From Department of Oncology-Pathology
Karolinska Institutet, Stockholm, Sweden

CANCER PROTEOGENOMICS – CONNECTING GENOTYPE TO MOLECULAR PHENOTYPE

Ioannis Siavelis

Karolinska Institutet

Stockholm 2022

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2022

ISBN 978-91-8016-873-1
Cover illustration: Depiction of the molecular dogma of biology using the first capitalized greek letter for DNA (Δεοξυριβονουκλεϊκό οξύ), RNA (Ριβονουκλεϊκό οξύ) and protein (Πρωτεΐνη).

Cancer proteogenomics - connecting genotype to molecular phenotype
Thesis for Doctoral Degree (Ph.D.)

By

Ioannis Siavelis

The thesis will be defended in public at Lecture Hall (Ground Floor), Karolinska Comprehensive Cancer Center (CCK), R8:01, Stockholm, 16th December 2022 at 9:00 am

Principal Supervisor:
Prof. Janne Lehtiö, Ph.D.
Karolinska Institutet
Department of Oncology and Pathology

Co-supervisor(s):
Dr. Henrik Johansson, Ph.D.
Karolinska Institutet
Department of Oncology and Pathology

Opponent:
Prof. Matthias Selbach, Ph.D.
Max-Delbrück Center for Molecular Medicine (MDC)
Department of Proteome Dynamics

Examination Board:
Prof. Roman Zubarev, Ph.D.
Karolinska Institutet
Department of Medical Biochemistry and Biophysics (MBB)

Doc. Mika Gustafsson, Ph.D.
Linköpings universitet
Department of Physics, Chemistry and Biology (IFM)

Dr. Anna Pasetto, Ph.D.
Karolinska Institutet
Department of Laboratory Medicine
Oslo University,
Oslo University Hospital

Popular science summary of the thesis

Our body consists of trillions of cells. We can think of cells as towns with buildings for different purposes, e.g. houses for accommodation, schools, and hospitals. All these buildings contribute to a town's good functioning. To construct a building, though, is a process with multiple steps including: first, a blueprint with instructions about the room arrangement; second, a construction site to lay the foundations of the building; third, the actual building. By analogy, buildings inside cells are called proteins. The blueprint that holds information about how to produce proteins is called DNA and the construction site that sets the stage for protein is called RNA. Taken together, the building analogue in biology is a flow of information from DNA to RNA and, finally, to protein that is called the central dogma of molecular biology.

Building a construction, however, is not without rules, but is regulated, among others, by the blueprint drawings and the engineering plans. Similarly, the central dogma is controlled by processes that dictate RNA and protein abundances inside cells like the DNA status. In science, we are now able to infer the control along the central dogma by measuring the total number of DNA, RNA and proteins using large amount of data called omics. The study of DNA, RNA and protein omics data is called proteogenomics.

In this thesis, we used proteogenomics to study the central dogma regulation in human cells that are malfunctioning. These cells are proliferating in a fast and abnormal way at the expense of nearby cells forming mases called tumors and, eventually, causing a disease known as cancer: the second most common cause of death in humans.

In **Papers I** and **II**, we studied two different types of cancers called breast cancer and acute lymphoblastic leukemia, respectively. Proteogenomics data suggested that an increase in DNA abundance by means of multiple DNA copies–a situation called copy number alteration–leads mainly to overproduction of RNA and, to a lesser extent, to overproduction of proteins. Interestingly, certain copy number alterations only affected RNA leaving protein levels unchanged. We found evidence that this compensation is related to the increased destruction of proteins that are interacting together to form structures called complexes. This phenomenon resembles building a block of flats where the area of one flat affects the area of its adjacent flat to maintain the symmetry of the overall construction.

In **Paper III**, we studied a lethal type of cancer called non-small cell lung cancer. We found that lung cancer patients could be separated into six distinct groups based on their protein abundances in the cancer cells and in the surrounding cells of the tumor, namely the tumor microenvironment. The six groups were related to specific cancer-promoting processes like energy production or proliferation but were also characterized by unique proteins produced from previously inactive DNA. In the building analogue, these proteins are like

exotic touristic attractions which are designed from previously low-profit architecture firms seeking new opportunities in prosperous times.

Notwithstanding, a typical town consists less of exotic buildings and more of buildings with similar architecture. These buildings are based on the same blueprint but, for example, differ in color, neon signs or other decorations on the outside. In biology, these buildings relate to proteoforms, that is proteins that originate from the same DNA but differ in their final form. **Paper IV** describes a tool to detect proteoforms from omics data. The tool finds proteoforms by 'demolishing' proteins and inferring proteoforms from the remnants. We found that a few of the identified proteoforms belonged to 'alternatively engineered' RNA forms called splice variants.

In summary, this thesis presents cancer proteogenomics data that track the central dogma of biology and its regulation. For a town mayor, it is more informative to know the architecture of the town than the building blueprints to intervene for the well-being of the citizens. Similarly for scientists, it is more informative to know proteins than DNA or RNA to intervene in a disease for the patients' best interest. It seems that the future in cancer research is a matter of collaboration between doctors, architects, and engineers.

Abstract

The central dogma of molecular biology describes the one-way road from DNA to RNA and finally to protein. Yet, how this flow of information encoded in DNA as genes (genotype) is regulated in order to produce the observable traits of an individual (phenotype) remains unanswered. Recent advances in high-throughput data, i.e., 'omics', have allowed the quantification of DNA, RNA and protein levels leading to integrative analyses that essentially probe the central dogma along all of its constituent molecules. Evidence from these analyses suggest that mRNA abundances are at best a moderate proxy for proteins which are the main functional units of cells and thus closer to the phenotype.

Cancer proteogenomic studies consider the ensemble of proteins, the so-called proteome, as the readout of the functional molecular phenotype to investigate its influence by upstream events, for example DNA copy number alterations. In typical proteogenomic studies, however, the identified proteome is a simplification of its actual composition, as they methodologically disregard events such as splicing, proteolytic cleavage and post-translational modifications that generate unique protein species – proteoforms.

The scope of this thesis is to study the proteome diversity in terms of: a) the complex genetic background of three tumor types, i.e. breast cancer, childhood acute lymphoblastic leukemia and lung cancer, and b) the proteoform composition, describing a computational method for detecting protein species based on their distinct quantitative profiles.

In **Paper I**, we present a proteogenomic landscape of 45 breast cancer samples representative of the five PAM50 intrinsic subtypes. We studied the effect of copy number alterations (CNA) on mRNA and protein levels, overlaying a public dataset of drug-perturbed protein degradation.

In **Paper II**, we describe a proteogenomic analysis of 27 B-cell precursor acute lymphoblastic leukemia clinical samples that compares high hyperdiploid versus ETV6/RUNX1-positive cases. We examined the impact of the amplified chromosomes on mRNA and protein abundance, specifically the linear trend between the amplification level and the dosage effect. Moreover, we investigated mRNA-protein quantitative discrepancies with regard to post-transcriptional and post-translational effects such as mRNA/protein stability and miRNA targeting.

In **Paper III**, we describe a proteogenomic cohort of 141 non-small cell lung cancer clinical samples. We used clustering methods to identify six distinct proteome-based subtypes. We integrated the protein abundances in pathways using protein-protein correlation networks, bioinformatically deconvoluted the immune composition and characterized the neoantigen burden.

In **Paper IV**, we developed a pipeline for proteoform detection from bottom-up mass-spectrometry-based proteomics. Using an in-depth proteomics dataset of 18 cancer cell lines, we identified proteoforms related to splice variant peptides supported by RNA-seq data.

This thesis adds on the previous literature of proteogenomic studies by analyzing the tumor proteome and its regulation along the flow of the central dogma of molecular biology. It is anticipated that some of these findings would lead to novel insights about tumor biology and set the stage for clinical applications to improve the current cancer patient care.

List of scientific papers

I. Johansson, H.J., Socciarelli, F., Vacanti, N.M., Haugen, M.H., Zhu, Y., **Siavelis, I.**, Fernandez-Woodbridge, A., Aure, M.R., Sennblad, B., Vesterlund, M., Branca, R.M., Orre, L.M., Huss, M., Fredlund, E., Beraki, E., Garred, Ø., Boekel, J., Sauer, T., Zhao, W., Nord, S., Höglander, E.K., Jans, D.C., Brismar, H., Haukaas, T.H., Bathen, T.F., Schlichting, E., Naume, B., Luders, T., Borgen, E., Kristensen, V.N., Russnes, H.G., Lingjærde, O.C., Mills, G.B., Sahlberg, K.K., Børresen-Dale, A.-L., Lehtiö, J., 2019.
Breast cancer quantitative proteome and proteogenomic landscape.
Nature Communications 10, 1600.

II. Yang, M., Vesterlund, M., **Siavelis, I.**, Moura-Castro, L.H., Castor, A., Fioretos, T., Jafari, R., Lilljebjörn, H., Odom, D.T., Olsson, L., Ravi, N., Woodward, E.L., Harewood, L., Lehtiö, J., Paulsson, K., 2019.
Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia.
Nature Communications 10, 1519.

III. Lehtiö, J., Arslan, T., **Siavelis, I.**, Pan, Y., Socciarelli, F., Berkovska, O., Umer, H.M., Mermelekas, G., Pirmoradian, M., Jönsson, M., Brunnström, H., Brustugun, O.T., Purohit, K.P., Cunningham, R., Asl, H.F., Isaksson, S., Arbajian, E., Aine, M., Karlsson, A., Kotevska, M., Hansen, C.G., Haakensen, V.D., Helland, Å., Tamborero, D., Johansson, H.J., Branca, R.M., Planck, M., Staaf, J., Orre, L.M., 2021.
Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune-evasion mechanisms.
Nature Cancer 2, 1224–1242.

IV. **Siavelis, I.**, Johansson, H.J., Stahl, M., Socciarelli, F., Mermelekas, G., Jafari, R., Lehtiö, J.
DEpMS: Differential Expression analysis of proteoforms from Mass Spectrometry-based bottom-up proteomics.
Manuscript

Contents

List of abbreviations

ADAMTS12	ADAM Metallopeptidase With Thrombospondin Type 1 Motif 12
ALK	Anaplastic lymphoma kinase
ALL	Acute lymphoblastic leukemia
ARHGDIB	Rho Guanosine Diphosphate Dissociation Inhibitor Beta
ARID1A	AT-Rich Interaction Domain 1A
ASCAT	Allele-specific copy number analysis of tumors
BAF	B allele frequency
BC	Breast cancer
BCP-ALL	B-Cell Precursor Acute Lymphoblastic Leukemia
BH	Benjamini-Hochberg
BRAF	B-Raf Proto-Oncogene, Serine/Threonine Kinase
BRCA1	Breast Cancer 1
BRCA2	Breast Cancer 2
CCNB1	Cyclin B1
CD8	Cluster of Differentiation 8
cDNA	Complementary DNA
CNA	Copy number alteration
CS	Consensus clustering
CTA	Cancer-testis antigen
CTCF	CCCTC-Binding Factor
CTNNA2	Catenin Alpha 2
CTNNB1	Catenin Beta-1
DDA	Data-dependent acquisition
DIA	Data-independent acquisition
DNA	Deoxyribonucleic acid
EGFR	Epidermal Growth Factor Receptor
ELM	Eukaryotic Linear Motif

EMT	Epithelial–mesenchymal transition
eQTL	Expression quantitative loci
ER	Estrogen receptor
ERBB2	Erb-B2 Receptor Tyrosine Kinase 2
ESI	Electrospray ionization
ETV6	ETS variant transcription factor 6
GC-content	Guanine-Cytosine content
HDAC6	Histone deacetylase 6
HER2	Human epidermal growth factor receptor 2
HiC	Chromatin conformation capture
HiRIEF	High-resolution isoelectric focusing
HPLC	High-performance liquid chromatography
HPV	Human papillomavirus
HSP90	Heat shock protein 90
IGFALS	Insulin-like growth factor-binding protein complex acid labile subunit
INPPL1	Inositol Polyphosphate Phosphatase Like 1
IPAW	Integrated proteogenomics analysis workflow
JAK1	Janus Kinase 1
KEAP1	Kelch-like ECH-associated protein 1
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIF20A	Kinesin Family Member 20A
KRAS	Kirsten ras oncogene proto-oncogene
LC	Liquid chromatography
LFQ	Label-free quantification
MET	MET Proto-Oncogene, Receptor Tyrosine Kinase
MFGE8	Milk fat globule-EGF factor 8
miRNA	MicroRNA
mRNA	Messenger RNA
MS	Mass spectrometry

MYO1E	Myosin IE
NCPs	Non-canonical proteins/peptides
NMF	Non-negative matrix factorization
NSCLC	Non-small cell lung cancer
OXPHOS	Oxidative phosphorylation
PACSIN2	Protein Kinase C And Casein Kinase Substrate In Neurons 2
PC	Principal component
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PEST	Proline (P), glutamic acid (E), serine (S), threonine (T)
pI	Isoelectric point
PIK3CA	Phosphatidylinositol 4,5-Bisphosphate 3-Kinase
pQTL	Protein quantitative trait loci
PR	Progesteron receptor
PTEN	Phosphatase And Tensin Homolog
RB1	Retinoblastoma transcriptional corepressor 1
RBM27	RNA Binding Motif Protein 27
RBP	RNA binding protein
RMATS	Replicate Multivariate Analysis of Transcript Splicing (MATS)
RNA	Ribonucleic acid
RUNX1	Runt-related transcription factor 1
SILAC	Stable isotope labeling by amino acids in cell culture
SMARCA1	SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily A, Member 1
SNP	Single nucleotide polymorphism
SOX9	SRY-Box Transcription Factor 9
STK11	Serine/threonine kinase 11
SYNE1	Spectrin Repeat Containing Nuclear Envelope Protein 1
TAD	Topological associated domains

TMB	Tumor mutational burden
TMT	Tandem mass tag
TP53	Tumor Protein P53
tRNA	Transfer-RNA
UMAP	Unifold Manifold Approximation and Projection
UPS	Ubiquitin proteasome system
VEGF	Vascular endothelial growth factor
WEE1	WEE1 G2 Checkpoint Kinase
WES	Whole exome sequencing
WGS	Whole genome sequencing

1 Introduction

1.1 The central dogma of molecular biology

The central dogma of molecular biology describes the flow of information from DNA (deoxyribonucleic acid) to RNA (ribonucleic acid) to proteins through the processes of replication [1], transcription [2] and translation [3]. Yet, how information –encoded in DNA as genes (genotype)– flows across the different molecular levels to be 'expressed' –decoded as observable characteristics (phenotype)– has been a matter of interest since the observation that inbred isogenic bean plants produced pods of varying size [4]. At the molecular level, phenotype can be ascribed to the protein composition, although a more precise definition will be endophenotype to clarify that proteins are still intermediates between gene expression and the actual phenotype [5] (**Figure 1**).

A prerequisite to understand the emergence of cellular phenotypes is the ability to efficiently probe the primary molecular constituents of the central dogma. From the observational experiments of Gregor Mendel in 1860s, we have now reached high-throughput data generation that profile the entirety of DNA, RNA and proteins (-omics data). Large scale efforts analyzing omics data led to the first draft sequence of the human genome in the beginning of 21st century [6,7]. The similar proteome draft lagged ten years [8,9], owing to the greater molecular complexity of proteins. These obstacles were surpassed by advances in sensitivity, standardization, and multiplexing in mass spectrometry (MS) proteomics methods [10–12].

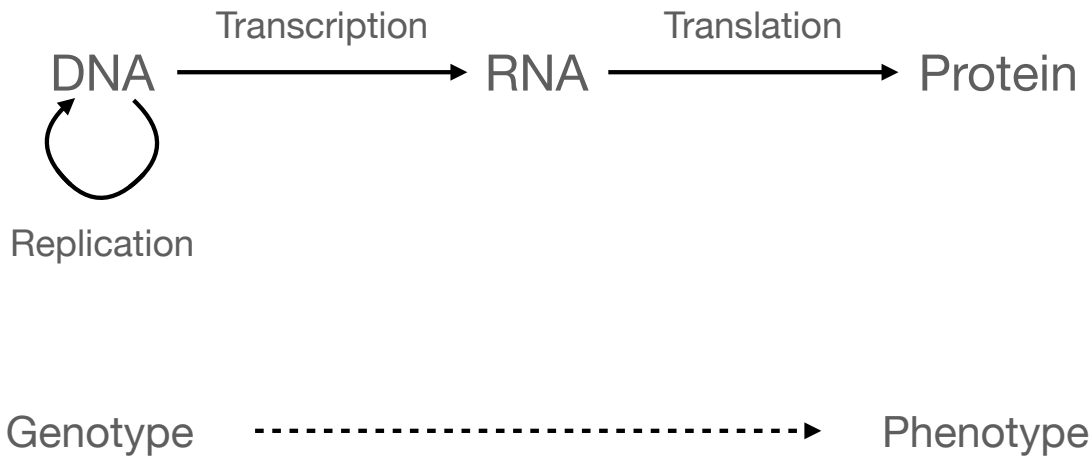

Figure 1. The central dogma of molecular biology. Information flow from DNA (genotype) to RNA to protein (phenotype).

