, plays multiple roles in B lymphocyte development and function and is critical for plasma cell differentiation. Its high expression in ABC-DLBCL reflects constitutive NF-κB activity and plasmacytic differentiation. Recently, mutations leading to "chronically active" BCR signaling have been described as a mechanism providing aberrant cellular survival signals in ABC-DLBCL (16). In these cases, the constitutive NF-κB activation in ABC-DLBCL depends on the multiprotein CARD11-BCL10-MALT1 (CBM) complex (17–19). Such findings may open the door for new therapeutic modalities that target components of BCR signaling upstream of NF-κB. Furthermore, improvements in DNA sequencing technologies have paved the way to the discovery of novel aspects of DLBCL pathology, such as impairments in chromatin methylation and evasion of T cell immune surveillance (20). This shows that the deployment of novel methodologies continuously enhances our understanding of the complex biology of lymphomas.

Despite the success of gene expression profiling in differentiating between tumor subtypes, the extracted transcriptional signatures do not always suffice to identify biological drivers of tumor pathogenesis. Furthermore, their adoption in the clinic, where protein-based assays are more commonly used, has been slow. A long standing aim of the proteomics community is to directly study human cancer at the protein rather than the transcript level (3). Here, we use high resolution shotgun proteomics combined with a super-SILAC quantitative approach in an attempt to segregate DLBCL subtypes. If applicable, the super-SILAC technology should be particularly accurate, robust, and reproducible because it provides an entire reference proteome consisting of thousands of heavy labeled proteins for comparison of a large number of tumor samples. We evaluate the super-SILAC spike-in approach for distinguishing cell lines derived from ABC- and GCB-DLBCL patients. Choosing such closely related disease entities sets a high bar for our quantitative proteomics technology. Furthermore, the fact that specific biological differences between ABC and GCB are already known allows us to evaluate proteomics results in light of those differences.

EXPERIMENTAL PROCEDURES

Cell Culture Sample Preparation—ABC-DLBCL cell lines (HBL1, OciLy3, RIVA, TMD8, and U2932) and GCB-DLBCL cell lines (BJAB, DB, HT, SUDHL-4, and SUDHL-6) were grown in RPMI medium (Invitrogen) supplemented with 20% fetal bovine serum. Cell lysis was performed using a buffer consisting of 4% SDS, 0.1 M DTT, and 0.1 M Tris-HCl pH 7.5 followed by incubation at 95 °C for 5 min. The lysates were sonicated using a Branson type sonicator and then centrifuged at 16,100 × *g* for 10 min.

Cell lines selected for inclusion in the super-SILAC mix were grown in RPMI medium containing $^{13}C_6{}^{15}N_2$-lysine (Lys^8) and $^{13}C_6{}^{15}N_4$-arginine (Arg^{10}) (Cambridge Isotope Laboratories) instead of the natural amino acids and supplemented with 20% dialyzed fetal bovine serum. The cells were cultured for at least six passages until they were fully labeled as assessed by quantitative mass spectrometry. Less than 1% of tryptic peptides contained unlabeled arginine or lysine in the nine labeled cell lines (Ramos, Mutu, BL-41, U2932, OciLy3, BJAB, L1236, L428, and DB) and less than 0.3% of identified peptides showed evidence of Arg to Pro conversion. Equal amounts of the heavy lysates were mixed to generate the super-SILAC mix.

Protein Digestion and Fractionation—The super-SILAC mix (100 μg) was combined with an equal amount of the unlabeled cells and further processed by the filter-aided sample preparation (FASP) method (21). In short, the sample was loaded on Microcon filters with a 30-kDa cutoff (Millipore, Billerica, MA), which allows the replacement of SDS with a urea containing buffer. The proteins were then alkylated with iodoacetamide followed by overnight trypsin digestion at 37 °C in 50 mM ammonium bicarbonate. Peptides were collected from the filter after centrifugation and elution with water (2×).

Using strong anion exchange chromatography, 40 μg of the peptide mixture from each replicate was fractionated (22). In summary, the strong anion exchange (SAX) was performed in tip columns prepared from 200-μl micropipet tips stacked with six layers of a 3M Empore anion exchange disk (1214-5012; Varian, Palo Alto, CA). We used Britton & Robinson universal buffer composed of 20 mM acetic acid, 20 mM phosphoric acid, and 20 mM boric acid and titrated with NaOH to the desired pH for column equilibration and fraction elution. After loading the peptides at pH 11 and collecting it, five additional fractions were collected consecutively with buffers of pH 8, 6, 5, 4, and 3. The eluted fractions were desalted on reversed phase C_{18} Empore disc StageTips (23). Peptide elution was performed twice with 20 μl of buffer B containing 80% ACN in 0.5% acetic acid. Organic solvents were removed by a SpeedVac concentrator to prepare the samples for analysis by LC-MS/MS.

Liquid Chromatography and MS for Fractionation Experiments—Eluted peptides were separated on an in-house-made 15-cm column with a 75-μm inner diameter packed with ReproSil-Pur C_{18}-AQ 3 μm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) using an Easy nanoflow HPLC system (Proxeon Biosystems, now Thermo Fisher Scientific). The HPLC was coupled via a nanoelectrospray ion source (Proxeon Biosystems) to an LTQ-Orbitrap Velos mass spec-

trometer (Thermo Fisher Scientific) (24). Approximately 2 μg of peptides were loaded in buffer A (0.5% (v/v) acetic acid) with a flow rate of 500 nl min^{-1} and eluted with a 200-min linear gradient at a flow rate of 200 nl min^{-1}. Four different gradients were applied for optimal separation based on average peptide hydrophobicity. A gradient of 2–25% buffer B to separate the pH 11 fraction; 7–25% buffer B for the pH 8 fraction; 7–30% buffer B for the pH 6 and 5 fractions; and 7–37% buffer B for the pH 4 and 3 fractions. After each gradient, the column was washed, reaching 90% buffer B followed by re-equilibration with buffer A.

The mass spectra were acquired with an automatic switch between a full scan and up to 10 data-dependent MS/MS scans. Target value for the full scan MS spectra were 1,000,000 and resolution was 30,000 at *m/z* 400. Up to the 10 most intense ions (minimum signal threshold of 5,000) were sequentially isolated and accumulated to a target value of 40,000 with a maximum injection time of 150 ms and were fragmented by higher energy collisional dissociation (25). For a subset of measurements, MS/MS target values were set to 50,000. The spectra of the fragmented ions were acquired in the Orbitrap analyzer with resolution of 7,500 at *m/z* 400.

Liquid Chromatography and MS for Single-shot Experiments—The peptides were separated on an in-house-made 50-cm column with a 75-μm inner diameter packed with 1.8 μm C_{18} resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany). The Thermo EASY-nLC 1000 system with a binary buffer system consisting of 0.5% formic acid (buffer A) and 80% acetonitrile in 0.5% formic acid (buffer B) was used for reverse phase chromatography. Peptides (~4 μg) were eluted with a 220-min linear gradient of buffer B up to 30% at a flow rate of 250 nl min^{-1}. The column temperature was kept at 40 °C by an in-house designed oven with a Peltier element (26). The LC was coupled to a Q Exactive mass spectrometer (27) (Thermo Fisher Scientific) via the nanoelectrospray source (Proxeon Biosystems, now Thermo Fisher Scientific). Mass spectra were acquired on the Q Exactive in a data-dependent mode with an automatic switch between a full scan and up to 10 data-dependent MS/MS scans. Target value for the full scan MS spectra was 3,000,000 with a maximum injection time of 20 ms and a resolution of 70,000 at *m/z* 400. The 10 most intense ions with charge two or more from the survey scan were selected with an isolation window of 1.6 Th and fragmented by higher energy collisional dissociation (25) with normalized collision energies of 25. The ion target value for MS/MS was set to 1,000,000 with a maximum injection time of 60 ms and a resolution of 17,500 at *m/z* 400. These settings lead to constant injection times of 60 ms, fully in parallel with transient acquisition of the previous scan, ensuring fast cycle times. Repeat sequencing of peptides was kept to a minimum by dynamic exclusion of the sequenced peptides for 25 s.

Data Analysis—The acquired raw files were analyzed by MaxQuant (28) (version 1.2.0.34). Andromeda, a probabilistic search engine incorporated into the MaxQuant framework (29), was used to search the peak lists against the IPI human database version 3.68 which contains 87,083 entries. Common contaminants were added to this database. The search included cysteine carbamidomethylation as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. The second peptide identification option in Andromeda was enabled (29). For statistical evaluation of the data obtained, the posterior error probability and false discovery rate were used. The false discovery rate was determined by searching a reverse database. A false discovery rate of 0.01 for proteins and peptides was required. Enzyme specificity was set to trypsin allowing N-terminal cleavage to proline. Two miscleavages were allowed, and a minimum of six amino acids per identified peptide were required. Peptide identification was based on a search with an initial mass deviation of the precursor ion of up to 6 ppm, and the allowed fragment mass deviation was set to 20 ppm. The mass accuracy of the precursor ions was improved by retention time-dependent mass recalibration (28). To match identifications across different replicates and adjacent fractions, the "match between runs" option in MaxQuant was enabled within a time window of 2 min. Quantification of SILAC pairs was performed by MaxQuant with standard settings using a minimum ratio count of 2. Bioinformatics analysis was done with Perseus tools available in the MaxQuant environment.

When needed for the analysis, the missing values were replaced using data imputation. The idea of our algorithm for imputation of missing values is that they should simulate signals of low abundant proteins. To accomplish this, we first determine the mean and standard deviation of all valid values in the matrix. Then we draw numbers for the missing entries from a suitable probability distribution in an independently, identically distributed way. For that purpose, we use a normal distribution with a mean and standard deviation adjusted in such a way as to simulate signals of low abundant proteins. This is necessary because the missing values are biased toward the detection limit of the LC-MS/MS measurement. Optimal values for the down shift parameter were adjusted in a way that the distribution of imputed values adjusts smoothly to the lower end of the distribution of measured values. We iteratively adjusted the values to avoid too large or too small down shifts. The former would result in a separation of imputed and measured values (a bi-modal total distribution), whereas the latter would introduce too much noise into the system and would potentially destroy protein signatures. The two values for downshifting and width adjustment are determined once but then apply to all the cell lines. These optimal values were different for the label-free and SILAC reference cases. For label-free data, we employed a width of 0.3 and a downshift of 1.8; in the super-SILAC data, the width was 0.3, and the downshift was 0.5, each in units of the standard deviation of the distribution of present values.

RESULTS AND DISCUSSION

Development of a Lymphoma Super-SILAC Mix—To accurately quantify proteome differences between lymphoma subtypes, we set out to generate a super-SILAC mix that would be optimally suited as an internal standard for a broad range of B-cell lymphomas. We considered commonly used cell lines derived from patients with different types of the disease. First we selected two lines, L428 and L1236, to represent Hodgkin's lymphoma. Of the non-Hodgkin's lymphomas, we selected three cell lines of patients with Burkitt's lymphoma, which is characterized by a c-Myc t(8;14) translocation. For DLBCL, we started out with the five ABC type cell lines and five GCB type cell lines that we wished to segregate by proteomics. From these, we chose two ABC type cell lines (Oci-Ly3 and U2932), as well as two GCB type cell lines (BJAB and DB).

Next, we wished to select an optimal subset of these nine representative cell lines (*green* in Fig. 1*A*). Instead of empirically testing different combinations, we reasoned that an in-depth proteome of each of the nine cell lines should be sufficient to mathematically determine the best combination. For this purpose, we performed a six-fraction FASP-SAX-based analysis, with 4-h gradients on an LTQ-Orbitrap Velos and higher energy collisional dissociation-based fragmentation (Fig. 1*A* and "Experimental Procedures"). This involved a single day of measurement time for each of the nine proteomes.

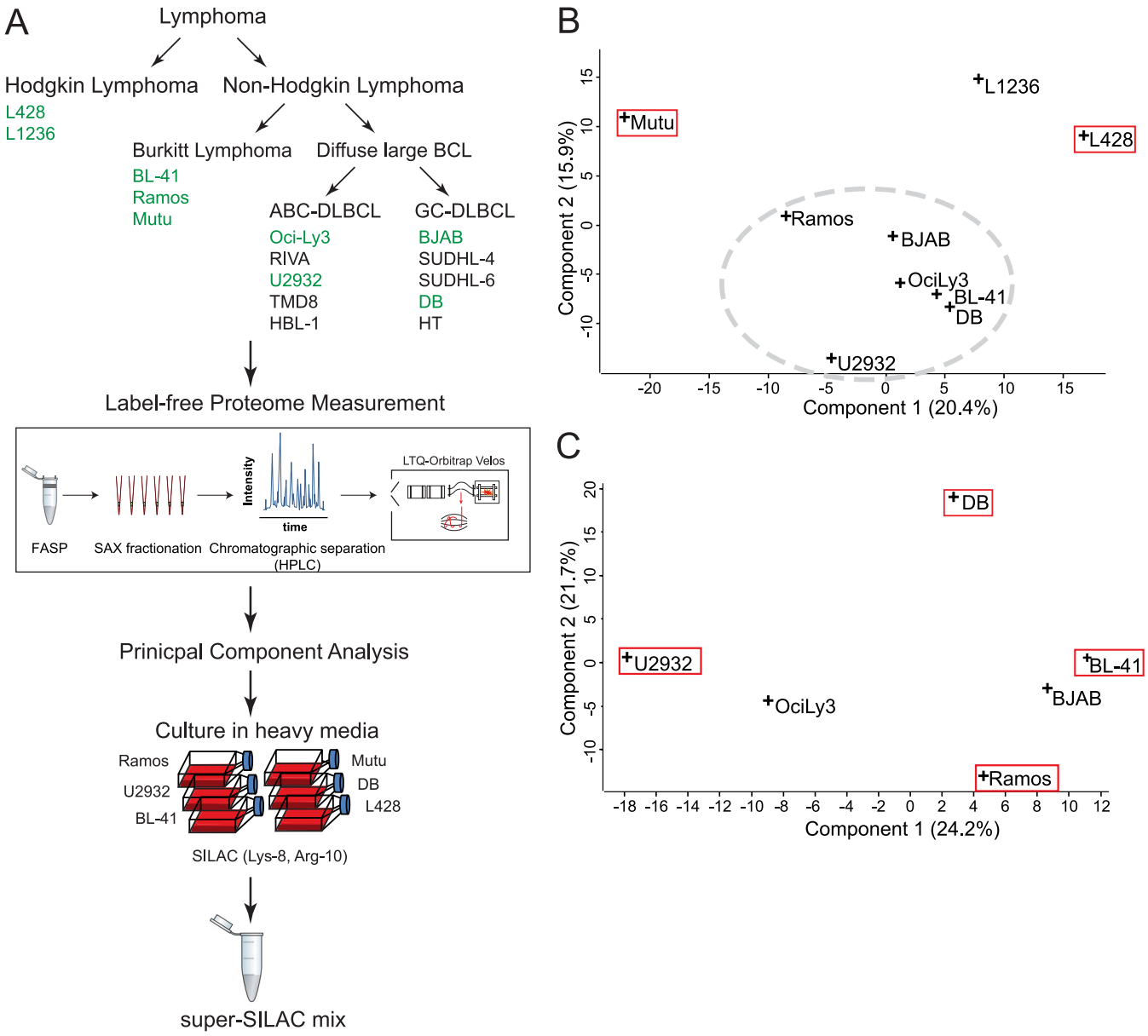


FIG. 1. **Rational construction of lymphoma super-SILAC mix.** *A*, label-free proteomics of nine B-cell lymphoma cell lines was performed after FASP-SAX processing and analyzed using high resolution precursor and fragment measurements on an Orbitrap Velos. They included two Hodgkin lymphoma cell lines (L428 and L1236) and seven non-Hodgkin lymphoma cell lines (Ramos, Mutu, BL-41, OciLy3, U2932, BJAB, and DB). *B*, PCA of nine B-cell lymphoma cell lines based on their protein expression profiles. The *red boxes* indicate cell lines selected for the super-SILAC mix. The *gray dashed ellipse* groups non-Hodgkin lymphoma cell lines to be further analyzed by a second PCA. *C*, PCA of the six non-Hodgkin lymphoma cell lines encircled in *B*. The *red boxes* indicate cell lines selected for the super-SILAC mix.

To compare the label-free proteomes of the cell lines to each other, we performed principal component analysis (PCA). PCA transforms large data sets into points in a data space of orthogonal components, such that the first component captures most of the variability. Because PCA analysis requires a complete data set (in this case label-free protein intensities for all identified proteins in all samples), we employed "data imputation" as described in "Experimental Procedures." We aimed to create a mixture of cells that capture the largest diversity. Therefore, we searched for those that were most distant from one another. Mutu(−), one of the Burkitt-derived cell lines, was the furthest outlier (Fig. 1*B*). L1236 and L428, the only Hodgkin's lymphoma cell lines, were also outliers. We therefore selected Mutu(−) and one of the two Hodgkin's lymphoma cell lines (L428). We then performed a second round of PCA on the remaining seven non-Hodgkin cell lines and selected the four outermost in the resulting PCA space (U2932, DB, BL-41, and Ramos) (Fig. 1*C*).

To produce the super-SILAC mix from the selected six cell lines, we grew them in heavy SILAC media and mixed them in equal proportions. For a first evaluation, we spiked the mix

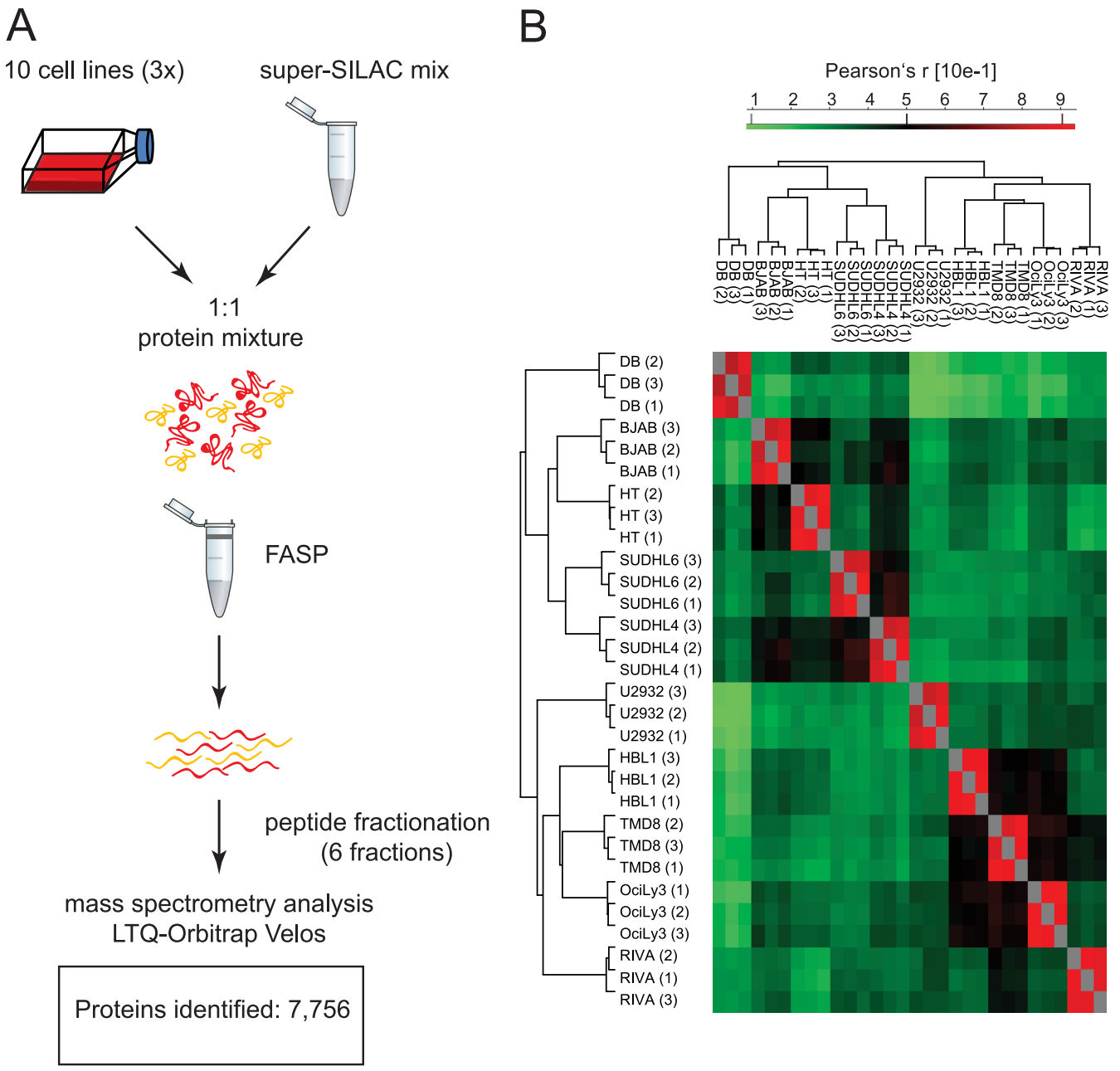


FIG. 2. **Proteomic workflow and overall results.** *A*, the super-SILAC mix developed on the basis of label-free proteome comparison was used as an internal standard for 10 different DLBCL cell lines. The samples were processed by FASP-SAX followed by triplicate 1-day proteome analyses. *B*, heat map of Pearson correlation coefficients showing reproducibility between replicates.

into lysate of an unlabeled ABC-DLBCL cell line (HBL-1) that was not part of the original selection. The histogram of fold changes between cell line proteins and super-SILAC proteins was narrow, with 96% of the values within a 4-fold range. To check the overall selection procedure, we also performed this experiment with a mix of all nine initially selected cell lines. The width of the distribution was essentially unchanged, indicating that the six-cell line mix already adequately represented the proteome. Finally, a three-cell line mix of only the largest outliers of the PCA analysis (Mutu, L428, and U2932; Fig. 1*B*) also performed surprisingly well, attesting to the usefulness of our selection procedure (supplemental Fig. 1).

In-depth Proteome Coverage Using the Lymphoma Super-SILAC Mix—We spiked the super-SILAC mix into five unlabeled ABC and five GCB cell lines and analyzed them as described above for the label-free experiment, except that each proteome was measured in triplicate (Fig. 2*A*). Joint analysis of the resulting 180 LC MS/MS files (30 days measuring time) in MaxQuant identified a total of 7,756 different protein groups, by far the largest B-cell lymphoma proteome

reported to date. Of these proteins, 6,263 were quantified in at least two replicates of the same cell line (supplemental Tables I and II).

At this depth of proteome coverage, we identified and quantified a large number of known members of the B-cell receptor-initiated signal transduction pathway (supplemental Fig. 2). Likewise, many transcription factors relevant to B-cell biology were measured. Altogether, we identified 285 proteins annotated by Gene Ontology to have sequence-specific DNA-binding transcription factor activity (supplemental Table II). This list included many transcription factors playing important roles in B-cells, such as basic leucine zipper transcription factor ATF-like (BATF), B-cell lymphoma 3 protein (BCL3), B-cell lymphoma 6 protein (BCL6), immunoglobulin transcription factors 1 and 2 (ITF1 and ITF2), Ets domain-containing PU.1, and the B-lineage specifying transcription factor PAX-5.

Next, we quantified all 30 proteome measurements against each other based on the ratios to the super-SILAC mix and calculated their Pearson correlation coefficients (r). Unsupervised clustering of the rows and columns of the matrix of the 30×30 coefficients co-clustered the triplicates in each case (Fig. 2*B*). Good reproducibility is further indicated by the high average Pearson coefficients of the triplicates ($r = 0.87$).

Segregation between DLBCL Subtypes—To investigate whether our proteomics measurements can segregate ABC from GCB proteomes and to determine an optimal data analysis strategy, we started by performing unsupervised hierarchical clustering of all proteome measurements. We required that proteins were present in at least 50% of the 30 measurements and filled any missing values by "data imputation" ("Experimental Procedures"). Again, replicate measurements were always clustered together. Intriguingly, the two major branches of the dendrogram precisely grouped all the ABC and all the GCB subtypes together and apart from each other. This indicates that these subtypes have quite different protein expression patterns at a global level that are capable of defining them as distinct entities.

The cluster indicated with *arrow B* in Fig. 3*A*, consists of 107 proteins, 70 of which are annotated as ribosomal, 12 of which are components of the 20 S proteasome, and 14 of which are components of the 26 S proteasome (CORUM annotation) (15). As shown in Fig. 3*B*, their expression varies little across the cell lines; thereby they serve as "loading controls" and validate correct normalization and imputation of the proteome samples by MaxQuant. This ensures that the variation of protein expression values between ABC and GCB can directly be attributed to biological differences between these cell types rather than experimental artifacts. Fig. 3*C* shows the differences in expression of two clusters in the upper part of Fig. 3*A* (indicated with *arrows C* and *D*) with large differential expression patterns between the two main branches of the dendrogram. The first cluster consists of 16 proteins that are up-regulated in the ABC subtype relative to GC. This cluster includes proteins such as CD44, FOXP1, IL4I1, VAV2, and BID (supplemental Table IV). The second cluster consists of 19 proteins that are up-regulated in the GCB subtype and includes proteins such as CD81, KIND3, WIP, INPP5B, PAG, and BRDG1 (supplemental Table V).

Principal component analysis was performed to project the SILAC-based proteome measurements into a two-dimensional data space. We first applied PCA for the subgroup of proteins that were quantified in each of the 30 proteome measurements (100% valid values; 3,007 protein groups). Component 1 of the PCA, which accounts for 20.5% of total variability (*horizontal axis* in the two-dimensional plot of Fig. 4*A*), clearly separates GCB (group on the *left side*) from ABC (group on the *right side*). Furthermore, Fig. 4*A* shows that the distance between the replicates is much smaller than the separation between the groups, supporting the robustness of the segregation.

The proteins that are most responsible for separating the proteomes in the PCA can be seen in the "loadings." The loadings of component 1, which capture the differences between the two groups, include the transcription factor IRF4, mentioned above as one of the main drivers of the functional differences between GCB and ABC lymphomas (Fig. 4*B*). In fact, high expression of IRF4 in ABC-DLBCL is tied to the constitutive activity of NF-κB that is required for survival of this subtype of lymphoma cells (15). This transcription factor, which was quantified in 30 of 30 proteomes, is the strongest differentiator in this unbiased large scale analysis. PTP1B was another one of the strongest loadings of component 1. PTP1B is a key tyrosine phosphatase implicated in the regulation of JAK/STAT signaling. The preferential expression of PTP1B in ABC-DLBCL is already known, and its overexpression has been suggested to contribute to the enhanced STAT6 dephosphorylation that is observed in these tumors upon IL-4 stimulation (30, 31).

The above analysis required quantification of the proteins in every proteome measurement, which could exclude many interesting proteins, such as those exclusively expressed in only one subtype. We therefore employed imputation of missing values to make a larger subset of the proteome amenable to PCA analysis. We first filtered for at least 50% valid values (4,991 proteins) and imputed the missing values. Incorporation of the information from these additional proteins led to an even stronger separation of the subtypes (Fig. 4*C*). The GCB cell lines appear to cluster more tightly together, whereas the ABC cell lines U2932 and RIVA are somewhat separated from the other ABC cell lines. The loadings in Fig. 4*D* reveal additional known markers such as the cell surface markers CD44 for ABC-DLBCL (quantified exclusively in ABC) and CD27 for GCB-DLBCL (Fig. 4*D*). The above analysis demonstrates that requiring less than 100% valid values and imputing missing values is a valid and robust strategy for segregation of subtype groups, as well as for finding individual differentiators by proteomics.

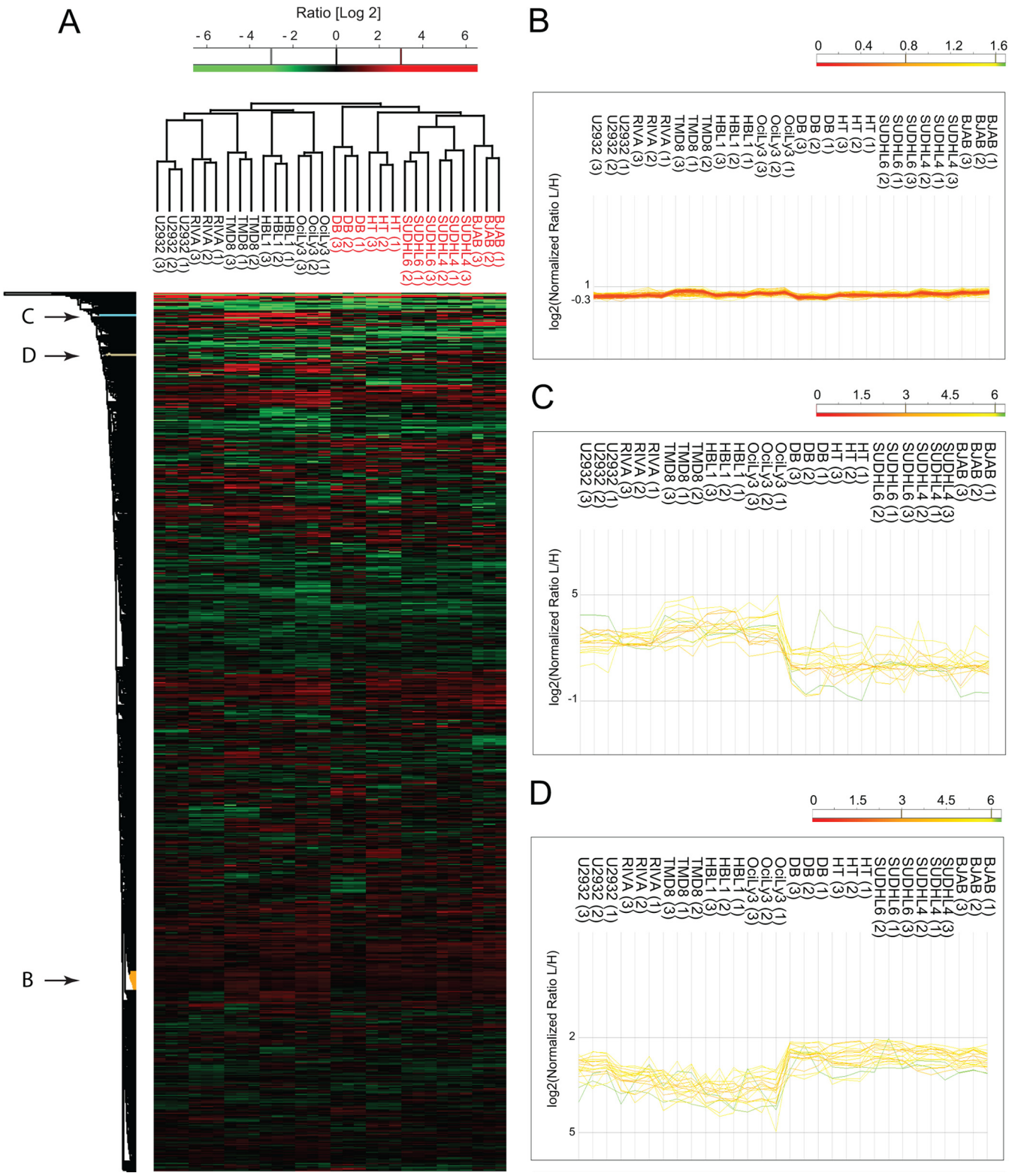


FIG. 3. **Unsupervised hierarchical clustering.** *A*, unsupervised clustering of protein expression profiles of 10 DLBCL cell lines after filtering for 50% valid values and imputation of missing values. *B*, expression patterns for a cluster enriched for ribosomal and proteasomal proteins. *C*, expression patterns for a cluster of proteins with higher expression levels in ABC relative to GCB. *D*, expression patterns for a cluster of proteins with higher expression levels in GCB relative to ABC.

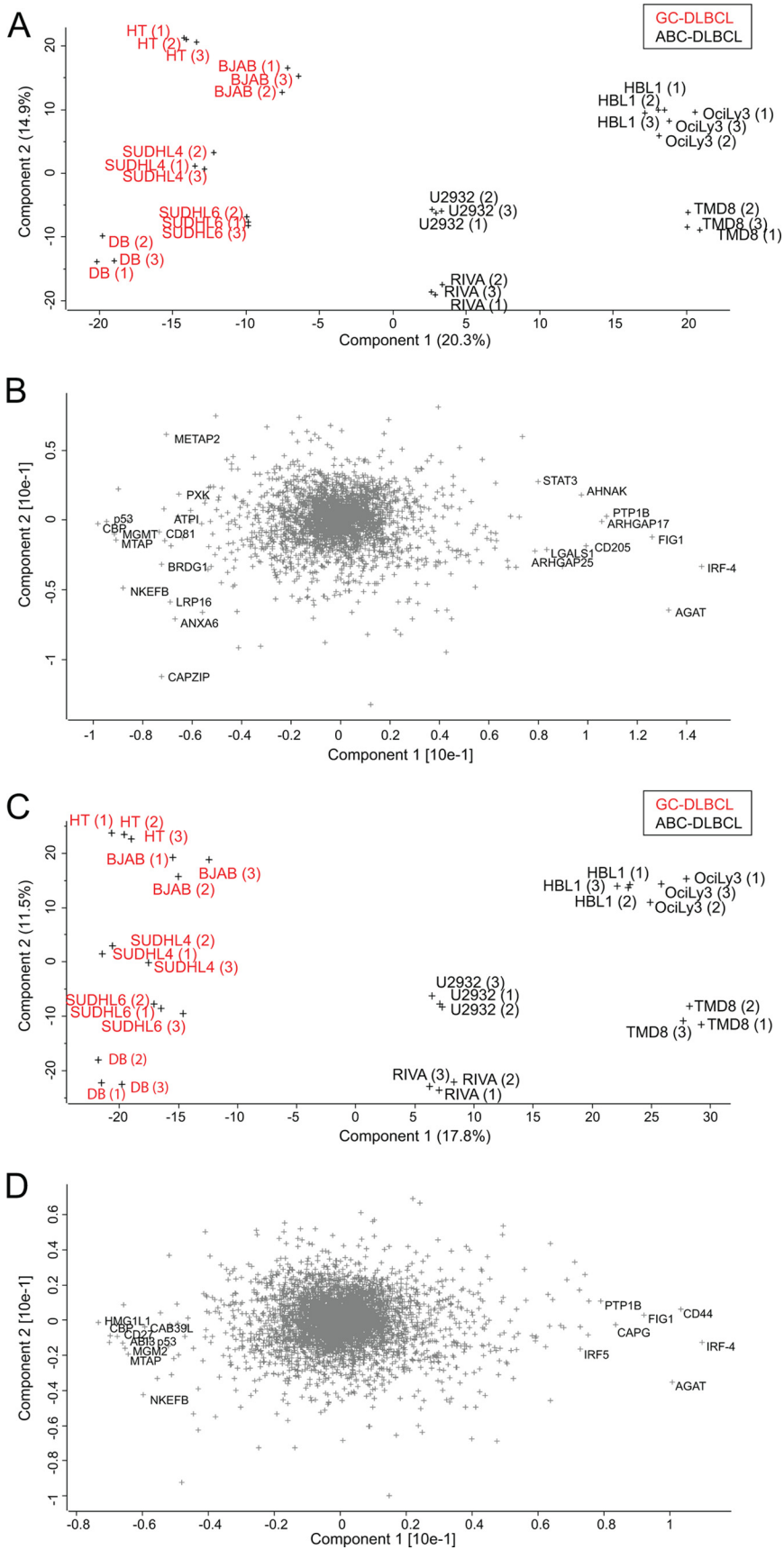


FIG. 4. **Principal component analysis.** *A*, the proteomes of 10 DLBCL cell lines measured in triplicate segregated into ABC-DLBCL and GCB-DLBCL subtypes after filtering for 100% valid values (3,007 proteins). *B*, loadings of *A* reveal proteins that strongly drive the segregation in PCA component 1. *C*, the same analysis as in *A* but after filtering for 50% valid values (4,991 proteins) and filling the missing values by data imputation results in even stronger separation. *D*, loadings of *C* uncover additional known and unknown markers that segregate the ABC and GCB subtypes.

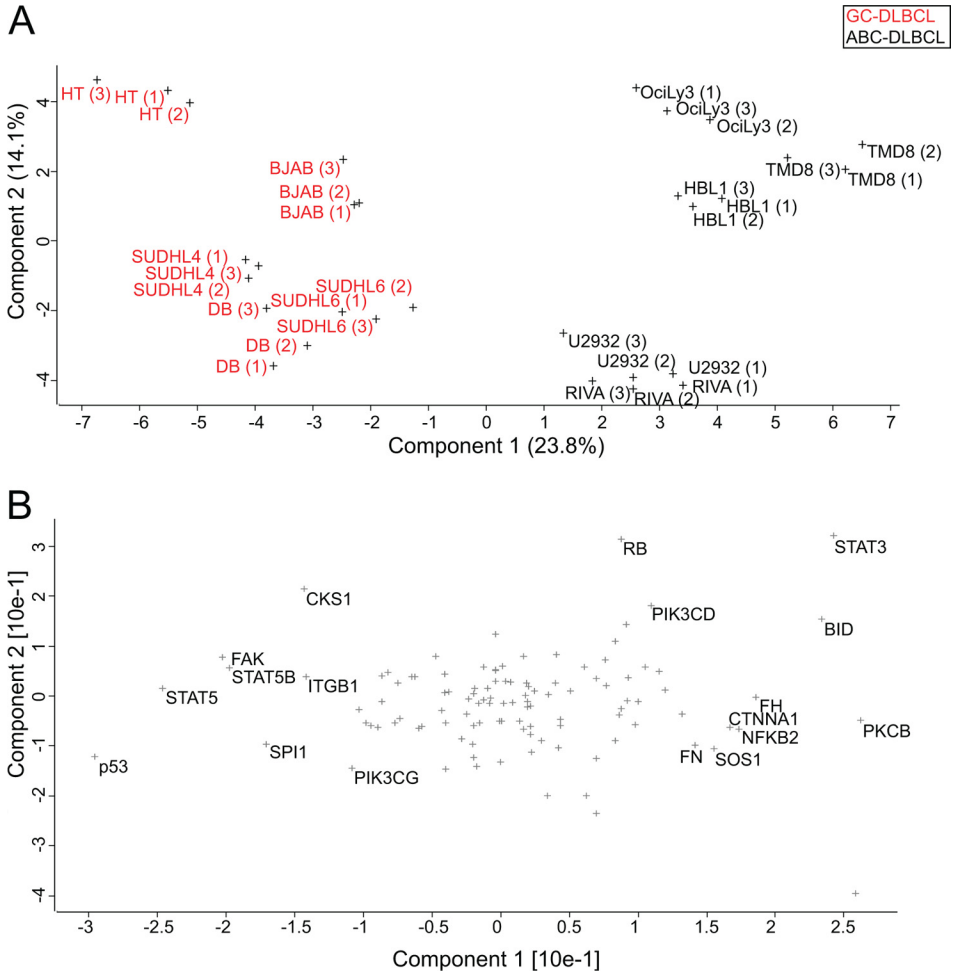
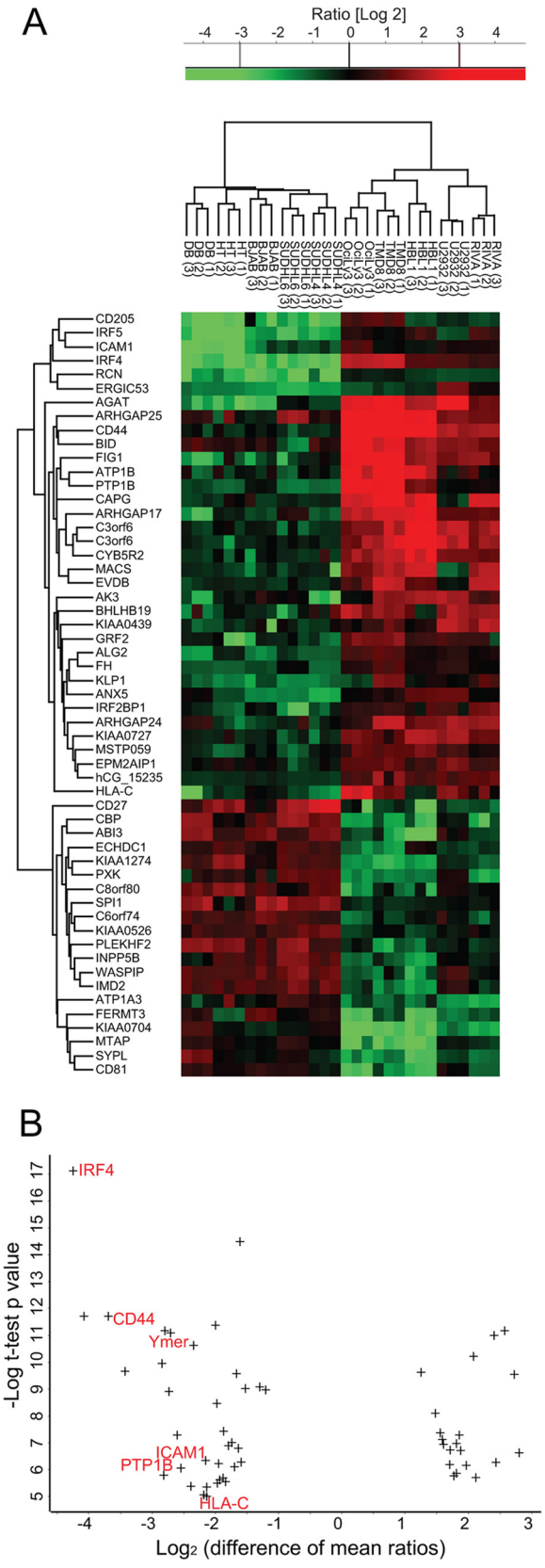


FIG. 5. **Category-based analysis of subtype differences.** *A*, PCA of 10 lymphoma cell lines after filtering for proteins annotated by KEGG to be involved in cancer (KEGG category: pathways in cancer). *B*, loadings of PCA in *A*.

Category-based Analysis of Subtype Differences—The above analyses were global and unbiased in that they considered the entire proteome. To determine whether specific groups of proteins by themselves could differentiate the subtypes, we extracted the proteins belonging to specific KEGG categories from the quantified proteome. We then performed PCA analysis on this subset as described above. Interestingly, the category "pathways in cancer" (108 quantified proteins) was able to clearly separate the groups, albeit to a lesser degree than the full proteome (Fig. 5*A*). The strongest loadings preferential for GCB in this category were p53, STAT5, STAT5B, and SPI1/PU.1 (Fig. 5*B*). SPI1 has a major role in maintaining germinal center B-cells through repressing the expression of plasma cell transcriptional regulators and thus blocking plasma cell differentiation (32). The strongest loadings preferential for ABC included the anti-apoptotic protein BCL2, overexpression of which is a known mechanism by which NF-κB driven tumors evade apoptosis. Surprisingly, the pro-apoptotic protein BID was also in this group. The loadings preferential for ABC include STAT3. It has been shown that NF-κB signaling in ABC induces the expression of IL-6 and IL-10, which act through JAK kinases and STAT3 as autocrine signals (33). The constitutive activity of STAT3 promotes proliferation and cell survival in the ABC subtype (34). This explains the synergistic effect of blocking JAK signaling and NF-κB signaling in killing ABC cells (33). PKCB is another interesting driver of the ABC subtype because its overexpression is a strong marker for refractory or fatal DLBCL and a recognized drug target (35). Our finding that it is preferentially expressed in the aggressive ABC subtype compared with the GCB subtype may therefore be of clinical interest. The observation that a small group of proteins can separate the subtypes prompted us to search for such groups in the entire quantified data set.

t Test Signature—To identify in a supervised manner a set of proteins that significantly distinguishes the ABC from the GCB subtype, we performed a *t* test between the cell lines using a permutation-based false discovery rate (0.05). This analysis resulted in a set of 55 proteins (Fig. 6*A*) that strongly segregated the subtypes as seen after PCA analysis (supplemental Fig. 3). Interestingly, cell lines of the GCB subtype collapse into a single cluster, indicating that the proteins most



strongly differentiating ABC and GCB subtypes do not distinguish different GCB cell lines from each other (variation between cell lines is equal to the variation between replicates). In contrast, replicates of ABC cell lines remained distinguishable, which indicates a larger degree of heterogeneity in ABC-type cell lines as we observed previously.

In the signature obtained by unbiased proteomic analysis, there are at least six proteins whose gene expression levels are already described to be different between the subtypes (ABC-like: IRF4, CD44, and PTP1B; GCB-like: CD27, SPI1, and WIP). Because the total number of signature proteins is small, this already validates our proteomic signature and encouraged us to further investigate the new proteins in our signature. These include the recently described GTPase Speckled-like pattern in the germinal center (SLIP-GC), whose expression is limited to germinal center B-cells and to lymphomas derived from the germinal center including diffuse large B-cell lymphomas (36). This finding supports the potential use of SLIP-GC as a potential marker that can be used to differentiate the two subtypes from each other. Another member of the signature set is the surface marker CD81, which has also very recently been reported to be highly expressed in normal germinal B-cells. Further assessment of the role of this cell surface marker in the risk stratification of patients with DLBCL has already been recommended (37). A further novel protein that has a higher expression level in our GCB-DLBCL signature is the signaling adaptor Cbp/PAG. In B-cell non-Hodgkin lymphomas, PAG and Lyn kinase constitute the core of an oncogenic signalosome that results in proliferative and pro-survival signals. The Lyn and PAG signalosome can interact with downstream kinases to mediate these signals in different lymphoma cell lines (38). Our finding that PAG is up-regulated in GCB suggests investigating the modality associated with PAG in this subtype in particular. Ymer, a protein that we previously identified as an effector of EGF signaling (10, 39), is relatively up-regulated in the ABC subtype. Ymer is also known as CCDC50 or C3orf6, and although not studied in the context of DLBCL, this protein has been shown to be required for cell survival in chronic lymphocytic leukemia and mantle cell lymphoma cells (40). It is involved in the control of NF-κB signaling, which is a characteristic of the ABC subtype where it is up-regulated (15). Therefore, in addition to validating differentiators known from gene expression profiling, our proteomic signature reveals a novel set of proteins,

FIG. 6. ***t* test signature.** *A*, *t* test analysis of the proteins from the two groups of cell lines resulted in a signature of 55 proteins that are most significantly different. The panel depicts a heat map of the ratios of these proteins after clustering. *B*, plot of the difference of mean ratios *versus* the significance of signature proteins. The proteins on the *left* are significantly up-regulated in the ABC relative to GCB subtype. The protein names highlighted in *red* indicate NF-κB regulated genes.

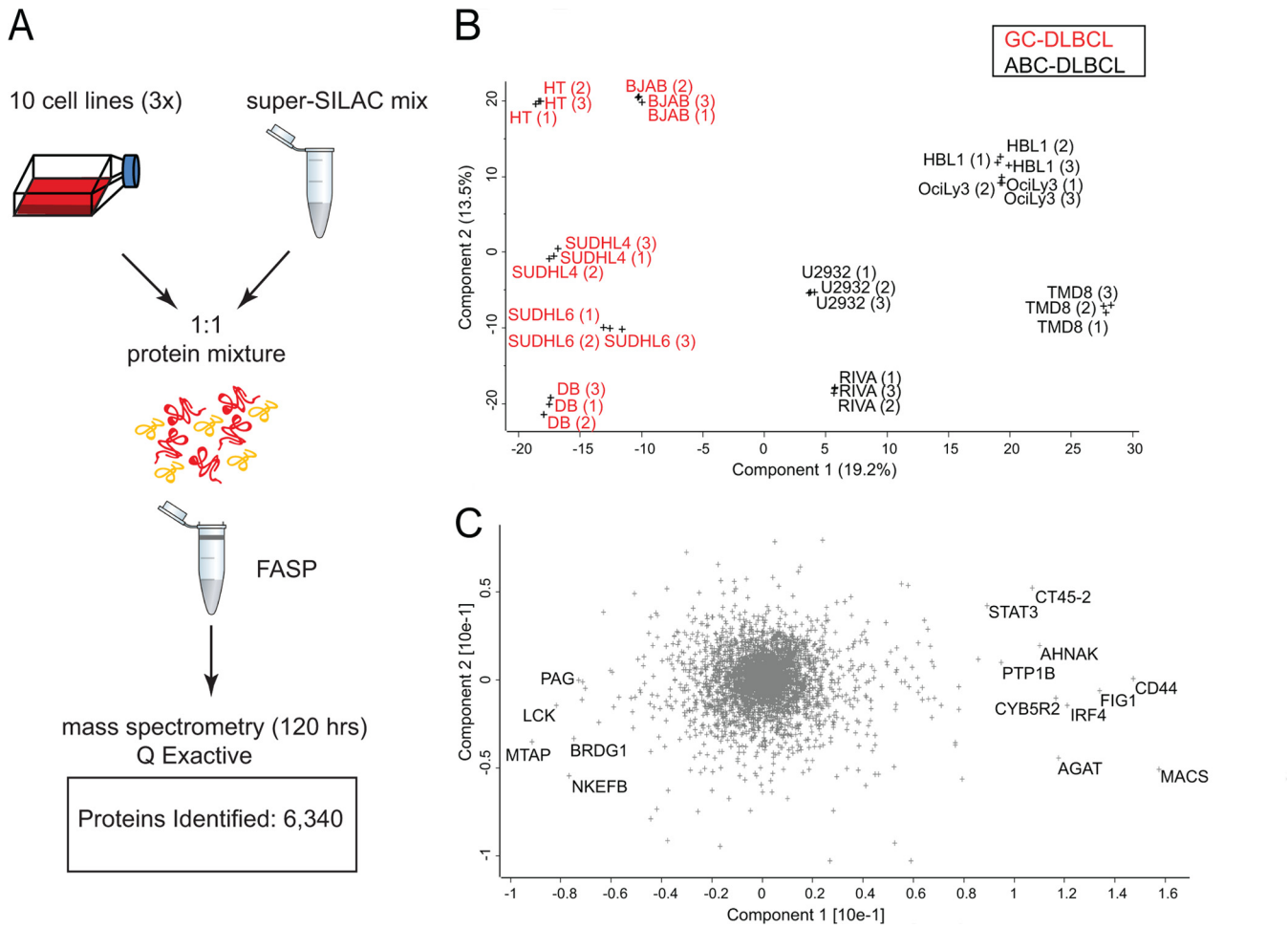


FIG. 7. **Single-shot proteome measurements to distinguish ABC from GCB.** *A*, unfractionated, FASP-processed peptide mixtures were directly loaded onto a relatively long column (50 cm) after StageTipping. The proteomes were analyzed in triplicates in 4-h runs by an UHPLC (EASY nLC 1000) system coupled to a benchtop quadrupole Orbitrap mass spectrometer (Q Exactive). *B*, principal component analysis of the single-shot measurements. *C*, loadings of PCA in *B* highlighting the proteins that strongly drive the segregation in PCA component 1.

some of which have been shown to be involved in lymphomagenesis and might be of clinical relevance.

In an attempt to identify annotated protein categories that are significantly and exclusively up-regulated in one lymphoma subtype relative to the other, we plotted the difference of mean ratios *versus* the statistical significance of our signature proteins. This revealed that all NF-κB-regulated proteins in the signature are significantly up-regulated in ABC relative to GCB subtype (Fig. 6*B*). They included IRF4, CD44, Ymer, PTP1B, ICAM, and HLA-C. Thus, our proteomic findings, similar to previous results of gene expression profiling, are consistent with high NF-κB activity in ABC-DLBCL as a hallmark that distinguishes this subtype from GCB-DLBCL.

BCR signaling has been shown to play an important role in lymphomagenesis where malignant B-cells exploit the normal regulatory roles of this pathway for their own purposes (41). We extracted proteins from our data set that are KEGG annotated to be involved in BCR signaling and found that we covered all but eight proteins in this category (59 of 67; supplemental Fig. 2). To better investigate small but reproducible protein changes, we normalized their expression levels by z-scoring across replicates and cell lines. Taking the median values for every subtype revealed four large clusters. The proteins highlighted in *red* in supplemental Fig. 4*A* consist of the BCR signaling proteins that exhibit the largest expression differences between the two subtypes and that are higher in the ABC subtype. Interestingly, this cluster included NF-κB1, as well as the two upstream regulatory proteins MALT1 and CARD11 (supplemental Fig. 4*B*). This is consistent with the role of the multiprotein CARD11-BCL10-MALT1 (CBM) complex in driving the constitutive NF-κB activity in the ABC subtype (17–19). Conversely, the proteins (highlighted in *green* in supplemental Fig. 4*A*) are BCR signaling proteins that are higher in GCB (supplemental Fig. 4*C*).

Rapid Lymphoma Classification in Single-shot Runs—Above, we have demonstrated that quantitative proteomics can readily segregate cell lines derived from patients in a robust manner. However, sample amounts and measurement

time of our workflow (Fig. 1) would still be an obstacle to clinical application. We therefore wanted to investigate the possibility of making the approach more practical by reducing the measurement time and the amount of sample consumed. Taking advantage of the higher speed and sensitivity of the newly introduced quadrupole Orbitrap mass spectrometer (Q Exactive) (27), we investigated whether we could reach the depth required to segregate the cell lines in a single-shot experiment, that is, without fractionation. The samples were prepared as before, except that FASP-prepared peptides were directly loaded on StageTips and eluted into the autosampler device. Single 4-h gradient runs were performed in triplicates for each of the 10 cell lines, and data files were processed together in MaxQuant. This resulted in the identification of 6,340 proteins and the quantification of 4,611 in at least two replicates of the same cell line (Fig. 7*A*) (supplemental Tables VI and VII). Filtering for 50% valid values resulted in 3,566 quantified proteins. Upon PCA analysis of the single measurements, we obtained a similar segregation of the two subtypes, and the loadings responsible for the PCA segregation showed a very strong overlap with the previously obtained loadings (Fig. 7*B*). Interestingly, the data obtained from singlets was sufficient to segregate the two subtypes. This shows that single-shot measurements can reach the depth required for robust separation of lymophoma subtypes, opening up for the analysis of several patient samples per day with sample requirements in the low microgram range.

Conclusion and Outlook—Here we have shown that high accuracy, quantitative proteomics based on a super-SILAC approach can robustly segregate closely related cancer subtypes directly at the level of expressed proteins. We developed and used a super-SILAC mix of labeled B-cell lymphoma cell lines as an internal standard to segregate subtypes of DLBCL. We selected the cell lines with the most distinct protein expression profiles to obtain the best coverage of different lymphoma-specific proteins. The mix was spiked into five ABC-DLBCL and five GCB-DLBCL cell lines, which allowed robust, unsupervised segregation of these two histologically indistinguishable lymphomas based on their protein expression profiles. We found that requiring protein quantification values to be present in half of the samples and replicates and imputing the remaining values led to the most robust segregation. The data also revealed a protein expression signature that differentiates the two subtypes. This signature confirmed known markers previously discovered by gene expression studies and highlighted novel ones. Interestingly, our straightforward PCA analysis of the proteome differences revealed proteins such as IRF4, CD44, STAT3, PTP1B, and CD27 as the strongest differentiators between subtypes. The fact that these and a number of other proteins, which all have a strong biological rationale to drive subtype differences, emerge as the top hits in an unbiased analysis, is very intriguing. Furthermore, unbiased and supervised segregation revealed a number of novel proteins, which can now be studied for their involvement in these lymphoma subtypes.

To our knowledge, this is the first high accuracy, quantitative proteomics study that unequivocally classified tumor cell lines on par with microarray-based methods. This ability of the super-SILAC proteomic approach to readily segregate between tumor subtypes now opens up the possibility of employing proteomics in many situations that have previously been studied with transcript-based approaches. Toward this goal, we already combined the super-SILAC quantitative approach with single-shot measurements on a benchtop quadrupole Orbitrap instrument. These measurements attained the depth and accuracy required to segregate the two subtypes as exemplified by a number of representative cell lines. Considering the significant reduction in measuring time and in required sample amount, it is conceivable that this workflow could be employed in routine settings to answer practical clinical questions such as tumor classification or drug efficacy.

Acknowledgments—We thank Daniel Krappmann, Berit Jungnickel, Ralf Küppers, Martin Janz, Stephan Mathas, and Georg Lenz for the provision of cell lines. We thank Tamar Geiger and Maria Robles (Charo) for helpful discussions.

The acquired raw data for the fractionated proteome is uploaded to Tranche (https://proteomecommons.org/tranche/) with the hash code: Ib9WXUuYMjCxiCk40HNWo3YM+xvP31WEgYg6RWMjm5295exN5kmoMbyekZXqMvXpG8rvLgDELBXBKcceiScd+R2WqDkAAAAAAACQyQ==.

* This work was supported by European Commission's 7th Framework Program PROSPECTS Grant Agreement HEALTH-F4-2008-201648 and a Deutsche Krebshilfe Onconet2 grant. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S This article contains supplemental materials.

¶ Supported by Deutsche Forschungsgemeinschaft Grant TRR54.

‖ To whom correspondence should be addressed. Tel.: 49-89-8578-2557; E-mail: mmann@biochem.mpg.de.

REFERENCES

1. Quackenbush, J. (2006) Microarray Analysis and Tumor Classification. *New Engl. J. Med.* **354,** 2463–2472
2. McDermott, U., Downing, J. R., and Stratton, M. R. (2011) Genomics and the Continuum of Cancer Care. *New Engl. J. Med.* **364,** 340–350
3. Hanash, S., and Taguchi, A. (2010) The grand challenge to decipher the cancer proteome. *Nat. Rev. Cancer* **10,** 652–660
4. Choudhary, C., and Mann, M. (2010) Decoding signalling networks by mass spectrometry-based proteomics. *Nat. Rev. Mol. Cell Biol.* **11,** 427–439
5. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422,** 198–207
6. Nilsson, T., Mann, M., Aebersold, R., Yates, J. R., 3rd, Bairoch, A., and Bergeron, J. J. (2010) Mass spectrometry in high-throughput proteomics: Ready for the big time. *Nat. Methods* **7,** 681–685
7. Cox, J., and Mann, M. (2011) Quantitative, high-resolution proteomics for data-driven systems biology. *Annu. Rev. Biochem.* **80,** 273–299
8. Vermeulen, M., and Selbach, M. (2009) Quantitative proteomics: A tool to assess cell differentiation. *Curr. Opin. Cell Biol.* **21,** 761–766
9. Mallick, P., and Kuster, B. (2010) Proteomics: A pragmatic perspective. *Nat. Biotechnol.* **28,** 695–709
10. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H.,

Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1,** 376–386
11. Mann, M. (2006) Functional and quantitative proteomics using SILAC. *Nat. Rev. Mol. Cell Biol.* **7,** 952–958
12. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R., and Mann, M. (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **7,** 383–385
13. Geiger, T., Wisniewski, J. R., Cox, J., Zanivan, S., Kruger, M., Ishihama, Y., and Mann, M. (2011) Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nat. Protoc.* **6,** 147–157
14. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–511
15. Davis, R. E., Brown, K. D., Siebenlist, U., and Staudt, L. M. (2001) Constitutive nuclear factor κB activity is required for survival of activated B cell-like diffuse large B cell lymphoma cells. *J. Exp. Med.* **194,** 1861–1874
16. Davis, R. E., Ngo, V. N., Lenz, G., Tolar, P., Young, R. M., Romesser, P. B., Kohlhammer, H., Lamy, L., Zhao, H., Yang, Y., Xu, W., Shaffer, A. L., Wright, G., Xiao, W., Powell, J., Jiang, J. K., Thomas, C. J., Rosenwald, A., Ott, G., Muller-Hermelink, H. K., Gascoyne, R. D., Connors, J. M., Johnson, N. A., Rimsza, L. M., Campo, E., Jaffe, E. S., Wilson, W. H., Delabie, J., Smeland, E. B., Fisher, R. I., Braziel, R. M., Tubbs, R. R., Cook, J. R., Weisenburger, D. D., Chan, W. C., Pierce, S. K., and Staudt, L. M. (2010) Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. *Nature* **463,** 88–92
17. Ngo, V. N., Davis, R. E., Lamy, L., Yu, X., Zhao, H., Lenz, G., Lam, L. T., Dave, S., Yang, L., Powell, J., and Staudt, L. M. (2006) A loss-of-function RNA interference screen for molecular targets in cancer. *Nature* **441,** 106–110
18. Ferch, U., Kloo, B., Gewies, A., Pfänder, V., Düwel, M., Peschel, C., Krappmann, D., and Ruland, J. (2009) Inhibition of MALT1 protease activity is selectively toxic for activated B cell-like diffuse large B cell lymphoma cells. *J. Exp. Med.* **206,** 2313–2320
19. Hailfinger, S., Lenz, G., Ngo, V., Posvitz-Fejfar, A., Rebeaud, F., Guzzardi, M., Penas, E. M., Dierlamm, J., Chan, W. C., Staudt, L. M., and Thome, M. (2009) Essential role of MALT1 protease activity in activated B cell-like diffuse large B-cell lymphoma. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 19946–19951
20. Pasqualucci, L., Trifonov, V., Fabbri, G., Ma, J., Rossi, D., Chiarenza, A., Wells, V. A., Grunn, A., Messina, M., Elliot, O., Chan, J., Bhagat, G., Chadburn, A., Gaidano, G., Mullighan, C. G., Rabadan, R., and Dalla-Favera, R. (2011) Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat. Genet.* **43,** 830–837
21. Wiśniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat. Methods* **6,** 359–362
22. Wiśniewski, J. R., Zougman, A., and Mann, M. (2009) Combination of FASP and StageTip-based fractionation allows in-depth analysis of the hippocampal membrane proteome. *J. Proteome Res.* **8,** 5674–5678
23. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal. Chem.* **75,** 663–670
24. Olsen, J. V., Schwartz, J. C., Griep-Raming, J., Nielsen, M. L., Damoc, E., Denisov, E., Lange, O., Remes, P., Taylor, D., Splendore, M., Wouters, E. R., Senko, M., Makarov, A., Mann, M., and Horning, S. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* **8,** 2759–2769
25. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4,** 709–712
26. Thakur, S. S., Geiger, T., Chatterjee, B., Bandilla, P., Frohlich, F., Cox, J., and Mann, M. (2011) Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Mol. Cell. Proteomics* **10,** M110.003699
27. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011) Mass spectrometry-based proteomics using Q exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol. Cell. Proteomics* **10,** M111.011015
28. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
29. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A peptide search engine integrated into the MaxQuant environment. *J Proteome Res.* **10,** 1794–1805
30. Lu, X., Malumbres, R., Shields, B., Jiang, X., Sarosiek, K. A., Natkunam, Y., Tiganis, T., and Lossos, I. S. (2008) PTP1B is a negative regulator of interleukin 4-induced STAT6 signaling. *Blood* **112,** 4098–4108
31. Lu, X., Nechushtan, H., Ding, F., Rosado, M. F., Singal, R., Alizadeh, A. A., and Lossos, I. S. (2005) Distinct IL-4-induced gene expression, proliferation, and intracellular signaling in germinal center B-cell-like and activated B-cell-like diffuse large-cell lymphomas. *Blood* **105,** 2924–2932
32. Schmidlin, H., Diehl, S. A., Nagasawa, M., Scheeren, F. A., Schotte, R., Uittenbogaart, C. H., Spits, H., and Blom, B. (2008) Spi-B inhibits human plasma cell differentiation by repressing BLIMP1 and XBP-1 expression. *Blood* **112,** 1804–1812
33. Lam, L. T., Wright, G., Davis, R. E., Lenz, G., Farinha, P., Dang, L., Chan, J. W., Rosenwald, A., Gascoyne, R. D., and Staudt, L. M. (2008) Cooperative signaling through the signal transducer and activator of transcription 3 and nuclear factor-κB pathways in subtypes of diffuse large B-cell lymphoma. *Blood* **111,** 3701–3713
34. Ding, B. B., Yu, J. J., Yu, R. Y., Mendez, L. M., Shaknovich, R., Zhang, Y., Cattoretti, G., and Ye, B. H. (2008) Constitutively activated STAT3 promotes cell proliferation and survival in the activated B-cell subtype of diffuse large B-cell lymphomas. *Blood* **111,** 1515–1523
35. Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., and Golub, T. R. (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8,** 68–74
36. Richter, K., Brar, S., Ray, M., Pisitkun, P., Bolland, S., Verkoczy, L., and Diaz, M. (2009) Speckled-like pattern in the germinal center (SLIP-GC), a nuclear GTPase expressed in activation-induced deaminase-expressing lymphomas and germinal center B cells. *J. Biol. Chem.* **284,** 30652–30661
37. Luo, R. F., Zhao, S., Tibshirani, R., Myklebust, J. H., Sanyal, M., Fernandez, R., Gratzinger, D., Marinelli, R. J., Lu, Z. S., Wong, A., Levy, R., Levy, S., and Natkunam, Y. (2010) CD81 protein is expressed at high levels in normal germinal center B cells and in subtypes of human lymphomas. *Human Pathol.* **41,** 271–280
38. Tauzin, S., Ding, H., Burdevet, D., Borisch, B., and Hoessli, D. C. (2011) Membrane-associated signaling in human B-lymphoma lines. *Exp. Cell Res.* **317,** 151–162
39. Kratchmarova, I., Blagoev, B., Haack-Sorensen, M., Kassem, M., and Mann, M. (2005) Mechanism of divergent growth factor effects in mesenchymal stem cell differentiation. *Science* **308,** 1472–1477
40. Farfsing, A., Engel, F., Seiffert, M., Hartmann, E., Ott, G., Rosenwald, A., Stilgenbauer, S., Döhner, H., Boutros, M., Lichter, P., and Pscherer, A. (2009) Gene knockdown studies revealed CCDC50 as a candidate gene in mantle cell lymphoma and chronic lymphocytic leukemia. *Leukemia* **23,** 2018–2026
41. Lenz, G., and Staudt, L. M. (2010) Aggressive lymphomas. *New Engl. J. Med.* **362,** 1417–1429

2.2 N-linked glycosylation enrichment for in-depth cell surface proteomics of diffuse large B-cell lymphoma subtypes

2.2.1 Project aim and summary

After segregating the two morphologically indistinguishable subtypes of diffuse large B-cell lymphoma (DLBCL) based on their global protein expression profiles, we wanted to investigate whether the same could be done by targeting a specific class of functionally relevant proteins.

The repertoire of cell surface proteins provides a unique molecular fingerprint to phenotype cells and cellular states. In particular, membrane proteins may specifically characterize cancer cells and aid in the development of targeted therapies because they are the upstream and unique members of signaling cascades. This is in contrast to downstream-activated signaling pathways whose members can be redundant. Therefore, we focused this study on membrane proteins, which are not only key players in cancer cell biology but are also located at the interface between a cancer cell and its environment.

It is routine to classify different cell types using a few antibodies directed against known proteins. However, discovering differentiating proteins distinguishing closely related tumor-subtypes derived from the same cell type requires a global and unbiased approach. Since glycosylation is a hallmark of membrane proteins, we took advantage of the recently developed N-glyco-FASP enrichment approach to explore the cell surface in a global manner. For quantification, we used a variant of the super-SILAC approach. In this way, we characterized the membrane proteins of five ABC-DLBCL and five GCB-DLBCL patient-derived cell lines. The attained depth and quantification accuracy allowed the correct segregation of the subtypes. Our study also established that related tumor subtypes can be classified by MS-based proteomics on the basis of PTM-bearing peptides. Our results also constitute the largest B-cell lymphoma membrane proteome to date. Importantly, many glycosites, which were identified as strong segregators in this study, were localized on proteins that we had suggested to be markers in our previous proteome

study of the same system. This further validates the clinical potential of these candidate proteins. Interestingly, the differences in the expression levels of membrane glycoproteins reflected tumor associated hallmarks or characteristics of the stage of B-cell development from which these cells are derived. For instance, differential activity of NF-κB signaling between the subtypes was apparent at the level of the membrane proteome.

2.2.2 Contribution

This project was a continuation of the first project. Under the supervision of Matthias Mann and Marc Schmidt-Supprian, I performed and optimized all sample preparation techniques and MS analysis methods. Under the supervision of Juergen Cox, I performed both data acquisition and analysis. I designed all figures and tables for the publication. The manuscript was written by me with the help of Matthias Mann and Marc Schmidt-Supprian.

2.2.3 Publication

This project was published in 2014 as a research article in Molecular and Cellular Proteomics:

N-linked Glycosylation Enrichment for In-depth Cell Surface Proteomics of Diffuse Large B-cell Lymphoma Subtypes

Sally J. Deeb, Juergen Cox, Marc Schmidt-Supprian, and Matthias Mann

Mol Cell Proteomics. 2014 January; 13(1)

This paper is available on line at http://www.mcponline.org

N-linked Glycosylation Enrichment for In-depth Cell Surface Proteomics of Diffuse Large B-cell Lymphoma Subtypes*[S]

Sally J. Deeb‡, Juergen Cox‡, Marc Schmidt-Supprian§, and Matthias Mann‡¶

Global analysis of lymphoma genome integrity and transcriptomes tremendously advanced our understanding of their biology. Technological advances in mass spectrometry-based proteomics promise to complete the picture by allowing the global quantification of proteins and their post-translational modifications. Here we use N-glyco FASP, a recently developed mass spectrometric approach using lectin-enrichment, in conjunction with a super-SILAC approach to quantify *N*-linked glycoproteins in lymphoma cells. From patient-derived diffuse large B-cell lymphoma cell lines, we mapped 2383 glycosites on 1321 protein groups, which were highly enriched for cell membrane proteins. This N-glyco subproteome alone allowed the segregation of the ABC from the GCB subtypes of diffuse large B-cell lymphoma, which before gene expression studies had been considered one disease entity. Encouragingly, many of the glycopeptides driving the segregation belong to proteins previously characterized as segregators in a deep proteome study of these subtypes (S. J. Deeb *et al.* MCP 2012 PMID 22442255). This conforms to the high correlation that we observed between the expression level of the glycosites and their corresponding proteins. Detailed examination of glycosites and glycoprotein expression levels uncovered, among other interesting findings, enrichment of transcription factor binding motifs, including known NF-kappa-B related ones. Thus, enrichment of a class of post-translationally modified peptides can classify cancer types as well as reveal cancer specific mechanistic changes. ***Molecular & Cellular Proteomics 13: 10.1074/mcp.M113.033977, 240–251, 2014.***

Correct classification of cancer subtypes is a long-standing aim for any heterogeneous diagnostic category and is a necessary basis for rational treatment. Diffuse large B-cell lymphoma (DLBCL)[1] is the most frequent subtype of malignant lymphomas and is clinically heterogeneous (1). The molecular characterization of DLBCL based on gene expression profiling led, for the first time, to the identification of distinct DLBCL entities with significant differences in their pathogenesis, response to conventional treatment and clinical outcomes (2). In fact, gene expression signatures correlated these subtypes to distinct stages of B-cell development. Germinal-center B-cell-like DLBCL (GCB) possesses a gene expression signature characteristic of germinal center B cells and has a favorable outcome compared with activated B-cell-like DLBCL (ABC) subtype which possesses a gene expression signature characteristic of B cells activated through their B-cell receptor (2). We have previously demonstrated the ability to segregate these subtypes based on their in-depth protein expression profiles in a cell line model derived from patients (3). Diagnosis in this system is particularly challenging because the two subtypes studied are histologically indistinguishable but could be differentiated by gene expression profiling (2).