1.2 Central dogma regulation

Beyond its profound simplicity, the DNA to protein pathway is under extensive regulation –imprinted on the molecular structure of DNA, RNA, proteins and their interactions – via mechanisms acting:

1) Transcriptionally: For example, DNA and histone modifications, Guanine-Cytosine (GC)-content, open/close chromatin state, topological associated domains (TADs), gene proximity to regulatory sequences (enhancer/silencer/promoter regions), composition of transcription factors, co-factors, chromatin modifiers and non-coding RNAs, and RNA polymerase status (kinetic model of transcription initiation, elongation, and termination) [2].

2) Post-transcriptionally: For example, 5'-end capping and 3'-end polyadenylation, splicing, nuclear export to cytoplasm, miRNAs and RNA-binding proteins, RNA modifications, and RNA decay [13,14].

3) Translationally: For example, 5' prime motifs, upstream open reading frames, mRNA length and secondary structure, Kozaq sequences, premature stop codons, miRNAs and RNA-binding proteins that dictate translational efficiency, codon-bias, translation initiation factor abundances, tRNA availability, as well as the abundance, composition, and phosphorylation status of ribosomes [15,16].

4) Post-translationally: For example, C- and N-terminal peptide structure, eukaryotic linear motifs (ELMs), PEST sequences enriched in proline (P), glutamic acid (E), serine (S),and threonine (T), peptide modifications (phosphorylation, acetylation), chaperone-assisted protein folding, intra- and extra-cellular transport and compartmentalization, and degradation via proteasome or lysosome [17].

These regulatory processes (**Figure 2**) span all scales of biology–single molecules, pathways, subcellular structures, cells, tissues, organs, and organisms–to ensure cellular functionality with robust phenotypes under:

1) Homeostatic conditions: Cells need to maintain steady internal conditions. Proteins such as transcription factors and signaling genes need to be dynamically adjusted while housekeeping proteins, e.g. ribosomal and cytoskeletal proteins, should be kept at a constant concentration. In that manner, the biological role is reflected on protein abundance.

2) Biological noise: Cells have to cope with external noise, e.g., cues that act on surface receptors , and with internal noise caused by stochastic transcription [18].

3) Metabolic constraints: Varying energy resources remodel cellular protein abundances by optimizing RNA translation, protein shuttling between organelles, protein degradation, and enzymatic reactions [19,20].

4) Environmental and physiological changes: For example, these include adaptation to stress conditions and ageing [21], and monitoring of circadian rhythm [22].

5) Genetic variations. Germline and somatic DNA alterations perturb gene expression levels that in turn need to be re-adjusted [23].

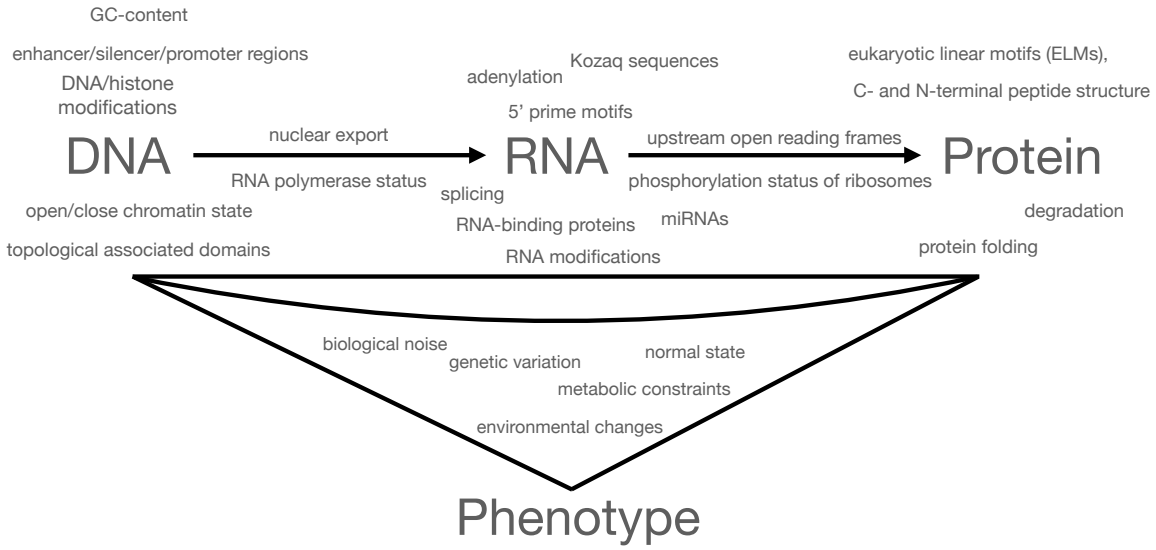

Figure 2. Central dogma regulation accommodates perturbations to produce sustainable phenotypes.

1.3 Central dogma quantification

How protein abundance relates to upstream molecules, especially RNA, has been a question tracing the history of molecular dogma regulation. Based on previous studies (extensively reviewed in [24–26]), experimental settings can vary by:

1) Experimental model: Bacteria, eukaryotic single-cell microorganisms, human cells, xenografts, and tissue samples in normal or disease state.

2) Level of analysis: Bulk/cell-population level or single-cell level.

3) State of the experimental system which can be divided into:

    a) Steady state, in which the sample's condition remains constant. This category includes popular experimental methods such as single nucleotide

polymorphism (SNP) arrays, RNA-sequencing and mass-spectrometry for analysis of genomic, transcriptomics, and proteomic data, respectively.

b) Dynamic state in which the sample's condition changes. This includes modified experimental methods such as RNA-labeling, pulsed-SILAC and fluorescence microscopy, all of which capture time-dependent changes of labeled molecules due to degradation or synthesis.

2) Computational method: Analyses of the relationship between genotype and phenotype can be based on correlations, linear and non-linear regression, expression and protein quantitative trait loci (eQTLs, pQTLs*),* kinetic models assuming synthesis and degradation rates for RNA or more complex models incorporating splice variants [27,28].

3) Dimension of analysis, which fundamentally distinguishes two kinds of estimation:

a) mRNA-protein covariation of quantified genes in a specific sample, assessing how inherent gene features regulate mRNA-protein relationship.

b) mRNA-protein covariation for a specific gene across samples, assessing how sample-specific features regulate the mRNA-protein relationship (**Figure 3**).

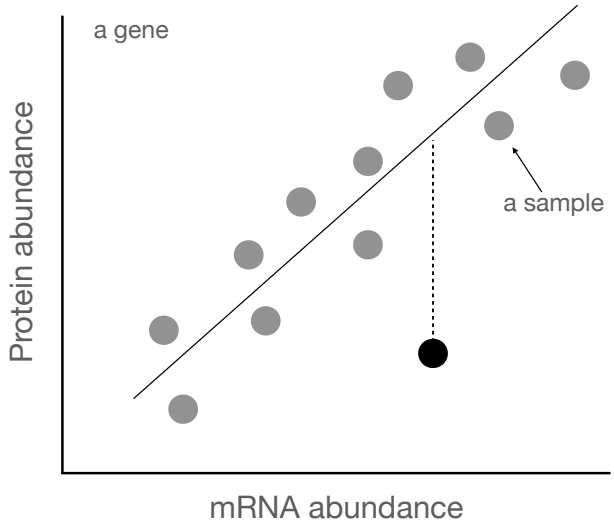

Figure 3. Correlation between mRNA and protein abundance for a gene across samples. Solid line is the expected protein abundance of a sample given its mRNA abundance. Displayed in black is a sample for which the gene's measured protein abundance is deviating from its expected value.

The above categorization suggests a framework to chronologically dissect previous mRNA-protein correlation studies.

1.3.1 Across-gene correlations

Steady-state, across-gene correlations between mRNA and proteins in a systematic manner were calculated as early as 1999 with an estimate of ~0.4 in yeast studies [29]. Up to 2009, technical limitations in depth and accuracy of proteome and transcriptome quantification had precluded high mRNA-protein correlations. As an example, in 2008, the overall correlation coefficient across 150 signature genes in hematopoietic cell lines was found to be 0.59 [30]. Similarly, in haploid versus diploid yeast cells, a moderate correlation (R = 0.46 to 0.68) was observed after excluding outliers [31]. In 2010, relative abundances were only partially predicted by mRNA in bone osteosarcoma, squamous cell carcinoma and brain glioblastoma cell lines (overall Spearman correlation 0.63, 0.60 and 0.58) [32], while at the same year, measurements in cellular lysate from the human Daoy medulloblastoma cell line yielded an estimate of 0.46 [33]; accounting for sequence features raised this correlation to 0.82, thus explaining two thirds of the protein variance. In 2011, Spearman's correlation between transcript and protein abundance values of 8609 genes in HeLa cells reached a value of 0.6 [34]. In the same year, a seminal paper by Schwanhäusser et al. [35] established a kinetic model of the central dogma of biology incorporating turnover rates for RNAs and proteins. RNA and protein levels were clearly correlated (R^2 = 0.41) and that association markedly improved when accounting for translation rate constants (R^2 = 0.95), although extrapolating their model to MCF7 cell lines explained 60% of the protein variance. Re-analysis of the same dataset with different models raised that variance to about 56%–84%, while the translation rate could only explain 9% of the protein abundance variability [36].

The year of 2014 was marked by the publication of the draft map of the human proteome based on mass spectrometry analysis by Wilheim et al. [8]. The authors compared RNA-seq data with their quantitative protein measurements of 12 human tissues reporting an average Spearman correlation of 0.41. Based on Schwanhäusser's kinetic model and by using protein to mRNA ratios as proxy for translation rates, they argued that one could accurately predict protein abundances from the measured mRNA abundance given a transcript-specific constant rate in each tissue. However, these assertions were proven to be influenced by the Simpson paradox, conflating across-sample variability with across-gene variability [37,38].

1.3.2 Across-sample correlations

Steady state, across-sample correlations were investigated already in 2009 for 1066 genes in 23 cell lines, with per gene correlations ranging from 0.25 to 0.52, depending on different experimental methods [39]. However, this type of analysis was mainly undertaken by subsequent quantitative trait loci (QTL) studies investigating how genetic variation affects RNA (eQTL) and protein abundances (pQTL) [40–50]. Despite the early MS instrumentation and low sample sizes, these studies demonstrated that:

1) Most protein associations lacked equivalent transcript associations or were estimated with reduced effect sizes

2) Uniquely identified pQTLs could be attributed to post-transcriptional mechanisms

3) Genomic variants affecting multiple transcript/proteins, the so-called hotspots, seemed to differ at the mRNA and protein level

4) pQTLs captured protein-protein interactions and disease-associated variants and

5) pQTLs could act as molecular fingerprints of environmental changes. In such cases, genomic information seemed to evade buffering and propagated at the protein level, imprinting unique phenotypic changes.

1.3.3 Correlations in time and space

As an extension to the analyses explained above, longitudinal studies explored time-dependent mRNA and protein profiles during perturbations or developmental stages [51–60]. Compared with steady-state experiments, longitudinal studies incorporated dynamical modeling to: describe the role of protein synthesis and degradation in varying conditions, categorize dynamic changes as either acute or lagging, propose post-transcriptional regulatory mechanisms (such as RNA-binding proteins), and underscore rewiring of the protein interactome.

On the spatial scale, single-cell analyses investigated location-dependent mRNA-protein correlations elucidating the intratumor heterogeneity of the molecular dogma [61,62]. Gene co-variations, apparent from the RNA level, suggested early post-transcriptional regulation by miRNAs, nuclear localization and complex membership [63].

1.3.4 Meta-analyses

Lastly, accumulated proteomics data in meta-analyses recapitulated that mRNA levels correlated moderately with protein abundances reaching an across-gene Spearman correlation of 0.58 (range: 0.43-0.66) and an across-sample correlation of 0.31 [64]. Consistently, in human tissue samples the explainable percentage of protein abundance by mRNA-protein ratios varied between ~55% to ~80%, indicating post-transcriptional and autoregulatory mechanisms of gene expression [65].

1.4 Central dogma in cancer

The evidence above suggests that mRNA is at least a moderate proxy of protein abundances across the diverse experimental settings, and distinct molecular phenotypes can arise irrespective of upstream events in the central dogma. A particular phenotype of interest is cancer, where cells deviate from a physiological to a neoplastic state through the accumulation of specific traits called hallmarks [66–68]. In this context, cancer can be conceptualized as a molecular phenotype that emerges from dysregulations of the central dogma. For example, cancer cells can highjack enhancers, generate alternative splice-variants, potentiate translation, and enhance proteasome degradation at the transcriptional, post-transcriptional, translational, and post-translational level, respectively. A genotype to phenotype view on cancer inevitably starts from DNA level with genome instability as one of the facilitating characteristics of cancer hallmarks to eventually describe the impact of DNA alterations –mutations, structural variants, and aneuploidy– on the proteome.

1.4.1 Copy number alterations

Aneuploidy, the abnormal chromosome stoichiometry in an organism, originates from chromosome missegregation or non-reciprocal translocations during mitosis. It manifests in 90% of solid tumors as whole chromosome, chromosomal arm-level, and somatic copy number alterations (CNAs) [69]. A typical solid tumor harbors approximately 3 gains and 5 losses of chromosome arms or longer chromosomal regions [70]. The consequences of aneuploidy, which can be detrimental or fitness-promoting, are highly-context specific, depending on cell type, genetic make-up, tumor stage and tumor microenvironment [71].

An initial distinction of the CNAs effects is whether local (in-cis) or distant (trans) effects are investigated. Chromosomal gains or losses can supposedly be linked to in-cis over-expression of oncogenes and loss of tumor suppressors respectively with interesting exceptions of rescue mechanisms as seen in mutations for which sequence-related genes can assume the function of the mutant protein via transcriptional adaptation [72]. However, effects of broad CNAs can equally originate from trans-combinatorial gene alterations via changes of transcription factor abundances or epigenetic events, e.g. disruption of DNA methylation, increase of transcriptional noise, resource overload fueled by uninterrupted proliferative signaling, stoichiometric imbalance, promiscuous protein-protein interactions, and pathway modulation [73]. As observed in cancer cell lines, CNAs pervade the protein level albeit with less effect, suggesting existence of compensatory mechanisms (**Figure 4**).

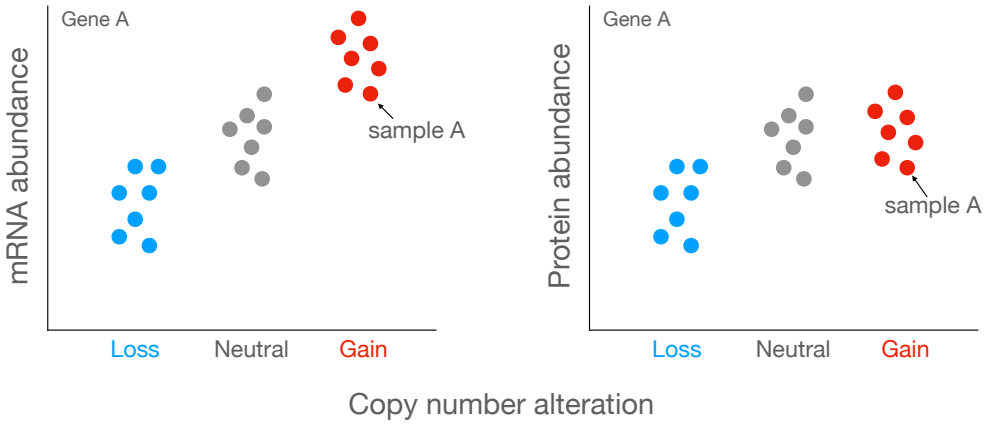

Figure 4. In-cis effect of CNAs. Abundances of mRNA (left) and protein (right) for the same gene (gene A) across three different CNA statuses (loss, neutral, gain). Protein-level buffering for samples with copy number gains is displayed.

1.4.1.1 CNA effect on protein complexes

Compensatory mechanisms most likely take place post-translationally through protein degradation as shown in disomic yeast strains in which translation efficiency remained unaltered for the buffered proteins [74]. This finding contrasts with the proportional synthesis model which poses that complex formation is defined during translation [75]; suggesting that genetic perturbations shift the control of complex stoichiometry from translational to post-translational mechanisms along with degradation rate modulation [76].

The studies above underscore the need of balanced expression of complex components for multi-subunit protein complexes in order to function properly [77]. Unbalanced production of proteins could impair specific cellular functions associated with the affected protein complexes as seen in yeast experiments in which lethality of β-tubulin gene gains is abrogated by additional gains of a-tubulin gene, thus restoring the stoichiometry of a/β-tubulin dimers. Consistently, fibroblasts from Down syndrome (trisomy 21) patients degrade complex members of the respective chromosome to maintain stoichiometry [78]. Meta-analysis in tumors found that CNAs are buffered by post-transcriptional regulation in protein complex members, further demonstrating that some complex subunits act as rate-limiting steps in complex assembly [79,80].

1.4.1.2 Systems-level CNA effects

Gene compensation due to protein complex formation represents a fragmented view of CNAs impact on cancer cells. A system-level view on cellular function suggests that aneuploidy compromises protein degradation by eliciting proteotoxic stress [81]. Proteotoxic stress describes the inability of the main cellular pathways of proteome

maintenance to fold proteins properly and control protein concentrations [82]. These pathways might be underloaded under normal conditions but are deployed during proteotoxic stress to cope with excessive protein degradation [83] and maintain the integrity of the proteome. Proteotoxic stress compensation pathways include:

1) the heat shock chaperone network that contributes to proper protein folding. For example, in a large group of cancers, MYC oncogene activation was shown to rewire the chaperone network to support a pro-survival heat shock protein HSP90/HSC70-driven integrated protein-protein interactome [84,85].
2) the ubiquitin proteasome system (UPS), a multi-step process of attaching ubiquitin to soon-to-be-degraded proteins by a macro-molecular protein complex called proteasome [86]. This is exemplified by the widespread driver mutations of key enzymes that are responsible for protein degradation in cancer [87].
3) autophagy-lysosome system [88]. Mechanisms by which autophagy promotes cancer include suppression of the p53 tumor suppressor protein and maintainenance of mitochondrial metabolic function [89].
4) aggresome, a proteasome-independent, histone deacetylase 6 (HDAC6)-driven pathway that targets misfolded proteins to the lysosome or to the chaperone network for degradation or refolding respectively [90].

In a pan-cancer context, aneuploidy has been shown to correlate with cell-cycle genes [91] in contrast to yeast models that showed decreased proliferation [74]. A general picture emerges that at the expense of an initial survival cost, aneuploidy educates cells to endure high levels of proteome imbalance and to sample mutations from a fitness landscape in order to assign heterogeneous roles in tumor cells [92]. How genetic alterations establish heterogenous molecular phenotypes is of great translational significance in clinical proteogenomics studies since compensating mechanisms can be pharmacologically targeted [93].

1.5 Clinical cancer proteogenomics studies

Proteogenomics combines MS-based proteomics with additional omics-level evidence to probe gene expression regulation [94]. Beyond the regulatory principles discussed above regarding the canonical proteome, proteogenomic analysis can further lead to the discovery of novel protein-coding regions by using customized, DNA/RNA-based databases upon which spectra from an MS experiment can be mapped to. Refined gene models by improved spectra mappability has led to the identification of single nucleotide polymorphisms, allele-specific expression variations, large structural chromosomal variations, alternative spliced transcripts, alternative translation initiation sites, and novel open reading frame events [95,96]. Importantly, interrogation of post-

translationally modified proteins, predominantly via phosphorylation, acetylation and methylation, highlighted the added value of investigating proteoform expression [97].

Series of clinical cancer proteogenomic studies [98–137] have been published in recent years assessing: the impact of DNA alterations on protein abundance; genome-wide mRNA-protein correlations; proteome-centric sample subtyping and pathway re-wiring.

In-depth clinical onco-proteogenomics studies investigate how genetic variation of mutations and somatic CNAs in cancer change the proteome landscape by interrogating all three major molecular components –DNA, RNA, and proteins. Examples provided below demonstrate that these proteome-oriented multi-omics analyses indicate:

1) Increased accuracy in predicting the impact of mutations on the proteome. In endometrial cancer, the effect of missense mutations remained undetected at the RNA level but changed protein concentration, as observed in high CTNNB1 and TP53 expression and low PIK3CA and SYNE1 expression for samples with corresponding hotspot mutations [115]. Moreover, truncating mutations in driver genes ***ARID1A, INPPL1, JAK1***, ***PTEN***, and ***RBM27*** led to decreased protein levels as expected. Conversely, in colon cancer, truncating mutations upregulated transcription factor SOX9 as most of the mutations occurred upstream of the ubiquitin-target site K398 which is responsible for proteasomal degradation [106].

2) Protein-level dampening of CNA-driven mRNA abundances: In colon cancer, focal amplifications had the strongest local cis-effects on both mRNA and protein level, suggesting that CNAs in regions of focal amplification nominate genes with high protein abundance [98]. However, many SCNA-driven events at the mRNA level were buffered at their protein counterpart attesting for limited phenotypic impact –results that were confirmed with more robust proteomics quantitation .

3) Identification of driver- and survival-related genes from hotspots in the CNA-protein correlation analysis. In high grade serous ovarian cancer, many hotspot-protein-level alterations were uniquely enriched for proteins involved in driver events of cell invasion, migration and tumor immunity. In addition, most influential CNAs strongly predicted patient survival and shared known prognostic proteins such as CTNNA2, ARHGDIB, and PACSIN2 [99].

4) RNA-protein decoupling as a surrogate for proteome homeostasis: In medulloblastoma, two independent proteomic data sets revealed a subset of patients with lower than expected across-gene median Spearman correlation compared with the rest of the samples; suggesting that the low correlation group

differed in translation and/or proteostasis. Enrichment analysis underscored translation-related functions and post-translational functions regulating the ubiquitin proteasome system, while the RNA-processing/metabolism gene sets consisted primarily of RNA binding proteins (RBPs) [102,103].

5) Fidelity of protein-protein interaction networks: In HPV-negative head and neck squamous cell carcinoma, although co-expression network analysis at the transcriptome and proteome identified the same functional modules (metabolic pathways and tumor microenvironment), the protein network increased prediction performance for KEGG pathways by above 10% compared with the corresponding mRNA network [125]. In rhabdomyosarcoma, protein-protein interaction networks linked upregulation of G2/M and unfolded protein response pathways to targeted therapy with HSP90 and WEE1 inhibitors [101].

6) Alternative modes of pathway activation: Two different ways of EGFR activation were identified in HPV-negative head and neck squamous cell carcinoma that became evident at the phosphorylation level: EGFR-amplification-driven events that phosphorylated proteins involved in cytoskeleton organization, and EGFR-amplification-independent events that resulted in increased pathway activity correlating with abundances of EGFR ligands rather than EGFR itself. This decoupling of EGFR copy number status from its actual activity argues that EGFR ligand abundances might be a better biomarker for anti-EGFR treatment; a finding revealed exclusively at the post-translational level that would be missed by genomic data [125].

7) Improved survival predictors: In prostate cancer, biomarkers generated from protein abundances were significantly superior in predicting biochemical relapse compared with CNAs, methylation status and mRNA levels [110]. In clear cell renal carcinoma, high tumor grade was associated with cell-cycle regulation and DNA repair at the mRNA level, while increased activity of proteins in the Krebs cycle and the electron transport chain (OXPHOS), and *N*-linked glycosylation were detected exclusively at the protein level [109].

8) Unique subtype discovery with phenocopy effects: In clear cell renal carcinoma, proteome-based subtypes were discovered with immune composition in CD8 T-cells related to metabolism and VEGF signaling pathway [109]. In pediatric brain cancer [120], pediatric craniopharyngiomas co-clustered with BRAF-mutated gliomas at the proteome level. Guilt-by-association principle suggests that a subgroup of craniopharyngioma patients could benefit from MEK/MAPK inhibitors that are currently considered beneficial for the mutated glioma cases.

The above findings suggest that overlaying multiple omics layers with proteomics can reveal unique cancer biology and indicate that while genomic and transcriptomic profiling is being incorporated in clinical decisions for cancer treatment [138], protein-level profiling is needed to ensure genetic events propagate to the cancer phenotype.

2 Research aims

The aim of this thesis was to determine the genotype-to-phenotype relationship in cancer by integrating MS-based proteomics with other high-throughput omics data. The output of these analyses generated new hypotheses about the composition of cancer proteome under the influence of genetic and transcriptomic variation. Specifically, in:

Paper I, we aimed to investigate the effects of copy number alterations on the mRNA and protein levels in breast cancer (BC) tissue samples.

Paper II, we aimed to investigate the effects of aneuploidy on the mRNA and protein levels in B-cell precursor acute lymphoblastic leukemia (BCP-ALL) bone marrow and peripheral blood samples.

Paper III, we aimed to identify proteome-based subtypes in non-small cell lung cancer (NSCLC), and determine their biology in conjunction with orthogonal data for potential clinical use in stratifying patients for treatment.

Paper IV, we aimed to computationally identify proteoforms –peptide-level evidence of central-dogma regulated events– using mass-spectrometry bottom-up proteomics.

3 Materials and methods

3.1 Ethical considerations

Papers I, **II**, and **III** involve handling sample material from patients and ethical approvals were acquired.

In particular, for **Paper I**, tumor and matched normal material were taken from operated breast cancer patients which have provided their written consent upon participating in the study. Ethical approvals were acquired from the regional committee for medical and health research, Regional Ethical Committee Southeast in Norway (approval number 2007.1125, 2016/433).

Paper II involved diagnostic samples from pediatric BCP-ALL cases that had been treated at Skåne University Hospital, Lund, Sweden. Informed consent was obtained, and the study was approved by the Ethics Committee of Lund University.

In **Paper III**, two different patient cohorts (discovery and validation) were conducted.

The discovery cohort comprised material from:

- operable tumors from early-stage lung cancer patients that were surgically treated at the Skåne University Hospital in Lund, Sweden
- biopsy material from inoperable lung cancer.

The study was approved by the Regional Ethical Review Board in Lund, Sweden (registration no. 2004/762 and 2014/32) and all experiments were conducted with patient consent. Information about the study was available for all patients through advertisements in the local news media in the region.

The validation cohort consisted of resected tumor samples from surgically treated lung cancer patients at the Oslo University Hospital in Oslo, Norway from 2006 to 2015 with signed informed consent. The study was approved by the Regional Ethical Committee for Medical and Health Research Ethics, South-East in Oslo, Norway (ref. S-06402b).

No ethical approvals were required for **Paper IV** since it involves commercially available cell lines and the published patient cohort from **paper III**.

All clinical studies adhered to the three fundamental principles of the Declaration of Helsinki regarding human research: respect of person's autonomy by obtaining informed consents, Hippocratic principle of "do no harm" by minimizing adverse outcomes in the medical procedures involved, and fairness in the patient selection by medical criteria without preconceived bias.

3.2 Summary of clinical cohorts and respective cancer types

Paper I, II and **III** are studies of three different cancer types.

3.2.1 Breast cancer

Paper I analyzed material from 45 patients with breast cancer chosen from a larger Norwegian cohort (OSLO2) [139].

Breast cancer is the most commonly diagnosed malignancy in women affecting the terminal lobular units of the collecting duct that are epithelial structures within the breast responsible for milk production during lactation. Histologically, breast cancer can be divided into preinvasive (in situ carcinoma) and invasive (lobular carcinoma and ductal carcinoma no special type). Overall survival has significantly improved over the past years due to an increased understanding of the molecular basis of the disease. Transcriptome-driven approaches have classified breast cancer into distinct intrinsic subtypes: luminal group A and B with estrogen receptor (ER) and/or progesterone receptor (PR) immunohistochemical positivity, HER2 enriched subtype (HER2) with amplification of the respective gene, basal-like tumors that are ER, PR, HER2 negative, and normal-like tumors that are well-differentiated tumors with low proliferation index [140]. Prognostic significance of the intrinsic subtypes has been shown by using an RNA microarray-based 50-gene signature called PAM50 [141]. PAM50 subtypes have been recently profiled by proteomics supporting the transcriptomic data [100,113,135]. Despite the established genetic predisposition of ***BRCA1*** and ***BRCA2*** mutations, breast cancer also demonstrates a rich landscape of copy number alterations with recurrent chromosomal gains and losses of prognostic significance [142,143].

3.2.2 Pediatric B-cell precursor acute lymphoblastic leukemia

Paper II describes a proteogenomic analysis of 27 acute lymphoblastic leukemia (ALL) pediatric patients.

ALL is a malignancy of lymphoblasts –immature lymphocytes and their progenitors– that infiltrate the bone marrow and other lymphoid organs. ALL is the most common pediatric cancer with approximately 60% of all cases occurring in children and adolescents younger than 20 years. ALL cases are classified as B-cell precursor (BCP)-ALL or T-ALL depending on the type of affected B- or T- lymphoblast, with BCP-ALL comprising most cases. Genomics stratifies BCP-ALL into genetic subgroups driven by gene fusions, mutations, and CNAs with prognostic and therapeutic differences. In BCP-ALL, the genetic groups of the fusion gene ETV6/RUNX1 and high hyperdiploidy (>50 chromosomes) are associated with favorable prognosis. The genomic landscape of high

hyperdiploid childhood ALL delineates a specific stable and clonal pattern of trisomies X, 4, 6, 10, 14, 17, and 18, and trisomy or tetrasomy 21, and lack of monosomies [144,145].

3.2.3 Non-small cell lung cancer

Paper III analyzed material from 423 non-small cell lung cancer (NSCLC) patients.

Lung cancer is the most lethal cancer worldwide that arises from proliferating cells of unknown origin that progressively acquire cancer hallmarks to survive in an immune enriched microenvironment [146]. Histologically, lung cancer is divided to small-cell and non-small cell lung cancer, with the latter representing approximately 85% of all new diagnoses. NSCLC is further subdivided into adenocarcinoma, squamous lung cancer, and large cell carcinoma. Survival strongly depends on the tumor stage and the eligibility for surgical resection with 5-year overall survival ranging from 83% for stage IA to 10% for stage IV. Unfortunately, most patients are diagnosed at an advanced stage due to non-specificity of symptoms. Implicated genetic events include translocations of the anaplastic lymphoma kinase gene (***ALK***), gene activating mutations in the epidermal growth factor receptor (***EGFR)***, and mutations of ***TP53***, ***RB1***, ***KRAS*** and ***STK11***. Lung cancer cells foster a tumor-friendly microenvironment by secreting growth factors and avoiding immune recognition. Recent studies have elucidated this druggable tumor-microenvironment interaction [147,148].

3.3 Experimental methods

3.3.1 Mass spectrometry-based proteomics

Proteomics is the quantitative study of the protein complement of cells. A typical bottom-up proteomics experiment starts with protein extraction and sample preparation that digests proteins into peptides with a sequence-specific protease. To make peptide mixtures less complex, prefractionation based on peptides' biophysical properties, such as the isoelectric point (pI) and pH, can be applied [95]. Further separation takes place by high-performance liquid chromatography (HPLC) systems with low flow rates. Mass spectrometers detect the presence and abundance of peptides using the mass and net charge of molecules [149]. Mass spectrometers consist of an ion source, a mass analyzer, and a detector. Initially, electrospray ionization (ESI) produces gaseous ions from peptides in the liquid phase. Then, mass analyzers separate ions by the mass-to-charge ratios (m/z) using the ions' idiosyncratic trajectories in the electrical field. The first mass analyzer is usually accompanied by a 'collision cell' which is another mass analyzer that fragments ions. At the final stage, intact peptide ions or fragment ions enter the detector and produce

spectra called MS1 or precursor ion spectra in the former case, and MS2 or MS/MS or product ion spectra in the latter.

Peptide ions had been typically analyzed by user-defined, rules such as m/z and intensity, to select as many peptides as possible for acquiring MS2 spectra, a method called data-dependent acquisition (DDA) [150]. However, this selection is semi-stochastic as there are more peptides than time needed to analyze them. Instead, in data-independent acquisition (DIA) methods, the mass analyzer selects a window of m/z values capturing more peptides ions at the expense of generating more complex MS2 spectra; relying on subsequent deconvolution usually with the help of a spectral library (**Figure 5**).