The cell surface proteome of B cells plays a very important role in mediating interactions with the surrounding environment and is of particular importance in determining their fate. The B-cell receptor, for instance, is the key functional player on the surface of B cells, responsible for their development, peripheral maintenance and antigen-specific functional response. Other cell surface proteins such as ICAM-1 (CD54) have important roles in mediating the binding of B cells to other cell types. Furthermore, CD40 and CD80 bind to T-cell proteins (CD40L and CD28, respectively) and mediate co-stimulatory signals required for B-cell (and T-cell) activation. The large repertoire of B-cell surface proteins and the complexity of regulation of B-cell activation make the B-cell surface an interesting niche to explore tumorigenic differences.

In classic approaches like flow-cytometry, antibodies directed against known proteins are commonly employed to

From the ‡Proteomics and Signal Transduction Department, Max-Planck Institute of Biochemistry, D-82152 Martinsried, Germany; §Molecular Immunology and Signal Transduction Group, Max-Planck Institute of Biochemistry, D-82152 Martinsried, Germany

Received August 29, 2013, and in revised form, October 31, 2013
Published, MCP Papers in Press, November 4, 2013, DOI 10.1074/mcp.M113.033977

Author contributions: S.J.D., M.S., and M.M. designed research; S.J.D. performed research; S.J.D. and J.C. analyzed data; S.J.D., M.S., and M.M. wrote the paper.

[1] The abbreviations used are: DLBCL, diffuse large B-cell lymphoma; ABC-DLBCL, activated B-cell–like diffuse large B-cell lymphoma; BCR, B-cell receptor; ER, endoplasmic reticulum; ECM, extracellular matrix; FASP, filter-aided sample preparation; GCB-DLBCL, germinal-center B-cell–like DLBCL; MS, mass spectrometry; PCA, principal component analysis; PTM, post-translational modification, SILAC, stable isotope labeling with amino acids in cell culture.

phenotype cells of different origin. This technology requires antibodies with high specificity and allows the multiplexing of up to 18–36 different differentiation markers at a time (4). However, classifying closely-related tumors derived from the same cell type where it is not known which proteins are expressed on the cell surface and to what levels is a more complex problem that first requires an unbiased quantitative in-depth approach to analyze membrane proteins. Taking into consideration that glycosylation is a hallmark of membrane proteins we wanted to investigate the possibility of enriching for glycosylated peptides as a handle to explore the cell surface proteome. In addition, we wanted to ask the question if closely related tumor subtypes such as different DLBCLs can be classified by mass spectrometry (MS)-based proteomics on the basis of PTM-bearing peptides.

The cell surface proteome has been investigated by different approaches. One early method was optimized for the global analysis of both membrane and soluble proteins. It used high pH, which favors the formation of membrane sheets and proteinase K that cleaves the exposed hydrophilic domains of membrane proteins nonspecifically (5). More recent methods targeting the cell surface were based on capturing and covalently labeling glycan moieties on cell surface proteins. Based on such an approach a study on the immune cells using the cell surface capture (CSC) technology which covalently labels extracellular glycan moieties on live cells resulted in the identification of 104 proteins in Jurkat T cells, 96 proteins in an experiment comparing Jurkat T cells and Ramos B cells and 341 proteins in an experiment to detect cell surface changes during differentiation of embryonic stem cells (6). Using the same technology, the combined analysis of 19 B-cell precursor acute lymphoblastic leukemia (BCP-ALL) cases resulted in the identification of 713 cell surface proteins (7).

As glycosylation is increasingly being recognized as one of the key post-translational modifications involved in tumorigenesis with the potential for defining biomarkers, several glycoproteomic studies were performed to study different cancer entities (8). In some of these studies, the primary focus was to specifically capture cell surface and membrane N-glycoproteins based on hydrazide chemistry or lectin affinity approaches. Membrane N-glycoproteins were investigated in colon carcinoma (9), thyroid cancer (10), and breast cancer (11). A more recent study in breast cell lines allowed to distinguish between normal, benign and cancerous ones as well as luminal from basal breast cancer cells based on their glycoprotein profiles (12).

Our laboratory has previously described an extension of the filter-aided sample preparation (FASP) method (13), in which lectins are placed on top of a filter where they selectively retain and enrich glycosylated peptides (14). This approach, termed N-glyco FASP, allows the characterization of thousands of glycosylation sites in complex biological samples such as cell lines, tissues and body fluids in evolutionary diverse species (15). For quantification, this method can also be combined with SILAC (14). In particular, for comparing a large number of unlabeled samples, the super-SILAC approach can be employed (16). It allows precise quantitative comparison of many samples whether cell lines or tissues by spiking in the same SILAC-labeled standard in each of them (16, 17). The standard is generated in such a way that it encompasses as many proteins as possible of the system in question. For that purpose we had previously selected six lymphoma cell lines for a lymphoma super-SILAC mix based on their maximally distinct protein expression profiles (3). Here we decided to take advantage of the depth of the N-glyco FASP method and quantitative accuracy of super-SILAC approach to explore their applicability in the characterization and classification of DLBCL patients. The segregation of these lymphomas based on their quantified glycoproteomes would effectively classify closely related cancer subtypes on the basis of their pattern of post-translational modifications (PTM), a long standing aim of clinical proteomics. Furthermore, differences between DLBCL subtypes in glycosylation patterns or the expression levels of cell surface glycoproteins may reflect tumor associated hallmarks or characteristics of the stage of B-cell development from which these cells are derived. Therefore, characterizing tumors at the protein and PTM level has the potential to increase our understanding of tumor biology. In particular, the segregating signatures of closely related tumor subtypes could shed light on the corresponding biology related to the developmental stage from which the tumors are derived.

EXPERIMENTAL PROCEDURES

Cell Culture—DLBCL cell lines (HBL1, OciLy3, RIVA, TMD8, U2932, BJAB, DB, HT, SUDHL-4, SUDHL-6) were grown in RPMI 1640 medium (Invitrogen, Carlsbad, CA) supplemented with 20% bovine serum and Penicillin/Streptomycin (1:1000). Four biological replicates for each cell line were prepared. Cells were lysed in 4% SDS, 0.1 M dithiotreitol, and 0.1 M Tris-HCL followed by incubation at 95 °C for 5 min. Lysates were sonicated using a Branson type sonicator and then clarified by centrifugation at 16,100 × *g* for 10 min.

Cell lines from which we generated the super-SILAC mix were labeled with heavy amino acids by growing them in RPMI medium containing $^{13}C_6{}^{15}N_2$ - Lysine (Lys8) and $^{13}C_6{}^{15}N_4$ - Arginine (Arg10) (Cambridge Isotope Laboratories, Andover, MA) instead of the natural amino acids and supplemented with 20% dialyzed fetal bovine serum. We used quantitative mass spectrometry to assess the level of incorporation of the heavy amino acids after at least six passages. Almost complete incorporation was achieved in the six cell lines from which we generated the super-SILAC mix (Ramos, Mutu, BL-41, U2932, L428, DB), as less than 1% of tryptic peptides contained unlabeled arginine or lysine and less than 0.3% of identified peptides showed evidence of Arg to Pro conversion (3). The super-SILAC mix was generated by mixing equal amounts of the heavy lysates from the six cell lines.

Protein Digestion and N-glyco Peptide Enrichment—Equal amounts of the super-SILAC mix and the unlabeled cells (300 μg) were mixed on a 30 KDa filter (Millipore, Billerica, MA) and further processed by the filter-aided sample preparation (FASP) method (13). Briefly, the SDS-containing lysis buffer was replaced with a urea

buffer and this was followed by alkylation with iodoacetamide. The samples were then digested overnight by trypsin at 37 °C in 50 mM ammonium bicarbonate followed by elution with water (2×).

For N-glycosylation enrichment, tryptic peptides were transferred to a new filtration unit. They were mixed with a mixture of lectins (ConA, WGA, and RCA lectins) and incubated for 60 mins. Concanavalin A (Con A) binds to mannose; wheat germ agglutinin (WGA) binds to sialic acid as well as N-acetylglucosamine; agglutinin RCA120 binds to galactose modified at the 3–0 position as well as terminal galactose. On washing the samples, with a buffer composed of 20 mM Tris/HCl pH 7.6, 1 mM $MnCl_2$, 1 mM $CaCl_2$, 0.5 M NaCl, the unbound peptides were eluted, whereas the captured glycopeptides remained on the filter. The captured peptides were treated with PNGase F in $H_2{}^{18}O$, which leaves a characteristic mass shift on the previously glycosylated site (18). This was followed by elution and measurement of the deglycosylated peptides (14).

LC-MS/MS Analysis—Deglycosylated peptides were separated by a nanoflow HPLC (Proxeon Biosystems, now Thermo Fisher Scientific) coupled on-line to an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher Scientific) with a nanoelectrospray ion source (Proxeon Biosystems). Peptides were loaded with a flow rate of 500 nl/min on a C_{18}-reversed phase column (20 cm long, 75 μm inner diameter). The column was packed in-house with ReproSil-Pur C18-AQ 1.8 μm resin (Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) in buffer A (0.5% acetic acid). Peptides were eluted with a linear gradient of 8–30% buffer B (80% acetonitrile and 0.5% acetic acid) at a flow rate of 200 nl/min over 145 min. This was followed by 20 min from 30 to 60% buffer B. After each gradient, the column was washed, reaching 90% buffer B followed by re-equilibration with buffer A. Data was acquired using a data-dependent "top 10" method, dynamically choosing the 10 most abundant precursor ions from the survey scan (mass range 300–1800 Th) in order to isolate them in the LTQ and fragment them by higher energy collisional dissociation (HCD) (19). Full scan MS spectra were acquired at a resolution of 30,000 at *m/z* 400 with a target value of 1,000,000 ions. The ten most intense ions were sequentially isolated and accumulated to a target value of 40,000 with a maximum injection time of 150 ms. The lower threshold for targeting a precursor ion in the MS scans was 5000 counts. Fragmentation spectra were acquired in the Orbitrap analyzer with a resolution of 7500 at *m/z* 400.

Data Analysis—MaxQuant software (version 1.2.6.20) was used to analyze mass spectrometric raw data. We searched the MS/MS spectra against the Uniprot database (81,213 entries, release 2012_07) by the Andromeda search engine incorporated in the MaxQuant framework (20, 21). Cysteine carbamidomethylation was set as a fixed modification and N-terminal acetylation, methionine oxidation and deamidation in $H_2{}^{18}O$ were set as variable modifications. A false discovery rate (FDR) of 0.01 was required for proteins and peptides. Enzyme specificity was set to trypsin allowing N-terminal cleavage to proline. A minimum of seven amino acids per identified peptide were required and two miscleavages were allowed. The initial allowed mass deviation of the precursor ion was up to 6 ppm and for the fragment masses it was up to 20 ppm. Mass accuracy of the precursor ions was improved by time-dependent recalibration algorithms of MaxQuant. The "match between runs" option was enabled to match identifications across different replicates. Quantification of SILAC pairs was performed by MaxQuant with standard settings with a minimum ratio count of two. We analyzed the MaxQuant output data with the Perseus tools, which are also available in the MaxQuant environment.

RESULTS AND DISCUSSION

Defining the Quantitative N-glycoproteome of Diffuse Large B-cell Lymphoma Cell Lines—We selected five ABC-DLBCL (HBL1, OciLy3, RIVA, TMD8, U2932) and five GCB-DLBCL (BJAB, DB, HT, SUDHL-4, SUDHL-6) cell lines derived from lymphoma patients. Our previous study had shown that these five ABC and five GCB cell lines can be segregated very clearly by principal component analysis based on their global protein expression profiles (3). We reached a depth of 7756 identified proteins, which allowed the extraction of a signature of 55 proteins that strongly distinguishes between these cancer subtypes. This finding confirms that these cell lines are good representatives of ABC and GCB lymphomas and therefore attractive models to investigate if closely related tumor subtypes can be characterized by a quantitative PTM-based approach.

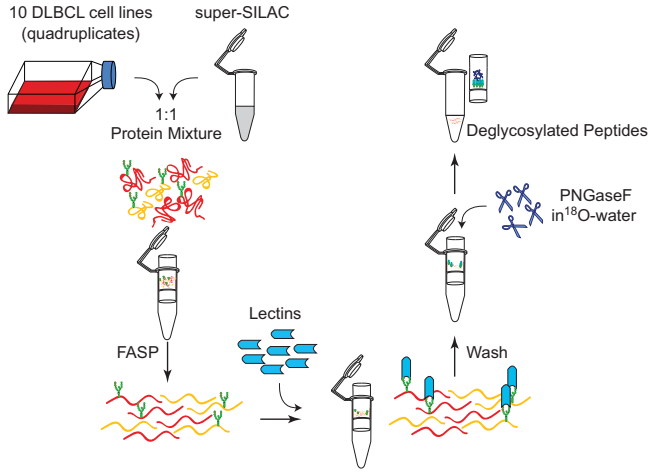


FIG. 1. **Workflow for lymphoma segregation.** A lymphoma super-SILAC mix was spiked in five ABC-DLBCL and five GCB-DLBCL cell lines in quadruplicate. After the samples were processed according to FASP-SAX protocol the eluted peptides were incubated with a mixture of lectins (ConA, WGA, and RCA lectins) and incubated for 60 min. The samples were then washed to get rid of the unbound peptides. The captured glycopeptides remain on the filter were they were treated with PNGase F in $H_2{}^{18}O$. The deglycosylated peptides are finally eluted and measured.

Enriching for glycoproteins would provide a handle for cell surface proteins, which may be especially informative for classification and discovery of biomarkers. For the purpose of an unbiased large-scale enrichment, we used the FASP-based *N*-linked glycopeptide capture method (N-Glyco-FASP) (Fig. 1 and "Experimental Procedures"). Briefly, the FASP-eluted glycopeptides were retained on a 30 kDa filter after mixing all peptides with a mixture of lectins, which aims at capturing the three N-glycan classes (high mannose, complex and hybrid). The large sizes of the glycopeptide-lectin complexes ensure their retention after washing away the non-glycosylated peptides. This method was shown to be efficient and unbiased in mapping the N-glycoproteome of mouse tissues and blood plasma (14) as well as in nonmammalian systems (15). Consequently, deglycosylation of captured N-glycopeptides was performed in ^{18}O-water using PNGase F. Deglycosylation in heavy water results in a mass shift of

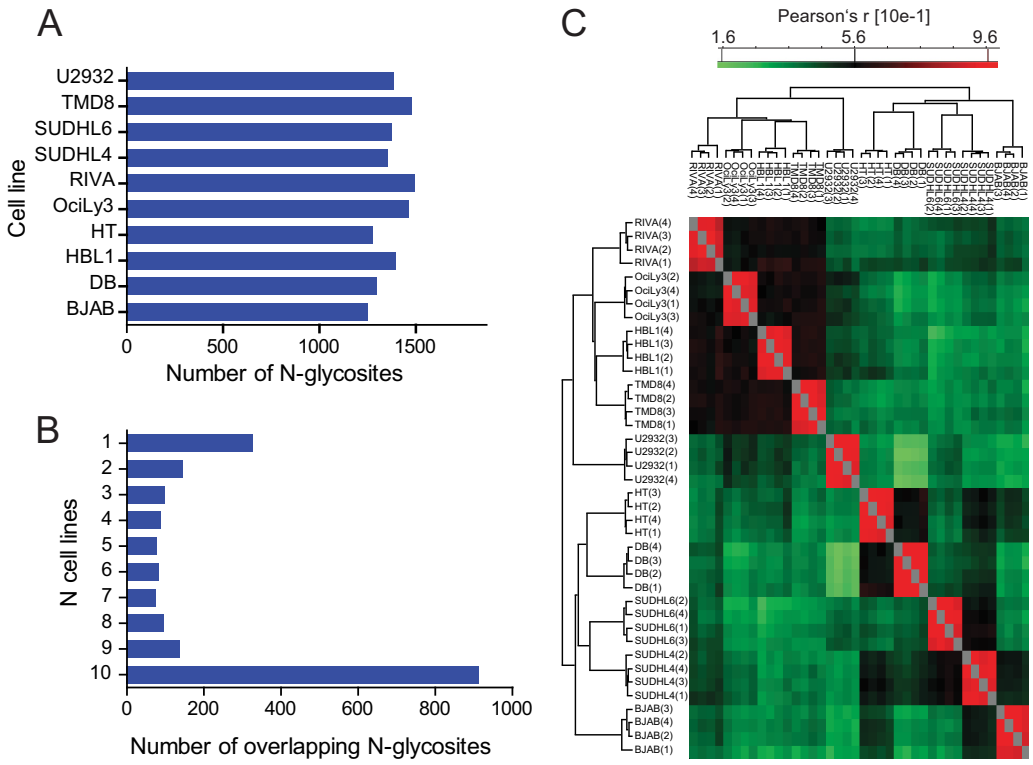


FIG. 2. **B-cell lymphoma N-glycosites quantified.** *A*, Number of N-glycosites quantified in each of the 10 cell lines. *B*, Overlap of N-glycosites quantified across the different cell lines. *C*, Heat map of Pearson's correlation (r) across all measurements.

2.9890 Da and makes it easily distinguished from spontaneous deamidation which results in a mass shift of 0.9858 Da. The resulting deglycosylated precursor peptides as well as MS/MS fragments were analyzed with high resolution, high mass-accuracy measurements on a linear ion trap Orbitrap mass spectrometer.

For an accurate quantitative comparison of the expression profiles of glycosites between ABC-DLBCL and GCB-DLBCL subtypes, we used a heavy-labeled super-SILAC mix of six lymphoma cell lines (3). The pooled lysates constituting the super-SILAC mix were spiked into each of the samples (five ABC-DLBCL and five GCB-DLBCL cell lines) in a 1:1 ratio before the first step of the glyco-enrichment experiment (Fig. 1). The resulting 10 samples were measured in quadruplicates with 165 min gradients. The total measuring time was less than 6 days.

Analysis using stringent filtering criteria in the MaxQuant software environment (20) resulted in the identification of 2383 glycosites, which mapped to 1321 protein groups (supplemental Tables S1 and S2). The median Andromeda identification score of deglycosylated peptides was 127 and the average localization probability of the glycosylation site to a single amino acid was 93%. Next, we filtered for sites with a localization probability greater than 0.75 (class I sites) and a score difference greater than 5 to the next best matching peptide in Andromeda. (Omitting the second filtering step would result in only two additional glycosites, namely beta-1,4-galactosyltransferase—score difference 4.7 and hypoxia up-regulated-protein 1—score difference 4.9.) Our analysis resulted in 2064 very high confidence sites mapped with single amino acid resolution to 1304 protein groups with average localization probability of 99.4% and only these were used in further analysis. Almost all of the high confidence sites were also quantified with at least two valid ratios (1967 of 2064).

In each of the cell lines we identified and quantified 1374 sites on average (Fig. 2*A*). There was excellent overlap between the cell lines as 913 sites were quantified across all 10 cell lines (Fig. 2*B*). From the ratios of the individual samples to the super-SILAC mix we calculated the Pearson correlation coefficients between the measurements. Without exception, quadruplicates co-clustered in a very tight manner (see color code of correlation coefficients in Fig. 2*C*), demonstrating reproducibility and precision of our quantitative strategy.

General Characteristics of the DLBCL-cell N-glycoproteome—Almost 96% (1973 sites) of the glycosites we identified match the canonical N-!P-[S/T] motif. The 91 sites that did not match show enrichment for cysteine ($p = 2.3E-11$) in the position of S/T, which confirms our previous observations (14) (Fig. 3*A*). Matching our data set to the Uniprot database shows that almost 82% (1687 sites) of the glycosites that we identified are annotated to be glycosylated (release 2012_07). However, for 1038 of these sites the annotation is based on prediction or similarity and therefore our results validate these

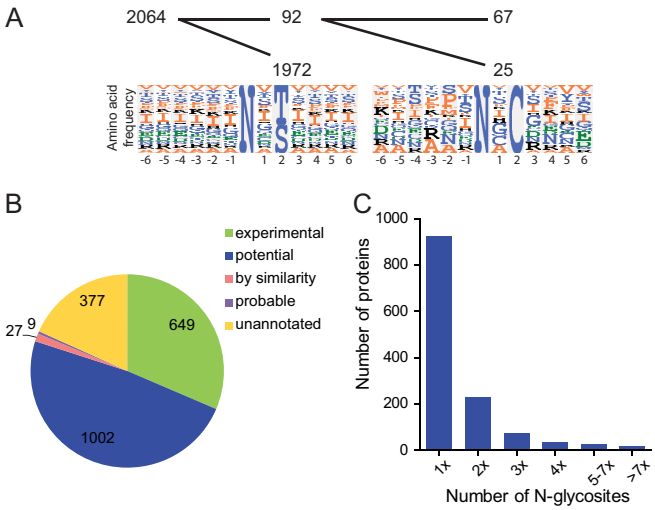


FIG. 3. **B-cell lymphoma N-glycosites identified.** *A*, Motif analysis of 2064 class I glycosites identified. *B*, Match to Uniprot database annotations. *C*, Number of N-glycosylation sites identified per protein.

sites as newly experimentally confirmed N-glycosylation sites (Fig. 3*B*). To our knowledge this dataset constitutes by far the largest human B-cell lymphoma N-glycoproteome reported to date and adds substantially to the human database of N-glycosylation sites. Among the 1304 protein groups to which the glycosites are mapped, 923 protein groups were identified with a single glycosylation site. Very few protein groups (24 protein groups) were identified with more than seven sites (Fig. 3*C*), and the maximum number of N-glycosylation sites was measured on alpha-2-macroglobulin receptor with 19 sites. Insulin-like growth factor 2 receptor was identified with 13. Other proteins of regulatory interest turned out to be heavily glycosylated as well; for instance lymphocyte antigen 75 (DEC205, CD205), receptor-type tyrosine-protein phosphatase eta (PTPRJ) and lysosome-associated membrane glycoprotein 1 (CD107a, LAMP-1) each has more than seven sites. The latter protein also highlights that in addition to the cell surface proteome, our approach enriches intracellular N-glycoproteins. For 1082 sites (52%) the annotated topological domain was extracellular and for 225 sites (11%) it was lumenal. Glycosylation on lumenal domains occurs on lysosomal or ER proteins, for instance (supplemental Fig. S1).

The Proteome Versus the N-glycoproteome—We next compared our data set of N-glycosylated peptides and proteins to our previously measured in-depth proteome of the same cell lines (3). That proteome contained 7756 protein groups and 517 of these also occur in the 1304 protein groups in the N-glycoproteome (matching the N-glcyoproteome to the proteome) (Fig. 4*A*). Strikingly, 787 proteins were exclusively identified by their N-glycosylated peptides, attesting to the enrichment capacity of the workflow. In eukaryotes, *N*-linked glycosylation occurs on secreted or membrane bound proteins, which are often of low abundance, making them more difficult to detect in highly complex samples such as total cellular lysates. This is supported by the fact that the intensities of de-glycosylated peptides of proteins only identified in the N-glyco experiment are shifted to low intensity values compared with those where the corresponding protein was identified in the in-depth proteome experiment (blue *versus* red bars in Fig. 4*B*).

Having extracted a large set of glycoproteins at high sensitivity, we explored which subsets of proteins are enriched in the N-glycoproteome. To obtain a general overview of the cellular localizations and molecular functions of the identified glycoproteins, we analyzed the 1304 proteins groups using Uniprot keywords. The keywords with the highest coverage were "glycoprotein" (89.1%), "membrane" (75.8%), "polymorphism" (71.4%), and "trans membrane" (70.4%) (supplemental Table S3). Compared with the proteome the two keywords with highest enrichment are glycoprotein and signal. Cell membrane proteins and proteins associated with the lysosome, Golgi apparatus and endoplasmic reticulum (ER) were also highly enriched in the glycoproteome compared with the proteome ($p < 1.5E-08$). Interestingly, the extracellular matrix (ECM), a category that was difficult to capture without N-glyco-enrichment, is well represented in the N-glycoproteome ($p < 3.5E-07$) (Fig. 4*C*). As these are suspension cells, the ability to capture this set of proteins via the N-glyco-FASP method comes from the fact that proteins destined for secretion are glycosylated via the classical secretion pathways after passing through the ER and Golgi system. ECM proteins are highly enriched in proteins involved in cancer pathways (FDR 3.5E-05, $p < 2E-07$) including pathways such as Wnt signaling (FDR 0.00085, $p < 1.5E-05$) as well as Hedgehog signaling (FDR 0.0022, $p < 6.5E-05$).

Extracellular matrix proteins are intensely studied in cancer progression because of their role in cellular functions such as adhesion, cell shape, migration, proliferation, polarity, differentiation and apoptosis. The enrichment of this class of proteins via the N-glycoproteome allows comparative analysis of different modes by which cancer cells can manipulate their environment. Molecular function categories which are overrepresented in the glycoproteome include receptors, secreted proteins, cell adhesion proteins, glycosyl transferases and metalloproteases (Fig. 4*C*). These functions are characteristic for glycoproteins and correlate with their extracellular or lumenal location (supplemental Fig. S1).

To investigate how the abundance of glycosites compares to that of proteins, we used the proteome dataset as a reference where no enrichment was performed and quantification is based solely on unmodified peptides. The use of a common super-SILAC mix in both proteome and glycoproteome measurements allows for normalization of technical variance within each experiment and for comparative analysis of proteome *versus* N-glycoproteome measurements. We analyzed the 1203 glycosites belonging to the proteins that matched between both experiments after filtering for at least two valid values in each set of proteome or glycoproteome measure-

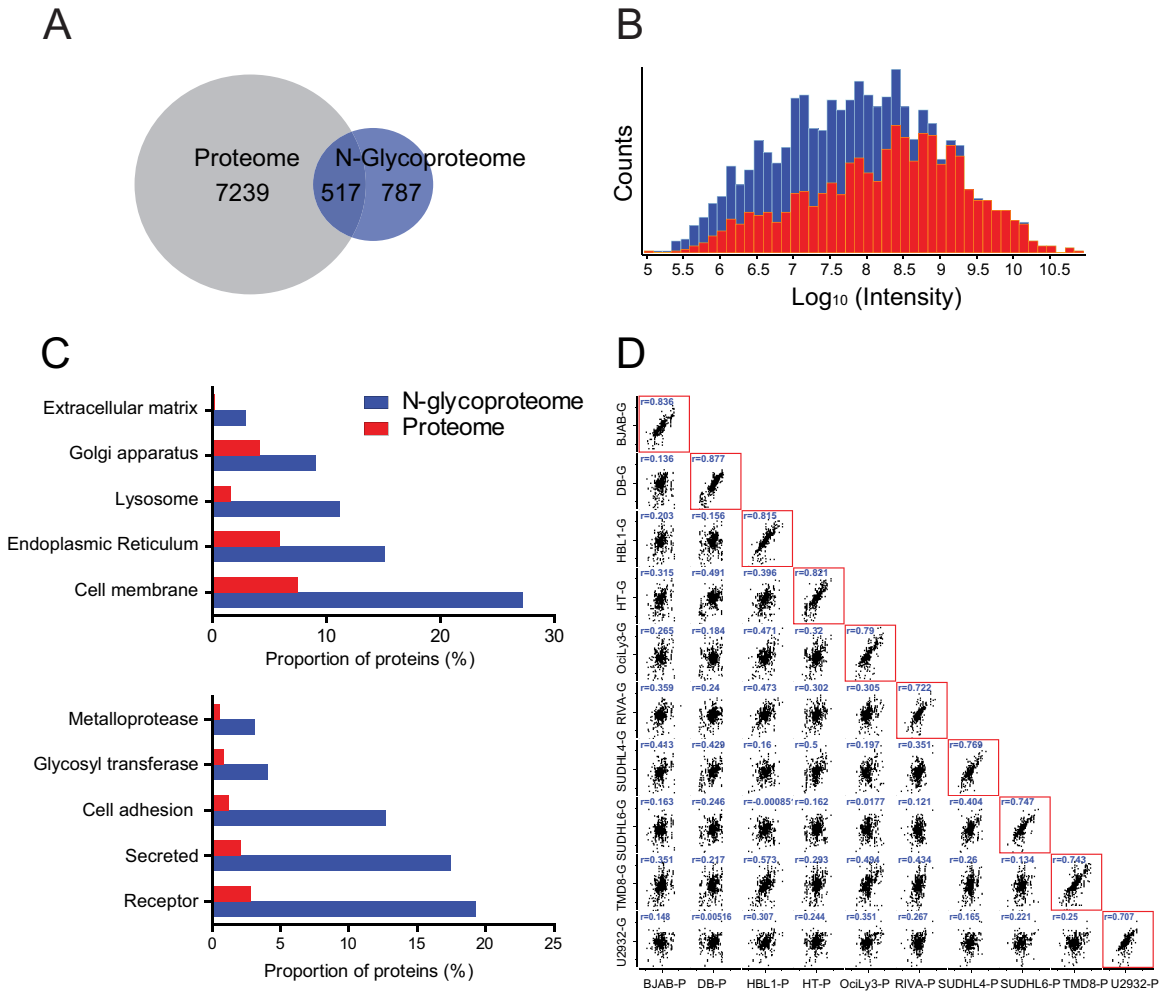


FIG. 4. **The proteome *versus* the N-glycoproteome.** *A*, The overlap in the protein groups identified in both proteome and N-glycoproteome. *B*, Intensity distribution of all glycosites quantified (blue bars) overlaid with intensity distribution of N-glycosites whose corresponding proteins groups were also identified in the proteome data set (red bars). *C*, Cellular components (upper panel) and molecular function (lower panel) categories that are overrepresented in the N-glycoproteome compared with the proteome based on keyword annotations in Uniprot database. *D*, Pearson correlations of glycosites ratios (G) against protein ratios (P) for each cell line *versus* the others.

ments. The ratios of the glycosites against protein ratios correlate well (Pearson $r = 0.78$ on average) (Fig. 4*D*). This is an indication that in the DLBCL cell lines changes in N-glycosylation of a protein are usually a reflection of the change of protein abundance, as expected of a largely cotranslational and stable modification such as N-glycosylation. This is not necessarily true for all sites and indeed highlights the cases in which glycosylation levels and protein expression levels are differentially regulated. With the aim of revealing additional biological differences between the subtypes and to simplify the analysis, we considered the GCB cell lines and ABC cell lines as one entity each and used the median expression level of glycosites and proteins across them. We calculated the log of glycosite to protein ratios and found that there were only few outliers (supplemental Fig. S2). The two strongest ones that are hyperglycosylated in the GCB subtype are glycosites on HLA proteins (HLA-A and HLA-E). Glycosylation on MHC class Ia, for instance, is required for recognition by allogeneic cytotoxic T lymphocytes and to mediate cytolysis (22). Additional interesting hits that are hyperglycosylated in the ABC subtype are three glycosites on ENTPD1 (CD39). CD39 was first described as a B-lymphocyte activation marker (23). It is a prototypic member of the ecto-nucleoside triphosphate diphosphohydrolase (E-NTPDase) family that hydrolyzes extracellular nucleoside diphosphates and triphosphates. Biological actions of CD39 are a consequence of this activity on extracellular nucleotides (24). It has been shown that *N*-linked oligosaccharides affect the enzymatic activity of CD39 (25) whose role in B lymphocytes is not yet clear but may contribute to the affinity maturation of antibody responses and to facilitate post-germinal center terminal B cell differentiation (24).

Segregation of DLBCL Subtypes Based on Glycopeptide Signatures—Principal component analysis (PCA) converts a

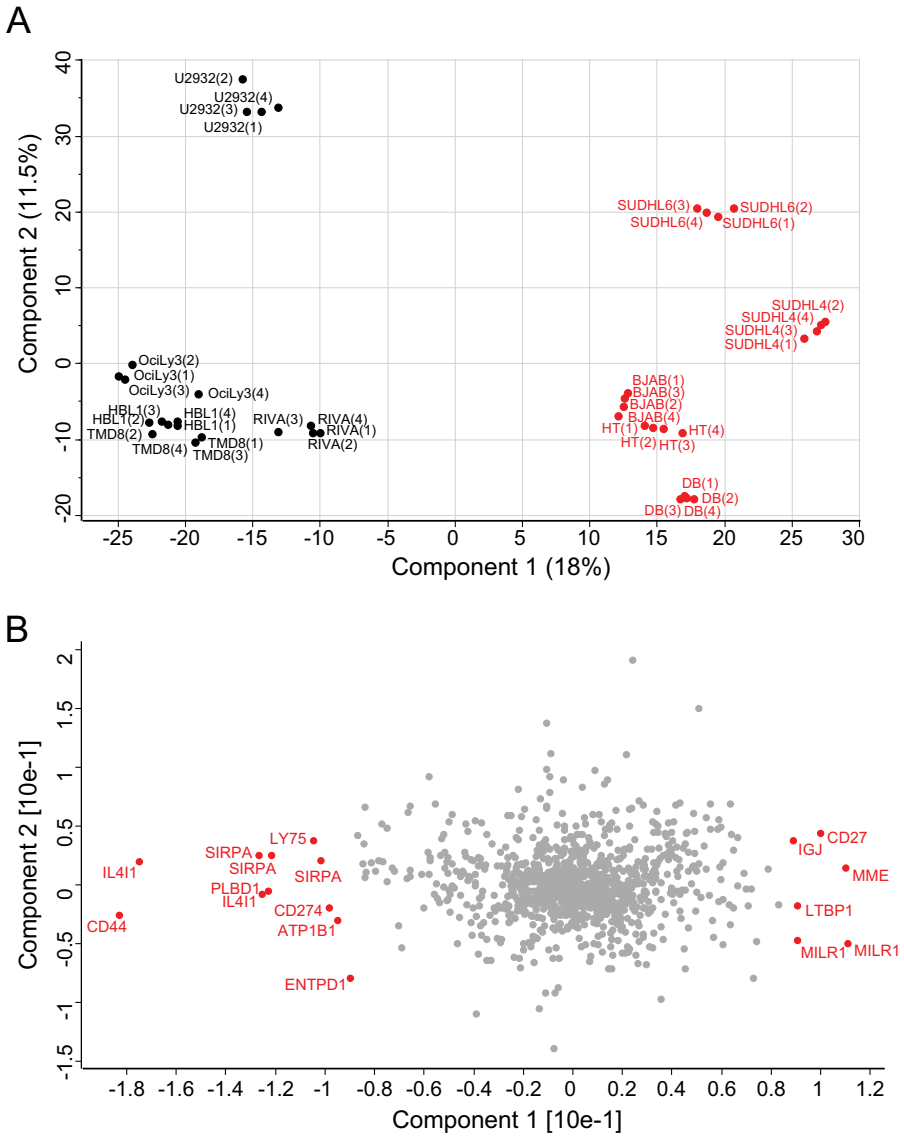


FIG. 5. **Principal component analysis.** *A*, The N-glycoproteome of 10 DLBCL cell lines measured in quadruplicates segregated into ABC-DLBCL (black) and GCB-DLBCL (red) subtypes based on component 1 which accounts for 18% of variability. *B*, Loadings of *A* reveal the main glycosites responsible for driving the segregation in component 1.

large number of data points, which in our case are the SILAC ratios of the glycosites, to a small set of uncorrelated variables—the principal components. Applying this classical statistical test to the glycoproteome resulted in the correct segregation of the cell lines into their corresponding subtypes already based on component 1 (Fig. 5*A*), the one which accounts for the largest variability in the system (18% in this case). The deglycosylated peptides most strongly driving the segregation (the "loadings") belonged to CD44, IL4I1, CD205, PLBD1, SIRPA, LTBP1, MILR1, MME, and CD27 (Fig. 5*B*). Reassuringly, several of these proteins were among the strongest drivers of segregation between ABC-DLBCL and GCB-DLBCL in our previous global proteome. Such proteins included CD27, which is more abundant in GCB-DLBCL and CD44, CD205 and IL4I1, which are more abundant in ABC-DLBCL. The fact that both our global proteome and glycoproteome studies resulted in these candidates increases their likelihood to be true markers of segregation.

One of the most differentiating markers is CD44, which is up-regulated in the ABC subtype relative to the GCB subtype. CD44 is increasingly linked to the progression of different cancer subtypes as well as to cancer-initiating cells (CICs) also known as cancer stem cells. In fact, CD44 is the most common marker of CICs (26). In the context of B cells, BCL-6 transcriptionally blocks the expression of a set of genes including CD44 that are induced when B cells are activated (27). The high abundance and important roles of BCL-6 in germinal center B-cells (28) therefore explain the low relative expres-

sion of CD44 in the GCB subtype, which is derived from germinal center B cells.

The second strongest driver of segregation is interleukin 4-induced protein 1 (IL4I1 or FIG1). The two deglycosylated peptides of this protein were both very strong segregators. Apart from providing additional positive control for our workflow, this finding indicates that the protein itself is highly up-regulated in the ABC subtype, which is indeed what we found in our previous proteome study, where IL4I1 was also a strong differentiator between subtypes. IL4I1 is normally activated by the IL4 receptor via STAT6. The expression of IL4I1 is regulated by NF-κB signaling through the activation of B-cells through the CD40 pathway (29). Accordingly, we find the glycosites of CD40 as well as CD80, the two receptors required for T-cell dependent activation, also up-regulated in the ABC subtype.

Consistent with our findings, differential IL-4 induced gene expression and intracellular signaling in the two subtypes have already been reported (30). IL4I1 has an immunomodulatory role as it has been identified as a secreted L-phenylalanine oxidase that is capable of inhibiting T cell proliferation through producing H_2O_2. It mediates an immunosuppressive effect *in vivo* through blocking the $CD8^+$ antitumor T-cell response (31). Expression of IL4I1 has also been reported to be a characteristic of primary mediastinal lymphoma, the third subtype of diffuse large B-cell lymphoma (32).

On the opposite side (higher expression in GCB), two of the strongest drivers are MME (CD10) and CD27, which is also consistent with the results of our proteome study. High MME expression is prognostic for GCB (2) (33). Down-regulation of MME is mediated through an NF-κB dependent mechanism, which explains the relatively lower level of expression of this protein in the ABC subtype which is characterized by activation of this pathway (34). CD27 is likewise suggested to be a marker with powerful prognostic value for DLBCL and has been included in several prediction algorithms. The serum level of CD27 is reported to be correlated with outcome of patients subjected to standard B-cell lymphoma (R-CHOP) treatment (35).

In contrast to the above mentioned drivers, allergin-1 (MILR1) and LTBP1 have not yet been associated with lymphoma classification. Allergin-1 is studied in the context of allergic responses where it has been shown to suppress IgE-mediated, mast cell-dependent anaphylaxis in mice. In this same study, it has been shown that allergin-1 was expressed on macrophages, neutrophils and dendritic cells as well as mast cells and/or basophils in both humans and mice (36). Interestingly, allergin-1 was also found to be expressed on human B cells (36). This broad expression pattern corresponds to the expression pattern of other immunoglobulin-like inhibitory receptors such as FcγRIIB, PIR-B, gp49B1, MAIR-I and SIRP-α. SIRP-α, which we also classified as a strong segregator of the two subtypes, is expressed on macrophages and dendritic cells and plays an important role in blocking phagocytosis through its interaction with CD47 (37), although its role on B cells is unknown. The activation of specific epitopes on the variable domain of CD47 resulted in a rapid induction of apoptosis in T cells (38). Thus, our data indicate that allergin-1 and SIRP-α might have important roles in nonallergic immune responses, possibly with relevance to the biology of lymphomas. LTBP1 belongs to the family of latent transforming growth factor beta (TGF-β) binding proteins, which are master regulators of TGF-β bioavailability. In addition, LTBPs are integral components of the fibronectin and microfibrillar extracellular matrix (ECM). In the context of breast cancer, elevated LTBP1 levels appear in two gene signatures predictive of enhanced metastatic behavior. The role of LTBP1 in metastasis is unclear but it has been suggested that LTBPs may provide a bridge between structural and signaling components of the epithelial to mesenchymal transition (EMT) (39).

Next, we wished to perform a global analysis of these proteins to discover pathways or protein classes that have a major contribution to the segregation. We annotated the proteins based on the gene set enrichment analysis (GSEA) database (40, 41), which consists *a priori* defined gene sets curated from publications or derived computationally as well as their promoter motifs. The two sets of genes up-regulated in the ABC subtype with highest enrichment were V$NFKB_Q6_01 and V$NFKAPPAB_01. The first gene set corresponds to genes with promoter regions [-2kb, 2kb] around transcription start site containing the computationally derived motif NNNNKGGRAANTCCCN, which does not match any known transcription factor. However, the second motif—also highly enriched—is GGGAMTTYCC, which matches NF-κB RELA. The proteins responsible for this enrichment included known NF-κB regulated proteins such as ATP1B, CPD, ICAM1, PFN1, CD83, LTB, IL4I1, WNT10A, and SLC12A2. In fact, the ABC subtype is characterized by constitutive activity of the NF-κB pathway. More specifically, it has been shown that NF-κB signaling in ABC leads to nuclear translocation of p50/RELA heterodimers and to a lesser extent p50/c-REL heterodimers (42). Hence, despite the relatively small size of the N-glycoproteome it can reveal biologically relevant differences between the subtypes.

Unsupervised Hierarchical Clustering and t Test Signature—When performing unsupervised hierarchical clustering of the deglycosylated peptides, we again obtained perfect segregation of ABC and GCB into the two major branches in the dendrogram (Fig. 6*A*). To extract a glycopeptides signature that significantly segregates the two subtypes we performed a *t* test with a false discovery rate of 0.05 and S0 of 0.1, which resulted in a signature of 38 glycosites (Fig. 6*B*). Many of these glycosites occurred in proteins that were members of the proteomic signature previously found to segregate the subtypes (3), reflecting the high correlation in the level of expression between the glycosites and the corresponding proteins noted above. Specifically, the previous proteomic

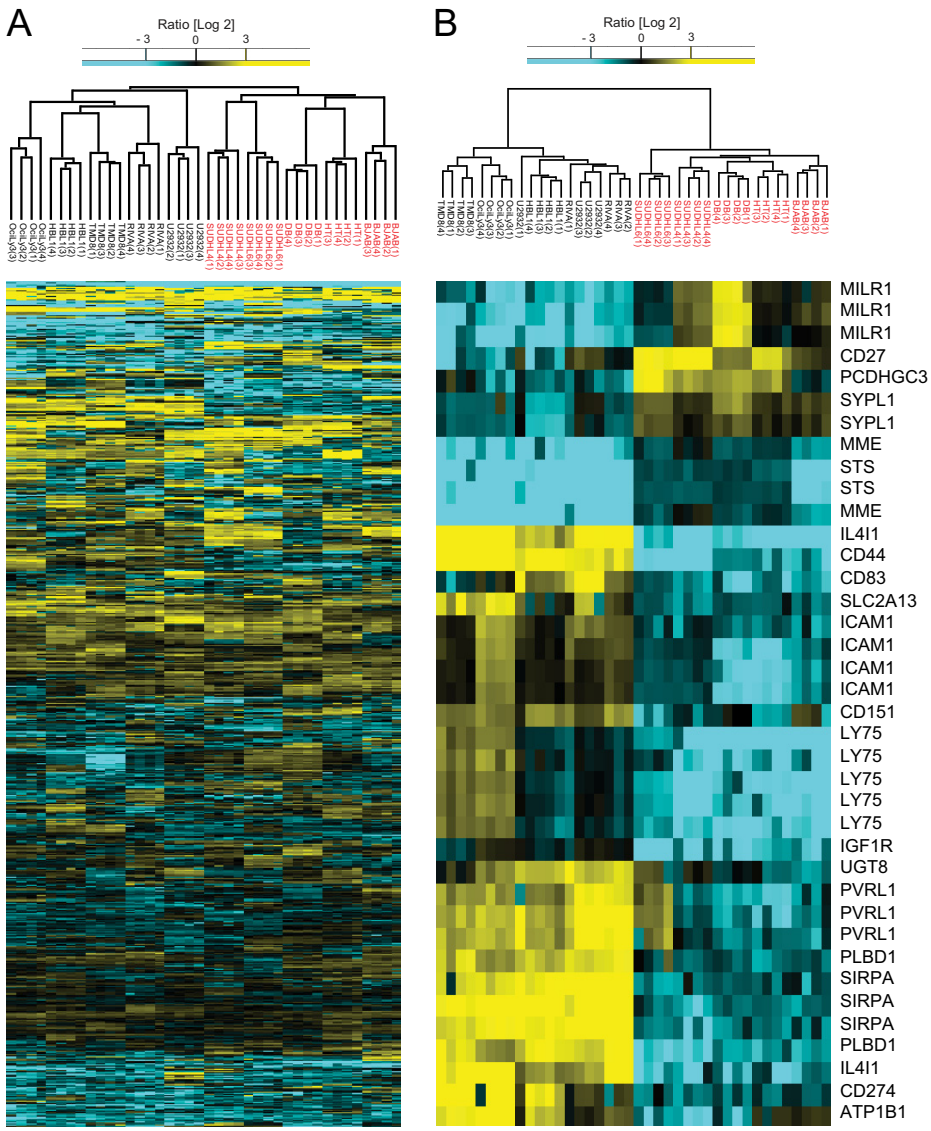


FIG. 6. **Unsupervised Hierarchical clustering and *t* test signature.** *A*, Unsupervised clustering (Euclidian distance) of 10 DLBCL cell lines based on their corresponding N-glycosites expression profile. *B*, Heat map of the 53 glycosites that were picked as most significantly different after performing *t* test analysis for the two subtypes.

signature contained 10 glycoproteins (Uniprot keyword annotation), which are CD27, ICAM1, RCN, CD205 (LY75), IL4I1 (FIG1), CD44, SYPL1, HLA-C, ATPIB, and EVDB. With the exception of EVDB, all of these were identified in the glycoproteome study. In our glycosites signature, glycosites belonging to seven of these nine glycoproteins were shown to be significantly different between the two subtypes. This is quite remarkable and reassuring especially taking into consideration the small size of the glycosites signature, which is composed of 20 glycoproteins. This large overlap, prompted us to evaluate how much information the N-glycoproteome alone would add to the characterization of the system. To this end we subtracted the glycosites on markers already identified in our global proteome study or in mRNA profiling studies. This left 20 glycosites on 12 proteins exclusive to the glycosignature. Remarkably, the PCA analysis segregated the two subtypes solely based on these exclusive glycosites (supplemental Figs. S3*A* and S3*B*).

Often in the glycosites signature, several deglycosylated peptides which belong to the same protein are significantly differentially expressed between the subtypes. This was most prominent in the case of CD205 (LY75) (5 peptides) and ICAM1 (4 peptides). CD205 belongs to the family of C-type lectin receptors (CLRs) which function as pattern recognition receptors recognizing carbohydrate ligands from infected microorganisms (43). CD205 was mainly studied in the context of dendritic cells where it is used as a docking site to deliver specific antigens (44). In the context of B cells it has been shown that CD205 modulates their phenotype causing up-regulation of co-stimulatory molecules on the cell surface (45).

It has also been shown that CD205 may have a role in promoting cell adhesion where blocking CD205 was suggested as a potential clinical strategy to interfere with early ovarian cancer metastasis (46). ICAM1 is an NF-κB regulated cell-surface receptor from the immunoglobulin superfamily whose serum levels correlate with a higher tumor burden and dissemination in DLBCL (47). ICAM1 has a role in cell adhesion, a costimulatory role to ensure a proper T cell response as well as a role in lymphoid trafficking and extravasation (47).

Performing a Fisher's exact test ($p < 0.01$; enrichment factor > 5) on our signature glycoproteins after adding GSEA annotations reveals an enrichment of interesting gene sets (supplemental Table S4). The one with the highest significance is in fact V$NFKAPPAB_01 and the second set corresponds to genes up-regulated on an inflammatory response in macrophages. With the aim of obtaining a broader view of such enrichments, we performed a less stringent *t* test with an FDR of 0.1 and S0 of 0.01 which resulted in 57 differentially expressed sites. The Fisher's exact test on this signature—requiring the same fivefold enrichment factor as before - added some new and interesting categories (supplemental Table S5). Two gene sets that correspond to IL6 regulated genes from two independent studies were enriched (DASU_IL6_SIGNALING_SCAR_UP and BROCKE_APOPTOSIS_REVERSED_BY_IL6). The role of IL6 in lymphomagenesis is interesting because upon transformation, B-cell lymphomas use IL6 paracrine signaling as a survival signal (48). In ABC-DLBCL in particular, NF-κB signaling was shown to induce the expression of IL6, which leads to activation of STAT3 in an autocrine manner. In fact, combination treatments that block both NF-κB signaling and STAT3 signaling are especially toxic to ABC-DLBCL as they work synergistically (49).

The evident biological relevance of the *t* test signature as well as the enrichment analysis highlight the potential of a PTM-based approach where in this case probing for membrane proteins revealed differential intracellular signaling.

CONCLUSION AND OUTLOOK

We have shown that the protein expression patterns of cell lines derived from ABC-DLBCL and GCB-DLBCL subtypes can unambiguously differentiate them (3). Taking proteomic approaches one step further, we wanted to investigate whether a specific set of functionally relevant proteins can also address this question. We focused this study on membrane proteins, which are key players in cancer cell biology and are located at the interface between a cancer cell and its environment. Taking into consideration the redundancy in the activated downstream signaling pathways, studying membrane proteins can be a more specific way of characterizing cancer cells which can help in classifying them and developing targeted therapies. We used the N-glyco-FASP protocol (14) as a tool to efficiently enrich for this class of proteins. The N-glyco-enrichment protocol does not use any chemical derivatization steps but does involve the extra steps of lectin enrichments and deglycosylation with PNGase F (14). As this could introduce additional variability, we performed this study in quadruplicates and used a lymphoma super-SILAC mix as an internal standard, which successfully minimized the effects of technical variations. High quantitative precision is necessary to differentiate the two subtypes especially in this case where quantification of N-glycosylation sites usually involves a single peptide (17).

Applying the N-glyco-FASP method on 10 lymphoma patient-derived cell lines resulted in a subset of the proteome highly enriched for membrane and secreted proteins. To our knowledge this is the largest membrane B-cell lymphoma proteome. This then enabled us to segregate the two closely related subtypes of DLBCL based on their N-glycoproteome expression profiles. Importantly, the loadings of component 1 which segregates the two subtypes in the principal component analysis include glycosites on proteins which we had suggested to be markers in our previous proteome study. This overlap further validates these easily accessible cell surface proteins as clinically interesting candidates. By implication, our novel candidates such as allergin-1 (MILR1), LTBP1 and SIRP-α should be very interesting targets for investigation as biological drivers of segregation. In addition to investigating these proteins on an individual basis, we tested bioinformatically for enrichments in the loadings of component 1. This revealed that one of the gene sets up-regulated in the ABC subtype corresponds to genes with promoter regions around transcription start site containing the motif for NF-κB RELA. Therefore our unbiased approach links the differences in N-glycoproteomes to differential transcriptional regulation between the subtypes. Differential activity of NF-κB signaling is considered to be one of the major pathways accounting for molecular differences between ABC and GCB subtypes of DLBCL. Hence, our approach can link differences in the glycoproteome to intrinsic biological differences between the subtypes using this small subset of proteins that was obtained in a straightforward and rapid manner.

Our study demonstrates that the enrichment of a single PTM can be used to differentiate between closely related tumor subtypes. This highlights the potential of targeting a particular set of proteins—in this case membrane proteins—that could be of very high clinical relevance in cancer classification and provision of targeted therapies.

In conclusion, the continuous development of mass spectrometry-based technologies generates more and more exciting tools to describe the biology of cancer cells and thereby unlock their secrets. This is especially true for post-translational modifications, which cannot be identified by genomics approaches. As a first step in this direction, we have here established that MS-based quantification of enriched glycosylated membrane proteins can distinguish between related lymphoma subtypes and to identify disease

segregating, novel cell surface targets on B-cell lymphoma cells. This approach might provide the basis for the future diagnosis of subtypes of B-cell lymphomas or any closely related tumor subtypes and even of normal cells where an unbiased global screening of the cell surface is required.

Acknowledgments—We thank Maria (Charo) Robles, Marlis Zeiler and Stefka Tyanova for helpful discussions.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the data set identifier PXD000332. Annotated spectra of formerly glycosylated peptides are provided in supplemental Fig. S4.

* This work was funded by the European Commission's 7th Framework Program PROSPECTS (grant agreement HEALTH-F4-2008-201648).

S This article contains supplemental Figs. S1 to S4 and Tables S1 to S5.

¶ To whom correspondence should be addressed: Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, D-82152 Martinsried, Germany. Tel.:+ 49-89-8578–2557; E-mail: mmann@biochem.mpg.de.

REFERENCES

1. Frick, M., Dörken, B., and Lenz, G. (2012) New insights into the biology of molecular subtypes of diffuse large B-cell lymphoma and Burkitt lymphoma. *Best Practice & Res. Clin. Haematol.* **25,** 3–12
2. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–511
3. Deeb, S. J., D'Souza, R., Cox, J., Schmidt-Supprian, M., and Mann, M. (2012) Super-SILAC allows classification of diffuse large B-cell lymphoma subtypes by their protein expression profiles. *Mol. Cell. Proteomics* **11,** 77–89
4. Bendall, S. C., Nolan, G. P., Roederer, M., and Chattopadhyay, P. K. (2012) A deep profiler's guide to cytometry. *Trends Immunol.* **33,** 323–332
5. Wu, C. C., MacCoss, M. J., Howell, K. E., and Yates, J. R. (2003) A method for the comprehensive proteomic analysis of membrane proteins. *Nat. Biotech.* **21,** 532–538
6. Wollscheid, B., Bausch-Fluck, D., Henderson, C., O'Brien, R., Bibel, M., Schiess, R., Aebersold, R., and Watts, J. D. (2009) Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat. Biotech.* **27,** 378–386
7. Mirkowska, P., Hofmann, A., Sedek, L., Slamova, L., Mejstrikova, E., Szczepanski, T., Schmitz, M., Cario, G., Stanulla, M., Schrappe, M., van der Velden, V. H. J., Bornhauser, B. C., Wollscheid, B., and Bourquin, J.-P. (2013) Leukemia surfaceome analysis reveals new disease-associated features. *Blood* **121,** e149–e159
8. Pan, S., Chen, R., Aebersold, R., and Brentnall, T. A. (2011) Mass Spectrometry Based Glycoproteomics—From a Proteomics Perspective. *Mol. Cell. Proteomics* **10**
9. Vercoutter-Edouart, A.-S., Slomianny, M.-C., Dekeyzer-Beseme, O., Haeuw, J.-F., and Michalski, J.-C. (2008) Glycoproteomics and glycomics investigation of membrane N-glycosylproteins from human colon carcinoma cells. *Proteomics* **8,** 3236–3256
10. Arcinas, A., Yen, T.-Y., Kebebew, E., and Macher, B. A. (2009) Cell Surface and Secreted Protein Profiles of Human Thyroid Cancer Cell Lines Reveal Distinct Glycoprotein Patterns. *J. Proteome Res.* **8,** 3958–3968
11. Whelan, S. A., Lu, M., He, J., Yan, W., Saxton, R. E., Faull, K. F., Whitelegge, J. P., and Chang, H. R. (2009) Mass Spectrometry (LC-MS/MS) Site-Mapping of N-Glycosylated Membrane Proteins for Breast Cancer Biomarkers. *J. Proteome Res.* **8,** 4151–4160
12. Yen, T.-Y., Macher, B. A., McDonald, C. A., Alleyne-Chin, C., and Timpe, L. C. (2011) Glycoprotein Profiles of Human Breast Cells Demonstrate a Clear Clustering of Normal/Benign versus Malignant Cell Lines and Basal versus Luminal Cell Lines. *J. Proteome Res.* **11,** 656–667
13. Wisniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat. Meth.* **6,** 359–362
14. Zielinska, D. F., Gnad, F., Wi, and Mann, M. (2010) Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints. *Cell* **141,** 897–907
15. Zielinska, Dorota F., Gnad, F., Schropp, K., Wi, and Mann, M. (2012) Mapping N-Glycosylation Sites across Seven Evolutionarily Distant Species Reveals a Divergent Substrate Proteome Despite a Common Core Machinery. *Mol. Cell* **46,** 542–548
16. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J. R., and Mann, M. (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Meth.* **7,** 383–385
17. Boersema, P. J., Geiger, T., Wiśniewski, J. R., and Mann, M. (2013) Quantification of the N-glycosylated Secretome by Super-SILAC During Breast Cancer Progression and in Human Blood Samples. *Mol. Cell. Proteomics* **12,** 158–171
18. Küster, B., and Mann, M. (1999) 18O-Labeling of N-Glycosylation Sites To Improve the Identification of Gel-Separated Glycoproteins Using Peptide Mass Mapping and Database Searching. *Anal. Chem.* **71,** 1431–1440
19. Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., and Mann, M. (2007) Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Meth.* **4,** 709–712
20. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372
21. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **10,** 1794–1805
22. Bagriaçik, E. Ü., Kirkpatrick, A., and Miller, K. S. (1996) Glycosylation of native MHC class Ia molecules is required for recognition by allogeneic cytotoxic T lymphocytes. *Glycobiology* **6,** 413–421
23. Maliszewski, C. R., Delespesse, G. J., Schoenborn, M. A., Armitage, R. J., Fanslow, W. C., Nakajima, T., Baker, E., Sutherland, G. R., Poindexter, K., and Birks, C. (1994) The CD39 lymphoid cell activation antigen. Molecular cloning and structural characterization. *J. Immunol.* **153,** 3574–3583
24. Dwyer, K., Deaglio, S., Gao, W., Friedman, D., Strom, T., and Robson, S. (2007) CD39 and control of cellular immune responses. *Purinergic Signal.* **3,** 171–180
25. Wu, J. J., Choi, L. E., and Guidotti, G. (2005) N-linked Oligosaccharides Affect the Enzymatic Activity of CD39: Diverse Interactions between Seven N-linked Glycosylation Sites. *Mol. Biol. Cell* **16,** 1661–1672
26. Zöller, M. (2011) CD44: can a cancer-initiating cell profit from an abundantly expressed molecule? *Nat. Rev. Cancer* **11,** 254–267
27. Shaffer, A. L., Yu, X., He, Y., Boldrick, J., Chan, E. P., and Staudt, L. M. (2000) BCL-6 represses genes that function in lymphocyte differentiation, inflammation, and cell cycle control. *Immunity* **13,** 199–212
28. Basso, K., and Dalla-Favera, R. (2012) Roles of BCL6 in normal and transformed germinal center B cells. *Immunol. Rev.* **247,** 172–183
29. Marquet, J., Lasoudris, F., Cousin, C., Puiffe, M.-L., Martin-Garcia, N., Baud, V., Chereau, F., Farcet, J.-P., Molinier-Frenkel, V., and Castellano, F. (2010) Dichotomy between factors inducing the immunosuppressive enzyme IL-4-induced gene 1 (IL4I1) in B lymphocytes and mononuclear phagocytes. *Eur. J. Immunol.* **40,** 2557–2568
30. Lu, X., Nechushtan, H., Ding, F., Rosado, M. F., Singal, R., Alizadeh, A. A., and Lossos, I. S. (2005) Distinct IL-4-induced gene expression, proliferation, and intracellular signaling in germinal center B-cell-like and activated B-cell-like diffuse large-cell lymphomas. *Blood* **105,** 2924–2932
31. Lasoudris, F., Cousin, C., Prevost-Blondel, A., Martin-Garcia, N., Abd-Alsamad, I., Ortonne, N., Farcet, J.-P., Castellano, F., and Molinier-Frenkel, V. (2011) IL4I1: an inhibitor of the CD8+ antitumor T-cell response in vivo. *Eur. J. Immunol.* **41,** 1629–1638
32. Copie-Bergman, C., Boulland, M.-L., Dehoulle, C., Möller, P., Farcet, J.-P., Dyer, M. J. S., Haioun, C., Roméo, P.-H., Gaulard, P., and Leroy, K. (2003) Interleukin 4–induced gene 1 is activated in primary mediastinal large B-cell lymphoma. *Blood* **101,** 2756–2761

33. Wright, G., Tan, B., Rosenwald, A., Hurt, E. H., Wiestner, A., and Staudt, L. M. (2003) A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc. Natl. Acad. Sci.* **100,** 9991–9996
34. Thompson, R. C., Herscovitch, M., Zhao, I., Ford, T. J., and Gilmore, T. D. (2011) NF-κB Down-regulates Expression of the B-lymphoma Marker CD10 through a miR-155/PU. 1 Pathway. *J. Biol. Chem.* **286,** 1675–1682
35. Goto, N., Tsurumi, H., Takemura, M., Kanemura, N., Kasahara, S., Hara, T., Yasuda, I., Shimizu, M., Yamada, T., Sawada, M., Takahashi, T., Yamada, T., Seishima, M., Moriwaki, H., and Takami, T. (2012) Serum soluble CD27 level is associated with outcome in patients with diffuse large B-cell lymphoma treated with rituximab, cyclophosphamide, doxorubicin, vincristine and prednisolone. *Leukemia Lymphoma* **53,** 1494–1500
36. Hitomi, K., Tahara-Hanaoka, S., Someya, S., Fujiki, A., Tada, H., Sugiyama, T., Shibayama, S., Shibuya, K., and Shibuya, A. (2010) An immunoglobulin-like receptor, Allergin-1, inhibits immunoglobulin E-mediated immediate hypersensitivity reactions. *Nat. Immunol.* **11,** 601–607
37. Brown, E. J., and Frazier, W. A. (2001) Integrin-associated protein (CD47) and its ligands. *Trends Cell Biol.* **11,** 130–135
38. Pettersen, R. D., Hestdal, K., Olafsen, M. K., Lie, S. O., and Lindberg, F. P. (1999) CD47 Signals T Cell Death. *J. Immunol.* **162,** 7031–7040
39. Anupama Chandramouli, J. S., Alicia Pinderhughes and Pamela Cowin (2011) Choreographing Metastasis to the Tune of LTBP. *J. Mammary Gland Biol. Neoplasia* **16,** 67–80
40. Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003) PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34,** 267–273
41. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550
42. Davis, R. E., Brown, K. D., Siebenlist, U., and Staudt, L. M. (2001) Constitutive Nuclear Factor κB Activity Is Required for Survival of Activated B Cell–like Diffuse Large B Cell Lymphoma Cells. *J. Exp. Med.* **194,** 1861–1874
43. Kingeter, L. M., and Lin, X. (2012) C-type lectin receptor-induced NF-[kgr]B activation in innate immune and inflammatory responses. *Cell Mol. Immunol.* **9,** 105–112
44. Jiang, W., Swiggard, W. J., Heufler, C., Peng, M., Mirza, A., Steinman, R. M., and Nussenzweig, M. C. (1995) The receptor DEC-205 expressed by dendritic cells and thymic epithelial cells is involved in antigen processing. *Nature* **375,** 151–155
45. McKay, P. F., Imami, N., Johns, M., Taylor-Fishwick, D. A., Sedibane, L. M., Totty, N. F., Hsuan, J. J., Palmer, D. B., George, A. J. T., Foxwell, B. M. J., and Ritter, M. A. (1998) The gp200-MR6 molecule which is functionally associated with the IL-4 receptor modulates B cell phenotype and is a novel member of the human macrophage mannose receptor family. *European J. Immunol.* **28,** 4071–4083
46. Giridhar, P., Funk, H., Gallo, C., Porollo, A., Mercer, C., Plas, D., and Drew, A. (2011) Interleukin-6 receptor enhances early colonization of the murine omentum by upregulation of a mannose family receptor, LY75, in ovarian tumor cells. *Clin. Exp. Metastasis* **28,** 887–897
47. Terol, M. J., Tormo, M., Martinez-Climent, J. A., Marugan, I., Benet, I., Ferrandez, A., Teruel, A., Ferrer, R., and García-Conde, J. (2003) Soluble intercellular adhesion molecule-1 (s-ICAM-1/s-CD54) in diffuse large B-cell lymphoma: association with clinical characteristics and outcome. *Ann. Oncol.* **14,** 467–474
48. Gilbert, L. A., and Hemann, M. T. (2012) Context-specific roles for paracrine IL-6 in lymphomagenesis. *Genes Development* **26,** 1758–1768
49. Lam, L. T., Wright, G., Davis, R. E., Lenz, G., Farinha, P., Dang, L., Chan, J. W., Rosenwald, A., Gascoyne, R. D., and Staudt, L. M. (2008) Cooperative signaling through the signal transducer and activator of transcription 3 and nuclear factor-κB pathways in subtypes of diffuse large B-cell lymphoma. *Blood* **111,** 3701–3713

2.3 Machine Learning Based Classification of Diffuse Large B-cell Lymphoma Patients by their Protein Expression Profiles

2.3.1 Project aim and summary

Previously, we have robustly segregated ABC-DLBCL and GCB-DLBCL cell lines based on their global protein and PTM-based expression profiles. In this study, we built on the cell line work to investigate the applicability of our high-resolution MS-based platform to subtype DLBCL patients on the basis of their tumor proteome.

Human tumors are usually preserved as formalin-fixed paraffin-embedded (FFPE) material. This makes them challenging to analyze because tissue proteomes are inherently complex and because proteins need to be efficiently extracted from FFPE. To deal with these challenges, we applied state-of-the-art technological advances in sample preparation, measurement and data analysis. The FFPE filter-aided sample preparation (FFPE-FASP) method for protein extraction and digestion in combination with a quadrupole Orbitrap mass spectrometer allowed us to reach an unprecedented depth of 9,000 proteins in DLBCL patient proteomes. We employed the lymphoma super-SILAC mix previously developed as a general spike-in standard. This resulted in accurate quantification that allowed the segregation of the subtypes either by the55 protein signature previously derived from cell lines as well as by the global protein expression profiles of the patient samples. Almost all the drivers of segregation in the global analysis correlated to known biology of the subtypes. To extract a signature of proteins with the highest segregating power, we developed a novel combination of statistical feature selection and machine learning. This analysis resulted in a signature of 20 proteins that perfectly segregated the subtypes providing a solid proof-of-principle for future applications in large patient cohorts.

2.3.2 Contribution

This project, a continuation of the first two projects, was also supervised by Matthias Mann and Marc Schmidt-Supprian. I performed and optimized all sample preparation techniques and MS analysis methods as well as data acquisition and analysis. Stefka Tyanova implemented the support vector machine algorithm. I designed all figures and tables for the publication. The manuscript was written by me with the help of Matthias Mann and Marc Schmidt-Supprian.

2.3.3 Manuscript

The included manuscript is currently in preparation.

Machine Learning Based Classification of Diffuse Large B-cell Lymphoma Patients by their Protein Expression Profiles

Sally J. Deeb[1], Stefka Tyanova[1,3], Marc Schmidt-Supprian[2], Juergen Cox[3] and Matthias Mann[1,*]

[1] Proteomics and Signal Transduction Group, Max-Planck Institute of Biochemistry, D-82152 Martinsried, Germany

[2] Department of Oncology and Hematology, 3te Medizinische Klinik, Technische Universitaet Muenchen, 81675 Munich, Germany

[3] Computational systems Biochemistry, Max-Planck Institute of Biochemistry, D-82152 Martinsried, Germany

[*]To whom correspondence may be addressed: Ph.:+ 49-89-8578-2557; E-mail:

mmann@biochem.mpg.de

Running title: Subtyping of lymphoma patients based on their proteomes

Abbreviations: ABC-DLBCL, activated B-cell–like diffuse large B-cell lymphoma; BCR, B-cell receptor; CLL, chronic lymphocytic leukemia; COO, cell-of-origin; DLBCL, diffuse large B-cell lymphoma; FFPE, formalin-fixed paraffin-embedded; GCB-DLBCL, germinal-center B-cell–like DLBCL; GEP, gene expression profiling; MS, mass spectrometry; PCA, principal component analysis; SILAC, stable isotope labeling with amino acids in cell culture; SVM, support vector machine.

Summary

Characterization of tumors at the molecular level has improved our knowledge of cancer causation and progression. Proteomic analysis of their signaling pathways promises to enhance our understanding of cancer aberrations at the functional level, but this requires accurate and robust tools. Here, we develop a state of the art quantitative mass spectrometric pipeline to characterize formalin-fixed paraffin-embedded (FFPE) tissues of patients with closely related subtypes of diffuse large B-cell lymphoma (DLBCL). We combined a super-SILAC approach with label-free quantification (hybrid LFQ), to address situations where the protein is absent in the super-SILAC standard yet present in the patient samples. Shotgun proteomic analysis on a quadrupole Orbitrap quantified almost 9000 tumor proteins in 20 patients. The quantitative accuracy of our approach allowed the segregation of DLBCL patients according to their cell-of-origin, using both their global protein expression patterns and the 55-protein signature obtained previously from patient-derived cell lines (Deeb *et al.* MCP 2012 PMID 22442255). Expression levels of individual segregation-driving proteins as well as categories such as extracellular matrix proteins behaved consistent with known trends between the subtypes. We employed machine learning (support vector machines) to extract candidate proteins with the highest segregating power. A panel of four proteins (PALD1, MME, TNFAIP8 and TBC1D4) classified the patients with very low error rates. Highly ranked proteins from the support vector analysis revealed differential expression of core signaling molecules between the subtypes, elucidating aspects of their pathobiology.

Clinical differences between human cancer subtypes have long been recognized by oncologists. However, comprehensive analyses of the underlying molecular differences have only become possible with the recent advent of powerful oligonucleotide-based technologies that allow global profiling of individual tumors (1). The potential benefits of improved molecular characterization are enormous (2). In fact, the molecular understanding of tumorigenesis and cancer progression is promising to enable a shift from non-specific cytotoxic drugs to drugs that are much more targeted towards cancer cells. An important step to achieve targeted therapies is to reliably identify the group of patients that are likely to benefit from a specific drug or treatment strategy. This ability to group cancer patients into clinically meaningful subtypes is a challenging task that requires well-designed and robust approaches.