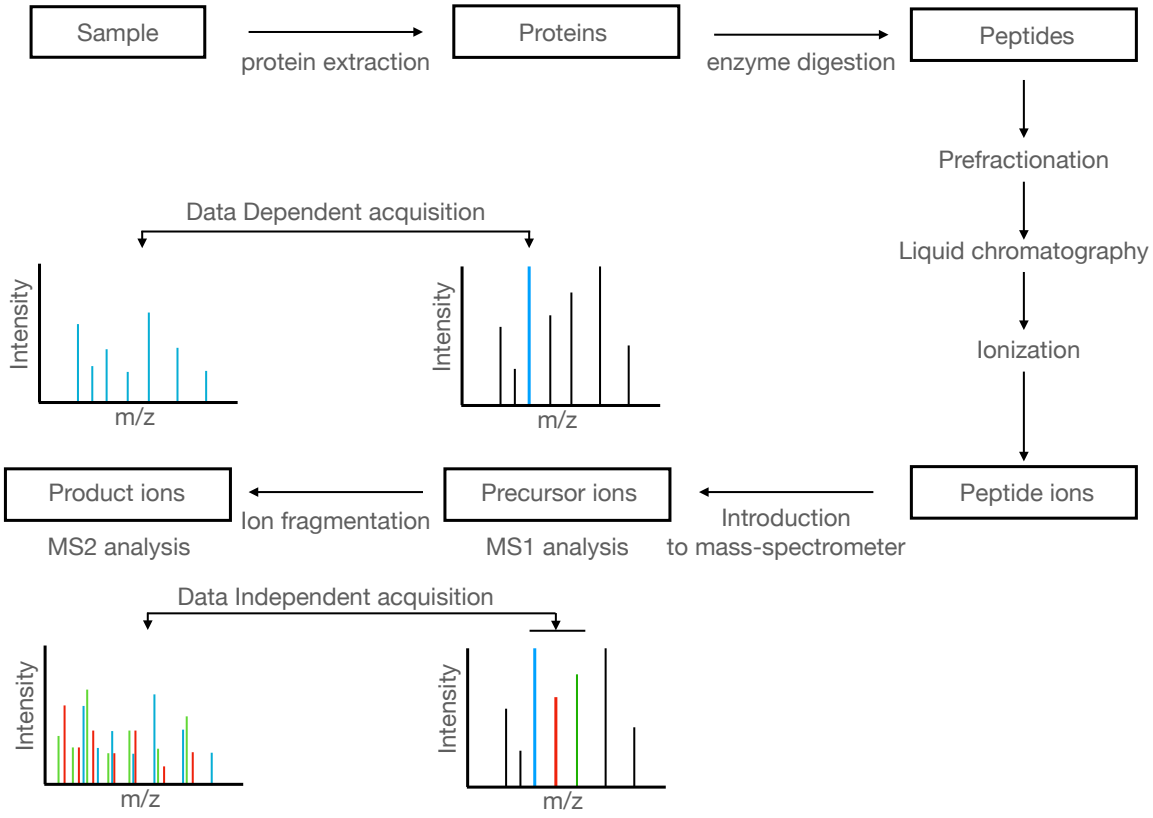

Figure 5. Typical MS/MS proteomics workflow with the two different modes of acquisition

Peptide quantification can be divided into two categories: label-free and label-based [151–153]. In label-free quantification (LFQ), the MS signals of the peptides (usually at the MS1 level) are acquired from the raw data. Despite experimental flexibility, this strategy suffers from quantification variance and is sensitive to instrument performance. Instead, label-based approaches use stable isotopes to produce peptides of similar physicochemical behavior, but with expected mass differences. Metabolic introduction of the isotopes can probe protein dynamics, while chemical labelling–tandem mass tag (TMT) labelling being the most popular method–uncovers the isotope distribution in the tag only after fragmentation. Chemical labelling allows

sample multiplexing with low variance in quantitation at the cost of ratio compression due to peptide co-fragmentation.

Mass spectrometers' final outputs are MS1 and MS2 spectra. Downstream processing involves matching the MS/MS spectra to peptide sequences using a reference database, algorithms to reconstruct proteins from peptides ('protein inference problem') [154], and peptide or protein-level quantification. Peptides are usually attributed to each respective protein by rule of Occam's Razor, which infers the minimum set of proteins that can be explained by the totality of peptides. Recent algorithms have been developed to match peptides to proteins using the quantitative pattern of the peptides in order to detect proteoforms –peptides of unique amino acid sequence and/or post-translational modifications [155,156].

So far described methods for peptide identification rely on reference databases from the canonical proteome which includes proteins that are functional, widely expressed, and conserved across species. Proteomics analysis to detect protein species beyond the canonical proteome using orthogonal genomics information is called proteogenomics and involves three steps: database construction, peptide search and peptide annotation [94].

Papers I, **II**, **III** and **IV** used TMT-labelled DDA LC-MS/MS proteomics coupled with pI-based prefractionation called high-resolution isoelectric focusing (HiRIEF). In **Paper III**, label-free DIA data were generated using HiRIEF and high-pH peptide prefractionation spectral libraries.

Papers I and **III** include a proteogenomics search using the IPAW pipeline [157].

Paper IV describes a tool to address the protein inference problem based on outlying peptide quantification profiles for proteoform inference.

3.3.2 DNA analysis

DNA analysis profiles the genotype –the DNA complement of a sample. DNA analysis can be divided into two broad categories: microarray-based and sequencing-based. In microarray-based analysis, a set of DNA probes are usually bound on a solid surface, to which sample DNA fragments can be hybridized. The probes are generally oligonoucleotides that are 'ink-jet printed' onto slides or synthesized ***in situ***. Labelled single-stranded DNA or antisense RNA fragments from a sample are hybridized to the DNA microarray proportionally to their abundance and later introduced into a scanner for measurement of total signal intensity and B allele frequency (BAF), a measure of allelic content.

DNA sequencing can be performed genome-wide (whole genome sequencing, WGS), exome-wide (whole exome sequencing, WES) or in a targeted way (panel sequencing) for specific DNA loci. Sequencing steps include: 1. Shearing, where DNA of a sample is cut into small pieces of known length; 2. Library preparation by PCR amplification and barcoding of the sheared DNA; 3. DNA sequencing where the prepared library is loaded onto the sequencer. The sequencer obtains sequence information of the fragment to be tested by capturing a fluorescent signal. This signal is then ***in silico*** converted into a sequencing peak.

Both microarray- and sequencing-based technologies are further analyzed downstream to correct for biases.

In **Paper I**, microarrays–Genome-Wide Human SNP Array 6.0 (Affymetrix)–were used for copy number calls. **Paper II** included WGS, WES and SNP arrays. For mutation calling in **Paper III**, targeted sequencing with a custom-designed panel of 370 cancer-related genes was used.

3.3.3 RNA analysis

RNA analysis studies the transcriptome, the total amount of transcripts in a sample. Like DNA analysis, RNA can be profiled with microarray and sequencing methods. The steps are common, except that the starting material can be enriched in mRNA and subsequently converted into complementary DNA (cDNA) before further processing. Downstream analysis detects differentially expressed and/or spliced transcripts between sample groups.

For transcriptome analysis, **Papers I and III** included RNA microarrays: Human Genome Survey Microarray version 2.0 (Applied Biosystem), and Illumina HT12 V4 microarrays (Illumina, San Diego, CA), respectively. **Paper II and IV** analyzed RNA sequencing data (ribosome-depleted and poly-A enriched, and ribosome-depleted, respectively).

3.3.4 Methylation analysis

Methylation analysis interrogates the methylation status of cytosines in the DNA of a sample. Broadly, methylation analysis is divided into microarray-based and sequencing-based methods. Both methods most commonly rely on bisulfite conversion of DNA to detect unmethylated cytosines. Bisulfite conversion changes unmethylated cytosines into uracils that can be identified by allele-specific oligonucleotides against thymines (after amplification) in the case of microarrays, and by sequenced thymine read counts in the case of sequencing.

For methylome analysis, a microarray-based technology–Illumina Infinium HumanMethylation450 BeadChip–was used in **Paper III**.

3.4 Computational methods

3.4.1 Consensus Clustering (CS)

Clustering is an unsupervised method of grouping data. Consensus clustering applies a clustering algorithm of choice in iteratively resampled original data to assess the stability of the discovered groups in terms of sampling variability [158]. Perturbations of the original data by resampling techniques can generate different clustering outcomes, and the agreement ('consensus') among them can be evaluated. A consensus matrix is a matrix that stores, for each pair of items, the proportion of clustering runs in which two items are clustered together. Perfect consensus corresponds to a consensus matrix with all the entries equal to either 0 or 1. Deviation from the perfect consensus matrix indicates a lack of stability of the putative clusters. Thus, for each of a series of cluster numbers ($K = 2, 3, \ldots, K_{max}$), a recipe for assessing the cluster stability requires to: 1) construct a consensus matrix, 2) compare the resultant consensus matrices, and 3) select the cluster number corresponding to the 'cleanest' matrix (i.e., a matrix containing 0's and 1's only). The cleanest matrix can then be clustered on its own for the final group assignment.

We used consensus clustering for subtyping patients in **Paper I** and **Paper III**.

3.4.2 Dimensionality reduction

Dimensionality reduction is a method of visualizing high-dimensional data by projection into a lower number of more intuitive dimensions.

Principal Component Analysis (PCA) reduces data by projecting them onto lower dimensions called principal components (PCs) to find the best representation of the original data using a limited number of PCs. The first PC is chosen to minimize the total distance between the data and their projection onto the PC. By minimizing this distance, the variance of the projected points is maximized. Subsequent PCs are selected in a similar way, with the additional requirement that they be uncorrelated with all previous PCs (orthogonal). Finally, the original data can be reconstructed by linear combinations of PCs with arbitrary (positive or negative) sign [159].

A more intuitive way of projecting data is to define linear combination of projections with positive-only sign. Non-negative matrix factorization (NMF) algorithm extracts such desirable projections from positive matrices [160]. A key parameter, rank ***k*** of the

decomposed matrices, can be used in NMF algorithm to cluster samples. Cluster stability can be assessed by a consensus model selection, similar to that of consensus clustering, that exploits the stochastic nature of the NMF algorithm.

Both previous methods rely on linear projections. Non-linear relationships can be visualized using advanced dimensionality reduction methods such as Unifold Manifold Approximation and Projection (UMAP) [161]. Intuitively, UMAP strives to approximate a low dimensional graph to be as similar as a high-dimensional graph based on the original data. Optimization to minimize the difference of the graph-based distances in the high and low dimensional space leads to better representations that capture both global and local relationships of the original data.

NMF clustering, and PCA coupled to UMAP dimensionality reduction were used in **Paper III**. PCA analysis was used in **Paper IV.**

3.4.3 Louvain Community Detection Algorithm

Louvain community detection algorithm partitions networks into 'meaningful' communities in the sense that nodes are densely connected within the community but sparsely connected with the nodes outside [162]. Quality of the partitions can be assessed by the modularity metric; a scalar value between -1 and 1 that quantifies the density of links inside communities as compared to links between communities. Louvain algorithm optimizes modularity in two phases that are performed iteratively:

- Assuming a network of N nodes, the first phase considers each node as a separate community. Then, for each node we evaluate the increase of modularity if we assign it to the community of its closest neighbor. The node is then assigned in the community for which the increase is maximum and positive. In case of no increase, the node remains in its original community. This process is applied repeatedly and sequentially for all nodes until no further improvement.
- The second phase begins with a new network whose nodes are now the previously found communities in the first phase and the links between new nodes are the sum of the links of the constituent old nodes. Once this second phase is completed, the first phase is iterated until there are no more changes and a maximum of modularity is attained.

An important caveat is that modularity optimization may fail to identify modules smaller than a scale which depends on the total size of the network and on the degree of the modules' interconnectedness (resolution limit).

Louvain community detection algorithm was applied in **Paper III** and **Paper IV.**

4 Results and Discussion

4.1 PAPER I: Breast cancer quantitative proteome and proteogenomic landscape

This paper presents a multi-omics analysis of 45 breast cancer tumor samples, representative of the five PAM50 intrinsic subtypes, with DNA, mRNA, protein, and metabolite quantification. At the protein level, tumors were clustered according to their respective intrinsic subtype. A finer clustering analysis found 6 core tumor clusters that subdivided basal-like and luminal B tumors by their immune composition. Network analysis with immunohistochemical validation pinpointed druggable EGFR and MET co-expression in basal- and normal-like subtypes. Integrative multi-omics analysis identified high correlations between mRNA and proteins of PAM50 genes among genome-wide moderate correlations, and attenuated effects of somatic CNAs on protein abundances relative to mRNA abundances. Furthermore, tumor-specific peptides from novel coding regions and single amino variants were identified in individual patients.

Specific to the thesis, we investigated the in-cis gene regulation across the CNA–mRNA–protein axis using SNP and RNA microarrays, and TMT-labelled MS-based data, respectively. Genomic gains and losses were defined using allele-specific copy number analysis of tumors (ASCAT) tool for threshold estimation (sample-specific ploidy $\pm$ 0.6) [163]. The genome-wide copy number landscape of gains and losses across 41 samples resembled that of previous large cohorts [142,164] **(Figure 6)**. At gene-level resolution, we split samples by the aberrant (gain/loss) and neutral copy number status and called significant CNA-mRNA and CNA-protein associations if they showed differential abundances using a Wilcoxon rank-sum test (Benjamini-Hochberg BH adjusted p-value $\leq$ 0.1, top 5% log2 fold change). Enrichment analysis of the significant CNA associations highlighted the HER2 subtype-driven amplified Chr17q region (***ERBB2*** locus) and the basal subtype-driven loss of hormone signaling. Unexpected fold changes –increased abundances within genomic loss regions– were identified for six genes (***IGFALS***, ***ADAMTS12***, ***KIF20A***, ***MFGE8***, ***MYO1E***, ***CCNB1***) suggesting compensatory routes of gene expression [165,166].

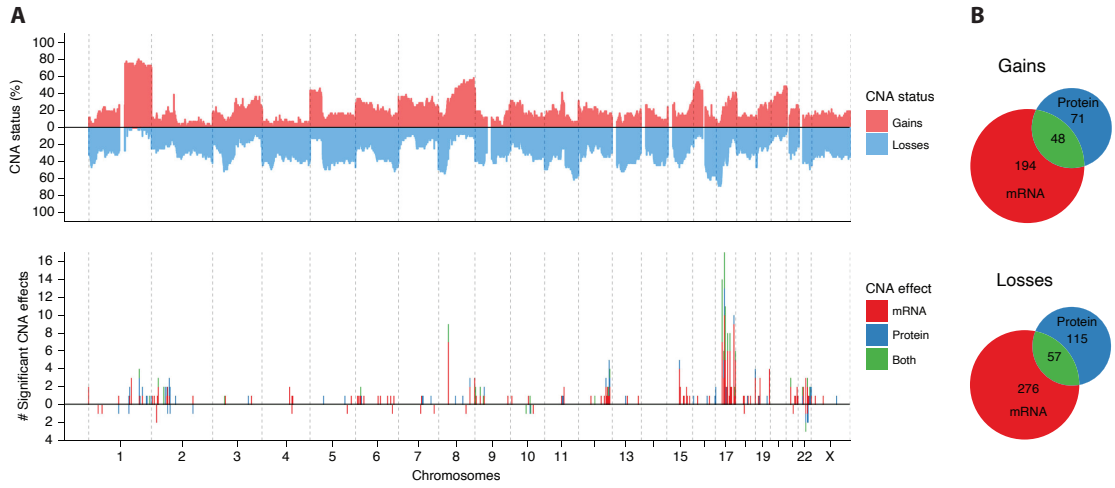

Figure 6. Somatic copy number alternation (CNA) landscape of gains and losses. A. Top, percentage of samples with gains/losses across the genome. Bottom, genome-wide significant mRNA and protein differential abundances for samples with copy number gains. B. Venn diagram of significant differential expressed genes at the mRNA and protein level for copy number gains and losses.

Furthermore, only ~20% of CNA-mRNA associations overlapped with the respective CNA-protein associations. We investigated this attenuated CNA effect on proteins by overlaying published protein ubiquitination data [167] as performed in a previous study [79]. Specifically, we defined genes with high CNA-mRNA and low CNA-protein abundance correlations as being highly attenuated using a Gaussian mixture model with two components **(Figure 7)**. Proteins in the high attenuation group were ubiquitinated significantly more that the rest of the genes across different time-points of drug-induced proteasome inhibition in HCT116 cell lines. Given that specific ubiquitination events target proteins for proteasome-mediated degradation, these data suggest that the discrepancy of CNA effects on mRNA and protein abundances may partly be attributed to protein degradation.