More than a decade ago, gene expression profiling discovered two subtypes of diffuse large B-cell lymphoma (DLBCL), which are morphologically indistinguishable (3). The subtyping was based on gene expression signatures that correspond to stages of B-cell development from which the tumor is derived. The germinal center B-cell-like DLBCL (GCB-DLBCL) transcriptome was dominated by genes characteristic of germinal center B-cells, whereas the transcriptome of activated B-cell-like DLBCL (ABC-DLBCL) more closely resembled activated B-cells *in vitro* (3). Importantly, the discovered subtypes defined prognostic categories (3, 4), opening up the possibility of differential treatment (5). Nonetheless, this cell-of-origin (COO) classification did not fully reflect the differences in overall survival after chemotherapy among patients. Follow-up studies - also using gene expression profiling - showed that a multivariate model constructed from three gene-expression signatures (germinal-center B-cell, stromal-1, and stromal-2) was a

better predictor of survival (6). Stromal-1 reflected extracellular matrix deposition and stromal-2, which had an unfavorable prognosis, reflected tumor blood vessel density.

In addition to DLBCLs, gene expression profiling also successfully sub-classified several other cancer types such as breast cancer (7). However, in colorectal adenocarcinoma there was no correlation between the subtypes derived from GEP and clinical phenotypes like patient survival or response to treatment (8). As RNA is a fragile molecule, one of the challenges of mRNA-based global expression studies is the required quality of the RNA sample (9). The problem is exacerbated when working with formalin-fixed paraffin-embedded (FFPE) tissues, which are frequently the only biopsy material available. The extraction of RNA from FFPE tissues is still a daunting task and snap-frozen tissues are preferred for microarray-based genome-wide GEP (10). For that reason and because proteins are standard marker molecules in pathology, in the last decade many approaches were developed to classify DLBCL patients on the basis of immunohistochemistry (IHC) of FFPE tissues. They attempted to simulate gene expression profiling in predicting the COO of tumors. However, gene expression profiling rather than IHC-based algorithms still best predicted prognosis in DLBCL patients treated with immunochemotherapy (11). Most recently, a targeted RNA (NanoString)–based test of 20 genes accurately assigned COO subtypes to DLBCL patients in FFPE (12) and has now been adopted as a diagnostic tool in a clinical trial to support the development of lenalidomide (Revlimid) as treatment for patients with DLBCL.

Proteins are the molecules that actually carry out biological function in a cell. Thus, proteomics has the potential to directly assess deregulated cellular processes and signaling pathways. In

the last decade, MS-based proteomics has developed tremendously in terms of sample preparation techniques, mass spectrometric instrumentation and data analysis. Enhanced sensitivity, accuracy and peptide sequencing speed of contemporary mass spectrometers allow the identification of thousands of proteins in a single experiment. This has already resulted in almost the complete coverage of complex biological samples such as human cancer cells (13, 14). We have shown that very large depth of complex proteomes can even be attained without pre-fractionation (single shot measurements) (15, 16). In addition, proteins and their post-translational modifications can be efficiently extracted from FFPE tissues (17). There have been complementary, enormous advances in data analysis and data management tools, facilitating the wide adoption of MS-based proteomics. In particular, these developments mean that characterizing small cohorts of human cancer patients in a reasonable amount of time is finally becoming feasible.

Previously, we have successfully subtyped DLBCL cell lines on the basis of their total protein expression patterns (18) and on their N-glycosylated peptide patterns (19). In this study, we decided to explore the applicability of our high-resolution MS-based platform to the problem of cancer subtyping from macro-dissected slices of FFPE tissue from patient samples. For quantification, we took advantage of the high accuracy of the super-SILAC approach (20) and combined it with label-free quantification of the proteins not present in the spiked-in standard. In addition to segregating cancer subtypes by our previously derived 55-protein signature and by the total protein expression patterns, we derived a novel combination of statistical feature selection and machine learning to define a small signature of differentiating proteins with the

highest segregating power. This analysis also allowed us to dissect important molecular differences between the subtypes.

EXPERIMENTAL PROCEDURES

Generation of the lymphoma super-SILAC mix – The super-SILAC mix was generated by combining equal amounts of heavy lysates from six lymphoma cell lines (Ramos, Mutu, BL-41, U2932, L428, and DB) as described (18). Stocks of this mix were prepared and used as standards that were spiked in each of the cell lines we previously studied and the 20 patient samples we analyzed in this study.

FFPE human tissues – FFPE samples of DLBCL were obtained from the Institute of Pathology, Charité - Universitätsmedizin Berlin. Analysis of the samples followed an informed consent approved by the local ethics committee.

Protein extraction from FFPE DLBCL tissues – For each patient sample, two FFPE slices of macro-dissected tissue were collected (13 µm thickness). They were processed for mass-spectrometry-based proteome analysis by extraction and digestion according to the Filter Aided Sample Preparation (FASP) protocol (FFPE-FASP) (17, 21). In short, FFPE tissue slices were incubated in 1 ml xylene (2x) with gentle agitation for 5 min at room temperature. After removing the paraffin, the samples were dried by incubating them in 1 ml absolute ethanol (2x). The dried samples were then lysed in a buffer consisting of 0.1 M Tris - HCl (pH 8.0), 0.1 M DTT and 4% SDS. After homogenization using a disperser, they were boiled at 99 °C using a heating block with agitation (600 rpm) for 60 min. The samples were then cleared by centrifugation.

Protein digestion and peptide fractionation – On a 30 KDa filter (Millipore, Billerica, MA, USA), 100 µg of each of the patient samples and the super-SILAC mix were mixed. The samples were further processed by the FASP method in which the SDS buffer is exchanged with a urea buffer (21). This was followed by alkylation with iodoacetamide and overnight digestion by trypsin at 37°C in 50 mM ammonium bicarbonate. The tryptic peptides were collected by centrifugation and elution with water (2x).

Strong anion exchange (SAX) chromatography was used to fractionate 40 µg of peptides from each patient sample (22). It was performed in tip-based columns from 200 µl micropipette tips stacked with 6 layers of a 3M Empore anion exchange disk (1214-5012; Varian, Palo Alto, CA). For the fractionation, a Britton & Robinson universal buffer (20 mM acetic acid, 20 mM phosphoric acid, and 20 mM boric acid) was used and titrated using NaOH to six buffers with the desired pHs (pH 11, 8, 6, 5, 4, and 3). Subsequently, six fractions from each sample were collected, followed by desalting the eluted fractions on reversed phase C18 Empore disc StageTips (23). The peptides were eluted from the StageTips using 20 µl of buffer B composed of 80% ACN in 0.5% acetic acid (2x). A SpeedVac concentrator prepared the samples for MS analysis by removing the organic solvents.

LC-MS/MS analysis – Peptides were separated by nanoflow HPLC (Thermo Fisher Scientific) coupled on-line to a quadrupole Orbitrap mass spectrometer (Q Exactive, Thermo Fisher Scientific) with a nanoelectrospray ion source. The peptides were eluted at a flow rate of 200 nl min^{-1} on an in-house made C_{18}-reversed phase column that was 50 cm long, 75 µm inner diameter and packed with ReproSil-Pur C18-AQ 1.8 µm resin (Dr. Maisch GmbH, Ammerbuch-

Entringen, Germany) in buffer A (0.5% acetic acid). For optimal separation based on average peptide hydrophobicity, four different linear gradients over a period of 205 min were applied. For pH 11 fraction, a gradient of 2–25% buffer B; for pH 8 fraction, a gradient of 7–25% buffer B; for pH 6 and 5 fractions, a gradient of 7–30% buffer B; for pH 4 and 3 fractions, a gradient of 7–37% buffer B. Each gradient was followed by column washing reaching 95% B and then re-equilibration with buffer A.

A data dependent 'top 10' method, in which the 10 most abundant precursor ions were selected for fragmentation, was used to acquire the data. For survey scans (mass range 300 – 1750 Th), the target value was 3,000,000 with a maximum injection time of 20 ms and a resolution of 70,000 at *m*/*z* 400. An isolation window of 1.6 Th was used for higher energy collisional dissociation with normalized collision energies of 25. For MS/MS scans, the target ion value was set to 1,000,000 with a maximum injection time of 60 ms and a resolution of 17,500 at *m*/*z* 400 and dynamic exclusion of 25s. This led to a constant injection time of 60 ms, which is fully in parallel with transient acquisition of the previous scan, ensuring fast cycle times.

The patient samples were received in two batches of 10 each, which were acquired with the same MS methods. For MS/MS, the 2nd batch a data dependent 'top 5' method was used where the 5 most intense ions from the survey scan were selected with an isolation window of 2.2 Th and dynamic exclusion of 45 s. The target ion value was set to 100,000 with a maximum injection time of 120 ms and a resolution of 17,500 at *m*/*z* 400.

Data analysis – We used the MaxQuant software environment (version 1.2.6.20) to analyze MS raw data. The MS/MS spectra were searched against the Uniprot database using the

Andromeda search engine incorporated in the MaxQuant framework (24, 25). Cysteine carbamidomethylation was set as a fixed modification and N-terminal acetylation and methionine oxidation as variable modifications. The maximum false discovery rate for both peptide and protein identifications was set to 0.01. Strict specificity for trypsin cleavage was required allowing cleavage N-terminal to proline. The minimum required peptide length was seven amino acids with a maximum of two miscleavages allowed. The initial precursor mass tolerance was 6 ppm and for the fragment masses it was up to 20 ppm. Time-dependent recalibration algorithm of MaxQuant was used to improve the precursor mass ions mass accuracy. The "match between runs" option was enabled, allowing the matching of identifications across measurements. Relative quantification of the peptides against their SILAC-labeled counterparts was performed with MaxQuant using a minimum ratio count of 1. We combine SILAC with label-free analysis ('hybrid algorithm') employing a minimum count of 1 (see RESULTS AND DISCUSSION). Perseus, which is a module of the MaxQuant software package, was used for the further statistical and bioinformatic analysis of the MaxQuant output data. Missing values were supplied by 'data imputation' to simulate signals of low abundant proteins under the assumption that they are biased toward the detection limit of the MS measurement (18).

RESULTS AND DISCUSSION

Workflow for quantitative proteome measurements of DLBCL FFPE patient samples – One of the most commonly used methods for tissue preservation involves fixing the sample in formalin followed by embedding it in paraffin, formalin-fixed paraffin-embedded (FFPE) tissues. It is

routinely used in tissue banks due to its compatibility with immunohistochemistry assays and its long-term preservation benefits in an economical format. However, FFPE cohorts have been challenging to use in gene expression studies due to the difficulty to isolate nucleic acids (26). Despite attempts to improve the quality of extracted RNA samples from FFPE tissues and to provide standardized protocols, currently snap-frozen tissues are greatly preferred in that workflow (10, 26). In clinical practice, tissue banks of frozen specimens are used for initial discovery studies but by far the largest sample numbers and almost all tumor specimens, are fixed in formalin. Taking advantage of the stability and ease-of-handling of proteins, we and others have recently shown that protein extraction from FFPE material is possible in a robust manner (17, 27). We did not observe quantitative or qualitative differences between FFPE and frozen tissues at the level of proteins or post-translational modifications (PTMs) (17). Our approach combined boiling in sodium dodecyl sulfate (SDS) with the filter aided sample preparation (FASP) method (21). The boiling step presumably reverses the cross-links induced upon fixation whereas the FASP method allows MS analysis of proteomic samples solubilized in high concentrations of SDS, which is advantageous for FFPE samples (27).

Here we macro-dissected two slices from each of 20 FFPE tumor samples from DLBCL patients (Fig. 1A). Peptides resulting from FASP preparation were subjected to six-step fractionation using a strong anion exchange chromatography (SAX) protocol followed by LC-MS analysis of each fraction (see EXPERIMENTAL PROCEDURES).

Accurate quantification is a requirement for the comparison of the protein expression profiles of the patient samples. For the 20 patient samples we used the same heavy-labeled super-

SILAC mix of six lymphoma cell lines that we had previously constructed to represent the lymphoma proteome as closely as possible and to cover as many of 'lymphoma-related' proteins as possible (18). Heavy lysates from each of the six cell lines were pooled together and spiked in a 1:1 ratio to each of the patient samples. To also quantify SILAC singlets for which the peptide is not found in the reference proteome but is seen in the samples, we introduce a new quantification algorithm in MaxQuant. This so called hybrid quantification algorithm is a generalization of the MaxLFQ algorithm for the accurate relative quantification of label-free data (28). The essence of the relative quantification step in MaxLFQ is that for each protein and for each sample pair the ratio is calculated for those peptide features that were determined in both samples. In the hybrid quantification algorithm one distinguishes the case in which a SILAC ratio to the reference is calculated in both samples for a given peptide feature from the case in which one or both ratios cannot be calculated. If both ratios are available, the ratio of ratios is used as input for the MaxLFQ quantification algorithm. In the other case and given that intensities are calculated in both samples for the light SILAC state, the ratio of these light intensities is taken. In the case that one or both light intensities are not present, the peptide feature does not take part in the quantification. All other steps of the MaxLFQ algorithm are applied in exactly the same way in the hybrid LFQ algorithm as well. The result of the hybrid algorithm is an intensity profile for each protein group over all samples, similar to the output of the conventional MaxLFQ algorithm. The whole intensity profile for a protein group can be multiplied with an arbitrary factor since only the relative intensity information is defined by the algorithm.

Combined analysis of the raw MS data by MaxQuant resulted in the identification of 9,012 protein groups across the 20 patient samples. We obtained quantitative results for 8,701 protein groups after employing the hybrid LFQ algorithm with an average of 6,278 protein groups in each of the 20 DLBCL patient samples. The average gain from the hybrid LFQ is 353 quantifications per sample compared to using SILAC ratios alone (supplemental Fig. 1). This relatively small percentage indicates that the vast majority of proteins were adequately quantifiable against the super SILAC standard already.

General characteristics of the proteome of 20 DLBCL FFPE patient samples – The achieved depth of the proteome resulted in good quantitative coverage of many signaling pathways and cellular processes that play a role in the development and progression of various cancers (Fig. 1B). These include processes such as DNA replication (94% coverage of annotated members) and apoptosis (77%). Importantly, there is almost complete coverage (91%) of the B-cell receptor signaling pathway, which can play a major role in lymphomagenesis, and high coverage of other blood cancer-associated proteins such as acute myeloid leukemia (83%) and chronic myeloid leukemia (83%).

Pairwise comparisons of all the samples against each other resulted in high Pearson coefficients between the samples (average r= 0.92) indicating both high quantitative accuracy between tumor measurements and high similarity in the global proteomes (see Fig. 2A for an example).

The dynamic range of MS signals for proteins from the patient samples proteomes spanned seven orders of magnitude with 94% of the proteins concentrated in four orders of magnitude (Fig. 2B). Overlaying 172 proteins that are annotated in the KEGG database as belonging to

pathways in cancer showed that cancer-related proteins spanned the entire dynamic range. This suggests that both highly and low abundant proteins can be important players in cancer cell biology (Fig. 2B).

Compared to the cell line system we previously analyzed, we here found 2,031 additional protein groups (Fig. 3A). We attribute this to technical factors, mainly the very fast and sensitive quadrupole-Orbitrap used in this study (29), in combination with the larger complexity of the patient samples. This interpretation is supported by the abundance distribution of the extra 2,031 protein groups, which was at the lower end of the total distribution (Fig. 3B). Furthermore, a Fisher exact test showed a high enrichment of extracellular proteins in this set of proteins. This is especially interesting as stromal signatures have already been shown to be important in lymphoma classification (6).

The 55-protein cell line-derived signature correctly classifies patients – We have previously derived a signature of 55 proteins that robustly segregated ABC-DLBCL and GCB-DLBCL in a cell line system (18). In addition to proteins that correlated to underlying known biological differences between the subtypes, the cell line signature also included new interesting candidates. To explore the potential of applying this signature to patients, we used the COO subtypes previously established by gene expression profiles on these samples (30). Matching the signature to the patient proteomes after filtering for 75% valid values resulted in quantitative values of 49 proteins in all of the patients. Remarkably, a principal component analysis (PCA) of these matches clearly segregated the two subtypes (Fig. 3C). Thus, our previous proteomic signature can directly be translated to patient samples and classify them

correctly, which is remarkable because it was derived entirely from a cell line based system. The loadings of component 1, which accounts for 25.7 % of the variability in this small subset of proteins, drive the correct segregation. However, this does not necessarily mean that the cell line signature is optimal to segregate the subtypes with the best possible accuracy. With the increased depth and faithfulness of the patient samples, a signature extracted from the patient proteomes themselves is worth investigating and evaluating, as addressed below.

Unsupervised segregation of patient samples based on their global protein expression profiles – To explore whether the global protein expression profiles of the patient samples would reveal intrinsic biological differences between the subtypes such as their different COO, we performed a principal component analysis based on the entire protein expression profile of each patient. As previously, we filtered for 75% valid values resulting in 5,480 protein groups quantified across the 20 patients. Components 1 versus 4 in the PCA provided a diagonal segregation of the patient samples according to their COO classification (Fig. 4A). The 'loadings' of such a PCA reveal the drivers causing the segregation (Fig. 4B). Among the proteins that are relatively upregulated in ABC-DLBCL are PTPN1 (PTP1B), IRF4, CCDC50 (Ymer), MNDA, SP140, IL16, RAB7L1, HCK, TNFAIP8, TNFAIP2, and HELLS. Reassuringly, many of these candidates reflect known biological differences between the subtypes. Strong drivers of segregation such as PTPN1, IRF4, CCDC50 as well as metabolic enzymes such as ARHGAP17 and CYB5R2 were already present in our previously derived cell line signature. This explains the applicability of the cell line-derived signature to segregate patient tissue proteomes and independently confirms the importance of these markers because they were picked up in two independent studies of this cancer system. For instance, IRF4, one of the strong drivers that we previously

highlighted, is a transcription factor that drives plasmacytic differentiation and its expression is directly regulated by NF-κB signaling, a pathogenic hallmark of ABC-DLBCL (31). A new drug (lenalidomide), which inhibits IRF4, selectively kills ABC-DLBCL cells and is currently in clinical trials (32).

The strongest drivers also include some interesting new candidates. One of the strong drivers that are upregulated in ABC-DLBCL is SP140, an interferon-inducible, nuclear lymphocyte-specific protein of unknown function. It is expressed in all human mature B cells and plasma cell lines, as well as in some T cells (33, 34). It possesses several chromatin related modules, which suggests a role of SP140 in chromatin-mediated regulation of gene expression (35). A genome-wide association study of single-nucleotide polymorphisms (SNPs) for chronic lymphocytic leukemia (CLL) showed that SP140 is a CLL risk locus. That study also identified IRF4 as another risk locus out of six loci in total (36), a remarkable overlap with our results. The myeloid cell nuclear differentiation antigen (MNDA) is another strong driver that emerged from the patient data. As the name indicates, MNDA is expressed constitutively in cells of the myeloid lineage, but it can also be expressed by normal and neoplastic B lymphocytes (37, 38). In a recent study that identified MNDA as a marker for nodal marginal zone lymphoma, the authors also analyzed the expression of MNDA in a cohort of 75 DLBCL cases. Interestingly, out of 34 cases in which it was highly expressed, 25 were of the ABC subtype (39). A highly interesting and novel segregator is IL16, a cytokine that is typically characterized as a chemoattractant of CD4+ cells to sites of inflammation. However, recent studies have suggested an important role of both the pro-molecule and the secreted form of IL-16 in the regulation of lymphocytic cancer cell proliferation (40). In fact, targeting IL-16 may be a novel therapeutic approach for T cell cancers

(cutaneous T cell lymphoma) and B cell cancers (multiple myeloma). In multiple myeloma, inhibition of IL16 production by siRNA or IL-16 bioactivity by neutralizing antibodies reduces cell proliferation by more than 80% (40).

On the other side of the diagonal segregation are drivers with higher protein levels in the GCB-DLBCL subtype. These include ABCC4, TBC1D4, LCK, CAV1, C3orf37 (HMCES), IGF2BP1 and TP53. TBC1D4 is a Rab GTPase-activating protein that promotes insulin-induced glucose transporter GLUT4 translocation to the plasma membrane, thus increasing glucose uptake (41). TBC1D4 has not yet been associated with lymphoma classification, but may be related to increased glucose uptake as observed in many cancer types and may indicate a difference between the cancer types in this respect (42). LCK is a lymphocyte cell-specific protein-tyrosine kinase studied extensively in the context of T-cells where it plays an important role in signal transduction after antigen binding. Dysregulation of LCK expression or LCK kinase activity has been implicated in T cell leukemia from mice to humans (43). LCK expression has also been reported in normal B-1 cells and in chronic lymphocytic leukemia B cells (44). It plays an important role in B-cell receptor signaling in CLL and specific LCK inhibitors have been suggested in the treatment of progressive CLL (45). Reassuringly, LCK has been shown to be present at high levels in normal germinal center cells (46). In addition, it was shown to be expressed in most lymphomas of germinal center origin (e.g. follicular lymphoma) and also many mantle cell lymphomas, chronic lymphocytic leukemia (CLL) and most T-cell neoplasms (46).

The diagonal segregation of the subtypes suggested that other biological factors compromised a more clear-cut COO segregation of the patients in the PCA. Enrichment analysis of protein

categories showed that *extracellular matrix region part* is one of the strongest cellular component categories (GOCC) significantly enriched in component 1 of the PCA (FDR=1.89E-33). Cancer module (CM) categories (GSEA) correspond to gene sets which are significantly changed in a variety of cancer conditions after mining a large compendium of cancer related microarray data (47). The most significantly enriched CM module in component 1 was MODULE_47 (FDR=6.55E-20) (Fig. 4C). This category included proteins such as ACTN1, BGN, COL1A1, COL1A2, COL6A1, COL6A2, COL6A3, COL6A4, FN1, LUM, POSTN and SERPINH1 (Fig. 4C). There is a large overlap between these drivers and the reported prognostically favorable stromal-1 signature, reflecting extracellular matrix deposition (6). In fact, the stromal signatures study showed that a multivariate model created from three gene-expression signatures - germinal-center B-cell (COO), stromal-1 (extracellular matrix deposition), and stromal-2 (tumor blood-vessel density) - was a better predictor of survival than the COO classification alone. Hence, survival of DLBCL patients after treatment is influenced by several biological attributes including the COO and the tumor microenvironment (6). In addition, expression levels of the ECM signature proteins we depicted in component 1 are on average higher in the GCB subtype. These findings confirm what has been previously reported (48) and show that our proteomic analysis captured the COO classification as well as other intrinsic biological differences between the subtypes.

Supervised characterization of ABC-DLBCL versus GC-DLBCL subtypes – After assigning a subtype to each patient sample based on GEP classification, we treated the samples as biological replicates of the same disease entity. We grouped patients belonging to the same subtype together and calculated the median expression value for each protein group. The proteomes of

GCB-DLBCL versus ABC-DLBCL had very high correlation (Pearson r = 0.98). Against this background of very high overall similarity, investigation of outliers from this tight cloud revealed markers that our unsupervised PCA analysis had already indicated as well as novel candidate markers which are connected to the known biology of the disease (Fig. 5A). This included TCL1A, FOXP1 and TLR9, which are upregulated in the ABC subtype. For instance, both TCL1A and FOXP1 are immunohistochemical markers of adverse outcome in DLBCL (49, 50). FOXP1 was also reported to occur in a subgroup of non-GC DLBCLs (51) and TCL1A has been suggested as tumor-associated antigen for immunotherapeutic strategies in common B-cell lymphomas (52).

Next we performed a 2D annotation enrichment analysis (53) using cancer modules (CM) for deriving differential cancer associated gene sets between these two closely related entities of DLBCL. As expected from the high proteome correlation, the subtypes are very similar in almost every cancer module annotated such as RNA splicing, protein biosynthesis and mitosis. However, MODULE_456 which corresponds to 'B lymphoma expression clusters' and MODULE_ 210 which corresponds to 'metallopeptidase activity' are different between the subtypes. MODULE_456 consists of 115 genes and is annotated to be significantly induced in B-cell lymphomas (p=2.7e-05) and specifically in GC-DLBCL (p=3.0e-05). This confirms what we observed in our analysis (Fig. 5B). The metallopeptidase and metalloendopeptidase gene sets comprising MODULE_210 consists of 28 genes and were significantly induced in microarrays of DLBCL (p=1.5e-06) and GC-DLBCL (p=5.1e-05) specifically (47). The proteins that we found in this gene set are particularly interesting given the role of MMPs in mediating tumor invasion.

The candidate differentially expressed proteins and categories clearly reflected relevant biological differences between ABC-DLBCL and GCB-DLBCL. However, these candidates can not necessarily be used as markers of classification. More sophisticated statistical tools are required to achieve a panel of candidate proteins that can be used for diagnostic purposes as discussed in the next section.

Support Vector Machines Feature Selection – In clinical studies, tumor and host variability combined with the large feature space of the data set (thousands of proteins compared to a relatively small number of patients) make it difficult to identify disease-relevant proteins. We addressed these challenges with a supervised learning method – Support Vector Machines (SVMs) - in combination with a test statistics based feature selection strategy. SVMs are a well-established machine learning technique that trains a predictor that best distinguishes between the known classes of the samples (in our case GC and ABC lymphoma subtypes). The principle of an SVM predictor is the definition of a so-called separation hyperplane that segregates the subtypes as clearly as possible in a training data set, which can be a subset of the measured samples. Using this 'machine learned' hyperplane, new samples of unknown subtype can be classified as GC or ABC depending on the side of the separation hyperplane on which each of these samples falls. The strength of SVMs lies in their ability to perform well in high dimensional data and in particular to efficiently find and assess sets of features with high predictive power.

We combined the SVM-based prediction with feature selection to optimize the performance of the classifier and to identify strongly discriminative features. The feature selection method employed p-values from standard ANOVA tests. As disease-relevant features that show large

quantitative differences between the two subtypes are more easily detectable and thus are clinically more relevant, we performed the ranking of the proteins such that it depended not only on the statistical significance of their differential expression between the different subtypes, but also on the actual size of this difference. The advantage of this method is that proteins with low p-values and high fold change receive higher ranks than those with low p-values and small fold change.

Feature selection was embedded in a cross-validation procedure to avoid the problem of over-fitting and wrong estimation of the classifier's performance. In each iteration (total 1000) of a random sampling cross validation, we used 90% of the data for training and feature ranking and the rest for testing and optimization of the number of features. The analysis resulted in a set of four ranked features that perform almost perfectly in the classification of the subtypes (1.4% error rate) (Fig. 6A). These top four candidates are: TBC1D4, PALD1, TNFAIP8 and MME (CD10). MME is part of previous immunohistochemistry-based classification algorithms (11). TBC1D4 plays a role in glucose uptake, TNFAIP8 is NF-κB regulated and involved in blocking apoptosis, and PALD1 is newly studied protein that may play a role in tumor invasiveness and metastasis.

Next, we were interested in comparing ranked features with the digital gene expression (NanoString)–based test of 20 genes that has been recently published and put into use in a clinical trial (12). The model is composed of 8 genes (TNFRSF13B, LIMD1, IRF4, CREB3L2, PIM2, CYB5R2, RAB7L1 and CCDC50) overexpressed in ABC-DLBCL, 5 housekeeping genes, and 7 genes (MME, SERPINA9, ASB13, MAML3, ITPKB, MYBL1 and SIPR2) overexpressed in GC-DLBCL. Gratifyingly, there is a 30% overlap in the differentially expressed genes in our data set.

For a broader selection of differential features, we used as an error rate cutoff, the point beyond which the correct unsupervised hierarchical clustering of the subtypes was lost. This resulted in 343 features (Fig. 6B). Interestingly, upon filtering for ECM, nuclear and plasma membrane proteins from these top 343 features, the last two categories maintained correct segregation on their own reflecting the cell-of-origin classification (Fig. 6C).

The set of 343 protein groups included 33 transcription factors, 14 protein kinases, and 12 oncogenes (supplemental Table I). Upon dividing the 343 protein groups into their two main clusters: one relatively upregulated in ABC-DLBCL and the second relatively upregulated in GCB-DLBCL we performed network analysis to investigate possible connections between them. Genes upregulated in the ABC-DLBCL subtype highlighted the CARD11-PKCB signaling core (supplemental Fig. 2A) that drives NF-κB signaling upon BCR signaling (54). The GCB-DLBCL subtype showed an LCK-PAG-P2K signaling module (supplemental Fig. 2B) which has been shown to be oncogenic in other lymphomas (55). In addition to an ECM core that we previously depicted to be upregulated on average in the GCB subtype, we also observe an MHCII network that has been previously reported to be on average higher in GCB (48).

CONCLUSIONS AND OUTLOOK

Previously, we had shown unambiguous segregation of patient-derived DLBCL cell lines into their COO subtypes based on their global protein expression profiles as well as an enriched set of membrane proteins (18, 19). In this study, we have analyzed 20 FFPE DLBCL patient samples, attaining a quantitative depth of more than 9000 proteins, which to our knowledge, is the largest lymphoma proteome available. Correct segregation of the subtypes based on their

protein expression profiles was possible after applying a cell line-derived signature from our previous studies or by using the whole set of proteins quantified in at least 75% of the samples. When global protein expression profiles were employed, the COO classification was not as clear cut as in the cell lines. This is most likely due to increased complexity of this system in which several important biological signatures (extracellular matrix and MHC II) also influence segregation. In fact, these signatures are known to be very valuable in the overall prediction of survival in DLBCL patients (48). Our results clearly show that global expression proteomics can segregate cancer types based on tumor samples from patients. Importantly for practical applications, our measurements only require small amounts of FFPE material, which are readily available in tissue banks or informal sample collections.

The high number of biologically relevant potential markers retrieved here argues well for future applications of proteomics to clinical questions such as tumor segregation. Our analysis highlighted both the COO signature and the ECM signature in line with the 'gold standard' predictor of survival which includes the COO classification as well as stromal signatures (6, 30). Nuclear and membrane proteins reflect the COO but ECM signature is more likely to reflect mechanisms the tumor develops to interact with its environment. Hence, they are at least partly independent signatures and patient survival depends on both.

In a classical view of biomarker development, global MS-based proteomics play a role primarily in the discovery phase (56). In post-discovery studies, MS-based or ELISA-based targeted approaches would then be employed on specific signature proteins. However, it is interesting to speculate that an untargeted approach could also be used in this phase. Recent technological

advances, especially at the peptide preparation, separation and MS-instrument levels, have led to powerful single-shot approaches that are positioned between in-depth shotgun proteomics employing fractionation and targeted approaches (15, 16). An appealing application of such single-shot systems would be the analysis of patient samples where measuring a large cohort is necessary for statistical analysis and validation. Considering the rate of MS developments, measuring a proteome of complex biological samples such as patient tissues comprehensive enough for tumor classification in about an hour should be achievable in the near future. Furthermore, simpler and robust sample preparation methods will allow easier sample handling and higher reproducibility (57). In conclusion, continuous MS-based technological advances hold great promise for future characterization and diagnosis of subtypes not only of B-cell lymphomas but any closely related tumor subtypes.

Acknowledgments—We thank Prof. Michael Hummel, Dr. Dido Lenze and Stefanie Mende for the provision of patient samples.

REFERENCES

1. van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. Ten years of next-generation sequencing technology. *Trends in Genetics* 30, 418-426.
2. Schilsky, R. L. (2010) Personalized medicine in oncology: the future is now. *Nat Rev Drug Discov* 9, 363-366.
3. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503-511.
4. Wright, G., Tan, B., Rosenwald, A., Hurt, E. H., Wiestner, A., and Staudt, L. M. (2003) A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proceedings of the National Academy of Sciences* 100, 9991-9996.
5. Roschewski, M., Staudt, L. M., and Wilson, W. H. (2014) Diffuse large B-cell lymphoma[mdash]treatment approaches in the molecular era. *Nat Rev Clin Oncol* 11, 12-23.
6. Lenz, G., Wright, G., Dave, S. S., Xiao, W., Powell, J., Zhao, H., Xu, W., Tan, B., Goldschmidt, N., Iqbal, J., Vose, J., Bast, M., Fu, K., Weisenburger, D. D., Greiner, T. C., Armitage, J. O., Kyle, A., May, L., Gascoyne, R. D., Connors, J. M., Troen, G., Holte, H., Kvaloy, S., Dierickx, D., Verhoef, G., Delabie, J., Smeland, E. B., Jares, P., Martinez, A., Lopez-Guillermo, A., Montserrat, E., Campo, E., Braziel, R. M., Miller, T. P., Rimsza, L. M., Cook, J. R., Pohlman, B., Sweetenham, J., Tubbs, R. R., Fisher, R. I., Hartmann, E., Rosenwald, A., Ott, G., Muller-Hermelink, H.-K., Wrench, D., Lister, T. A., Jaffe, E. S., Wilson, W. H., Chan, W. C., and Staudt, L. M. (2008) Stromal Gene Signatures in Large-B-Cell Lymphomas. *New England Journal of Medicine* 359, 2313-2323.
7. van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530-536.
8. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330-337.
9. Raspe, E., Decraene, C., and Berx, G. (2012) Gene expression profiling to dissect the complexity of cancer biology: Pitfalls and promise. *Seminars in Cancer Biology* 22, 250-260.
10. Perry, A. M., Cardesa-Salzmann, T. M., Meyer, P. N., Colomo, L., Smith, L. M., Fu, K., Greiner, T. C., Delabie, J., Gascoyne, R. D., Rimsza, L., Jaffe, E. S., Ott, G., Rosenwald, A., Braziel, R. M., Tubbs, R., Cook, J. R., Staudt, L. M., Connors, J. M., Sehn, L. H., Vose, J. M., López-Guillermo, A., Campo, E., Chan, W. C., and Weisenburger, D. D. (2012) A new biologic prognostic model based on immunohistochemistry predicts survival in patients with diffuse large B-cell lymphoma. *Blood* 120, 2290-2296.
11. Gutiérrez-García, G., Cardesa-Salzmann, T., Climent, F., González-Barca, E., Mercadal, S., Mate, J. L., Sancho, J. M., Arenillas, L., Serrano, S., Escoda, L., Martínez, S., Valera, A., Martínez, A., Jares, P., Pinyol, M., García-Herrera, A., Martínez-Trillos, A., Giné, E., Villamor, N., Campo, E., Colomo, L., López-Guillermo, A., and Balears, f. t. G. p. l. E. d. L. d. C. I. (2011) Gene-expression profiling and not immunophenotypic algorithms predicts prognosis in patients with diffuse large B-cell lymphoma treated with immunochemotherapy. *Blood* 117, 4836-4843.
12. Scott, D. W., Wright, G. W., Williams, P. M., Lih, C.-J., Walsh, W., Jaffe, E. S., Rosenwald, A., Campo, E., Chan, W. C., Connors, J. M., Smeland, E. B., Mottok, A., Braziel, R. M., Ott, G., Delabie, J., Tubbs, R. R., Cook, J. R., Weisenburger, D. D., Greiner, T. C., Glinsmann-Gibson, B. J., Fu, K., Staudt, L. M.,

Gascoyne, R. D., and Rimsza, L. M. (2014) *Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue*.
13. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol* 7.
14. Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011) The quantitative proteome of a human cell line. *Molecular Systems Biology* 7.
15. Nagaraj, N., Alexander Kulak, N., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., Vorm, O., and Mann, M. (2012) System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap. *Molecular & Cellular Proteomics* 11.
16. Mann, M., Kulak, Nils A., Nagaraj, N., and Cox, J. (2013) The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. *Molecular cell* 49, 583-590.
17. Ostasiewicz, P., Zielinska, D. F., Mann, M., and Wiśniewski, J. R. (2010) Proteome, Phosphoproteome, and N-Glycoproteome Are Quantitatively Preserved in Formalin-Fixed Paraffin-Embedded Tissue and Analyzable by High-Resolution Mass Spectrometry. *Journal of Proteome Research* 9, 3688-3700.
18. Deeb, S. J., D'Souza, R. C. J., Cox, J., Schmidt-Supprian, M., and Mann, M. (2012) Super-SILAC Allows Classification of Diffuse Large B-cell Lymphoma Subtypes by Their Protein Expression Profiles. *Molecular & Cellular Proteomics* 11, 77-89.
19. Deeb, S. J., Cox, J., Schmidt-Supprian, M., and Mann, M. (2014) N-linked Glycosylation Enrichment for In-depth Cell Surface Proteomics of Diffuse Large B-cell Lymphoma Subtypes. *Molecular & Cellular Proteomics* 13, 240-251.
20. Geiger, T., Cox, J., Ostasiewicz, P., Wisniewski, J., and Mann, M. (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat Methods* 7, 383 - 385.
21. Wisniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009) Universal sample preparation method for proteome analysis. *Nat Methods* 6, 359-362.
22. Wiśniewski, J. R., Zougman, A., and Mann, M. (2009) Combination of FASP and StageTip-Based Fractionation Allows In-Depth Analysis of the Hippocampal Membrane Proteome. *Journal of Proteome Research* 8, 5674-5678.
23. Rappsilber, J., Ishihama, Y., and Mann, M. (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* 75, 663-670.
24. Cox, J., and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26, 1367-1372.
25. Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., and Mann, M. (2011) Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J Proteome Res*.
26. Hewitt, S. M., Lewis, F. A., Cao, Y., Conrad, R. C., Cronin, M., Danenberg, K. D., Goralski, T. J., Langmore, J. P., Raja, R. G., Williams, P. M., Palma, J. F., and Warrington, J. A. (2008) Tissue Handling and Specimen Preparation in Surgical Pathology: Issues Concerning the Recovery of Nucleic Acids From Formalin-Fixed, Paraffin-Embedded Tissue. *Archives of Pathology & Laboratory Medicine* 132, 1929-1935.
27. Shi, S.-R., Liu, C., Balgley, B. M., Lee, C., and Taylor, C. R. (2006) Protein Extraction from Formalin-fixed, Paraffin-embedded Tissue Sections: Quality Evaluation by Mass Spectrometry. *Journal of Histochemistry & Cytochemistry* 54, 739-743.
28. Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014) MaxLFQ allows accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction. *Molecular & Cellular Proteomics*.