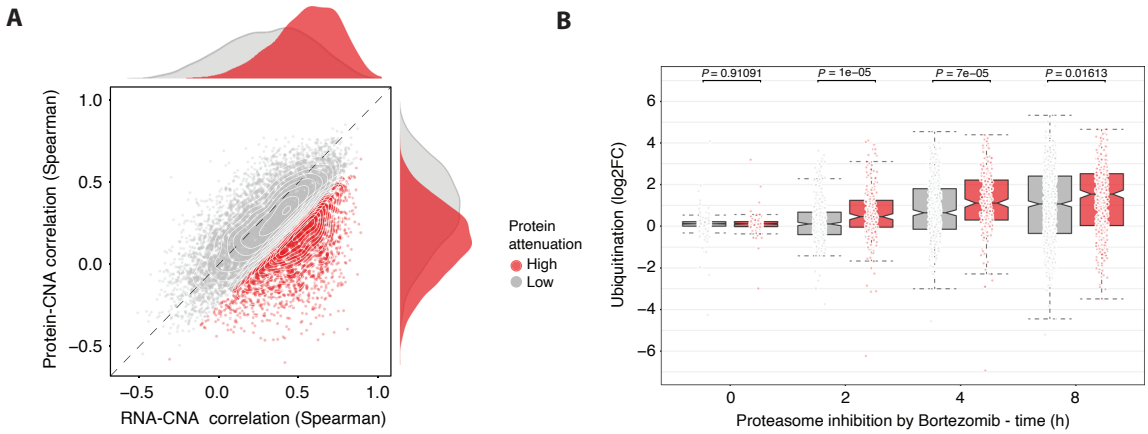

Figure 7. Scatter plot of per-gene CNA-RNA and CNA-protein Spearman correlations to identify CNA effects attenuated at the protein level. B. Boxplot of ubiquitination fold change after proteasome inhibition between bortezomib treated and untreated HCT116 cells.

In summary, the integrative analysis of CNA, mRNA and protein data recapitulated previous breast cancer genomic and proteogenomic landscape studies [100, 113] and attributed the dampened CNA effect along the gene expression axis to protein degradation. Limitations of these analyses include: the small sample size that precluded detection of small effect sizes; investigation of in-cis rather than trans effects; simplified copy number status calls instead of finer levels such as homologous deletions and very high amplifications; focus on protein degradation disregarding alternative, putative overlapping mechanisms of protein abundance compensation.

4.2 PAPER II: Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia

This paper presents a combined proteogenomics and chromatin conformation capture (Hi-C) analysis of primary samples from patients with high hyperdiploid and ETV6/RUNX1-positive pediatric acute lymphoblastic leukemia (ALL). To investigate ploidy effects, high hyperdiploid versus ETV6/RUNX1-positive comparisons at the mRNA and protein level showed reduced abundances of CTCF and cohesin complex subunits. These proteins participate in the hierarchical organization of the genome by establishing topologically associating domains (TADs), i.e., genomic regions that interact with each other much more frequently than with regions located in adjacent sequences. TAD borders were associated with genome-wide transcriptional dysregulation in the hyperdiploid group, for which selected cases displayed loss of TAD boundary strength and reduced insulation in Hi-C assays. At the chromosome level, cytogenetic analysis confirmed aberrant metaphase chromosome morphology in hyperdiploid cases.

Related to the thesis, we studied the in-cis effect of CNAs on mRNA and protein levels. Contrary to breast cancer, which is characterized by both short (focal) and chromosome arm-length copy number alterations, hyperdiploid ALL displays whole chromosome gains. For regions with whole chromosome gains, we calculated differential mRNA and protein abundance using ETV6/RUNX1-positive samples as the non-aberrant control group. A proxy of differential abundance, Cohen's d effect size, scaled linearly with the copy number status (**Figure 8**), with higher values for chromosomes X and 21 as demonstrated before at the mRNA level [145]. Despite the common trend, lower effect sizes were identified at the protein level suggesting protein compensation mechanisms.

We used per-gene mRNA–protein Spearman correlation as a surrogate for protein-level regulation and compared correlation distributions in groups devised by mRNA and protein stability, miRNA targeting, subcellular localization as well as differential expression, differential ubiquitination, and degradation kinetic profiles (**Figure 9**). In particular, we found that differentially expressed genes at the mRNA and protein level between hyperdiploid and ETV6/RUNX1-positive samples had higher mRNA-protein correlation than non-differentially expressed genes suggesting functional significance as previously described [168]. Moreover, in line with the breast cancer study (Paper I), significantly ubiquitinated proteins had lower mRNA-protein correlations likely due to increased protein degradation. This was further illustrated when data of protein degradation kinetics were used to assign proteins into exponential and non-exponential degradation profiles [76]. A part of proteins with non-exponential profiles are considered to belong to excessively produced protein complex subunits that are

degraded faster to maintain complex stoichiometry, and this tighter regulation at the protein level is depicted on their lower mRNA-protein correlations in this study.

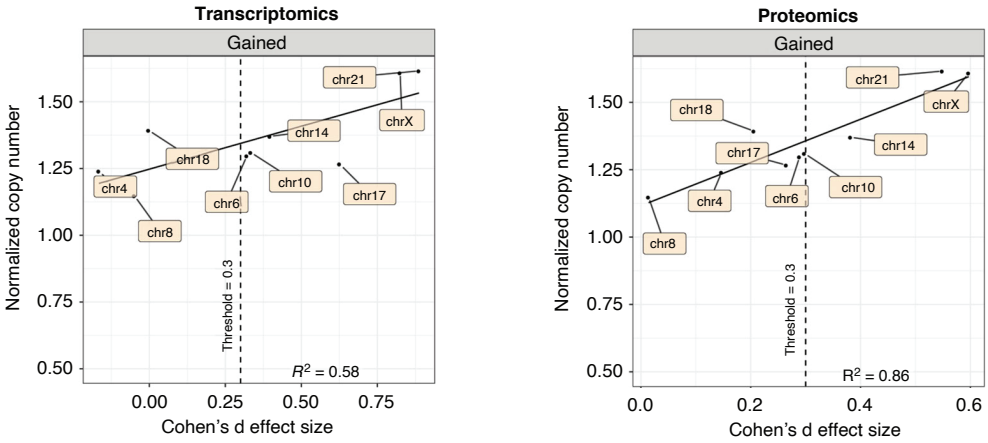

Figure 8. Cohen's d effect size analysis. Comparisons of mRNA and protein levels between high hyperdiploid and ETV6/RUNX1-positive samples were performed for gained chromosomes. Cohen's d effect sizes above 0.3 are called significant. Linear regression coefficient of determination is displayed.

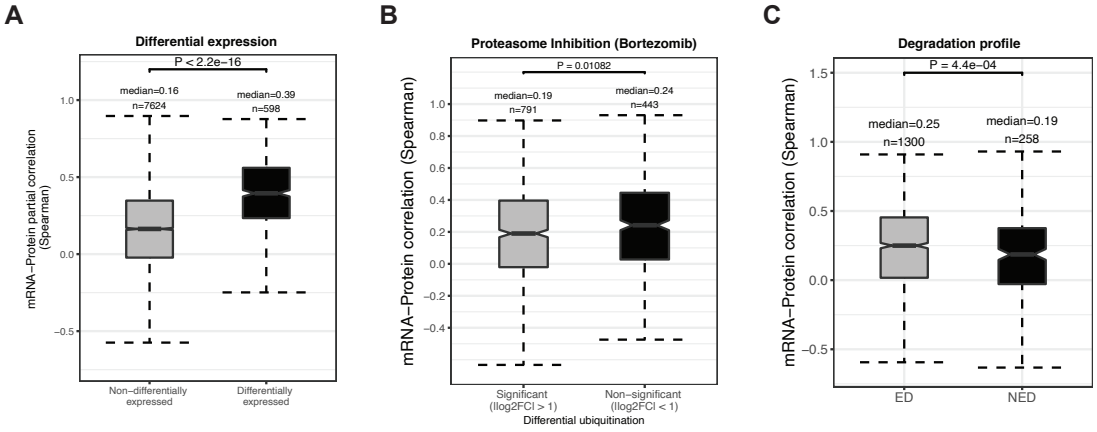

Figure 9. Protein-level regulation. A. Comparison of mRNA-protein correlation distribution for differentially abundant and non-differentially abundant mRNAs and proteins. B. Categorization of mRNA-protein correlation into low (black) and high (grey) ubiquitination of the corresponding protein based on data from bortezomib-induced proteasomal inhibition. C. Comparison of mRNA-protein correlation distribution for proteins with an exponential (ED) and non-exponential degradation (NED) kinetic profile. Number of observations, medians, first and third quartiles, and whiskers extending to 1.5 times the interquartile range are displayed. Unpaired, two-sided Wilcoxon rank sum test was used to calculate P-values.

In summary, CNAs in hyperdiploid ALL samples resulted in increased mRNA and protein production akin to the ploidy status. Concordance between mRNA and protein levels was higher in differentially expressed genes and less ubiquitinated proteins with slower degradation rate. A limitation of this analysis is that these findings describe general, pervasive gene expression regulatory mechanisms without ALL specificity.

4.3 PAPER III: Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune-evasion mechanisms

This paper presents a proteogenomics study of 141 early-stage tumors from non-small cell lung cancer (NSCLC) patients layering data from targeted DNA sequencing, DNA methylation and RNA microarrays, and TMT-labelled proteomics. Clustering analysis identified six subtypes driven by histology and distinct immune microenvironment that was immunohistochemically validated. A mechanistic hypothesis for immune evasion was proposed for the subtype enriched in ***STK11*** mutations regarding the LAG3-FGL1 immune checkpoint receptor-ligand pair. Immune-cold tumors also displayed higher expression of protein species that could serve as potential neoantigens. Finally, using machine learning-based classification, we identified the proposed subtypes in external validation datasets.

In the context of this thesis, we evaluated the unbiased grouping of the NSCLC proteome with unsupervised clustering algorithms. Spearman correlation-based consensus clustering on the overlapping proteome divided NSCLC proteomic samples into six subtypes (**Figure 10**).

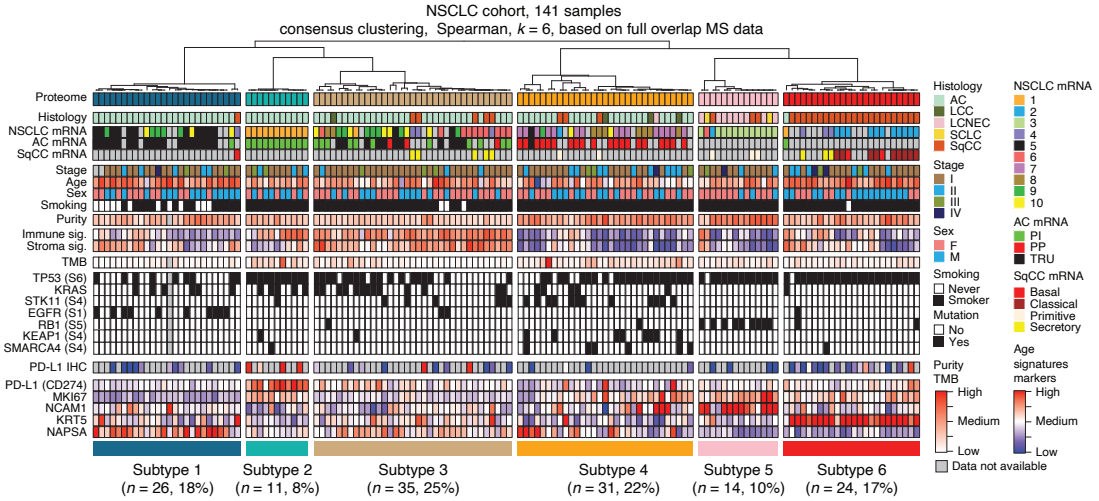

Figure 10. Heatmap of the consensus clustering analysis on the complete proteome of 141 NSCLC samples. Histology, mRNA subtypes, clinical parameters, tumor microenvironment signatures, common mutations and protein levels of selected markers are annotated.

This division was corroborated by an alternative clustering method using non-negative matrix factorization (NMF). For the discovered subtypes, available mutation data uncovered unique links between genotype and molecular phenotypes. Specifically, mutation enrichment was identified for subtype 1 (***EGFR***), subtype 4 (***STK11***, ***KEAP1*** and ***SMARCA1***), subtype 5 (***RB1***), and subtype 6 (***TP53***). In line with the mutations, network analysis indicated increased abundance of proteins involved in epithelial-to-mesenchymal (EMT) transformation, metabolic pathways,

E2F1/MYC signaling, and p53 pathway for the respective subtypes. Moreover, ***in-silico*** deconvolution of the tumor microenvironment identified decreased immune infiltration for most samples in these subtypes. This line of evidence indicates that enriched mutations may shape tumor composition by activating specific downstream oncogenic pathways.

To further characterize tumor microenvironment composition beyond the mutation impact, we investigated the distribution of protein species identified as cancer-testis antigens (CTAs) and non-canonical proteins/peptides (NCPs) across the subtypes. These protein species can be considered as potential neoantigens that allow cancer cells to be targeted by the immune system [169]. CTAs and NCPs were found to be overexpressed in the high-proliferative, low-immune infiltrated subtypes 4, 5 and 6, and were considered as high-scoring in the composite index for tumor neoantigen burden that incorporated the total number of mutations (TMB) (**Figure 11**). Abundances of CTAs and NCPs were negatively correlated with global methylation suggesting an epigenetic component in their expression regulation [170].

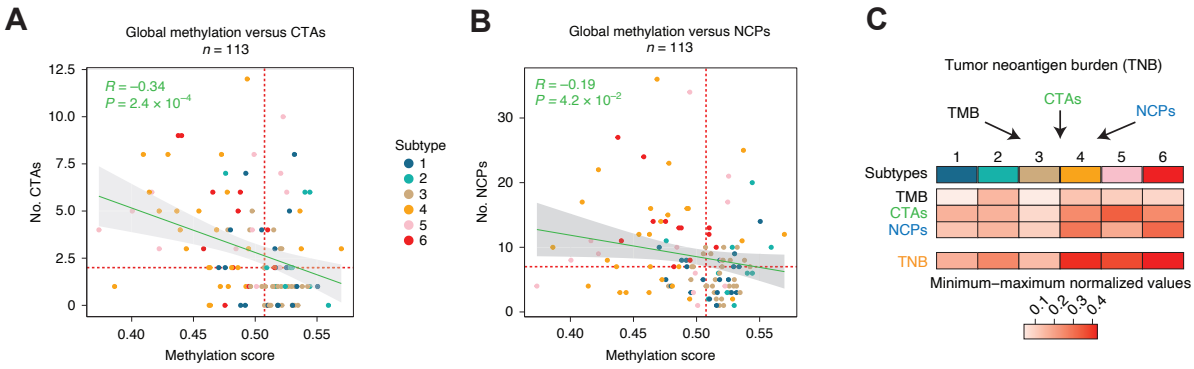

Figure 11. Scatter plot of global methylation with respect to the number of CT antigens per sample (A) and the number of NCPs per sample (B) (n = 113 samples). (C). Heatmap of tumor neoantigen burden per subtype as a composite score of tumor mutational burden (TMB), Cancer testis antigens (CTAs) and non-canonical proteins/peptides (NCPs).

In summary, we identified 6 NSCLC proteome-based subtypes with distinct genotype-phenotype interplay between enriched mutations and oncogenic pathways and tumor microenvironment composition. Subtype-specific non-canonical phenotypes were also linked to the DNA methylation status. These findings need to be further evaluated to determine causal relationships between mutations and downstream effects and avoid confounders of bulk profiling with single-cell technologies. Moreover, the antigenicity of the tumor neoantigen burden should be tested in additional ***in silico*** and ***in vitro*** experiments.

4.4 PAPER IV: DEpMS: Differential Expression analysis of proteoforms from Mass Spectrometry-based bottom-up proteomics

This paper presents a bioinformatics approach to discover proteoforms from bottom-up proteomics datasets. In typical proteomics analysis, the peptide association with its corresponding protein is lost during the protein digestion step. At the protein inference step, detected peptides are summarized under a common protein identifier using the minimum list of proteins that support all peptide evidence. By doing this, important information is lost for proteoforms, which comprise the complement of molecular forms a gene's protein product can take. A solution to overcome this limitation and detect putative proteoforms is to use the peptide quantitative pattern –the vector of peptide abundances across samples– before the protein inference step [155,156,171]. Here, we build on this idea by developing a tool that selects peptides with outlying quantitative patterns across samples, groups them by their similarity, and finally, tests them between conditions or across their shared co-variance pattern. We call this method Detection of proteoforms from Mass Spectrometry-based bottom-up proteomics (DEpMS).

We applied DEpMS in a TMT-labelled MS proteomics dataset of 18 cell lines profiled in triplicates with diverse embryonal origin to increase the proteoform diversity (ABMS, AntiBody validation using Mass Spectrometry dataset) (**Figure 12**). Group-based DepMS from pairwise comparisons between cell lines identified 768 putative proteoforms that corresponded to 717 proteins ($q \leq 0.01$). The majority of them were cell line specific (71.5%) with most proteins consisting of 2 proteoforms (93.4%). Splicing analysis of corresponding RNA sequencing data with the RMATS statistical tool [172] identified hundreds of splicing events predominantly of the exon skipping type (Benjamini-Hochberg BH p-value ≤ 0.01, $|\Delta\psi| > 0.1$). Common events at the RNA and protein level were found for 55 unique proteins (7.7% of the total proteoforms). This indicates that RNA splicing events only partially explain the diversity of the identified proteoforms, suggesting alternative mechanisms of proteoform generation [173], technical artifacts [174] or gene expression regulation events [175].

We further applied DEpMS in the clinical cohort of non-small cell lung cancer patients described in Paper III. Unsupervised Principal component analysis (PCA) of the outlying peptides using the covariance pattern imprinted in the first two principal components separated squamous cell carcinoma (subtype 6) from the EGFR mutant-enriched adenocarcinoma lung cancer (subtype 1), and nominated 73 proteoforms from 72 unique proteins ($q \leq 0.01$). All of them except one case were identified in separate principal components. Highlighted in **Figure 13** is an identified putative proteoform from collagen type VI alpha 3 chain (COL6A3) protein with the quantitative profile of its three associated peptides shown on top of the heatmap

(left) and the corresponding gene track (right). The peptides map to exon 6 of ***COL6A3***, which has been recently shown to be a tumor-stroma-specific splice variant in multiple solid tumors in quantitative immunopeptidomics[176]. The analysis here suggests that a subgroup of ***EGFR***-mutated lung cancer adenocarcinoma patients may not be eligible for T-cell receptor-based immunotherapy due to low protein abundance of the ***COL6A3*** splice variant.

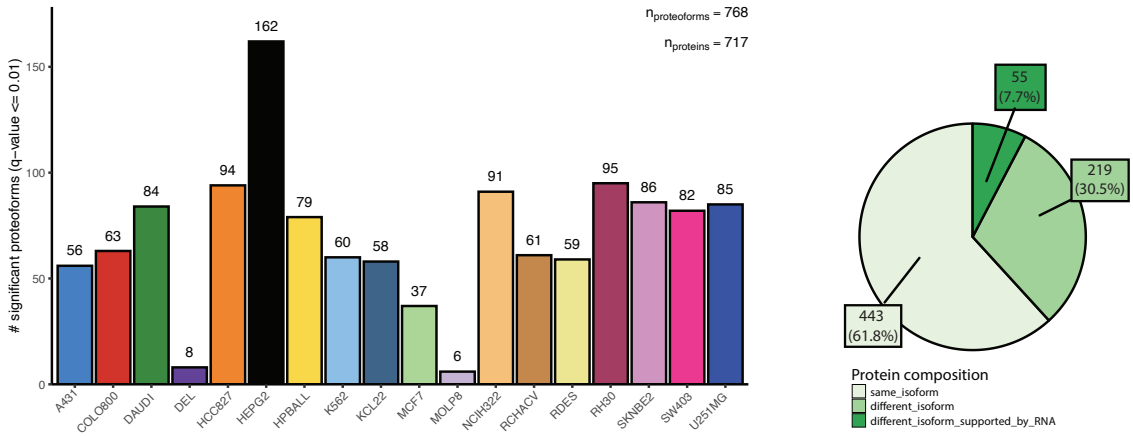

Figure 12. DEpMS analysis on ABMS data. Left, bar plot of the identified proteoforms per cell line (q ≤0.01). Right, pie chart of protein composition in significant proteoforms. Proteoforms are assigned to an isoform group if the constituent peptides support alternative gene models with RNA evidence.

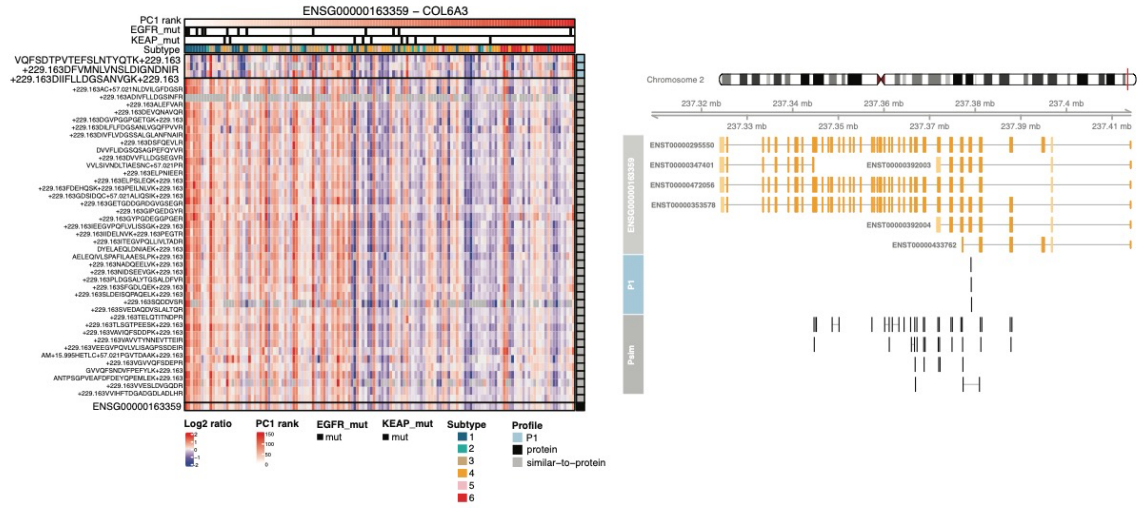

Figure 13. COL6A3 proteoform detection from PCA-based DEpMS on NSCLC data. Left, heatmap of COL6A3 peptides with the top three corresponding to a distinct proteoform (Profile P1).Right, gene track with the COL6A3 gene models and the P1 proteoform peptides overlapping splice variant-specific exon 6.

In summary, this study describes a tool for nominating proteoforms from MS-based bottom-up proteomics. Since the added proteoforms expand the identifications at the protein level and rely on outlying peptide profiles that could be generated from unreliable spectra, false discoveries are enriched. Orthogonal data or post-processing inspection may be needed to assess the quality of the proteoforms [177]

5 Conclusions

In summary, herein presented work has investigated how cancer proteome integrates changes along the axis of the central dogma of molecular biology. Specifically:

In **Paper I**, we studied the effect of CNAs on mRNA and protein abundances in breast cancer and found antithetic gene expression regulation: pervasive expression of the breast cancer driver gene ***ERBB2*** in the amplified Chr17q region, but genome-wide dampened abundances for proteins targeted by the proteasome.

In **Paper II**, we studied the gene expression regulation via post-transcriptional and post-translational mechanisms in the aneuploidy setting of B-cell high hyperdiploid childhood acute lymphoblastic leukemia. We found concordant dosage effects of copy number gained chromosomes at both mRNA and protein level, but mRNA-protein deregulation due to differential ubiquitination and protein degradation kinetic profile.

In **Paper III**, we investigated proteome composition and organization in NSCLC tumor samples that indicated six distinct immune microenvironment-related subtypes. The six subtypes displayed unique mutation enrichments and novel peptide expression associated with DNA hypomethylation.

In Paper IV, we devised a bionformatics approach to dissect proteoform diversity from bottom-up proteomics datasets. We found that splice-variants represent ~8% of the total pool of proteoform identifications and showcase COL3A6 splice variant as an example of a proof-of-concept proteoform discovery in a NSCLC clinical proteomics cohort.

6 Points of perspective

Proteogenomic studies have showed that mRNA measurements are at least moderate predictors of proteins and have moved the so far genomics-driven cancer biology field forward by uncovering new regulatory relationships between DNA, RNA and proteins. These findings can lead to novel cancer subtypes, potential biomarkers and druggable targets, but need to be validated in larger cohorts for clinical implementation. Thus, current clinical proteogenomic studies should be cautiously interpreted as hypothesis generating studies.

Overcoming limitations in all aspects of MS-based experiments [178] –sample preparation, liquid chromatography, mass spectrometer instrumentation and downstream analysis– will enable faster, more accurate and in-depth proteomics output at a cohort scale similar to plasma proteomics [179]. Of particular interest is the problem of downstream analysis with unassigned spectra that might stem from post-translationally modified peptides that are currently missed without modified database searches. Large consortia based on top-down proteomics that circumvent this problem by analyzing intact proteins will hold a pivotal role in the future for illuminating the human proteome diversity [173].

With an eye on advances in RNA sequencing technologies and the way these have progressed from bulk samples to single-cell analysis, one may envisage that future proteogenomics studies would follow a similar way. Already single-cell proteomics technologies have been developed that enable the proteome-wide profiling of proteins [180] and team efforts are currently underway to study the spatiotemporal variation of the human proteome at the single cell level [181]. This could reveal exciting insights into tumor evolution across time and space including mechanisms of metastasis and drug resistance. In the near future, one might as well imagine single-cell methods for tracking protein degradation be combined with genomic engineering methods to probe the direct effect of chromosomal aberrations on the proteome, unbiased from bulk measurements [182].

From the bioinformatic perspective, data integration could move proteogenomic analyses beyond parsimonious linear correlations by analyzing data in an multi-dimensional way accustomed to the omics at hand [183,184]. Among various sophisticated approaches, analysis based on graph neural networks seems the most promising for the proteogenomics field since gene regulation can naturally be cast as a graph of molecular interactions [185]. Such machine learning applications, though, should be used with caution avoiding biases of model misspecification and overfitting [186].

The potential of current and future proteogenomic studies for clinical translation seems particularly promising [187]. Proteogenomic findings could help bridge genomics with proteomics for personalized medicine [188] building upon the unique characteristics of proteins as integrators of upstream events in the central dogma and mediators of pharmacological actions. Within the drug therapy field, due to being closer to cancer phenotype, more rational drug combinations could be chosen to alleviate drug resistance or produce synergistic effects. Most importantly, identification of sample-specific non-canonical peptides/proteins that are able to elicit immune responses could increase immunotherapeutic options with immune checkpoint inhibitors, neoantigen vaccines or engineered T-cells [189,190]. Finally, diagnostic, predictive and prognostic cancer biomarkers could be developed and validated by the high quantification accuracy of targeted MS proteomics [191].

In this thesis, proteogenomics investigated the interplay between DNA, mRNA and proteins in cancer suggesting proteomes of unique composition in protein species, sculpted by gene compensation mechanisms, and integrated as tumor-organ phenotypes. This body of evidence can help generate compelling hypotheses about tumor biology with clinical translational potential.

7 Acknowledgements

Without question, being a PhD student has been an exciting period in my life. So exciting that I tried to live it to the fullest (aka longest ☺). Six and a half years after the start of the PhD studies, I feel honored and grateful to have contributed to such interesting projects together with so many talented people. The least I can do is to thank you here.

First, I would like to thank my main supervisor, **Janne Lehtiö**, for recruiting me to his group and introducing me to proteomics and central dogma regulation. Thank you for being eager to hear my thoughts and concerns, for always being supportive (financially too) and, particularly, for willing to share your 'beast mode' of scientific thinking when conceptualizing novel ideas like the residual analysis of mRNA-to-protein ratios, despite a busy schedule.

Second, I would like to thank my co-supervisor, **Henrik Johansson**, for letting me take part in the breast cancer project and for giving constructive feedback on the copy number analysis in the early days under Erik's guidance, and later, on the proteoform project. Paying attention to the detail and critically evaluating the results helped me evolve as a scientist. Thank you for being understanding, calm and reassuring when things did not go as planned and for letting me try my own ideas.

I would also like to thank **Mattias Vesterlund** for all the help and guidance with the leukemia projects, always having a clear plan about how to proceed and open to new suggestions.

A huge thanks goes to **Lukas Örre** for having me in the lung cancer project. Thank you for all the two-hour-and-more-long meetings with Taner in which we exchanged results and discussed how to move further. Those Thursday afternoons were legendary!

And a great thanks to **Rozbeh Jafari** for giving me the opportunity to work in the ALL cell line project with the awesome members of the ALL team. A special thank you for accepting to be my chairperson, as well.

I would also like to thank all the collaborators in Lund and Oslo for making the projects happen.

Also, thank you **Giorgos** for the phosphoproteome explanations, the fun atmosphere, the jokes that are not to be translated, and for letting me be the bell ringer for your departure every evening. Thank you **Jorrit** for the help in solving computer problems of byzantine complexity to me and for all the fun in the lab. Your humor cannot be caged. Thank you **David** for all the help with the lung cancer sequencing results and conversations about (and beyond) genomics. Special thanks to **Matthias Stahl** for letting me continue the proteoform project. Thank you **Aida**, a true problem solver and meme inspiration in the early days. A great thanks to **Helena** for always willing to help

and care. And thank you **Eduardo** for including me in the peritoneal dialysis project– PDE still echoes in my ears. Thank you **Fabio** for being the best of neighbors and showing me pathology slides from time to time. Thanks a ton for the help in the PhD defense application process, too. Thank you **Yi** for including me in the interesting sarcoma project and for the nice company during the Christmas holidays last year. Hope you and Fabio are having a great time in the States.

Thank you **Taner** for the company during lectures and projects in the lab and in life outside the lab. Studying/working with you was so fun and exciting. I hope working in the industry now is equally satisfying. And thank you **Olena** for all the support during the past months, but also for all the funny memes and the movie nights with Taner.

A thanks that cannot be contained in the written words of these acknowledgements goes to **Haris.** Thank you for all the lunches and dinners, the weird discussions and jokes, the time spending together inside and outside the lab. Above all, thank you for being so supportive, caring, and uplifting. Thank you for proofreading the kappa, as well.

And thanks to the rest of the Lehtiö lab people, past and present members, **Rui**, **Maria**, **Mann**, **Markus**, **Ali**, **Ann-Sofi**, **Sebastian**, **Isabelle**, **Luay**, **Nidhi**, **Noora**, **Yaroslav, Yanbo** and **Xiaofang (**with your lovely cat**).** Always great to have you around**.**

I would also like to thank my friend **Spyridoula** for the enthusiastic discussions about Sweden back in the long gone 2012, and for spending time together when I first arrived in Stockholm. This early period in Sweden would have been much less fun without her.

A special thanks to my mentor **Anita Göndör** for her support and the discussion about career prospects that helped me decide what to pursue further in the future.

Huge thanks to my friends in Greece, **Lefteris, Maria** and little **Anthia** for being in my life.

Finally, I want to thank my family for their support all these years. This thesis is dedicated to them.

8 References

1. Burgers PMJ, Kunkel TA. Eukaryotic DNA Replication Fork. Annual Review of Biochemistry. 2017;86(1):417–38.

2. Cramer P. Organization and regulation of gene transcription. Nature. 2019;573(7772):45–54.

3. Sonenberg N, Hinnebusch AG. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell. 2009;136(4):731–45.

4. Roll-Hansen N. Sources of Wilhelm Johannsen's Genotype Theory. Journal of the History of Biology. 2008 Nov;42(3):457–93.

5. Nussinov R, Tsai CJ, Jang H. Protein ensembles link genotype to phenotype. PLOS Computational Biology. 2019;15(6):e1006648.

6. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science. 2001;291(5507):1304–51.

7. Consortium IHGS. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931–45.

8. Wilhelm M, Schlegl J, Hahne H, Gholami AM, Lieberenz M, Savitski MM, et al. Mass-spectrometry-based draft of the human proteome. Nature. 2014;509(7502):582–7.

9. Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, et al. A draft map of the human proteome. Nature. 2014;509(7502):575–81.

10. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science. 1989;246(4926):64–71.

11. Hu Q, Noll RJ, Li H, Makarov A, Hardman M, Cooks RG. The Orbitrap: a new mass spectrometer. Journal of Mass Spectrometry. 2005;40(4):430–43.