29. Michalski, A., Damoc, E., Hauschild, J. P., Lange, O., Wieghaus, A., Makarov, A., Nagaraj, N., Cox, J., Mann, M., and Horning, S. (2011) Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Mol Cell Proteomics* 10, M111 011015.
30. Pfeifer, M., Grau, M., Lenze, D., Wenzel, S.-S., Wolf, A., Wollert-Wulf, B., Dietze, K., Nogai, H., Storek, B., Madle, H., Dörken, B., Janz, M., Dirnhofer, S., Lenz, P., Hummel, M., Tzankov, A., and Lenz, G. (2013) PTEN loss defines a PI3K/AKT pathway-dependent germinal center subtype of diffuse large B-cell lymphoma. *Proceedings of the National Academy of Sciences* 110, 12420-12425.
31. Davis, R. E., Brown, K. D., Siebenlist, U., and Staudt, L. M. (2001) Constitutive Nuclear Factor κB Activity Is Required for Survival of Activated B Cell–like Diffuse Large B Cell Lymphoma Cells. *The Journal of Experimental Medicine* 194, 1861-1874.
32. Yang, Y., Shaffer Iii, Arthur L., Emre, N. C. T., Ceribelli, M., Zhang, M., Wright, G., Xiao, W., Powell, J., Platig, J., Kohlhammer, H., Young, Ryan M., Zhao, H., Yang, Y., Xu, W., Buggy, Joseph J., Balasubramanian, S., Mathews, Lesley A., Shinn, P., Guha, R., Ferrer, M., Thomas, C., Waldmann, Thomas A., and Staudt, Louis M. (2012) Exploiting Synthetic Lethality for the Therapy of ABC Diffuse Large B Cell Lymphoma. *Cancer Cell* 21, 723-737.
33. Dent, A., Yewdell, J., Puvion-Dutilleul, F., Koken, M., de The, H., and Staudt, L. (1996) LYSP100-associated nuclear domains (LANDs): description of a new class of subnuclear structures and their relationship to PML nuclear bodies. *Blood* 88, 1423-1426.
34. Bloch, D. B., de la Monte, S. M., Guigaouri, P., Filippov, A., and Bloch, K. D. (1996) Identification and Characterization of a Leukocyte-specific Component of the Nuclear Body. *Journal of Biological Chemistry* 271, 29198-29204.
35. Zucchelli, C., Tamburri, S., Quilici, G., Palagano, E., Berardi, A., Saare, M., Peterson, P., Bachi, A., and Musco, G. (2014) Structure of human Sp140 PHD finger: an atypical fold interacting with Pin1. *FEBS Journal* 281, 216-231.
36. Di Bernardo, M. C., Crowther-Swanepoel, D., Broderick, P., Webb, E., Sellick, G., Wild, R., Sullivan, K., Vijayakrishnan, J., Wang, Y., Pittman, A. M., Sunter, N. J., Hall, A. G., Dyer, M. J. S., Matutes, E., Dearden, C., Mainou-Fowler, T., Jackson, G. H., Summerfield, G., Harris, R. J., Pettitt, A. R., Hillmen, P., Allsup, D. J., Bailey, J. R., Pratt, G., Pepper, C., Fegan, C., Allan, J. M., Catovsky, D., and Houlston, R. S. (2008) A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nat Genet* 40, 1204-1210.
37. Miranda, R. N., Briggs, R. C., Shults, K., Kinney, M. C., Jensen, R. A., and Cousar, J. B. (1999) Immunocytochemical analysis of MNDA in tissue sections and sorted normal bone marrow cells documents expression only in maturing normal and neoplastic myelomonocytic cells and a subset of normal and neoplastic B lymphocytes. *Human Pathology* 30, 1040-1049.
38. Joshi, A. D., Hegde, G. V., Dickinson, J. D., Mittal, A. K., Lynch, J. C., Eudy, J. D., Armitage, J. O., Bierman, P. J., Bociek, R. G., Devetten, M. P., Vose, J. M., and Joshi, S. S. (2007) ATM, CTLA4, MNDA, and HEM1 in High versus Low CD38–Expressing B-Cell Chronic Lymphocytic Leukemia. *Clinical Cancer Research* 13, 5295-5304.
39. Kanellis, G., Roncador, G., Arribas, A., Mollejo, M., Montes-Moreno, S., Maestre, L., Campos-Martin, Y., Rios Gonzalez, J. L., Martinez-Torrecuadrada, J. L., Sanchez-Verde, L., Pajares, R., Cigudosa, J. C., Martin, M. C., and Piris, M. A. (2009) Identification of MNDA as a new marker for nodal marginal zone lymphoma. *Leukemia* 23, 1847-1857.
40. Richmond, J., Tuzova, M., Cruikshank, W., and Center, D. (2014) Regulation of Cellular Processes by Interleukin-16 in Homeostasis and Cancer. *Journal of Cellular Physiology* 229, 139-147.
41. Sano, H., Kane, S., Sano, E., Mîinea, C. P., Asara, J. M., Lane, W. S., Garner, C. W., and Lienhard, G. E. (2003) Insulin-stimulated Phosphorylation of a Rab GTPase-activating Protein Regulates GLUT4 Translocation. *Journal of Biological Chemistry* 278, 14599-14602.

42. Hanahan, D., and Weinberg, Robert A. (2011) Hallmarks of Cancer: The Next Generation. *Cell* 144, 646-674.
43. Yu, C. L., Jove, R., and Burakoff, S. J. (1997) Constitutive activation of the Janus kinase-STAT pathway in T lymphoma overexpressing the Lck protein tyrosine kinase. *The Journal of Immunology* 159, 5206-5210.
44. Majolini, M. B., D'Elios, M. M., Galieni, P., Boncristiano, M., Lauria, F., Del Prete, G., Telford, J. L., and Baldari, C. T. (1998) Expression of the T-Cell–Specific Tyrosine Kinase Lck in Normal B-1 Cells and in Chronic Lymphocytic Leukemia B Cells. *Blood* 91, 3390-3396.
45. Talab, F., Allen, J. C., Thompson, V., Lin, K., and Slupsky, J. R. (2013) LCK Is an Important Mediator of B-Cell Receptor Signaling in Chronic Lymphocytic Leukemia Cells. *Molecular Cancer Research* 11, 541-554.
46. Paterson, J., Tedoldi, S., Craxton, A., Jones, M., Hansmann, M., Collins, G., Roberton, H., Natkunam, Y., Pileri, S., Campo, E., Clark, E., Mason, D., and Marafioti, T. (2006) The differential expression of LCK and BAFF-receptor and their role in apoptosis in human lymphomas. *Haematologica* 91, 772-780.
47. Segal, E., Friedman, N., Koller, D., and Regev, A. (2004) A module map showing conditional activity of expression modules in cancer. *Nat Genet* 36, 1090-1098.
48. Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B., Giltnane, J. M., Hurt, E. M., Zhao, H., Averett, L., Yang, L., Wilson, W. H., Jaffe, E. S., Simon, R., Klausner, R. D., Powell, J., Duffey, P. L., Longo, D. L., Greiner, T. C., Weisenburger, D. D., Sanger, W. G., Dave, B. J., Lynch, J. C., Vose, J., Armitage, J. O., Montserrat, E., López-Guillermo, A., Grogan, T. M., Miller, T. P., LeBlanc, M., Ott, G., Kvaloy, S., Delabie, J., Holte, H., Krajci, P., Stokke, T., and Staudt, L. M. (2002) The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. *New England Journal of Medicine* 346, 1937-1947.
49. Ramuz, O., Bouabdallah, R., Devilard, E., Borie, N., Groulet-Martinec, A., Bardou, V. J., Brousset, P., Bertucci, F., Birg, F., Birnbaum, D., and Xerri, L. (2005) Identification of TCL1A as an immunohistochemical marker of adverse outcome in diffuse large B-cell lymphomas. *International journal of oncology* 26, 151-157.
50. Banham, A. H., Connors, J. M., Brown, P. J., Cordell, J. L., Ott, G., Sreenivasan, G., Farinha, P., Horsman, D. E., and Gascoyne, R. D. (2005) Expression of the FOXP1 Transcription Factor Is Strongly Associated with Inferior Survival in Patients with Diffuse Large B-Cell Lymphoma. *Clinical Cancer Research* 11, 1065-1072.
51. Barrans, S. L., Fenton, J. A. L., Ventura, R., Smith, A., Banham, A. H., and Jack, A. S. (2007) *Deregulated over expression of FOXP1 protein in diffuse large B-cell lymphoma does not occur as a result of gene rearrangement*.
52. Weng, J., Rawal, S., Chu, F., Park, H. J., Sharma, R., Delgado, D. A., Fayad, L., Fanale, M., Romaguera, J., Luong, A., Kwak, L. W., and Neelapu, S. S. (2012) *TCL1: a shared tumor-associated antigen for immunotherapy against B-cell lymphomas*.
53. Cox, J., and Mann, M. (2012) 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC Bioinformatics* 13, S12.
54. Hara, H., Wada, T., Bakal, C., Kozieradzki, I., Suzuki, S., Suzuki, N., Nghiem, M., Griffiths, E. K., Krawczyk, C., Bauer, B., D'Acquisto, F., Ghosh, S., Yeh, W.-C., Baier, G., Rottapel, R., and Penninger, J. M. (2003) The MAGUK Family Protein CARD11 Is Essential for Lymphocyte Activation. *Immunity* 18, 763-775.
55. Tauzin, S., Ding, H., Burdevet, D., Borisch, B., and Hoessli, D. C. (2011) Membrane-associated signaling in human B-lymphoma lines. *Experimental Cell Research* 317, 151-162.
56. Rifai, N., Gillette, M. A., and Carr, S. A. (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat Biotech* 24, 971-983.

57. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N., and Mann, M. (2014) Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat Meth* 11, 319-324.

Figure legends

FIG. 1. **Proteomic workflow and coverage of 20 FFPE DLBCL patient samples.** *A,* Two slices of macro-dissected patient FFPE tissues were processed according to the FASP-FFPE protocol. The super-SILAC approach was employed for quantitative measurements using a quadrupole Orbitrap mass spectrometer (Q Exactive). Quantification was based on SILAC ratios combined with label free quantifications in cases where no SILAC pairs where detected. The data was analyzed using the MaxQuant software resulting in the identification of more than 9000 proteins. *B,* Percentage coverage of signaling pathways and cellular processes in the quantified patients proteome.

FIG. 2. **Quantified FFPE DLBCL patient proteomes.** *A*, Pearson's correlation coefficient (r) of two representative patient samples (TRR003 and TRR013). *B*, Dynamic range of patients' proteomes highlighting KEGG annotated proteins to be involved in 'pathways in cancer'.

FIG. 3. **DLBCL patient samples' proteome versus the DLBCL cell lines' proteome.** *A*, Overlap in the protein groups between the patient proteomes and the cell line proteomes. *B*, The distribution of proteins exclusively quantified in the patient samples (red) in comparison to the total distribution (blue). *C*, Principal component analysis of patient samples using the 55-protein segregating signature derived from cell lines.

FIG. 4. **Principal component analysis of patient samples using their global protein expression profiles.** *A*, The global proteomes of 20 DLBCL patient samples segregated diagonally into ABC-DLBCL (13 samples) and GCB-DLBCL subtypes (7 samples) based on component 1 which

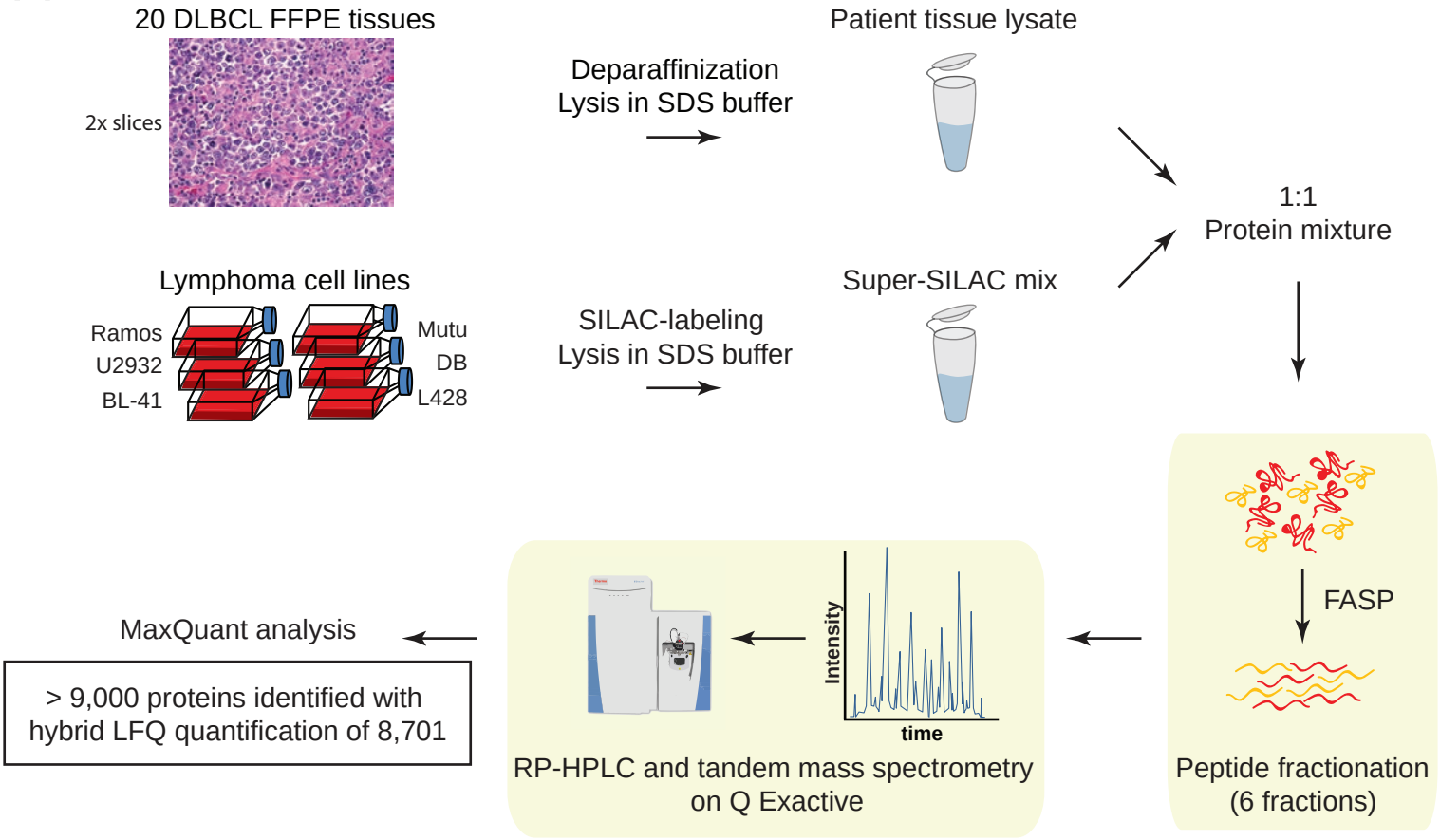
accounts for 11.9% of variability versus component 4 which accounts for 7.4% of the variability. *B*, Loadings of **A** highlighted in red reveal the main proteins driving the COO diagonal segregation. *C,* Cancer module 47 which is composed of extracellular proteins and collagens is highly enriched in component 1.

FIG. 5. **ABC-DLBCL versus GCB-DLBCL.** *A,* Pearson correlation of ABC-DLBCL versus GCB-DLBCL after taking median expression values of protein groups across patients in each subtype. *B,* 2D annotation enrichment of ABC-DLBCL against GCB-DLBCL using cancer modules annotated in GSEA.

FIG. 6. **Support vector machine analysis for optimal feature selection.** *A*, Support vector machine feature selection employing p-values of standard ANOVA tests resulted in a set of 4features with 1.4 percent error. B, Unsupervised hierarchical clustering of top 343 protein candidates or features determined by support vector machine analysis. C, Unsupervised hierarchical clustering of extracellular matrix, plasma membrane and nuclear proteins in 343 top protein candidates.

Figure 1

A


20 DLBCL FFPE tissues
2x slices
Lymphoma cell lines
Ramos
U2932
BL-41
Mutu
DB
L428
Deparaffinization
Lysis in SDS buffer
SILAC-labeling
Lysis in SDS buffer
Patient tissue lysate
Super-SILAC mix
1:1
Protein mixture
FASP
Peptide fractionation
(6 fractions)
Intensity
time
RP-HPLC and tandem mass spectrometry
on Q Exactive
MaxQuant analysis
> 9,000 proteins identified with
hybrid LFQ quantification of 8,701

B

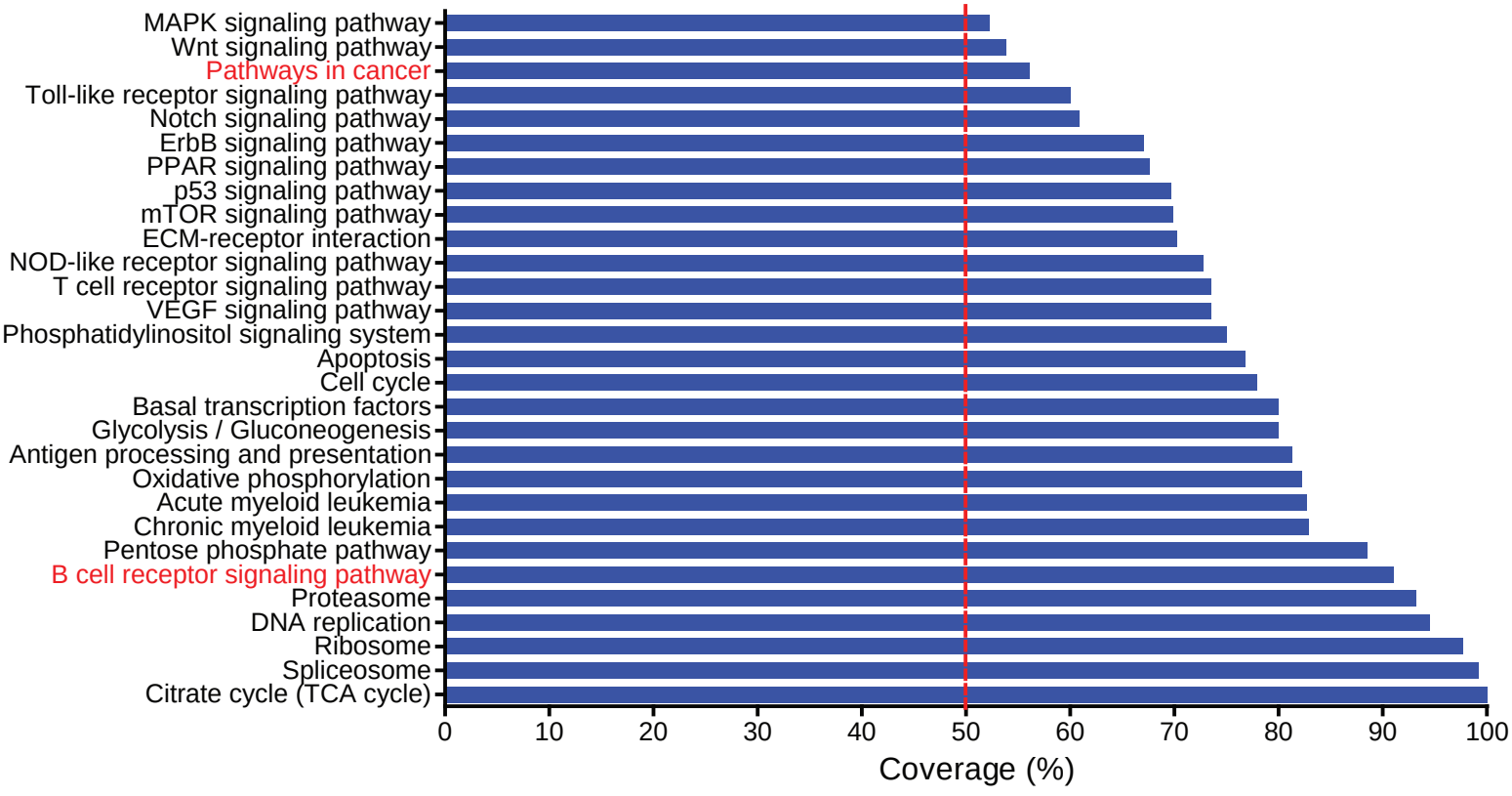

MAPK signaling pathway
Wnt signaling pathway
Pathways in cancer
Toll-like receptor signaling pathway
Notch signaling pathway
ErbB signaling pathway
PPAR signaling pathway
p53 signaling pathway
mTOR signaling pathway
ECM-receptor interaction
NOD-like receptor signaling pathway
T cell receptor signaling pathway
VEGF signaling pathway
Phosphatidylinositol signaling system
Apoptosis
Cell cycle
Basal transcription factors
Glycolysis / Gluconeogenesis
Antigen processing and presentation
Oxidative phosphorylation
Acute myeloid leukemia
Chronic myeloid leukemia
Pentose phosphate pathway
B cell receptor signaling pathway
Proteasome
DNA replication
Ribosome
Spliceosome
Citrate cycle (TCA cycle)
0
10
20
30
40
50
60
70
80
90
100
Coverage (%)

Figure 2

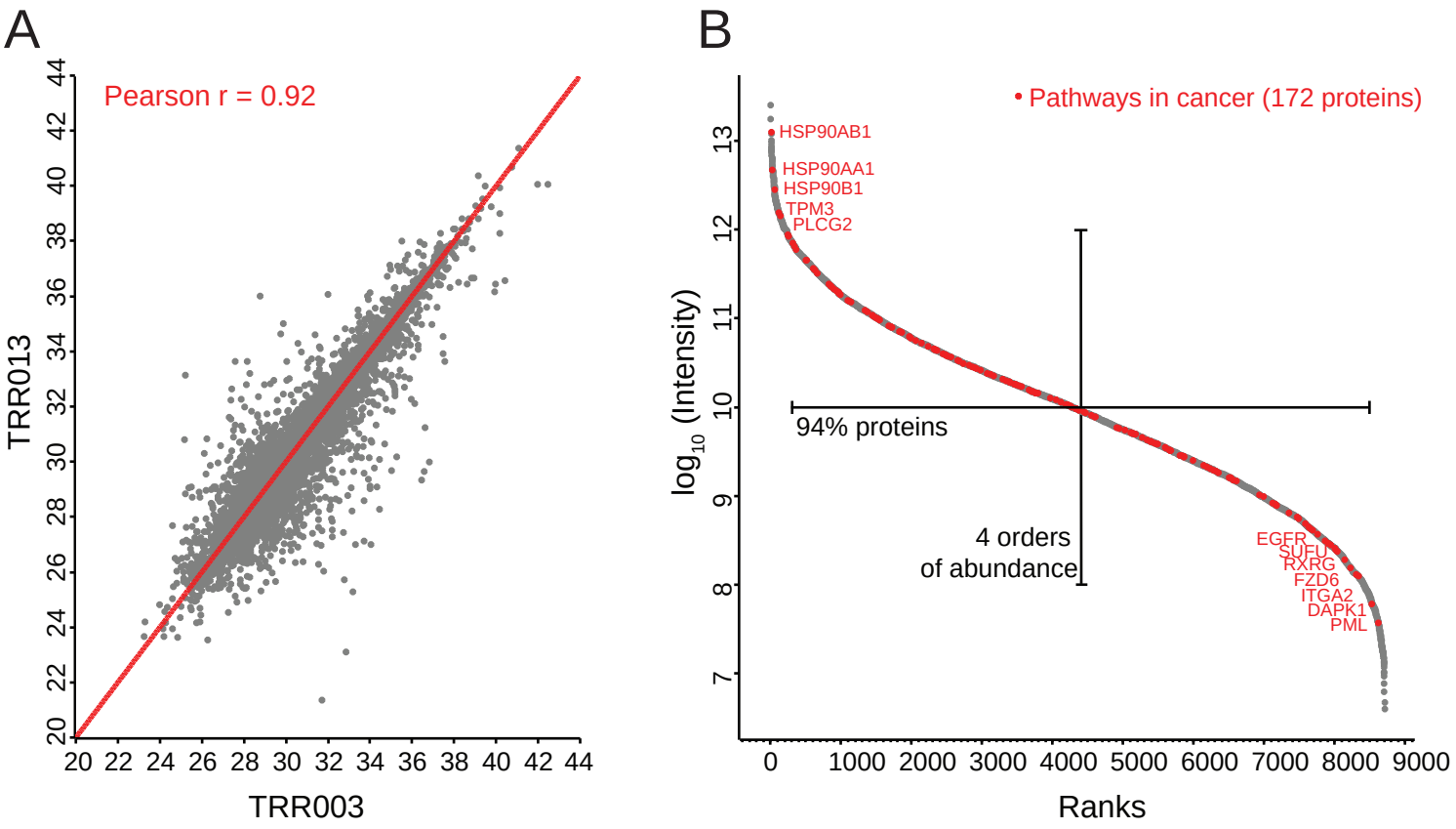
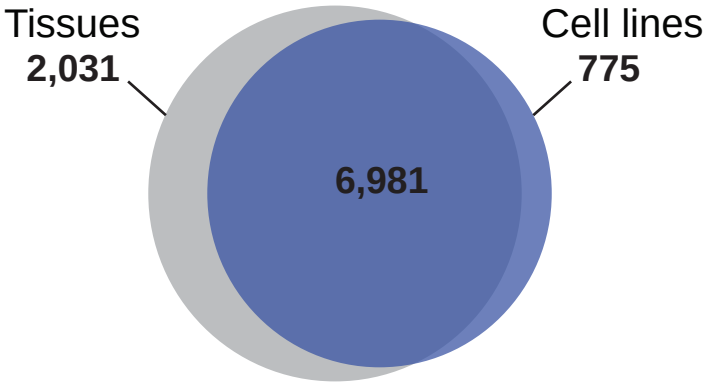
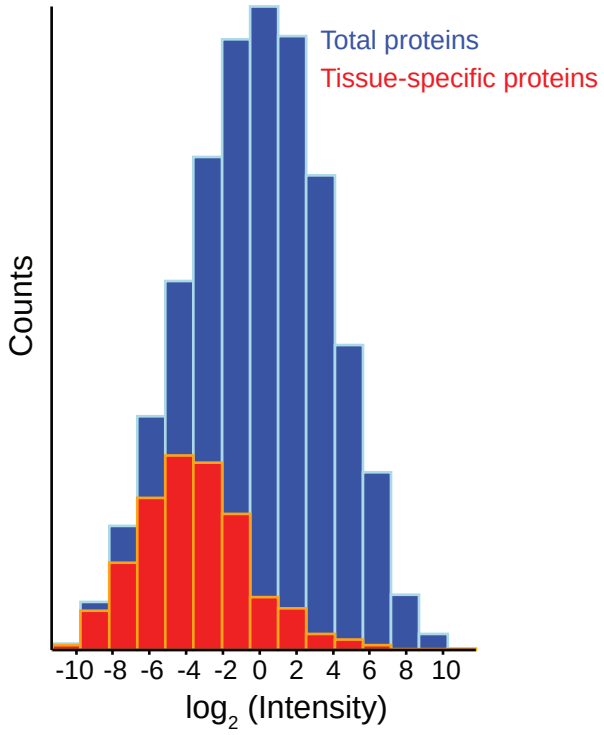

A
Pearson r = 0.92
TRR013
TRR003
20 22 24 26 28 30 32 34 36 38 40 42 44
B
• Pathways in cancer (172 proteins)
HSP90AB1
HSP90AA1
HSP90B1
TPM3
PLCG2
94% proteins
4 orders
of abundance
EGFR
SUFU
RXRG
FZD6
ITGA2
DAPK1
PML
log10 (Intensity)
7 8 9 10 11 12 13
0 1000 2000 3000 4000 5000 6000 7000 8000 9000
Ranks

Figure 3

A


Tissues
2,031
Cell lines
775
6,981

B


Total proteins
Tissue-specific proteins
Counts
-10 -8 -6 -4 -2 0 2 4 6 8 10
log2 (Intensity)

C

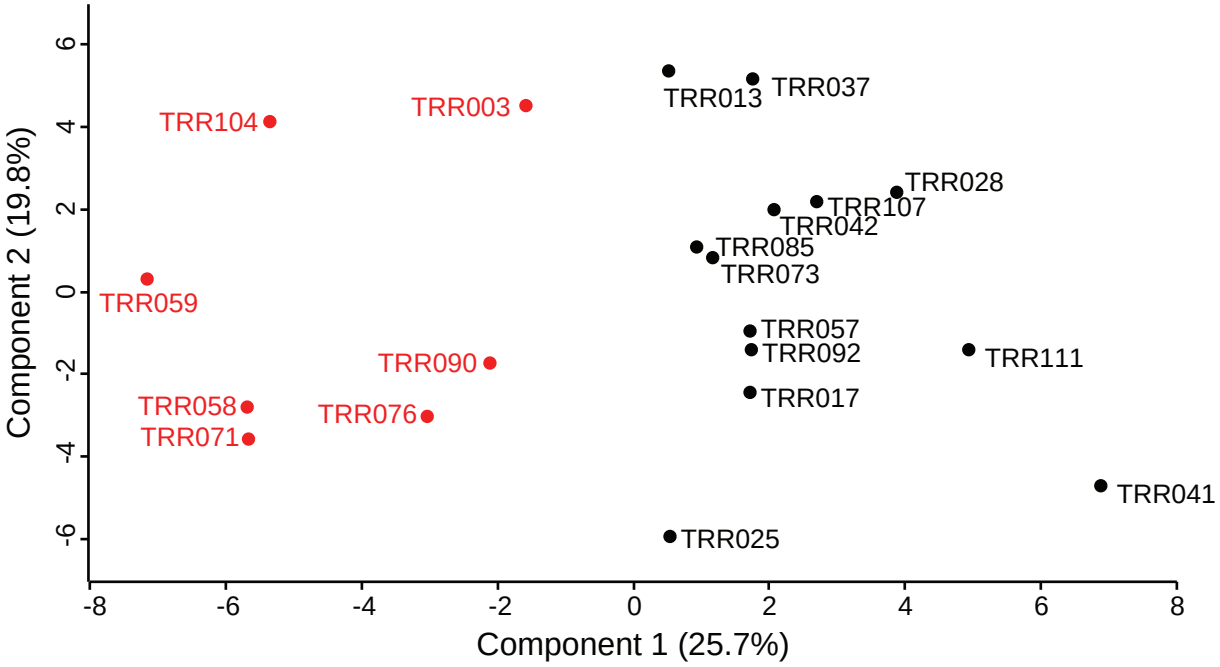
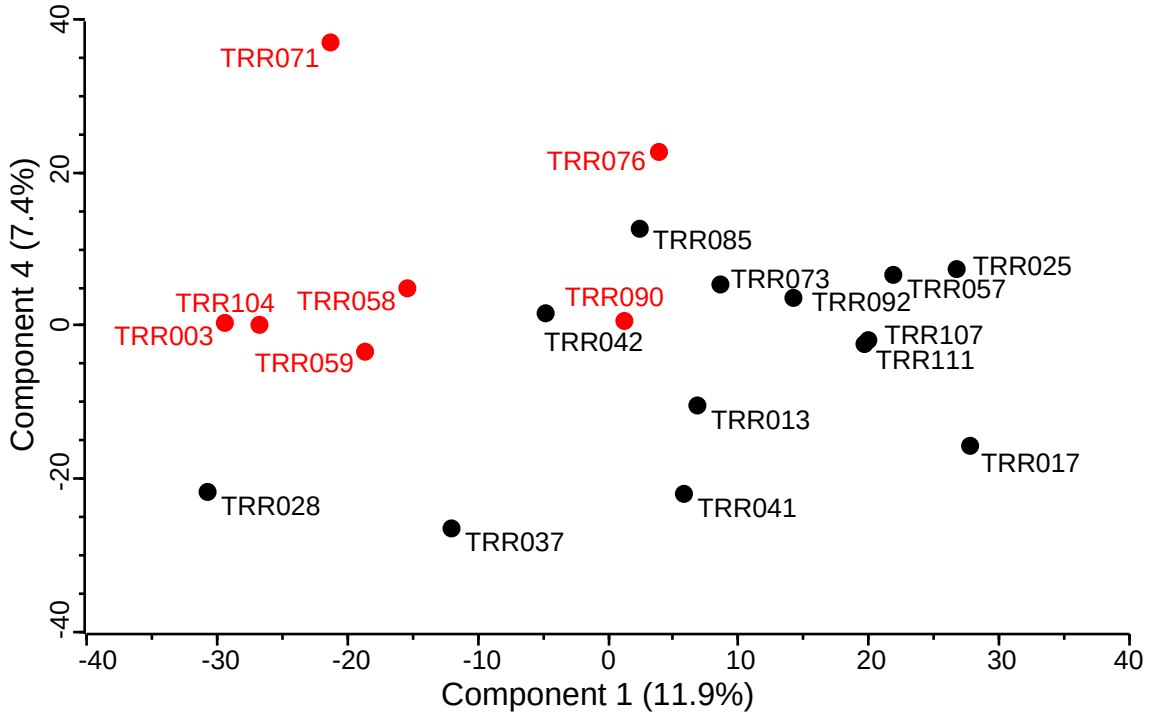
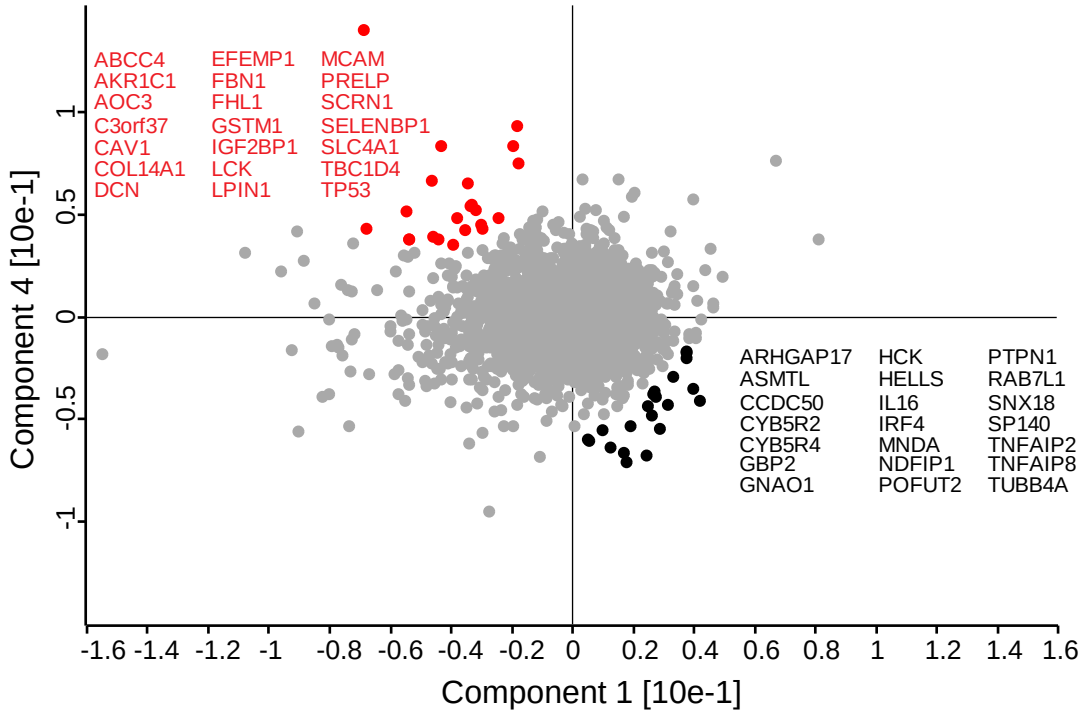