12. Ong SE, Mann M. Mass spectrometry–based proteomics turns quantitative. Nature Chemical Biology. 2005;1(5):252–62.

13. Corbett AH. Post-transcriptional regulation of gene expression and human disease. Current Opinion in Cell Biology. 2018 Jun;52:96–104.

14. Barbieri I, Kouzarides T. Role of RNA modifications in cancer. Nature Reviews Cancer. 2020 Apr;1–20.

15. Genuth NR, Barna M. Heterogeneity and specialized functions of translation machinery: from genes to organisms. Nature Reviews Genetics. 2018 Jun;19(7):1–22.

16. Robichaud N, Sonenberg N, Ruggero D, Schneider RJ. Translational Control in Cancer. Cold Spring Harbor Perspectives in Biology. 2019 Jul;11(7):a032896–17.

17. Labbadia J, Morimoto RI. The biology of proteostasis in aging and disease. Annual Review of Biochemistry. 2015;84(1):435–64.

18. Eling N, Morgan MD, Marioni JC. Challenges in measuring and understanding biological noise. Nature Reviews Genetics. 2019;20(9):1–13.

19. Dill KA, Ghosh K, Schmit JD. Physical limits of cells and proteomes. Proceedings of the National Academy of Sciences. 2011 Nov;108(44):17876–82.

20. Marguerat S, Bähler J. Coordinating genome expression with cell size. Trends in genetic. 2012 Nov;28(11):560–5.

21. Guang MHZ, Kavanagh EL, Dunne LP, Dowling P, Zhang L, Lindsay S, et al. Targeting Proteotoxic Stress in Cancer: A Review of the Role that Protein Quality Control Pathways Play in Oncogenesis. Cancers. 2019 Jan;11(1):66–21.

22. Takahashi JS. Transcriptional architecture of the mammalian circadian clock. Nature Reviews Genetics. 2017;18(3):1–16.

23. Calabrese C, Davidson NR, lu DD x0011F, Fonseca NA, He Y, Kahles A x000E9, et al. Genomic basis for RNA alterations in cancer. Nature. 2020 Jan;1–50.

24. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature Reviews Genetics. 2012;13(4):227–32.

25. Liu Y, Beyer A, Aebersold R. On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell. 2016 Apr;165(3):535–50.

26. Buccitelli C, Selbach M. mRNAs, proteins and the emerging principles of gene expression control. Nature Reviews Genetics. 2020;204:1–15.

27. Magnusson R, Rundquist O, Kim MJ, Hellberg S, Na CH, Benson M, et al. RNA-Sequencing And Mass-Spectrometry Proteomic Time-Series Analysis of T-Cell Differentiation Identified Multiple Splice Variants Models That Predicted Validated Protein Biomarkers In Inflammatory Diseases. Frontiers Mol Biosci. 2021;9:916128.

28. Protein velocity and acceleration from single-cell multiomics experiments. Genome Biology. 2020 Feb 17;21(1):1–6.

29. Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between Protein and mRNA Abundance in Yeast. Molecular and Cellular Biology. 1999;19(3):1720–30.

30. Tian Q, Stepaniants SB, Mao M, Weng L, Feetham MC, Doyle MJ, et al. Integrated Genomic and Proteomic Analyses of Gene Expression in Mammalian Cells. Molecular & Cellular Proteomics. 2004 Oct;3(10):960–9.

31. de Godoy LMF, Olsen JV, Cox J, Nielsen ML, Hubner NC, Fröhlich F, et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature. 2008 Sep;455(7217):1251–4.

32. Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, et al. Defining the transcriptome and proteome in three functionally different human cell lines. Molecular Systems Biology. 2010 Dec;6(1):1–9.

33. Vogel C, de Sousa Abreu R, Ko D, Le SY, Shapiro BA, Burns SC, et al. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. Molecular Systems Biology. 2010 Aug;6:1–9.

34. Nagaraj N, Wiśniewski JR, Geiger T, Cox J, Kircher M, Kelso J, et al. Deep proteome and transcriptome mapping of a human cancer cell line. Molecular Systems Biology. 2011 Nov;7(1):1–8.

35. Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature. 2011 May;473(7347):337–42.

36. Li JJ, Bickel PJ, Biggin MD. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. PeerJ. 2014;2:e270-26.

37. Franks A, Airoldi E, Slavov N. Post-transcriptional regulation across human tissues. PLOS Computational Biology. 2017 May;13(5):e1005535-20.

38. Fortelny N, Overall CM, Pavlidis P, Freue GVC. Can we predict protein from mRNA levels? Nature. 2017 Jul;547(7664):E19–20.

39. Gry M, Rimini R, Strömberg S, Asplund A, Pontén F, Uhlen M, et al. Correlations between RNA and protein expression profiles in 23 human cell lines. BMC Genomics. 2009;10(1):365–14.

40. Foss EJ, Radulovic D, Shaffer SA, Ruderfer DM, Bedalov A, Goodlett DR, et al. Genetic basis of proteome variation in yeast. Nature Genetics. 2007;39(11):1369–75.

41. Fu J, Keurentjes JJB, Bouwmeester H, America T, Verstappen FWA, Ward JL, et al. System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nature Genetics. 2009 Jan;41(2):166–7.

42. Skelly DA, Merrihew GE, Riffle M, Connelly CF, Kerr EO, Johansson M, et al. Integrative phenomics reveals insight into the structure of phenotypic diversity in budding yeast. Genome Research. 2013 Sep;23(9):1496–504.

43. Wu L, Candille SI, Choi Y, Xie D, Jiang L, Li-Pook-Than J, et al. Variation and genetic control of protein abundance in humans. Nature. 2013 Jul;499(7456):79–82.

44. Wu Y, Williams EG, Dubuis S, Mottis A, Jovaisaite V, Houten SM, et al. Multilayered Genetic and Omics Dissection of Mitochondrial Activity in a Mouse Reference Population. Cell. 2014 Sep;158(6):1415–30.

45. Battle A, Khan Z, Wang SH, Mitrano A, Ford MJ, Pritchard JK, et al. Impact of regulatory variation from RNA to protein. Science. 2015;347(6222):664–7.

46. Chick JM, Munger SC, Simecek P, Huttlin EL, Choi K, Gatti DM, et al. Defining the consequences of genetic variation on a proteome-wide scale. Nature. 2016 Jun;534(7608):500–5.

47. Mirauta BA, Seaton DD, Bensaddek D, Brenes A, Bonder MJ, Kilpinen H, et al. Population-scale proteome variation in human induced pluripotent stem cells. Elife. 2020;9:e57390.

48. Emilsson V, Ilkov M, Lamb JR, Finkel N, Gudmundsson EF, Pitts R, et al. Co-regulatory networks of human serum proteins link genetics to disease. Science. 2018;361(6404):769–73.

49. Blein-Nicolas M, Negro SS, Balliau T, Welcker C, Cabrera-Bosquet L, Nicolas SD, et al. A systems genetics approach reveals environment-dependent associations between SNPs, protein coexpression, and drought-related traits in maize. Genome Res. 2020;30(11):1593–604.

50. Grossbach J, Gillet L, Clément-Ziza M, Schmalohr CL, Schubert OT, Schütter M, et al. The impact of genomic variation on protein phosphorylation states and regulatory networks. Mol Syst Biol. 2022;18(5):e10712.

51. Lee MV, Topper SE, Hubler SL, Hose J, Wenger CD, Coon JJ, et al. A dynamic model of proteome changes reveals new roles for transcript alteration in yeast. Molecular Systems Biology. 2011;7(1):514.

52. Vogel C, Silva GM, Marcotte EM. Protein Expression Regulation under Oxidative Stress. Molecular & Cellular Proteomics. 2011 Dec;10(12):M111.009217-12.

53. Lackner DH, Schmidt MW, Wu S, Wolf DA, Bähler J. Regulation of transcriptome, translation, and proteome in response to environmental stress in fission yeast. Genome Biology. 2012 Apr;13(4):R25-14.

54. Jovanovic M, Rooney MS, Mertins P, Przybylski D, Chevrier N, Satija R, et al. Dynamic profiling of the protein life cycle in response to pathogens. Science. 2015 Mar;347(6226):1259038–1259038.

55. Cheng Z, Teo G, Krueger S, Rock TM, Koh HWL, Choi H, et al. Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. Molecular Systems Biology. 2016 Jan;12(1):855.

56. van den Berg PR, Budnik B, Slavov N, Semrau S. Dynamic post-transcriptional regulation during embryonic stem cell differentiation. bioRxiv. 2017 Mar;1–35.

57. Liu TY, Huang HH, Wheeler D, Xu Y, Wells JA, Song YS, et al. Time-Resolved Proteomics Extends Ribosome Profiling-Based Measurements of Protein Synthesis Dynamics. Cell Systems. 2017 Jun;4(6):636-644.e9.

58. Rendleman J, Cheng Z, Maity S, Kastelic N, Munschauer M, Allgoewer K, et al. New insights into the cellular temporal response to proteostatic stress. Elife. 2018 Oct 12;7.

59. Eisenberg AR, Higdon A, Keskin A, Hodapp S, Jovanovic M, Brar GA. Precise Post-translational Tuning Occurs for Most Protein Complex Components during Meiosis. Cell reports. 2018 Dec;25(13):3603-3616.e3.

60. Rendleman J, Cheng Z, Maity S, Kastelic N, Munschauer M, Allgoewer K, et al. New insights into the cellular temporal response to proteostatic stress. Elife. 2018;7:e39054.

61. Genshaft AS, Li S, Gallant CJ, Darmanis S, Prakadan SM, Ziegler CGK, et al. Multiplexed, targeted profiling of single-cell proteomes and transcriptomes in a single reaction. Genome Biology. 2016 Sep;1–15.

62. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nature biotechnology. 2017 Aug;35(10):936–9.

63. Tarbier M, Mackowiak SD, Frade J, Catuara-Solarz S, Biryukova I, Gelali E, et al. Nuclear gene proximity and protein interactions shape transcript covariations in mammalian single cells. Nature Communications. 2020;11(1):5445.

64. Jarnuczak AF, Najgebauer H, Barzine M, Kundu DJ, Ghavidel F, Perez-Riverol Y, et al. An integrated landscape of protein expression in human cancer. Scientific Data. 2021;8(1):115.

65. Moulana A, Scanteianu A, Jones D, Stern AD, Bouhaddou M, Birtwistle M. Gene-Specific Predictability of Protein Levels from mRNA Data in Humans. bioRxiv. 2018 Aug;1–39.

66. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000 Jan;100(1):57–70.

67. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4;144(5):646–74.

68. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discovery. 2022;12(1):31–46.

69. Holland AJ, Cleveland DW. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nat Rev Mol Cell Biol. 2009 Jul;10(7):478–87.

70. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010 Feb;463(7283):899–905.

71. Ben-David U, Amon A. Context is everything: aneuploidy in cancer. Nature Reviews Genetics. 2019 Sep;21(1):1–19.

72. El-Brolosy MA, Kontarakis Z, Rossi A, Kuenne C, Günther S, Fukuda N, et al. Genetic compensation triggered by mutant mRNA degradation. Nature. 2019 Mar 30;568(7751):193–7.

73. Moriya H. Quantitative nature of overexpression experiments. Molecular biology of the cell. 2015 Nov;26(22):3932–9.

74. Dephoure N, Hwang S, O'Sullivan C, Dodgson SE, Gygi SP, Amon A, et al. Quantitative proteomic analysis reveals posttranslational responses to aneuploidy in yeast. Elife. 2014 Jul;3:36–27.

75. Li GW, Burkhardt D, Gross C, Weissman JS. Quantifying Absolute Protein Synthesis Rates Reveals Principles Underlying Allocation of Cellular Resources. Cell. 2014 Apr;157(3):624–35.

76. McShane E, Sin C, Zauber H, Wells JN, Donnelly N, Wang X, et al. Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. Cell. 2016;167(3):803-815.e21.

77. Taggart JC, Zauber H, Selbach M, Li GW, McShane E. Keeping the Proportions of Protein Complex Components in Check. Cell Systems. 2020 Feb;10(2):125–32.

78. Liu Y, Borel C, Li L, Müller T, Williams EG, Germain PL, et al. Systematic proteome and proteostasis profiling in human Trisomy 21 fibroblast cells. Nature Communications. 2017 Oct;8(1):1–15.

79. Goncalves E, Fragoulis A, Garcia-Alonso L, Cramer T, Saez-Rodriguez J, Beltrao P. Widespread Post-transcriptional Attenuation of Genomic Copy-Number Variation in Cancer. Cell Systems. 2017 Oct;5(4):386-398.e4.

80. Cheng P, Zhao X, Katsnelson L, Camacho-Hernandez EM, Mermerian A, Mays JC, et al. Proteogenomic analysis of cancer aneuploidy and normal tissues reveals divergent modes of gene regulation across cellular pathways. eLife. 2022;11:e75227.

81. Zhu J, Tsai HJ, Gordon MR, Li R. Cellular Stress Associated with Aneuploidy. Developmental Cell. 2018 Feb;44(4):420–31.

82. Wolff S, Weissman JS, Dillin A. Differential Scales of Protein Quality Control. Cell. 2014 Mar;157(1):52–64.

83. Harper JW, Bennett EJ. Proteome complexity and the forces that drive proteome imbalance. Nature. 2016 Sep;537(7620):328–38.

84. Joshi S, Wang T, Araujo TLS, Sharma S, Brodsky JL, Chiosis G. Adapting to stress - chaperome networks in cancer. Nature Reviews Cancer. 2018 Sep;18(9):562–75.

85. Rodina A, Wang T, Yan P, Gomes ED, Dunphy MPS, Pillarsetty N, et al. The epichaperome is an integrated chaperome network that facilitates tumour survival. Nature. 2016 Oct;538(7625):397–401.

86. Mészáros B, Kumar M, Gibson TJ, Uyar B, Dosztányi Z. Degrons in cancer. Science signaling. 2017 Mar;10(470):eaak9982.

87. Martínez-Jiménez F, Muinos F, López-Arribillaga E, Lopez-Bigas N, Gonzalez-Perez A. Systematic analysis of alterations in the ubiquitin proteolysis system reveals its contribution to driver mutations in cancer. Nature Cancer. 2020;1(1):1–24.

88. Levine B, Kroemer G. Biological Functions of Autophagy Genes: A Disease Perspective. Cell. 2019 Jan;176(1–2):11–42.

89. White E. The role for autophagy in cancer. The Journal of clinical investigation. 2015 Jan;125(1):42–6.

90. Bazzaro M, Lin Z, Santillan A, Lee MK, Wang MC, Chan KC, et al. Ubiquitin Proteasome System Stress Underlies Synergistic Killing of Ovarian Cancer Cells by Bortezomib and a Novel HDAC6 Inhibitor. Clinical Cancer Research. 2008 Nov;14(22):7340–7.

91. Davoli T, Uno H, Wooten EC, Elledge SJ. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. Science. 2017;355(6322):eaaf8399.

92. Payne JL, Wagner A. The causes of evolvability and their evolution. Nature Reviews Genetics. 2018;20(1):1–15.

93. Chamberlain PP, Hamann LG. Development of targeted protein degradation therapeutics. Nature Chemical Biology. 2019;15(10):937–44.

94. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nature Methods. 2014 Nov;11(11):1114–25.

95. Branca RMM, Orre LM, Johansson HJ, Granholm V, Huss M, Pérez-Bercoff A, et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. Nature Methods. 2014;11(1):59–62.

96. Zhu Y, Orre LM, Johansson HJ, Huss M, Boekel J, Vesterlund M, et al. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nature Communications. 2018;9(1):1–14.

97. Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, Boja ES, et al. How many human proteoforms are there? Nature chemical biology. 2018 Feb;14(3):206–14.

98. Zhang B, Wang J, Wang X, Zhu J, Liu Q, Shi Z, et al. Proteogenomic characterization of human colon and rectal cancer. Nature. 2014;513(7518):382–7.

99. Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, et al. Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. Cell. 2016;166(3):755–65.

100. Mertins P, Mani DR, Ruggles KV, Gillette MA, Clauser KR, Wang P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016;534(7605):55–62.

101. Stewart E, McEvoy J, Wang H, Chen X, Honnell V, Ocarz M, et al. Identification of Therapeutic Targets in Rhabdomyosarcoma through Integrated Genomic, Epigenomic, and Proteomic Analyses. Cancer Cell. 2018 Sep;34(3):411-426.e19.