Component 2 (19.8%)
Component 1 (25.7%)
TRR104
TRR003
TRR013
TRR037
TRR028
TRR107
TRR042
TRR085
TRR073
TRR059
TRR057
TRR092
TRR111
TRR090
TRR058
TRR071
TRR076
TRR017
TRR041
TRR025

Figure 4

A


TRR071
TRR076
TRR085
TRR025
TRR073
TRR057
TRR092
TRR104
TRR058
TRR090
TRR003
TRR042
TRR107
TRR059
TRR111
TRR013
TRR017
TRR028
TRR041
TRR037
Component 4 (7.4%)
Component 1 (11.9%)
40
20
0
-20
-40
-40
-30
-20
-10
0
10
20
30
40

B


ABCC4
AKR1C1
AOC3
C3orf37
CAV1
COL14A1
DCN
EFEMP1
FBN1
FHL1
GSTM1
IGF2BP1
LCK
LPIN1
MCAM
PRELP
SCRN1
SELENBP1
SLC4A1
TBC1D4
TP53
ARHGAP17
ASMTL
CCDC50
CYB5R2
CYB5R4
GBP2
GNAO1
HCK
HELLS
IL16
IRF4
MNDA
NDFIP1
POFUT2
PTPN1
RAB7L1
SNX18
SP140
TNFAIP2
TNFAIP8
TUBB4A
Component 4 [10e-1]
Component 1 [10e-1]
1
0.5
0
-0.5
-1
-1.6 -1.4 -1.2 -1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6

C

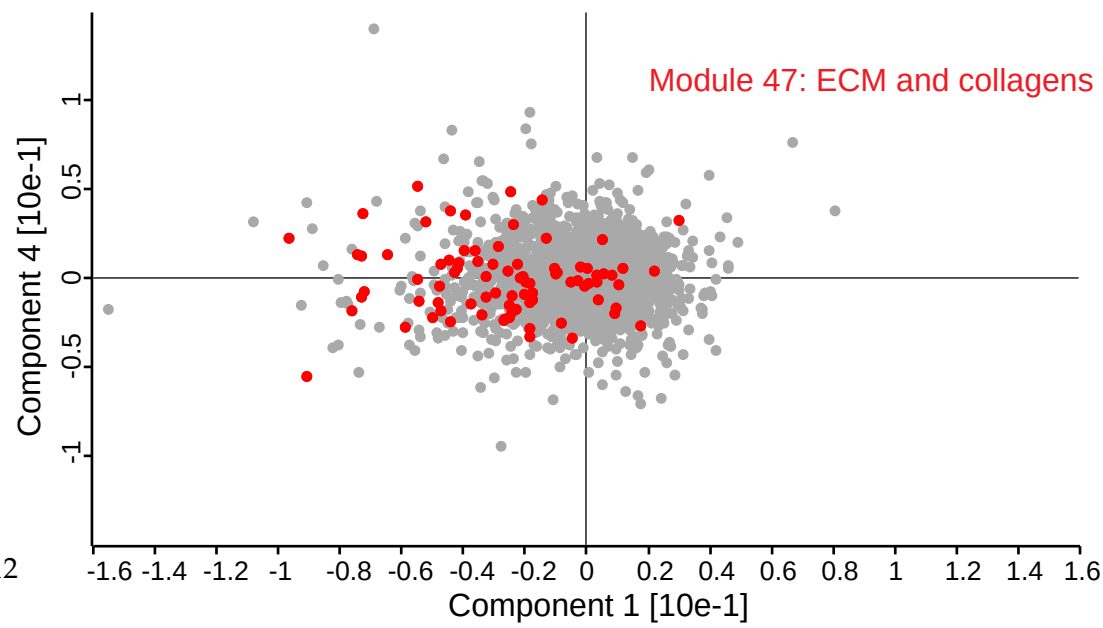
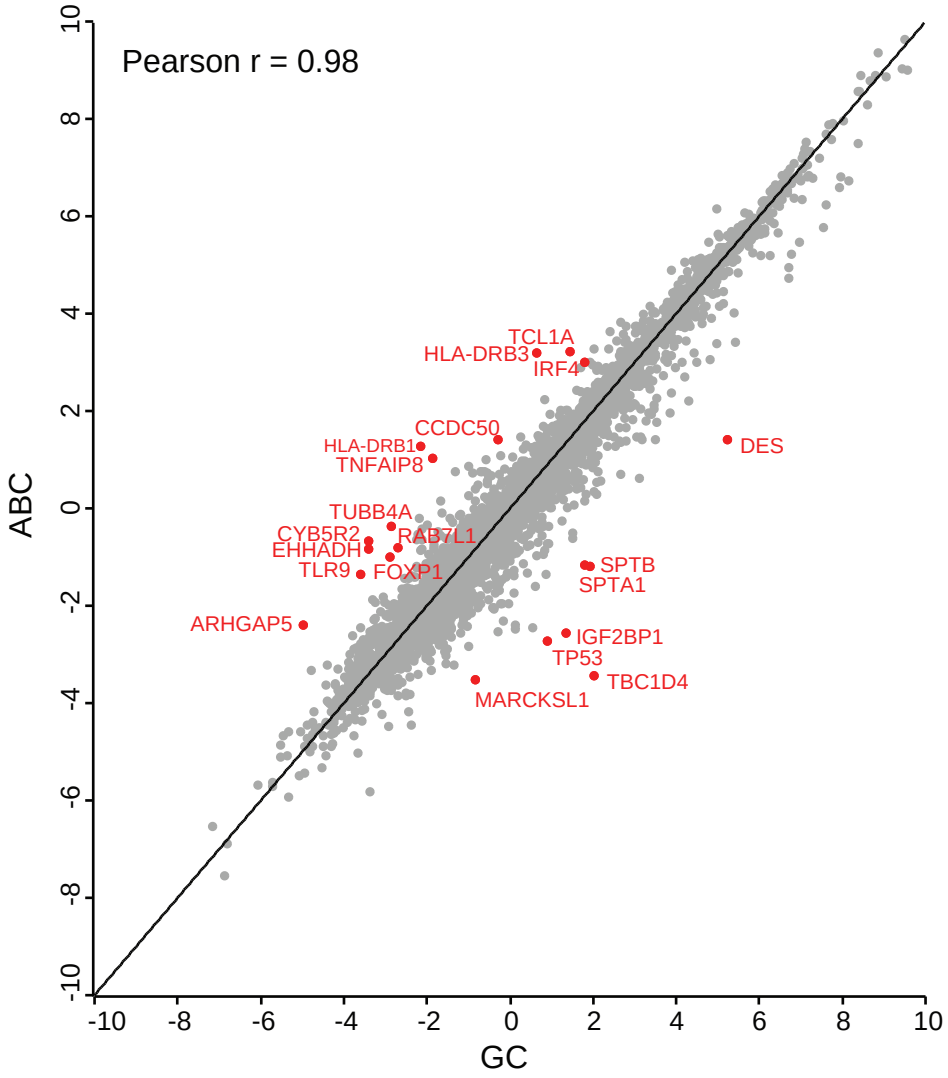

Module 47: ECM and collagens
Component 4 [10e-1]
Component 1 [10e-1]
1
0.5
0
-0.5
-1
-1.6 -1.4 -1.2 -1 -0.8 -0.6 -0.4 -0.2 0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6

Figure 5

A


Pearson r = 0.98
ABC
GC
TCL1A
HLA-DRB3
IRF4
CCDC50
HLA-DRB1
TNFAIP8
DES
TUBB4A
CYB5R2
RAB7L1
EHHADH
TLR9
FOXP1
SPTB
SPTA1
ARHGAP5
IGF2BP1
TP53
TBC1D4
MARCKSL1

B

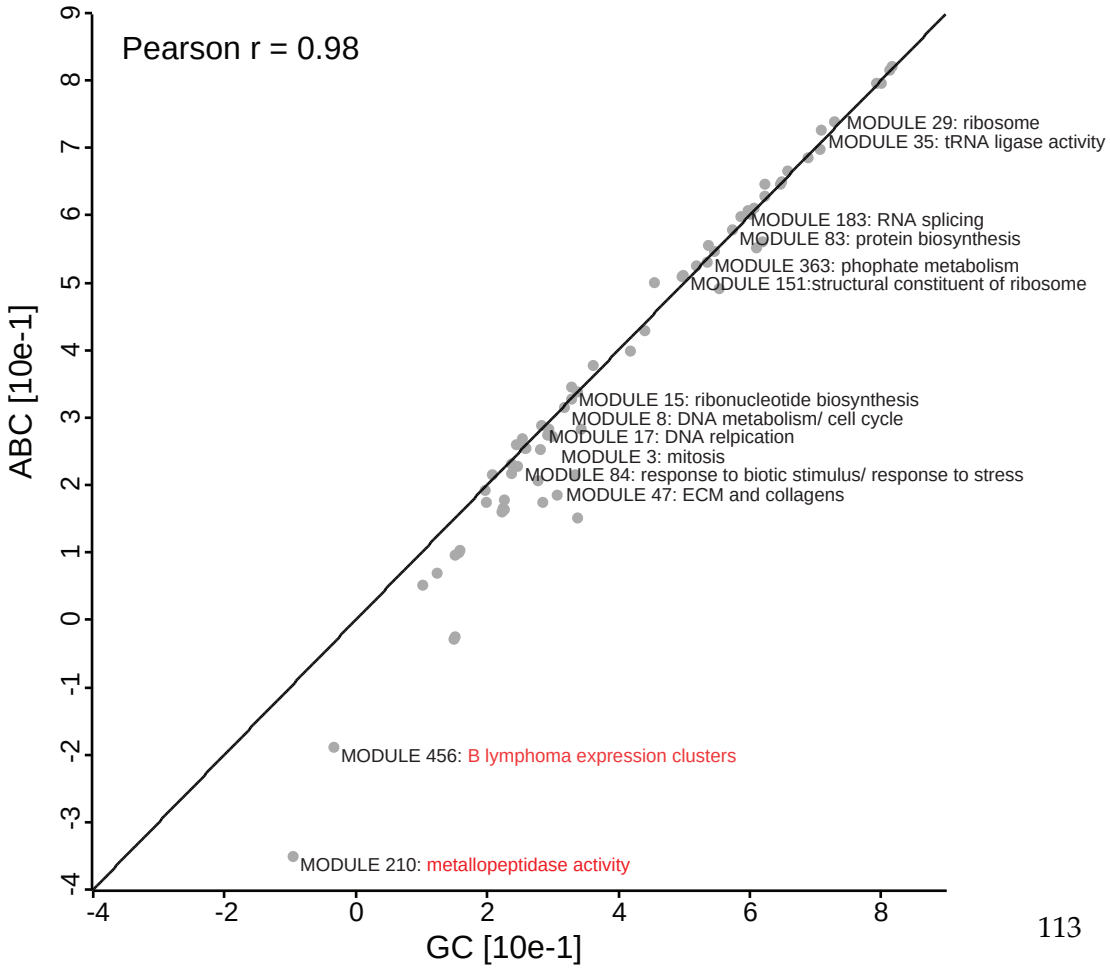
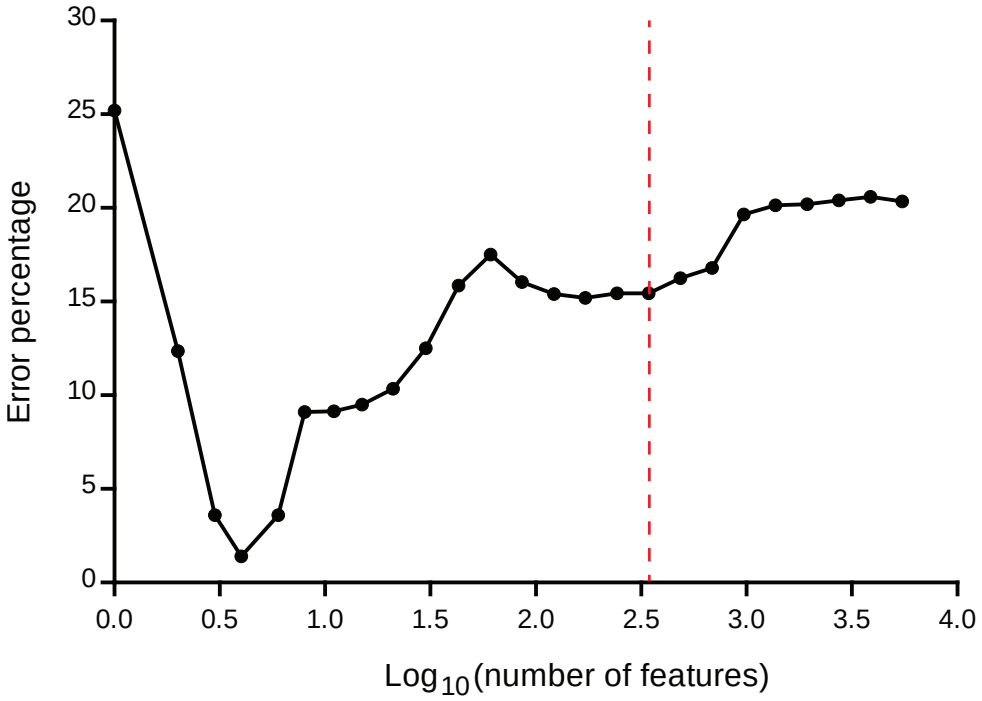
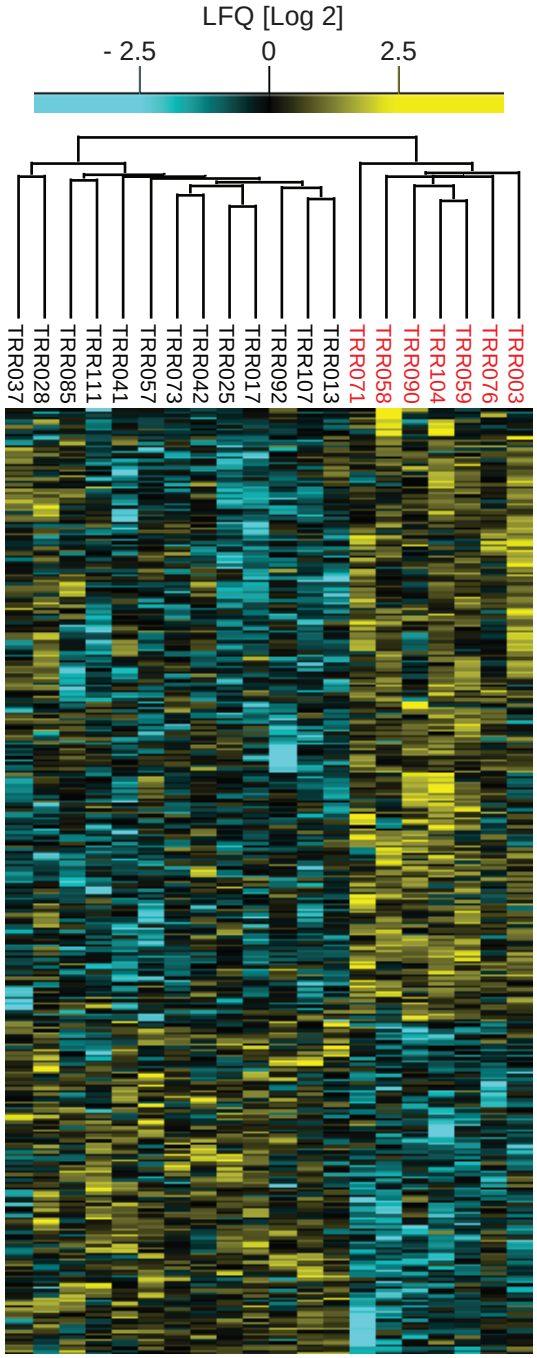

Pearson r = 0.98
ABC [10e-1]
GC [10e-1]
MODULE 29: ribosome
MODULE 35: tRNA ligase activity
MODULE 183: RNA splicing
MODULE 83: protein biosynthesis
MODULE 363: phophate metabolism
MODULE 151:structural constituent of ribosome
MODULE 15: ribonucleotide biosynthesis
MODULE 8: DNA metabolism/ cell cycle
MODULE 17: DNA relpication
MODULE 3: mitosis
MODULE 84: response to biotic stimulus/ response to stress
MODULE 47: ECM and collagens
MODULE 456: B lymphoma expression clusters
MODULE 210: metallopeptidase activity

Figure 6

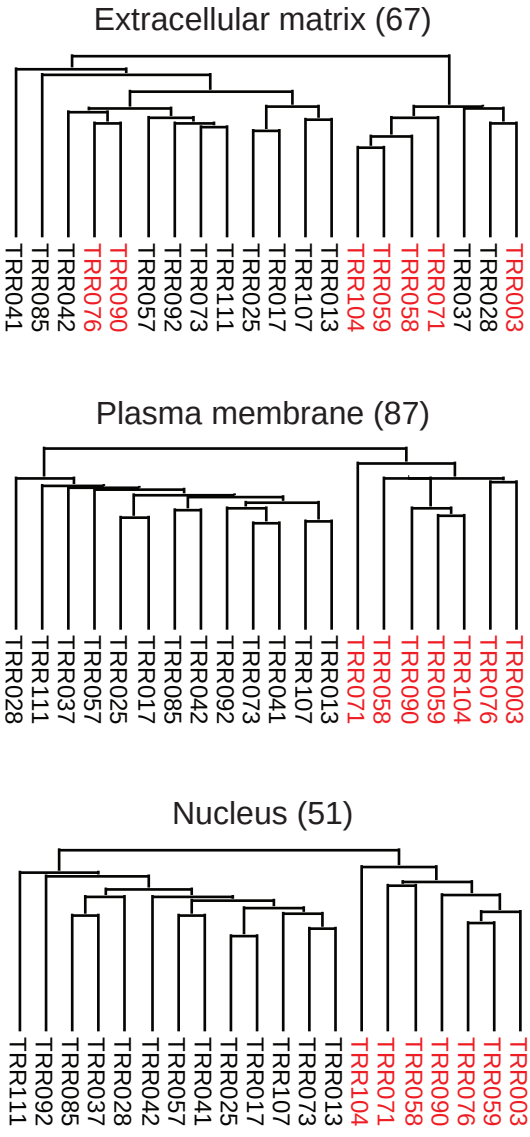
A


Error percentage
30
25
20
15
10
5
0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Log10(number of features)

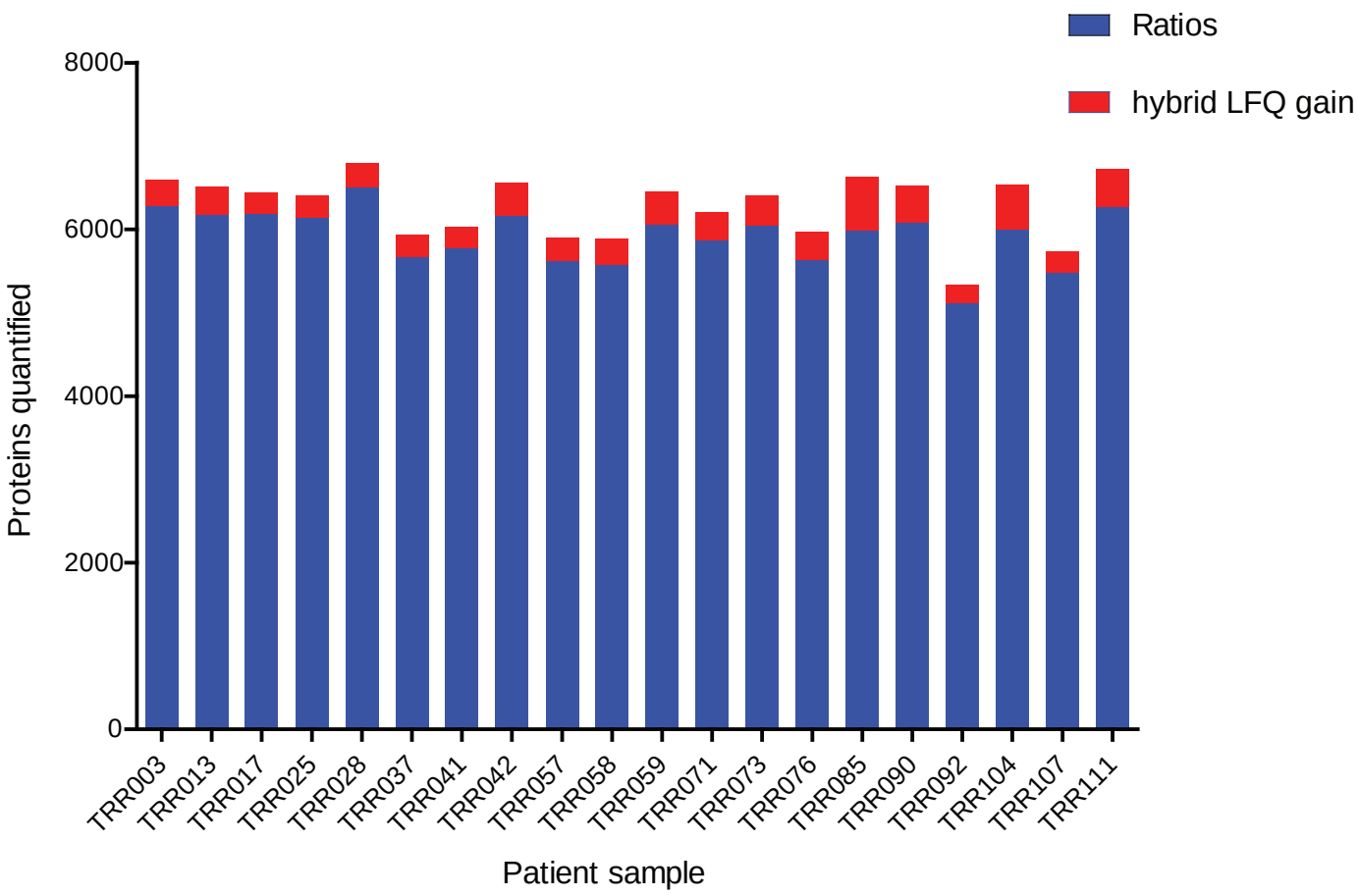
B


LFQ [Log 2]
- 2.5
0
2.5
TRR037
TRR028
TRR085
TRR111
TRR041
TRR057
TRR073
TRR042
TRR025
TRR017
TRR092
TRR107
TRR013
TRR071
TRR058
TRR090
TRR104
TRR059
TRR076
TRR003

C

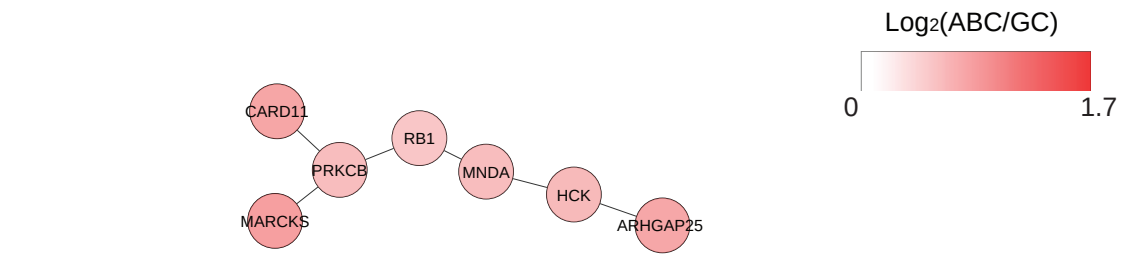

Extracellular matrix (67)
TRR041
TRR085
TRR042
TRR076
TRR090
TRR057
TRR092
TRR073
TRR111
TRR025
TRR017
TRR107
TRR013
TRR104
TRR059
TRR058
TRR071
TRR037
TRR028
TRR003
Plasma membrane (87)
TRR028
TRR111
TRR037
TRR057
TRR025
TRR017
TRR085
TRR042
TRR092
TRR073
TRR041
TRR107
TRR013
TRR071
TRR058
TRR090
TRR059
TRR104
TRR076
TRR003
Nucleus (51)
TRR111
TRR092
TRR085
TRR037
TRR028
TRR042
TRR057
TRR041
TRR025
TRR017
TRR107
TRR073
TRR013
TRR104
TRR071
TRR058
TRR090
TRR076
TRR059
TRR003

Supplementary figure 1

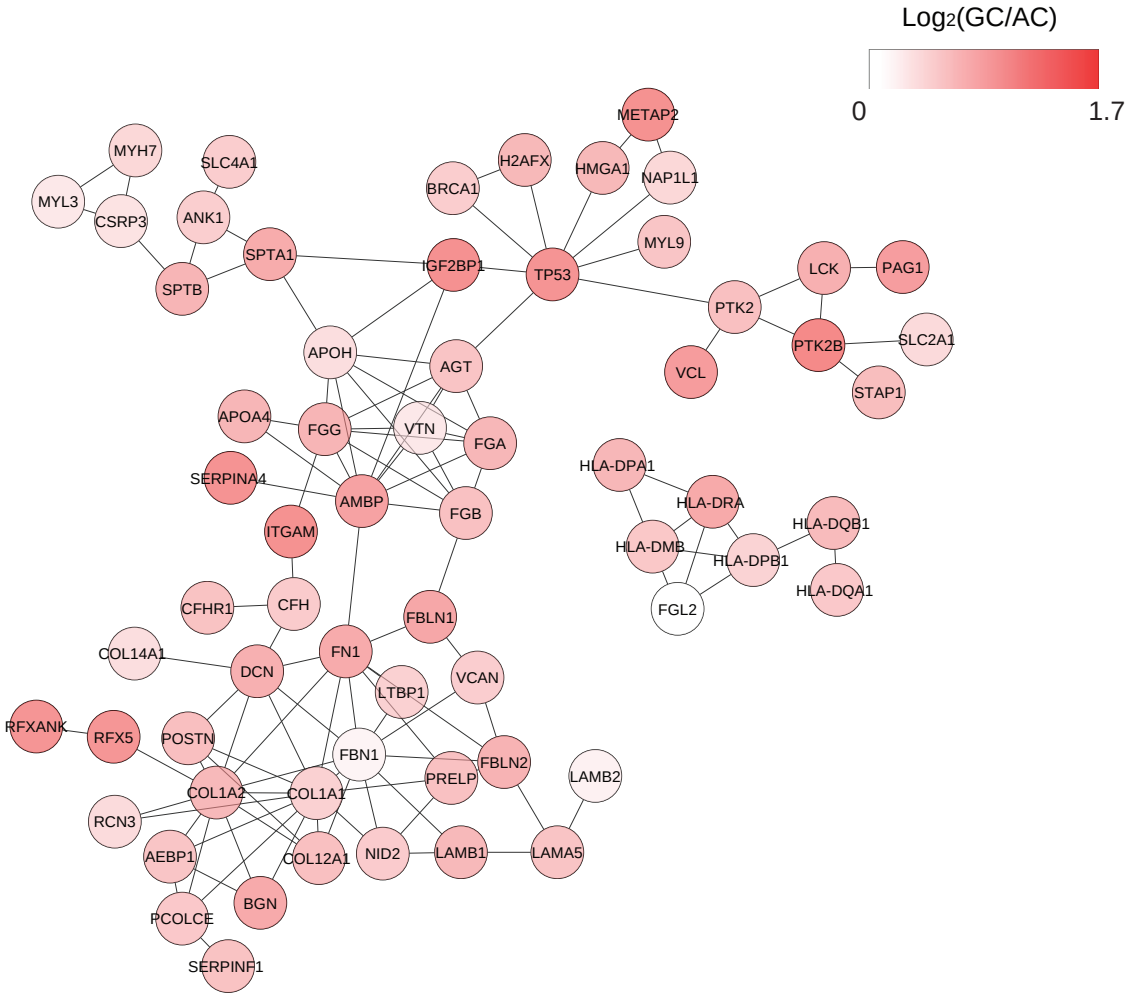


Supplementary figure 2

A



B



Supplementary table I

	cytokines and growth factors	transcription factors	cell differentiation markers	protein kinases	translocated cancer genes	oncogenes	tumor suppressors
tumor suppressors	0	3	0	0	0	0	5
oncogenes	0	4	0	2	10	12	
translocated cancer genes	0	3	0	2	10		
protein kinases	0	0	0	14			
cell differentiation markers	0	0	11				
transcription factors	0	33					
cytokines and growth factors	3						

3. OUTLOOK

One of the most celebrated success stories of personalized medicine is the treatment of Philadelphia–positive chronic myelogenous leukemia (CML) patients with Imatinib. The success rate of Imatinib in the treatment of CML was extraordinary. In one of the first clinical studies, complete hematologic response (white blood cell count returns to within normal range) was observed in 53 of 54 CML patients within the first four weeks of therapy [22]. Equally impressive results were obtained in a five-year follow-up study. After 60 months of treatment, the estimated overall survival of patients who received Imatinib as initial therapy was 89% [155].

The background to this success was the discovery of the atypical small chromosome in CML cells which was later named the "Philadelphia chromosome", after the city in which it was discovered. Importantly, nearly all leukemic cells from patients with CML carried the Philadelphia chromosome [156]. Following the discovery, decades of research revealed the molecular biology of this aberrant chromosome. In 1973, it was discovered that it was a translocated chromosome [157]. New staining techniques, at that time, revealed what may be a translocation between the long arms of chromosomes 22 and 9. It took another 10 years until the actual genes involved in the Philadelphia chromosome were determined. In 1983, the human c-*abl* oncogene was located to the translocated region of chromosome 9 that becomes part of the Philadelphia chromosome [158]. A year later, the same group was able to identify the breakpoint cluster region (*bcr*), the region in which all the chromosome 22 breakpoints seemed to occur in CML patients [159]. It was not until the 1990 that the function of BCR-ABL was identified. The BCR-ABL fusion results in the production of an abnormal tyrosine kinase protein that is not properly regulated [160]. This kinase is highly active causing the cells to proliferate at an abnormally high rate resulting in the accumulation of immature white blood cells. Eventually, the knowledge resulting from 30 years of basic scientific research had set the stage for the development of a drug against CML. In collaboration with industry, compounds, which might fit the ATP-binding site of BCR-ABL based on computer

models, were screened. This resulted in the discovery of Imatinib (Gleevec), which showed the remarkable results in vitro [161] and later on in clinical trials that were described above.

Imatinib has even been called a miracle drug and it has raised hope for more similarly effective drugs based on personalized medicine approaches. However, it has now become clear that Imatinib was an exceptional case. CML is caused by a single aberrant protein related to a consistent chromosomal translocation. In fact, the *bcr–abl* oncogene is present in 95% of patients with CML [161]. It has been implicated as the cause of this disease and therefore, all efforts could be focused on targeting this oncogenic driver. Unfortunately, this does not seem the case for most other cancers. Nevertheless, the story remains an excellent example of how basic knowledge of oncogenic aberrations can be translated into successful targeted therapeutics.

In contrast to CML, analysis of primary tumor samples using whole-genome and exome sequencing has revealed tremendous molecular complexity and genetic heterogeneity for DLBCL, which is the focus of this thesis [121, 162]. Substantial variation of mutated genes was observed from patient to patient and also between published studies [162]. In the time of molecular definition of diseases, a major breakthrough would be the identification of combinations of novel agents to target the oncogenic drivers of each subset of disease [121]. This would likely be based on a principled understanding of how to manage the subgroups differently. Hence, translational molecular investigations will be essential to achieve the goal of precision medicine and expand the number of curable DLBCL patients.

With the increased speed and accessibility of sequencing technologies, massive resources have been devoted to genome sequencing of DLBCL and other human tumors. One of the largest initiatives is The Cancer Genome Atlas (TCGA), aimed at sequencing the genomes of common human cancers. In the last couple of years, these studies have produced massive datasets for the cancer biology community; however, the amount of new biology revealed by such studies has been disappointing so far [163]. The majority of

mutated genes and altered pathways were already known [162]. In glioblastoma, for instance, eight genes were significantly amplified or mutated. These genes included the epidermal growth factor receptor (EGFR), phosphatidylinositol 3-kinase (PI3K), and the cell cycle regulator retinoblastoma (Rb). In fact, all eight genes were previously known to play an important role in cancer [164]. Furthermore, this study showed that mutations in the tumor-suppressor p53 were particularly common [164], a finding which is hardly surprising. Extensive genomic sequencing of ovarian cancer [165] and colorectal cancer [151] obtained nearly the same gene and pathway information. Despite revealing little new biology regarding cancer treatment, sequencing data provides valuable information on real human cancers confirming studies that relied on cancer models. In addition, an unbiased catalog of the different types of mutations can reveal important insights into tumor evolution [166, 167] as well as combinations of mutations that occur in specific tumor types. Furthermore, recent sequencing studies have revealed genetic heterogeneity between cells of the same tumor, further complicating attempts to translate genetic data into therapy [168]. One of the key messages of these studies is that alterations in signaling pathways are central to the molecular biology of cancers [163].