102. Forget A, Martignetti L, Puget S, Calzone L, Brabetz S, Picard D, et al. Aberrant ERBB4-SRC Signaling as a Hallmark of Group 4 Medulloblastoma Revealed by Integrative Phosphoproteomic Profiling. Cancer Cell. 2018;34(3):379-395.e7.

103. Archer TC, Ehrenberger T, Mundt F, Gold MP, Krug K, Mah CK, et al. Proteomics, Post-translational Modifications, and Integrative Analyses Reveal Molecular Heterogeneity within Medulloblastoma Subgroups. Cancer Cell. 2018;34(3):396-410.e8.

104. Johansson HJ, Socciarelli F, Vacanti NM, Haugen MH, Zhu Y, Siavelis I, et al. Breast cancer quantitative proteome and proteogenomic landscape. Nature Communications. 2019;10(1):1600.

105. Yang M, Vesterlund M, Siavelis I, Moura-Castro LH, Castor A, Fioretos T, et al. Proteogenomics and Hi-C reveal transcriptional dysregulation in high hyperdiploid childhood acute lymphoblastic leukemia. Nature Communications. 2019;10(1):1519.

106. Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, et al. Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. Cell. 2019;177(4):1035-1049.e19.

107. Mun DG, Bhin J, Kim S, Kim H, Jung JH, Jung Y, et al. Proteogenomic Characterization of Human Early-Onset Gastric Cancer. Cancer Cell. 2019;35(1):111-124.e10.

108. Stewart PA, Welsh EA, Slebos RJC, Fang B, Izumi V, Chambers M, et al. Proteogenomic landscape of squamous cell lung cancer. Nature Communications. 2019;10(1):3578.

109. Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, et al. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. Cell. 2019;179(4):964-983.e31.

110. Sinha A, Huang V, Livingstone J, Wang J, Fox NS, Kurganovs N, et al. The Proteogenomic Landscape of Curable Prostate Cancer. Cancer Cell. 2019;35(3):414-427.e6.

111. Gao Q, Zhu H, Dong L, Shi W, Chen R, Song Z, et al. Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. Cell. 2019;179(2):561-577.e22.

112. Ajani J, Zhao S, Wang R, Song S, Kobayashi M, Hao D, et al. Proteogenomic Landscape of Gastric Adenocarcinoma Peritoneal Metastases. Preprint from Research Square. 2020.

113. Krug K, Jaehnig EJ, Satpathy S, Blumenberg L, Karpova A, Anurag M, et al. Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. Cell. 2020 Nov;183(5):1436-1456.e31.

114. Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K, et al. Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. Cell. 2020 Jul;182(1):200-225.e35.

115. Dou Y, Kawaler EA, Zhou DC, Gritsenko MA, Huang C, Blumenberg L, et al. Proteogenomic Characterization of Endometrial Carcinoma. Cell. 2020 Feb;1–47.

116. Chen YJ, Roumeliotis TI, Chang YH, Chen CT, Han CL, Lin MH, et al. Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. Cell. 2020 Jul;182(1):226-244.e17.

117. Xu JY, Zhang C, Wang X, Zhai L, Ma Y, Mao Y, et al. Integrative Proteomic Characterization of Human Lung Adenocarcinoma. Cell. 2020 Jul;182(1):245-261.e17.

118. McDermott JE, Arshad OA, Petyuk VA, Fu Y, Gritsenko MA, Clauss TR, et al. Proteogenomic Characterization of Ovarian HGSC Implicates Mitotic Kinases, Replication Stress in Observed Chromosomal Instability. Cell Reports Medicine. 2020 Aug;1(1):100004.

119. Hu Y, Pan J, Shah P, Ao M, Thomas SN, Liu Y, et al. Integrated Proteomic and Glycoproteomic Characterization of Human High-Grade Serous Ovarian Carcinoma. Cell Reports. 2020 Oct;33(3):108276.

120. Petralia F, Tignor N, Reva B, Koptyra M, Chowdhury S, Rykunov D, et al. Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer. Cell. 2020 Nov;183(7):1962-1985.e31.

121. Li C, Sun YD, Yu GY, Cui JR, Lou Z, Zhang H, et al. Integrated Omics of Metastatic Colorectal Cancer. Cancer Cell. 2020 Nov;38(5):734-747.e9.

122. Betancourt LH, Gil J, Kim Y, Doma V, Çakır U, Sanchez A, et al. The human melanoma proteome atlas—Defining the molecular pathology. Clinical and Translational Medicine. 2021;11(7):e473.

123. Lehtiö J, Arslan T, Siavelis I, Pan Y, Socciarelli F, Berkovska O, et al. Proteogenomics of non-small cell lung cancer reveals molecular subtypes associated with specific therapeutic targets and immune-evasion mechanisms. Nature Cancer. 2021;2(11):1224–42.

124. Yanovich-Arad G, Ofek P, Yeini E, Mardamshina M, Danilevsky A, Shomron N, et al. Proteogenomics of glioblastoma associates molecular patterns with survival. Cell Reports. 2021;34(9):108787.

125. Huang C, Chen L, Savage SR, Eguez RV, Dou Y, Li Y, et al. Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. Cancer cell. 2021;

126. Cao L, Huang C, Zhou DC, Hu Y, Lih TM, Savage SR, et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. Cell. 2021;184(19):5031-5052.e26.

127. Wang LB, Karpova A, Gritsenko MA, Kyle JE, Cao S, Li Y, et al. Proteogenomic and metabolomic characterization of human glioblastoma. Cancer cell. 2021;

128. Liu W, Xie L, He YH, Wu ZY, Liu LX, Bai XF, et al. Large-scale and high-resolution mass spectrometry-based proteomics profiling defines molecular subtypes of esophageal cancer for therapeutic targeting. Nature Communications. 2021;12(1):4961.

129. Li L, Jiang D, Zhang Q, Liu H, Qin Z, Xu F, et al. Integrative proteogenomic characterization of early esophageal cancer. Preprint from Research Square. 2021.

130. Satpathy S, Krug K, Beltran PMJ, Savage SR, Petralia F, Kumar-Sinha C, et al. A proteogenomic portrait of lung squamous cell carcinoma. Cell. 2021;184(16):4348-4371.e40.

131. Qu Y, Feng J, Wu X, Bai L, Xu W, Zhu L, et al. A proteogenomic analysis of clear cell renal cell carcinoma in a Chinese population. Nat Commun. 2022;13(1):2052.

132. Nassiri F, Liu J, Patil V, Mamatjan Y, Wang JZ, Hugh-White R, et al. A clinically applicable integrative molecular classification of meningiomas. Nature. 2021;1–7.

133. Jayavelu AK, Wolf S, Buettner F, Alexe G, Häupl B, Comoglio F, et al. The proteogenomic subtypes of acute myeloid leukemia. Cancer Cell. 2022;

134. Herbst SA, Vesterlund M, Helmboldt AJ, Jafari R, Siavelis I, Stahl M, et al. Proteogenomics refines the molecular classification of chronic lymphocytic leukemia. Nature Communications. 2022;13(1):6226.

135. Anurag M, Jaehnig EJ, Krug K, Lei JT, Bergstrom EJ, Kim BJ, et al. Proteogenomic markers of chemotherapy resistance and response in triple negative breast cancer. Cancer Discovery. 2022;12(11):2586–605.

136. Dong L, Lu D, Chen R, Lin Y, Zhu H, Zhang Z, et al. Proteogenomic characterization identifies clinically relevant subgroups of intrahepatic cholangiocarcinoma. Cancer Cell. 2022;40(1):70-87.e15.

137. Fujita M, Chen MJM, Siwak DR, Sasagawa S, Oosawa-Tatsuguchi A, Arihiro K, et al. Proteo-genomic characterization of virus-associated liver cancers reveals potential subtypes and therapeutic targets. Nature Communications. 2022;13(1):6481.

138. Rodon J, Soria JC, Berger R, Miller WH, Rubin E, Kugel A, et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. Nature medicine. 2019;2:1–19.

139. (OSBREAC) OBCRC, Aure MR, Jernström S, Krohn M, Vollan HKM, Due EU, et al. Integrated analysis reveals microRNA networks coordinately expressed with key proteins in breast cancer. Genome Medicine. 2015;7(1):21.

140. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000 Aug;406(6):747–52.

141. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. Journal of Clinical Oncology. 2009;27(8):1160–7.

142. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486(7403):346–52.

143. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nature Genetics. 2013;45(10):1127–33.

144. Paulsson K, Forestier E, Lilljebjörn H, Heldrup J, Behrendtz M, Young BD, et al. Genetic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(50):21719–24.

145. Paulsson K, Lilljebjörn H, Biloglav A, Olsson L, Rissler M, Castor A, et al. The genomic landscape of high hyperdiploid childhood acute lymphoblastic leukemia. Nature Genetics. 2015 Jun;47(6):672–6.

146. Gridelli C, Rossi A, Carbone DP, Guarize J, Karachaliou N, Mok T, et al. Non-small-cell lung cancer. Nature Reviews Disease Primers. 2015;1(1):15009.

147. Altorki NK, Markowitz GJ, Gao D, Port JL, Saxena A, Stiles B, et al. The lung microenvironment: an important regulator of tumour growth and metastasis. Nature Reviews Cancer. 2018;19(1):1–23.

148. Anagnostou V, Niknafs N, Marrone K, Bruhm DC, White JR, Naidoo J, et al. Multimodal genomic features predict outcome of immune checkpoint blockade in non-small-cell lung cancer. Nature Cancer. 2020;1(1):1–32.

149. Sinha A, Mann M. A beginner's guide to mass spectrometry–based proteomics. The Biochemist. 2020 Sep;42(5):64–9.

150. Ludwig C, Gillet L, Rosenberger G, Amon S, Collins BC, Aebersold R. Data-independent acquisition-based SWATH- MSfor quantitative proteomics: a tutorial. Molecular Systems Biology. 2018;14(8):1–23.

151. Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. Proceedings of the National Academy of Sciences. 2003;100(12):6940–5.

152. Bantscheff M, Lemeer S, Savitski MM, Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Analytical and Bioanalytical Chemistry. 2012;404(4):939–65.

153. Sinitcyn P, Rudolph JD, Cox J. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. Annual Review of Biomedical Data Science. 2018;1(1):207–34.

154. Nesvizhskii AI, Aebersold R. Interpretation of Shotgun Proteomic Data. Molecular & Cellular Proteomics. 2005;4(10):1419–40.

155. Bludau I, Frank M, Dörig C, Cai Y, Heusel M, Rosenberger G, et al. Systematic detection of functional proteoform groups from bottom-up proteomic datasets. Nature Communications. 2021;12(1):3810.

156. Dermit M, Peters-Clarke TM, Shishkova E, Meyer JG. Peptide Correlation Analysis (PeCorA) Reveals Differential Proteoform Regulation. Journal of Proteome Research. 2021;20(4):1972–80.

157. Zhu Y, Orre LM, Johansson HJ, Huss M, Boekel J, Vesterlund M, et al. Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. Nature Communications. 2018;9(1):1–14.

158. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning. 2003;52(1–2):91–118.

159. Lever J, Krzywinski M, Altman N. Principal component analysis. Nature Methods. 2017;14(7):641–2.

160. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences. 2004;101(12):4164–9.

161. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv.org. 2018;stat.ML:arXiv:1802.03426.

162. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008;2008(10):P10008-12.

163. Loo PV, Nordgard SH, Lingjærde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences. 2010;107(39):16910–5.

164. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Sep;490(7418):61–70.

165. Myhre S, Lingjærde OC, Hennessy BT, Aure MR, Carey MS, Alsner J, et al. Influence of DNA copy number and mRNA levels on the expression of breast cancer related proteins. Mol Oncol. 2013 Jun;7(3):704–18.

166. Mohanty V, Wang F, Mills GB, Network CR, Chen K. Uncoupling of gene expression from copy number presents therapeutic opportunities in aneuploid cancers. Cell Reports Medicine. 2021;2(7):100349.

167. Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, Possemato A, et al. Systematic and Quantitative Assessment of the Ubiquitin-Modified Proteome. Molecular cell. 2011 Oct;44(2):325–40.

168. Koussounadis A, Langdon SP, Um IH, Harrison DJ, Smith VA. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. Scientific Reports. 2015 Jun;5(1):10775.

169. Yarchoan M, Johnson BA, Lutz ER, Laheru DA, Jaffee EM. Targeting neoantigens to augment antitumour immunity. Nature Reviews Cancer. 2017 Apr;17(4):209–22.

170. Jones PA, Ohtani H, Chakravarthy A, Carvalho DDD. Epigenetic therapy in immune-oncology. Nature Reviews Cancer. 2019 Feb;19(3):1–11.

171. Forshed J, Johansson HJ, Pernemalm M, Branca RMM, Sandberg A, Lehtiö J. Enhanced Information Output From Shotgun Proteomics Data by Protein Quantification and Peptide Quality Control (PQPQ)*. Molecular & Cellular Proteomics. 2011;10(10).

172. Shen S, Park JW, Lu Z xiang, Lin L, Henry MD, Wu YN, et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences. 2014;111(51).

173. Smith LM, Agar JN, Chamot-Rooke J, Danis PO, Ge Y, Loo JA, et al. The Human Proteoform Project: Defining the human proteome. Science advances. 2021;7(46):eabk0734.

174. Bogdanow B, Zauber H, Selbach M. Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides*. Molecular & Cellular Proteomics. 2016;15(8):2791–801.

175. Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO. Alternative splicing and the evolution of phenotypic novelty. Philosophical Transactions of the Royal Society B: Biological Sciences. 2017;372(1713):20150474.

176. Kim GB, Fritsche J, Bunk S, Mahr A, Unverdorben F, Tosh K, et al. Quantitative immunopeptidomics reveals a tumor stroma–specific target for T cell therapy. Science Translational Medicine. 2022;14(660):eabo6135.

177. The M, Tasnim A, Käll L. How to talk about protein-level false discovery rates in shotgun proteomics. PROTEOMICS. 2016;16(18):2461–9.

178. Dupree EJ, Jayathirtha M, Yorkey H, Mihasan M, Petre BA, Darie CC. A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. Proteomes. 2020;8(3).

179. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558(7708):1–24.

180. Specht H, Emmott E, Petelski AA, Huffman RG, Perlman DH, Serra M, et al. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. Genome Biology. 2021;22(1):50.

181. Burnum-Johnson KE, Conrads TP, Drake RR, Herr AE, Iyengar R, Kelly RT, et al. New Views of Old Proteins: Clarifying the Enigmatic Proteome. Molecular & Cellular Proteomics. 2022;21(7):100254.

182. Taylor AM, Shih J, Zhang X, Schumacher SE, Wang C, Liu J, et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell. 2018 Mar;33(4):1–36.

183. Ritchie MD, Holzinger ER, Li R. Methods of integrating data to uncover genotype-phenotype interactions. Nature Reviews Genetics 2015;16(2):85–97.

184. Vitrinel B, Koh HWL, Kar FM, Maity S, Rendleman J, Choi H, et al. Exploiting Interdata Relationships in Next-generation Proteomics Analysis. Molecular & Cellular Proteomics. 2019;18(8 suppl 1):S5–14.

185. Li MM, Huang K, Zitnik M. Graph representation learning in biomedicine and healthcare. Nature Biomedical Engineering. 2022;1–17.

186. Palmblad M, Böcker S, Degroeve S, Kohlbacher O, Käll L, Noble WS, et al. Interpretation of the DOME Recommendations for Machine Learning in Proteomics and Metabolomics. J Proteome Res. 2022;21(4):1204–7.

187. Zhang B, Whiteaker JR, Hoofnagle AN, Baird GS, Rodland KD, Paulovich AG. Clinical potential of mass spectrometry-based proteogenomics. Nature Reviews Clinical Oncology. 2018 Nov;16(4):1–13.

188. Tamborero D, Dienstmann R, Rachid MH, Boekel J, Lopez-Fernandez A, Jonsson M, et al. The Molecular Tumor Board Portal supports clinical decisions and automated reporting for precision oncology. Nature Cancer. 2022;3(2):251–61.

189. Wen B, Li K, Zhang Y, Zhang B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. Nature Communications. 2020;11(1):1–14.

190. Hiam-Galvez KJ, Allen BM, Spitzer MH. Systemic immunity in cancer. Nature Reviews Cancer. 2021;1–15.

191. Nakayasu ES, Gritsenko M, Piehowski PD, Gao Y, Orton DJ, Schepmoes AA, et al. Tutorial: best practices and considerations for mass-spectrometry-based protein biomarker discovery and validation. Nature Protocols. 2021;16(8):3737–60.