A cancer cell has many hundreds of pathways and networks at its disposal to exploit in various ways and to develop resistance to targeted or untargeted therapies. They can divert and diversify when faced with a roadblock i.e. a targeted drug [169]. The community now recognizes even more that cancer is not a disease of single mutations or genes and that it instead involves the dysregulation of multiple signaling pathways and networks which correspond to cellular processes such as proliferation, apoptosis, differentiation and migration [153]. Biological processes compromised by the tumor to ensure its survival include sustaining proliferative signaling, evading growth suppressors, resisting cell death, achieving replicative immortality, inducing angiogenesis, and activating invasion and metastasis. Two additional hallmarks of potential generality are reprogramming of energy metabolism and evading immune destruction [154]. A recently developed network-based stratification (NBS) method,

already discussed above, allowed the identification of clinically meaningful patient subsets integrating somatic mutations profiles with gene networks. Patients with mutations in similar network regions are clustered together resulting in stratification of ovarian, uterine and lung cancer cohorts from TCGA into clinically informative subtypes**.** This shows that although tumors appear different at the level of genes, groupings appear at the level of impacted biological networks and systems [149].

Finally, it is clearly the signaling proteins that are responsible for a tumor's phenotype. Robust and reliable tools for analyzing the status of tumor signaling networks and the state of tumor proteome are therefore needed. Only MS-based proteomics provides a direct way to study compromised signaling pathways and cancer hallmark processes. This thesis has provided early applications of modern MS-based technology to characterize tumors at the cell line and patient levels. While the three projects presented here provide promising proof-of-principle, larger patient cohorts will be required to validate MS-based proteomics as an indispensable tool to enhance the molecular understanding of tumors. However, it is already clear that it is the only technology with the potential to characterize tumors at the level of PTMs. With even more technological advancements in the horizon, such a reality does not seem to be so far-fetched.

ABBREVIATIONS

ABC-DLBCL	Activated B-cell–like diffuse large B-cell lymphoma
AQUA	Absolute quantitation
BCR	B-cell receptor
CID	Collision induced dissociation
CLL	Chronic lymphocytic leukemia
COO	Cell-of-origin
DLBCL	Diffuse large B-cell lymphoma
ELISA	Enzyme linked immune assays
ESI	Electrospray ionization
FASP	Filter-aided sample preparation
FDR	False discovery rate
FFPE	Formalin-fixed paraffin-embedded
GCB-DLBCL	Germinal-center B-cell–like DLBCL
GEP	Gene expression profiling
HCD	Higher energy collision dissociation
HPLC	High performance liquid chromatography
LC	Liquid chromatography
LTQ	Linear trap quadrupole
m/z	Mass-to-charge ratio
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
NBS	Network-based stratification
PCA	Principal component analysis
PTM	Post-translational modification
RF	Radio-frequency
SAX	Strong anion exchange chromatography
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SILAC	Stable isotope labeling with amino acids in cell culture
SVM	Support vector machine

REFERENCES

1. Consortium, T.U., *The Universal Protein Resource (UniProt) 2009.* Nucleic Acids Research, 2009. **37**(suppl 1): p. D169-D174.
2. Zhou, G., et al., *2D Differential In-gel Electrophoresis for the Identification of Esophageal Scans Cell Cancer-specific Protein Markers.* Molecular & Cellular Proteomics, 2002. **1**(2): p. 117-123.
3. O'Farrell, P.H., *High resolution two-dimensional electrophoresis of proteins.* Journal of Biological Chemistry, 1975. **250**(10): p. 4007-4021.
4. Zangar, R.C., S.M. Varnum, and N. Bollinger, *Studying Cellular Processes and Detecting Disease with Protein Microarrays.* Drug Metabolism Reviews, 2005. **37**(3): p. 473-487.
5. Janzi, M., et al., *Serum Microarrays for Large Scale Screening of Protein Levels.* Molecular & Cellular Proteomics, 2005. **4**(12): p. 1942-1947.
6. Uhlén, M., et al., *A Human Protein Atlas for Normal and Cancer Tissues Based on Antibody Proteomics.* Molecular & Cellular Proteomics, 2005. **4**(12): p. 1920-1932.
7. Aebersold, R. and M. Mann, *Mass spectrometry-based proteomics.* Nature, 2003. **422**(6928): p. 198-207.
8. Adkins, J.N., et al., *Toward a Human Blood Serum Proteome: Analysis By Multidimensional Separation Coupled With Mass Spectrometry.* Molecular & Cellular Proteomics, 2002. **1**(12): p. 947-955.
9. Jacobs, J.M., et al., *Utilizing Human Blood Plasma for Proteomic Biomarker Discovery†.* Journal of Proteome Research, 2005. **4**(4): p. 1073-1085.
10. Veenstra, T.D., et al., *Biomarkers: Mining the Biofluid Proteome.* Molecular & Cellular Proteomics, 2005. **4**(4): p. 409-418.
11. Bendall, S.C., et al., *Single-Cell Mass Cytometry of Differential Immune and Drug Responses Across a Human Hematopoietic Continuum.* Science, 2011. **332**(6030): p. 687-696.
12. Saffert, R.T., et al., *Comparison of Bruker Biotyper Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometer to BD Phoenix Automated Microbiology System for Identification of Gram-Negative Bacilli.* Journal of Clinical Microbiology, 2011. **49**(3): p. 887-892.
13. Bizzini, A., et al., *Performance of Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry for Identification of Bacterial Strains Routinely Isolated in a Clinical Microbiology Laboratory.* Journal of Clinical Microbiology, 2010. **48**(5): p. 1549-1554.
14. Olsen, J.V. and M. Mann, *Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry.* Molecular & Cellular Proteomics, 2013. **12**(12): p. 3444-3452.
15. Huang, P.H., A.M. Xu, and F.M. White, *Oncogenic EGFR Signaling Networks in Glioma.* Sci. Signal., 2009. **2**(87): p. re6-.
16. Iwai, L.K., et al., *Quantitative Phosphoproteomic Analysis of T Cell Receptor Signaling in Diabetes Prone and Resistant Mice.* Journal of Proteome Research, 2010. **9**(6): p. 3135-3145.
17. Baker, E., et al., *Mass spectrometry for translational proteomics: progress and clinical implications.* Genome Medicine, 2012. **4**(8): p. 63.
18. Rifai, N., M.A. Gillette, and S.A. Carr, *Protein biomarker discovery and validation: the long and uncertain path to clinical utility.* Nat Biotech, 2006. **24**(8): p. 971-983.

19. Danesh, J., et al., *C-Reactive Protein and Other Circulating Markers of Inflammation in the Prediction of Coronary Heart Disease.* New England Journal of Medicine, 2004. **350**(14): p. 1387-1397.
20. Antman, E.M., et al., *Cardiac-Specific Troponin I Levels to Predict the Risk of Mortality in Patients with Acute Coronary Syndromes.* New England Journal of Medicine, 1996. **335**(18): p. 1342-1349.
21. Catalona, W.J., et al., *Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: A prospective multicenter clinical trial.* JAMA, 1998. **279**(19): p. 1542-1547.
22. Druker, B.J., et al., *Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia.* New England Journal of Medicine, 2001. **344**(14): p. 1031-1037.
23. Qian, W.-J., et al., *Advances and Challenges in Liquid Chromatography-Mass Spectrometry-based Proteomics Profiling for Clinical Applications.* Molecular & Cellular Proteomics, 2006. **5**(10): p. 1727-1744.
24. Anderson, N.L. and N.G. Anderson, *The Human Plasma Proteome: History, Character, and Diagnostic Prospects.* Molecular & Cellular Proteomics, 2002. **1**(11): p. 845-867.
25. Hu, Y., et al., *Comparative Proteomic Analysis of Intra- and Interindividual Variation in Human Cerebrospinal Fluid.* Molecular & Cellular Proteomics, 2005. **4**(12): p. 2000-2009.
26. Wattiez, R. and P. Falmagne, *Proteomics of bronchoalveolar lavage fluid.* Journal of Chromatography B, 2005. **815**(1–2): p. 169-178.
27. Xie, H., et al., *A Catalogue of Human Saliva Proteins Identified by Free Flow Electrophoresis-based Peptide Separation and Tandem Mass Spectrometry.* Molecular & Cellular Proteomics, 2005. **4**(11): p. 1826-1830.
28. Nagaraj, N. and M. Mann, *Quantitative Analysis of the Intra- and Inter-Individual Variability of the Normal Urinary Proteome.* Journal of Proteome Research, 2010. **10**(2): p. 637-645.
29. Lee, H.-J., et al., *Biomarker discovery from the plasma proteome using multidimensional fractionation proteomics.* Current Opinion in Chemical Biology, 2006. **10**(1): p. 42-49.
30. Granger, J., et al., *Albumin depletion of human plasma also removes low abundance proteins including the cytokines.* PROTEOMICS, 2005. **5**(18): p. 4713-4718.
31. Sedlaczek, P., et al., *Comparative analysis of CA125, tissue polypeptide specific antigen, and soluble interleukin-2 receptor α levels in sera, cyst, and ascitic fluids from patients with ovarian carcinoma.* Cancer, 2002. **95**(9): p. 1886-1893.
32. Mor, G., et al., *Serum protein markers for early detection of ovarian cancer.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(21): p. 7677-7682.
33. Sabatine, M.S., et al., *Multimarker Approach to Risk Stratification in Non-ST Elevation Acute Coronary Syndromes: Simultaneous Assessment of Troponin I, C-Reactive Protein, and B-Type Natriuretic Peptide.* Circulation, 2002. **105**(15): p. 1760-1763.
34. Nagaraj, N., et al., *Deep proteome and transcriptome mapping of a human cancer cell line.* Mol Syst Biol, 2011. **7**.
35. Nagaraj, N., et al., *System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-shot Ultra HPLC Runs on a Bench Top Orbitrap.* Molecular & Cellular Proteomics, 2012. **11**(3).
36. Mann, M., et al., *The Coming Age of Complete, Accurate, and Ubiquitous Proteomes.* Molecular cell, 2013. **49**(4): p. 583-590.

37. Fenn, J., et al., *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64-71.
38. Karas, M. and F. Hillenkamp, *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.* Analytical Chemistry, 1988. **60**(20): p. 2299-2301.
39. Wilm, M. and M. Mann, *Analytical Properties of the Nanoelectrospray Ion Source.* Analytical Chemistry, 1996. **68**(1): p. 1-8.
40. Mann, M., *Proteomics for biomedicine: a half-completed journey.* EMBO Molecular Medicine, 2012. **4**(2): p. 75-77.
41. Shevchenko, A., et al., *Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels.* Analytical Chemistry, 1996. **68**(5): p. 850-858.
42. Wisniewski, J.R., et al., *Universal sample preparation method for proteome analysis.* Nat Methods, 2009. **6**(5): p. 359-62.
43. Wilm, M., et al., *Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry.* Nature, 1996. **379**(6564): p. 466-469.
44. Gygi, S.P., et al., *Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology.* Proceedings of the National Academy of Sciences, 2000. **97**(17): p. 9390-9395.
45. Rappsilber, J., Y. Ishihama, and M. Mann, *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics.* Anal Chem, 2003. **75**(3): p. 663-70.
46. Wiśniewski, J.R., A. Zougman, and M. Mann, *Combination of FASP and StageTip-Based Fractionation Allows In-Depth Analysis of the Hippocampal Membrane Proteome.* Journal of Proteome Research, 2009. **8**(12): p. 5674-5678.
47. Deeb, S.J., et al., *Super-SILAC Allows Classification of Diffuse Large B-cell Lymphoma Subtypes by Their Protein Expression Profiles.* Molecular & Cellular Proteomics, 2012. **11**(5): p. 77-89.
48. Hein, M.Y., et al., *Chapter 1 - Proteomic Analysis of Cellular Systems*, in *Handbook of Systems Biology*, A.J.M. Walhout, M. Vidal, and J. Dekker, Editors. 2013, Academic Press: San Diego. p. 3-25.
49. Pinkse, M.W.H., et al., *Selective Isolation at the Femtomole Level of Phosphopeptides from Proteolytic Digests Using 2D-NanoLC-ESI-MS/MS and Titanium Oxide Precolumns.* Analytical Chemistry, 2004. **76**(14): p. 3935-3943.
50. Rush, J., et al., *Immunoaffinity profiling of tyrosine phosphorylation in cancer cells.* Nat Biotech, 2005. **23**(1): p. 94-101.
51. Zielinska, D.F., et al., *Precision Mapping of an In Vivo N-Glycoproteome Reveals Rigid Topological and Sequence Constraints.* Cell, 2010. **141**(5): p. 897-907.
52. Olsen, J.V., et al., *Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks.* Cell, 2006. **127**(3): p. 635-648.
53. Huang, P.H., et al., *Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma.* Proceedings of the National Academy of Sciences, 2007. **104**(31): p. 12867-12872.
54. Olsen, J.V., et al., *Quantitative Phosphoproteomics Reveals Widespread Full Phosphorylation Site Occupancy During Mitosis.* Sci. Signal., 2010. **3**(104): p. ra3-.
55. Witze, E.S., et al., *Mapping protein post-translational modifications with mass spectrometry.* Nat Meth, 2007. **4**(10): p. 798-806.
56. Udeshi, N.D., et al., *Refined Preparation and Use of Anti-diglycine Remnant (K-ε-GG) Antibody Enables Routine Quantification of 10,000s of Ubiquitination Sites in Single Proteomics Experiments.* Molecular & Cellular Proteomics, 2013. **12**(3): p. 825-831.

57. Wagner, S.A., et al., *Proteomic Analyses Reveal Divergent Ubiquitylation Site Patterns in Murine Tissues.* Molecular & Cellular Proteomics, 2012. **11**(12): p. 1578-1585.
58. Wu, R., et al., *A large-scale method to measure absolute protein phosphorylation stoichiometries.* Nat Meth, 2011. **8**(8): p. 677-683.
59. Thakur, S.S., et al., *Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation.* Mol Cell Proteomics, 2011. **10**(8): p. M110 003699.
60. Köcher, T., R. Swart, and K. Mechtler, *Ultra-High-Pressure RPLC Hyphenated to an LTQ-Orbitrap Velos Reveals a Linear Relation between Peak Capacity and Number of Identified Peptides.* Analytical Chemistry, 2011. **83**(7): p. 2699-2704.
61. Beck, M., et al., *The quantitative proteome of a human cell line.* Molecular Systems Biology, 2011. **7**(1).
62. Wisniewski, J.R., et al., *Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma.* Mol Syst Biol, 2012. **8**.
63. Ostasiewicz, P., et al., *Proteome, Phosphoproteome, and N-Glycoproteome Are Quantitatively Preserved in Formalin-Fixed Paraffin-Embedded Tissue and Analyzable by High-Resolution Mass Spectrometry.* Journal of Proteome Research, 2010. **9**(7): p. 3688-3700.
64. Andrews, G.L., et al., *Performance Characteristics of a New Hybrid Quadrupole Time-of-Flight Tandem Mass Spectrometer (TripleTOF 5600).* Analytical Chemistry, 2011. **83**(13): p. 5442-5446.
65. Michalski, A., et al., *Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer.* Mol Cell Proteomics, 2011. **10**(9): p. M111 011015.
66. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
67. Schwartz, J.C., M.W. Senko, and J.E.P. Syka, *A two-dimensional quadrupole ion trap mass spectrometer.* Journal of the American Society for Mass Spectrometry, 2002. **13**(6): p. 659-669.
68. Makarov, A., *Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis.* Analytical Chemistry, 2000. **72**(6): p. 1156-1162.
69. Makarov, A., et al., *Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer.* Analytical Chemistry, 2006. **78**(7): p. 2113-2120.
70. Hu, Q., et al., *The Orbitrap: a new mass spectrometer.* Journal of Mass Spectrometry, 2005. **40**(4): p. 430-443.
71. Makarov, A., et al., *Dynamic Range of Mass Accuracy in LTQ Orbitrap Hybrid Mass Spectrometer.* Journal of the American Society for Mass Spectrometry, 2006. **17**(7): p. 977-982.
72. Olsen, J.V., et al., *Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-trap.* Molecular & Cellular Proteomics, 2005. **4**(12): p. 2010-2021.
73. Cox, J., A. Michalski, and M. Mann, *Software Lock Mass by Two-Dimensional Minimization of Peptide Mass Errors.* Journal of the American Society for Mass Spectrometry, 2011. **22**(8): p. 1373-1380.
74. Scigelova, M. and A. Makarov, *Orbitrap Mass Analyzer – Overview and Applications in Proteomics.* PROTEOMICS, 2006. **6**(S2): p. 16-21.
75. Olsen, J.V., et al., *Higher-energy C-trap dissociation for peptide modification analysis.* Nat Meth, 2007. **4**(9): p. 709-712.

76. Olsen, J.V., et al., *A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed.* Molecular & Cellular Proteomics, 2009. **8**(12): p. 2759-2769.
77. Michalski, A., et al., *Ultra High Resolution Linear Ion Trap Orbitrap Mass Spectrometer (Orbitrap Elite) Facilitates Top Down LC MS/MS and Versatile Peptide Fragmentation Modes.* Molecular & Cellular Proteomics, 2012. **11**(3).
78. Geiger, T., J. Cox, and M. Mann, *Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation.* Mol Cell Proteomics, 2010. **9**(10): p. 2252-61.
79. Bateman, K.P., et al., *Quantitative–Qualitative Data Acquisition Using a Benchtop Orbitrap Mass Spectrometer.* Journal of the American Society for Mass Spectrometry, 2009. **20**(8): p. 1441-1450.
80. Bantscheff, M., et al., *Quantitative mass spectrometry in proteomics: a critical review.* Anal Bioanal Chem, 2007. **389**(4): p. 1017-31.
81. Gygi, S.P., et al., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.* Nat Biotech, 1999. **17**(10): p. 994-999.
82. Boersema, P.J., et al., *Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates.* PROTEOMICS, 2008. **8**(22): p. 4624-4632.
83. Thompson, A., et al., *Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS.* Analytical Chemistry, 2003. **75**(8): p. 1895-1904.
84. Ross, P.L., et al., *Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents.* Molecular & Cellular Proteomics, 2004. **3**(12): p. 1154-1169.
85. Ong, S.-E., et al., *Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics.* Molecular & Cellular Proteomics, 2002. **1**(5): p. 376-386.
86. Schwanhäusser, B., et al., *Global analysis of cellular protein translation by pulsed SILAC.* PROTEOMICS, 2009. **9**(1): p. 205-209.
87. Doherty, M.K. and R.J. Beynon, *Protein turnover on the scale of the proteome.* Expert Review of Proteomics, 2006. **3**(1): p. 97-110.
88. Hubner, N.C. and M. Mann, *Extracting gene function from protein–protein interactions using Quantitative BAC InteraCtomics (QUBIC).* Methods, 2011. **53**(4): p. 453-459.
89. Geiger, T., et al., *Super-SILAC mix for quantitative proteomics of human tumor tissue.* Nat Methods, 2010. **7**: p. 383 - 385.
90. Kirkpatrick, D.S., S.A. Gerber, and S.P. Gygi, *The absolute quantification strategy: a general procedure for the quantification of proteins and post-translational modifications.* Methods, 2005. **35**(3): p. 265-273.
91. Hanke, S., et al., *Absolute SILAC for Accurate Quantitation of Proteins in Complex Mixtures Down to the Attomole Level.* Journal of Proteome Research, 2008. **7**(3): p. 1118-1130.
92. Zeiler, M., et al., *A Protein Epitope Signature Tag (PrEST) Library Allows SILAC-based Absolute Quantification and Multiplexed Determination of Protein Copy Numbers in Cell Lines.* Molecular & Cellular Proteomics, 2012. **11**(3).
93. Pratt, J.M., et al., *Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes.* Nat. Protocols, 2006. **1**(2): p. 1029-1043.
94. Neubert, T.A. and P. Tempst, *Super-SILAC for tumors and tissues.* Nat Meth, 2010. **7**(5): p. 361-362.
95. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-342.

96. Schilsky, R.L., *Personalized medicine in oncology: the future is now.* Nat Rev Drug Discov, 2010. **9**(5): p. 363-366.
97. Sawyers, C.L., *The cancer biomarker problem.* Nature, 2008. **452**(7187): p. 548-552.
98. Collins, I. and P. Workman, *New approaches to molecular cancer therapeutics.* Nat Chem Biol, 2006. **2**(12): p. 689-700.
99. Pegram, M. and D. Slamon, *Biological rationale for HER2/neu (c-erbB2) as a target for monoclonal antibody therapy.* Semin Oncol, 2000. **27**(5 Suppl 9): p. 13-9.
100. Lynch, T.J., et al., *Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib.* New England Journal of Medicine, 2004. **350**(21): p. 2129-2139.
101. Paez, J.G., et al., *EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy.* Science, 2004. **304**(5676): p. 1497-1500.
102. Pao, W., et al., *EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(36): p. 13306-13311.
103. Gonzalez de Castro, D., et al., *Personalized Cancer Medicine: Molecular Diagnostics, Predictive biomarkers, and Drug Resistance.* Clin Pharmacol Ther, 2013. **93**(3): p. 252-259.
104. Fisher, B., et al., *A Randomized Clinical Trial Evaluating Tamoxifen in the Treatment of Patients with Node-Negative Breast Cancer Who Have Estrogen-Receptor–Positive Tumors.* New England Journal of Medicine, 1989. **320**(8): p. 479-484.
105. Fisher, B., et al., *Tamoxifen and Chemotherapy for Lymph Node-Negative, Estrogen Receptor-Positive Breast Cancer.* Journal of the National Cancer Institute, 1997. **89**(22): p. 1673-1682.
106. Yoshida, H., et al., *Accelerated Degradation of PML-Retinoic Acid Receptor α (PML-RARA) Oncoprotein by All-trans-Retinoic Acid in Acute Promyelocytic Leukemia: Possible Role of the Proteasome Pathway.* Cancer Research, 1996. **56**(13): p. 2945-2948.
107. Pao, W., et al., *<italic>KRAS</italic> Mutations and Primary Resistance of Lung Adenocarcinomas to Gefitinib or Erlotinib.* PLoS Med, 2005. **2**(1): p. e17.
108. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-536.
109. Paik, S., et al., *A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer.* New England Journal of Medicine, 2004. **351**(27): p. 2817-2826.
110. Drukker, C.A., et al., *A prospective evaluation of a breast cancer prognosis signature in the observational RASTER study.* International Journal of Cancer, 2013. **133**(4): p. 929-936.
111. Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotech, 2008. **26**(10): p. 1135-1145.
112. Garraway, Levi A. and Eric S. Lander, *Lessons from the Cancer Genome.* Cell, 2013. **153**(1): p. 17-37.
113. Alymani, N.A., et al., *Predictive biomarkers for personalised anti-cancer drug use: Discovery to clinical implementation.* European Journal of Cancer, 2010. **46**(5): p. 869-879.
114. Mauro, M.J., *Defining and Managing Imatinib Resistance.* ASH Education Program Book, 2006. **2006**(1): p. 219-225.
115. Muramatsu, M., et al., *Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme.* Cell, 2000. **102**(5): p. 553-563.
116. Lenz, G. and L.M. Staudt, *Aggressive Lymphomas.* New England Journal of Medicine, 2010. **362**(15): p. 1417-1429.

117. Lenz, G. and L.M. Staudt, *Aggressive lymphomas.* N Engl J Med, 2010. **362**(15): p. 1417-29.
118. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**(6769): p. 503-511.
119. Shaffer, A.L., et al., *Signatures of the Immune Response.* Immunity, 2001. **15**(3): p. 375-385.
120. Klein, U., et al., *Transcriptional analysis of the B cell germinal center reaction.* Proceedings of the National Academy of Sciences, 2003. **100**(5): p. 2639-2644.
121. Roschewski, M., L.M. Staudt, and W.H. Wilson, *Diffuse large B-cell lymphoma[mdash]treatment approaches in the molecular era.* Nat Rev Clin Oncol, 2014. **11**(1): p. 12-23.
122. Armitage, J.O., *My Treatment Approach to Patients With Diffuse Large B-Cell Lymphoma.* Mayo Clinic Proceedings, 2012. **87**(2): p. 161-171.
123. Friedberg, J.W., *Relapsed/Refractory Diffuse Large B-Cell Lymphoma.* ASH Education Program Book, 2011. **2011**(1): p. 498-505.
124. Dunleavy, K., et al., *Dose-Adjusted EPOCH-Rituximab Therapy in Primary Mediastinal B-Cell Lymphoma.* New England Journal of Medicine, 2013. **368**(15): p. 1408-1416.
125. Iqbal, J., et al., *BCL2 Translocation Defines a Unique Tumor Subset within the Germinal Center B-Cell-Like Diffuse Large B-Cell Lymphoma.* The American Journal of Pathology, 2004. **165**(1): p. 159-166.
126. Lenz, G., et al., *Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways.* Proceedings of the National Academy of Sciences, 2008. **105**(36): p. 13520-13525.
127. Rodon, J., et al., *Development of PI3K inhibitors: lessons learned from early clinical trials.* Nat Rev Clin Oncol, 2013. **10**(3): p. 143-153.
128. Morin, R.D., et al., *Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin.* Nat Genet, 2010. **42**(2): p. 181-185.
129. Davis, R.E., et al., *Constitutive Nuclear Factor κB Activity Is Required for Survival of Activated B Cell–like Diffuse Large B Cell Lymphoma Cells.* The Journal of Experimental Medicine, 2001. **194**(12): p. 1861-1874.
130. Lim, K.-H., Y. Yang, and L.M. Staudt, *Pathogenetic importance and therapeutic implications of NF-κB in lymphoid malignancies.* Immunological Reviews, 2012. **246**(1): p. 359-378.
131. Rui, L., et al., *Malignant pirates of the immune system.* Nat Immunol, 2011. **12**(10): p. 933-940.
132. Rawlings, D.J., et al., *Integration of B cell responses through Toll-like receptors and antigen receptors.* Nat Rev Immunol, 2012. **12**(4): p. 282-294.
133. Lenz, G., et al., *Oncogenic CARD11 Mutations in Human Diffuse Large B Cell Lymphoma.* Science, 2008. **319**(5870): p. 1676-1679.
134. Rawlings, D.J., K. Sommer, and M.E. Moreno-Garcia, *The CARMA1 signalosome links the signalling machinery of adaptive and innate immunity in lymphocytes.* Nat Rev Immunol, 2006. **6**(11): p. 799-812.
135. Davis, R.E., et al., *Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma.* Nature, 2010. **463**(7277): p. 88-92.
136. Ngo, V.N., et al., *Oncogenically active MYD88 mutations in human lymphoma.* Nature, 2011. **470**(7332): p. 115-119.
137. Compagno, M., et al., *Mutations of multiple genes cause deregulation of NF-[kgr]B in diffuse large B-cell lymphoma.* Nature, 2009. **459**(7247): p. 717-721.

138. Rosenwald, A., et al., *The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma.* New England Journal of Medicine, 2002. **346**(25): p. 1937-1947.
139. Lenz, G., et al., *Stromal Gene Signatures in Large-B-Cell Lymphomas.* New England Journal of Medicine, 2008. **359**(22): p. 2313-2323.
140. Hewitt, S.M., et al., *Tissue Handling and Specimen Preparation in Surgical Pathology: Issues Concerning the Recovery of Nucleic Acids From Formalin-Fixed, Paraffin-Embedded Tissue.* Archives of Pathology & Laboratory Medicine, 2008. **132**(12): p. 1929-1935.
141. Perry, A.M., et al., *A new biologic prognostic model based on immunohistochemistry predicts survival in patients with diffuse large B-cell lymphoma.* Blood, 2012. **120**(11): p. 2290-2296.
142. Puvvada, S., S. Kendrick, and L. Rimsza, *Molecular classification, pathway addiction, and therapeutic targeting in diffuse large B cell lymphoma.* Cancer Genetics, 2013. **206**(7–8): p. 257-265.
143. Hans, C.P., et al., *Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray.* Blood, 2004. **103**(1): p. 275-282.
144. Gutiérrez-García, G., et al., *Gene-expression profiling and not immunophenotypic algorithms predicts prognosis in patients with diffuse large B-cell lymphoma treated with immunochemotherapy.* Blood, 2011. **117**(18): p. 4836-4843.
145. Dunleavy, K., et al., *Differential efficacy of bortezomib plus chemotherapy within molecular subtypes of diffuse large B-cell lymphoma.* Blood, 2009. **113**(24): p. 6069-6076.
146. Yang, Y., et al., *Exploiting Synthetic Lethality for the Therapy of ABC Diffuse Large B Cell Lymphoma.* Cancer Cell, 2012. **21**(6): p. 723-737.
147. Hernandez-Ilizaliturri, F.J., et al., *Higher response to lenalidomide in relapsed/refractory diffuse large B-cell lymphoma in nongerminal center B-cell–like than in germinal center B-cell–like phenotype.* Cancer, 2011. **117**(22): p. 5058-5066.
148. Chin, L. and J.W. Gray, *Translating insights from the cancer genome into clinical practice.* Nature, 2008. **452**(7187): p. 553-563.
149. Hofree, M., et al., *Network-based stratification of tumor mutations.* Nat Meth, 2013. **10**(11): p. 1108-1115.
150. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-218.
151. *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-337.
152. Raspe, E., C. Decraene, and G. Berx, *Gene expression profiling to dissect the complexity of cancer biology: Pitfalls and promise.* Seminars in Cancer Biology, 2012. **22**(3): p. 250-260.
153. Kreeger, P.K. and D.A. Lauffenburger, *Cancer systems biology: a network modeling perspective.* Carcinogenesis, 2010. **31**(1): p. 2-8.
154. Hanahan, D. and Robert A. Weinberg, *Hallmarks of Cancer: The Next Generation.* Cell, 2011. **144**(5): p. 646-674.
155. Druker, B.J., et al., *Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia.* New England Journal of Medicine, 2006. **355**(23): p. 2408-2417.
156. Nowell, P.C., *A minute chromosome in human chronic granulocytic leukemia.* Science, 1960. **132**: p. 1497.
157. Rowley, J.D., *A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining.* Nature, 1973. **243**(5405): p. 290-293.

158. Heisterkamp, N., et al., *Localization of the c-abl oncogene adjacent to a translocation break point in chronic myelocytic leukaemia.* Nature, 1983. **306**(5940): p. 239-242.
159. Groffen, J., et al., *Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22.* Cell, 1984. **36**(1): p. 93-99.
160. Lugo, T., et al., *Tyrosine kinase activity and transformation potency of bcr-abl oncogene products.* Science, 1990. **247**(4946): p. 1079-1082.
161. Druker, B.J., et al., *Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells.* Nat Med, 1996. **2**(5): p. 561-566.
162. Zhang, J., et al., *Genetic heterogeneity of diffuse large B-cell lymphoma.* Proceedings of the National Academy of Sciences, 2013. **110**(4): p. 1398-1403.
163. Yaffe, M.B., *The Scientific Drunk and the Lamppost: Massive Sequencing Efforts in Cancer Discovery and Treatment.* Sci. Signal., 2013. **6**(269): p. pe13-.
164. *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-1068.
165. *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-615.
166. Negrini, S., V.G. Gorgoulis, and T.D. Halazonetis, *Genomic instability [mdash] an evolving hallmark of cancer.* Nat Rev Mol Cell Biol, 2010. **11**(3): p. 220-228.
167. Stratton, M.R., *Exploring the Genomes of Cancer Cells: Progress and Promise*. Science, 2011. **331**(6024): p. 1553-1558.
168. Swanton, C., R. Burrell, and P.A. Futreal, *Breast cancer genome heterogeneity: a challenge to personalised medicine?* Breast Cancer Research, 2011. **13**(1): p. 104.
169. Hutchinson, L., *Predicting cancer's next move.* Nat Rev Clin Oncol, 2014. **11**(2): p. 61-62.

ACKNOWLEDGEMENTS

Looking back at the journey of my PhD, I can't but be thankful to be able to do great science and to enjoy it at the same time. It is clear for me that it is the people that surrounded me and believed in me that made this journey an amazing one. For that, I would like to deeply thank the following people:

Prof. Dr. Matthias Mann for giving me the opportunity to work in his lab, to learn state-of-the-art technology and to apply it to challenging questions. Thank you Matthias for your trust and for being a visionary who is never short of exciting and thrilling ideas. Your strong belief in your goals and your absence of fear to push the limits are definitely things I aspire. Thank you for giving me the opportunity to get in contact with top-notch scientific environments and to explore the world and meet amazing people from all over the globe.

Prof. Dr. Marc Schmidt-Supprian for being my co-advisor and for the great support. Thank you Marc for all the scientific input and for taking part in my thesis advisory committee. Thank you also for encouraging me at the beginning of my PhD to pick my own projects and work on what I find interesting.

Prof. Dr. Hubert Serve for being a member of my thesis advisory committee and for all your scientific and clinical input.

Dr. Dietmar Martin, Prof. Dr. Klaus Förstemann, Prof. Dr. Karl-Peter Hopfner, and Prof. Dr. Reinhard Fässler for your support and for being part of my thesis committee.

Alison Dalfovo and Theresa Schneider for your help with all the administrative issues. It is never easy to deal with German documents. Thank you for all the help!

I want to thank the IMPRS coordination office for selecting me and for giving me the opportunity to visit the MPI and meet the Mann lab. Thank you also for all the support and help during my transition to Germany.

My office mates for all the fruitful discussions as well as all the laughs and the good times. Jürgen Cox for the support in data analysis and bioinformatics. The office girls, Francesca, Stefka and Tikira, for all the scientific and non-scientific discussions ☺. It was great having you around! Thank you for all the support especially during the time I was writing my thesis.

Korbi and Igor– thank you for all the help with the machines! I learned so much from you and I highly appreciate it!

All of my IMPRS girls: Joanna, Valeria, Vanessa, Natalia and Ania for all the fun times!

Marco for all the interesting discussions and support!

Eva for all the walks and talks ☺

Gaby for being a great friend and for staying in touch even after you left the lab! You know that I enjoy all the fun times we spend together. I am looking forward for more of them ;)

Charo for being there and for being a great support when it comes to all walks of life. Thank you also for your two amazing little ones: Ana and Leo <3 I looked forward to see them on Friday afternoons and play with them.

Marlis, my bestie! I cherish all the time we spent together and all the hours of just talking about all sorts of topics. I had great time cooking, partying, and even doing 'sports' with you ☺. It was great having you around. You were more like a sister and someone I can always count on. I know that our friendship will not end with our PhDs. I wish you all the best in your next step which I hope will be in Munich ☺

My brothers, Wael and Ramee, for being there for me and for all the fun times!

My little sister, reem, for being supportive, encouraging and a sweetheart. There were days when just talking to you would make me laugh and forget everything.

I can't but thank a special person to whom I am addicted. Nadim, thank you for being there and for being who you are.

My parents, to whom I owe everything, thank you for believing in me and supporting me in every step I made. Thank you for your outmost trust and love. I wouldn't have made it without you.